Network-Based Distributional Semantic Models

Elias Iosif

Department of Electronic and Computer Engineering

Technical University of Crete

A thesis submitted
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

2013

To Elena

Acknowledgements

First, I would like to express my sincere gratitude to my advisor Prof. Alexandros Potamianos for his continuous support and guidance in this exciting research area. His scientific and mentoring brilliance were offered to me with true generosity: countless hours of personal discussions and brainstorming in combination with inspiring group meetings. The research efforts of this thesis would not be possible without his significant contribution.

In addition, I would like to thank all the committee members for participating in the review and examination of this thesis. More specifically, I would like to thank Prof. Euripides Petrakis for the many fruitful discussions we had during the years of my graduate studies. Also, I thank the Head of the Software and Knowledge Engineering Laboratory, NCSR “Demokritos” Dr Vangelis Karkaletis for giving me the opportunity to work with his group in an excellent research environment. I wish to thank Prof. Marco Baroni, Center for Mind/Brain Sciences (CIMeC) of the University of Trento, for his hospitality during my short visits at CIMeC and his constructive comments on my work. I also thank Prof. Michail Lagoudakis for the encouragement to apply my research interests in the context of the graduate courses he taught. Last but not least, I feel that a significant part of my personality was shaped during the long period I spent at the ECE Department of the Technical University of Crete: I respectfully thank all my teachers. I gratefully acknowledge the financial support I received from the PortDial project (“Language Resources for Portable Multilingual Spoken Dialog Systems") supported by the EU Seventh Framework Programme (FP7), grant number 296170, and and the Basic Research Programme, Technical University of Crete, Project Number 99637: “Unsupervised Semantic Relationship Acquisition by Humans and Machines: Application to Automatic Ontology Creation”.

Friendship and research can have a perfect match: many thanks to Maria Giannoudaki, Kalliopi Zervanou, Vassiliki Kouloumenta, Nikolaos Malandrakis, Theodosis Moschopoulos, Pavlos Papadopoulos and Orfeas Tsergoulas.

During my journey to knowledge I was not alone. I respectfully thank my grandparents Dimitris & Ioanna and my mother Maria, my brother Michalis and my little sisters Konstantina & Ioanna for their endless love. This journey always had the memory of my father as a guardian angel. Now, the journey continues with a new member on board, my wife Elena, to whom this thesis is dedicated.

Abstract

In this thesis, the unsupervised creation of language-agnostic Distributional Semantic Models (DSMs) using web harvested data is investigated for the problem of semantic similarity estimation. Semantic similarity can be regarded as the building block for numerous tasks of Natural Language Processing, e.g., affective text analysis and paraphrasing. The first part of the thesis deals with the construction of typical DSMs following the well-established Vector Space Model. More specifically, corpora are created by harvesting web documents following a query-based approach. Two families of similarity metrics are applied, while related parameters are investigated. Similarity metrics are evaluated against human similarity ratings achieving state-of-the-art results that are comparable with knowledge-based metrics. Despite its good performance, the aforementioned methodology suffers from quadratic query complexity with respect to the size of the lexicon. A methodology of linear query complexity is proposed, which is applied for corpus creation with respect to a lexicon consisting of thousands of nouns. Using this corpus, we propose a novel network-based implementation of DSMs, which is based on the notion of semantic neighborhoods. Semantic neighborhoods are considered as a parsimonious representation of corpus statistics, while they capture two main types of lexical relations: semantic and associative. The problem of the automatic classification of associative and semantic relations is also addressed, motivated by findings from the literature of psycholinguistics and corpus linguistics. Moreover, three novel neighborhood-based similarity metrics are proposed, motivated by the hypotheses of attributional and maximum sense similarity. The proposed metrics are shown to outperform the baseline approaches for the task of semantic similarity estimation between words. Inspired by evidence for cognitive organization of concepts, based on the degree of concreteness, we further investigate the performance and organization of network DSMs for abstract vs. concrete nouns. Finally, the framework of network DSMs is extended for the creation of multimodal networks using textual and visual features, and the estimation of semantic similarity beyond word level (noun compounds). Very good results are achieved for both extensions, showing the flexibility of the network-based framework.

Περίληψη

Η παρούσα διατριβή πραγματεύεται την κατασκευή κατανεμημένων σημασιολογικών μοντέλων (Distributional Semantic Models - DSMs) χρησιμοποιώντας κειμενικά δεδομένα που έχουν συλλεγεί από τον παγκόσμιο ιστό. Μερικά από τα κύρια και πιο ενδιαφέροντα χαρακτηριστικά της κατασκευής των μοντέλων αυτών είναι η μη χρήση τεχνικών επίβλεψης (unsupervised) και η μη εξάρτηση σε γλωσσολογικά χαρακτηριστικά, γεγονός που τα καθιστά -από πλευράς υλοποίησης- ανεξάρτητα από τη φυσική γλώσσα ως προς την οποία εφαρμόζονται (language-agnostic). Η κύρια εφαρμογή των ανωτέρω μοντέλων αφορά στην εκτίμηση της σημασιολογικής ομοιότητας (semantic similarity).

Η συμβολή της σημασιολογικής ομοιότητας είναι ιδιαιτέρως σημαντική για ένα πλήθος εφαρμογών του τομέα της Επεργασίας Φυσικού Λόγου. Παραδείγματα τέτοιων εφαρμογών περιλαμβάνουν την ανάλυση του συναισθηματικού περιεχομένου κειμενικών δεδομένων και τεχνικές παράφρασης. Το πρώτο πειραματικό μέρος της διατριβής αφορά στην κατασκευή σύνηθων κατανεμημένων σημασιολογικών μοντέλων σύμφωνα με το καθιερωμένο Vector Space Model.

Μία από τις κύριες κατευθύνσεις αυτής της προσπάθειας είναι η δημιουργία σωμάτων κειμένων (corpora) μέσω της ανάκτησης εγγράφων του παγκοσμίου ιστού αποστέλνωντας επερωτήσεις (queries) προς μηχανές αναζήτησης. Επιπλέον, μελετώνται δύο βασικοί τύποι μετρικών σημασιολογικής ομοιότητας σε συνάρτηση με ένα πλήθος παραμέτων. Οι χρησιμοποιούμενες μετρικές αποτιμώνται ως προς τη συσχέτισή τους με βαθμολογίες σημασιολογικής ομοιότητας που έχουν ληφθεί από ανθρώπους. Η επίδοσή τους παρατηρήθηκε να είναι συγκρίσιμη με εκείνη που επιτυγχάνουν οι τρέχουσες τεχνολογίες αιχμής, καθώς και με την επίδοση ενός άλλου τύπου μετρικών που βασίζεται στην άντληση πληροφορίας από πηγές γνώσης (knowledge-based metrics). Παρά την αξιόλογη επίδοσή της, η πιο πάνω μεθοδολογία κρίνεται πρακτικώς δύσχρηστη αναφορικά με τον υπολογισμό της σημασιολογικής ομοιότητας μεταξύ όλων των ζευγών λέξεων οι οποίες δύνανται να περιέχονται σε ένα λεξικό. Το μειονέκτημα τούτο οφείλεται στην τετραγωνική πολυπλοκότητα της δημιουργίας επερωτήσεων ως προς το μέγεθος του χρησιμοποιούμενου λεξικού. Στο δεύτερο πειραματικό μέρος της εργασίας, προτείνεται μία μεθοδολογία για την αντιμετώπιση του προαναφερθέντος μειονεκτήματος, σύμφωνα με την οποία η δημιουργία επερωτήσεων υπέχει γραμμική πολυπλοκότητα σε σχέση με το λεξικό αναφοράς. Η προτεινόμενη μεθολογία εφαρμόζεται για τη δημιουργία ενός

σώματος κειμένου από δεδομένα του παγκόσμιου ιστού ως προς ένα λεξικό το οποίο αποτελείται από χιλιάδες ουσιαστικών. Χρησιμοποιώντας το πιο πάνω σώμα κειμένου, μια νέα, βασισμένη σε δίκτυα, υλοποίηση των κατανεμημένων σημασιολογικών μοντέλων προτείνεται, κεντρική ιδέα της οποίας είναι οι σημασιολογικές γειτονιές (semantic neighborhoods). Οι σημασιολογικές γειτονιές μπορούν να θεωρηθούν ως μια φειδωλή, αλλά συνάμα περιεκτική, αναπαράσταση της λεκτικής στατιστικής πληροφορίας που εμπεριέχεται στο σώμα κειμένου. Επιπλέον, δυο βασικοί τύποι λεξιλογικών σχέσεων ενυπάρχουν στις γειτονιές αυτές: σημασιολογικές και συσχετιστικές (associative). Η αυτόματη κατηγοριοποίηση των βασικών αυτών σχέσεων διερευνάται, σύμφωνα με κάποια ευρήματα της βιβλιογραφίας της ψυχογλωσσολογίας (psycholinguistics) και της εφαρμοσμένης σε σώματα κειμένων γλωσσολογίας (corpus linguistics). Επιπρόσθετα, τρεις νέες μετρικές σημασιολογικής ομοιότητας βασισμένες σε δίκτυα προτείνονται, έχοντας ως θεωρητικό υπόβαθρο τις υποθέσεις αναφορικά με την ομοιότητα χαρακτηριστικών (attributional similarity) και τη μέγιστη εννοιολογική ομοιότητα (maximum sense similarity). Η επίδοση των προτεινόμενων σημασιολογικών μετρικών παρατηρείται να υπερβαίνει εκείνη των βασικών (baseline) μετρικών ως προς την εκτίμηση της ομοιότητας μεταξύ λέξεων. Η προτεινόμενη υλοποίηση των κατανεμημένων σημασιολογικών μοντέλων, καθώς και οι αντίστοιχες μετρικές ομοιότητας, διερευνώνται περαιτέρω ως προς δύο τύπους ουσιαστικών, η διάκριση των οποίων προέρχεται από το πεδίο της γνωσιακής επιστήμης: αφηρημένα (abstract) και συμπαγή (concrete). Το κύριο έναυσμα για τη διάκριση αυτή αποτελούν οι ενδείξεις σχετικά με τη διαφοροποιημένη οργάνωση στο ανθρώπινο γνωσιακό σύστημα των ανωτέρω τύπων βάσει του βαθμού σημασιολογικής συμπάγειας. Τέλος, τα προτεινόμενα κατανεμημένα σημασιολογικα μοντέλα και οι μετρικές ομοιότητας κατασκευάζονται και αποτιμώνται σε κάποιες περαιτέρω εφαρμογές. Πιο συγκεκριμένα, τα αποκλειστικώς βασισμένα σε κειμενικά δεδομένα μοντέλα, επεκτείνονται σε πολυτροπικά (multimodal) χρησιμοποιώντας κειμενικά και οπτικά (visual) χαρακτηριστικά. Επιπλέον, μελετάται η επέκταση των προτεινόμενων μοντέλων με στοχό την αναπαράσταση των σημασιολογικών γειτονιών πολυλεκτικών όρων αποτελούμενων από ουσιαστικά, καθώς και η εκτίμηση της σημασιολογικής ομοιότητας αυτών. Πολύ καλά αποτελέσματα επιτυγχάνονται για τις ανωτέρω εφαρμογές καταδεικνύοντας την προσαρμοστικότητα των προτεινόμενων μοντέλων.

Contents

List of Figures

Chapter 1

Introduction

1.1 Corpus-based Lexical Semantics

A broad definition of *meaning* according to the Webster's dictionary is "what is intended to be understood, signified, indicated, etc" Guralnik [1976]. The vital role of natural languages in human communication is itself a strong proof that the meaning of conceptual (mental) entities, e.g., ideas or opinions, can be conveyed through appropriate lexicalizations, i.e., use of words. In principle, even a single word provided in isolation, i.e., out of context, can carry a certain amount of semantic content. Briefly, this thesis deals with corpus-based computational models built upon the foundations of *lexical semantics*, which is a discipline of linguistics referring to the meaning of words. The idea that the semantic properties of words can be revealed through the context in which they exist (or could exist) is supported by the linguistic theory of *contextual approach* Cruse [1986]; Haas [1962, 1964]. The notion of *context* refers to linguistic, as well as to extra-linguistic context. Without ignoring the effect of extra-linguistic factor, the linguistic context can be regarded as a sufficient carrier of the semantics of words for three reasons: (i) very often the relations between words and extra-linguistic contexts are formulated within linguistic context, e.g., "There are many books here", (ii) potentially any extra-linguistic context can be lexicalized, and (iii) linguistic context can be utilized more easily Cruse [1986]. The theoretic foundations of this thesis rely on the key idea of the contextual approach: the meaning of a word can be reflected (at some extend) with regard to its linguistic environment. This is summarized by the famous statement "You shall know a word by the company it keeps" Firth [1957].

According to the contextual approach words should be considered with respect to other words with which they co-occur within a specified context. Sentences, paragraphs, and documents are examples of such contextual types. Word co-occurrence can be divided into two

main types: (i) positional, and (ii) relational Evert [2005]. The positional type refer to the co-occurrence of words by considering their proximity within context. Structural relations, e.g., grammatical dependencies, form the basis for the definition of relational co-occurrences. The consideration of grammatical relations is justified by the ultimate goal of grammar: to convey the intended meaning Cruse [1986]. Early research efforts were focused on positional co-occurrence due to the lack of computational tools needed for the extraction of relation co-occurrences Stevens et al. [1965]. Nowadays, both positional (e.g., Agirre et al. [2009]) and relational types (e.g., Baroni and Lenci [2010]) are incorporated in computational models. However, the former type constitutes a language-agnostic paradigm of text processing, which is directly applicable for the case of under-resourced languages.

Within the contextual framework the relations between words can be broadly distinguished into two categories under the perspective of structuralism Harris. [2001] point of view: *syntagmatic* and *paradigmatic* Cruse [1986]; Sahlgren [2006]. Syntagmatic relations refer to words that co-occur within the same context. Paradigmatic relations concern words that exist in the same context but without co-occurring. The latter is a type of substitutional relation suggesting that if two words are paradigmatically related then the one may substitute the other without altering the meaning of the context. This is a rather technical distinction that does not encode the semantic relations between words, i.e., lexical semantics. Lexical semantics is a well-investigated area for linguistics including relations ranging from (several variants of) synonymy and antonymy to more complex hierarchical configurations, such as hypernymy/hyponymy and meronymy. In principle, the enumeration of semantic relations between words seems to be an extremely difficult task. For example, relations such as “Cause-Effect”, “Instrument-Agency” may be of great importance for certain domains or disciplines, e.g., cognitive sciences. Such relations can be regarded as relations with strong *associative* characteristics McNamara [2005], even if a formal linguistic definition is not available.

The computational models presented in this work are focused on the estimation of word *semantic similarity*, i.e., how much similar the meaning of two words is. The notion of semantic similarity is built upon the existence of semantic relatedness. In other words, we aim to measure the strength of relatedness that hold between words on the basis of their similarity in meaning. This relatedness pertains a diverse range of lexical relations, e.g., two words may be regarded at some extent as semantically similar even if they are not synonyms. However, it should be stressed out that the estimation of semantic similarity does not necessarily address the recognition of the underlying types of semantic relations.

The computational models that adopt the aforementioned contextual approach are refer to as *Distributional Semantic Models (DSMs)* Baroni and Lenci [2010]. In the first paragraph of this section we used a dictionary entry in order to define the meaning of “meaning”. Such

dictionary definitions can be used for the representation of word semantics, however, the used language is of rather technical style. (Moreover, the exploitation of dictionaries does not constitute a generic computational framework, while a number of drawbacks exist, e.g., development cost, coverage limitations.) A more generic approach is the use of real-life examples of language usage, i.e., corpora. Of course, an "adequate" number of such examples is required in order to have a "sufficient" amount of contextual evidence. Although a number of corpora is available, this is not the case for the less-resourced languages. In this work, we address the task of corpora creation using simple techniques for harvesting web data. The world wide web covers a plethora of domains, authoring styles and languages, and is fertile ground for data harvesting.

Up to this point we have described the half of the machinery needed for estimating word similarity based on lexical semantics: (i) the theoretic framework, i.e., the contextual approach, and (ii) the required resources from which the relevant (contextual) features can be extracted. The missing tools include: (i) a formal scheme for the representation of contextual features, (ii) appropriate measurements of semantic similarity that can be defined with respect to the representation scheme. The Vector Space Model (VSM) Turney and Pantel [2010] is the most widely-used implementation of DSMs. VSM can be regarded as a high-dimensional space, which is typically formulated as a matrix. The (distributed) semantics of each target word are encoded by a vector (matrix row) that contains the words (matrix columns) that co-occur with it within a specified context. This is a spatial representation enabling the definition of similarity in terms of proximity. The theoretical argument of similarity-as-proximity was articulated by the *geometric metaphor of meaning* Lackoff and Johnson [1997]: semantically similar words are supposed to live "near" each other within a space (not necessarily VSM). Similarity measurements are missioned with the quantification of the "nearness" notion. Both syntagmatic and paradigmatic relations may be incorporated in such measurements. From the syntagmatic perspective, the similarity between words is based on their direct co-occurrences. Following the paradigmatic consideration, word similarity can be estimated as a function of contextual commonality. This is the well-established *distributional hypothesis of meaning* according to which similarity of meaning is implied by the contextual similarity Harris [1954]. Of course, both syntagmatic and paradigmatic relations may be integrated into a single similarity measurement. The proposal of a network-based implementation of DSMs upon which a number of novel similarity measurements are defined is among the main research efforts presented in this thesis.

The discipline of cognitive semantics investigates fundamental questions about the underlying structures and mechanisms that drive semantic tasks such as the attributions of features (properties) to objects. A long list of related issues include the acquisition, representation,

organization and retrieval of the relevant information needed for the completion of semantic tasks. Clearly, it is hard to have a principled mapping between the models employed in cognitive semantics and computational linguistics. But, seems intuitive to claim (even an unjustified) connection: words correspond to cognitive objects, i.e., concepts. Broadly speaking, the DSMs themselves constitute a representation for cognitive semantics built upon linguistic evidence. Also, a great number of problems addressed within the framework of DMSs are characterized by cognitive aspects, e.g., estimation of word semantic similarity. Thus, we believe that the literature of cognitive semantics can serve as a valuable source for inspiration for the community of computational linguistics.

1.2 Measuring Similarity

In this section, a number of issues regarding the definition of similarity are briefly presented. These are given by the perspectives of cognitive psychology and information theory, while the respective similarity measures are defined. Also, it is discussed why such measures can not be regarded as true metrics for the case of word semantic similarity.

1.2.1 Features of Similarity: The Tverskian Contrast Model

The work of Amos Tversky is acknowledged as a pioneering one in the field of cognitive and mathematical psychology Tversky [2003]. In his seminal paper "Features of Similarity" Tversky [1977] the notion of similarity is considered as the organizing principle that drives cognitive tasks such as concept formulation/classification and generalization. Several empirical formulations of similarity were observed including pair rating, object sorting, substitution errors, correlated occurrences. The motivation of this work was the proposal of an alternative theory regarding similarity meant to overcome the unjustified assumptions (with respect to empirical evidence from the psychological studies) of the dominating geometric models. As it was mentioned, objects are represented as points in a dimensional space, while their similarity is expressed in terms of proximity using a distance measurement. The core of the Tverskian theoretical approach of similarity is the matching between features, which is also known as *contrast model*. More specifically, this model is based on the contribution of common and distinctive features. In addition, asymmetries regarding distinctive features are also modeled. In general, the term feature refers to the value of a variable of the following types: binary, nominal, ordinal and cardinal. A generic measurement for estimating the similarity between

two objects [1] A and B was defined as follows:

$$\mathcal{S}_T(A,B) = \frac{f_T(F_A \cap F_B)}{f_T(F_A \cap F_B) + \lambda_1 f_T(F_A - F_B) + \lambda_2 f_T(F_B - F_A)}, \tag{1.1}$$

for $\lambda_1, \lambda_2 \geq 0$. In (1.1), F_A and F_B represent the set of features of A and B, respectively. Set $(F_A \cap F_B)$ is the intersection of F_A and F_B, representing the features that are shared (common) between w_i and w_j. Sets $(F_A - F_B)$ and $(F_B - F_A)$ include the features that are different (distinctive) between F_A and F_B, The $f_T(.)$ notation stands for any function that operates over a set. Constants λ_1 and λ_2 weight the contribution of sets $(F_A - F_B)$ and $(F_B - F_A)$, respectively, to the similarity computation.

The contrast model implies asymmetry in similarity, i.e., $\mathcal{S}_T(A,B) \neq \mathcal{S}_T(B,A)$. This hypothesis was confirmed by experimental findings based on judgmental (e.g., rating) and behavioral tasks (e.g., choice). Moreover, Tversky investigated the role of common and the distinctive features for different tasks. Regarding judgmental tasks it was found that common features are weighted more during similarity judgment than difference judgment.

1.2.2 An Info-Theoretic Definition

In Lin [1998], an information-theoretic definition of similarity measures was provided, which shares a number of characteristics with the theoretical contrast model of Tversky Tversky [1977]. Here, the main point of this work are briefly presented. The motivation for this work was that some weak aspects of well-established similarity measurements. For example, the definition for a number of measure types is domain-specific such as network measures that require an appropriate structure. In addition, it was observed that often the underlying assumptions were not clearly mentioned. Lin's work is based on the following two generic characteristics regarding the definition of similarity measures:

1. **Universality.** Similarity is defined by the perspective of information theory. Such definition is expected to apply in any domain/task that can be probabilistically modeled.

2. **Theoretical justification.** The definition of similarity should originate from a set of appropriate assumptions. This foundation is expected to drive the derivation of the respective formula(s).

Moreover, a number of intuitions were proposed regarding the similarity between two "objects" A and B, $\mathcal{S}_L(A,B)$, as follows:

[1] Here, the notion of "object" is used with a broad sense. Examples of such objects are words, images, etc.

1. $\mathcal{S}_L(A, B)$ is related with the commonality of A and B: more commonality indicates more similarity.

2. $\mathcal{S}_L(A, B)$ is related with the differences of A and B: more differences indicate less similarity.

3. Regardless of the commonality, the maximum value of $\mathcal{S}_L(A, B)$ is obtained only if A and B are identical.

The following assumptions were made

1. The commonality between A and B is denoted as $\mathcal{I}(\mathcal{C}(A, B))$, where $\mathcal{C}(A, B)$ stands as a proposition about the commonality of A and B, while $\mathcal{I}$ denotes the information of a proposition. The information of a proposition can be computed by taking the negative logarithm of the proposition's probability, e.g,, $\mathcal{I}(\ \mathcal{C}(A, B)) = -\log p(\ \mathcal{C}\ (A, B))$.

2. The difference between A and B is defined as $\mathcal{I}(\mathcal{D}(A, B)) - \mathcal{I}(\mathcal{C}(A, B))$, where $\mathcal{I}(\mathcal{D}(A, B))$ stands for a proposition describing A and B.

3. The similarity between A and B is a function f_L of their respective commonalities and differences: $\mathcal{S}_L(A, B) = f_L(\ \mathcal{I}(\mathcal{C}(A, B)), \mathcal{I}(\mathcal{D}\ (A, B)))$.

4. $\mathcal{S}_L(A, B) = 1$ if A and B are identical.

The following theorem was reached by Lin's analysis.

Theorem 1 *The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.*

$$\mathcal{S}_L(A, B) = \frac{\log p(\mathcal{C}(A, B))}{\log p(\mathcal{D}(A, B))} \tag{1.2}$$

An interpretation of this theorem states that knowing the commonality of A and B, $\mathcal{S}_L(A, B)$ corresponds to the amount of the extra information needed for the determination of those objects. The generic definition of similarity proposed in (1.2) was applied on a number of tasks including similarity computation using ordinal values, string similarity based on character insertions/deletions, and estimation of word similarity exploiting contextual features (by adopting the distributional hypothesis of meaning) or taxonomic characteristics (use of the WordNet hierarchy) Lin [1998].

1.2.3 Measurability Without Metricity

In essence, the required elements for the computation of similarity between two objects is a set of appropriate features and a similarity measure. Two schemes are widely-used for the transformation of a similarity measure $\mathcal{S}$ into a dissimilarity (or distance) measure $\mathcal{D}$: (i) $\mathcal{D} = 1 - \mathcal{S}$ if the upper bound of $\mathcal{S}$ equal to 1, and (ii) $\mathcal{D} = \frac{1}{1-\mathcal{S}}$ Lin [1998]; Sahlgren [2006]. This is a rather technical manipulation under the (cognitive) assumption that similarity and dissimilarity are linearly related. These elements resemblances the definition of a metric space Pekalska and Duin [2005]: a pair (F, D), where X is a set and D is a distance function (metric) $D : F \times F \rightarrow \mathbb{R}^0_+$. For all $x, y, z \in F$ the following are satisfied:

1. **Reflexivity:** $D(x, x) = 0$.
2. **Symmetry:** $D(x, y) = D(y, x)$.
3. **Definiteness:** $D(x, y) = 0 \Rightarrow x = y$.
4. **Triangle inequality:** $D(x, y) \leq D(x, z) + D(y, z)$.

DSMs do not constitute true metric spaces, since the triangle inequality is violated, e.g., $D(\text{“tree”}, \text{“forest”}) \not\leq D(\text{“tree”}, \text{“flower”}) + D(\text{“forest”}, \text{“flower”})$, where D is a dissimilarity measure. Strictly speaking, any similarity/dissimilarity function defined over DSMs can not be regarded as a true similarity/dissimilarity metric; “measure” is a more precise term rather than “metric”. However, it has been a commonplace in the literature of DSMs (including the present thesis included) the interchangeable use of both terms .

1.3 Applications

Semantic similarity computation between words is closely related the problem of word sense disambiguation (WSD) Agirre and Edmonds [2007]. WSD methods can be divided into two main categories: (i) supervised approaches that apply machine learning for learning sense labels for a set of words with respect to a given context (sense labeling), and (ii) unsupervised approaches that automatically discriminate (discover) word senses without label assignment. For both categories the key criterion is the semantic similarity between the target word and the candidate senses. The similarity at the word level is among the essential features for computing semantic textual similarity (STS), i.e., similarity between larger segments of text such as sentences. STS was investigated at various levels: lexical Androutsopoulos and Malakasiotis [2010], syntactic Zanzotto et al. [2009], and semantic Bos and Markert [2005]; Rinaldi et al. [2003]. Machine translation evaluation metrics were also applied for similarity estimation at

the lexical level similarity Finch et al. [2005]; Perez and Alfonseca [2005] including BLEU Papineni et al. [2002] that is based on word n-gram overlap. Recently, the task of sentence similarity estimation has attracted the great interest of the research community as shown by the participation in the respective task of the SemEval 2012 Agirre et al. [2012]. The top performing systems utilized numerous types of features and similarity metrics in combination with domain adaptation techniques. The success of those systems can be mainly attributed to the efficient incorporation of machine learning, while many questions remain open regarding the underlying models of compositional meaning. STS is closely related to the problems of paraphrasing, which is bidirectional and based on semantic equivalence Madnani and Dorr [2010] and textual entailment, which is directional and based on relations between semantics Dagan et al. [2006]. There is a variety of applications for semantic similarity, both at word and sentence, including information extraction Szpektor and Dagan [2008], question answering Harabagiu and Hickl [2006], machine translation Mirkin et al. [2009].

The analysis of affective text, i.e., analysis of emotional context, is a recent research area that pertains several applications of Natural Language Processing, e.g., opinion mining and sentiment analysis Balog et al. [2006]; Hu and Liu [2004]. The assignment of affective scores to words constitutes the building block for affective text analysis. For a given set of words, semantic and affective similarity are related under the hypothesis that "semantic similarity can be translated to affective similarity" Malandrakis et al. [2013]. In Turney and Littman [2002], the affective score of a new word was estimated using a fixed set of words (also known as "seeds") for which their affective scores, as well as their respective semantic similarities were known. In particular, the affective score for a new word is computed by algebraic combinations of the similarities and ratings of seed words. Unlike the utilization of a fixed set of seeds Turney and Littman [2002], the automatic selection of seeds was investigated in Malandrakis et al. [2011] in combination with several kernels, i.e, functions for controlling the contribution of semantic similarity. Handle multi-word terms. In Malandrakis et al. [2013], the problem of sentence-level affective rating was investigated using a hierarchical (use of n-grams) compositional framework in which multiword terms were also considered in order to capture non-compositional semantics.

Semantic similarity has been also employed in the field of spoken dialogue systems (SDS). Grammar induction depends on the availability of semantic classes that correspond to the domain concepts. The basis for the creation of such classes is the semantic similarity between the candidate terminals, following an agglomerative algorithm Meng and Siu [2002]; Pargellis et al. [2004]. Various measurements of similarity have been compared in Pargellis et al. [2001, 2004]. Variations of the aforementioned algorithm include combination of similarity metrics Iosif et al. [2006] and soft-clustering Iosif and Potamianos [2007b]. Word similarity has been

also employed in class-based n-gram language modeling Brown et al. [1992]. In Niesler et al. [1998], various class-based n-gram language models were interpolated with word-based models. Classes included clusters of words of same to part-of-speech, as well as semantically similar words. The interpolated models obtained higher performance in terms of perplexity and word error rate.

The last decade similarity-based approaches are combined with data mined from the world wide web. The use of web as a corpus appears to be a working solution to the data sparseness problem In Cimiano et al. [2004], linguistic patterns were employed for the identification of ontological relations. The goal was the extraction of relevant instances for certain concepts of a given the domain ontology. Relatedness measures were applied using web-based statistics. For example, "South Africa" was found to be an appropriate instance for concept "country" rather than "hotel" In Moschopoulos et al. [2013], the relatedness between actors in policy networks were estimated using a variety of features including web page counts (number of hits), outlinks, and lexical information extracted from web documents or web snippets. The features were evaluated for both positive and negative (antagonistic) actor relations. The web has been also exploited for a variety of other applications, such as social networks extraction Jin et al. [2007], collaborative filtering Mobasher et al. [2007], sentiment analysis Godbole et al. [2007], music genre classification Schedl et al. [2006].

1.4 Contributions

The first part of the present thesis deals with a web-based methodology for the estimation of semantic similarity between words and biomedical terms. One of the main characteristics of this approach is that is fully unsupervised, i.e, no knowledge-resources are required. The web covers a plethora of domains, authoring styles and languages, and is considered as fertile ground for automatic semantic knowledge acquisition. Web data are accessed via the submission of appropriate queries to web search engines. More specifically, two different types of web data are investigated for the estimation of semantic similarity. First, the number of hits returned by web search engines are utilized as statistics of word co-occurrence. A number of well-established co-occurrence-based metrics are applied and compared for the estimation of word semantic similarity. The second data type deals with web documents, which are downloaded for corpus creation. This is conditioned on word-pairs, explicitly requesting the co-occurrence of word-pairs in the same document through the use of conjunctive AND queries. This corpus is exploited for the constructing a typical DSM based on the distributional hypothesis of meaning. In addition, a number of parameters are investigated including the window size applied for the extraction of contextual features and various schemes for weighting those features.

Overall, the performance of this methodology is shown to be comparable to that of supervised resource-based algorithms.

The core of the thesis concerns an efficient and scalable methodology for corpora creation from web data in combination with a novel network-based (also fully unsupervised) implementation of DSMs. Despite the success of the aforementioned methodology a limitation regarding scalability is introduced by the utilization of AND queries: given a lexicon of size N, the required query complexity for corpus creation is quadratic $\mathcal{O}(N^2)$. In order to tackle this limitation, a method of linear query complexity with respect to N is proposed for corpus creation. The key idea is the employment thousands of individual queries and the aggregation of the harvested data. This strategy smooths the domination of very frequent words, while enables the better representation of rare words within the corpus. Next, a semantic network is created encoding the relevant corpus statistics. This builds upon the formulation of semantic neighborhoods, which capture diverse information at the syntactic, semantic and pragmatic level. Motivated by maximum sense and attributional similarity three novel network-based similarity metrics are proposed. Combinations of co-occurrence-based and contextual metrics are investigated for the computation of semantic neighborhoods and the related network similarity metrics. The performance of the main network metrics is also investigated for the case of abstract and concrete nouns for both English and Greek. Moreover, the proposed network-based DSMs is extended towards: (i) the creation of multimodal networks through the integration of visual and textual features, and (ii) the estimation of semantic similarity between compositional expressions.

Beside the estimation of semantic similarity that pertains a wide range of lexical relations, we deal with the automatic discrimination of two fundamental types of (lexical) relations, namely, associative and semantic. These relation types play an important role for the disciplines of lexical semantics and psycholinguistics. Three different types of discriminative features extracted from web-harvested data are proposed. The best performing feature is motivated by findings in cognitive science and psycholinguistics about the asymmetry of the semantic priming.

1.5 Organization of the Thesis

Two models for estimating semantic similarity between words are presented in Chapter 2. The first type relies on the exploitation of knowledge resources, such as WordNet and Wikipedia. A number of different approaches are described including taxonomic and information content-based methods. A completely different approach for the representation of word semantics and the measurement of semantic similarity is adopted by the second type: the DSMs framework

whereas the distributional hypothesis of meaning is adopted. Particular attention is given to Vector Space Models (VSM), which constitute the main implementation of DSMs.

In Chapter 3, we focus on the creation of problem of fully unsupervised web-based DSMs used for estimating the semantic similarity computation between words and biomedical terms. In order to estimate the semantic similarity between words/terms two families of unsupervised, web-based similarity metrics are investigated. The first type considers only the number of hits returned by a web search engine. The second is fully corpus-based: the top-ranked documents returned by a web query are downloaded and the contextual similarity is employed for the estimation of semantic similarity. In addition, various schemes for the weighting of contextual features are investigated. The proposed methodology requires no expert knowledge or language resources and, as a result, it can be regarded as language-agnostic.

A new network-based implementation of unsupervised DSMs is proposed in Chapter 4. First, a corpus of document snippets is harvested from the web. Then, a semantic network is constructed encoding the semantic relations between words in the corpus. Co-occurrence and context features are used to measure the strength of relations. The network is regarded as a parsimonious representation of the information encoded in the corpus. We then the notion of semantic neighborhood is defined, as well as three associated metrics of semantic similarity. The proposed semantic similarity metrics are motivated by the maximum sense similarity, attributional similarity and metric space assumptions. In addition, the main network-based metrics are further investigated for the case of abstract and concrete nouns.

In Chapter 5, we deal with two basic types of lexical relations, namely, associative and semantic. More specifically, the automatic classification of associative and semantic is addressed. Lexical relations such as synonymy, hypernymy/hyponymy, constitute the fundamental types of semantic relations. Associative relations are harder to define, since they include a long list of diverse relations, e.g., “Cause-Effect”, “Instrument-Agency”. From the perspective of cognitive scientists, associative relatedness is triggered by the co-occurrence of words , while the definition of semantic relatedness is controversial. In particular, two novel features are proposed for the discrimination of these relations using information automatically extracted from the web, while syntactic patterns are also investigated

In Chapter 6, the main network-based similarity metrics proposed in Chapter 4 are applied to three problems: (i) the integration of visual with textual features for the creation of multimodal semantic networks, (ii) the estimation of semantic similarity between compositional noun compounds, based on the utilization of semantic neighborhoods and the adaptation of network similarity metrics, and (iii) the creation of a simple noun taxonomy.

This thesis concludes with Chapter 7, where on going and future research directions are also discussed. The further investigation of semantic neighborhoods is of immediate interest

including normalization issues and the application of algorithms from graph theory. Also, the adaptation of the proposed network metrics to the problem of sentence-level semantic similarity estimation constitutes an interesting extension of the network-based DSMs.

Chapter 2

Models of Lexical Semantic Similarity

In this chapter, two widely-used models for estimating semantic similarity between words are presented. The first type relies on the exploitation of knowledge resources, such as WordNet and Wikipedia. Distributional semantic models (DSMs) constitute a different type of models that adopt the distributional hypothesis of meaning, according to which the (corpus-based) contextual environment of words is considered.

2.1 Knowledge-based Models

2.1.1 WordNet-based

WordNet is a lexical network of words of certain parts of speech: nouns, verbs, adjectives and adverbs, Words are organized into a hierarchy of relations including hyponymy/hypernymy (IsA), meronymy, antonymy and the ComplementOf relation. Every sense of each word, w, is assumed to correspond to a lexicalized concept, c, which is defined with respect to (i) a set of synonyms (synsets), (ii) a definition (gloss), and (iii) an example of usage. For instance, the first sense of the noun "car" in WordNet3.0 has ("automobile", "machine", "motorcar") as synset, "a motor vehicle with four wheels; usually propelled by an internal combustion engine" as gloss, and the sentence "he needs a car to get to work" as an example of usage. Each concept stands as a node in this network, while concepts are linked via the aforementioned relations formulating the edges between nodes. The following definitions will be adopted for the description of WordNet-based similarity metrics Budanitsky and Hirst [2006]:

1. The shortest path between two concepts, c_i and c_j, has length $l(c_i, c_j)$. The length is computed in terms of nodes or edges.

2. The depth of a concept c_i, $d(c_i)$, is defined as the length of the path from the hierarchy root r to c_i. That is $d(c_i) = l(r, c_i)$.

3. The most specific common subsumer for two concepts, c_i and c_j, is denoted as $m(c_i, c_j)$.

4. Let $S(c_i, c_j)$ be a metric of semantic similarity between concepts c_i and c_j. The semantic similarity between two words w_i and w_j is computed as follows:

$$S(w_i, w_j) = \max_{c_i \in C_i, c_j \in C_j} \{S(c_i, c_j)\}, \tag{2.1}$$

where C_i and C_j denote the set of WordNet concepts that stand as senses for words w_i and w_j, respectively.

2.1.1.1 Length of taxonomic paths

A straightforward approach for the computation of word semantic similarity in the framework of a hierarchy is to take into account the length of the path that exist between their senses (concepts). According to Resnik [1995] the similarity between two concepts is inversely proportional to the length of their respective path. This hypothesis was adopted in Rada et al. [1989] for estimating the similarity of medical terms using the MeSH (Medical Subject Headings) hierarchy. More specifically, the number of edges that link two terms was assumed to indicate the semantic distance of terms. This approach was evaluated for an information retrieval task yielding good performance. The success of this simple approach was explained in Lee et al. [1993]: the utilization of $l(c_i, c_j)$ over a semantic network that is built upon a variety of relations fails to capture the semantic similarity between c_i and c_j, however, the performance of $l(c_i, c_j)$ improves when only IsA relations are considered. Hirst and St-Onge [1998] have proposed the following measurement for the computation of semantic similarity between c_i and c_j:

$$S_{HS}(w_i, w_j) = \alpha - l(c_i, c_j) - \beta t(c_i, c_j), \tag{2.2}$$

where α and β are fixed constants (set as $\alpha = 8$, $\beta = 1$), while $t(c_i, c_j)$ denotes the number of times the direction of $l(c_i, c_j)$ changes. Three types of directions were defined across $l(c_i, c_j)$: (i) horizontal (antonymy), (ii) upward (hypernymy, meronymy), and (iii) downward (hyponymy, holonymy).

A drawback of the above length-based approach is the implicit assumption that every edge correspond to the same amount of (semantic) distance. However, according to Resnik [1995] this does not hold since there exists a considerable variability regarding the semantic distances represented by edges. This was observed for sub-networks, which tend to be denser that others.

A number of scaling approaches have been proposed for addressing this issue as follows.

