Technical University of Crete

School of Electrical and Computer Engineering

Corpus-Based Methods for Learning Models of Metaphor in Modern Greek

Konstantinos Pechlivanis

Thesis Committee

Professor Vasilios Digalakis (ECE)

Dr Stasinos Konstantopoulos (N.C.S.R. Demokritos - I.I.T.)

Associate Professor Michail G. Lagoudakis (ECE)

Chania, May 2017

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Υπολογιστικές Μέθοδοι Εκμάθησης Μοντέλων Εντοπισμού Μεταφορικής Σημασίας από Σώματα Νεοελληνικών Κειμένων

Κωνσταντίνος Πεχλιβάνης

Εξεταστική Επιτροπή

Καθ. Βασίλειος Διγαλάκης (ΗΜΜΥ)

Δρ. Στασινός Κωνσταντόπουλος (Ε.Κ.Ε.Φ.Ε. ΔΗΜΟΚΡΙΤΟΣ - Ι.Π.Τ.)

Αναπλ. Καθ. Μιχαήλ Γ. Λαγουδάκης (ΗΜΜΥ)

Χανιά, Μάιος 2017

"Metaphors have a way of holding the most truth in the least space."

(Orson Scott Card)

Abstract

In this thesis, we propose a method for detecting metaphorical usage of content terms based on the hypothesis that metaphors can be detected by being characteristic of a different domain than the one they appear in. We formulate the problem as one of extracting knowledge from text classification models, where the latter have been created using standard text classification techniques without any knowledge of metaphor. We then extract from such models a measure of how characteristic of a domain a term is, providing us with a reliable method of identifying terms that are surprising for the context within which they are used.

In order to investigate our research proposal we started with compiling-crawling a corpus of articles from three Greek newspapers that offer content on-line. In order to have an initial classification, we mapped the sections of these three newspapers to domains from the top level of the relevant taxonomy of the International Press Telecommunications Council (IPTC). The training set is only annotated with the broad thematic categories assigned by the newspapers' editors.

In order to evaluate our method, we have manually annotated 89 articles with metaphorical term usage. The manual annotation was carried out by an initial annotator, with an expert annotator resolving inconsistencies to create the golden corpus. The annotation task was designed and elaborated using Ellogon platform.

In our experiments, we report results using Term Frequency - Inverse Document Frequency (TF-IDF) to identify the literal (characteristic) domain of terms and we analyse the interaction between TF-IDF and other typical word features, such as Part of Speech tags and Document Frequency. Terms could be words or N-grams. The classification of terms is accomplished using an adapted version of Maximum Likelihood Classifier.

Our method makes single-term binary decisions about metaphorical usage. We applied Precision, Recall and F_1-*score* as evaluation metrics. We compared our system to a naive baseline method and to relevant work as well. Although our model seems to be over-general, producing many false positives, the overall F_1-*score* outperforms both the baseline method and the related previous work.

keywords: metaphor detection, natural language processing, information extraction, feature extraction, text mining, distributional semantics, machine learning, crawling, topic, categorization, annotation, term weighting, term frequency, inverse term frequency, document frequency, classification

Περίληψη

Σκοπός της διπλωματικής εργασίας είναι η ανάπτυξη μεθόδων αναγνώρισης μεταφορικής και γενικά μη-κυριολεκτικής χρήσης όρων, βασιζόμενοι στην υπόθεση ότι μια λέξη που χρησιμοποιείται μεταφορικά ανήκει σε διαφορετική κατηγορία από αυτή του κειμένου στο οποίο εμφανίζεται. Η ιδέα βασίζεται στην λογική εξόρυξης πληροφορίας από γλωσσικά μοντέλα, τα οποία χρησιμοποιούν γνωστές μεθόδους ταξινόμησης, χωρίς να απαιτείται προγενέστερη γνώση των μεταφορών ή άλλων σημασιολογικών πόρων. Στόχος αυτών των μοντέλων είναι η εξαγωγή του βαθμού κατά τον οποίο ένας όρος είναι χαρακτηριστικός σε κάποια κατηγορία. Αυτό συντελεί στον εντοπισμό λέξεων οι οποίες δεν ανήκουν σημασιολογικά στο κείμενο στο οποίο εμφανίζονται.