The approach of Sussna [1997] was motivated by the finding that sibling-concepts that are positioned at the lower levels of the hierarchy tend to be more similar in contrast to those that lie at the upper levels. A dual directionality was assigned to each edge enabling the distinction of the corresponding relation as forward, r, i.e., relation between c_i and c_j, and backward, $r^{'}$, i.e., relation between c_j and c_i. Also, for each relation r a range of weights was defined $[\min_r, \max_r]$. The core idea is to normalize the weight of each edge that correspond to a relation r and leaves from c_i, $q(c_i \rightarrow r)$, by the total number of edges of r type that also leave from c_i denoted as $e_r(c_i)$:

$$q(c_i \rightarrow r) = \max_r - \frac{\max_r - \min_r}{e_r(c_i)}. \tag{2.3}$$

The semantic distance between two adjacent concepts c_i and c_j, $D_s(c_i, c_j)$, is defined as the sum of their respective weights computed across the two directions of the underlying relation, i.e., r and $r^{'}$, normalized by the maximum concept depth:

$$D_s(c_i, c_j) = \frac{q(c_i \rightarrow r) + q(c_j \rightarrow r^{'})}{2 \max\{d(c_i), d(c_j)\}}. \tag{2.4}$$

The semantic distance between two concepts c_i and c_j that lie in arbitrary positions in the network is estimated by summing the semantic distances of the adjacent concepts (according to (2.4)) that exist in the shortest path that links c_i and c_j.

The approach proposed in Wu and Palmer [1994] incorporates the depth of the most specific common subsumer of c_i and c_j, $d(m(c_i, c_j))$, as a scaling factor into the computation of semantic similarity between c_i and c_j:

$$S_{WP}(c_i, c_j) = \frac{2d(m(c_i, c_j))}{l(c_i, m(c_i, c_j)) + l(c_j, m(c_i, c_j)) + 2d(m(c_i, c_j))}. \tag{2.5}$$

In Leacock and Chodorow [1998], the length of the shortest path between c_i and c_j, $l(c_i, c_j)$, is normalized by the maximum depth that exists in the hierarchy and the semantic similarity between c_i and c_j is computed as:

$$S_{LC}(c_i, c_j) = -\log \frac{l(c_i, c_j)}{2 \max_{c \in C}\{d(c)\}}, \tag{2.6}$$

where C denotes the set of all concepts that are included in the network. Unlike other approaches this type of normalization is not conditioned on the concepts under investigation, i.e., it depends only on the used network.

2.1.1.2 Information content

The core idea of information content-based approaches is the augmentation of typical path-based methods of similarity computation through the incorporation of corpus-based statistics.

The influential work of Resnik [1995] is motivated by the hypothesis that two concepts are similar to the degree of their shared information. In the framework of WordNet this can be implemented by considering the position of the most specific subsumer of the concepts under comparison. An excerpt of the WordNet hierarchy including "nickel", "dime" and "credit card" is shown in Fig. 2.1. Solid lines denote IsA relations, while dashed lines represent the intermediate nodes not shown for the sake of space. The position of most specific common subsumer varies according to the concepts of interest: "coin" for "nickel" vs. "dime", and "medium of exchange" for "dime" vs. "credit card". Unlike "coin", "credit card" is an abstract concept, which is positioned at the upper levels of the hierarchy. According to Resnik [1995] the similarity between two concepts is estimated using the probability of occurrence of their respective most specific subsumer, defined as

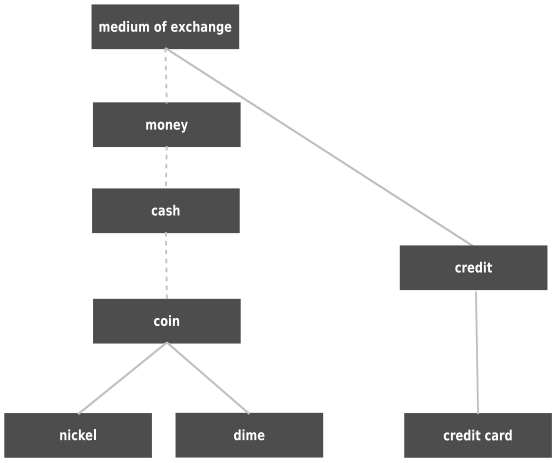


Figure 2.1: Excerpt of the WordNet hierarchy for "nickel", "dime" and "credit card". IsA relations are denoted by solid lines, while dashed lined stand for omitted intermediate nodes Budanitsky and Hirst [2006]; Resnik [1995].

$$S_R(c_i, c_j) = -\log p(m(c_i, c_j)), \tag{2.7}$$

where $-\log p(m(c_i, c_j))$ stands as the *information content* of $m(c_i, c_j)$. In Resnik [1995], the Brown Corpus Francis and Kučera [1982] was used for estimating the probability of concept c_i as

$$p(c_i) = \frac{\sum_{w \in W_i} f(w)}{N}, \tag{2.8}$$

where W_i denotes the set of words for which concept c_i is a subsumer, function $f(.)$ is the frequency for word w, and N is the total number of WordNet words that are included in the corpus. For the example depicted by Fig. 2.1, according to the counting approach of (2.8) the probability of "coin" is determined by the corpus frequencies of "nickel" and "dime". According to (2.8) and (2.7) the similarity between two concepts is increased as the specificity of their most specific subsumer increases. In contrast to the path-based similarity approaches, the metric defined by (2.8) utilize the structure of the underling hierarchy by identifying only the most specific common subsumers, ignoring the links between the concepts under comparison. According to Budanitsky and Hirst [2006], this strategy has some potentials drawbacks as it can be exemplified by the equal similarity scores of ("money","credit") and ("dime","credit card"), since both pairs have the same most specific subsumer, i.e., "medium of exchange".

The similarity metric of (2.7) was extended by the approach proposed in Jiang and Conrath [1997] including the information content of the individual concepts c_i and c_j. In particular, the proposed metric is a dissimilarity measurement defined as follows

$$D_J(c_i, c_j) = 2 \log p(m(c_i, c_j)) - (\log p(c_i) + \log p(c_j)). \tag{2.9}$$

Interestingly, in Pedersen [2010] it was shown that the use of (sense) untagged corpora by information content similarity metrics yielded higher performance compared to the largest sense-tagged corpus (SemCor) for the task of semantic similarity computation between words.

2.1.1.3 Gloss-based

This category of similarity metrics relies on the exploitation of word glosses (i.e., definitions) typically included in dictionaries and related resources. The key idea was initiated by Lesk [1986] for the task of word sense disambiguation. The core of the Lesk algorithm is the assignment of a sense to a target word provided within context by examining the overlap of its glosses with respect to the glosses of the other co-occurring (in the particular context) words. For example, the glosses for the first WordNet senses of "fruit" and "tree" are "the ripened reproductive body of a seed plant" and "a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms", respectively. There is a non-empty overlap because both glosses include "plant" (stop words are excluded). Each gloss was represented as a bag-of-words, while the target word was assigned the sense for which gloss overlap was maximized. This idea was based on the assumption that words that co-exist within the same context tend to refer to the same topic. The original algorithm was applied over three dictionaries, namely, Webster's 7^{th} Collegiate, the Collins English Dictionary, and the Oxford Advanced Learner's Dictionary of Current English A number

of other early approaches have been also employed dictionaries for word sense disambiguation, e.g., Niwa and Nitta [1994]; Wilks et al. [1990].

The availability of word glosses in WordNet has attracted the interest of recent research efforts, extending the early dictionary-based approaches. In Banerjee and Pedersen [2002], the original Lesk algorithm was extended by the idea of extended gloss overlap for the definition of a semantic similarity metric. In particular, the WordNet gloss of a word of interest was expanded by considering the glosses of other (directly) related words. The latter were identified by exploiting the WordNet hierarchy and taking into account relations such as hypernymy/hyponymy, meronymy, etc. The motivation for this extension was the observation that the length of glosses is of limited size [1], providing no adequate vocabulary for the task of similarity estimation. In addition, the computation of gloss overlap was extended in Banerjee and Pedersen [2002] by considering the common n-grams of glosses under comparison, in contrast to the original Lesk algorithm where only unigrams were used. The extended gloss overlap algorithm was found to outperform the Lesk algorithm with respect to the SenseEval-2 word sense disambiguation task. In Patwardhan and Pedersen [2006], a corpus-based methodology for estimating semantic similarity was combined with the work of Banerjee and Pedersen [2002] regarding the gloss extension. First, each word of interest was represented by a vector consisting of second-order co-occurrences derived from a corpus. The idea of second-order vector was proposed by Schütze [1998] in order to tackle the sparsity of VSM. All the glosses included in WordNet were aggregated for the creation of a corpus from which the aforementioned vectors were built. Second, each vector was augmented by following the approach suggested in Banerjee and Pedersen [2002]. The similarity between two words was estimated as the cosine of their respective vectors. The gloss vector approach was reported to outperform the extended gloss overlap algorithm for the task of semantic similarity estimation between words, and also obtained higher results for the SenseEval-2 word sense disambiguation task. Gloss-based vector were also employed in Inkpen and Hirst [2003] for the disambiguation of near synonyms .

2.1.2 Wikipedia-based

WordNet constitutes the most widely-used knowledge resource for a large number of approaches dealing with several semantic tasks including the estimation of semantic similarity between words. Recently, the collective effort of the Wikipedia project has created a large and continuously updated resource of encyclopedic knowledge that attracted the interest of several research communities including computational semantics.

[1] The average length of WordNet glosses was reported to be seven words Banerjee and Pedersen [2002].

Wikipedia concepts (i.e., articles) were used by Explicit Semantic Analysis (ESA) proposed in Gabrilovich and Markovitch [2007] for estimating word semantic similarity. More specifically, an index of the Wikipedia articles was constructed with respect to the words of interest. Each word was represented by a vector with each dimension corresponding to an article, term frequency-inverse document frequency was applied for weighting the vector elements. The similarity between words was estimated as the cosine of their vectors. An extension of the ESA algorithm was proposed by Hassan and Mihalcea [2009] dealing with the weighting of vector elements. Based on the observation that the ESA algorithm seemed to exhibit a bias towards larger articles, the values of vector elements were normalized according to the length of the corresponding article. In addition, the vector elements were further normalized taking into account the depth of the corresponding concepts in the Wikipedia category tree, assigning greater weight to more specific (i.e., positioned at lower tree levels) concepts. The similarity between words was estimated by a Lesk-like measurement Lesk [1986] based on the overlap of concept vectors. Temporal Semantic Analysis (TSA) was suggested in Radinsky et al. [2011] as an extension of the ESA algorithm. The key idea of the TSA algorithm was the modeling of the "temporal" characteristics of words (given an article collection) for the estimation of their semantic similarity. Such characteristics were defined with respect to the publishing dates of New York Times articles. The motivation for this approach was the observation that semantically similar words tend to appear in articles published around a certain date, although they may not co-exist within the same article.

Wikipedia was used mainly as an article collection by the aforementioned approaches, ignoring the underlying structure, i.e., the links between articles. This structure was utilized by a number of research efforts for representing the semantic of words and computing their semantic similarity. In Milne and Witten [2008], a Wikipedia concept (or word) was represented as a vector consisting of incoming or outgoing concepts. By incoming concepts are meant the concepts that point to the target concepts, while the outgoing concepts are the concepts to which the target concept points to. The vector cosine was used for estimating the similarity between concepts. The link-based representation was also employed in Liu and Chen [2010], where the concept similarity was computed using an overlap-based measurement like the Lesk algorithm Lesk [1986] or the taxonomic-based metric proposed in Wu and Palmer [1994].

The exploitation of multiple knowledge sources was motivated by the observation that the type of semantics of the (lexical) content of each resource are different Zesch et al. [2008]. Several fusion schemes for representing the semantic information have been followed for resources like WordNet, Wikipedia, and Wiktionary, e.g., separate representations that are combined Szarvas et al. [2011] or a joint representation Zhang et al. [2011].

2.1.3 Network-Based Approaches

The WordNet- and Wikipedia-based approaches discussed in Section 2.1.1 and Section 2.1.2, respectively, can be considered as network-based ones since the used resources are structured. However, the presented methods seem to make a simple use of the underlying network, rather than adapting other graph-based algorithms or related models taken from the literature of cognitive science. For example, regarding WordNet the major network feature is the length of the path existing between concepts, while only the direct links between concepts are used for the case of Wikipedia. In this section, a number of approaches are presented that apply more sophisticated algorithms and models for the task of semantic similarity estimation.

In Gouws et al. [2010], a the network was constructed using the links between Wikipedia concepts, i.e., articles. The similarity between two words (concepts) w_i and w_j was estimated by applying the spreading activation model Collins and Loftus [1975] over the network. Initially, a non-zero activation value was assigned to the node corresponding to w_i, while the activation values for the other nodes were set to zero. Spreading activation was triggered in order to propagate the initial activation value of w_i to w_j through their links. After the termination of the spreading process activation values were accumulated in w_2 and the rest nodes of the network. In order to represent the semantics of w_i a vector having as elements the (final) activation values of the network nodes was built. The same procedure was repeated for the case of w_2. The semantic similarity between w_i and w_j was estimated as the sum of the final activation values of w_i and w_j or as the cosine of their respective vectors. The approach proposed in Wojtinnek et al. [2012] was also motivated by the spreading activation model where a very large network was constructed by exploiting the links between Wikipedia concepts. For each node a vector was created including a number of strongly connected nodes. The similarity between two words was estimated as the cosine of their respective vectors. The Wikipedia-based network was found to yield better performance compared to a structured approach for network creation based on the British National Corpus. Another research effort that was inspired by the spreading activation model is the work of Harrington [2010] where a semantic network was created from an unstructured corpus as opposed to the exploitation of structured resources like Wikipedia. However, the links between words were identified using a set of linguistic tools for named entity recognition, parsing, and semantic analysis. A number WordNet-based similarity metrics were adapted by the WikiRelate! system Strube and Ponzetto [2006] to the Wikipedia structure. More specifically, three types of metrics were used, namely, path-based (Leacock and Chodorow [1998]; Rada et al. [1989]; Wu and Palmer [1994]), information content-based (Resnik [1995]), and gloss-based (Banerjee and Pedersen [2002]). In Hughes and Ramage [2007], a network was built using WordNet links and statistics from the sense-tagged SemCor

corpus. This network was considered as a Markov chain and random walks were applied for computing stationary distributions for the words of interest. A variant of the Kullback-Leibler divergence was proposed for estimating the similarity between words. A WordNet-derived network was also used in Agirre et al. [2006] where the personalized PageRank algorithm Haveliwala et al. [2002] was applied for the computation of a probability distribution for every target word. Word similarity was estimated via the cosine similarity between the vectorized distributions.

2.2 Distributional Semantic Models (DSMs)

The fundamental idea of distributional semantic models (DSMs) is the representation of word meaning by considering the context in which the word occurs, also known as the *distributional hypothesis of meaning*. This idea originates from early works in theoretical linguistics Firth [1957]; Harris [1954] and even philosophy Wittgenstein [1953], and it is also summarized by the famous statement of Firth "you shall know a word by the company it keeps". Although Wittgenstein was mainly interested in the paralinguistic aspects of language, e.g., social factors, his argument "meaning is use" is consistent with the underlying assumption of DSMs.

Word-occurrence is the building block of high-dimensional spaces for context representation known as vector space models (VSM) Turney and Pantel [2010]. Such models are defined with respect to a specified vocabulary; one dimension is allocated for each vocabulary item. This constitutes a spatial representation in which the notion of semantic similarity is approximated in terms of proximity Sahlgren [2006]. The *geometric metaphor of meaning* has been theoretically investigated in Lackoff and Johnson [1980, 1997]. The core idea is that words with similar meaning exist "near" each other, while the dissimilar ones are positioned "far apart". According to the geometric metaphor of meaning, words are represented as points in a space, while their similarity is considered as the proximity between the corresponding points Sahlgren [2006]. One of the first exprerimental studies of the distributional hypothesis of meaning is Rubenstein and Goodenough [1965], suggesting that "words which are similar in meaning occur in similar contexts". This statement was re-visited in Schütze and Pedersen [1995], considering the data sparseness problem, as "words with similar meanings will occur with similar neighbors if enough text material is available". The linguist Zelling Harris initially believed that it is possible to typologize the entire spectrum of semantics based solely on their distributional properties Harris [1968, 1970]. Later, he revised this belief acknowledging the effect of extralinguistic factors. The core idea of his work is that the differences in meaning are mediated by differences in distributional features: "...if we consider words or morphemes A and B to be more different in meaning that A and C, then will often find that the distributions

of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution". The earliest validation of the distributional hypothesis was conducted in Rubenstein and Goodenough [1965] where the contextual similarities of 65 noun pairs where compared to synonymy scores given by students. It is worth to quote their main conclusions: (a) "there is a positive relationship between the degree of synonymy (semantic similarity) existing between a pair of words and the degree to which their contexts are similar", and (b) "it may safely inferred that a pair of words is highly synonymous if their contexts show a relatively great amount of overlap. Inference of degree of synonymy from less amounts of overlap, however, is apparently uncertain since words of low or medium synonymy differ relatively little in overlap". Moreover, Rubenstein and Goodenough, noted that the generalization of the above conclusions is dependent on factors like vocabulary size and homogeneity of content. Three decades later the experiment of Rubenstein and Goodenough was repeated in Miller and Charles [1998] (30 out of the 65 pairs were used) where similar conclusions were reached.

VSM are typically formulated as matrices constituting a formal implementation of DSMs. There are two common types of such matrices, namely, word-context and word-document matrices, constructed for computing similarity between words Landauer and Dumais [1997] and documents/queries Salton et al. [1975], respectively. This work is related with the word-context type, where each target word is represented by a vector (matrix row) that contains the words (matrix columns) that co-occur with it within a specified context (also referred as contextual features). Beyond its simplicity, the popularity of this representation can be attributed to the fact that it is well-aligned with the distributional hypothesis of meaning. The construction of VSM includes the following parameters: (i) extraction of contextual features, (ii) schemes for weighting the extracted contextual features, (iii) optional techniques for dimensionality reduction, and (iv) metrics for the computation of similarity (or distance) between the target words.

2.2.1 Extraction of Contextual Features

The primary input of DSMs is a corpus (or a set of corpora) whose lexical content is assumed to capture the semantics for the target words. Let the following sentences serve as an (toy) example of such a corpus.

```
Cars are motor vehicles with four wheels;
usually propelled by an internal combustion engine.
A tree is a tall perennial woody plant having
a main trunk and branches forming a distinct elevated crown.
They built a large plant to manufacture
a special type of engine for cars.
```

```
He reads his newspaper at breakfast.
```

Following the distributional hypothesis of meaning the first issue to be defined is the context according to which the contextual features for the target words will be extracted. The definition of context seems to depend on the problem under investigation Sahlgren [2008]. For example, within the framework of Information Retrieval this is typically defined at the document level where the task at hand is the topical similarity. However, a more narrow contextual scope may be used for the problem of word similarity e.g., sentential level or even few words in the left and right context of the target word Clark [2013]. The word-context matrix for the example corpus

	breakfast	cars	crown	large	motor	...	tall	trunk	vehicles
engine	0	2	0	0	1	...	0	0	1
newspaper	1	0	0	0	0	...	0	0	0
plant	0	1	1	1	0	...	1	1	0
tree	0	0	1	0	0	...	1	1	0

Table 2.1: Example of word-context matrix.

presented above is presented in Table 2.1 for which the context was defined at the sentence level. A sample of contextual features (columns) is illustrated for few target words (rows). The value of each matrix element stand for the (raw) feature frequency computed within the context of the corresponding target word. More weighting schemes for scoring the contextual features are presented in Section 2.2.2. A number of corpus pre-processing steps are required for such representation, e.g., sentence splitting for this example. Other pre-preprocessing steps may include lemmatization, stemming, normalization, filtering of stop words etc. Note that there are no standard procedures for the pre-processing of corpora towards the construction of VSM. e.g., inclusion of stop words as contextual features. Once the word-context representation is completed the meaning of each target word is reflected by its contextual features. According to theoretic linguistics the relation between target words and contextual features is characterized as *syntagmatic*, while the words that tend to occur in similar contextual environment are defined as *paradigmatically* related Cruse [1986]; Sahlgren [2006].

2.2.1.1 Unstructured and Structured Models

There are two main approaches regarding the extraction of the contextual features, namely, unstructured and structured Baroni and Lenci [2010]. This distinction deals with the consideration (or not) of syntactic relations between the target words and their contextual features.

Unstructured approaches do not consider the linguistic structure of context with respect to the target words. A contextual window of fixed size (K words) is centered on the target

word and the surrounding lexical features that fall within it are extracted Bullinaria and Levy [2007]; Iosif and Potamianos [2010]. Specifically, the right and left contexts of length K are considered for each occurrence of the target w in the corpus, i.e.,

$$[v_{K,L} \ ... \ v_{2,L} \ v_{1,L}] \ w \ [v_{1,R} \ v_{2,R} \ ... \ v_{K,R}],$$

where $v_{i,L}$ and $v_{i,R}$ represent the i^{th} word to the left and to the right of w respectively. The feature vector for target w is defined as $T_{w,K} = (t_{w,1}, t_{w,2}, ..., t_{w,N})$, where $t_{w,i}$ is a non-negative integer and K is the context window size. Note that the length of the feature vector is equal to the vocabulary size N, i.e., all words in the vocabulary are considered as features (unless a selection is applied, e.g, exclusion of stop words). The i^{th} feature value $t_{w,i}$ reflects the occurrence of vocabulary word v_i within the left or right context window K of (all occurrences of) the term w. As it was mentioned, the (optimal) size of the contextual window may vary according to the task under investigation, e.g., from immediate context used for computing the semantic similarity between words Agirre et al. [2009]; Iosif and Potamianos [2010], to larger context size used for estimating the reaction time during lexical priming Lund and Burgess [1996]. An extension of the unstructured approach is the employment of second-order co-occurrence statistics for the creation of contextual feature vectors Schütze [1998]. Two words are characterized by second-order co-occurrence if they do not co-occur directly, but both co-exist with a third word. Schütze applied this extension for the problem of word sense discrimination in order to tackle the sparsity of first-order VSM and improve its robustness.

The basic idea behind structured models is the utilization of syntactic relationships as features for the creation of semantic spaces. Typical examples of such relations are argument structures (subject/object) and modifications (adjective-noun) extracted by shallow or full parsing Pado and Lapata [2007]. Syntactic relations can be represented as 2-tuples of the arguments Grefenstette [1994] or as n-tuples in order to incorporate direct and indirect dependencies Pado and Lapata [2007]. The paradigm of "one task, one model" of structured DSMs was advanced in Baroni and Lenci [2010] by the arrangement of tuples into a third-order tensor. This enables the creation of different semantic spaces for different semantic tasks (e.g., estimation of semantic similarity between words, categorization of concepts, computation of verb selectional preferences, etc.), while the extraction of dependency tuples is task-independent ("the same distributional information can be shared across tasks"). In Agirre et al. [2006], unstructured DSMs were shown to obtain slightly higher performance than structures ones.

2.2.1.2 Exemplar Models

The issue of polysemy is raised as a drawback regarding the representation of semantic spaces adopted by traditional DSMs Erk and Padó [2010]. Typically, in DSMs a single vector is

used for representing the contextual features of a target word. For the case of a polysemous target word the single-vector representation conflates features that correspond to the different (in-corpus) senses of the target. This can be illustrated by observing the contextual features of "plant" in the example of Table 2.1. The use of exemplar models was proposed as an alternative implementation of DSMs for addressing the problem of polysemy Erk and Padó [2010]; Reddy et al. [2011]. Instead of a single vector, a set of exemplars is utilized for the representation of a target word. The set of exemplars is defined as the set of (corpus) sentences in which the target occurs Erk and Padó [2010]. For example, the word "light" has a set of $316,126$ exemplars in the ukWaC corpus Reddy et al. [2011]. Each exemplar can be represented as a unstructured (i.e., bag-of-words) or structured (i.e., encoding syntactic relations) vector. For a target word given within context (e.g., sentential) polysemy is modeled by the activation/selection of the relevant exemplars with respect to a "point of comparison", where the latter can be regarded as another exemplar Erk and Padó [2010]. In Erk and Padó [2010], the activation was controlled by setting a threshold regarding the similarity of exemplars estimated by measurements such cosine similarity or Jaccard coefficient. An approach for exemplar selection was proposed in Reddy et al. [2011] for the case of noun-noun compounds, e.g., "traffic light". The basic idea was to constrain the exemplars of the one constituent to include words semantically related with the other constituent, e.g., exemplars of "light" may include the word "car". The exemplar-based approach seems to be aligned with semantic tasks where the words of interest are provided within context. A related example is the selection of paraphrases for a target word that occurs in a given context Erk and Padó [2010]. In Reddy et al. [2011], the exemplar model was applied in the framework of semantic compositionality for the task of similarity computation between noun-noun compounds. However, the polysemy mechanism of exemplar models is not obvious how is applied (i.e., which exemplars to use and how) and benefits out-of-context semantic tasks in comparison with the typical implementations of DSMs.

2.2.2 Weighting of Contextual Features

In this section, a number of widely-used measurements are presented for the weighting of the contextual features. Following the definitions of Curran [2003] let (w, r, v) denote the co-occurrence of target word w and feature v under relation r within the specified context. Note that r can be any relation defined according to the used structured model. For the case of unstructured DSMs r simply signifies the co-occurrence of w and v. Also, let $f(w, r, v)$ denote the unnormalized corpus frequency of (w, r, v).

Identity. This is the simplest weighting scheme assigning 1 if relation r exists between w

and v without taking into account the frequency of the relation:

$$t_{w,i} = 1. \tag{2.10}$$

Freq. This scheme employs the raw frequency of (w, r, v):

$$t_{w,i} = f(w, r, v). \tag{2.11}$$

RelFreq. In this scheme, $f(w, r, v)$ is normalized by the frequency of target word w:

$$t_{w,i} = \frac{f(w, r, v)}{f(w, \star, \star)}. \tag{2.12}$$

Note that in the denominator the $\star$ symbol is used as a placeholder (for any relation r or contextual feature v).

Tf–Idf. This scheme was inspired by the term frequency-inverse document frequency scheme that is widely used in Information Retrieval:

$$t_{w,i} = \frac{f(w, r, v)}{n(\star, r, v)}. \tag{2.13}$$

Here, the notion of inverse document frequency is adapted for the construction of word-content matrices (instead of word-document matrices). For this purpose, $n(\star, r, v)$ denotes the number of different relations in which the contextual feature v is involved.

Gref94. This is an extension of the Tf-Idf scheme proposed in Grefenstette [1994]:

$$t_{w,i} = \frac{\log_2(f(w, r, v) + 1)}{\log_2(n(\star, r, v) + 1)}. \tag{2.14}$$

The logarithm was introduced in order to reduce the domination of high frequencies.

In this thesis, we have experimented with the most of the aforementioned weighting schemes, as well as with some variations of the Tf-Idf. More details are given in Table 3.1 of Section 3.2.

2.2.3 Dimensionality Reduction

The typical dimensionality of the word-context matrix is tens of thousands, while the vast majority of its elements are equal to zero. The most widely-used technique for reducing the dimensions of such matrices is the Singular Value Decomposition (SVD), which is based on linear algebra. SVD can be applied both to word-document and word-context matrices, while the estimation of similarity between documents and words is referred to as Latent Semantic Indexing (LSI) and Latent Semantic Analysis (LSA), respectively. The incorporation of SVD within the framework of Information Retrieval was introduced by Deerwester et al. [1990]. One of the earliest applications of SVD for the estimation of similarity between words is described in Landauer and Dumais [1997].

The key idea behind SVD is the factorization of the original matrix $\mathbf{X}$ with respect to three matrices. The latter are utilized in order to formulate a low–rank approximation of $\mathbf{X}$. This approximation usually results into a significant dimensionality reduction: from tens of thousands to few hundreds. In particular, the original matrix $\mathbf{X}$ is expressed as a product of the matrices $\mathbf{U\Sigma V}^{\mathrm{T}}$ Turney and Pantel [2010]. The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthogonal having unit length $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$. The singular values are included in $\mathbf{\Sigma}$ that is diagonal, while $\mathbf{X}$ and $\mathbf{\Sigma}$ have the same rank r. Consider a matrix $\mathbf{\Sigma}_{\mathrm{k}}$ formulated by the top k singular values, where $k < r$. Also, let $\mathbf{U}_{\mathrm{k}}$ and $\mathbf{V}_{\mathrm{k}}$ be the matrices that are created by the corresponding columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. The original matrix $\mathbf{X}$ can be approximated by $\hat{\mathbf{X}} = \mathbf{U}_{\mathrm{k}}\mathbf{\Sigma}_{\mathrm{k}}\mathbf{V}_{\mathrm{k}}^{\mathrm{T}}$ (with rank equal to k) given that the error $||\ \hat{\mathbf{X}} - \mathbf{X}\ ||_{\mathrm{F}}$ is minimized. $||\ .\ ||_{\mathrm{F}}$ stands for the Frobenius norm Golub and Loan [1996].

A number of different perspectives regarding the application of SVD are briefly described in Turney and Pantel [2010]. The low-dimensional approximation is considered to (i) capture the latent meaning of words, (ii) reveal higher-order co-occurrences, (iii) reduce to the "noise" introduced by non-informative contextual features, and (iv) tackle the sparsity problem.

A criticism about the application of SVD within the framework of VSM is raised in Turney and Pantel [2010]. This deals with the underlying assumption according to which the contextual features follow a Gaussian distribution, which is not true. A number of most recent approaches that attempt to address this issue are reported in Turney and Pantel [2010] including Probabilistic Latent Semantic Indexing (PLSI) Hofmann [1999] and Latent Dirichlet Allocation (LDA) Blei et al. [2003].

2.2.4 Semantic Similarity Metrics

In this section, two types of similarity metrics are presented, namely co-occurrence-based and context-based. According to the first type the similarity between words is estimated by using

directly the co-occurrence of the words of interest, i.e, syntagmatic relatedness. The context-based metrics rely on the distributional hypothesis of meaning according to which the semantic similarity is implied by the paradigmatic relatedness.

2.2.4.1 Co-occurrence-based metrics

Co-occurrence metrics use association ratios between words that are computed using their co-occurrence frequency in a specified context. The definitions that follow consider the "web as a corpus", i.e., the word co-occurrence is regarded at the document level. The basic assumption of this approach is that high association ratios indicate a semantic relation between words[1] Church and Hanks [1990]. For the documents indexed by a search engine we define the notations shown in Table 2.2 Feldman et al. [1998]. We use the notation $\{D\}$ for a set

Notation	Description
$\{D\}$	set of all documents indexed by search engine
$\|D\|$	number of documents in $\{D\}$
w	a word or term
$\{D\|w\}$	subset of $\{D\}$, documents indexed by w
$\{D\|w_1, w_2\}$	subset of $\{D\}$, documents indexed by w_1 and w_2
$\|D\|w\|$	number of documents in $\{D\}$ indexed by w
$\|D\|w_1, w_2\|$	number of documents in $\{D\}$ indexed by w_1 and w_2

Table 2.2: Definitions for document sets indexed by search engines.

of documents, $|D|$ for document set cardinality, $\{D|w\}$ for the set of documents that contain the word w and $\{D|w_1, w_2\}$ for the set of documents that contain both word w_1 *and* w_2. In this work, four co-occurrence measures are used to compute semantic similarity between word or term pairs, namely: Jaccard coefficient, Dice coefficient, mutual information (as defined in Bollegala et al. [2007]), and Google-based Semantic Relatedness Gracia et al. [2006].

Jaccard and Dice coefficients. The Jaccard coefficient is a measure for calculating the similarity (or diversity) between sets. The variation of the Jaccard coefficient used in this work is defined as:

$$J(w_1, w_2) = \frac{|D|w_1, w_2|}{|D|w_1| + |D|w_2| - |D|w_1, w_2|} \tag{2.15}$$

[1] It is interesting to note that web-based co-occurrence metrics often outperform more elaborate corpus-based metrics. This shows that overcoming the data sparseness problem is sometimes more important than building an accurate estimator. For example an improved n-gram language probability estimation using web n-gram occurrence can be found in the literature Zhu and Rosenfeld [2001].

In probabilistic terms, (2.15) finds the maximum likelihood estimate of the ratio of the probability of finding a document where words w_1 and w_2 co-occur over the probability of finding a document where either w_1 or w_2 occurs[1]. If w_1 and w_2 are the same word then the Jaccard coefficient is equal to 1 (absolute semantic similarity). If two words never co-occur in a document collection then the Jaccard coefficient is 0. The Dice coefficient is related to the Jaccard coefficient and is computed as:

$$C(w_1, w_2) = \frac{2\,|\,D\,|\,w_1, w_2\,|}{|\,D\,|\,w_1\,| + |\,D\,|\,w_2\,|} \tag{2.16}$$

Again, the Dice coefficient is equal to 1 if w_1 and w_2 are identical, and 0 if two words never co-occur.

Mutual information. If we assume that the number of documents indexed by the words w_1, w_2 are random variables X, Y, respectively, then the pointwise mutual information (MI) between X and Y measures the mutual dependence between the occurence of words w_1 and w_2 Church and Hanks [1990]. The maximum likelihood estimate of MI is:

$$I(X, Y) = \log \frac{\frac{|D|w_1, w_2|}{|D|}}{\frac{|D|w_1|}{|D|}\frac{|D|w_2|}{|D|}} \tag{2.17}$$

Mutual information measures the information that variables X and Y share. It quantifies how the knowledge of one variable reduces the uncertainty about the other. For instance, if X and Y are independent, then knowing X does not give any information about Y and the mutual information is 0. For $X = Y$, the knowledge of X provides the value of Y with certainty and the mutual information is 1. Note that the number of relevant documents is normalized by the total number of documents indexed by the search engine, $|D|$, giving a maximum likelihood estimate of the probability of finding a document in the web that contains this word.

Google-based Semantic Relatedness. Motivated by Kolmogorov complexity, Cilibrasi and Vitanyi Cilibrasi and Vitanyi [2007]; Vitanyi [2005] proposed a page-count-based similarity measure, called the Normalized Google Distance, defined as:

$$G_0(w_1, w_2) = \frac{\max\{A\} - \log\,|\,D\,|\,w_1, w_2\,|}{\log\,|\,D\,| - \min\{A\}}, \tag{2.18}$$

[1]The normalization terms $|\,D\,|$ (total number of documents) at the nominator and denominator cancel each other out.

where $A = \{\log \mid D \mid w_1 \mid, \log \mid D \mid w_2 \mid\}$. As the semantic similarity between two words increases, the distance computed by (2.18) decreases. Thus, this metric can be considered as a dissimilarity measure. Note that the metric is also unbounded, ranging from 0 to ∞. In Gracia et al. [2006], a variation of Normalized Google Distance is proposed that defines a similarity measurement. This variation is typically referred to as "Google-based Semantic Relatedness":

$$G(w_1, w_2) = e^{-2G_0(w_1, w_2)} \tag{2.19}$$

where $G_0(w_1, w_2)$ is computed according to (2.18). Note that the Google-based Semantic Relatedness is bounded taking values between 0 and 1.

Beside the use of the web as a corpus for obtaining number of hits, the aforementioned co-occurrence-based metrics can be also defined with respect to any text corpus. In such cases, the word frequencies can be considered at the level of several corpus units, e.g., sentences, paragraphs. In this thesis, we adopt both perspectives, i.e., employing web-based hits, as well as word frequencies computed over a text corpus. The performance of these perspectives for the task of semantic similarity computation between words is presented in Chapter 4.

2.2.4.2 Context-based metrics

Unlike co-occurrence-based metrics, the semantic similarity is estimated through the paradigmatic relatedness for the case of context-based metrics. This approach follows the distributional hypothesis of meaning suggesting that "similarity of context implies similarity of meaning".

Cosine similarity. It is reported to be the most widely-used similarity metric with respect to VSM Clark [2013]; Turney and Pantel [2010]. For a given weighting scheme the similarity between two words , w_1 and w_2, is estimated as the cosine of their corresponding feature vectors, $T_{w_1,K}$ and $T_{w_2,K}$, as follows Iosif and Potamianos [2010]:

$$Q^H(w_1, w_2) = \frac{\sum_{i=1}^{N} t_{w_1,i} t_{w_2,i}}{\sqrt{\sum_{i=1}^{N} (t_{w_1,i})^2} \sqrt{\sum_{i=1}^{N} (t_{w_2,i})^2}} \tag{2.20}$$

where H is the context window length and N is the vocabulary size. The cosine similarity metric assigns 0 similarity score when w_1, w_2 have no common context (completely dissimilar words), and 1 for identical words.

Besides cosine similarity a number of other metrics have been employed for estimating the semantic similarity between words. In the next paragraphs, a number of info-theoretic and geometric measurements are briefly presented. We follow the formulation proposed in Meng and Siu [2002]; Pargellis et al. [2004] where the contextual feature vectors were transformed

into probability distributions. In particular, the contextual probabilities were estimated using n-gram language modeling.

Kullback-Leibler (KL) divergence. This is a measure of the difference between two probability distributions, while it is also known as *relative entropy* Cover and Thomas [1991]. Suppose two distributions, P and Q of a discrete random variable. Their KL divergence is computed as

$$D_{KL}(P\|Q) = \sum_{y \in Y} P(y) \log \frac{P(y)}{Q(y)} \tag{2.21}$$

over all values $y \in Y$. KL metric is not symmetric, i.e., $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$, In addition, KL can be regarded as a measurement of dissimilarity since equals to 0 when the two distributions are the same, and greater than zero otherwise. Given that the KL metric measures the dissimilarity between two distributions, the greater their divergence is, the easier (on average) their discrimination is Kullback [1959]; Lee [1997]. From another point of view, if the difference between distributions P and Q is large, then P and Q is dissimilar, so, it is inefficient (on average) to use Q instead of P Kullback [1959]; Lee [1997]. Within the framework of DSMs the KL divergence can be applied for the estimation of word semantic dissimilarity given that the contextual vectors of target words are transformed into probability distributions. For example, this was performed in Meng and Siu [2002]; Pargellis et al. [2004] by considering the immediate context of target words, i.e., estimating probabilities of the left and right bigrams of targets. Let the bigram probability distributions of target words w_1 and w_2, denoted as W_1 and W_2, respectively. The semantic dissimilarity between w_1 and w_2 was estimated using the KL divergence of the corresponding right bigram conditional probability distributions W_1 and W_2, as:

$$D^R_{KL}(W_1\|W_2) \equiv D^R_{KL}(w_1, w_2) = \sum_{v_{1,R} \in V} p(v_{1,R} \mid w_1) \log \frac{p(v_{1,R} \mid w_1)}{p(v_{1,R} \mid w_2)}, \tag{2.22}$$

where $v_{1,R}$ denotes the first word that occurs in the right contexts of word w_i ($i = 1$ *or* 2), $p(v_{1,R} \mid w_i)$ is the bigram conditional probability of the bigram "$w_i\ v_{1,R}$". Also, note that the two bigram distributions, W_1 and W_2, are compared over the whole vocabulary V [1]. In similar fashion with (2.22), the divergence between W_2 and W_1 for the right contexts is computed as:

$$D^R_{KL}(W_2\|W_1) \equiv D^R_{KL}(w_2, w_1) = \sum_{v_{1,R} \in V} p(v_{1,R} \mid w_2) \log \frac{p(v_{1,R} \mid w_2)}{p(v_{1,R} \mid w_1)}. \tag{2.23}$$

[1] The backoff strategy can be followed in language modeling for estimating the probability of bigrams that do not occur within the corpus.

Regarding the left context-dependent divergence (2.22) and (2.23) are formulated as:

$$D^L_{KL}(W_1 \| W_2) \equiv D^L_{KL}(w_1, w_2) = \sum_{v_{1,L} \in V} p(v_{1,L} \mid w_1) \log \frac{p(v_{1,L} \mid w_1)}{p(v_{1,L} \mid w_2)} \tag{2.24}$$

and

$$D^L_{KL}(W_2 \| W_1) \equiv D^L_{KL}(w_2, w_1) = \sum_{v_{1,L} \in V} p(v_{1,L} \mid w_2) \log \frac{p(v_{1,L} \mid w_2)}{p(v_{1,L} \mid w_1)}, \tag{2.25}$$

respectively. The symmetric left and right contextual dissimilarity between words w_1 and w_2 is defined as Pargellis et al. [2004]:

$$D^{L,R}_{KL}(w_1, w_2) = D^L_{KL}(w_1, w_2) + D^L_{KL}(w_2, w_1) + D^R_{KL}(w_1, w_2) + D^R_{KL}(w_2, w_1). \tag{2.26}$$

The KL metric is unbounded, since the bigram probabilities that appear in the denominators may take values close to zero. Due to this, the computation of the KL divergence was observed to be dominated by few, infrequent bigrams Pargellis et al. [2004].

Information-radious (IR). This metric is similar to the KL divergence (also known as Jensen–Shannon divergence Lin [1991]), however it is bounded, since the denominator is the average of the probability distributions:

$$D_{IR}(P \| Q) = \sum_{y \in Y} P(y) \log \frac{P(y)}{\frac{1}{2}(P(y) + Q(y))}. \tag{2.27}$$

As in KL, the divergence of bigram conditional probability distributions W_1 and W_2 (and vice versa) are defined as follows:

$$D^R_{IR}(W_1 \| W_2) \equiv D^R_{IR}(w_1, w_2) = \sum_{v_{1,R} \in V} p(v_{1,R} \mid w_1) \log \frac{p(v_{1,R} \mid w_1)}{\frac{1}{2}(p(v_{1,R} \mid w_1) + p(v_{1,R} \mid w_2))} \tag{2.28}$$

and

$$D^R_{IR}(W_2 \| W_1) \equiv D^R_{IR}(w_2, w_1) = \sum_{v_{1,R} \in V} p(v_{1,R} \mid w_2) \log \frac{p(v_{1,R} \mid w_2)}{\frac{1}{2}(p(v_{1,R} \mid w_1) + p(v_{1,R} \mid w_2))} \tag{2.29}$$

for the right contexts, respectively. Similarly, for the left contexts we have

$$D^{L}_{IR}(W_1 \| W_2) \equiv D^{L}_{IR}(w_1, w_2) = \sum_{v_{1,L} \in V} p(v_{1,L} \mid w_1) \log \frac{p(v_{1,R} \mid w_1)}{\frac{1}{2}(p(v_{1,L} \mid w_1) + p(v_{1,L} \mid w_2))} \tag{2.30}$$

and

$$D^{L}_{IR}(W_2 \| W_1) \equiv D^{L}_{IR}(w_2, w_1) = \sum_{v_{1,L} \in V} p(v_{1,L} \mid w_2) \log \frac{p(v_{1,L} \mid w_2)}{\frac{1}{2}(p(v_{1,L} \mid w_1) + p(v_{1,L} \mid w_2))}. \tag{2.31}$$

The symmetric left and right contextual dissimilarity between w_1 and w_2 is computed as Pargellis et al. [2004]:

$$D^{L,R}_{IR}(w_1, w_2) = D^{L}_{IR}(w_1, w_2) + D^{L}_{IR}(w_2, w_1) + D^{R}_{IR}(w_1, w_2) + D^{R}_{IR}(w_2, w_1). \tag{2.32}$$

Each of the four terms of the above summation has an upper bound value equal to $\log(2)$, so the maximum score of absolute dissimilarity is $4\log(2)$.

Manhattan-norm (MN). This is a geometric measurement Bullinaria and Levy [2007] defined as follows:

$$D_{MN}(P \| Q) = \sum_{y \in Y} \mid P(y) - Q(y) \mid . \tag{2.33}$$

In particular, the MN metric relies on the absolute difference between the bigram conditional probability distributions W_1 and W_2. Due to the absolute function the MN metric is symmetric:

$$D_{MN}(P \| Q) \equiv D_{MN}(Q \| P)$$

The contextual distance between the bigram conditional probability distributions W_1 and W_2 is

$$D^{R}_{MN}(W_1 \| W_2) \equiv D^{R}_{MN}(w_1, w_2) = \sum_{v_{1,R} \in V} \mid p(v_{1,R} \mid w_1) - p(v_{1,R} \mid w_2) \mid \tag{2.34}$$

for the right contexts. In similar manner, the distribution distance for the left context is defined as

$$D^{L}_{MN}(W_1 \| W_2) \equiv D^{L}_{MN}(w_1, w_2) = \sum_{v_{1,L} \in V} \mid p(v_{1,L} \mid w_1) - p(v_{1,L} \mid w_2) \mid . \tag{2.35}$$

The symmetric (left and right) contextual dissimilarity between w_1 and w_2 is computed as

Pargellis et al. [2004]:

$$D_{MN}^{L,R}(w_1, w_2) = D_{MN}^{L}(w_1, w_2) + D_{MN}^{R}(w_1, w_2). \tag{2.36}$$

Each of both terms of (2.36) has a lower and upper bound of zero and two, respectively. Thus, two words of identical contextual distributions will have a zero value of MN distance, while a distance score equal to four indicates absolute dissimilarity.

Cosine similarity (re-formulated). In this paragraph, the cosine similarity metric defined in (2.20) is re-formulated for the case when the contextual feature vectors are transformed into probability distributions:

$$S_{CS}(P\|Q) = \frac{\sum_{y\in Y} P(y)Q(y)}{\sqrt{\sum_{y\in Y} P(y)^2 \sum_{y\in Y} Q(y)^2}}. \tag{2.37}$$

Note that the CS metric is symmetric:

$$S_{CS}(P\|Q) \equiv S_{CS}(Q\|P)$$

The similarity between the bigram conditional probability distributions W_1 and W_2 is computed as:

$$S_{CS}^{R}(W_1\|W_2) \equiv S_{CS}^{R}(w_1, w_2) = \frac{\sum_{v_{1,R}\in V} p(v_{1,R} \mid w_1)p(v_{1,R} \mid w_2)}{\sqrt{\sum_{v_{1,R}\in V} p(v_{1,R} \mid w_1) \sum_{v_{1,R}\in V} p(v_{1,R} \mid w_2)}} \tag{2.38}$$

for the right contexts. Similarly, for the left context we have

$$S_{CS}^{L}(W_1\|W_2) \equiv S_{CS}^{L}(w_1, w_2) = \frac{\sum_{v_{1,L}\in V} p(v_{1,L} \mid w_1)p(v_{1,L} \mid w_2)}{\sqrt{\sum_{v_{1,L}\in V} p(v_{1,L} \mid w_1) \sum_{v_{1,L}\in V} p(v_{1,L} \mid w_2)}}. \tag{2.39}$$

The symmetric (left and right) contextual similarity between words w_1 and w_2 is computed as Pargellis et al. [2004]:

$$S_{CS}^{L,R}(w_1, w_2) = S_{CS}^{L}(w_1, w_2) + S_{CS}^{R}(w_1, w_2). \tag{2.40}$$

Each of both terms of Equation 2.40 has a lower and upper bound of zero and one, respectively. Thus, two words of identical contextual distributions will have similarity score equal to two.

Note that the incorporation of contextual probability distributions in the aforementioned

metrics is not limited to the case of bigram probabilities, i.e., higher-order n-gram probabilities may be used. For example, in Pargellis et al. [2004] both bigram and trigram probabilities were employed for the representation of the contextual distributions. A comparison of the presented context-based metrics were presented in several works, e.g., Bullinaria and Levy [2007]; Pargellis et al. [2001, 2004]. Certainly it is hard to draw generic conclusions about the relative performance of the several similarity metrics due to factors that vary across different studies, e.g., corpora, implementation of VSM, evaluation tasks, etc. However, in the aforementioned studies the cosine similarity was reported to be among the best performing metrics.

Chapter 3

DSMs I: Semantic Similarity Computation Using Web Documents

3.1 Introduction

Numerous information retrieval and natural language processing applications require knowledge of semantic similarity between words or terms. For example, by adding semantically similar words to a web query (query expansion) it is likely to increase the relevance[1] of retrieved documents Gauch and Wang [1997]. Moreover, semantic similarity measures are used in many natural language processing (NLP) tasks, such as language modeling Fosler-Lussier and Kuo [2001], grammar induction Siu and Meng [1999], word sense disambiguation Dagan et al. [1997], speech understanding and spoken dialogue systems Fosler-Lussier and Kuo [2001]. In Iosif et al. [2006]; Pargellis et al. [2004], several unsupervised statistical metrics are presented and applied to the automatic induction of semantic classes for both semantically homogeneous and heterogeneous corpora.

The majority of the semantic similarity metrics employed today use hand-crafted language resources Jiang and Conrath [1997]; Leacock and Chodorow [1998]; Li et al. [2003]; Petrakis et al. [2006]. The use and updating of resources, such as thesauri or ontologies, is a time consuming and tedious task, demanding human labor and often expert knowledge. Also, language resources are not ubiquitous and are unavailable for many languages. As a result, such methods are of little utility for applications where human and language resources are sparse. In addition, these methods cannot be applied for words or terms that are not included in the resource repository, e.g., scientific terms, out of vocabulary words, neologisms. To overcome

[1] In Flank [1998]; Mihalcea and Moldovan [2000]; Voorhees [1994], it is shown that query expansion using related words acquired from WordNet increases the recall of retrieved documents.

this problem knowledge resources are often constructed for specific domains where general-purpose ontologies do not offer adequate term coverage. For example, in addition to WordNet, domain-specific ontologies, e.g., MeSH, are used for applications in the (bio)medical domain Petrakis et al. [2006]. Improving term coverage remains an open research issue; algorithms are proposed in the literature on how to pool multiple knowledge resources or add terms to existing language resources, e.g., ontology merging techniques and cross-ontology similarity metrics. The work of Budanitsky and Hirst Budanitsky and Hirst [2006] provides a thorough review of different metrics that use the WordNet resource for computing semantic similarity.

The web has a multilingual character; new words, neologisms and occasionalisms (hapax legomena), are added frequently and efficiently. Thus, it is the obvious place for mining semantic relationships for unseen words. Also, the web contains both general-purpose words, found in news articles and blogs, as well as, scientific terminology, found in documents written by experts. Overall, the web covers a plethora of domains, authoring styles and languages, and is fertile ground for automatic semantic knowledge acquisition. The web has been exploited for a variety of NLP applications. In Zhu and Rosenfeld [2001], web page counts returned by a search engine were used to estimate the probability of n-gram language models. In Dekang et al. [2003], the web page counts of fixed lexical patterns were used to identify synonymy and antonymy between nouns. An extension of this approach was proposed in Chklovski and Pantel [2004]; web queries of lexico-syntactic patterns were used for discovering relationships between verbs. The web is also an invaluable source for constructing text corpora. For example, in Terra and Clarke [2003], a large corpus of web pages was constructed and used for word sense disambiguation. Other applications where automatically-constructed web corpora have been used to train statistical models include machine translation Popovic and Ney [2005] and question-answering systems Dumais et al. [2002].

Recently there has been much research interest in developing web-based similarity measures. Typically such approaches use the results returned by one or more web search engines using one or multiple queries. Web-based similarity measures can be broadly divided into three categories: (i) measures that rely only on the number of the returned hits, (ii) measures that download a number of the top-ranked documents and then apply text processing techniques, (iii) measures that combine both approaches. Web-based similarity computation algorithms have been used in a diverse range of applications, such as automatic annotation of web pages Cimiano et al. [2004], social networks construction Mika [2005]; Mori et al. [2006] and music genre classification Geleijnse and Korst [2006]; Schedl et al. [2006]. However, in most cases, the form of the web query and/or the feature extraction process is application-dependent, e.g., if one is interested in movie genre classification it is useful to include the term “movie” in the submitted query.

In this chapter, we focus on the problem of fully unsupervised web-based semantic similarity computation between words or terms; no hand-crafted rules or resources are employed. Web search engines are used for text corpus mining and context-based similarity distances are automatically computed on this corpus. The proposed algorithm requires no expert knowledge or language resources and, as a result, it can be readily applied to different languages. In order to calculate the semantic similarity between words, we investigate two families of unsupervised, web-based similarity metrics . The first type considers only the number of hits returned by a web search engine, as in Bollegala et al. [2007] and Gracia et al. [2006]. The second is fully text-based, downloads the top-ranked documents returned by a web query and compares the context around words of interest to estimate semantic similarity. The following are the original contributions of this work:

1. Several contextual similarity algorithms are proposed and evaluated over large collections of downloaded documents.

2. The metrics are evaluated both on the Miller-Charles and on a medical term dataset, i.e., in this work we investigate both word and term similarity. The two evaluation domains are also semantically different: ordinary words of general use vs medical terms.

3. We demonstrate the effect of feature and document selection on semantic similarity computation. For example, it is shown that non-content words (stop-words) are important features for word similarity computation but poor features for term similarity computation.

4. We show that the proposed fully unsupervised method based on context similarity can compete with state-of-the-art supervised similarity metrics that employ elaborate language resources.

The remainder of this chapter is organized as follows. In Section 3.2, an overview of related work in the area of semantic similarity computation is presented. In Section 3.3, the semantic similarity computation algorithm is described along with the experimental procedure. In Section 3.4, the evaluation results are reported for the proposed algorithms for two evaluation datasets. The results are compared with state-of-the-art semantic similarity algorithms that employ knowledge resources such as WordNet and MeSH. The results are further discussed in Section 3.5, and implications of feature selection and document selection for context-based similarity metrics are presented. Finally, we conclude with Section 3.6, where promising directions for further research are also mentioned.

3.2 Related work

Metrics that measure semantic similarity between words or terms can be classified into four main categories depending if knowledge resources are used or not: (i) supervised *resource-based metrics*, consulting only human-built knowledge resources, such as ontologies, (ii) supervised *knowledge-rich text-mining metrics*, i.e., metrics that perform text mining but also rely on knowledge resources, (iii) unsupervised *co-occurrence metrics*, i.e., unsupervised metrics that assume that the semantic similarity between words or terms can be expressed by an association ratio which is a function of their co-occurrence and (iv) unsupervised *text-based metrics*, i.e., metrics that are fully text-based and exploit the context or proximity of words or terms to compute semantic similarity. The last two groups of metrics do not use any language resources or expert knowledge and depend only on web search engines. In this sense, these metrics are referred to as "unsupervised"; no semantically labeled human-annotated data is required to compute the semantic distance between words or terms. Resource-based and knowledge-rich text-mining metrics, however, use such data, and are henceforth referred to as "supervised" metrics.

Several resource-based methods have been proposed in the literature that use, e.g., WordNet, for semantic similarity computation. Edge counting methods consider the length of the paths that link the words, as well as the word positions in the taxonomic structure Leacock and Chodorow [1998]; Li et al. [2003]. Information content methods compute similarity between words by combining taxonomic features that exist in the used resource, e.g., number of subsumed words, with frequencies computed over textual corpora Jiang and Conrath [1997]. Hybrid methods combine synsets[1] with word neighborhoods and other features Petrakis et al. [2006]. In the work of Bollegala et al. Bollegala et al. [2007], a hybrid method, among others, is defined that combines page counts, returned by a search engine, and lexico-syntactic patterns, extracted from the returned snippets using a number of synonymous nouns acquired from WordNet.

Co-occurrence-based metrics attempt to implement computational models for the notion of "word association" which is used in psycholinguistics. This notion describes the procedure of lexical decision of human associative memory. In Church and Hanks [1990], an association ratio is proposed using the information theoretic metric of mutual information in order to identify patterns which can be used for the construction of semantic classes. In Bollegala et al. [2007], several association metrics are applied, using a search engine in order to obtain co-occurrence counts for a word pair. If the pair of interest consists of the words w_1 and w_2,

[1] A synset is a set of words (or terms) that are considered to be synonymous. This notion is widely used in lexical resources like WordNet.

their co-occurrence frequency is taken to be equal to the number of hits returned by a search engine, given a query of the form "w_1 AND w_2".

Text-based metrics typically use contextual features to compute semantic similarity. Context-based metrics operate under the assumption that words with similar contexts have similar meaning. One of the first studies of this hypothesis is the work of Rubenstein and Goodenough stating that "words that are similar in meaning occur in similar contexts" Rubenstein and Goodenough [1965]. Using this assumption, the semantic similarity between two words can be estimated by measuring the difference between the probability distributions of their contextual features. Various context-based metrics have been proposed in the literature, such as: Kullback-Leibler, information radius and Manhattan norm Pargellis et al. [2004]; Siu and Meng [1999]. The contextual probability distributions can be estimated (and smoothed) using n-gram language models Jelinek [1998]. Another representation of the contextual environment of a word is the *bag-of-words* model Lewis [1998]. According to this model, the contextual features of a word form the elements of a vector. Assuming independence among the features, the similarity of two words is computed as the product of their feature vectors using cosine similarity Iosif et al. [2006]; Pangos et al. [2005]. More recently, context-based similarity metrics construct document collections by querying web search engines and downloading a number of the returned top-ranked documents, in order to compute semantic similarity between words or terms Iosif and Potamianos [2007a].

Scheme	Acronym	$t_{w,i}$ (if $c(v_i) > 0$)
Binary	B	1
Term frequency	TF	$\frac{c(v_i)}{c(w)}$
Add-one TF	TF1	$\frac{c(v_i)+1}{c(w)+\alpha_w}$
Log of TF	LTF	$\frac{\log(c(v_i))}{\log(c(w))}$
Add-one LTF	LTF1	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha_w)}$
TF-inverse document freq.	TFIDF	$\frac{c(v_i)}{c(w)} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i \rvert}$
Log of TFIDF	LTFIDF	$\frac{\log(c(v_i))}{\log(c(w))} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i \rvert}$
Add-one LTFIDF	LTF1IDF	$\frac{\log(c(v_i)+1)}{\log(c(w)+\alpha_w)} \log \frac{\lvert D \rvert}{\lvert D \rvert v_i \rvert}$

Table 3.1: Context Feature Weighting Schemes

The various feature weighting schemes used in this work for computing the value of $t_{w,i}$ are presented in Table 3.1. The weighting schemes can be classified into binary and frequency-

based. The binary metric assigns weight $t_{w,i} = 1$ when the i^{th} word in the vocabulary exists at the left or right context of at least one instance of the word w, and 0 otherwise. Frequency-based weighting schemes compute the (normalized) frequency of occurrence of context words. Various frequency-based weighting schemes popular in natural language processing and web applications are proposed and evaluated, specifically, term-frequency (TF), logarithmic term-frequency (LTF), term frequency inverse document frequency (TFIDF), logarithmic TFIDF, and add-one smoothing of these methods. As shown in Table 3.1, for frequency-based metrics the value of $t_{w,i}$ is computed as a function of the counts $c(v_i)$, i.e., the number of occurrences of the i^{th} vocabulary word v_i within the left or right context of all occurrences (in a corpus) of term w. Note that the counts $c(v_i)$ are normalized by $c(w)$, the number of occurrences of word w in the corpus. For the case of add-one term frequency (TF1), the probability of occurrence is smoothed by adding one to the counts and normalizing them by $c(w) + \alpha_w$, where α_w is defined as the total number of unique words that appear in the context(s) of w.

Logarithmic term frequency (LTF) weighting is similar to term frequency (TF), the main difference being the non-linear scaling of counts and the assignment of weight 0 to singletons, i.e., context words appearing only once. By applying the logarithmic weighting scheme, the highly frequent contextual features are not allowed to dominate the computation of similarity score (unlike the linear term frequency weighting scheme). Also, the non-linearity introduced by the logarithmic scheme could be a simple way to approach the non-linear process by which the human memory builds the semantic associations between words (assuming that the contextual features are taken into account during the cognitive process). Logarithmic add-one smoothing (LTF1) takes singletons into account with a positive weight of $\frac{\log(2)}{\log(c(w)+\alpha_w)}$, a straightforward generalization of TF1.

The term-frequency inverse document frequency (TFIDF) metric is a popular metric in information retrieval that assigns more weight to semantically salient words, effectively reducing the effect of stop words and non-content words. Similarly, in this work, the logarithm of the inverse document frequency of context words in each document is computed as $\log \frac{|D|}{|D|v_i|}$ and used to multiply the TF estimate. Note that $|\, D|v_i\,|$ denotes all documents indexed by context word v_i. Similarly the logarithmic TFIDF (LTFIDF) multiplies the LTF estimate with the inverse document frequency. Finally, the add-one smoothing version of this metric is computed (LTF1IDF).

3.3 Corpus based similarity computation

We experimented with (i) page-count-based, and (ii) text-based similarity metrics, described in Chapter 2. For the page-count metrics the Yahoo! search engine was used to determine

the frequency occurrence and co-occurrence of words or terms w_1 and w_2. Specifically, the total number of hits for the queries "w_1", "w_2" and "w_1 AND w_2" were used to compute the Jaccard, Dice, Mutual Information and Google metrics.

For the contextual similarity metrics, for each pair of words or terms (w_1, w_2) a few hundred documents were downloaded using "w_1 AND w_2" (e.g., "boy AND lad") queries. The *URLs* for the top ranked documents were retrieved using the Yahoo! search engine via the Yahoo! Search API. AND queries retrieve documents containing both terms, as opposed to generic "w_1 OR w_2" queries that download documents containing either term. In Iosif and Potamianos [2007a], preliminary experiments have shown that AND queries significantly outperform OR queries for context-based semantic similarity computation. Once the documents are downloaded, the left and right contexts of all occurrences of w_1 and w_2 are examined and the corresponding feature vectors are constructed according to the following experimental parameters:

1. Number of web documents, $|D|$: how many web documents are used.

2. Contextual window size, K: the left and right contexts of w_1 and w_2 are examined according to the value of contextual window size. The window size is applied within the sentence boundaries.

3. Stop words filtering (yes/no): consideration (or not) of stop words in the feature vectors.

4. Type of weighting scheme: the values of vector features are set according to one of the weighting schemes presented in Table II.

The semantic similarity between words or terms w_1, w_2 is then computed as the cosine similarity (see (2.20)) of their corresponding contextual feature vectors following the unstructured approach described in Section 2.2.1.1.

3.4 Evaluation

In this section, we present a comparative evaluation of the similarity algorithms, in terms of correlation, with respect to the human ratings of: (i) the Miller-Charles dataset of common words, and (ii) the MeSH dataset of medical terms. Both the page-count-based similarity metrics are evaluated, as well as, the fully text-based similarity algorithms. The proposed algorithms are also compared with metrics that use knowledge resources, e.g., the WordNet ontology for the Miller-Charles dataset, and the MeSH ontology for the MeSH dataset of medical terms.

3.4.1 Corpus description

For evaluation purposes we used two datasets: (i) the Miller-Charles dataset [1] of common nouns Miller and Charles [1998], and (ii) a dataset of medical terms included in the MeSH ontology. The first dataset consists of 28 noun pairs of general use that were rated according to their semantic similarity by 38 human subjects. The assigned similarity scores range from 0 (not similar) to 4 (perfect synonymy). The selection of this dataset was mainly motivated by its wide use that enabled us to compare our work with a variety of other approaches.

The MeSH dataset includes 34 medical terms pairs that have been rated for similarity by experts. MeSH is the acronym for "Medical Subject Headings" and is a taxonomic hierarchy containing medical terms proposed by the National Library of Medicine, USA. The MeSH dataset contains 36 pairs of MeSH terms rated by human experts, e.g., "asthma-pneumonia" and "anemia-appendicitis". In this work, a subset of 34 pairs was used due to the limited amount of web documents available for the 5^{th} and 36^{th} pair. The MeSH dataset along with the human-rated similarity scores were taken from the work of Petrakis et al. Petrakis et al. [2006]. Petrakis et al. asked Dr. Qi at Dalhousie University to construct a set of MeSH term pairs. Then medical experts were asked to submit similarity scores for the MeSH term pairs using a web-based tool. In total, 8 experts took part in the above procedure assigning similarity scores from 0 (no similarity) to 4 (absolute similarity). Pairs with standard deviation of similarity scores higher than the user defined threshold of $t = 0.8$ were excluded from the evaluation [2]. The MeSH dataset was selected in order to investigate similarity between terms that are rated by experts rather than naive subjects.

3.4.2 Evaluation metric

Let $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ be the random vectors that store the similarity scores given by human subjects and the computational metric, respectively, for each of the $i = 1, 2, ..., n$ word pairs. The correlation coefficient between the scores produced by humans and machine is estimated using the Pearson correlation, as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

[1] The experimentation with the relatively small Miller-Charles (MC) dataset took place in the very early phases of the work described in this thesis. Then, we were mainly interested about the performance of similarity metrics for common words and technical terms. Later, we became aware of the other two standard English datasets: RG Rubenstein and Goodenough [1965], and WS353 Finkelstein et al. [2002]. Given the fact that the pairs of the MC dataset are included within RG and WS353, we expect that the main conclusions of this chapter to also apply for these two datasets.

[2] For more details see
http://www.intelligence.tuc.gr/similarity/datasets/MeSHDataset.pdf.

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, for $i = 1, 2, ..., n$.

3.4.3 Evaluation of page-count-based metrics

The correlation scores between the page-count-based semantic similarity metrics and human ratings are presented in Table 3.2 for the two tasks. The similarity metrics based on the Jaccard

Dataset	J	C	I	G
Miller-Charles	0.41	0.41	0.69	0.66
MeSH	0.26	0.29	0.30	0.41

Table 3.2: Correlation of page-count metrics.

(J) and Dice (C) coefficients achieve comparable correlation performance, which is expected given the similarities between the two metrics. The Mutual Information (I) and Google-based Semantic Relatedness (G) achieve significantly[1] better performance than Jaccard and Dice, especially for the Miller-Charles dataset. Overall, the achieved correlation for the words of general use (Miller-Charles dataset) is significantly higher than that of the medical terms (MeSH dataset), for all metrics.

3.4.4 Evaluation of context-based metrics

Next we present the performance of the context-based metrics for the various feature weighting schemes shown in Table 3.1 and for different contextual window sizes K. The performance of each metric is shown as a function of the number of downloaded documents. The correlation scores for the Miller-Charles dataset are shown in Fig. 3.1(a) and (b), and for the MeSH dataset in Fig. 3.1(c) and (d).

In Fig. 3.1(a), the correlation scores for the Miller-Charles dataset are shown using several weighting schemes. Performance is shown as a function of the context window K (ranging from 1 to 20) for a total number of 100 downloaded documents. For most metrics, highest correlation is achieved with context size $K = 1$, i.e., considering only the immediate context of one word to the left and one to the right. For larger context windows, performance degrades fast especially for the TFIDF weighting schemes. The highest correlation score of 0.72 is achieved by the LTF scheme with the binary weighting scheme being a close second. Note that the linear frequency-based weighting schemes, i.e., TF and TFIDF, perform poorly, compared to their logarithmic counterparts, especially, for large context sizes.

[1] When comparing the performance of similarity metrics in this work, the term "significantly better" is used to indicate statistical significance at a level higher than 95% using the paired t-test.

In Fig. 3.1(b), the performance of the binary (B), LTF and LTFIDF weighting schemes are shown for a context window size of $H = 1$ as a function of the number of downloaded documents (ranging from 10 to 1000). The correlation improves with the number of documents and the performance bound is not reached even at 1000 documents. Good correlation performance is achieved with as few as 30 documents, however, it is clear that the performance of the similarity metric is not robust if fewer than 100 documents are used. Overall, the LTF scheme performs best up to approximately 500 documents, while the binary scheme provides better performance for a larger number of documents. Also note, that the performance gap between LTF and LTFIDF is bridged for a large number of documents. Overall, the highest correlation score of 0.88 is achieved using the binary weighting scheme and 1000 documents.

In Fig. 3.1(c), the correlation score for the MeSH dataset is shown. The weighting schemes, context window size and number of documents (100) are the same as in (a), and thus the two plots are directly comparable. The main differences in performance for the MeSH dataset compared to the Miller-Charles dataset are: (i) the relative performance of the weighting schemes, i.e., for the MeSH dataset the LTFIDF weighting scheme significantly outperforms all other schemes (note the especially poor performance of the binary weighting scheme), and (ii) the optimum context window size, i.e., for the MeSH dataset best correlation scores are achieved for context window size between $H = 2$ and $H = 5$, as opposed to $H = 1$ for the Miller-Charles. In addition, the degradation of performance for large context windows is much more graceful for the MeSH dataset. The best correlation score is 0.67 for $H = 3$ and the LTFIDF weighting scheme.

In Fig. 3.1(d), the performance of the binary (B), LTF, and LTFIDF weighting schemes are shown as a function of the number of downloaded documents. The window size used is $H = 1$ in order for plots (b) and (d) to be directly comparable[1]. As in (b), correlation increases as more documents are considered. However, in (d), the highest score is achieved for 800 documents and the performance degrades somewhat for 1000 documents. The LTFIDF weighting scheme significantly outperforms the other two metrics, while the binary scheme performs the worst, i.e., the relative metric performance is reversed in (d) compared to (b). Finally, note that the absolute performance of the metrics for the MeSH dataset is worse than for the Miller-Charles dataset; this is consistent with results reported in the literature.

Add-one smoothing schemes LTF1 and LTF1IDF that do not discard contextual singletons achieve almost identical correlation scores to LTF and LTFIDF respectively. Thus, the results for LTF1 and LTF1IDF are not included in the plots.

[1] Although for 100 documents the best correlation score is obtained for $H = 3$, for large number of documents comparable performance is obtained for context window sizes of one, two or three.

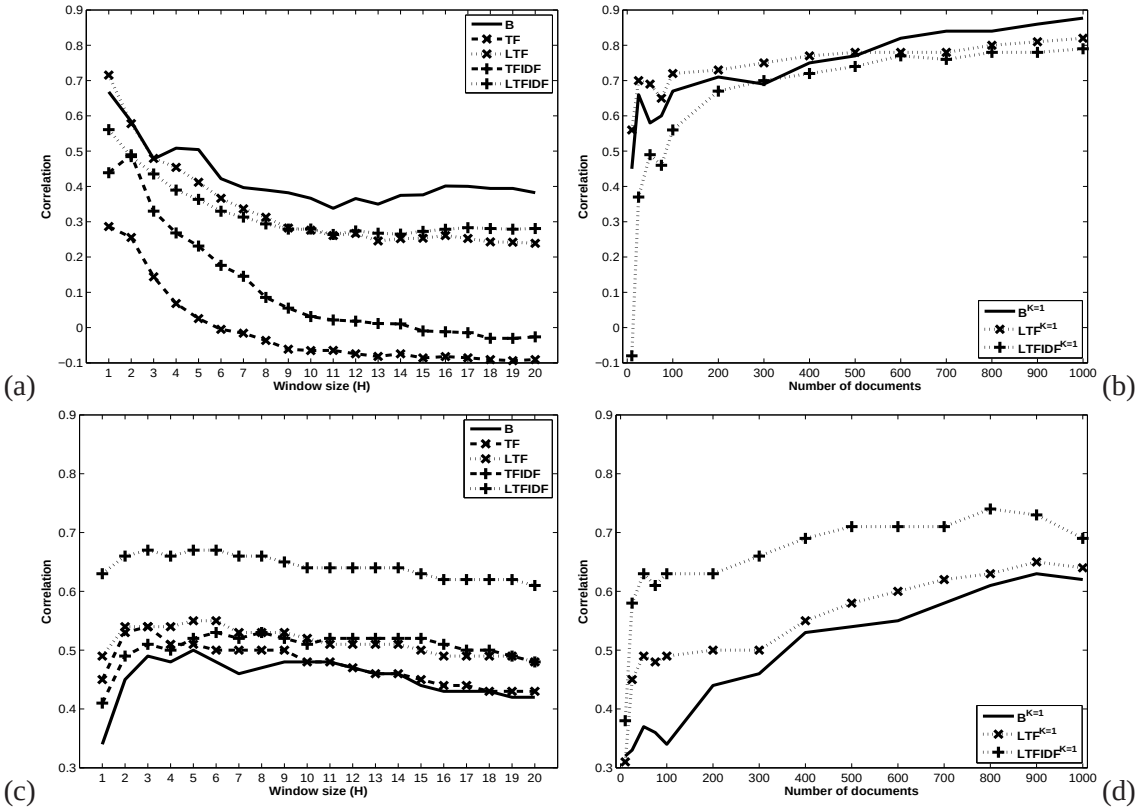


Figure 3.1: Correlation scores between context-based similarity computation and human ratings for: (a),(b) the Miller-Charles dataset, and (c),(d) the MeSH dataset. Performance of the various weighting schemes as a function of context window size is shown in (a),(c) for 100 documents. Performance as a function of number of documents is shown in (b),(d) for $H = 1$.

3.4.5 Stop-word filtering

Motivated by the differences in performance between the term weighting schemes for the word and term tasks, we investigate next how stop-word filtering affects performance. For this purpose we classified the contextual words into stop-words (sw) and non-stop-words and computed the semantic similarity scores for the three possible setups: (i) only stop-words are considered in the similarity computation algorithm, (ii) any word not being a stop-word is considered, and (iii) all words are considered (same setup as that used for Fig. 3.1 (a),(c)). The correlation score was computed for 100 documents using context window size $H = 1$. For each dataset the best weighting scheme was used, i.e., LTF for Miller-Charles and LTFIDF for MeSH. The results are shown in Table 3.3.

For the Miller-Charles dataset, the inclusion of stop-words boosts overall performance, in fact, similarity computation using only stop-words as features outperforms somewhat similarity

Dataset	Type of context		
	only stop-words (sw)	w/o sw	both
Miller-Charles	0.68	0.64	0.72
MeSH	0.25	0.66	0.63

Table 3.3: Correlation for different types of context.

computation using only non stop-words! For the MeSH terms dataset, however, stop-word-based similarity computation performs very poorly. In fact, including stop-words seems to be hurting overall performance; from 0.66 when stop-words are excluded to 0.63. For a more detailed discussion on stop-word filtering and feature selection see Section 3.5.

3.4.6 Unsupervised vs supervised metrics

Next the performance of the proposed unsupervised algorithms is compared with semantic similarity computation algorithms found in the literature. In addition to page-count similarity metrics, we also consider metrics that consult knowledge resources, i.e., supervised similarity computation algorithms. The metrics considered here, along with the main characteristics of each metric, are summarized in Table 3.4 and Table 3.5, for the Miller-Charles and MeSH datasets, respectively.

The Li Li et al. [2003], Jiang Jiang and Conrath [1997], X-Similarity Petrakis et al. [2006], and Leacock-Chodorow Leacock and Chodorow [1998] metrics exploit the semantic hierarchical structure of ontologies, WordNet or MeSH, to compute semantic similarity as described in Section 3.2. The correlation scores for these metrics can be found in Petrakis et al. [2006]. For the Miller-Charles dataset, correlation scores of 0.82, 0.83 and 0.74, were reported for the Li, Jiang and X-Similarity metrics, respectively. For the MeSH dataset, the following correlation scores were reported: 0.70 (Li), 0.71 (Jiang), 0.74 (Leacok-Chodorow) and 0.71 (X-Similarity).

The performance of the web-based metrics is summarized as follows. For the Miller-Charles dataset, the resource-based SemSim metric proposed in Bollegala et al. [2007], achieves a correlation score of 0.83 that is similar to the ontology-based methods above. The fully unsupervised Sahami Sahami and Heilman [2006] metric is shown to have a moderate correlation of 0.58 (results are reproduced from the implementation and evaluation in Bollegala et al. [2007]). Moderate correlation scores are achieved also by the metrics that consider only the page counts returned by a query, especially for mutual information and Google. The unsupervised context-based metric using the binary weighting scheme and context window $H = 1$ achieves the highest correlation (0.88) among the unsupervised metrics for 1000 documents. Note that the

Metric	Use of ($\surd$: yes, X: no)						Need of external knowledge	Correlation
	WWW Search engine	Page counts	Snippets	Lexico-Syntactic patterns	WordNet ontology	Down. docs		
Jaccard (J)	$\surd$	$\surd$	X	X	X	X	X	0.41
Dice (C)	$\surd$	$\surd$	X	X	X	X	X	0.41
Mutual info. (I)	$\surd$	$\surd$	X	X	X	X	X	0.69
Google-based sem. relat. (G)	$\surd$	$\surd$	X	X	X	X	X	0.66
Sahami	$\surd$	X	$\surd$	X	X	X	X	0.58
SemSim	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	X	$\surd$	0.83
Li	X	X	X	X	$\surd$	X	$\surd$	0.82
Jiang	X	X	X	X	$\surd$	X	$\surd$	0.83
X-Similarity	X	X	X	X	$\surd$	X	$\surd$	0.74
Proposed $Q^{H=1}$ (B: 1000 docs)	$\surd$	X	X	X	X	$\surd$	X	0.88

Table 3.4: Properties and performance of similarity metrics for the Miller-Charles dataset.

performance of the context-based metrics is comparable to that of the resource-based metrics for semantic similarity computation between words. In fact, the reported correlation score of 0.88 is among the highest reported in the literature[1] for this dataset.

For the MeSH dataset, all page-count-based metrics have poor results. The best correlation (0.69) among the unsupervised metrics for 1000 documents is obtained by the context-based metric with window $Q^{H=1}$ using the LTFIDF scheme. The performance of the context-based LTFIDF metric is worse but comparable to that of the supervised methods.

3.5 Discussion

In this section, the evaluation results are further analyzed and explained. Specifically, we investigate the performance of the supervised and unsupervised semantic similarity computation algorithms and explain the difference in performance for words and terms. Issues such as feature selection and document selection are addressed.

3.5.1 Corpus creation and document selection

A shortcoming of web-based methods for similarity computation is that, as far as the algorithm is concerned, the search engine is a "black box". This is especially relevant for page-count based metrics, such as Jaccard and Google, where the number of returned hits is very much

[1]The highest (to our knowledge) reported correlation score for the Miller-Charles dataset is equal to 0.89 Li et al. [2003]. The proposed algorithm exploits the shortest path length and depth between the words of interest in the WordNet hierarchy.

Metric	Use of ($\surd$: yes, X: no)				Need of external knowledge	Correlation
	WWW Search engine	Page counts	MeSH ontology	Down. docs		
Jaccard (J)	$\surd$	$\surd$	X	X	X	0.26
Dice (C)	$\surd$	$\surd$	X	X	X	0.29
Mutual information (I)	$\surd$	$\surd$	X	X	X	0.30
Google-based sem. relat. (G)	$\surd$	$\surd$	X	X	X	0.41 0.41
Li	X	X	$\surd$	X	$\surd$	0.70
Jiang	X	X	$\surd$	X	$\surd$	0.71
LeacokChodorow	X	X	$\surd$	X	$\surd$	0.74
X-Similarity	X	X	$\surd$	X	$\surd$	0.71
Proposed $Q^{H=1}$ (LTFIDF: 1000 docs)	$\surd$	X	X	$\surd$	X	0.69

Table 3.5: Properties and performance of similarity metrics for the MeSH dataset.

search engine dependent and changes over time. For the context-based approach, the assumption is that a search engine is a reliable provider of representative examples of language usage. Although this is a reasonable assumption, the relative ranking of documents returned by a search engine might affect the algorithm's performance, given that only the top ranking documents are downloaded. Another factor that affects performance is the type of web query used, as well as, the way the query is "interpreted" by the search engine.

In this work, we use a conjunction query to search for documents in which the words or terms of interest co-exist, i.e., "w_1 AND w_2". We have also noted that using a conjunction query works much better in practice than using a disjunction, i.e., "w_1 OR w_2" Iosif and Potamianos [2007a]. There are two possible explanations for the significantly better performance of corpora created using AND vs OR queries. First, co-occurrence is by itself a feature used in semantic similarity computation, e.g., page-count based similarity. Second, the created corpus is semantically more homogeneous and stylistically more consistent. Specifically, by examining occurrences of w_1 and w_2 in the same document, the topic and authoring style are the same for the context of both words. In Pangos et al. [2005], it was shown that context-based similarity metrics work much better in semantically homogeneous domains, e.g., travel reservation, than in semantically broad domains, e.g., news. Similar observations have been made for unsupervised word sense disambiguation algorithms that also employ context-based metrics, specifically, "the sense of a target word is highly consistent within any given document" Yarowsky [1995].