Εξετάζοντας την ερευνητική μας πρόταση, αρχικά, συλλέξαμε σώματα κειμένων από τρεις ελληνικές εφημερίδες που μοιράζονται το περιεχόμενό τους στο διαδίκτυο. Με σκοπό την απόκτηση μια αρχικής ταξινόμησης για κάθε άρθρο, υιοθετήσαμε την ταξινόμηση που παρέχει το International Press Telecommunications Council (IPTC) χρησιμοποιώντας τις πιο ευρείες κατηγορίες. Η μοναδική επισημείωση στα δεδομένα εκπαίδευσης είναι οι κατηγορίες των άρθρων, οι οποίες έχουν ανατεθεί από τους εκδότες των εφημερίδων.

Για την αξιολόγηση της μεθόδου μας έχουμε επισημειώσει 89 άρθρα. Η διαδικασία επισημείωσης περιλαμβάνει των εντοπισμό των όρων που χρησιμοποιούνται μεταφορικά. Η επισημείωση εκπονήθηκε αρχικά από έναν επισημειωτή και στη συνέχεια, ένας ειδικευμένος επισημειωτής διόρθωσε τις ανακολουθίες που προέκυψαν, με σκοπό τη δημιουργία ενός σώματος κειμένων για τη δοκιμή του συστήματος. Η διαδικασία επισημείωσης σχεδιάστηκε και εκπονήθηκε με τη χρήση της πλατφόρμας του Ellogon.

Στα πλαίσια αυτής τη έρευνας, κάναμε χρήση της μετρικής Συχνότητα Όρων - Αντίστροφη Συχνότητα Εγγράφων (TF-IDF) με σκοπό τον εντοπισμό της χαρακτηριστικής κατηγορίας στην οποία ανήκει ένας όρος. Επιπλέον, αναλύσαμε την αλληλεπίδραση μεταξύ της μετρικής TF-IDF με άλλα χαρακτηριστικά λέξεων, όπως το μέρος του λόγου στο οποίο ανήκει, καθώς και τη συχνότητα εμφάνισής του στα διαφορετικά έγγραφα. Ένας όρος αποτελεί μια λέξη ή ένα n-γράμμα. Η κατηγοριοποίηση των όρων έγινε με τη χρήση μιας προσαρμοσμένης μορφής του Ταξινομητή Μέγιστης Πιθανοφάνειας.

Η αξιολόγηση του συστήματος έγινε με την χρήση των μετρικών Precision, Recall και F_1-*score*. Η απόφαση μια επιτυχημένης ανίχνευσης λαμβάνει χώρα για κάθε όρο ξεχωριστά, ελέγχοντας αν είναι μη-κυριολεκτικής σημασίας. Τέλος, συγκρίναμε τα αποτελέσματα του συστήματος με ένα απλοϊκό μοντέλο, καθώς και με μια σχετική δουλειά που είχε υλοποιηθεί παλιότερα. Παρόλο που το μοντέλο μας δείχνει να είναι υπεργενικευμένο, ξεπερνάει σε απόδοση τα προαναφερθέντα.

Λέξεις κλειδιά: αναγνώριση μεταφοράς, επεξεργασία φυσικής γλώσσας, εξαγωγή πληροφορίας, εξαγωγή χαρακτηριστικών, εξόρυξη κειμένου, κατανεμημένες σημασιολογίες, μηχανική μάθηση, κατηγοριοποίηση, επισημείωση, στάθμιση όρων, συχνότητα όρων, αντίστροφη συχνότητα εγγράφων, συχνότητα εγγράφων, ταξινόμηση

Acknowledgements

First and foremost, I would like to thank my supervisors and mentors, Vasilis Digalakis and Stasinos Konstantopoulos for their enthusiasm and support on this thesis work, giving me the opportunity to work on with the extremely interesting discipline of Computer Science. Their constant feedback and bibliographical pointers were always helpful. I want to make special reference to Eirini Florou, not only for her contribution and patience in our cooperation, but also for taking the time to construct the rules of Stemmer, in order to complete this work.