As far as the relative ranking of documents by the search engine is concerned we have not observed any statistical significant effect on the performance of the context-based metrics.

Specifically, we have tested the performance of the algorithm on document deciles, i.e., documents ranked 1-100, 101-200, up to 901-1000. We have found no significant effect of rank in any of these experiments, either for the word or the term task. More research is necessary (e.g., bottom ranked documents) to verify that indeed search engine ranking does not affect context-based semantic similarity performance.

During the application of a context-based similarity metrics over the collection of downloaded documents, we assumed that the lexical features of each document have the same importance (or weight) in the similarity computation formula. In practice, however, documents are different in many ways, e.g., authoring style, author's expertise, balance between graphical and textual content. It is not uncommon in web-based applications to assert the "quality" of a document and exclude (or weight less) low quality documents. In our case, we experimented with a variety of "grammaticality" metrics[1] in order to establish the quality of a document. The following metrics were used to compute document grammaticality: (i) the average number of words in a paragraph, assuming that a document consisting of paragraphs of larger size is of "higher quality", (ii) the fraction of document vocabulary that is included in the document compared with a Wall Street Journal corpus, i.e., selecting documents with a more formal way of writing, and (iii) the perplexity of the text in the document computed using n-gram language model built from a Wall Street Journal; this feature is looking for documents with richer vocabulary and more complex syntax. The computed metrics were then used to weigh the contribution of the features extracted from each document. None of the proposed algorithms provided consistent performance improvement compared to the baseline results. This is an indication that the performance of context-based similarity metrics is not affected much by document writing style or document "quality".

3.5.2 Feature selection for word and term similarity

The evaluation results showed that co-occurrence (page-count-based metrics) at document level can provide only rough estimates of semantic similarity. This trend is more pronounced for the specialized medical terms of the MeSH dataset. Context-based metrics achieved higher correlation scores compared to page-count-based metrics for both tasks. Overall, context seems to be the most important feature for semantic similarity computation, followed by co-occurrence. Moreover, evaluation results showed that performance improves as the number of downloaded documents increases, which is in agreement with the statement of Schütze and Pedersen Schütze and Pedersen [1995] "words with similar meanings will occur with similar neighbors if enough text material is available". Although it is clear that contextual similarity

[1]The notion of grammaticality is used here in a broad sense rather than the exact linguistic sense of conforming to a syntactic grammar.

implies semantic similarity, the amount of context to take into account in this process, as well as, the relative weighting of the contextual features needs further investigation and is discussed next.

We have investigated various aspects of feature selection for context-based similarity metrics, namely, context window size, the use of stop-words and the relative weighting of context words. Specifically, we found that when using the very immediate context (window size one) better performance was achieved for the Miller-Charles dataset, while a context window size between two and five words was optimal for the MeSH dataset. In addition, stop-words were valuable features for the Miller-Charles dataset, but provided little or no information for the MeSH dataset. Finally, term frequency (TF)-based feature weighting provided good results for the Miller-Charles dataset, while term-frequency inverse document frequency (TFIDF)-based weighting provided best results for the MeSH dataset. In essence, optimal feature selection was quite different for word and term similarity computation.

Putting together the observations from these experiments one may draw general conclusions about feature selection for context-based similarity computation between common words or between specialized terms. Note that the immediate context of nouns, which consists mainly of stop-words and very frequent contextual features, encodes syntactic dependencies. Stop-words mainly include various function words, such as articles and conjunctions, which are fragments of local (within sentences) syntactic patterns in which the target words participate. Longer context, that consists mainly of content words and features with high TFIDF weights, encodes mostly semantic dependencies. Thus for common nouns, syntax seems to be the most salient feature, while for terms, semantics are more important. More research is necessary to better understand how to tune the feature selection process for specific domains, as well as, how to better combine different types of features, e.g., fusion of syntax and semantic-based features. Note, however, that the generic feature selection and weighting algorithms presented in this work for word and term semantic similarity computation already provide good baseline performance. Also preliminary experiments indicate that the proposed algorithms perform well for other languages, e.g., Greek.

A final note on the comparison between the supervised resource-based and unsupervised context-based semantic similarity computation algorithms. In this work, we have shown for the first time that unsupervised metrics achieve comparable performance to supervised resource-based ones. Comparing, however, the best results of supervised and unsupervised algorithms should be done with care, as in both cases, there is a long list of parameters that are being "tuned" for the specific dataset, i.e., there is a danger of model overfiting. Extensive experiments on additional datasets, as well as, optimization of parameters on held-out data is required in order to draw general conclusions about the detailed performance of the algorithms. Inde-

pendent of their relative performance, however, the proposed unsupervised algorithms should prove a valuable tool for populating existing ontologies with new members[1], as well as, create ontologies for new languages. Finally, we believe that the development of computational similarity metrics can serve as an additional research tool in the field of human cognition and language acquisition research.

3.6 Conclusions

We presented and compared two families of unsupervised, web-based metrics for semantic similarity computation between words, namely, page-count and context-based metrics. Page-count metrics consider only hits returned by a search engine, while the proposed context-based semantic similarity algorithms download the top ranked documents returned by a web query and compute the frequency of occurrence of contextual features. The proposed algorithms do not consult any external knowledge resource and can be generalized and applied to other languages. The performance of the unsupervised algorithms was evaluated and compared with resource-based semantic similarity computation algorithms on the Miller-Charles dataset and the MeSH dataset of medical terms.

The page-count-based metrics produced low to mid correlation with human semantic similarity scores. Good correlation scores were obtained using the context-based metrics, achieving performance of up to 0.88 and 0.74 for the Miller-Charles and MeSH datasets, respectively. The performance achieved is comparable to that of supervised resource-based semantic similarity computation algorithms. The following conclusions can be drawn for the performance of unsupervised similarity computation algorithms: (i) context is a better feature for semantic similarity computation than co-occurrence considered at document level, (ii) for the Miller-Charles dataset best results are obtained for a contextual window size of one, including stop-words as features and the LTF or binary weighting schemes, (iii) for the MeSH dataset best results are obtained for a contextual window size of two to five, excluding stop-words as features and the LTFIDF feature weighting scheme, (iv) logarithmic weighting of contextual feature outperforms linear weighting for both tasks, and, (v) performance of context-based metrics improves as the number of documents increases (with the exception of the last two data-points for the MeSH dataset). Preliminary experiments on document selection did not show significant correlation with performance. Overall, the proposed context-based algorithm provides good performance, is fully automatic, requires little computation-power and small to

[1]Note that two out of the 30 noun-pairs in the Miller-Charles dataset Miller and Charles [1998] were not included in the original versions of WordNet forcing researchers to evaluate on a 28 pair subset Bollegala et al. [2007].

medium amounts of web text, and can be generalized and applied to other languages. In order to use the proposed algorithms in practice for ontology creation, one may use a combination of page-count and contextual metrics, i.e., use page-count metrics to identify candidates and contextual metrics to refine the similarity scores.

This work is a first step towards our understanding of the potential of context-based metrics for semantic similarity computation. A variety of issues related to document selection, feature selection and feature fusion have to be further investigated. In addition, a better understanding of acquisition of semantics by humans could lead to improved semantic similarity computation algorithms.

Chapter 4

DSMs II: Similarity Computation Using Semantic Networks

4.1 Introduction

Semantic similarity is the building block for numerous applications of natural language processing (NLP), such as grammar induction Meng and Siu [2002] and affective text categorization Malandrakis et al. [2011]. DSMs Baroni and Lenci [2010] are based on the distributional hypothesis of meaning Harris [1954] assuming that semantic similarity between words is a function of the overlap of their linguistic contexts. DSMs are typically constructed from co-occurrence statistics of word tuples that are extracted from a text corpus or from data harvested from the web. A wide range of contextual features are also used by DSMs exploiting lexical, syntactic, semantic, and pragmatic information. DSMs have been successfully applied to the problem of semantic similarity computation. According to Baroni and Lenci [2010], the success of contextual DSM features is due to their ability to encode the attributes of word senses. According to *attributional similarity* Turney [2006], semantic similarity between words is based on the commonality of their sense attributes. A closely related assumption is that the semantic similarity of two words can be estimated as the similarity of their two closest senses Resnik [1995], henceforth, referred to as the *maximum sense similarity* assumption.

In this chapter, we investigate a new unsupervised approach for the construction of DSMs with application to lexical semantic similarity computation [1] . First, a corpus of snippets (short pieces of text containing words of interest) is harvested from the web. Then, a semantic network is constructed encoding the semantic relations between words in the corpus. Co-occurrence and context features are used to measure the strength of relations. The network is a parsimonious

[1]The core of this chapter is also presented in Iosif and Potamianos [2013b].

representation of the information encoded in the corpus. We then define the notion of semantic neighborhood and associated metrics of semantic similarity that exploit this notion. The proposed semantic similarity metrics are motivated by the maximum sense similarity, attributional similarity and metric space assumptions. The similarity metrics are evaluated against human similarity ratings using standard datasets, achieving state-of-the-art results. This work builds upon our prior research in Iosif and Potamianos [2010, 2012], while the following are the original contributions:

1. An efficient and scalable methodology is proposed, for corpus creation using web-harvested data. Unlike the quadratic query complexity of our previous algorithm Iosif and Potamianos [2010], the proposed method has linear query complexity with respect to the size of the lexicon.

2. Three unsupervised language-agnostic similarity computation algorithms are proposed that exploit the semantic neighborhoods. The best performing neighborhood-based metrics outperform well-established approaches that rely on elaborate knowledge resources.

3. We demonstrate the effectiveness of co-occurrence-based similarity metrics when corpus-based frequencies are incorporated in comparison to the use of web hits Iosif and Potamianos [2010]. This is further investigated with respect to the textual proximity of co-occurring words.

4. The assumption that the semantic similarity between two words can be estimated as the similarity of their two closest senses is validated using (sense-untagged) web data.

5. The computation of semantic neighborhoods introduced in Iosif and Potamianos [2012] is extended by applying a number of co-occurrence-based similarity metrics in addition to the context-based metrics. Word co-occurrence is shown to be more salient than contextual features regarding the discovery of senses via semantic neighborhoods.

The remainder of the work is organized as follows: In Section 4.2, we review related work in the areas of semantic similarity computation and word sense disambiguation. The procedure and motivation behind harvesting a corpus of snippets from the web is detailed in Section 4.3. In Section 4.4, we define our semantic network and propose three novel similarity metrics that utilize the notion of semantic neighborhood. The corpora and experimental procedures are described in Section 4.5, while the evaluation results are reported in Section 4.6. Last, Section 4.8 concludes this work.

4.2 Related Work

Semantic similarity metrics can be divided into two broad categories: (i) metrics that rely on knowledge resources, and (ii) corpus- or web-based metrics that do not require any external knowledge source. A representative example of the first category are metrics that exploit the WordNet ontology Miller [1990]. For computing word similarity these metrics incorporate features such as the length of paths between them Leacock and Chodorow [1998]; Wu and Palmer [1994] or the information content of their least subsumer that is estimated from a corpus Jiang and Conrath [1997]; Resnik [1995]. WordNet glosses have been also exploited for extracting contextual information Banerjee and Pedersen [2002]; Patwardhan and Pedersen [2006]. An in depth review of the major WordNet-based metrics can be found in Budanitsky and Hirst [2006]. Corpus-based metrics are formalized as DSMs Baroni and Lenci [2010] and are based on the distributional hypothesis of meaning Harris [1954]. DSMs can be categorized into unstructured (unsupervised) that employ a bag-of-words model Iosif and Potamianos [2010] and structured that rely on syntactic relationships between words Baroni and Lenci [2010]; Grefenstette [1994]. Web-based metrics employ search engines to estimate the frequency of word co-occurrence Gracia et al. [2006]; Turney [2001]; Vitanyi [2005] or construct corpora Bollegala et al. [2007]; Iosif and Potamianos [2010]. The identification and extraction of other types of relations has been mainly studied through the use of linguistic patterns. Lexico-syntactic patterns were applied in the influential work of Hearst Hearst [1992], for the identification of hyponymy, followed by numerous similar approaches, e.g., Caraballo [1999].

Recently, motivated by the graph theory, several aspects of the human languages have been modeled using network-based methods. In Mihalcea and Radev [2011]; Radev and Mihalcea [2008], an overview of network-based approaches is presented for a number of NLP problems. Different types of language units can be regarded as vertices of such networks, spanning from single words to sentences. Typically, network edges represent the relations of such units capturing phenomena such as co-occurrence, syntactic dependencies, and lexical similarity. An example of a large co-occurrence network is presented in Widdows and Dorow [2002] for the automatic creation of semantic classes. In Ferrer-I-Cancho and Solé [2001], it is reported that the co-occurrence networks of words that co-exist at very short proximity, exhibit a number of small-world properties and are highly clustered. Similar observations regarding the structural properties of co-occurrence networks were also made in Véronis [2004], where the HyprLex algorithm was proposed for sense discovery. In Agirre et al. [2006], an extension of the main ideas presented in Véronis [2004] was proposed for word sense disambiguation (WSD). In particular, the PageRank algorithm Brin and Page [1998] was employed for identifying hubs over a co-occurrence network.

Semantic similarity computation is closely related to WSD. WSD methods can be divided into two main categories: (i) supervised approaches that apply machine learning for learning sense labels for a set of words with respect to a given context (sense labeling), and (ii) unsupervised approaches that automatically discriminate (discover) word senses without label assignment. A detailed survey of WSD is provided in Agirre and Edmonds [2007]; Ide and Véronis [1998]; Navigli [2009]. The employment of network-based metrics for the computation of semantic similarity has attracted less attention compared to WSD. WordNet-based similarity metrics can be regarded as a special case of network metrics, since they are built on the top of a manually created network. To the best of our knowledge few network-based metrics are reported in the literature that integrate network creation with semantic similarity computation. In Lemaire and Denhière [2004], a co-occurrence network was constructed, and the similarity between two words was estimated as the product of weights of the shortest path between them with moderate performance results.

Following the paradigm of the vector space model (VSM) that constitutes the main implementation of DSMs, our approach is based on corpus-based co-occurrence statistics for the creation of a semantic network. One important difference with prior work in this area is that no language-specific tools, e.g., dependency parsers Baroni and Lenci [2010], or human annotations, e.g., Wikipedia hyperlinks Wojtinnek et al. [2012], are used here. For example, in Wojtinnek et al. [2012] the English Wikipedia was used for the disambiguation of target words (Wikipedia concepts) and a very large network was constructed by exploiting the hyperlinks between them. For each node (word) a vector was created including a number of strongly connected nodes selected by an algorithm inspired by spreading activation theory Collins and Loftus [1975]. The similarity between two words was estimated as the cosine of their respective vectors. In our work, two types of metrics are investigated for weighting the strength of the link between a reference noun and its neighbors, namely, co-occurrence-based and context-based. Co-occurrence-based metrics were previously used for the weighting of contextual features Agirre et al. [2009]; Baroni and Lenci [2010] and the creation of co-occurrence networks Widdows and Dorow [2002]. To the best of our knowledge context-based metrics have never been applied for any of the aforementioned tasks. Our work is also motivated by cognitive consideration and theories of semantics. The network-based metrics proposed here are motivated by two well-founded hypotheses regarding semantic similarity, namely, maximum sense similarity Resnik [1995] and attributional similarity Turney [2006]. Our work extends the traditional VSM approach into a two tier system: corpus statistics are parsimoniously encoded in a network, while the task of similarity computation is shifted (from corpus-based techniques) to operations over network neighborhoods. The proposed network creation process constitutes a new paradigm for implementing DSMs that enables the direct exploitation of neighborhood

semantics, e.g., definition of metrics that adopt different hypotheses regarding semantic similarity, investigation of neighborhood structural properties.

4.3 Corpus Creation Using Targeted Web Queries

In this section, we investigate the creation of corpora from web-harvested data via the formulation of targeted queries. There are two main types of web queries that can be used for corpus creation: (i) conjunctive queries (AND), and (ii) individual queries (IND) [1] . Assuming N words in our lexicon, in the first case all pairwise AND conjunctions are formed and the corresponding queries are posed to a web engine, e.g., "w_i AND w_j". Corpus creation via AND queries leads to quadratic query complexity $\mathcal{O}(N^2)$ in the number of words in the lexicon. Alternatively, one can download documents or snippets with linear query complexity $\mathcal{O}(N)$ using IND queries, i.e., "w_i".

The main advantage of AND queries is that they construct a corpus that is conditioned on word-pairs, explicitly requesting the co-occurrence of word-pairs in the same document. Co-occurrence is a strong indicator of similarity and corpora created via AND queries have been shown to provide very good semantic similarity estimates Iosif and Potamianos [2010]. To better understand the role of co-occurrence as a feature in semantic similarity computation, we need to revisit the very definition of semantic similarity, as it pertains to words and their senses. According to the information-theoretic approach proposed in Resnik [1995], the similarity of two concepts can be estimated as the similarity of their two closest senses. This is also in agreement with our "common sense" (cognitive) model of semantic similarity: when two words are mentioned, their closest senses are activated[2]. We believe that an important contribution of the co-occurrence feature to semantic similarity computation is that *co-occurrence acts as a semantic filter that only retains the two closest senses*. See Section 4.6.2 for the experimental justification of this claim.

Unfortunately attempting to build corpora and DSMs using conjunctive AND queries does not scale to thousands of words due to quadratic query complexity [3]. We are thus forced to investigate the alternative of using IND queries and face the sense disambiguation issues associated with such corpora. Corpora created via IND queries are similar to a typical text corpus with one important difference: the frequency of occurrence of the words in our lexicon

[1] Word co-occurrence statistics estimated on a web-harvested corpus may be biased due to optimizations applied by web search engines when ranking documents and selecting snippets. This is especially true for query words in corpora resulting from conjunctive AND queries, but less so for corpora harvested via IND queries.

[2] The maximum sense similarity assertion is widely employed by many top-performing similarity metrics, such as the WordNet-based metrics Budanitsky and Hirst [2006].

[3] Although a work-around could be found, e.g., using cross-products of all term statistics in a search engine index and n-gram counts.

can be manipulated [1] to deviate from Zipf's law. Assuming that the same number of snippets is downloaded for each word in our lexicon (using IND queries), we expect that rare words will be well-represented within the corpus. As a result, the corpus will be more "informative", i.e., the entropy rate of a unigram (zeroth order Markov process) model will be higher.

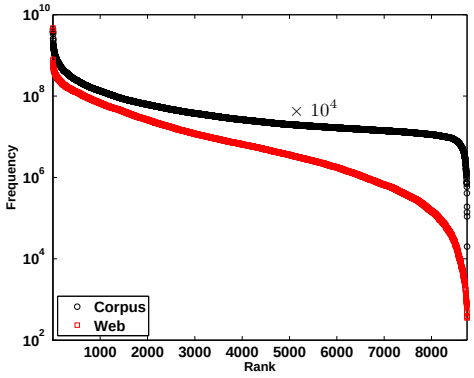


Figure 4.1: Frequency of $8,752$ nouns vs. their rank. The frequencies were computed using 1) corpus counts (black curve), and 2) web hits (red curve). For comparison purposes the corpus frequencies were multiplied by 10^4.

The normalization word-frequency effect can be illustrated by plotting the empirical distribution of the frequency of the words in the lexicon. Using a lexicon of $8,752$ nouns, the noun frequencies are plotted as a function of their rank in Fig.4.1. More specifically, we created a web corpus by posing an IND query for each noun and retrieving the $1,000$ top-ranked snippets (see Section 4.5). The corpus frequencies were multiplied by 10^4 in order to facilitate the comparison with the red curve showing web hits frequency. According to the Zipf's law Zipf [1965], the frequency of a word w decreases non-linearly as its rank increases:

$$f(w) = \frac{c}{r(w)^{\gamma}}, \tag{4.1}$$

where $f(w)$ and $r(w)$ are the frequency and the rank of word w, respectively, while c and γ are corpus-dependent. It is clear that the frequency difference between the high-ranked and the low-ranked words is somewhat normalized, i.e., smaller (absolute) γ, for the case of corpus frequencies, as opposed to the use of web hits. For the example of Fig.4.1, γ equals to -0.54 and -0.90 for corpus frequencies and web hits, respectively. These values were computed for the ranks lying between $1,000$ and $6,000$ using a least squares linear model. This normalization is expected to smooth the domination of very frequent words at the denominator of

[1] For this example this sort of "manipulation" is caused by requesting fixed number of snippets for each word of the lexicon.

co-occurrence-based metrics, such as (2.15)–(2.17). The performance of web hits and corpus counts is presented in Table 4.2.

4.4 Semantic Network

Next, we construct a semantic network encoding the relevant corpus statistics. The network is defined as an undirected (under a symmetric similarity metric) graph $F = (V, E)$ whose the set of vertices V are all words in our lexicon L, and the set of edges E contains the links between the vertices. The links (edges) between words in the network are determined and weighted according to the pairwise semantic similarity of the vertices.

The network is a parsimonious representation of corpus statistics as they pertain to the estimation of semantic similarities between word-pairs in the lexicon. In addition, the network can be used to *discover relations that are not directly observable in the data*; such relations emerge via the systematic covariation of similarity metrics. Semantic neighborhoods play an important role in this process. The members of the semantic neighborhoods of words are expected to contain features capturing diverse information at the syntactic, semantic and pragmatic level.

4.4.1 Semantic Neighborhoods

For each word (reference word) that is included in the lexicon, $w_i \in L$, we consider a subgraph of F, $F_i = (N_i, E_i)$, where the set of vertices N_i includes in total n members of L, which are linked with w_i via edges E_i. The F_i subgraph is referred to as the semantic neighborhood of w_i Iosif and Potamianos [2012]. The members of N_i (neighbors of w_i) are selected according to a semantic similarity metric (co-occurrence-based defined in Section 2.2.4.1, or context-based defined in by (2.20) in Section 2.2.4.2) with respect to w_i, i.e., the n most similar words to w_i are selected. Note that the semantic network is not a metric space under (the proposed co-occurrence or context-based) semantic similarity because the triangle inequality is, in general, not satisfied. Next, we propose three semantic similarity metrics that utilize the notion of semantic neighborhood.

4.4.2 Maximum Similarity of Neighborhoods

This metric is based on the hypothesis that the similarity of two words, w_i and w_j, can be estimated by *the maximum similarity of their respective sets of neighbors*, defined as follows:

$$M_n(w_i, w_j) = \max\{\alpha_{ij}, \alpha_{ji}\}, \tag{4.2}$$

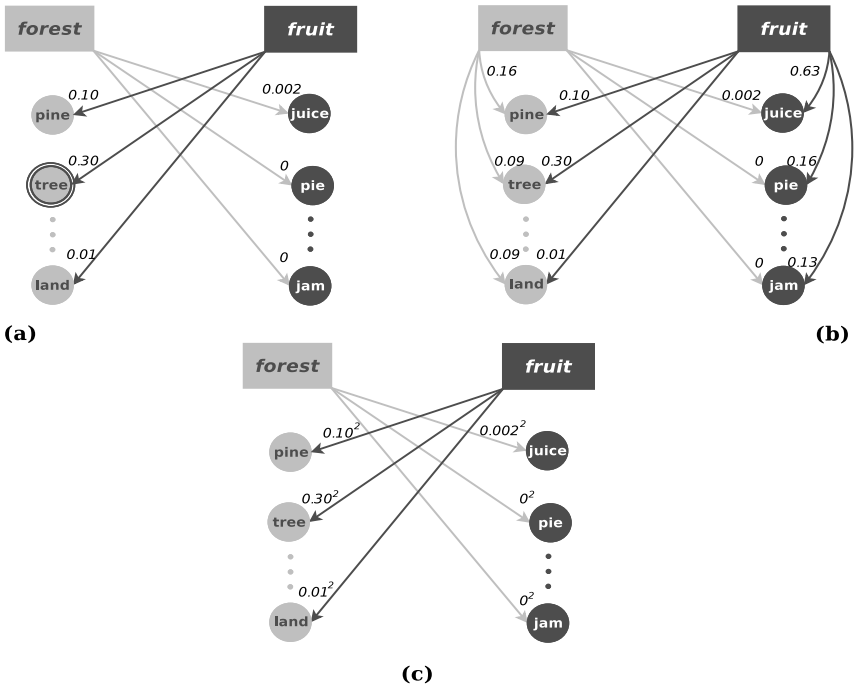


Figure 4.2: Pictorial view of neighborhood-based metrics. Two reference nouns, "forest" and "fruit", are depicted along with their neighborhoods: {pine, tree, ..., land} and {juice, pie, ..., jam}, respectively. Arcs represent the similarities between reference nouns and neighbors. The similarity between "forest" and "fruit" is computed according to (a) maximum similarity of neighborhoods, (b) correlation of neighborhood similarities, and (c) sum of squared neighborhood similarities.

where

$$\alpha_{ij} = \max_{x \in N_j} S(w_i, x), \quad \alpha_{ji} = \max_{y \in N_i} S(w_j, y).$$

α_{ij} (or α_{ji}) denotes the maximum similarity between w_i (or w_j) and the neighbors of w_j (or w_i) that is computed according to a similarity metric S (see for example (2.15)–(2.17), (2.19), (2.20)). N_i and N_j are the set of neighbors for w_i and w_j, respectively. The definition of M_n is motivated by the maximum sense similarity assumption[1]. As discussed above, semantic neighborhoods encode diverse information. Here the underlying assumption is that the most salient information in the neighbors of a word are semantic features denoting senses of this word. In other words, we assume that semantic neighborhoods (and semantic networks, in general) can be used to mine for word senses[2]. The M_n metric takes values in the interval $[0, 1]$, where 1

[1] This metric utilizes the similarities between w_i and x, $\forall\, x \in N_j$, as well as between w_j and y, $\forall\, y \in N_i$. This is slightly different than considering all the pairwise similarities between the members of N_i and N_j.

[2] See also Navigli and Crisafulli [2010] for word sense discovery via semantic networks.

stands for absolute similarity. Also, $M_n(w_i, w_j) = M_n(w_j, w_i)$, i.e., M_n is symmetric. An example illustrating the computation of similarity between "forest" and "fruit" is depicted by Fig.4.2(a) [1] . $M_n(\textit{"forest"}, \textit{"fruit"}) = 0.30$ because the similarity between "fruit" and "tree" (among all neighbors of "forest") is the largest.

4.4.3 Correlation of Neighborhood Similarities

The similarity between w_i and w_j is defined as follows:

$$R_n(w_i, w_j) = \max\{\beta_{ij}, \beta_{ji}\}, \tag{4.3}$$

where

$$\beta_{ij} = \rho(C_i^{N_i}, C_j^{N_i}), \;\; \beta_{ji} = \rho(C_i^{N_j}, C_j^{N_j})$$

and

$$C_i^{N_i} = (S(w_i, x_1), S(w_i, x_2), \ldots, S(w_i, x_n)), \quad \text{where } N_i = \{x_1, x_2, \ldots, x_n\}.$$

Note that $C_j^{N_i}$, $C_i^{N_j}$, and $C_j^{N_j}$ are defined similarly as $C_i^{N_i}$. The ρ function stands for the Pearson's correlation coefficient, N_i is the set of neighbors of word w_i, and S is a similarity metric. Here, we aim to exploit the entire semantic neighborhoods for the computation of semantic similarity, as opposed to M_n where a single neighbor is utilized. The motivation behind this metric is attributional similarity, i.e., we assume that semantic neighborhoods encode attributes (or features) of a word. Neighborhood correlation similarity in essence compares the distribution of semantic similarities of the two words on their semantic neighborhoods. Thus, this metric is expected to provide more robust similarity estimates compared to M_n, especially when few data are available. The ρ function incorporates the covariation of the similarities of w_i and w_j with respect to the members of their semantic neighborhoods. The underlying assumption is that two semantically similar words are expected to have co-varying similarities with respect to their neighbors. Moreover, the ρ function normalizes this covariance by the standard deviations of the similarities of w_i and w_j. The similarity scores computed by R_n metric ranges in the interval $[-1, 1]$, where -1 and 1 denote zero and absolute similarity, respectively. R_n is symmetric, since $R_n(w_i, w_j) = R_n(w_j, w_i)$. The similarity computation process is exemplified in Fig.4.2(b) for the words w_1 ="forest" and w_2 = "fruit". The similarity vectors between the neighbors N_1 of "forest" and each of the words are computed: $C_1^{N_1} = (0.16, 0.09, \ldots, 0.09)$,

[1] We also investigated a variation regarding the creation of semantic neighborhoods including within the neighborhoods the target words. This variation was observed to yield almost identical performance with the proposed approach.

$C_2^{N_1} = (0.10, 0.30, \ldots, 0.01)$. Similarly, $C_1^{N_2}$, $C_2^{N_2}$ are computed for the neighbors of "fruit" and combined to estimate $R_n(\textit{"forest"}, \textit{"fruit"}) = -0.04$.

4.4.4 Sum of Squared Neighborhood Similarities

The similarity between w_i and w_j is defined as follows:

$$E_n^{\theta}(w_i, w_j) = \left(\sum_{x \in N_j} S^{\theta}(w_i, x) + \sum_{y \in N_i} S^{\theta}(w_j, y) \right)^{\frac{1}{\theta}}, \tag{4.4}$$

where N_i is the set of neighbors of word w_i, and S is any similarity metric. Similar to (4.3) all neighbors contribute to the computation of the final similarity score, here this is performed by summing the squares ($\theta = 2$) of similarities between w_i and w_j's neighbors. The same calculation is repeated for w_j and the neighbors of w_i to make $E_n^{\theta}(w_i, w_j)$ symmetric. This is illustrated by Fig.4.2(c) for the computation of similarity between "forest" and "fruit" for $\theta = 2$. That is $E_n^{\theta=2}(\textit{"forest"}, \textit{"fruit"}) = \sqrt{(0.10^2 + 0.30^2 + \cdots + 0.01^2) + (0.002^2 + 0^2 + \cdots + 0^2)}$ $= 0.22$.

The $E_n^{\theta=2}$ metric is unbounded since the yielding similarity scores range within $[0, \infty)$. This range is smoothed in a non-linear way by taking the square root of the accumulated squares of similarities. As in (4.3), the motivation underlying $E_n^{\theta=2}$ metric is the attributional similarity, i.e., neighbors stand as attributes (or features). However, what is different here is the utilization of the attributional similarity as indicator for semantic similarity, i.e., the accumulation of word–to–neighbor similarities. The contribution of each word–to–neighbor similarity is non-linearly weighted using the square of the respective similarity score. The motivation behind using $\theta > 1$ is that more similar words in the neighborhoods should be weighted more in the final similarity decision[1]. Qualitatively, the $E_n^{\theta=2}$ weighting scheme takes the middle road between selecting the maximum pairwise similarity in (4.2) and the "linear" weighting of pairwise similarity in (4.3). Note that as θ goes to ∞, E_n^{θ} and M_n become equivalent.

4.5 Evaluation Datasets, Corpora and Experimental Procedure

4.5.1 Evaluation Datasets

The performance of similarity metrics was evaluated against human ratings from three standard datasets of noun pairs, namely: 1) MC Miller and Charles [1998], 2) RG Rubenstein and

[1]Despite the resemblance between the $E_n^{\theta=2}$ metric and the Euclidean distance, no assumption is adopted here about the semantic neighborhoods being metric spaces under S.

Goodenough [1965], and 3) WS353 Finkelstein et al. [2002]. The first dataset consists of 28 noun pairs. For the second and the third dataset we present results for the subset of 57 and 272 pairs, respectively, that are also included in SemCor3[1] corpus. The Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against the ground truth. Let $X = (x_1, x_2, ..., x_m)$ and $Y = (y_1, y_2, ..., y_m)$ be the vectors that contain the similarity scores given by human subjects and the computational metric, respectively, for each of the $i = 1, 2, ..., m$ word pairs of the datasets. Pearson's correlation coefficient is computed as follows:

$$\rho_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, for $i = 1, 2, ..., m$. This coefficient was selected instead of Spearman's rank correlation coefficient in order to retain the initial scaling of similarities in the evaluation metric, as opposed to the alternation of this scaling through the transformation of similarities into ranks.

4.5.2 Experimental Corpora and Procedure

We created the following corpora of web snippets using AND or IND queries posed via the Yahoo! Search API. 1) Corpus1: Using AND queries $1,000$ snippets were acquired for each pair of nouns, for the MC dataset. The major aspect of this corpus is the (explicitly requested) co-occurrence of nouns for which the similarity is computed. 2) Corpus2: Using IND queries $1,000$ snippets were acquired for each (unique) noun of the MC dataset. Unlike Corpus1, the creation of Corpus 2 is not driven by the co-occurrence constraint. 3) Corpus3: The same IND queries were used as for the case of Corpus2, but the queries were augmented with lexical descriptors denoting senses (see Section 4.6.2 for details). Corpus3 can be regarded as an extension of Corpus2, in which the acquired data are intended to (uniformly) cover the different senses of nouns. The aforementioned corpora are exploited (see Section 4.6.2) for investigating the effect of word co-occurrence and their senses to the computation of context-based similarity (for the MC dataset only). 4) Corpus4: This is a corpus created using IND queries, consisting of approximately $8,752,000$ snippets. More specifically, $1,000$ snippets were acquired for each noun taken from a set of $8,752$ English nouns of the SemCor3 corpus. Corpus4 is used for the creation of the semantic network as described in Section 4.4.

For Corpus4 the baseline performance of co-occurrence and context-based similarity metrics was computed (see also below for parameter definition). Then the semantic neighborhoods were defined and the maximum/correlation neighborhood similarities were computed. A de-

[1] http://www.cse.unt.edu/~rada/downloads.html

tailed list of experiments was conducted trying to investigate the performance of the following list of parameters: 1) the size of the contextual window, H, used in Q^H, M_n, R_n, $E_n^{\theta=2}$ 2) the metric used for the selection of neighbors: co-occurrence-based (J, D, I, G) or context-based similarity (Q^H), 3) the S metric used in M_n, R_n, $E_n^{\theta=2}$: co-occurrence-based (J, D, I, G) or context-based similarity (Q^H), 4) the neighborhood size (number of neighbors n), used in M_n, R_n, $E_n^{\theta=2}$ metrics, 5) the corpus size, i.e., number of snippets per word (50, 100, 200, 500, 1,000) used to construct the network, and 6) the network size, that is the number of concepts (nouns of lexicon) that constitute the network: 9, 88, 176, 876, $1,751$, $4,376$, $6,127$, and $8,752$. The results are presented next.

4.6 Results

The performance of the context-based metric and the co-occurrence-based metrics is compared in Section 4.6.1 (baseline performance). In Section 4.6.2, we compare the performance of the baseline context-based metric for corpora created via AND and IND queries, and we show that senses play an important role in achieving good performance. In Section 4.6.3, we present the performance of the proposed neighborhood-based metrics, defined in Section 4.4, that utilize the large corpus created via IND queries (Corpus4) and the corresponding semantic network.

4.6.1 Baseline

We consider as baseline the performance of the following metrics: 1) context-based similarity metric Q^H defined by (2.20) in Section 2.2.4.2, 2) co-occurrence-based metrics, defined in Section 2.2.4.1, relying on counts that were computed either using the web as a corpus (number of hits), or the corpus of snippets harvested with respect to the $8,752$ nouns (Corpus4). The

Dataset	Contextual window size			
	$H=1$	$H=2$	$H=3$	$H=5$
MC	**0.53**	0.35	0.29	0.20
RG	**0.52**	0.41	0.37	0.29
WS353	**0.30**	0.21	0.17	0.13

Table 4.1: Performance of context-based metric Q^H for several values of H.

baseline scores for the context-based similarity metric Q^H are presented in Table 4.1, for several values of the contextual window size H. The best correlation scores are obtained for $H=1$ across all datasets, while the performance drops as the size of the contextual window increases. Even for $H=1$, moderate correlation scores are achieved for the MC and RG datasets, while

the baseline performance is poor for the WS353 dataset. These results indicate the inability of naive context-based similarity metrics to exploit contextual features, despite the availability of a large corpus. The baseline performance for co-occurrence-based metrics that incorporate

Dataset	Co-occurrence-based metrics using							
	Web counts				Corpus counts			
	J	D	I	G	J	D	I	G
MC	-0.20	0.24	**0.35**	0.33	0.59	0.59	0.78	**0.85**
RG	-0.01	0.21	**0.28**	0.31	0.60	0.60	0.77	**0.81**
WS353	-0.02	0.10	0.19	**0.20**	0.18	0.22	0.60	**0.61**

Table 4.2: Performance of co-occurrence-based metrics using web and corpus counts: Jaccard (J), Dice (D), Mutual info. (I), and Google-based sem. rel. (G).

web counts[1] (hits) or corpus counts (Corpus4) is shown in Table 4.2. Regarding corpus counts, the co-occurrence of nouns is considered at the snippet boundary. We observe that the employment of corpus counts leads to significantly higher correlation scores, compared to using web counts. For example, the correlation improves from 0.33 to 0.85 using the G metric for the case of the MC dataset. This observation is consistent for all metrics across all three datasets. For corpus counts, the best performance is achieved by Google-based Semantic Relatedness, G, while the Mutual information, I, is a close second. Jaccard, J, and Dice, D, coefficients have lower but comparable performance.

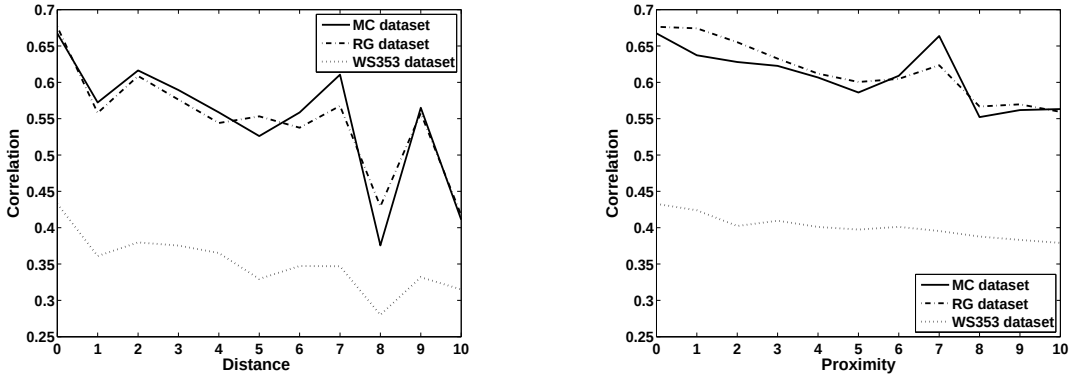
Dataset	Number of snippets				
	50	100	200	500	1000
MC	43%	28%	10%	3%	0%
RG	40%	24%	12%	8%	5%
WS353	20%	13%	8%	3%	3%

Table 4.3: Percentage of highly related pairs that have zero co-occurrence corpus counts as a function of downloaded snippets.

Despite the high performance of co-occurrence metrics using corpus counts, their applicability is strongly depended on the corpus size. The percentage of highly related noun pairs that have zero co-occurrence (corpus) counts are presented in Table 4.3, for several numbers of downloaded snippets. We assumed that two nouns are highly related if the corresponding similarity score (normalized between 0 and 1) provided by human subjects is greater than 0.5. The reported number of snippets were randomly selected from the initial (full) corpus. We observe that for the RG and WS353 datasets, even for the maximum number of snippets ($1,000$

[1]The Exalead web search engine was used (`http://www.exalead.com/search/`).

per noun) 3–5% of the highly related pairs do not co-occur.



(a) (b)

Figure 4.3: Correlation performance of the co-occurrence-based metric I vs. word (a) distance and (b) proximity (within web documents).

The poor performance of co-occurrence-based metrics that rely on web counts may be attributed to the fact that the co-occurrence of words is estimated at the document level, rather than at the level of snippet or sentence (for corpus counts). The key difference between web and corpus counts is the proximity of the co-occurring words, as well as, the different corpus statistics shown in Fig.4.1. In order to investigate the role of proximity we formulated NEAR queries that constrain the distance between two words in an AND web query[1]. The performance of the I metric using web counts is presented as a function of the distance and proximity of co-occurring words (within documents) in Fig.4.3(a) and Fig.4.3(b), respectively. The distance, δ, between two co-occurring words denotes that exactly δ tokens interfere between them. The proximity, π_δ, of two words allows to π tokens to appear between them, where $0 \leq \pi_\delta \leq \delta$. We observe that imposing a distance/proximity constraint significantly improves the achieved correlation compared to the baseline of web co-occurrence counts in Table 4.2. For example, the correlation for the RG dataset improves from 0.28 (see Table 4.2) to 0.68 for $\delta = 0$ and π_0. Despite the clear improvement in the performance of web-based count performance (when applying a proximity constraint), corpus-based counts still outperform web-based counts. The second reason behind the superior performance of corpus-based counts is the (normalized) word frequency statistics[2] of the snippet corpus (vs. web) shown in Fig. 4.1.

[1] This was performed by using the NEAR operator which is supported by the Exalead search engine. For example, the "w_i NEAR/2 w_j" query returns the number of hits for which words w_i and w_j co-occur at proximity equal to 2.

[2] For a theoretical analysis of how word-frequency normalization in a web snippet corpus reduces the estimation error of co-occurrence similarity metrics see Iosif and Potamianos [2013a].

4.6.2 Incorporating Word Senses Through Web Queries

We compare the performance of context-based similarity metrics for web corpora created via AND or IND queries. All results reported in this section are for the MC dataset. Baseline similarity scores here are computed using the $Q^{H=1}$ metric[1] defined by (2.20) in Section 2.2.4.2. The correlation scores for the context-based similarity metric using AND and IND queries (dot-

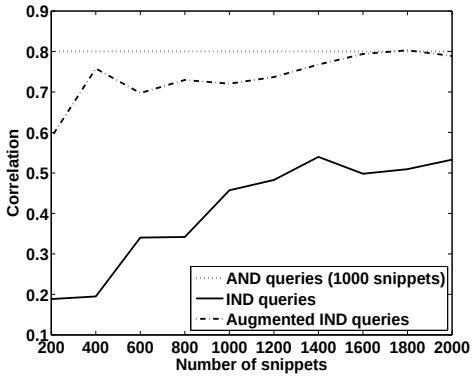


Figure 4.4: Correlation performance for context-based similarity for web corpora created via AND queries (dotted), IND queries (solid), and IND queries augmented with sense descriptors (dashed-dotted) for the MC dataset.

ted and solid line, respectively) are shown in Fig. 4.4 as a function of the number of snippets. The performance for AND queries is a single point and was obtained at $1,000$ queries[2] (shown here as reference). It is clear that context-based similarity metrics perform much better when using AND rather than IND queries.

Our hypothesis is that the very good performance of AND queries is due to co-occurrence acting as a semantic filter that retains the two closest senses of the two words. Moreover, the poor performance of IND queries is due to the limited coverage of senses within the top snippets. In order to verify this hypothesis we perform (sense) filtering explicitly following three steps: 1) identify all senses of the words of interest using WordNet[3], 2) use conjunctive AND queries between a word and each of its word senses to obtain relevant snippets that (mainly) contain the desired sense, e.g., the IND query for "magician" becomes "magician AND illusionist" (augmented), given the first WordNet sense of this word, and 3) compute the context-based similarity between all possible pairs of word senses and select the maximum

[1] For the rest experiments in this paper we use a context window of $H = 1$. Our experiments, as well as, our prior work Iosif and Potamianos [2010] indicate that $H = 1$ provides the best results for the problem of similarity computation.

[2] The maximum number of IND queries is greater than the number of AND queries, due to the use of two individual queries, instead of a single conjunctive query. Web search engines return up to $1,000$ snippets per query.

[3] WordNet is used here simply to validate this hypothesis.

similarity. Step 3 makes the implicit assumption that word similarity should be computed between the two closest senses Budanitsky and Hirst [2006], i.e., if s_{ik} is the kth sense of the word w_i the maximum sense context-based similarity Q' between words w_i, w_j is defined as:

$$Q'(w_i, w_j) = \max_{k,l} Q(s_{ik}, s_{jl}), \tag{4.5}$$

where Q is defined by (2.20) in Section 2.2.4.2. The performance of the augmented IND queries is shown in Fig.4.4 with a dashed-dotted line. It is clear that the use of the augmented IND queries significantly outperforms simple IND queries and approaches the performance of AND queries as the number of snippets increases.

Overall, the presented results suggest that the exploitation of word senses is essential for the accurate computation of semantic similarity. We have also experimentally demonstrated that context-based semantic similarity estimates are more accurate if we consider the two closest senses, i.e., the maximum pair-wise sense similarity score. A major road-block in top-down corpus creation using IND queries is the lack of sense coverage in the corpus. We show next that by creating a corpus by posing IND queries for thousands of words, as well as, by employing the notion of semantic neighborhood we can overcome this roadblock and obtain excellent semantic similarity estimates.

4.6.3 Semantic Network

Next, we investigate the computation of semantic networks using different types of similarity metrics. Next, we present the evaluation results for the proposed neighborhood-based similarity metrics, defined by (4.2)–(4.4), for different ways of defining the semantic neighborhoods.

4.6.3.1 Semantic Neighborhoods

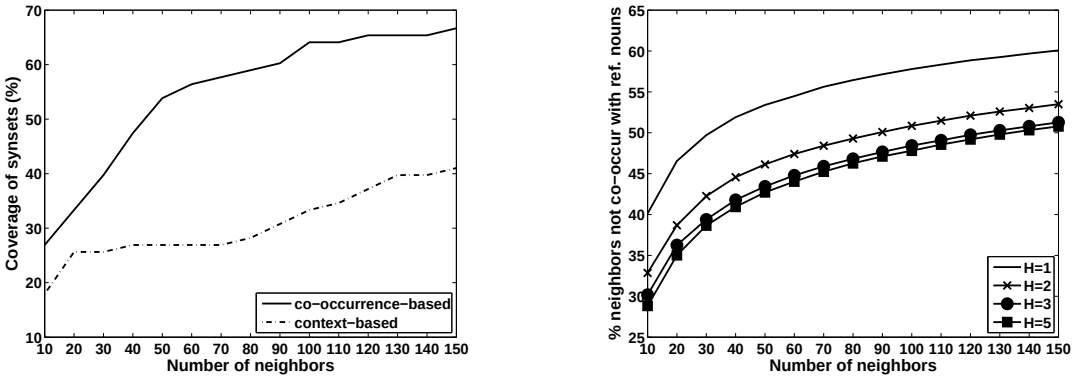
The semantic neighborhood of each word is estimated using one of the co-occurrence-based metrics defined in Section 2.2.4.1, or the context-based similarity metric Q^H defined by (2.20) in Section 2.2.4.2. Our semantic network consists of $8,752$ nouns. Given a (reference) noun w, let $A(w)$ and $B(w)$ be the neighborhood sets of w computed using co-occurrence-based and context-based metrics. The intersection of $A(w)$ and $B(w)$, $A(w) \cap B(w)$, as well as their differences, $A(w) - B(w)$ and $B(w) - A(w)$, are shown in Table 4.4 for ten nouns that are included in the experimental datasets. The co-occurrence-based metric D defined in (2.16) was applied for the computation of $A(w)$, while the context-based metric $Q^{H=1}$ defined by (2.20) in Section 2.2.4.2 was used for the computation of $B(w)$. For both metrics, the 50 top-ranked neighbors were considered. The neighbors that are emphasized using bold fonts denote

Reference noun (w)	Neighbors selected by		
	D and $Q^{H=1}$ $(A(w) \cap B(w))$	D only $(A(w) - B(w))$	$Q^{H=1}$ only $(B(w) - A(w))$
automobile	**auto**, **vehicle**, **car**, engine	accident, mechanic, starter, convertible	bus, aviation, tractor, lighting
brother	son, father, nephew, dad	twin, priest, police, girl	guy, lawyer, neighbor, pianist
car	**vehicle**, travel, service, price	accident, driver, **automobile**, fuel	business, city, game, quality
coast	island, **beach**, resort, sea	bay, boat, tsunami, port	lake, summer, entertainment, weather
food	water, health, service, industry	meal, kitchen, snack, gourmet	product, market, quality, life
forest	**land**, tree, vegetation, wildlife	rain, fire, pine, wood	nature, region, environment, property
fruit	tree, **plant**, taste, juice	vine, jam, acidity, pie	meal, wood, food, garden
hill	mountain, tree, park, forest	slope, **mound**, walk, snowball	island, city, resort, summer
journey	**trip**, destination, adventure, **travel**	discovery, quest, voyage, road	vision, goal, holiday, culture
slave	nigger, slavery, servant, manumission	gladiator, labor, freedom, master	beggar, democracy, society, aristocracy

Table 4.4: Excerpts of semantic neighborhoods for ten nouns using the co-occurrence-based metric Dice (D) and/or the context-based metric $Q^{H=1}$.

(lexicalized) senses of the respective reference nouns.

We observe that the discovery of a number of senses via the neighborhoods is feasible for some nouns, e.g., "automobile" and "car". This is more clear for $A(w) \cap B(w)$ compared to $A(w) - B(w)$ and $B(w) - A(w)$. However, sense discovery appears to be difficult for other nouns, such as "food" and "slave", for which their respective senses can not be easily described by single words. In addition to synonymy, taxonomic relations are encoded by the neighbors of $A(w) \cap B(w)$, e.g., IsA(vehicle, car), PartOf(automobile, engine). Relations of associative nature, e.g., ProducedBy(industry, food), are also denoted by some neighbors of $A(w) \cap B(w)$. Essentially, the main difference between $A(w) - B(w)$ and $B(w) - A(w)$ is that the former includes members that tend to formulate more direct associative relations with the reference nouns. In some cases these relations appear in the corpus as bigrams, such as "car accident" and "hill slope". Members of $B(w) - A(w)$ seem to correspond to relations of a broader semantic/pragmatic scope, such as (food, life) and (journey, culture).



(a) (b)

Figure 4.5: (a) Percentage of WordNet synonyms included in the semantic neighborhoods vs. number of neighbors. The neighborhoods were computed using 1) co-occurrence-based metric D (solid line), and 2) context-based metric $Q^{H=1}$ (dash–dotted line). The reference nouns were taken from the RG dataset. (b) Percentage of neighbors that do not co-occur with the reference nouns vs. number of neighbors. In total, $1,000$ reference nouns were randomly selected from the lexicon. The neighborhoods were computed by the context-based metric Q^H. The percentage is shown for different values of H.

Given the importance of senses for the computation of semantic similarity, we attempt to quantify the performance of co-occurrence and context-based metrics with respect to the discovery of senses through their neighborhoods. The percentage of synonyms of reference nouns (taken from the RG dataset) that are included in the neighborhoods are presented in Fig.4.5(a) as a function of the neighborhood size. The sets of synonyms for each reference noun were created by consulting the WordNet synsets. The semantic neighborhoods were computed using either the co-occurrence metric D, or the context metric $Q^{H=1}$. In general, more synonyms are captured by the D metric compared to the $Q^{H=1}$ metric. This distinction is greater for neighborhoods that include more than 50 members.

Moreover, we investigate the effect of the context window H with respect to the selection of neighbors that do not co-occur with the reference nouns. The percentage of such neighbors computed by Q^H is depicted in Fig.4.5(b) for several sizes of the neighborhoods, and for four values of the contextual window size H. The percentages were computed for $1,000$ nouns that were randomly selected from the network. The best results are consistently obtained when using immediate context, i.e., $H=1$, which can be attributed to the best performance of this window value for the case of context-based similarity computation Iosif and Potamianos [2010]. This is also shown here in Table 4.1.

4.6.3.2 Neighborhood-based Metrics

The computation of semantic similarity consists of two basic steps: 1) computation of seman-

Dataset	Neighbor selection	Similarity computation	Abbreviation for neighbor sel./ similarity comp.	Metrics		
				$M_{n=100}$	$R_{n=100}$	$E^{\theta=2}_{n=100}$
MC	co-occur.	co-occur.	(CC/CC)	0.90	0.72	**0.90**
MC	co-occur.	context	(CC/CT)	**0.91**	0.28	0.46
MC	context	co-occur.	(CT/CC)	0.52	**0.78**	0.56
MC	context	context	(CT/CT)	0.51	0.77	0.29
RG	co-occur.	co-occur.	(CC/CC)	**0.87**	0.67	**0.86**
RG	co-occur.	context	(CC/CT)	0.86	0.32	0.53
RG	context	co-occur.	(CT/CC)	0.58	**0.72**	0.61
RG	context	context	(CT/CT)	0.57	0.69	0.33
WS353	co-occur.	co-occur.	(CC/CC)	**0.64**	0.50	**0.64**
WS353	co-occur.	context	(CC/CT)	**0.64**	0.14	0.20
WS353	context	co-occur.	(CT/CC)	0.47	0.56	0.48
WS353	context	context	(CT/CT)	0.46	**0.57**	0.11

Table 4.5: Correlation for neighborhood-based metrics. Four combinations of the co-occurrence-based metric Dice (D) and the context-based metric $Q^{H=1}$ were used for the definition of semantic neighborhoods and the computation of similarity scores.

tic neighborhoods, and 2) computation of similarity scores (the S metric in (4.2) and (4.3)), allowing for the following combinations.

- Compute neighborhoods and similarity scores using a co-occurrence-based metric (CC/CC).
- Compute neighborhoods using a co-occurrence-based metric; compute similarity scores using a context-based metric (CC/CT).
- Compute neighborhoods using a context-based metric; compute similarity scores using a co-occurrence-based metric (CT/CC).
- Compute neighborhoods and similarity scores using a context-based metric (CT/CT).

For the above approaches, the co-occurrence-based metric[1] D and the context-based metric $Q^{H=1}$ were used. The correlation results for the neighborhood-based metrics $M_{n=100}$, $R_{n=100}$, and $E^{\theta=2}_{n=100}$ for neighborhood size of 100 are presented in Table 4.5 (see the next paragraph for the choice of n). The use of a co-occurrence metric for neighbor selection achieves the highest results for all datasets, for $M_{n=100}$ and $E^{\theta=2}_{n=100}$, while, the context-based metric appears

[1] D achieved slightly higher performance than other co-occurrence metrics (not shown here for the sake of space).

to be better for selecting neighbors for the correlation-based neighborhood metric $R_{n=100}$. The choice of the semantic similarity metric is of secondary importance for the $M_{n=100}$ and $R_{n=100}$ metrics, provided that the appropriate metric is used for neighborhood creation. For the $E^{\theta=2}_{n=100}$ metric however, only the (CC/CC) combination performs well. The results are significantly higher compared to the context-based baselines (see Table 4.1). The best $M_{n=100}$ and $E_{n=100}$ metrics also outperform the metrics that rely on web or corpus counts. Overall, utilizing network neighborhoods for estimating semantic similarity can achieve very good performance, and the type of metric (feature) used to select the neighborhood is a key performance factor.

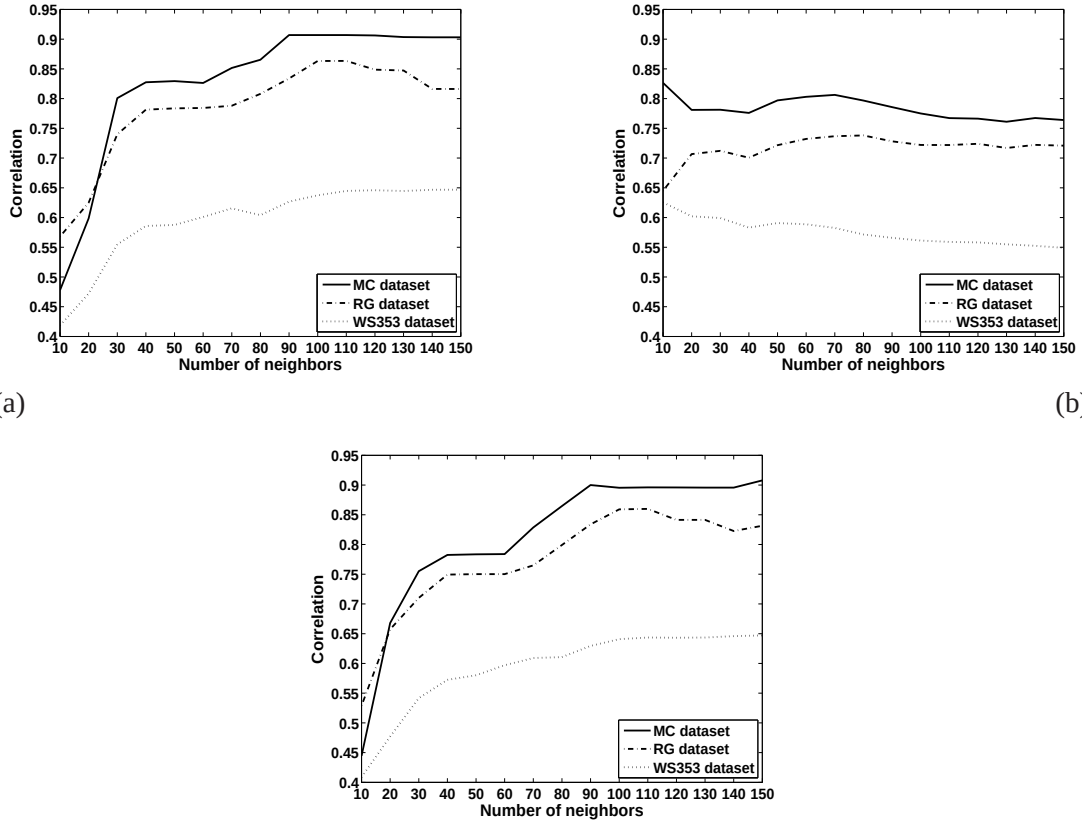


Figure 4.6: Performance vs. number of neighbors for neighborhood-based metrics: (a) maximum similarity of neighborhoods M_n: (CC/CT), (b) correlation of neighborhood similarities R_n: (CT/CC), and (c) sum of squared neighborhood similarities E^{θ}_n: (CC/CC).

Next, we investigate the performance of the metrics as a function of neighborhood size n. The performance of the M_n metric using co-occurrence-based metric D for neighbor selection,

and $Q^{H=1}$ for similarity computation is shown in Fig.4.6(a). We observe that performance increases with n peaking around $n=80-100$. The performance remains high also for $n > 100$. The performance of the R_n metric using $Q^{H=1}$ for neighbor selection and D for similarity computation is shown in Fig.4.6(b). The performance of R_n is relatively flat as a function of neighborhood size, achieving good performance even for small neighborhoods. The performance of the E_n^θ metric using D for both neighborhood selection and similarity estimation is shown in Fig.4.6(c). M_n and E_n^θ exhibit comparable performance, while both appear to be better than R_n for high values of n.

Metric	Neighbor selection	Similarity computation	Dataset	Number of snippets per noun				
				50	100	200	500	1,000
Baseline	*not applicable*	co-occur. (corpus-based)	MC	0.24	0.31	0.43	0.57	**0.59**
			RG	0.35	0.42	0.56	**0.62**	0.60
			WS353	0.26	0.26	**0.27**	**0.27**	0.22
Baseline	*not applicable*	context	MC	0.35	0.52	**0.57**	0.54	0.53
			RG	0.38	0.45	0.50	**0.55**	0.52
			WS353	0.30	0.33	**0.34**	0.32	0.30
$M_{n=100}$	co-occur.	context	MC	0.54	0.61	0.71	0.88	**0.91**
			RG	0.54	0.60	0.73	0.83	**0.86**
			WS353	0.53	0.54	0.56	0.62	**0.64**
$R_{n=100}$	context	co-occur.	MC	0.28	0.49	0.67	0.73	**0.78**
			RG	0.42	0.60	0.68	0.69	**0.72**
			WS353	0.50	0.48	0.54	0.55	**0.56**
$E_{n=100}^{\theta=2}$	co-occur.	co-occur.	MC	0.56	0.61	0.69	0.83	**0.90**
			RG	0.57	0.61	0.72	0.81	**0.86**
			WS353	0.53	0.54	0.57	0.61	**0.64**

Table 4.6: Performance with respect to the number of corpus snippets per noun for the baseline and the neighborhood-based metrics.

The correlation scores for the best performing neighborhood metrics ($M_{n=100}$, $R_{n=100}$ and $E_{n=100}^{\theta=2}$ for the (CC/CT), (CT/CC) and (CC/CC) approaches, respectively) are presented in Table 4.6 as a function of the number of snippets downloaded for each word in the network. The performance of the corresponding baseline metrics are also shown in Table 4.6, i.e., the D metric relying on corpus counts, and $Q^{H=1}$. We observe that the neighborhood metrics outperform the baseline performance for all datasets. All three neighborhood metrics consistently obtain better correlation performance as the number of snippets increases. Unlike neighborhood metrics, the performance of baseline metrics is not shown to improve as the number of snippets increases and plateaus around 300–500 snippets.

Num. of concepts in net.	Metrics								
	$M_{n=100}$			$R_{n=100}$			$E^{\theta=2}_{n=100}$		
	MC	RG	WS353	MC	RG	WS353	MC	RG	WS353
9	0.68	0.63	0.55	0.87	0.70	**0.60**	0.75	0.42	**0.66**
88	0.68	0.63	0.55	**0.88**	**0.79**	**0.60**	0.70	0.45	0.61
176	0.68	0.63	0.54	0.86	0.78	**0.60**	0.70	0.44	0.60
876	0.68	0.69	0.58	0.83	0.74	0.59	0.73	0.60	0.62
1,751	0.75	0.73	0.62	0.80	0.71	0.58	0.80	0.66	0.64
4,376	**0.95**	0.82	**0.68**	0.78	0.70	0.57	**0.95**	0.75	**0.66**
6,127	0.91	**0.86**	0.65	0.77	0.72	0.57	0.90	0.72	0.64
8,752	0.91	**0.86**	0.64	0.78	0.72	0.56	0.90	**0.86**	0.64

Table 4.7: Performance of the neighborhood metrics for various network sizes.

Next, we investigate the performance of the neighborhood metrics with respect to the number of concepts (nouns) included in the network. The concepts were randomly selected; results are presented in Table 4.7 in the form of average correlation computed over ten runs. We experimented with various network sizes varying from 9 (0.1% of network) up to $8,752$ (100% of network) words. Regarding $M_{n=100}$ and $E^{\theta=2}_{n=100}$ metrics, performance improves as the network grows with best results around 4–5K words. Conversely $R_{n=100}$ perform best for small networks[1].

4.6.4 Fusion of Neighborhood Metrics

Next, we investigate the fusion of the best performing neighborhood metrics, M_n, R_n, and $E^{\theta=2}_n$, using the (CC/CT), (CT/CC), and (CC/CC) combinations, respectively (see Table 4.5). The fusion was performed as a weighted linear combination of their respective similarity scores. The largest dataset, i.e., WS353, was used for learning the weights of similarities using 10–fold cross validation. Then, the weights learned on (all of) WS353 were applied to the CM and RG datasets. Three different algorithms implemented in Weka[2] were applied for learning the weights, namely, linear regression, regression using Support Vector Machines (SVM), and regression trees. The performance of the fusion of metrics is presented in Table 4.8 for $n=100$,

[1]Note that since the neighborhood size is set to be (up to) $n = 100$ for all experiments for the first two rows (with network size of 9 or 88 words) all available words in the network are used to construct the neighborhoods, i.e., the set of neighbors is the same for all words considered. The superior performance of $R_{n=100}$ for small network size is a strong indication that using a common set of words to compare semantic similarities on, works better than using each word's semantic neighbor. The approach of using a common set of "seed words" has been successfully applied to affective text analysis Malandrakis et al. [2011]; Turney and Littman [2002] and warrants further research also for semantic similarity computation.

[2]`http://www.cs.waikato.ac.nz/ml/weka/`

along with the performance of the best individual neighborhood metric[1].

Metric/ Fusion algorithm	Dataset		
	MC	RG	WS353
Best individual neighborhood metric	0.91	**0.86**	0.64
Linear regression	0.91	**0.86**	0.65
Regression using SVM	0.91	**0.86**	0.65
Regression trees	**0.94**	0.82	**0.73**

Table 4.8: Performance for the fusion of neighborhood metrics.

We observe that the performance of fusion using linear and SVM–based regression is almost identical to the performance of the best individual neighborhood metric. Performance gains are obtained using regression trees for the CM (from 0.91 to 0.94) and WS353 dataset (from 0.64 to 0.73). However, performance is worse on the RG dataset. This trend is probably due to the different distribution of the similarity scores in the datasets (MC dataset for example contains only highly similar or dissimilar word pairs, while RG contains more uniformly distributed similarity scores).

4.6.5 Semantic Concreteness

Typically, the *degree of semantic concreteness* of a word is not taken into account in distributional models. However, evidence from neuro- and psyco-linguistics demonstrates significant differences in the cognitive organization of abstract and concrete nouns. For example, Kiehl et al. [1999] and Noppeney and Price [2004] show that concrete concepts are processed more efficiently than abstract ones (aka "the concreteness effect"), i.e., participants in lexical decision tasks recall concrete stimuli faster than abstract. According to dual code theory, Paivio [1971], the stored semantic information for concrete concepts is both verbal and visual, while for abstract concepts stored information is only verbal. Neuropsychological studies show that people with acquired dyslexia (deep dyslexia) face problems in reading abstract nouns aloud Coltheart [2000], *verifying that concrete and abstract concepts are stored in different regions of the human brain anatomy* Kiehl et al. [1999]. The reversal concreteness effect is also reported for people with semantic dementia with a striking impairment in semantic memory Papagno et al. [2009].

Motivated by this evidence, we study the semantic network organization and performance of DSMs for estimating the semantic similarity of abstract vs concrete nouns [2] . Specifically,

[1] We observed that the fusion algorithms exhibited similar (relative) performance for also other values of n (not reported here).

[2] Part of the work described in this section was conducted in collaboration with Maria Giannoudaki (ECE

we investigate the validity of the maximum sense and attributional similarity assumptions in network-based DSMs for abstract and concrete nouns (for both English and Greek).

4.6.5.1 Experimental Procedure

Lexica and corpora creation: For English we used a lexicon consisting of $8,752$ English nouns taken from the SemCor3[1] corpus. In addition, this lexicon was translated into Greek using Google Translate[2], while it was further augmented resulting into a set of $9,324$ entries. For each noun an individual query was formulated and the $1,000$ top ranked results (document snippets) were retrieved using the Yahoo! search engine[3]. A corpus was created for each language by aggregating the snippets for all nouns of the lexicon.

Network creation: For each language the semantic neighborhoods of lexicon noun pairs were computed following the procedure described in Section 4.4 using either co-occurrence D or context-based $Q^{H=1}$ metrics [4].

Network-based similarity computation: For each language, the semantic similarity between noun pairs was computed applying either the max-sense M_n or the attributional R_n network-based metric. The underlying semantic similarity metric (the S metric in (4.2) and (4.3)) can be either D or Q^H. Given that for both neighborhood creation and network-based semantic similarity estimation we have the option of D or Q^H, a total of four combinations emerge for this two-phase process: (i) D/D, i.e., use co-occurence metric D for both neighborhood selection and network-based similarity estimation, (ii) D/Q^H, (iii) Q^H/D, and (iv) Q^H/Q^H.

4.6.5.2 Evaluation Datasets

The performance of network-based similarity metrics was evaluated for the task of semantic similarity between nouns. The Pearson's correlation coefficient was used as evaluation metric to compare estimated similarities against the ground truth (human ratings). The following datasets were used:

English (WS353): Subset of WS353 dataset Finkelstein et al. [2002] consisting of 272 noun pairs (that are also included in the SemCor3 corpus).

Greek (GIP): In total, 82 native speakers of modern Greek were asked to score the similarity

Department, Technical University of Crete): creation of the GIP dataset, characterization of abstract/concrete nouns from the WS353 and GIP datasets, and analysis of the neighborhoods of abstract/concrete nouns. This work is also presented in Iosif et al. [2013].

[1] http://www.cse.unt.edu/~rada/downloads.html

[2] http://translate.google.com/

[3] http://www.yahoo.com//

[4] We have also experimented with other values of context window H not reported here for the sake of space. However, the highest performance was achieved for $H = 1$.

of the noun pairs in a range from 0 (dissimilar) to 4 (similar). The resulting dataset consists of 99 nouns pairs (a subset of pairs translated from WS353).

Abstract vs Concrete: From each of the above datasets two subsets of pairs were manually selected, where both nouns in the pair are either abstract or concrete, i.e., pairs consisting of one abstract and one concrete nouns were ruled out. More specifically, 74 abstract and 74 concrete noun pairs were selected from WS353, for a total of 148 pairs. Regarding GIP, 18 abstract and 18 concrete noun pairs were selected, for a total of 36 pairs.

4.6.5.3 Results

The performance of the two proposed network-based metrics, M_n and R_n, for neighborhood size of 100, is presented in Table 4.9 with respect to the English (WS353) and Greek (GIP) datasets. Baseline performance (i.e., no use of the network) is also shown for co-occurrence-based metric D and context-based metric Q^H. For the max-sense similarity $M_{n=100}$ metric,

Language: dataset	Number of pairs	Baseline		Network metric	Neighbor selection / Similarity computation			
		D	Q^H		D/Q^H	D/Q^H	Q^H/D	Q^H/Q^H
English: WS353	272	0.30	0.22	$M_{n=100}$	**0.64**	**0.64**	0.47	0.46
				$R_{n=100}$	0.50	0.14	**0.56**	**0.57**
Greek: GIP	99	0.25	0.13	$M_{n=100}$	**0.51**	**0.51**	0.04	0.04
				$R_{n=100}$	-0.11	0.03	**0.66**	0.11

Table 4.9: Pearson correlation with human ratings for neighborhood-based metrics for English and Greek datasets. Four combinations of the co-occurrence-based metric D and the context-based metric Q^H were used for the definition of semantic neighborhoods and the computation of similarity scores. Baseline performance is also shown.

the use of the co-occurrence metric D for neighbor selection yields the best correlation performance for both languages. For the attributional similarity $R_{n=100}$ metric, best performance is achieved when using the context-based metric D for the selection of neighbors in the network. As explained in Iosif and Potamianos [2013b], the neighborhoods selected by the D metrics tend to include words that denote word senses (yielding best results for similarity), while neighborhoods computed using the Q^H metric are semantically broader including word attributes (yielding best results for attributional similarity). The network-based DSMs results are also significantly higher compared to the baseline for both languages. The best results achieved by D/Q^H for the $M_{n=100}$, and Q^H/D for the $R_{n=100}$ are consistent with the results reported in Iosif and Potamianos [2013b] for English. The best performing metric for English is $M_{n=100}$ (max-sense) while for Greek $R_{n=100}$ (attributional). Overall, utilizing network neigh-

borhoods for estimating semantic similarity can achieve good performance[1], and the type of metric (feature) used to select the neighborhood is a key performance factor.

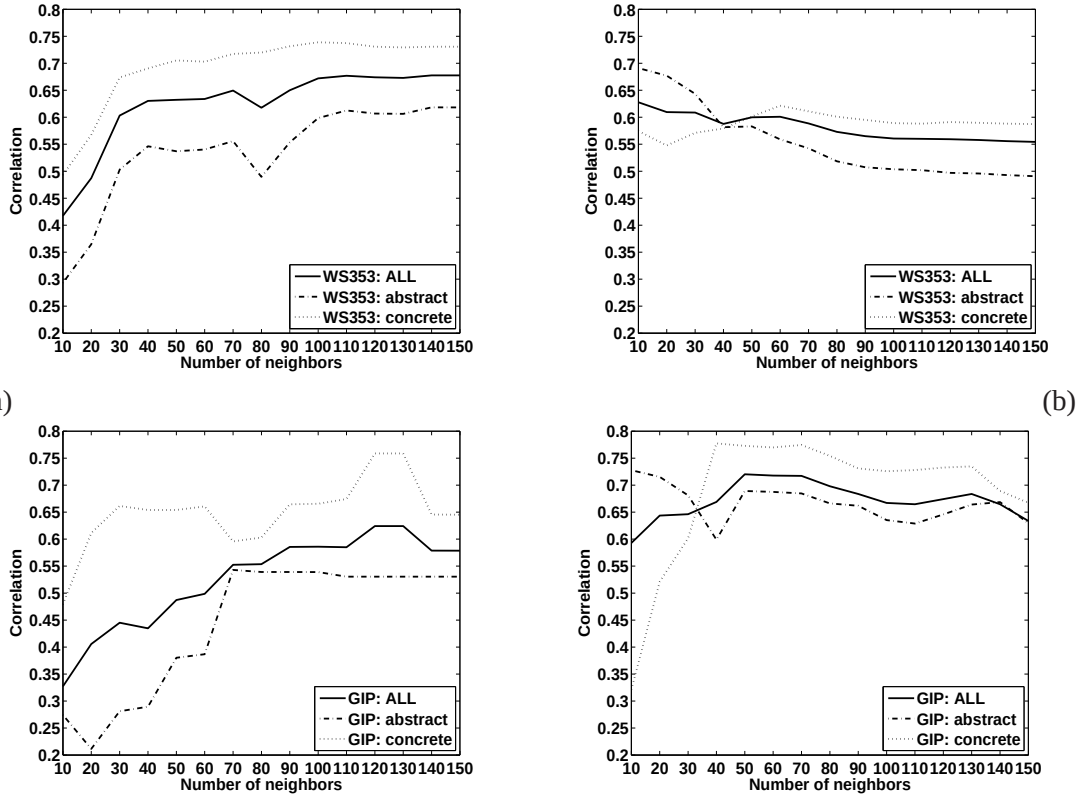


Figure 4.7: Correlation as a function of number of neighbors for network-based metrics. Max-sense M_n (D/Q^H) for datasets: (a) English and (c) Greek. Attributional R_n (Q^H/D) for datasets: (b) English and (d) Greek.

Next, we investigate the performance of the network metrics with respect to the neighborhood size n for the abstract and concrete noun pairs included in English and Greek datasets. The performance of the max-sense M_n (D/Q^H) metric is shown in Fig. 4.7(a),(c) for the (subsets of) WS353 and GIP, respectively. The performance over the whole (abstract and concrete) dataset is shown with a solid line. Similarly the results for the attributional R_n (Q^H/D) metric are shown in Fig. 4.7(b),(d). The main conclusions for these experiments (for both languages) are: 1) The correlation performance for concrete noun pairs is higher than for abstract noun pairs. 2) For concrete nouns the max-sense M_n metric achieves best performance, while for

[1] The best correlation score for the WS353 dataset does not exceed the top performance (0.68) of unsupervised DSMs Agirre et al. [2006]. However, we have found that the proposed network metrics obtain state-of-the-art results for other standard datasets, e.g., 0.87 for Rubenstein and Goodenough [1965] and 0.91 for Miller and Charles [1998].

abstract nouns the attributional R_n metric is the top performer. 3) For the R_n network metric, very good performance is achieved for abstract noun pairs for a small neighborhood size n (around 10), while for concrete nouns larger neighborhoods are needed (up to 40 and 30 neighbors, for English and Greek, respectively). In order to further investigate the network or-

Neighbor selection metric	Number of reference nouns	Type of reference nouns	Type of neighbors (abstract/concrete)			
			English (WS353)		Greek (GIP)	
			abstract	concrete	abstract	concrete
D	15	abstract	**76%**	24%	**82%**	18%
D	15	concrete	36%	**64%**	23%	**77%**
Q^H	15	abstract	**82%**	18%	**91%**	9%
Q^H	15	concrete	31%	**69%**	31%	**69%**

Table 4.10: Distribution of abstract vs concrete nouns in (abstract/concrete noun) neighbourhoods.

ganization for abstract vs concrete nouns, we manually inspected the top twenty neighbors of 30 randomly selected nouns (15 abstract and 15 concrete) and classified each neighbor as either abstract or concrete. The distributions of abstract/concrete neighbors are shown in Table 4.10 as a function of neighbor selection metric (D vs Q^H) and reference noun category. It is clear, that the neighborhoods of abstract nouns contain mostly abstract concepts, especially for the Q^H neighbor selection metric (similarly the neighborhoods of concrete nouns contain mainly concrete concepts). The neighbors of concrete nouns mainly belong to the same semantic class (e.g., "vehicle", "bus" for "car") often corresponding to relevant senses. The neighbors of the abstract nouns have an attributive function, reflecting relative attributes and/or aspects of the referent nouns (e.g., "religion", "justice" for "morality").

4.6.6 Comparison with Other Approaches

A comparison between our best results[1] and the performance of other similarity metrics is summarized in Table 4.11. The primary criterion for the selection of the presented metrics is the type of the exploited resources and corpora. This enables the comparison of knowledge– and data–driven approaches, while the latter often are the only feasible choice for under-resourced languages. The approaches that are presented in Table 4.11 can be distinguished into two main categories: (i) use of knowledge resources, such as WordNet, (ii) use of large corpora, e.g., Wikipedia and corpora harvested from the web. In addition, we consider a third category dealing with the integration of (i) and (ii) within a machine learning-based framework.

[1] As mentioned in Section 4.5 regarding the RG and WS353 datasets, we used their respective subsets covered by SemCor3. The same subsets were also used for the evaluation of the WordNet-based metrics.

Metric / System[a]	Resources / Corpora	ML[b]	Dataset		
			MC	RG	WS353
Wup	WordNet	no	0.76	0.78	0.34
Res	WordNet + SemCor	no	0.77	0.80	0.37
Vector	WordNet + SemCor	no	0.85	0.79	0.47
WikiRelate!	Wikipedia	no	0.45	0.53	0.48
AAHKPS1	4 billion web docs	no	0.88	0.89	0.66
TypeDM	ukWaC + Wikipedia + BNC	no	–	0.82	–
IP	28,000 web docs: AND queries	no	0.88	–	–
IP_s	web doc snippets: AND queries	no	0.80	0.81	0.57
AAHKPS2	WordNet + 4 billions web docs	yes	0.92	0.96	0.78
SSS	WordNet + 9 million web doc snippets	yes	0.88	–	–
Proposed	$\sim$ 9 million web doc snippets: IND queries				
($M_{n=100}$)		no	0.91	0.87	0.64
($E^{\theta=2}_{n=100}$)		no	0.91	0.86	0.64
(Fusion)		yes	0.94	0.82	0.73

[a] The metrics/systems shown in full uppercase, e.g. IP, were abbreviated using the first letter of authors' last names.

[b] Use of machine learning.

Table 4.11: Performance of several metrics/systems.