Also, I will not forget the supportive friends Aggelos Makrigiorgos, Alexandros Mavrommatis, Vaggelis Michelioudakis, and Giorgos Pierris, who share their moments with me, during my master studies in Athens.

I would also like to thank my family for their support and the encouragement all of these years, without them nothing would had been possible.

Finally, I would also like to thank "Lefkaditika Nea" and "Thraki" for granting us permission to use their articles for our research and "Avgi" for offering its content under a creative commons license.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Metaphor

1.1.1 What is metaphor

A metaphor is a literary figure of speech in which an implied comparison is made between two unlike things that actually share an important property. We can say that two different things are compared to each other or contrasted against each other. This comparison allows us to use fewer words and lets the reader or the listener to find the shared properties that are implied in the metaphor. The word metaphor itself is a “metaphor”, coming from a Greek word meaning to “transfer” or “carry across”. Metaphors “carry” meaning from one word or idea to another.

The simplest form of metaphor is: “The [first thing] is a [second thing].” Let’s see the following example:

- His **home** was a **prison**.

In the above sentence, we understand immediately that his home had some of negative characteristics. Mainly, we imagine, he could not leave his home. He was trapped inside. In this sentence, “prison” is a metaphor. Look at another example:

- **George** is a **sheep**.

What is one characteristic of sheep? They follow each other. So we can imagine that George is a follower, not a leader. In this sentence “sheep” is a metaphor.

The above metaphors have the simplest forms as they are nouns. But there are other ways of making metaphors, for example with verbs or adjectives. The Table 1.1 contains relevant examples.

Table 1.1: Metaphor vs Original sense of the word.

Metaphor example	Original sense of the word
The committee **shot** her ideas **down** one by one.	Anti-aircraft guns shoot down planes.
He **broke into** her conversation.	Burglars break into buildings.
He lost his job after a **heated** argument with his boss.	We have a heated swimming pool.
He was dressed rather vulgarly in a **loud** checked suit.	I can't hear because the radio is too loud.

Especially, in the case of Greek language there is a variety of words which may appear either in a literal or in a non literal context (see Table 1.2).

Table 1.2: Non literal vs Literal.

Non literal example	Literal example
Glukia zoi ('Sweet life')	Glukia karamela ('Sweet candy')
Karfono ta matia mou ('Fix my eyes on')	Karfono to karfi ('Nail the nail')
De sikono astia ('I'm not joking around')	Sikono varos ('Lift something heavy')
Kolimpo ston idrota ('Swim in the sweat')	Kolimpo sti thalasa ('Swim in the sea')

1.1.2 What is not a metaphor

A metaphor is sometimes confused with a *simile* which compares two subjects using "like" or "as". An example of *simile* would be: "He was as sly as a fox". While a metaphor would be "He was a fox". Within this work, we will not take into account either *simile* or *metonymy*, as we are going to concentrate entirely on metaphor detection. *Metonymy* enables us to use one part or aspect of an experience to stand for some other part (or the whole) of that experience. Some typical types of *metonymy* are:

- Part for Whole *e.g.* Many hands make light work.
- Whole for Part *e.g.* Australia beat Canada at cricket.
- Place for Institution *e.g.* The White House isn't saying anything.

- Producer for Products *e.g.* I like Shakespeare most.
- Potentiality for Actuality: A potential event (*e.g.* the ability, possibility, permission, obligation to undertake an action) is metonymically linked to its actual occurrence, for example: "He was able to finish his dissertation".

Unlike metaphor which involves two domains of experience, *metonymy* only requires one. Unlike metaphor which is based on similarity, *metonymy* requires contiguity.

Apart from *simile* or *metonymy* we should distinguish the metaphor and the *delexical* or light verbs. *Delexical* verb is a verb which has very little meaning in itself and is used with an object that carries the main meaning of the structure. *Delexical* verbs are treated in this works as defined by Collins Cobuild [1]. Moreover, the *delexical* verbs have a corresponding single verb (*e.g.* have a listen – to listen). Especially, in Greek language there is a variety of *delexical* verbs like *kano* ('do') , *vazo* ('put'), *pairno* ('take'), *dino* ('give'). These verbs are used with an object replacing a single verb in order to declare the action or the state. For example, may appear the *delexical* verbs *kano erotisi* ('do a question') or *dino iposhesi* ('give a promise'), although there are the single verbs *roto* ('ask') and *iposhome* ('promise'), correspondingly.

If there is a name entity, like *Chrisi Ammoudia*('Golden Beach'), which is simultaneously a metaphor expression, we don't consider it as metaphor, although it is, as this metaphor is the name of a village.

In the phrase "to grasp the concept" the physical action "to grasp" is used as a metaphor for "to understand" (which is non-physical). But this phrase has been used so often that most English speakers do not have an image of the physical action in their mind. This metaphor has "died"; it is a *"dead metaphor"*. Respectively, in the case of Greek language, there are cases like *ipostirizo tin apopsi* ('stand by my opinion'), *podi trapeziou* ('table's leg') where neither the speaker nor the hearer can understand the non literal meaning of the phrase.

Finally, it could be another one type of metaphors, the *implicit metaphor*. *Implicit metaphor* is due to an underlying cohesive grammatical and/or semantic link in the discourse which points to recoverable metaphorical material. For example: "Naturally, to embark on such as step is not necessarily to succeed in realizing it". In principle, it does not call for a non-literal indirect or direct comparison. Note, however, that it refers back to the metaphorically used lexical unit step. Since, an analysis of the discourse would need to show “step” instead of “it”, it becomes *implicitly metaphorical.* Although, this type of metaphors is very frequent, it is not included in the material we investigate in the remainder of this work.

1.1.3 Kinds of metaphor

Humans often use metaphor to describe abstract concepts through reference to more concrete or physical experiences. Metaphorical expressions may take a great variety of forms, ranging from conventional metaphors, which we produce and comprehend every day to poetic and novel ones. Rhetorical theorists and other scholars of language have discussed numerous dimensions of metaphors. However, we are going to concentrate on detecting four kinds of metaphor:

1. **Indirect, Lexical Metaphor** (type 1): metaphor at the level of a single word sense. A term identifies metaphorically another literal term. The literal term may be a single term or even a subordinate nominal clause. Characteristic examples include:

   (1) magika nisia
   magic islands
   magical islands

   (2) Omirikoi kavgades
   Homeric quarrels
   fierce quarrels

2. **Multi-word Metaphorical Expression** (type 2). A new meaning is obtained by combining all the constituents of the certain phrase. However, each constituent may be replaced by another term with similar meaning. Characteristic examples include:

   (3) Evale to heri tou
   put-3ps the hand his-CLITIC
   He helped

(4) evale freno
put-3ps brake
He slowed down

3. **Idiomatic Metaphorical Expressions** (type 3). The certain expression consists entirely of the specific constituents and no one of them can be replaced by another word with similar meaning. Characteristic examples include:

(5) richno mavri metra
drop black stone
He is gone forever

(6) echei mesanichta
has midinight
He knows nothing

4. **Direct, IS-A metaphor** (type 4): The core of these expressions is a copula that connects the subject of the certain verb with its complement. Characteristic examples include:

(7) O chronos einai chrima
time is money

(8) einai alepou
is fox
He is foxy

In the case of the multi-word metaphorical expressions, the meaning of the expression is totally different of the literal meaning of its constituents. For example, from the constituents of the phrase "My heart swelled with a sea of tears" is obtained a new meaning. In the case of the multi-word metaphorical expressions of the type 2, each constituent can be replaced by another term with similar meaning, in contrast with the idiomatic metaphorical expressions of the type 3 which are multi-word metaphorical expressions but they consist entirely of the specific constituents and no one of them can be replaced by another word with similar meaning. We can say that an idiomatic metaphorical expression is a lexeme made up of a sequence of two or more single lexemes. The lexeme of the multi-word metaphorical expression has properties that are not predictable from the properties of the individual lexemes or their normal mode of combination. Let's see a typical English example of the type 3 of metaphors which is "kick the bucket". This phrase means to die rather than to hit a bucket with one's foot. In the case of Greek language a typical example of idiomatic metaphorical expression may be considered the phrase "krouo to kodona tou kindinou". This phrase means "efisto tin prosochi kapoiou se oriseno thema" rather than "chtipo to koudouni tou kindinou". We should refer that the certain phrases have a typical structure and it is very rare to appear another structure or their lexemes have predicates. The components of the multi-word metaphorical expression must obey the following functions:

- Substitutability: A component of the certain expression cannot be replaced by a synonymous, hyperonymos or a synoponymous word.
- Deletion: A component of the multi-word metaphorical expression cannot be deleted.
- Category transformation: A component of the multi-word expression cannot change lexical category.
- Permutation: It is not possible to change the position of a term within the multi-word expression.
- Semantic Opacity: The sequence of the words cannot be understood, for example "ta ekane gis madiam" - *prokalese megali katastrofi* ('caused great destruction'). Opaque expressions are derived from historical or mythological events, from religious and literary tradition or referred to earlier habits and conditions that are currently ignored by the midfielder speaker of a language [2].
- Translatability: As the multi-word metaphorical expressions are original creation of each language, the translation of these expressions to another language cannot be a word to word literal translation.

In the case of the lexical metaphor and the IS-A metaphor only a term has a non literal meaning and enhances other terms with properties which constitute typical characteristics of the metaphorically used term. Regarding to lexical metaphors and IS-A metaphors the metaphorical term cannot be replaced by another term with similar meaning as this metaphor links the properties of the certain terms and changing the properties of the terms is plausible to change the metaphor, too. We can examine two examples; *O Giannis einai gaidouri* ('John is donkey') and *O Giannis einai alepou* ('John is fox'), although the donkey and the fox are animals each of them gives a different property to the subject. Especially, in the first case the subject is unconcerned while in the second the same subject is very intelligent.

We should refer that in the case of the direct, IS-A metaphor may be not a contrast between the contextual and a more basic meaning. The contextual meaning is also and the basic meaning. The comparison is expressed through direct language use. The direct language use may or may not be signalled with words like "like" or "san" in the case of Greek language. The following paragraph constitutes an example[1] of a direct, IS-A metaphor without a signal:

"They [system developers] seem to think that you can ask a businessman what his requirements are and get an answer that amounts to a draft system specification. A doctor doesn't ask his patient what treatment to prescribe. The patient can explain only what the problem is. It is the doctor that provides the remedy. A user may have a deep knowledge of business problems, but knowing little about computers, has no idea how they should be tackled."

The underlined lexical units have been marked as direct metaphor. Their direct metaphorical use is not signalled. Nevertheless, there is a comparison between two different domains (medical and systems development).

1.2 Thesis Contribution

Metaphor is a figure of non-literal usage of words that greatly impacts the interpretation of language. Computational treatment of metaphor aims at *detecting* metaphor and, more ambitiously, at *interpreting* it into literal semantics. The accuracy of automatic language analysis and interpretation systems is, naturally, bound to be affected by how a system treats metaphors. In a classification task, for example, even the (weaker) metaphor

[1] from Amsterdam Metaphor Corpus http://www.vismet.org/metcor/

detection task can improve results by disqualifying metaphorically used terms from being including in a text's features.

In the work described here, we are interested in detecting *novel* metaphorical usage of content terms, excluding idiomatic metaphorical expressions. This task is motivated by *text categorization* applications, where metaphorical terminology can lead to misclassifications. Furthermore, we are interested in developing methods that can be applied to languages that lack rich semantic resources, such as subcaterization frame dictionaries or semantic network dictionaries. Although the ability to correctly interpret metaphors would be useful, in such a setting even detecting metaphors is a challenging task and can still be applied to exclude or reduce the weight of metaphorical terms in text categorization models.

In this work we push further in the direction of minimizing the resources required in order to train the system, to present a method that only relies on having text placed in very broad thematic categories. Our use cases primarily stem in newspaper content categorization: newspaper content is organized in very broad thematic sections that can be used to detect out-of-topic, typically metaphorically used, terms. Metaphorically used terms can then be excluded or treated exceptionally in subsequent classification of the content in finer categories.