Three basic types of WordNet-based metrics are included in category (i): path length-based (Wup), information content-based (Res), and metrics that exploit the synset glosses (Vector). Wup Wu and Palmer [1994] is a purely taxonomic metric based on the notion of the least common subsumer (LCS), i.e., the most specific concept that is the parent node of two words. The similarity between two words, w_i and w_j, is estimated as the depth (distance from root node) of their LCS, normalized by their individual depths Pedersen [2010]. Wup is extended by the Res metric Resnik [1995] according to which the similarity of w_i and w_j is estimated as $Res(w_i, w_j) = -\log P(LCS(w_i, w_j))$, where $P(LCS(w_i, w_j))$ is the probability of the LCS of w_i and w_j estimated over a sense-tagged corpus Pedersen [2010]. The lexical information that is included in the WordNet glosses is utilized by the Vector metric Patwardhan and Pedersen [2006] for the construction of co-occurrence vectors extracted from a sense-tagged corpus. The similarity between w_i and w_j is estimated as the similarity of their respective vectors. In this work, we applied the aforementioned WordNet-based metrics using the WordNet::Similarity module[2] , which incorporates the SemCor corpus Pedersen and Michelizzi [2004]. More specifically, the similarity between two words was estimated according to

[2] http://search.cpan.org/dist/WordNet-Similarity/

(4.5) following the maximum sense similarity assumption Budanitsky and Hirst [2006]; Resnik [1995]. Regarding category (ii), the WikiRelate! system Strube and Ponzetto [2006] includes various taxonomy-based metrics that are typically applied to the WordNet hierarchy. The basic idea behind WikiRelate! is to adapt these metrics to a hierarchy extracted from the links between the pages of the English Wikipedia. A very large corpus is exploited by AAHKPS1 consisting of four billion web documents that were acquired via crawling Agirre et al. [2009]. For the computation of semantic similarity several variations of structured and unstructured DSMs were applied. An example of structured DSMs is the TypeDM model Baroni and Lenci [2010], where a number of lexico-syntactic patterns were extracted from the concatenation of three different corpora, namely, the web-harvested ukWaC corpus[1] , the dump of the English Wikipedia, and the British National Corpus (BNC). Our previous work, IP, is an example of corpus creation using a relatively small number of web documents Iosif and Potamianos [2010]. The basic idea was the use of conjunctive AND queries in order to retrieve documents in which the pair words co-occur. Also, we have replicated [2] our previous work using snippets instead of entire web documents (IP_s).

The third category that appears in Table 4.11 includes the following machine learning-based metrics/systems: AAHKPS2 and SSS. The basic approach behind AAHKPS2 Agirre et al. [2009] is the use of regression in order to combine similarity scores that were computed using different resources and corpora. A corpus of four billion web documents was exploited and results were derived using 10-fold cross validation. A different approach was followed by the SSS system Spanakis et al. [2009] according to which the WordNet was exploited in order to create thousands of word pairs denoting relations such as synonymy, meronymy, etc. These pairs were used for the formulation of web queries in order to create a corpus of snippets from which numerous lexico-syntactic patterns were extracted. The word similarity was estimated by a regression model considering the pattern frequencies as training features. The WS353 dataset was used for training excluding the pairs of the MC dataset, which were used for testing.

As it was expected, the exploitation of knowledge resources leads to high performance. The superiority of the Vector metric over the other WordNet-based metrics constitutes a successful paradigm regarding the exploitation of contextual features given that the word senses are knwon. The performance of the DSM-based approaches, i.e., AAHKPS1, TypeDM, and IP, is higher compared to the WordNet metrics. This observation is more interesting regarding the case of IP, where a relatively small corpus of web documents is used. Overall, the highest results are obtained by the machine learning-based approaches AAHKPS2 for the RG and WS353 datasets, and the fusion of $M_{n=100}$, $R_{n=100}$, and $E^{\theta=2}_{n=100}$ for the MC dataset. How-

[1] http://wacky.sslmit.unibo.it/

[2] As in IP, the top $1,000$ search results were acquired for each pair.

ever, we believe that further validation is needed for the machine learning approaches given the limited size of the datasets and the dangers of overfitting. Overall, the proposed $M_{n=100}$ and $E^{\theta=2}_{n=100}$ metrics can be regarded among the best-performing unsupervised data-driven metrics, built upon an efficient and scalable approach for corpus creation using web data.

4.7 Scalable and Efficient Corpus Indexing and Similarity Estimation

In this section, we briefly discuss some technical issues about the scalable and efficient creation of very large semantic networks. This discussion was motivated by the experience gained during the experimental work of this chapter. The implementation ideas that follow were conducted after [1] the completion of the experimental work (and respective results) presented in the previous sections of this chapter. Also, note that the comparison mentioned in the following paragraphs are meant to summarize our hands-on experience rather than to serve as a formal benchmark.

Scalable and efficient corpus indexing and similarity computation algorithms are essential for constructing very large semantic networks. The characterization "very large" is beyond the used lexicon of approximately 9K nouns. The initial step deals with the definition of lexicon(s) exhibiting "adequate" coverage for the language(s) of interest. The exploitation of typical dictionaries enable a straightforward solution regarding the the coverage requirement, since they aim to include the vast majority of words (or their canonical forms, i.e., lemmas) for a given language. Although a variety of such dictionaries exist for languages like English, this is not the case for under-resources languages such as a Greek. In order to overcome the fragmentation of dictionary availability across languages we used the lexicons that underly the GNU Aspell spell checkers [2] . In particular, we used the Aspell dictionaries for six languages, namely, (i) English, (ii) German, (iii) Italian, (iv) Spanish, (v) Greek, and (vi) Turkish. The use of those lexicons has a number of advantages that are well-aligned with our goal: (i) free availability for numerous languages, (ii) inflectional and derivational morphemes are included, which are missing from typical concise dictionaries, (iii) inclusion of known proper names, (iv) plain format, i.e., list of entries. Unfortunately, for some languages (e.g., Italian) the available Aspell dictionaries were enriched through an automatic procedure in an attempt to improve the cover-

[1] Nikolaos Malandrakis (Signal Analysis and Interpretation Laboratory, University of Southern California) and Ioannis Klasinas conducted the processing of Aspell dictionaries and Wikipedia dumps, as well as the harvesting of web data for corpus creation. Vassiliki Prokopi implemented a Java-based corpus indexing prototype for comparison purposes. The work of N. Malandrakis, I. Klasinas, and V. Prokopi was funded by the PortDial project (www.portdial.eu).

[2] http://aspell.net/

age of derivational morphemes. This resulted into extremely large dictionaries (i.e., more than one million of entries) due to the inclusion of auto-generated pseudo-words. The introduced redundancy may be acceptable for the purposes of spell checking, however, stands as a serious obstacle regarding the scalability of our approach. In order to alleviate this problem we filtered the Aspell dictionary entries with the vocabulary extracted from a large and authoritative resource of textual data. For the aforementioned languages the respective 2012 Wikipedia dumps were used. For each language the final lexicon was defined by taking the intersection of the corresponding Aspell dictionary and the Wikipedia vocabulary. For example, the resulting lexicon for English contains 125K (approx.) entries, while the largest lexicon was computed for Greek consisting of 407K (approx.) entries.

Given the lexicon for a particular language a corpus of web data (documents or document snippets) can be created using IND queries as described in Section 4.3. Once the corpus is created, the major technical challenge regards the corpus indexing that is essential for computing and storing the co-occurrence statistics needed by the co-occurrence-based and context-based similarity metrics defined in Section 2.2.4.1 and Section 2.2.4.2, respectively. The large size of the lexicons raise the demand for non-sparse indexing. Based on the observation that the vast majority of words do not co-occur, we proceeded with the storage of co-occurrence frequencies for the co-occurring words only. As a toy example, consider a very short lexicon that consist of seven entries only, which are assigned (e.g., based on alphabetical ordering) a unique numerical (integer) identifier from 1 to 7. Also, assume a small corpus that contains few instances of the lexicon words. The corpus index aims to store the absolute (non-zero) co-occurrence frequency of each word with respect to the other lexicon words. For this example, the following format was adopted and stored in a 7-line ASCII file.

```
1,1 5,1
2,3 3,1 5,1 6,1 7,1
2,1 3,2 7,1
4,1
1,1 2,1 5,1
2,1 6,1
2,1 3,1 7,1
```

The co-occurrence frequencies for a particular word can be located by jumping to that line whose numbering matches the word identifier, e.g., for the 3-rd word of the lexicon go to the 3-rd line of the index file. Each line of the index file includes a non-fixed number (because zero co-occurrence frequencies are not stored) of space-separated fields. Each field follows the format: “identifier of co-occurring word”,“,”,“absolute co-occurrence frequency”. For example,

the 3-rd word of the lexicon co-occurs with the 2-nd and the 7-th word one time (and two times with itself). The above non-sparse indexing is quite generic and it can be adapted to different considerations of word co-occurrence. For example, the frequencies can be counted within the sentence boundaries for the case of co-occurrence-based similarity metrics, while the frequency counting should be restricted within the selected contextual window (see parameter H defined in Section 2.2.4.2) regarding the context-based cosine similarity. Once the index construction is completed the encoded frequencies can be directly used for estimating the pairwise similarities between the lexicon entries. The same format can be also used for storing the estimated similarities: the co-occurrence frequencies are simply substituted by the similarity scores.

Especially during the index construction (as well as for the computation of similarities) appropriate data structures are required in order to store (and further process) the co-occurrence counts. Associative arrays (also referred to as hash tables) constitute a commonly used structure for such tasks. In particular, a hash of hashes (e.g., each value associated with a key is a hash) is useful for storing and processing the co-occurrence frequency (or similarity score) between two words. Regarding the exploitation of such structures we experimented with a small number of widely-used languages, namely Perl, Java and C++. For the case of Perl and Java we observed (working in a high-end desktop machine equipped with 32GB of RAM) an unaffordable memory overhead when using the respective built-in hashes caused by the large size of lexicons. The memory requirements were significantly were reduced using the SparseHash [1] library written in C++, which provides hash implementations optimized for low memory overhead.

4.8 Conclusions

We have investigated the estimation of semantic similarity using semantic networks, following an unsupervised [2] corpus-based approach. We have shown that it is possible to achieve state-of-the-art performance by encoding corpus statistics into a semantic network and then using the notion of semantic neighborhood to define novel semantic similarity metrics. The maximum neighborhood similarity metric performed the best when the semantic neighborhood was defined using co-occurrence metrics. We have also shown experimentally the importance of sense coverage and the validity of the maximum sense similarity assumption for context-based similarity metrics.

The fact that co-occurrence proved to be a good feature for selecting neighbors for the

[1] http://code.google.com/p/sparsehash/

[2] Despite the fact that the presented metrics have a number of experimental parameters, the characterization "unsupervised" refers to the notion of "language-agnostic".

maximum similarity metric implies that co-occurrence is a good feature for sense discovery. Moreover, we have studied the effect of word proximity for the estimation of semantic similarity, showing that very good performance is obtained when words co-occur at sentential level. The success of context-based similarity for neighborhood selection for the correlation metric implies that context is a good feature for discovering attributes in a network. In addition, the use of a corpus in which the not so common words are well-represented and a large lexicon creates an informative corpus that efficiently encodes the semantics of polysemous words and leads to good performance. More research and experimentation is needed to verify these claims. Overall, the achieved results are amongst the highest reported in the literature for unsupervised corpus-based metrics. Last but not least, the proposed approach is efficient, scalable and requires linear web query complexity with respect to the lexicon size. Future work deals with the incorporation of network features, such as centrality measurements, for the creation of semantic neighborhoods. Further research is needed with larger multilingual networks to verify the universality of the proposed metrics.

Moreover, we investigated the performance of network-based DSMs for semantic similarity estimation for abstract and concrete noun pairs of English and Greek. We observed a "concreteness effect", i.e., performance for concrete nouns was better than for abstract noun pairs. The assumption of maximum sense similarity as encoded by the M_n metric consistently yielded higher performance for the case of concrete nouns, while the semantic *similarity of abstract nouns was better estimated via the attributional similarity assumption* as implemented by the R_n metric. The results are consistent with the initial hypothesis that differences in cognitive organization may warrant different network organization in DSMs. In addition, abstract concepts were best modeled using an attributional network DSM with small semantic neighborhoods. This is a first step towards the better understanding of the network organization of DSMs for different categories of concepts. In terms of computation algorithms of semantic similarity, it might prove advantageous to define a metric that combines the maximum sense and attributional assumptions based on the semantic concreteness of the words under investigation. Further research on more data and languages is needed to verify the universality of the findings.

Chapter 5

Associative and Semantic Features Extracted From Web-Harvested Corpora

5.1 Introduction

We address the problem of automatic classification of associative and semantic relations between words, and particularly those that hold between nouns. Lexical relations such as synonymy, hypernymy/hyponymy, constitute the fundamental types of semantic relations Cruse [1986]. Associative relations are harder to define, since they include a long list of diverse relations, e.g., "Cause-Effect": onion–tears, "Instrument-Agency": hammer–carpenter. From the perspective of cognitive scientists, associative relatedness is triggered by the co-occurrence of words McNamara [2005], while the definition of semantic relatedness is controversial. The boundary between semantic and associative relations is not always clear, since highly associated words tend to be semantically related, e.g., (cat,dog). In McRae and Jones [2013], a short review of this argument is provided. However, a simple protocol is widely-used in order to smooth this fuzziness for dataset creation (for more details see Section 5.4). Previous research efforts have investigated semantic relations, such as the identification of synonyms, Iosif and Potamianos [2010], hyponyms, Caraballo [1999]. Also, the identification of other relations has attracted the research interest, e.g., the Task 8 of SemEval'10 dealt with the classification of various relations Hendrickx et al. [2010]. To our knowledge there have been very few computational efforts for the discrimination between associative and semantic relations, e.g., Turney [2008].

Such classification can be beneficial for a wide range of language technologies. For ex-

ample, in statistical language modeling, class-based language models Brown et al. [1992] have long been used to extend the coverage of the model – words in classes should typically be semantically related (i.e., sister hyponyms of the same hypernym). However, trigger models Lau et al. [1993] try to find words that change the probability distribution over other words, which is more of an associative relationship (e.g., postman – letter). Other technologies might use relationships in a different way: spoken dialogue systems often have an ontology of semantically related concepts (which one can attempt to learn from corpus data Pargellis et al. [2004]); query expansion techniques for information retrieval have also utilized semantically related concepts Fang [2008]. On the other hand, information extraction tasks may benefit from knowing associative relationships between words, since the contextual information leading to a decision to extract some piece of information is more likely to be associative in nature.

We propose an automated computational approach that discriminates between associative and semantic relations [1] . Text-based lexical and hit-based features are extracted from the web in order to classify given pairs of concepts as semantic or associative. These features do not rely on manually selected syntactic patterns, such as Hearst's patterns for the identification of "is-a" relations and semantic role labeling, but are rather motivated by general cognitive and linguistic principles. Specifically, we propose two novel features: (a) the degree of priming (co-occurrence asymmetry) as a function of the distance between the two words in text, and (b) the rate of change of context-based lexical similarity as a function of the context window size. Evaluation proceeds on a dataset containing 238 associative and semantic relations, which they were appropriately assembled by cognitive scientists in order to exclude any fuzzy relations.

5.2 Related Work

Semantic similarity metrics can be divided into two broad categories: (i) metrics that rely on knowledge resources, and (ii) corpus- or web-based metrics that do not require any external knowledge source. A representative example of the first category are metrics that exploit the WordNet ontology Miller [1990]. For computing the similarity between words these metrics incorporate features such as the length of paths between the two words Jiang and Conrath [1997]; Resnik [1995] or the information content of their least subsumer, estimated from a corpus Leacock and Chodorow [1998]; Wu and Palmer [1994]. WordNet glosses are also used as features in Patwardhan and Pedersen [2006]. A study that reviews in depth the major WordNet-based metrics is provided in Budanitsky and Hirst [2006]. Corpus-based metrics usually extract

[1]The work described in this chapter is also presented in Iosif et al. [2012]. A subset of the experimental features used in this chapter was developed in collaboration with Maria Giannoudaki (ECE Department, Technical University of Crete): linguistic patterns discussed in Section 5.3.3.

contextual features from text for computing semantic similarity. Web-based methods employ search engines to estimate the frequency of word co-occurrence Gracia et al. [2006]; Turney [2001]; Vitanyi [2005] or construct corpora Bollegala et al. [2007]; Iosif and Potamianos [2010]. The identification and extraction of other types of relations has been mainly studied through the use of linguistic patterns. Lexico-syntactic patterns were applied in the influential work of Hearst Hearst [1992], for the identification of hyponymy, followed by numerous similar approaches, e.g., Caraballo [1999]. Pattern-based approaches were also employed for the meronymy relation Girju et al. [2003].

5.3 Associative and semantic features

In this section, we propose two novel features for discriminating between associative and semantic relations using information automatically extracted from the web. Syntactic patterns are also investigated as features.

5.3.1 Hit-based priming coefficient

Hit-based metrics (summarized in Section 2.2.4.1) employ co-occurrence counts without taking into account: (i) the order of appearance of each word, and (ii) the distance (i.e., the number of words that intervene) between occurrences of the two words. Next, we motivate the use of these features for classifying associative and semantic relations.

The use of word order is motivated by findings in cognitive science and psycholinguistics, about the asymmetry of the priming phenomenon with respect to word pairs. In psycholinguistics, the notion of priming refers to the cognitive processing that takes place when two words in a certain order are presented to a human subject. In this framework, the first word p ("prime") serves as a stimulus that facilitates (or primes) the cognitive processing of the second word t ("target") McNamara [2005]. The selection of prime and target is determined experimentally for each word pair based on human response time, where response time is assumed to be inversely proportional to the strength of priming (or relatedness). Once the prime and target are defined, their usual order (p, t) is known as "forward", while the reverse order (t, p) is called "backward". It has been found that the difference between forward and backward priming is statistically significant for many related word pairs, e.g., responses to the pair ('light','bulb') were reported to be quicker than the responses to the pair ('bulb','light') Koriat [1981]; McNamara [2005]. Similar observations regarding the asymmetry of order of appearance within co-occurrence were also reported in the NLP literature Church and Hanks [1990]. However, data related to this phenomenon have been analyzed without further exploration of the cognitive

aspects of the problem.

Our goal is to define a "priming coefficient", i.e., a single metric that characterizes the degree of asymmetry in the forward and backward co-occurrence counts. Since priming is sensitive to ordering, we compute"forward" and "backward" co-occurrence counts (as a function of the distance between words) for each word pair. We expect that word pairs (p,t) with strong priming should appear much more often in the forward rather than the backward order. We expect priming to be a good discriminator between associative and semantic relations as psycholinguistics have suggested that priming effects can be of different magnitude for these different relations Ferrand and New [2003]; Plaut [1995].

Instead of using raw co-occurrence counts to estimate the priming coefficient, we propose to use the normalized hit-based metrics defined in Section 2.2.4.1 We introduce a variation of hit-based metrics that computes separately forward and backward co-occurrence counts, conditioned on the distance d between words. For a word pair (w_i, w_j), the forward relatedness $R_{f,m}$ is defined as

$$R^A_{f,m}(w_i, w_j) = A(w_i, w_j; d = m), \tag{5.1}$$

computed only for forward co-occurrence counts with distance d that is equal to m words. Function $A(.)$ denotes any of the hit-based metric defined in Section 2.2.4.1 Similarly, backward relatedness is defined as:

$$R^A_{b,m}(w_i, w_j) = A(w_j, w_i; d = m). \tag{5.2}$$

Total relatedness Λ^A_m is defined as the sum of the forward and backward relatedness

$$\Lambda^A_m(w_i, w_j) = R^A_{f,m}(w_i, w_j) + R^A_{b,m}(w_i, w_j) \tag{5.3}$$

for metric A, word pair (w_i, w_j) and distance equal to m. Finally, the priming coefficient Ψ^A_m is defined as the normalized absolute difference between forward and backward relatedness

$$\Psi^A_m(w_i, w_j) = \frac{|\, R^A_{f,m}(w_i, w_j) - R^A_{b,m}(w_i, w_j)\,|}{R^A_{f,m}(w_i, w_j) + R^A_{b,m}(w_i, w_j)}. \tag{5.4}$$

The priming coefficient is equal to 0 when the forward and backward co-occurrence counts are equal (no priming) and 1 when a word pair only appears with the forward (or backward) order (very strong priming).

5.3.2 Slope of text-based similarity

In Section 2.2.4.2, a context-based metric was defined by (2.20) that has been used in the literature for estimating the strength of semantic relations between words. In general, the strength of both semantic and associative relations covers a wide range from weak to strong; as a result, the relation strength by itself is a poor discriminator of the semantic vs associative class.

Based on observations in psycholinguistics Ferrand and New [2003] and computational linguistics Hearst [1992], words that are semantically similar, especially synonyms and words that belong to the same semantic class, can be identified by lexico-syntactic patterns from their immediate vicinity. For this case, context-based semantic similarity metrics are also shown to better correlate with human judgements when small contextual windows are used to compute similarity Iosif and Potamianos [2010]. Associative relations often imply a shared pragmatic context that is also evident from lexical similarity in the not-so-immediate vicinity. Thus, the relevance of lexical features extracted from context is expected to be a function of the contextual window size. According to the above considerations, we assume that the migration from syntactic to pragmatic features by increasing the size of H, will affect differently the context similarity of associative and semantic relations. For this purpose, we compute the difference of semantic similarity scores across different sizes of H. In particular, we focus on window sizes that differ exactly by one (first-order differences). Consider two words w_i and w_j. The difference of their similarity scores with respect to window sizes, H_x and H_y, is computed as:

$$S^{H_x}_{H_y}(w_i, w_j) = S^{H_x}(w_i, w_j) - S^{H_y}(w_i, w_j), \tag{5.5}$$

for $H_x - H_y = 1$. The similarities $S^{H_x}(w_i, w_j)$ and $S^{H_y}(w_i, w_j)$ are computed according to (2.20) defined in Section 2.2.4.2.

5.3.3 Linguistic patterns

We also examine whether specific syntactic patterns can discriminate between associative and semantic relations. By manual inspection of our data we have summarized the most common patterns for associative ([A1],[A2]) and semantic ([S1],[S2]) relations, respectively:

[A1] Complex Noun Phrases (NPs): $[NP_{term1|term2}[NP_{term1|term2}]]$, e.g., "**Ocean wave** energy is captured directly from surface waves or from pressure fluctuations."

[A2] Terms co-occurring in argument positions: $[NP_{term1}[VP[NP_{term2}]]]$, e.g., "...why do **giraffes** have long **necks**..."

[S1] The two terms in coordinative constructions: $[NP_{term1}]$ AND|OR $[NP_{term2}]$, e.g., "**Beet and radish** roots are similar in shape, but beets are usually larger than radishes."

[S2] The two terms in extended coordinative constructions, involving one additional NP between the NPs of interest: $[NP_{term1|term2}]$, $[NP]$ AND|OR $[NP_{term1|term2}]$, e.g., "... professional **carpet, upholstery and rug** cleaners in the Chicago ... "

Overall, associative noun pairs are expected to surface as arguments of the same phrase: in pattern A1 one NP is contained into the other, while in pattern A2 both NPs are manifested in the argument positions of the same VP (subject and object of the verb). Semantically related noun pairs form NPs that are structurally independent of each other; when they co-occur in close proximity they are usually connected with conjunctions.

5.4 Experimental Dataset

There are relatively few datasets containing rated associative or semantic relations between word pairs or terms, most of them containing fewer than 50 pairs. Lack of a standardized dataset of adequate size is a barrier to computational approaches that require fair amounts of data for training and testing. In this work, we have merged datasets taken from three different studies from the literature of psycholinguistics Chiarello et al. [1990]; Ferrand and New [2003]; Perea and Gotor [1997] for a total of 238 relations, equally split between 119 associative and 119 semantic relations (Table 5.1). All three datasets were designed for psycholinguistic exper-

Dataset No	# of semantic rel.	# of associative rel.
1	42	42
2	48	48
3	29	29
Total	119	119

Table 5.1: Experimental datasets.

iments related to priming, and contain only "pure" associative and semantic relations, avoiding word pairs that lie in the boundaries of the two relations. Well-established lists of free association norms, e.g., Nelson et al. [1998]; Palermo and Jenkins [1964], were used for the selection of "pure" associatively related pairs. Such lists are constructed by collecting the responses of human subjects when stimuli words are presented to them and they are asked to give the very first word they recall. Regarding "pure" semantic relations, the relevant pairs were selected according to the following criteria: (i) the words of each pair are members of the same semantic category and they have high scores of semantic relatedness, and (ii) they are not included

in lists of free association norms. The scores incorporated by the first criterion typically are estimated by collecting ratings given by human subjects. This way, pairs exhibiting strong associative and semantic relatedness, e.g., "cat–dog", were not included in the datasets. The semantically related pairs in datasets 1 and 3 consist exclusively of words that belong to the same semantic category, i.e., hyponyms of the same hypernym. The semantically related pairs in dataset 2 consist of words with various degrees of synonymy. Some indicative examples

Dataset No	Semantic rel.	Associative rel.
1	brass–iron	onion–tears
1	velvet–linen	hammer–nail
1	bacon–steak	pilot–plane
2	boat–ship	board–wood
2	work–labor	nucleus–center
2	fume–steam	hour–clock
3	clarinet–flute	drill–hole
3	pancake–waffle	cow–milk
3	rug–carpet	suitcase–trip

Table 5.2: Examples of dataset relations.

of the relations included in the experimental datasets are presented in Table 5.2. In theory a number of associative pairs may also exhibit a semantic relation. For example, "hammer" and "nail" can be considered as co-hyponyms, i.e., sharing the same hypernym. However, their associative relatedness is much stronger.

5.5 Experimental procedure

We compute hit-based metrics and text-based metrics through web search engines as described below:

5.5.1 Hit-based metrics

The number of word co-occurrences is estimated by Yahoo! search API[1] that returns the number of web hits given a particular query. We wish to compute the number of hits for the word pair (w_i, w_j) under the following constraints (i) w_i precedes w_j, and (ii) their distance, defined in Section 5.3.1 as the number of intervening words, is equal to m, i.e., $m = 2$. This is achieved by the query "$w_i \star\star w_j$" for $m = 2$. The "$\star$" symbol is a special search metacharacter, matching any word Bollegala et al. [2010]. Using this query formulation, we retrieve the number of

[1] http://developer.yahoo.com/search/

hits for both forward and backward ordering of the words up to a particular distance m. Once the number of hits is retrieved, the total relatedness $\Lambda_m^A(w_i, w_j)$ is computed according to (5.3), for each of the hit-based metrics A. Similarly, the priming coefficient $\Psi_m^A(w_i, w_j)$ is computed according to (5.4). For each word pair, $\Lambda_m^A(w_i, w_j)$ and $\Psi_m^A(w_i, w_j)$ are computed, using the four hit-based metrics $A = \{J, C, I, G\}$ and for distance values $m = 0, \ldots, 10$.

5.5.2 Text-based metrics

For the computation of text-based semantic similarity between the words of associative and semantic relations, we need to build a corpus from the web. For each word pair (w_i, w_j), we download 1000 snippets of web documents using the Yahoo! Search API. The web search is performed according to the conjunctive query "w_i AND w_j", ensuring that both words co-occur in the same snippet, for reasons explained in Iosif and Potamianos [2010]. Once the snippets are retrieved, we compute for each word pair: (i) the semantic similarity score, $S^H(w_i, w_j)$, according to (2.20) defined in Section 2.2.4.2, and (ii) the difference of similarities across different window sizes, $S_{H_y}^{H_x}(w_i, w_j)$, according to (5.5). The similarities are computed using B and LTF weighting schemes (see Table 3.1) for contextual window sizes $H = 1, \ldots, 10$.

5.6 Evaluation Results

In this section, we present results for associative vs semantic relation classification using the dataset described in Section 5.4. We used the support vector machine classifier provided by Weka[1]; similar results were obtained using naive Bayes classifier (not reported here due to lack of space). Note that for the case of individual features, e.g., priming coefficient, the classifiers were fed with scalars, i.e., the values for the respective feature. The evaluation was performed according to a 10-fold validation procedure. The evaluation results are reported in terms of classification accuracy.

In Fig. 5.1(a), the classification accuracy is shown for: (i) total relatedness, $\Lambda_m^J(w_i, w_j)$, computed according to (5.3) (solid line), and (ii) priming coefficient, $\Psi_m^J(w_i, w_j)$, computed according to (5.4) (dotted line). Classification accuracy is plotted as a function of m, the distance between words. It is clear that total relatedness achieves very poor accuracy that lies close to chance. The poor performance at $m = 0$ is an indication that the asymmetry of priming at the bigram level can not discriminate associative and semantic relations. The priming coefficient obtains good accuracy around 80% for most values of m, excluding the value $m = 1$. The discriminative ability of the priming coefficient improves for distance

[1] http://www.cs.waikato.ac.nz/ml/weka/

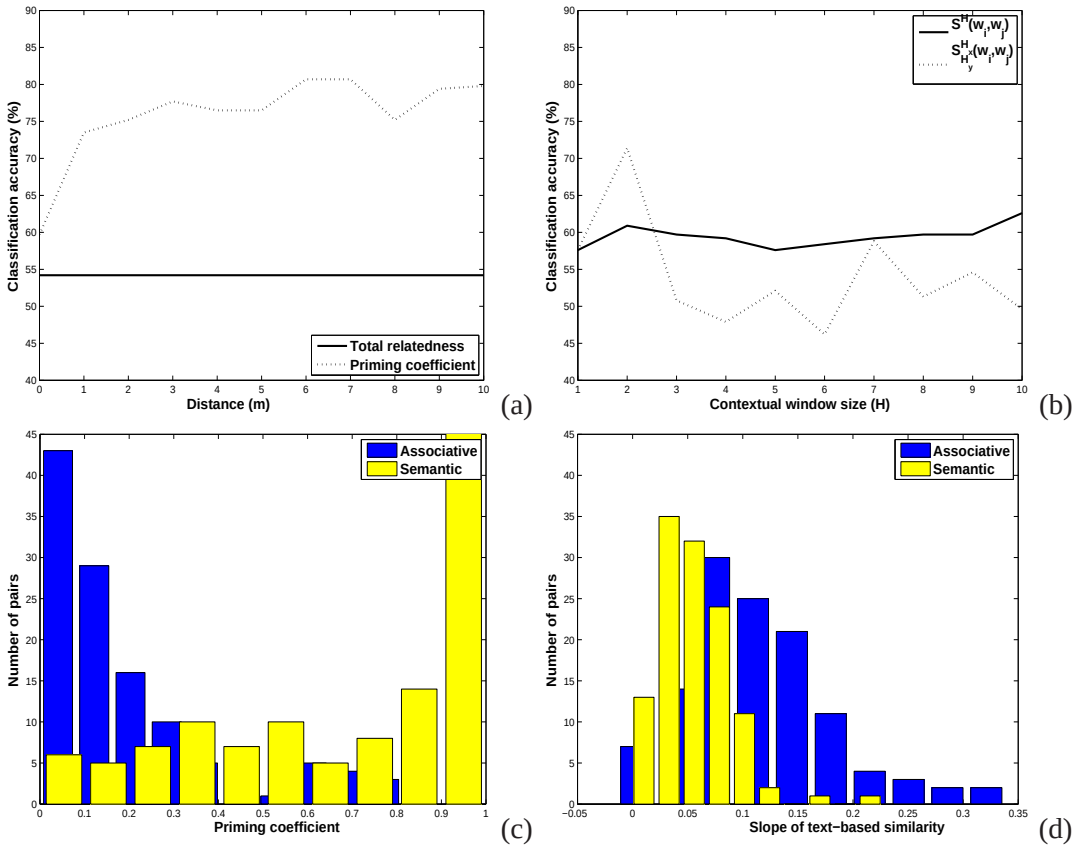


Figure 5.1: Classification accuracy for: (a) total relatedness $\Lambda^J_m(w_i, w_j)$, and priming coefficient $\Psi^J_m(w_i, w_j)$ as a function of distance m for the Jaccard (J) hit-based metric, (b) semantic similarity $S^H(w_i, w_j)$ and slope $S^{H_x}_{H_y}(w_i, w_j)$ metrics as a function of the window size H, using the binary B weighting scheme. Histograms for associative and semantic pairs: (c) priming coefficient $\Psi^J_5(w_i, w_j)$, (d) similarity slope $S^{H=2}_{H=1}(w_i, w_j)$.

values around 6 or 7 (although the differences in performance are not statistically significant). In Table 5.3, the classification precision is summarized for a number of hit-based metrics for the

Hit-based metrics	Accuracy	
	Total related.	Priming coef.
J	53.2%	86.5%
C	52.7%	86.5%
I	56.5%	85.7%
G	62.9%	86.5%

Table 5.3: Classification accuracy for total relatedness and priming coefficient.

total relatedness and priming coefficient. These results were obtained by joining the individual features for distances $m = 0, \ldots, 10$ into a single vector. Again, significantly higher results,

up to 86.5%, are achieved by the priming coefficient. There is no significant difference among the hit-based metrics.

In Fig. 5.1(b), the classification accuracy as a function of the window size H is shown for: (i) context-based similarity $S^H(w_i, w_j)$ computed according to (2.20) defined in Section 2.2.4.2 (solid line), and (ii) similarity slope $S^{H_x}_{H_y}(w_i, w_j)$ computed according to (5.5) (dotted line). For both of them, the binary B weighting scheme was used. Context-based similarity $S^H(w_i, w_j)$ is shown to be a relatively poor discriminator of associative vs semantic relations, and the achieved accuracy remains low, $55\% - 62\%$, for all values of H. The similarity slope $S^{H_x}_{H_y}(w_i, w_j)$ metric also performs poorly with the exception of window $H = 2$; performance for $S^{H=2}_{H=1}(w_i, w_j)$ exceeds 70% accuracy. Classification accuracies for both metrics $S^H(w_i, w_j)$, $S^{H_x}_{H_y}(w_i, w_j)$ and for both B and LTF weighting schemes are presented in Table 5.4. Results are computed for the joined feature vector containing values computed for contextual window sizes $H = 1, \ldots, 10$. For comparison, we have also included the accuracy for three WordNet-based similarity metrics, namely Leacock-Chodorow Leacock and Chodorow [1998], Resnik Resnik [1995], and Vector Patwardhan and Pedersen [2006]. These metrics were computed using the WordNet::Similarity package, developed by Pedersen and it is freely available through CPAN [1]. The $S^H(w_i, w_j)$ similarity metrics achieve relatively low accuracy, below 63%. WordNet-based metrics display diverse performance ranging from 54.2% for the Vector metric to 75.6% for the Resnik metric. The accuracy achieved by the slope $S^{H_x}_{H_y}(w_i, w_j)$ metric is up to 71.8% for the B weighting scheme. It is interesting to note that the best performing WordNet-based metric (Resnik) has substantial differences with the Vector metric, since it exploits the taxonomic paths of the WordNet hierarchy. The Leacock-Chodorow metric also relies on taxonomic features and it is shown to achieve performance comparable to the Resnik metric. The Vector metric, which yields the worst performance among the WordNet metrics, is quite close to the context-based metrics ($S^H(w_i, w_j)$), since both approaches utilize

Metrics of semantic similarity	Accuracy
$S^H(w_i, w_j)$, B scheme	62.6%
$S^H(w_i, w_j)$, LTF scheme	62.6%
$S^{H_x}_{H_y}(w_i, w_j)$, B scheme	**71.8%**
$S^{H_x}_{H_y}(w_i, w_j)$, LTF scheme	64.3%
WN: Leacock-Chodorow	71.0%
WN: Resnik	**75.6%**
WN: Vector	54.2%

Table 5.4: Accuracy for context-based similarity, similarity slope and WordNet-based (WN) similarity metrics.

[1] http://search.cpan.org/

lexical features. This is a weak indication that the incorporation of taxonomy-based similarity achieves better discrimination between associative and semantic relations.

To further investigate the behaviour of the best performing features, we have plotted their histograms for associative and semantic word pairs. In Fig. 5.1(c), we show the histogram for the priming coefficient $\Psi_5^J(w_i, w_j)$. The priming coefficient for the associative relations tends to be lower than that of semantic relations, especially for larger values of distance m. The histograms of the values of $S_{H=1}^{H=2}(w_i, w_j)$ metric are shown in Fig. 5.1(d). Both histograms have positive means, i.e., context-based semantic similarity increases when going from window size one to size two. However, the increase for associative relations is higher.

We have also combined the best performing features: (i) Ψ_m^G priming coefficient using the G hit-based metric, and (ii) $S_{H_y}^{H_x}(w_i, w_j)$ text-based metric using B scheme, by simply taking the union of their feature sets. This combination achieved slightly higher accuracy of 87.8%. Finally, we report results separately on dataset 2 that contains synonyms as semantic pairs, and compare the results with datasets 1 and 3. The results are presented in Table 5.5. Note that

Features	Set 1, 3	Set 2	All sets
$\Psi_m^G(w_i, w_j)$	87.7%	82.8%	86.5%
$S_{H_y}^{H_x}(w_i, w_j)$	76.1%	56.9%	71.8%
Both features			87.8%

Table 5.5: Accuracy for datasets 1, 3 vs dataset 2.

the accuracy drops for dataset 2 (synonyms) for both the priming coefficient and, especially, the similarity slope. This is an indication that synonyms might be harder to separate from associative pairs; however, due to the limited size of dataset 2 (29 assoc. and 29 sem. relations) no general conclusions can be drawn.

Also, some preliminary results on the classification between semantic and associative relations using linguistic patterns (on the same web corpus) are provided. The most accurate pattern for associative relations is A1 (complex NPs) achieving classification accuracy of 66%. For semantic relations the S1 pattern with coordinative constructions performs better, although its performance is below 60%. When all four patterns were used classification accuracy of up to 73.5% is achieved.

Last, in order to further validate our best performing feature, Ψ_m^G, we used some types of relations taken from the field of semantic role labeling, assuming that they can serve as associative ones. Regarding semantic relations we retained the relations of dataset 1. In particular, we considered four distinct types of relations taken from the SemEval2010–Task 8, "Multi-Way Classification of Semantic Relations Between Pairs of Nominals" Hendrickx et al. [2010]: (i) "Cause–Effect", (ii) "Instrument–Agency", (iii) "Component–Whole", and (iv) "Member–

Collection". For each type of the above relations, we created a distinct dataset including the semantic relations of dataset 1 and an equal number of randomly selected examples. For each dataset we evaluated the proposed priming coefficient regarding the classification of semantic and associative relations. For all datasets the classification accuracy is very similar and exceeds 80%, even for medium values of distance ($m = 3$). These results provide an additional confirmation regarding the good performance of the proposed feature, while they are consistent with the results obtained for the datasets assembled by cognitive scientists.

5.7 Conclusions

Motivated by findings in the psycholinguistics and computational linguistics literature, we investigated the problem of automatically classifying relations between words into either associative or semantic, using information extracted from the web. Two new features were proposed designed specifically for this classification task, namely, the priming coefficient measuring the asymmetry in the order of appearance of the word pair and the first-order difference (slope) of the context-based semantic similarity with respect to the contextual window size. For associative relations the priming coefficient takes significantly smaller values as the distance between the two words increases, while for semantic relations priming is less affected by word distance. For words that are semantically related their contextual similarity is higher for immediate rather than for distant context (small vs. large contextual windows); for associative relations context similarity is less affected by window size. The priming coefficient is shown to be a good feature for discriminating between the two classes, achieving classification accuracy up to 86%. The slope of the contextual similarity achieves good classification results, up to 72% accuracy. Overall, we have shown that it is feasible to classify associative and semantic relations without using lexical or syntactic patterns, but rather general linguistic properties measured through lexical corpus statistics, e.g., order of appearance, co-occurrence, distance, contextual similarity. We make available [1] a resource containing more than 9.000 priming coefficients, computed for the pairs of the experimental datasets.

Further research is needed with larger datasets to verify the universality of these claims. Also special cases of associative and semantic relations should be investigated and the relative performance of the proposed features should be evaluated. The proposed features could be also relevant for investigating the differences between various types of semantic relationships, as well as for studying the priming phenomenon across different languages within the proposed computational framework.

[1] http://www.telecom.tuc.gr/~iosife/downloads.html

Chapter 6

Applications of Network-based DSMs

6.1 Introduction

In this section, we investigate the application of the network-based similarity metrics for three problems. In Section 6.2, visual and textual features are integrated for the creation of multimodal networks [1] . In Section 6.3, the semantic neighborhoods are utilized for determining the semantics of compositional expressions. Specifically, the two main network-based metrics are adapted for estimating the semantic similarity between noun–noun pairs. In Section 6.4, the network similarity metrics are applied to the construction of a simple noun taxonomy.

6.2 Network-based DSMs of Words and Images

Using the distributional hypothesis of meaning for estimating semantic similarity between words is limited to a single modality, i.e., text. However, according to cognitive science the semantic properties of concepts are also determined by the features of other modalities, e.g., color Barsalou et al. [2008]. In this section, a preliminary attempt is presented towards the creation of multimodal network-based DMSs, using features extracted from text and images. The key idea is to use both types of features for the definition of semantic neighborhoods and the computation of network-based similarity metrics as described in Chapter 4. It should be noted that the key idea of this section heavily relies on the work proposed in Bruni et al. [2011]. The contribution of this section deals with the adaptation of the aforementioned idea on the network-based framework.

[1]The visual features used in experiments of this section were kindly provided by Elia Bruni (Center for Mind/Brain Sciences (CIMeC) of the University of Trento). For more details see Bruni et al. [2011].

6.2.1 The Visual Analogue of Bag–of–Words Models

The "Bag-of-Words" (BoW) model is a widely-used approach for implementing unstructured DSMs. The main characteristic of BoW model is that the syntactic relations between the target words and their contextual features are not taken into account. In essence, the contextual features under BoW model can be regarded as a lexicon, which is represented as set of lexical entries. The BoW model often is utilized for the construction VSM [1], where the contextual features are encoded in a matrix, while the similarity between words is estimated using functions of vectors (i.e., rows of matrix). The notion of "Bag-of–Visual–Words" (BoVW) model was inspired by the BoW model in an attempt to represent images with respect to a common "visual lexicon" Bruni et al. [2011]; Csurka et al. [2004]; Sivic and Zisserman [2003]. Given an image collection the following steps are followed for the construction of the BoVW model Bruni et al. [2011]: 1) Salient local regions, e.g., 10×10 pixels, are identified and represented as vectors. Local regions were reported to be more robust to occlusions and spatial variation: compared to global regions Fei-Fei and Perona [2005]. Note that, multiple vectors may be used for each region. In such cases, vectors contain different type of features. The local regions are also referred to as keypoints. 2) The identified keypoints are projected into a space that is shared between the images of the collection. Next, the projections are clustered, while each cluster is considered as a visual word. 3) Every image is then represented as vector that includes such visual words.

6.2.2 Multimodal Network Creation

In this section, the creation of semantic networks is briefly described. What is new here is the network creation based on visual features, since the approach of Chapter 4 was followed regarding the text modality. In particular, the visual features we experimented with were based on the work described in Bruni et al. [2011] in which the VLFeat implementation Vedaldi and Fulkerson [2013] was applied. The extraction procedure is summarized as follows: A standard detector, Difference of Gaussian (DoG) Lowe [2004], was employed for the identification of keypoints and their assignment into visual words. The Scale-Invariant Feature Transform (SIFT) was used for the representation of keypoints by a 128-dimensional vector. SIFT exhibits a number of useful properties: invariance to image scale, orientation, noise, distortion, as well as partial invariance changes of illumination. The k-means algorithm was applied for clustering the detected keypoints into 2000 clusters, i.e., visual words. In order to obtain a more granular analysis of the images the number of visual words was increased by 16. This was performed using a one-level 4×4 pyramid of spatial histograms. This way each image

[1] Note that the BoW model can be also used without adopting the VSM.

was represented by a vector with dimensions corresponding to 32K visual words.

Recall from Chapter 4 that the network creation consists of two main steps: 1) computation of semantic neighborhoods, and 2) computation of similarity scores. Regarding text modality two types of similarity metrics (in conjunction with the respective features) were applied: co-occurrence-based (CC) and context-based (CT). Here, a third type of feature, i,e., visual (VS), is available for both steps, while the similarity between images is estimated using the cosine of their respective features. In this section, we focus only to the cases where VS is used (i) either for Step 1 or Step 2, (ii) for both steps. When VS is used for one step only, the other step can be performed using either CC or CT. In total, there are 5 combinations.

6.2.3 Experimental Networks and Datasets

An important step for the creation of the intended multimodal network is the mapping of the noun lexicon within the image collection. Of course, it is quite difficult to find an one–to–one mapping, i.e., a certain image represents a certain noun. In Bruni et al. [2011], the image collection of the ESP-Game dataset was used, in which each image was annotated with a textual description (set of tags), i.e., multiple tags were allowed per image. In order to obtain a visual representation for each tag the following approach was followed: Each tag was associated to the set of images that were tagged with it. Then, each tag was represented by a vector of visual words computed by summing the vectors of the corresponding images. In total, 11K unique tokens were used as tags. For the creation multimodal networks we used the intersection between our existing noun network consisting of ($8,752$ members) and the set of tags used in Bruni et al. [2011]. This resulted to a set of $3,450$ nouns for which a semantic network was created as described in Section 6.2.2. For evaluation purposes,we used the noun pairs of (i) Rubenstein-Goodenough (RG) Rubenstein and Goodenough [1965] and (ii) WordSim353 (WS353) Finkelstein et al. [2002] datasets which were included the set of $3,450$ nouns: 35 and 175 pairs, respectively.

6.2.4 Evaluation Results

In this section, the performance of the network-based similarity metrics is evaluated against human ratings using Pearson's correlation coefficient.

The baseline performance is presented in Table 6.1 with respect to the co-occurrence-based (CC), contextual (CT) and visual features (VS). For the case of CC the bast performing co-occurrence-based metric was applied: Google-based Semantic Relatedness defined by (2.18). The cosine similarity, defined by (2.20), was employed for the CT and VS features. An immediate context window $K = 1$ was used for CT. It is clear that the co-occurrence feature metric

Feature used for similarity metric	Subset of RG dataset (35 pairs)	Subset of WS353 dataset (175 pairs)
CC	**0.85**	**0.61**
CT	0.67	0.25
VS	0.47	0.33

Table 6.1: Performance of baseline metrics using co-occurrence-based (CC), contextual (CT) and visual features (VS).

outperforms the other types of features for both datasets. No clear winner emerges between the CT and VS features.

Type of feature for		Number of neighbors				
Selection of neighbors	Similarity computation	10	50	100	150	200
Subset of RG dataset (35 pairs)						
CC	VS	0.64	**0.79**	**0.79**	0.70	0.67
CT	VS	0.78	0.76	0.69	0.66	0.65
VS	CC	**0.58**	0.55	0.29	0.36	0.29
VS	CT	0.48	0.42	0.25	0.33	0.27
VS	VS	0.43	0.40	0.23	0.27	0.26
Subset of WS353 dataset (175 pairs)						
CC	VS	0.44	0.59	0.66	**0.70**	0.64
CT	VS	0.44	0.47	0.38	0.32	0.34
VS	CC	**0.47**	0.41	0.37	0.32	0.32
VS	CT	0.34	0.33	0.34	0.28	0.28
VS	VS	0.37	0.30	0.31	0.27	0.27

Table 6.2: Performance for M_n neighborhood-based metric for several number of neighbors.

The performance of the neighborhood-based metric M_n, defined by (4.2), is presented in Table 6.2 for different neighborhood sizes. This is shown for all combinations of textual (CC or CT) and visual (VS) features used for neighbor selection and similarity computation. The main point of interest here is to investigate the performance of the visual features when used either for neighbor selection or computation of the final similarity score (or for both steps). It is clear that the highest performance is obtained when textual features (in particular CC) are used for neighbor selection. Best results are obtained when 50–150 neighbors are taken into consideration. For both datasets the achieved correlation is higher compared to the baseline metric relying of visual features alone. For the case of the WS353 dataset, this

correlation score outperforms all baselines. In addition, seems that small neighborhood sizes yield better results when visual features are used for neighbor selection. Regarding WS353, when visual features are used for both neighborhood selection and similarity computation, the achieved correlation is slightly higher than the respective baseline. The performance of

Type of feature for		Number of neighbors				
Selection of neighbors	Similarity computation	10	50	100	150	200
Subset of RG dataset (35 pairs)						
CC	VS	**0.44**	0.34	0.33	0.35	0.31
CT	VS	0.29	0.37	0.35	0.33	0.33
VS	CC	0.73	0.86	**0.89**	0.88	0.86
VS	CT	0.64	0.67	0.61	0.59	0.58
VS	VS	0.40	0.45	0.44	0.35	0.27
Subset of WS353 dataset (175 pairs)						
CC	VS	0.18	0.24	0.21	0.22	0.23
CT	VS	0.21	**0.28**	0.27	0.25	0.22
VS	CC	0.62	**0.67**	**0.67**	0.65	0.64
VS	CT	0.33	0.26	0.22	0.20	0.19
VS	VS	0.17	0.34	0.33	0.31	0.29

Table 6.3: Performance for R_n neighborhood-based metric for several number of neighbors.

the neighborhood-based metric R_n, defined by (4.3), is presented in Table 6.3 for different number of neighbors. The results are shown for all combinations of textual (CC or CT) and visual (VS) features used for neighbor selection and similarity computation. For both datasets this performance is better compared to all baselines. As in the case of the M_n metric and the WS353 dataset, the achieved correlation is slightly higher that the respective baseline when visual features are used for both neighborhood selection and similarity computation. Unlike the M_n metric, the best performance of R_n is achieved when visual features are used for neighbor selection (for 50–100 neighbors). The poor performance of the M_n metric observed when the neighborhoods are computed according to visual features may be attributed to the following reasons: visual features may include semantically irrelevant neighbors, and only one neighbor is used by the M_n metric for estimating the final similarity. The R_n metric appears to be more robust than M_n for the case of noisy neighborhoods, due to the use of more neighbors according to a averaging-related scheme.

The first research attempt for combining the textual and visual features for the task of semantic similarity estimation was proposed in Feng and Lapata [2010]. The key idea was to

train topic models over a corpus of news articles containing both images and textual content. In particular, word semantics were represented as probability distributions over a set of topics. The similarity between words was estimated using a number of divergence metrics including the Kullback-Leibler and the Jensen-Shannon divergence. For evaluation purposes 254 pairs of the WS353 dataset were used. The bimodal model yielded higher correlation score (0.31) that the text-based model (0.24) [1] . A simpler model for combining textual and visual features was proposed in Bruni et al. [2011] and applied for the same task (estimation of similarity between nouns using 260 pairs taken from the WS353 dataset). The cosine similarity was used as similarity metrics for all modalities. The textual and visual feature vectors were combined by a simple concatenation after normalizing the values of the individual vectors. Overall, the highest correlation (Spearman) score was reported for the bimodal vector (0.51). The textual features yielded higher performance than the visual features: 0.44 and 0.32, respectively.

The proposed network-based DSM was shown to be portable to the use of visual features, given the availability of BoVW model. However, the exploitation of visual features only (for both neighborhood and similarity computation) did not provide a significant improvement of the baseline performance. The most interesting observation regards the synergistic use of textual and visual features. The best performing modality for neighborhood and similarity computation appear to depend on the used similarity metric. More specifically, for the case of the M_n metric, the highest performance was obtained using the textual (co-occurrence-based) features for neighborhood selection and the visual features for similarity computation. Regarding the R_n metric, the highest performance was achieved using the the visual and the textual (co-occurrence-based) features for neighborhood and similarity computation, respectively. Our current findings constitute a promising starting point which is in agreement with the literature of cognitive science: the human semantic knowledge is built and organized on the basis of both verbal and non-verbal information Barsalou et al. [2008]; Rogers and McClelland [2004]. However, the design parameters of the proposed model should be investigated in more depth. Despite the fact that our model enables the incorporation of bimodal features, the underlying steps (i.e., neighborhood and similarity computation) are performed according to unimodal features. This simplification may deviate from the human cognitive system, e.g., multimodal features may be used for either neighborhood selection and similarity computation. In addition, it should be stressed out that the textual and visual features were combined and used for the particular task of word semantic similarity estimation. Our observations need to be further validated with respect to other semantic tasks.

[1] Scores refer to Pearson's correlation coefficient. Note that in Feng and Lapata [2010] the performance of the visual features was not reported.

6.3 Network-based DSMs for Noun–Noun Expressions

A limitation of network-based DSMs presented in Chapter 4 is that the estimation of semantic similarity is considered at the word level only. However, the computation of semantic similarity between phrases or sentences are important for a number of applications, such as grammar induction, paraphrasing and textual entailment. This is related with the principle of *compositionality* stating that the the semantics of a complex (multi-word) expression is determined by the semantics of its constituents. In this section, we investigate the task of semantic similarity estimation between compositional short phrases (more specifically, noun–noun (NN)) within the framework of network-based DSMs. Although, NN are considerably shorter larger textual fragments, e.g., entire sentences, we believe that this is a reasonable way to proceed, i.e., from shorter to larger fragments. In particular, the following two main issues are addressed:

1. How to represent the semantics of compositional expressions.

2. How to exploit the above representation for estimating the semantic similarity between two complex expressions.

One of the main aspects of this effort is the decomposition of the problem into the two aforementioned sub-problems: representation of semantics and similarity estimation. This is somewhat different compared to the majority of compositional models that are mainly focused on the second part investigating functions applied on feature vectors Mitchell and Lapata [2010]. A recent approach that is similar to our perspective is discussed in Turney [2012]. The main idea proposed by Peter Turney is the exploitation of two distinct models for the problem of estimating the semantic similarity between two compositional phrases. The first model (referred to as the "domain space"), and it is meant for representing the semantics of the constituents. The purpose of the second model (referred to as the "function space") is to represent the modifications of meaning that take place for the case of compositional phases, e.g., for "traffic light", "traffic" modifies the meaning of "light". The domain space is built following the typical procedure of VSM, however, only nouns are considered as contextual features. The motivation for this filtering is that the domain or topic of a word is determined by the nearby occurring nouns. The function space is created as the domain space by considering only verb-based patterns that occur in close proximity with the target word. The motivation for employing only verb patterns is hypothesis that the function/role a word is captured by the verbs that occur near it.

6.3.1 Representation of Compositional Semantics and Similarity Metrics

As an example of a simple compositional complex expression consider a NN denoted as $c_i = (c_{i1} \;\; c_{i2})$, where c_{i1} and c_{i2} are its respective constituents. Also, let N_{i1} and N_{i2} be the semantic neighborhoods of c_{i1} and c_{i2}, respectively. Following the principle of compositionality we hypothesize that the semantics of c_i are determined by the semantic neighborhoods of its constituents Baldwin [2006]; Frege [1884]. Also, we expect that the meaning of the NN would be more specific compared to the meaning of its parts. More specifically, we assume that the composed meaning will be related with the shared meaning of the constituents. Hence, we assume that the semantics of c_i can be captured by considering the overlap between the semantic neighborhoods N_{i1} and N_{ij}. If the neighborhoods are represented as sets, we define a hybrid neighborhood $N_i^{'}$ for c_i computed by taking the intersection of N_{i1} and N_{i2}, i.e., $N_i^{'} = N_{i1} \cap N_{i2}$. Extension model Murphy [2002] can be considered as a cognitive analogue for the proposed neighborhood intersection, under the assumption that neighbors reflect semantic features. The basic idea is that each individual concept is represented by a set of semantic features, while the composed semantics of multiple concepts is driven by "extending" the individual set of features (i.e., gradually considering more features) until the convergence to a "sufficient" overlap of features. This is also related with the work in the framework of prototype theory regarding the definition of composite concepts on the basis of simple ones Osherson and Smith [1981]. According to Osherson and Smith [1981], the composite semantics of two concepts can be determined by a function that takes into account: (i) the degree according to which the one concept falls in the (semantic) extension of the other, and (ii) the relatedness between the one concept and the prototype (concept) of the other.

It should be stressed out that although the aforementioned approach may be valid for the particular case of NN, different models may apply for other types of multi-word expressions. For example, see Baroni and Zamparelli [2010] for the distributional representation of adjective–nouns (AN), where the semantics of adjectives are modeled via a linear function from a vector (noun representation) to another vector (AN representation). A survey of compositional models and the underlying theoretical background, with references to different types of multi-word expression, is presented in Baroni et al. [2013]; Mitchell and Lapata [2010]; Turney [2012].

Using the above considerations we show how the network-based metrics M_n and R_n, (defined in Chapter 4 by (4.4.2) and 4.4.3, respectively) can be adapted for the estimation of semantic similarity between two (compositional) NN: $c_i = (c_{i1} \;\; c_{i2})$ and $c_j = (c_{j1} \;\; c_{j2})$.

Compositional Maximum Similarity of Neighborhoods

The similarity between $c_i = (c_{i1}\ \ c_{i2})$ and $c_j = (c_{j1}\ \ c_{j2})$ can be estimated as:

$$M_n^{'}(c_i, c_j) = \max\{\phi_{ij}, \phi_{ji}\}, \tag{6.1}$$

where

$$\phi_{ij} = \max_{x \in N_j^{'}} \frac{1}{2}\Big(S(c_{i1}, x) + S(c_{i2}, x)\Big), \quad \phi_{ji} = \max_{y \in N_i^{'}} \frac{1}{2}\Big(S(c_{j1}, y) + S(c_{j2}, y)\Big).$$

ϕ_{ij} (or ϕ_{ji}) denotes the maximum similarity between c_i (or c_j) and the neighbors of c_j (or c_i). $N_i^{'}$ and $N_i^{'}$ are the hybrid (i.e., the result of intersection) neighborhoods of c_i and c_j, respectively. S is a similarity metric as defined in Chapter 4.

Compositional Correlation of Neighborhood Similarities

The similarity between $c_i = (c_{i1}\ \ c_{i2})$ and $c_j = (c_{j1}\ \ c_{j2})$ can be estimated as:

$$R_n^{'}(c_i, c_j) = \max\{\kappa_{ij}, \kappa_{ji}\}, \tag{6.2}$$

where

$$\kappa_{ij} = \rho(C_i^{N_i^{'}}, C_j^{N_i^{'}}), \quad \kappa_{ji} = \rho(C_i^{N_j^{'}}, C_j^{N_j^{'}})$$

and

$$C_i^{N_i^{'}} = \Big(\frac{1}{2}\big(S(c_{i1}, x_1)+S(c_{i2}, x_1)\big), \frac{1}{2}\big(S(c_{i1}, x_2)+S(c_{i2}, x_2)\big), \ldots, \frac{1}{2}\big(S(c_{i1}, x_m)+S(c_{i2}, x_m)\big)\Big)$$

where

$$N_i^{'} = \{x_1, x_2, \ldots, x_m\}.$$

Note that $C_j^{N_i^{'}}$, $C_i^{N_j^{'}}$, and $C_j^{N_j^{'}}$ are defined similarly to $C_i^{N_i^{'}}$. The ρ function stands for the Pearson's correlation coefficient, $N_i^{'}$ is the hybrid neighborhood of NN c_i, and S is a similarity metric as defined in Chapter 4.

6.3.2 Experiments and Evaluation Results

For the evaluation of the composition network-based metrics we used a dataset of NN pairs Mitchell and Lapata [2010]. In total, 108 pairs are included, which were rated by human subject regarding their semantic similarity in a 1–7 scale. A number of examples of NN pairs

Pair	Similarity degree
marketing director–assistant manager	high
telephone number–phone call	high
capital market–future development	medium
research contract–training programme	medium
bedroom window–education officer	low
league match–family allowance	low

Table 6.4: Examples of NN pairs Mitchell and Lapata [2010].

along with their respective similarity degree is presented in Table 6.4. In our experiments we used the pairs whose constituents were included in our network of $8,752$ nouns: 92 out of 108 pairs. The performance of the compositional network-based similarity metrics, was evaluated against human ratings using Spearman's correlation coefficient [1] . Note that the average inter-annotator agreement computed in terms of correlation coefficient can be regraded as the upper bound for the performance of metrics. Regarding the 92 pairs, this bound equals to 0.66

Type of feature for		Number of neighbors				
Selection of neighbors	Similarity computation	10	50	100	150	200
$M_n^{'}$ metric						
CC	CC	0.19	0.52	0.57	**0.62**	0.60
CC	CT	0.05	0.31	0.24	0.37	0.36
CT	CC	0.23	0.39	0.28	0.25	0.17
CT	CT	0.07	0.21	0.12	0.08	0.01
$R_n^{'}$ metric						
CC	CC	-0.04	-0.01	0.39	0.26	0.40
CC	CT	-0.02	0.10	0.08	-0.08	-0.17
CT	CC	0.10	0.36	0.59	**0.60**	0.55
CT	CT	-0.16	-0.02	0.07	0.08	0.14

Table 6.5: Performance for $M_n^{'}$ and $R_n^{'}$ neighborhood-based metrics for several number of neighbors. Performance upper bound: 0.66 (average inter-annotator agreement).

As in the case of word-level semantic similarity estimation (see Chapter 4) two basic steps are required: 1) computation of semantic neighborhoods, and 2) computation of similarity scores (the S metric in (6.1) and (6.2)), resulting into the following combinations:

[1] Spearman's correlation was used in order to compare the results with the literature. Note, that almost identical performance was observed when using Pearson's correlation coefficient.

- Compute neighborhoods and similarity scores using a co-occurrence-based metric (CC/CC).
- Compute neighborhoods using a co-occurrence-based metric; compute similarity scores using a context-based metric (CC/CT).
- Compute neighborhoods using a context-based metric; compute similarity scores using a co-occurrence-based metric (CT/CC).
- Compute neighborhoods and similarity scores using a context-based metric (CT/CT).

For the above approaches, the co-occurrence-based metric G (Google-based Semantic Relatedness defined by (2.19)) and the context-based metric $Q^{H=1}$ (defined by (2.20)) were used.

The performance of $M_n^{'}$ and $R_n^{'}$ is presented in Table 6.5 for a different sizes of neighborhoods. First, it is interesting to observe that both metrics perform quite well give the performance upper bound (0.66). However, the highest score (0.62) is achieved by $M_n^{'}$. For both metrics the highest correlation is reached for 150 neighbors. This is observed when the (CC/CC) and (CT/CC) combinations are used for $M_n^{'}$ and $R_n^{'}$, respectively. This is consistent with the best performance of (CC/CC) and (CT/CC) combinations for M_n and R_n, respectively, for the case of single word similarity estimation (see Table 4.5 in Chapter 4). However, the (CC/CT) combination appears to yield poor performance when applying the $M_n^{'}$ metric, as opposed to the corresponding metric (i.e., M_n) for the case of single words. This is an indication that the context plays a lesser role for longer expressions.

In Mitchell and Lapata [2010], nine compositional models were applied for estimating the similarity between NN pairs. In particular, the underlying idea of these models regarding the representation of the (compositional) semantics of a multi-word expression is the application of a function over the vectors that represent the semantics of constituents (i.e., single words). The latter were constructed according to the typical VSM. Given a NN phrase, every composition model resulted into a (single) vectorial representation. The similarity between two NN phrases was estimated as the cosine of their respective vectors. The models used in Mitchell and Lapata [2010] were evaluated using a set of 108 NN pairs. The highest performance (0.49 Spearman's correlation coefficient [1]) was reported for a model based

[1] A different methodology was followed when computing the correlation coefficient. The usual approach (used in the literature regarding the standard datasets of similarity tasks) is to average the human scores, and then compute the correlation coefficient between the averaged scores and the scores estimated by the experimental models. In Mitchell and Lapata [2010], it is reported only that the human scores were not averaged. In Turney [2012], more details are given about the evaluation based on a personal communication between Peter Turney and Jeff Mitchell. The basic idea is: x people rated y pairs, yielding $x \times y$ ratings, which were vectorized. However, one score for each pair was estimated by the used models. In order to make the computation of correlation feasible, the model's scores of were duplicated x times.

on vector multiplication, which was equal to the upper bound of performance (i.e., the average inter-annotator agreement). In Turney [2012], the similarity between two NN was computed within the framework of a dual-space model that combines domain and function models: $S(c_i, c_j) = g(S_d(c_{i1}, c_{j1}) + S_d(c_{i2}, c_{j2} + S_f(c_{i1}, c_{j1}) + S_f(c_{i2}, c_{j2}))$, where the $g(.)$ function denotes the geometric mean. $S_d(.)$ and $S_f(.)$ stand for the similarities estimated over the domain and function models, respectively. In Turney [2012], the same dataset was used as in Mitchell and Lapata [2010], however, the upper bound of performance was computed as the leave-one-out (Spearman's) correlation between the human ratings. The aforementioned dual model yielded 0.54 correlation, which it was reported to be equal to the upper bound of performance.

Overall, both co-occurrence and contextual features used for neighbor selection yield comparable performance. However, the utilization of these features appears to depend on the used similarity metric: co-occurrence-based for the $M_n^{'}$ metric and context-based for the $R_n^{'}$ metric. A limitation of the proposed approach is that the word order in not taken into account for the computation of the hybrid neighborhood, as well as for estimating the semantic similarity between NN. The sensitivity of (computational) semantic models to word order is an open research issue Turney [2012].

6.4 Network-based DSMs for Simple Taxonomy Creation

In this section, the network-based similarity metrics described in Chapter 4 are applied for the creation of simple taxonomy of nouns. In particular, the ESSLLI dataset Baroni et al. [2008] was used, which constitutes a three-level taxonomy depicted by 6.1. The lowest level of the taxonomy consists of (instances of) the following six concepts: (i) "birds", (ii) "land animals", (iii) "greens", (iv) "fruits", (v) "tools", and (vi) "vehicles". The middle level includes the concepts (i) "animals", (ii) "vegetables", and (iii) "artifacts", while the upper level is distinguished into "living beings" and "objects".

The original ESSLLI dataset consists of 44 nouns (instances). We used the subset of those nouns that was covered by the network of $8,752$ nouns presented in Chapter 4: 31 nouns (instances). For each taxonomic level, the network-based metrics were applied for the construction of a similarity matrix upon which the k-means clustering algorithm was incorporated. The purity of clusters, P, was used as evaluation metric, defined as Baroni and Lenci [2010]:

$$P = \frac{1}{c} \sum_{i=1}^{k} \max_{j}(c_i^j), \tag{6.3}$$

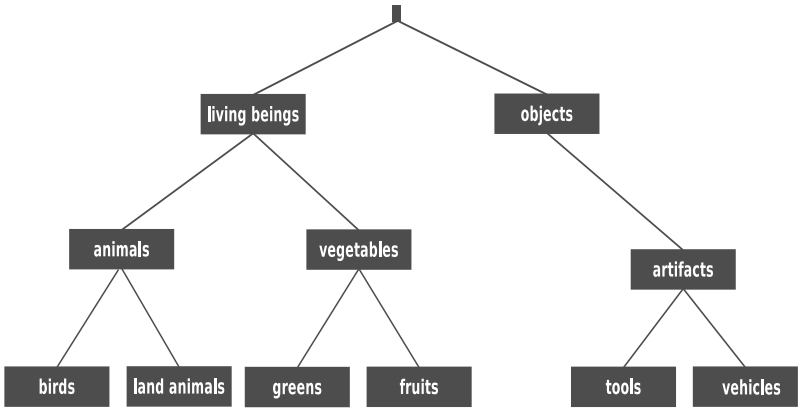


Figure 6.1: Taxonomy of ESSLLI dataset Baroni et al. [2008].

where c_i^j is the number of nouns assigned to the i^{th} cluster that belong to the j^{th} groundtruth class. The number of clusters is denoted by k, while c is the total number of nouns included in the dataset. Purity expresses the fraction of nouns that belong to the true class, which is most represented in the cluster Baroni and Lenci [2010], taking values in the range $[0, 1]$, where 1 stands for perfect clustering.

The performance of M_n and R_n is presented in Table 6.6 for different sizes of neighborhoods, with respect to the three taxonomic levels: top–medium –low. For the above approaches, the co-occurrence-based (CC) metric G (Google-based Semantic Relatedness defined by (2.19)) and the context-based (CT) metric $Q^{H=1}$ (defined by (2.20)) were used. The performance of baseline co-occurrence-based and context-based metrics is also shown. We observe that both network-based metrics outperform the baseline purity, while their performance is comparable. For both M_n and R_n the highest purity scores are achieved for neighborhoods including 50 – 150 members. Regarding the M_n metric, the highest results are obtained when CT is used for the selection of neighbors, for which the use of CC and CT for the computation of similarity have comparable performance. For the case R_n the best purity is yielded when CT is utilized for both neighbor selection and similarity computation. Overall, the achieved purity scores are comparable with the best performance reported in the literature Baroni and Lenci [2010], where structured DSMs where employed for creating the similarity matrix upon which the k-means clustering algorithm was applied. Unlike the task of similarity estimation (see Section 6.2 and Section 6.3, as well as Chapter 4), for this task the CT feature appears to perform better than CC regarding the estimation of similarity. This is also observed for the baseline performance. This difference may be attributed to the similarity matrix with which

Type of feature for		Number of neighbors				
Sel. of neigh.	Sim. comp.	10	50	100	150	200
Baseline CC: 0.65-0.68-0.71 Baseline CT: 0.65-0.87-0.77						
M_n metric						
CC	CC	0.55-0.68-0.74	0.84-0.68-0.81	0.58-0.74-0.84	0.81-0.84-0.74	0.77-0.94-0.77
CC	CT	0.55-0.61-0.58	0.68-0.68-0.77	0.55-0.81-0.81	0.84-0.87-0.74	0.61-0.90-0.74
CT	CC	0.81-0.90-0.87	**1-0.94-0.90**	1-0.94-0.84	1-0.94-0.87	0.90-0.90-0.81
CT	CT	0.77-0.94-0.84	0.97-0.94-0.84	1-0.90-0.81	0.94-0.90-0.84	0.97-0.90-0.84
R_n metric						
CC	CC	0.58-0.61-0.45	0.61-0.61-0.61	0.81-0.77-0.77	0.77-0.81-0.74	0.81-0.74-0.77
CC	CT	0.58-0.71-0.71	0.58-0.65-0.61	0.61-0.65-0.61	0.61-0.61-0.68	0.61-0.58-0.65
CT	CC	0.68-0.61-0.55	0.71-0.87-0.68	0.71-0.81-0.87	0.65-0.81-0.87	0.65-0.81-0.87
CT	CT	0.87-0.81-0.71	**1-0.94-0.90**	1-0.94-0.81	1-0.94-0.81	1-0.94-0.81

Table 6.6: Performance for M_n and R_n neighborhood-based metrics for different number of neighbors. Purity is shown for all three levels of the taxonomy: top - medium - low.

the clustering algorithm is fed. In general, it was observed that the use of CT features tend to result into similarity matrices that are less sparse compared to the use of CC features.

6.5 Conclusions

The integration of textual and visual features for the creation of multimodal semantic networks yielded promising results. In particular, we improved on the unimodal baseline performance when both types of features were used by the proposed network-based similarity metrics (WS353 dataset). However, the end–to–end use of visual features seems to result into noisy networks that do not achieve top performance. The network-based DSMs appear to the compositionality framework (at least for the case of noun–noun). The simple idea of taking the intersection of semantic neighborhoods in order to represent the semantics of compositional expressions for the case of NN proved quite effective. Also, the adaptation of the word-level network similarity metrics yielded quite high performance; close to the upper bound as indicated by the average inter-annotator agreement. However, the applicability of this approach needs to be further investigated using larger and more complex expressions, e.g., AN.

The network-based similarity metrics were also applied to the construction of a simple three-level taxonomy of nouns, obtaining quite good performance. For this task, the utilization of contextual features appeared to perform slightly better compared to the co-occurrence based features. This suggests that the relative performance of features may vary according to the task under investigation. Last but not least, larger taxonomies can be used for the justification of the

aforementioned observations.

Chapter 7

Conclusions and Future Research

In this section, the main contributions and conclusions of this work are discussed. Also, the related ongoing work is briefly presented, while interesting future directions are outlined.

7.1 Main Contributions and Conclusions

The main contributions of this work deal with the creation of language-agnostic DSMs using web harvested data. The "language agnostic" characterization refers to the fact that no language-specific features (and related tools) are employed by the underlying algorithms. It was shown that the web is a valuable source for corpora creation. More specifically, a query-based approach was employed for harvesting web data. The scalability of this approach was also investigated with respect to a large lexicon. It was shown that the massive aggregation of web data enables the representation of less-dominant word senses within corpora. A number of parameters of DSMs were investigated for the task of semantic similarity estimation between words, including the extraction and weighting of contextual features. In addition, a network-based implementation of DSMs was proposed in combination with three novel similarity metrics, motivated by the assumptions of maximum sense similarity and attributional similarity. The network-based DSMs were extended towards creation of multimodal networks based on textual and visual features. and the estimation of similarity beyond the word level. Finally, motivated by the literature of cognitive science we investigated the discrimination of associative versus semantic relations, and the performance of the proposed network similarity metrics with respect to the word concreteness/abstraction. Next, we discuss the main conclusions of this thesis.

7.1.1 DSMs based on Vector Space Model

The typical implementation of DSMs is the widely-used VSM, which relies on the word-context matrix. In this framework, we investigated a number of query types submitted to web search engines for corpora creation. Very good results were obtained using conjunctive AND queries according to which the co-occurrence of word-pairs in the same document was explicitly requested. The word co-occurrence was regarded as a semantic filter that retains the relatedness of word senses. Towards this direction we demonstrated that the similarity estimates are more accurate if we consider the two closest senses, i.e., the maximum pair-wise sense similarity score. In the contrary, the employment of individual queries for corpora creations was shown to yield poor results. This was attributed to the lack of sense coverage in the corpus.

We investigated a number of parameters for context-based similarity metrics including the context window size and the relative weighting of contextual features. For the case of nouns, we found that the very immediate context encoding mostly syntactic dependencies yielded better performance. Also, the simple binary weighting of contextual features was observed to have comparable performance with other frequency-based weighting schemes. Unlike (single-word) common nouns, the highest performance regarding the case of multi-word medical terms [1] was obtained for larger context window. The use of narrow window size implicitly imposes a strong syntactic filtering, according to which certain types of contextual features are captured, (mainly) including function words (e.g., articles and conjunctions), nouns and adjectives. Such feature types may exhibit different roles in the representation of noun semantics. Unlike common nouns biomedical terms are not expected to have multiple senses. The high performance of larger window size (compared to common nouns) indicates that the pragmatic information is essential for the their semantic representation. The binary scheme used for feature weighting in the cosine similarity metric is related to a simple cognitive model presented in Ingram [2007]. According to this model the similarity between two concepts can be estimated by considering the overlap of their semantic attributes, which are assigned binary values.

In addition to context-based similarity estimation, we investigated several well-established similarity metrics that rely directly on the co-occurrence of the words under investigation. We observed that the critical factor is the proximity of the co-occurring words. Many research efforts consider word co-occurrence at the document level due to the straightforward use of number of hits returned by web search engines. However, smaller linguistic contexts, such as sentences, appear to better reflect the semantic relatedness of the co-occurring words. We believe that co-occurrence and the close proximity of words helps the development of word as-

[1] A non-compositional approach was followed for the multi-word medical terms.

sociations, which constitute the primary information type upon which more complex semantic representations are built. In Kahneman [2011], a dual-system framework is described regarding the processing of knowledge during semantic tasks: a system of associative relations is rapidly activated by experienced stimuli, while the associations are further processed by a subsequent system of semantic nature. The close proximity of co-occurring words seems to agree with the limited capacity of the immediate (also referred to as working) human memory: experimental findings suggest that certain number of information chunks can be efficiently processed by the (typical) human memory system Miller [1955].

7.1.2 Network-based DSMs

Despite the good performance of corpora constructed by conjunctive AND queries, the underlying methodology is not efficient for estimating the pairwise similarities between the entries of a lexicon of size N. This is due to the quadratic query complexity $\mathcal{O}(N^2)$. In order to tackle this limitation in scalability, we proposed a method for corpora creation exhibiting linear query complexity with respect to N. The key idea was the employment thousands of individual queries (one query for each entry of the lexicon) and the aggregation of the harvested data. Such a corpus has one basic difference compared to typical corpora: the frequency of word occurrence deviates from Zipf's law. This idiosyncrasy of the resulting corpus was beneficial for the task of similarity estimation: the domination of very frequent words was smoothed, while rare words were better represented within the corpus. This observation is important given that the senses of a word do not occur equiprobably. More specifically, the conditional probabilities of word senses appear [1] to follow a power-law. The sense coverage of corpora depend on the methodology followed for their creation, while it is critical for DSMs. The aforementioned aggregation of data enables the discovery of less frequent senses for polysemous words, given that a large lexicon L is used. The basic idea is that instances of a word w_i can be found implicitly, i.e., within data retrieved for w_j, where $w_i \neq w_j$. This applies when w_i is a polysemous and w_j stands as an infrequent (lexicalized) sense of w_i. Given this corpus, the typical context-based cosine similarity metric yielded poor performance due to sense disambiguation issues. The notion of semantic neighborhoods was introduced in order to better capture the semantics of the words of interest. The members of neighborhoods were found to encode a variety of lexical relations including synonymy, taxonomic relations, as well as a long list of associative relations. More specifically, we investigated both co-occurrence-based and context-based metrics for the creation of semantic neighborhoods. We observed that the neighbors captured by co-occurrence-based metrics tended to formulate more direct associative relations with the

[1] This observation was made after analyzing the SemCor3 sense-tagged corpus Iosif and Potamianos [2013a].

reference nouns. The presence of relations of a broader semantic/pragmatic scope was stronger for the neighborhoods computed by contextual metrics. In addition, the neighborhoods created using co-occurrence-based metrics were found to have greater synonym coverage.

Based on semantic neighborhoods, three novel metrics of semantic similarity were proposed. These metrics were motivated by the assumptions of maximum sense similarity and attributional similarity. According to the first assumption the most salient information in the neighborhood of a word are semantic features denoting senses of this word. We believe that the space of semantic neighborhood can be break down into multiple "sub-spaces" of lower dimensionality. In addition, we assume that such sub-spaces reflect the semantic of word senses. For the semantic neighborhoods used in this thesis, this claim is currently supported by a number of preliminary results [1] for the task of word similarity estimation. The integration of the multiple low-dimensional spaces into a global representation constitute an equally challenging subsequent step. The motivation behind the second assumption is that neighborhoods encode semantic attributes of words. Also, the underlying assumption is that two semantically similar words are expected to have co-varying similarities with respect to their neighbors. The cognitive analogue of this assumption can be found in the PDP approach Rogers and McClelland [2004] according to which semantically related items (i.e., concepts lexicalized by words) are characterized by "coherent sets of multiple properties that all covary reliably". However, our sequential/flat approach does not follow the PDP framework, since each neighborhood is treated as a single set (of properties), which is not semantically coherent. In total four combinations of co-occurrence and context features were investigated for the computation of the proposed metrics consisting of two phases: computation of semantic neighborhoods and similarity score. It was observed that the best performing types of features vary with respect to the underlying assumption. For example, the highest results regarding the maximum sense similarity assumption were obtained when the semantic neighborhoods were defined using co-occurrence metrics. In general, the best performing neighborhood size was observed to depend on the adopted assumption: larger neighborhoods for the maximum sense similarity assumption compared to the assumption of attributional similarity. Overall, the proposed network-based metrics outperformed the respective baselines.

In addition, the proposed approach for network-based DSMs was extended across textual and visual features. The integration of textual and visual features yielded promising results, e.g., for the WS353 dataset the unimodal baseline metrics were outperformed when both types of features were used. This observation seems to be in agreement with the cognitive assump-

[1] Work conducted by Georgia Athanasopoulou (ECE Department, Technical University of Crete) based on existing and novel dimensionality reduction algorithms. Also, similar ideas are discussed in Karlgren et al. [2008] even without strong experimental evidence.

tion suggesting that knowledge from different modalities is fused into a common semantic representation Rogers and McClelland [2004]. However, such a fusion is not truly performed within the proposed bimodal network, since each of the underlying steps (i.e., neighborhood and similarity computation) rely on unimodal features. We believe that the study of multimodal features should include the aspect of semantic abstraction/concreteness, for which the influence of verbal and non-verbal experience is assumed to differ.

The network-based DSMs were also adapted within the compositionality framework. The semantic neighborhoods were exploited for representing the semantics of compositional expressions The main network-based metrics were adapted to the composite neighborhoods for estimating the semantic similarity between two complex expressions. Very good performance was achieved, however, the applicability of this approach needs to be further investigated with respect to more complex expressions. The proposed network-based metrics were also applied for the creation of a simple three-level taxonomy of nouns. The excellent performance of similarity metrics with respect to the upper level (living beings vs. object) is in agreement with the cognitive theories about the process of knowledge acquisition. More specifically, coarse-grained distinctions are reported to be acquired before finer-grained distinctions, while basic categories are expected to be "maximally informative and distinctive" Mandler [2002]; Rogers and McClelland [2004]. A difference regarding our network-based model is that it can not demonstrate the developmental process of semantic representation as in the case of the PDP framework Rogers and McClelland [2004]. Nevertheless, this process can be simulated by our approach given the appropriate selection of the underlying lexicon, i.e., by using a gradually enriched lexicon aimed to surrogate the development of (the typical) mental lexicon that takes place from childhood to adulthood.

7.1.3 Cognitive Aspects of Lexical Semantics

Motivated by findings in the psycholinguistics and computational linguistics literature, we investigated the problem of automatically classifying relations between words into either associative or semantic, using information extracted from the web. We proposed the priming coefficient, which was shown to be a good feature for discriminating between the two classes. The priming coefficient was also applied with respect to discrimination between synonymy and some types of relations taken from the field of semantic role labeling: "Cause–Effect", "Instrument–Agency", "Component–Whole", and "Member–Collection". Quite high results were obtained for binary classifications. Moreover, the performance of the proposed network-based similarity metrics was investigated for the case of abstract and concrete nouns. A "concreteness effect" was observed , i.e., performance for concrete nouns was better than for ab-

stract noun pairs. In addition, abstract concepts were best modeled using an attributional network DSM with small semantic neighborhoods.

7.1.4 Summary

Regarding the typical DSMs, the feature of co-occurrence was shown to yield higher performance than contextual features for the task of similarity estimation, given that co-occurrence was considered at the sentence level. However, the relative performance of these two features was reversed for categorical tasks. The proposal of network-based DSMs constitutes the major contribution of this thesis, where both co-occurrence and context features were used for defining the semantic neighborhoods and estimating similarities. The exploitation of word co-occurrence resulted into semantically more coherent neighborhoods. Given such neighborhoods, the network metric relying on the assumption of maximum sense similarity obtained the highest results for the task of similarity estimation. For the case of single-word nouns the performance of both co-occurrence and contextual features appeared to be comparable when used for the step of similarity estimation. However, for the case of longer phrases (noun–nouns) good performance was obtained only when using co-occurrence-based features. Moreover, a "concreteness effect" was observed for the aforementioned metric: higher performance was achieved for semantically concrete nouns as opposed to the case of abstract nouns. The integration of visual and textual features within the network-based DSMs led to promising results: the performance of network-based metrics slightly exceeded the unimodal baselines. However, the saliency of the two feature types was shown to be uneven, since semantically more relevant neighborhoods were computed by the textual features.

7.2 Ongoing Research and Future Directions

Minimum Error Similarity. According to the assumption of maximum sense similarity the similarity of two words can be estimated as the minimum pairwise similarity of their senses. Although words often co-occur with their closest senses, word occurrences correspond to all senses. So, the denominator of the typical co-occurrence-based metrics is overestimated causing underestimation error for similarities between polysemous words.

Knowing that the probability of word occurrence follows the Zip's law, we empirically investigated the validity of this observation for the case of word senses. Consider the probabilities of senses $p(s_{ik}), k = 1, ..., N_i$ for a certain word w_i. To do so, we estimated the average probability of word senses $\langle \hat{p}(s_{ik}) \rangle_i$ for certain values of N_i, across the words of a lexicon L. This was performed using maximum likelihood estimation for the polysemous nouns of

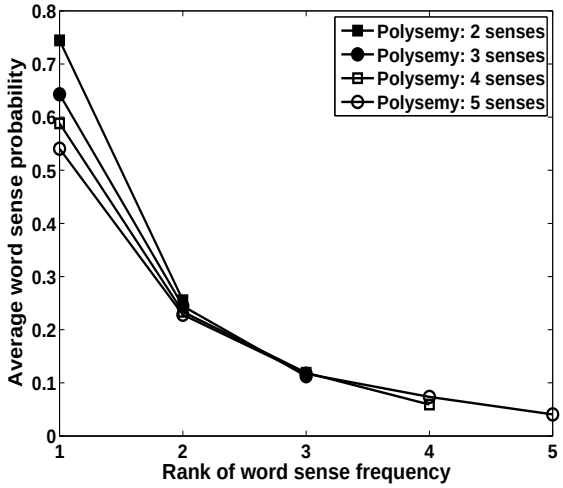


Figure 7.1: Average probability of word senses for different degrees of polysemy (number of senses) as a function of the rank of word sense frequency.

the SemCor3 sense-tagged corpus occurring with N_i ranging between 2 and 5. The average probability of word senses is depicted in Fig. 7.1 as a function of the rank of sense frequency for several values of N_i (degrees of polysemy). It is obvious that the senses of a word do not occur equiprobably. In the contrary, the probabilities of word senses appear to follow a power-law. For example, for $N_i = 3$, the most frequent sense (on average) corresponds to the 64% of the word probability mass. Moreover, for N_i greater than 3 the depicted distribution of less frequent senses seems not to follow Zipf law. Motivated by the above observation and adopting the assumption of maximum sense similarity, a modified version of the point–wise mutual information was defined as $\log \frac{p(w_i \wedge w_j)}{(p(w_i)p(w_j))^{\gamma}}$, where $p(w_i)$ and $p(w_j)$ are the occurrence probabilities of words w_i and w_j, respectively, while the probability of their co-occurrence is denoted by $p(w_i \wedge w_j)$. The exponential weights γ was introduced in order to reduce the contribution of $p(w_i)$ and $p(w_j)$ in the similarity metric. The effect of γ for the task of noun similarity is shown in Fig. 7.2 as a function of the Pearson's correlation coefficient with respect to human ratings. The three standard datasets (MC, RG, and WS353) were used, while the highest correlation score was obtained for $\gamma = 0.90$ for all datasets. Based on the above considerations we aim to investigate a machine leaning-based approach in order to learn the optimal weight using generic features, such as word occurrence and co-occurrence frequencies. The correlation coefficient with the human ratings can be used as the basis of the error criterion. Overall, this line of research constitutes one of the main dimensions of our future work Iosif and Potamianos [2013a].

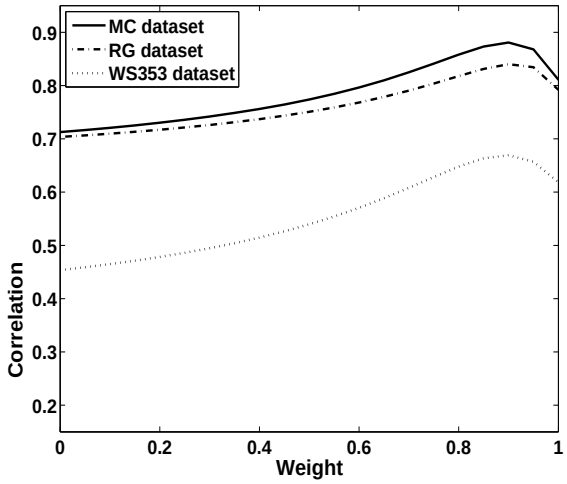


Figure 7.2: The effect in performance of weight γ for the task of similarity rating with respect to MC, RG, and WS353 datasets.

Graph-based Algorithms. In principle, several graph-based algorithms can be applied on the proposed network-based implementation of DSMs. However, the basic question concerns the interpretation of such algorithms within the framework of lexical semantics, i.e., what type of semantic information can be revealed. We believe that such approaches can be considered at two broad levels, namely, global and local. The entire network is considered for the case of global analysis, which is expected to provide useful cues regarding the overall structure. During a preliminary analysis, the PageRank and HITS link analysis algorithms were applied over the entire network. For example, it was interesting to observe that words such as "business" and "money" exhibited quite high PageRank scores. However, a series of relevant questions remain open including the interpretation and exploitation of hub and authority scores with respect to DSMs. Local analysis can be formulated at the neighborhood level motivated by the observation that local properties differ across words. Different words are expected to have different neighborhood statistics, e.g., the distribution of similarities between the targets and the respective neighbors. We suggest that such differences should be taken into account for relevant tasks, e.g., the estimation of semantic similarity. We have investigated two simple normalization schemes for network creation using both co-occurrence-based and context-based similarities. For the case of context-based similarities an improvement was obtained for the tasks of noun similarity estimation and classification. Another example of local analysis is the identification of cliques within neighborhoods. A direct utilization of this analysis is the discovery of word senses under the assumption that each clique denotes the different (at some

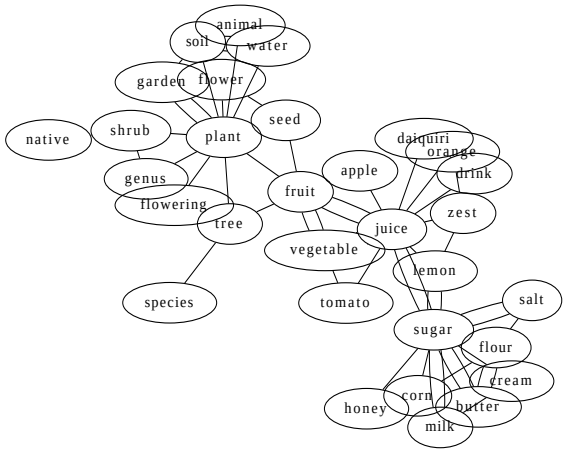


Figure 7.3: Example of cliques for the neighborhood of “fruit”.

extent) senses of the target. An example of such cliques is presented in Fig.7.3, which were derived from the semantic neighborhood of “fruit”. On this basis, various similarity metrics can be investigated following the maximum sense similarity assumption or any other related approaches. Overall, we believe that the exploitation of the network structure enables a number of different perspectives that may differ across tasks, ranging from the simple visualization of concepts up to the discovery of salient semantic features.

Multilingual Networks. The proposed approach for the creation of network-based DSMs can be extended to multilingual linked networks. A common reference point can be formulated using a shared lexicon. Such a lexicon can be regarded as a list of concepts for each language of interest, i.e., parallel monolingual lexicons. The construction of each monolingual network can be a distinct process depending only on the respective lexicon. The most challenging aspect of this idea is the linking of the individual networks. The linking procedure can be possibly driven by the mapping between concepts as defined in the shared lexicon. Of course, a number of issues should be carefully addressed, such as the mappings other than one-to-one. Clearly, the linking complexity is quadratic with respect to the number of languages. Thus, the case of bilingual networks seems a realistic starting point. We expect that the linked network will be semantically richer compared to its constituents. This enables the investigation of relevant network similarity metrics under the hypotheses that semantic features are encoded by semantic neighborhoods and that a fraction of them are universal (i.e., exist in both networks). In addition, more sophisticated analysis can be performed about the semantic diversity (and other related aspects) of the linked networks. Such diversities may occur due to various extra-linguistic factors, such as cultural differences.

Compositionality. The preliminary work on the estimation of similarity between nouns–nouns worths to be extended to the case of larger textual fragments, i.e., sentences. Under the framework of network-based DSMs, one of the primary questions is about the computation of semantic neighborhoods for such large fragments. For example, how to define the (composite) semantic neighborhood for a certain sentence. The consideration of all respective neighborhoods (i.e., the neighborhoods of component words) via set operations like intersection, is likely to result into poorly populated neighborhoods, especially for relatively large fragments. Also, the grammatical dependencies between words are ignored by such a flat approach. A hierarchical approach seems to be a more principled solution to this problem. The key idea is the definition of semantic neighborhoods at different sentence levels. A range of relevant tools, from unsupervised chunkers up to trainable parsers, can be employed for the hierarchical representation of sentences. However, a number of other issues, which are independent to the framework of network-based DSMs, remains open. The most critical problem deals with the alignment of the sentences under investigation, i.e., which specific parts of the sentences should be taken into account in the process of similarity estimation.

Some preliminary efforts were conducted for the case of grammar induction for spoken dialogue systems. The basic idea behind grammar induction is the estimation of similarity between phrases that are meant to reflect the semantics of grammar rules. For example, consider a grammar rule denoting the semantics of "departure city". Phrases such as "fly out of <City>" and "depart from <City>" can constitute the right part of this rule, where <City> is a label that stands for the (terminal) concept of "city". One may argue that the estimation of similarity for such phrases would be easier than the case sentences, due to their shorter length. However, an additional factor of difficulty arises due to the fact that non-content words reflect key domain-specific semantics. For example, the semantic divergence between "fly from <City>" and "fly to <City>" quite fine-grained to be captured by a language-agnostic approach that consults no domain-specific knowledge resources. We investigated both non-compositional (i.e., treat the entire phrase as a chunk) and compositional approaches for estimating the similarity between such phrases. The compositional approach was based on the averaging of the pairwise similarities of component words. For both approaches, baseline (i.e., no network) context-based similarity metrics were employed yielding moderate to poor performance. Interestingly, the highest performance was obtained by a different type of metrics based on the character n-gram overlap of phrases. The compositional aspects of this application-specific task remains an open question that warrants further research.

Affective text analysis. The analysis of the emotional content of textual data can be applied

to several tasks related to sentiment analysis and opinion mining. Nowadays, the importance of such applications is greater given the blooming of social media. Semantic similarity is an essential feature for the affective rating of words and sentences under the assumption that "semantic similarity can be translated to affective similarity" Malandrakis et al. [2013]. In Malandrakis et al. [2013], the estimation of continuous valence scores was investigated with respect to single-words, as well as entire sentences. Given a particular word (target), the key idea for estimating its valence rating is based on the linear combination of similarities computed between the target and a set of seed words and the (known) valence ratings of seeds. In Malandrakis et al. [2013], several well-established (baseline) co-occurrence-based and context-based similarity metrics were compared, where the highest results were reported for the Google-based semantic relatedness and cosine similarity using a narrow context window. Two of the proposed network-based similarity metrics (M_n and R_n) were also incorporated in the early phases of the above work [1] regarding the word-level rating. No significant difference in performance was observed between the baseline and the network metrics. We wish to investigate in more depth the contribution of the network-based metrics following the recent experimental procedure of Malandrakis et al. [2013] (including also the E_n metric) both at the word- and sentence-level. Overall, we consider the affective analysis of text one of the most interesting applications of the proposed work for future work.

Cognitive Aspects. Another exciting direction for future research is the incorporation of our findings about semantic priming and concreteness/abstraction within the framework of network-based DSMs. Such findings may shed more light to the semantics encoded by semantic neighborhoods. In particular, it would be interesting to discriminate strong associative from other semantic relations that hold between the target words and their respective neighbors. In the same fashion, it would be useful to estimate the degree of semantic concreteness (or abstraction) for the nodes of the network. Given the aforementioned features, we can investigate the design of feature-specific networks and similarity metrics, as well as combinations of them. Of course, numerous other cognitive aspects, e.g., typicality, can be also investigated. Overall, the spirit of this paragraph can be summarized by modifying the statement of Frederick Jelinek about statistical language modeling "put language back into language modeling" as "put cognition (back) into DSMs".

[1] The majority of this work was conducted by Nikolaos Nikolaos Malandrakis (Signal Analysis and Interpretation Laboratory, University of Southern California) when he was with the ECE Department, Technical University of Crete.

References

E. Agirre and P. Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Springer, 2007. 7, 57

E. Agirre, D. Martínez, O. L. de Lacalle, and A. Soroa. Two graph-based algorithms for state-of-the-art WSD. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 585–593, 2006. 21, 24, 56, 79

E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–27, 2009. 2, 24, 57, 82

E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proc. of the Sixth International Workshop on Semantic Evaluation (SemEval)*, pages 385–393, 2012. 8

I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187, 2010. 7

T. Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other? COLING/ACL Workshop on Multiword Expressions, 2006. 106

K. Balog, G. Mishne, and M. De Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In *Proc. of EACL*, pages 207–210, 2006. 8

S. Banerjee and T. Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proc. Third International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145, 2002. 18, 20, 56

M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010. 2, 23, 24, 54, 56, 57, 82, 110, 111

M. Baroni and R. Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, 2010. 106

M. Baroni, S. Evert, and A. Lenci. Bridging the gap between semantic theory and computational simulations. In *Proc. of ESSLLI Distributional Semantic Workshop*, 2008. xii, 110, 111

M. Baroni, R. Bernardi, and R. Zamparelli. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies (to appear)*, 2013. 106

L. W. Barsalou, A. Santos, K. Simmons W, and C. D. Wilson. Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, and A. C. Graesser, editors, *Symbols, Embodiment, and Meaning*, pages 245–283. Oxford University Press, 2008. 99, 104

D. M. Blei, A. Y. Ng, and M. I . Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 27

D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proc. of International Conference on World Wide Web*, pages 757–766, 2007. 28, 38, 39, 47, 52, 56, 89

D. Bollegala, Y. Matsuo, and M. Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In *Proc. of International World Wide Web Conference*, pages 151–160, 2010. 93

J. Bos and K. Markert. Recognising textual entailment with logical inference. In *Proc. of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, page 628–635, 2005. 7

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the Seventh International Conference on World Wide Web*, pages 107–117, 1998. 56

P. Brown, P. deSouza, R. Mercer, V. Della Pietra, and J. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992. 9, 88

E. Bruni, G. B. Tran, and M. Baroni. Distributional semantics from text and images. In *Proc. of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 22–32, 2011. 99, 100, 101, 104

A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32:13–47, 2006. xi, 13, 16, 17, 37, 56, 58, 69, 82, 88

J. Bullinaria and J. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007. 24, 33, 35

S. A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proc. of the annual meeting of the Association for Computational Linguistics: HLT*, pages 120–126, 1999. 56, 87, 89

C. Chiarello, C. Burgess, L. Richards, and A. Pollock. Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain and Language*, 38(1):75–104, 1990. 92

T. Chklovski and P. Pantel. VERBOCEAN: Mining the web for fine-grained semantic verb relations. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 33–40, 2004. 37

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. 28, 29, 39, 89

R. L. Cilibrasi and P. Vitanyi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007. 29

P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proc. of International Conference on World Wide Web*, pages 462–471, 2004. 9, 37

S. Clark. Vector space models of lexical meaning. Chapter to appear in the forthcoming Wiley-Blackwell Handbook of Contemporary Semantics 2nd edition by S. Lappin and C. Fox (eds), 2013. 23, 30

A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975. 20, 57

M. Coltheart. Deep dyslexia and right-hemisphere reading. *Brain and Language*, 71:299–309, 2000. 76

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991. 31

D. A. Cruse. *Lexical Semantics*. Cambridge University Press, 1986. 1, 2, 23, 87

G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. of the Workshop on Statistical Learning in Computer Vision*, pages 1–22, 2004. 100

J. R. Curran. *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh, 2003. 25

I. Dagan, L. Lee, and F. C.N. Pereira. Similarity-based methods for word sense disambiguation. In *Proc. of the Association for Computational Linguistics*, pages 56–63, 1997. 36

I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc, editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer Berlin / Heidelberg, 2006. 8

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391–407, 1990. 27

L. Dekang, Z. Shaojun, Q. Lijuan, and Z. Ming. Identifying synonyms among distributionally similar words. In *Proc. of International Joint Conference on Artificial Intelligence*, pages 1492–1493, 2003. 37

S. Dumais, M. Banko E. Brill, J. Lin, and A. Ng. Web question answering: Is more always better? In *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298, 2002. 37

K. Erk and S. Padó. Exemplar-based models for word meaning in context. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 92–97, 2010. 24, 25

S. Evert. *The Statistics of Word Cooccurrences Word Pairs and Collocations*. PhD thesis, University of Stuttgart, 2005. 2

H. Fang. A re-examination of query expansion using lexical resources. In *Proc. of the annual meeting of the Association for Computational Linguistics*, pages 139–147, 2008. 88

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2005. 100

R. Feldman, I. Dagan, and H. Hirsh. Mining text using keyword distributions. *Journal of Intelligent Information Systems*, 10(3):281–300, 1998. 28

Y. Feng and M. Lapata. Visual information in semantic representation. In *Proc. of the HLT-NAACL*, pages 91–99, 2010. 103, 104

L. Ferrand and B. New. Semantic and associative priming in the mental lexicon. In P. Bonin, editor, *Mental lexicon: Some words to talk about words*, pages 25–43. Hauppauge, NY: Nova, 2003. 90, 91, 92

R. Ferrer-I-Cancho and R. V. Solé. The small world of human language. *Proceedings of The Royal Society of London, Series B, Biological Sciences*, 268:2261–2266, 2001. 56

A. Finch, S. Y. Hwang, and E. Sumita. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proc. of the 3rd International Workshop on Paraphrasing*, page 17–24, 2005. 8

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131, 2002. 43, 64, 77, 101

J. R. Firth. *Studies in Linguistic Analysis: A Synopsis of Linguistic Theory 1930-1955*. Oxford Philological Society, 1957. 1, 21

S. Flank. A layered approach to NLP-based information retrieval. In *Proc. of International Conference on Computational Linguistics*, pages 397–403, 1998. 36

E. Fosler-Lussier and H.-K. J. Kuo. Using semantic class information for rapid development of language models within ASR dialogue systems. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pages 553–556, 2001. 36

W. N. Francis and H. Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, 1982. 16

G. Frege. *Die Grundlagen Der Arithmetik*. Breslau, Germany: W. Koebner, 1884. 106

E. Gabrilovich and S. Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proc. of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611, 2007. 19

S. Gauch and J. Wang. A corpus analysis approach for automatic query expansion. In *Proc. of International Conference on Information and Knowledge Management*, pages 278–284, 1997. 36

G. Geleijnse and J. Korst. Tagging artists using co-occurrences on the web. In *Proc. of Philips Symposium on Intelligent Algorithms*, pages 171–182, 2006. 37

R. Girju, A. Badulescu, and D. Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 1–8, 2003. 89

N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proc. of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. 9

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996. 27

S. Gouws, G.-J. van Rooyen, and H. A. Engelbrecht. Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. In *Proc. of the 2nd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 46–54, 2010. 20

J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the web: A multiontology disambiguation method. In *Proc. of International Conference on Web Engineering*, pages 241–248, 2006. 28, 30, 38, 56, 89

G. Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994. 24, 26, 56

D. B. Guralnik, editor. *Webster's New World Dictionary of the American Language*. Collins World, 1976. 1

W. Haas. The theory of translation. *Philosophy*, 37:208–228, 1962. 1

W. Haas. Semantic value. In *Proc. of the IXth International Congress of Linguists*, pages 1066–1072, 1964. 1

S. Harabagiu and A. Hickl. Methods for Using Textual Entailment in Open-Domain Question Answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, 2006. 8

B. Harrington. A semantic network approach to measuring relatedness. In *Proc. of the 23rd International Conference on Computational Linguistics*, pages 356–364, 2010. 20

R. Harris. *Saussure and His Interpreters*. New York University Press, 2001. 2

Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954. 3, 21, 54, 56

Z. Harris. *Mathematical Structures of Language*. Interscience Publishers, 1968. 21

Z. Harris. Distributional structure. Papers in Structural and Transformational Linguistics, 1970. 21

S. Hassan and R. Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, 2009. 19

T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *Proc. of the 11th International World Wide Web Conference*, pages 432–442, 2002. 21

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. of Conference on Computational Linguistics*, pages 539–545, 1992. 56, 89, 91

I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the 5th International Workshop on Semantic Evaluation*, pages 33–38, 2010. 87, 97

G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 305–332. MIT Press, 1998. 14

T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 50–57, 1999. 27

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. of SIGKDD*, pages 168–177, 2004. 8

T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 581–589, 2007. 20

N. Ide and J. Véronis. Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998. 57

J. Ingram. *Neurolinguistics: An Introduction to Spoken Language Processing and Its Disorders*. Cambridge University Press, 2007. 115

D. Inkpen and G. Hirst. Automatic sense disambiguation of the near-synonyms in a dictionary entry. In *Proc. of the 4th Conference on Intelligent Text Processing and Computational Linguistics*, pages 258–267, 2003. 18

E. Iosif and A. Potamianos. Unsupervised semantic similarity computation using web search engines. In *Proc. of the International Conference on Web Intelligence*, pages 381–387, 2007a. 40, 42, 49

E. Iosif and A. Potamianos. A soft-clustering algorithm for automatic induction of semantic classes. In *Interspeech*, 2007b. 8

E. Iosif and A. Potamianos. Unsupervised semantic similarity computation between terms using web documents. *IEEE Transactions on Knowledge and Data Engineering*, 22(11): 1637–1647, 2010. 24, 30, 55, 56, 58, 68, 71, 82, 87, 89, 91, 94

E. Iosif and A. Potamianos. SemSim: Resources for normalized semantic similarity computation using lexical networks. In *Proc. of the Eighth International Conference on Language Resources and Evaluation*, pages 3499–3504, 2012. 55, 60

E. Iosif and A. Potamianos. Minimum error semantic similarity using text corpora constructed from web queries. *IEEE Transactions on Knowledge and Data Engineering* (submitted to), 2013a. 67, 116, 120

E. Iosif and A. Potamianos. Similarity computation using semantic networks created from web-harvested data. *Natural Language Engineering* (DOI: 10.1017/S1351324913000144), 2013b. 54, 78

E. Iosif, A. Tegos, A. Pangos, E. Fosler-Lussier, and A. Potamianos. Combining statistical similarity measures for automatic induction of semantic classes. In *Proc. IEEE/ACL Workshop on Spoken Language Technology*, pages 86–89, 2006. 8, 36, 40

E. Iosif, M. Giannoudaki, E. Fosler-Lussier, and A. Potamianos. Associative and semantic features extracted from web-harvested corpora. In *Proc. of the Eighth International Conference on Language Resources and Evaluation*, pages 2991–2998, 2012. 88

E. Iosif, A. Potamianos, M. Giannoudaki, and K. Zervanou. Semantic similarity computation for abstract and concrete nouns using network-based distributional semantic models. In

Proc. of the 10th International Conference on Computational Semantics, pages 328–334, 2013. 77

F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998. 40

J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of International Conference on Research on Computational Linguistics*, pages 19–33, 1997. 17, 36, 39, 47, 56, 88

Y. Jin, Y. Matsuo, and M. Ishizuka. Extracting social networks among various entities on the web. In *Proc. of Fourth European Semantic Web Conference*, pages 251–266, 2007. 9

D. Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. 116

J. Karlgren, A. Holst, and M. Sahlgren. Filaments of meaning in word space. In *European Conference on Information Retrieval*, 2008. 117

K. A. Kiehl, P. F. Liddle, A. M. Smith, A. Mendrek, B. B. Forster, and R. D. Hare. Neural pathways involved in the processing of concrete and abstract nouns. *Human Brain Mapping*, 7:225–233, 1999. 76

A. Koriat. Semantic facilitation in lexical decision as a function of prime-target association. *Memory and Cognition*, 9:587–598, 1981. 89

S. Kullback. *Information Theory and Statistics*. John Wiley, 1959. 31

G. Lackoff and M. Johnson. *Metaphors We Live By*. University of Chigago Press, 1980. 21

G. Lackoff and M. Johnson. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. Basic Books, 1997. 3, 21

T. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997. 22, 27

R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pages 45–48, 1993. 88

C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification in WordNet. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press, 1998. 15, 20, 36, 39, 47, 56, 88, 96

J. H. Lee, M. H. Kim, and Y. J. Lee. Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):108–207, 1993. 14

L. Lee. *Similarity-based Approaches to Natural Language Processing*. PhD thesis, Harvard University, 1997. 31

B. Lemaire and G. Denhière. Incremental construction of an associative network from a corpus. In *Proc. of the 26th Annual Meeting of the Cognitive Science Society*, pages 825–830, 2004. 57

M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of the 5th Annual International Conference on Systems Documentation*, pages 24–26, 1986. 17, 19

D. Lewis. Naive Bayes at forty: The independence assumption in information retrieval. In *Proc. of European Conference on Machine Learning*, pages 4–15, 1998. 40

Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, 2003. 36, 39, 47, 48

D. Lin. An information-theoretic definition of similarity. In *Proc. of International Conference on Machine Learning*, 1998. 5, 6, 7

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. 32

H. Liu and Y. Chen. Computing semantic relatedness between named entities using Wikipedia. In *Proc. of the International Conference on Artificial Intelligence and Computational Intelligence*, pages 388–392, 2010. 19

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 100

K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208, 1996. 24

N. Madnani and B. J. Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387, 2010. 8

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Kernel models for affective lexicon creation. In *Proc. Interspeech*, pages 2977–2980, 2011. 8, 54, 75

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech and Language Processing* (DOI: 10.1109/TASL.2013.2277931), 2013. 8, 124

J. M. Mandler. On the foundations of the semantic system. In E. M. Forde and G. Humphreys, editors, *Category Specificity in Mind and Brain*, pages 315–340. Psychology Press, 2002. 118

T. P. McNamara. *Semantic priming: Perspectives from Memory and Word Recognition*. Psychology Press, 2005. 2, 87, 89

K. McRae and M. Jones. Semantic memory. In D. Reisberg, editor, *The Oxford Handbook of Cognitive Psychology*. Oxford University Press, 2013. 87

H. Meng and K.-C. Siu. Semi-automatic acquisition of semantic structures for understanding domain-specific natural language queries. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):172–181, 2002. 8, 30, 31, 54

R. Mihalcea and D. Moldovan. Semantic indexing using WordNet senses. In *Proc. of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, pages 35–45, 2000. 36

R. Mihalcea and D. Radev. *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press, 2011. 56

P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. of International Semantic Web Conference*, pages 522–536, 2005. 37

G. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235–312, 1990. 56, 88

G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1998. 22, 43, 52, 63, 79

G. A. Miller. The magical number seven, plus or minus two. Some limits on our capacity for processing information. *Psychological Review*, 101(2):343–352, 1955. 116

D. Milne and I. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proc. of the Workshop on Wikipedia and Artificial Intelligence*, pages 25–30, 2008. 19

S. Mirkin, L. Specia, N. Cancedda, I. Dagan, M. Dymetman, and S. Idan. Source-language entailment modeling for translating unknown terms. In *Proc. of the 47th Annual Meeting of ACL and the 4th Int. Joint Conference on Natural Language Processing of AFNLP*, pages 791–799, 2009. 8

J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34:1388–1429, 2010. 105, 106, 107, 108, 109, 110

B. Mobasher, X. Jin, and Y. Zhou. Semantically enhanced collaborative filtering on the web. In *Proc. of 1st European Web Mining Forum*, 2007. 9

J. Mori, T. Tsujishita, Y. Matsuo, and M. Ishizuka. Extracting relations in social networks from the web using similarity between collective contexts. In *Proc. of International Semantic Web Conference*, pages 487–500, 2006. 37

T. Moschopoulos, E. Iosif, L. Demetropoulou, A. Potamianos, and S. Narayanan. Towards the automatic extraction of policy networks using web links and documents. *IEEE Transactions on Knowledge and Data Engineering (to appear)*, 2013. 9

G. L. Murphy. *The Big Book of Concepts*. The MIT Press, 2002. 106

R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69, 2009. 57

R. Navigli and G. Crisafulli. Inducing word senses to improve web search result clustering. In *Proc. of Conference on Empirical Methods in Natural Language Processing*, pages 116–126, 2010. 61

D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/, 1998. 92

T. R. Niesler, E. W. D Whittaker, and P.C. Woodland. Comparison of part-of-speech and automatically derived category-based language models for speech recognition. In *ICASSP*, 1998. 9

Y. Niwa and Y. Nitta. Co-occurrence vectors from corpora versus distance vectors from dictionaries. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 304–309, 1994. 18

U. Noppeney and C. J. Price. Retrieval of abstract semantics. *NeuroImage*, 22:164–170, 2004. 76

D. N. Osherson and E. E. Smith. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 11:237–262, 1981. 106

S. Pado and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007. 24

A. Paivio. *Imagery and Verbal Processes*. New York, Holt, Rinehart and Winston, 1971. 76

D. S. Palermo and J. J. Jenkins. *Word Association Norms: Grade School Through College*. University of Minessota Press, 1964. 92

A. Pangos, E. Iosif, A. Potamianos, and E. Fosler-Lussier. Combining statistical similarity measures for automatic induction of semantic classes. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pages 278–283, 2005. 40, 49

C. Papagno, R. Capasso, and G. Miceli. Reversed concreteness effect for nouns in a subject with semantic dementia. *Neuropsychologia*, 47(4):1138–1148, 2009. 76

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002. 8

A. Pargellis, E. Fosler-Lussier, A. Potamianos, and C.-H. Lee. A comparison of four metrics for auto-inducing semantic classes. In *Proc. IEEE of Automatic Speech Recognition and Understanding Workshop*, pages 218–221, 2001. 8, 35

A. Pargellis, E. Fosler-Lussier, C. H. Lee, A. Potamianos, and A. Tsai. Auto-induced semantic classes. *Speech Communication*, 43(3):183–203, 2004. 8, 30, 31, 32, 33, 34, 35, 36, 40, 88

S. Patwardhan and T. Pedersen. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proc. of the EACL Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, 2006. 18, 56, 81, 88, 96

T. Pedersen and S. Patwardhan J. Michelizzi. WordNet::Similarity - measuring the relatedness of concepts. In *Proc. of AAAI*, pages 1024–1025, 2004. 81

Ted Pedersen. Information content measures of semantic similarity perform better without sense-tagged text. In *Proc. of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 329–332, 2010. 17, 81

E. Pekalska and R. P. W. Duin. *The Dissimilaity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2005. 7

M. Perea and A. Gotor. Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition*, 62(2):223–240, 1997. 92

D. Perez and E. Alfonseca. Application of the bleu algorithm for recognizing textual entailments. In *Proc. of the PASCAL Challenges Worshop on Recognising Textual Entailment*, 2005. 8

E. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou. X-Similarity: Computing semantic similarity between concepts from different ontologies. *Journal of Digital Information Management*, 4(4):233–238, 2006. 36, 37, 39, 43, 47

D. C. Plaut. Semantic and associative priming in distributed attractor network. In *Proc. of 17th Annual Conference of the Cognitive Science Society*, pages 37–42, 1995. 90

M. Popovic and H. Ney. Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. In *Proc. of the 10th Annual Conference of the European Association for Machine Translation*, pages 212–218, 2005. 37

R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989. 14, 20

D. Radev and R. Mihalcea. Networks and natural language processing. *AI Magazine*, 29(3): 116–126, 2008. 56

K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proc. of the 20th International Conference on World Wide Web*, pages 337–346, 2011. 19

S. Reddy, I. Klapaftis, D. McCarthy, and S. Manandhar. Dynamic and static prototype vectors for semantic composition. In *Proc. of the 5th International Joint Conference on Natural Language Processing*, pages 705–713, 2011. 25

P. Resnik. Using information content to evaluate semantic similarity in a taxanomy. In *Proc. of International Joint Conference for Artificial Intelligence*, pages 448–453, 1995. xi, 14, 16, 20, 54, 56, 57, 58, 81, 82, 88, 96

F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, and D. Molla. Exploiting paraphrases in a question answering system. In *Proc. of the 2nd International Workshop on Paraphrasing*, pages 25–32, 2003. 7

T. T. Rogers and J. L. McClelland. *Semantic Cognition. A Parallel Distributed Processing Approach*. The MIT Press, 2004. 104, 117, 118

H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965. 21, 22, 40, 43, 63, 79, 101

M. Sahami and T. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proc. of International Conference on World Wide Web*, pages 377–386, 2006. 47

M. Sahlgren. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces*. PhD thesis, Stockholm University, 2006. 2, 7, 21, 23

M. Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–53, 2008. 23

G. Salton, A. Wong, and S. C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. 22

M. Schedl, T. Pohle, P. Knees, and G. Widmer. Assigning and visualizing music genres by web-based co-occurrence analysis. In *Proc. of International Conference on Music Information Retrieval*, pages 260–265, 2006. 9, 37

H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998. 18, 24

H. Schütze and J. Pedersen. Information retrieval based on word senses. In *Proc. of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995. 21, 50

K. C. Siu and H. Meng. Semi-automatic acquisition of domain-specific semantic structures. In *Proc. of European Conference on Speech Communication and Technology*, pages 2039–2042, 1999. 36, 40

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. of the Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, 2003. 100

G. Spanakis, G. Siolas, and A. Stafylopatis. A hybrid web-based measure for computing semantic relatedness between words. In *Proc. of the 21st International Conference on Tools with Artificial Intelligence*, pages 441–448, 2009. 82

M. E. Stevens, V. E. Giuliano, and L. B. Heilprin., editors. *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*. National Bureau of Standards Miscellaneous Publication, 1965. 2

Michael Strube and Simone Paolo Ponzetto. WikiRelate! computing semantic relatedness using Wikipedia. In *Proc. of 21st National Conference on Artificial intelligence*, pages 1419–1424, 2006. 20, 82

M. J. Sussna. *Text Retrieval Using Inference in Semantic Metanetworks*. PhD thesis, University of California, 1997. 15

G. Szarvas, T. Zesch, and I. Gurevych. Combining heterogeneous knowledge resources for improved distributional semantic models. In *Proc. of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 289–303, 2011. 19

I. Szpektor and I. Dagan. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 849–856, 2008. 8

E. Terra and C. L. A. Clarke. Frequency estimates for statistical word similarity measures. In *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 165–172, 2003. 37

P. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, 2006. 54, 57

P. Turney. Domain and function: a dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44(1):533–585, 2012. 105, 106, 109, 110

P. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada, 2002. 8, 75

P. D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. of the European Conference on Machine Learning*, pages 491–502, 2001. 56, 89

P. D. Turney. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proc. of International Conference on Computational Linguistics*, pages 905–912, 2008. 87

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010. 3, 21, 27, 30

A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. 4, 5

A. Tversky. *Preference, Belief, and Similarity: Selected Writings*. Bradford Books, 2003. 4

A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. `www.vlfeat.org/`, 2013. 100

J. Véronis. Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language*, 18(3):223–252, 2004. 56

P. Vitanyi. Universal similarity. In *Proc. of Information Theory Workshop on Coding and Complexity*, pages 238–243, 2005. 29, 56, 89

E. Voorhees. Query expansion using lexical-semantic relations. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994. 36

D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 1093–1099, 2002. 56, 57

Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. Providing machine tractable dictionary tools. *Machine Translation*, 5:99–154, 1990. 18

L. Wittgenstein. *Philosophical Investigations*. Blackwell, 1953. 21

P.-R. Wojtinnek, S. Pulman, and J. Völker. Building semantic networks from plain text and Wikipedia with application to semantic relatedness and noun compound paraphrasing. *International Journal of Semantic Computing (IJSC). Special Issue on Semantic Knowledge Representation.*, 6(1):67–91, 2012. 20, 57

Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Proc. of the annual meeting on Association for Computational Linguistics*, pages 133–138, 1994. 15, 19, 20, 56, 81, 88

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, 1995. 49

F. Zanzotto, M. Pennacchiotti, and A. Moschitti. A machine learning approach to textual entailment recognition. *Natural Language Engineering*, 15(4):551–582, 2009. 7

T. Zesch, C. Muller, and I. Gurevych. Using wiktionary for computing semantic relatedness. In *Proc. of the 23rd National Conference on Artificial Intelligence (AAAI)*, pages 861–866, 2008. 19

Z. Zhang, A. Gentile, and F. Ciravegna. Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pages 991–1002, 2011. 19

X. Zhu and R. Rosenfeld. Improving trigram language modeling with the world wide web. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pages 533–536, 2001. 28, 37

George K. Zipf. *The Psycho-Biology of Language*. MIT Press, 1965. 59