In the work described here, we pursue the same core hypothesis as Schulder and Hovy [3], namely, that metaphors can be detected by being characteristic of a different domain than the one they appear in. Unlike Schulder and Hovy, we do not rely on any manually chosen seed terms, neither do we formulate the problem as a classification problem where metaphor annotation is the output of a classifier.

Instead, we formulate the problem as one of *extracting knowledge* from text classification models, where the latter have been created using standard text classification techniques without any knowledge of metaphor. We then extract from such models a measure of how characteristic of a domain a term is, providing us with a reliable method of identifying terms that are *uncharacteristic* of the context within which they are used.

By doing this, we build upon the rich text classification literature and the robustness of its statistical methods in identifying domain-characteristic terms versus terms that are generally frequent across all domains. Furthermore, we provide a methodology that does not rely on any seeds or other semantic resources at the level of individual terms, but only on building a classifier that predicts very broad thematic annotations for complete articles, such as the “politics”, “sports”, and similar categories readily available in any newspaper corpus. This methodology does not require sentence structure information or semantic resources and can be applied to less-resourced languages, is robust to noisy data, and is efficient enough to be applied on large-scale corpora.

In the experiments presented here, we instantiate this generic methodology using the *Term Frequency – Inverse Document Frequency (tf-idf)* of the terms appearing in a document as features for *language models* that predict the domain based on the document terms. tf-idf balances between the frequency of a term in a particular context (tf) and its frequency across all contexts (idf) and is very well suited for identifying “surprising” words in a given context.

In our experiments we used a *Maximum TF-IDF Classifier* that uses term weighting as its only metric. We, then, assume that the “native” or literal-usage domain of a term is the domain where the term has the highest weight.

1.3 Motivation

Metaphors exist in our everyday life, and if we watch closely, we can see that we are drowning in a sea of metaphors. The metaphor has been a mean of stimulating emotions and motivating individuals, social and political groups, even entire cultures throughout history. The way metaphors are used is by replacing complicated or foreign ideas with familiar - yet generally unrelated concepts which however share an important attribute with the intended idea.

Descriptions of feelings are often given with a more metaphorical language than descriptions of behavior. For instance, to better interpret the emotionally tumultuous experience of a serious health problem writers often turn to metaphor to express this emotion. In case of cancer for example, the word “road” can be used as a metaphor to express the emotional experiences of waiting for or passing through steps in treatment. Metaphors affect the way we see, think, act, argue, learn, and communicate. Statistically humans utter about one metaphor every six minutes or every ten to twenty five words. Thus, there is interest for utilizing computers to automatically detect and classify metaphors within specific corpora including those composed of classic published literature, newspapers, emails, web blogs, political speeches, and religious sermons.

In Natural Language Processing (NLP), detecting metaphors and other non-literal figures of speech is imperative to interpret their meaning correctly. What metaphor detection does is that it focuses on and captures the motivation of speakers to express emotions and abstract concepts. As it's observed, a metaphor is often used to express the speakers' emotional experiences; we therefore model a speaker's motivation in using metaphor by detecting emotion and cognition words in metaphorical and literal sentences and their respective contexts. This leads to the automatic detection, classification, and mapping of metaphors of particular interest to psychotherapists, advertising agencies, clinical researchers, law enforcement agencies, and intelligence analysts as well as the traditional literature researcher, while presenting a significant issue for any individual attempting to explain cognitive thought, or anyone trying to train a computer to process and understand natural language.

1.4 Thesis Outline

In the remainder of this thesis, we first study the background (Section 2). Background contains details about the *feature extraction* and the creation of the feature matrix given the *tf-idf* term weighting method. Also, it deals with the *classification* of the terms and the related work that exist in metaphor detection until now.

The next chapter contains the extensively presentation of the corpus (Section 3). In this chapter we describe the whole corpus and the crawling of it as well. Moreover, we explain the topic *categorization* of the articles and the article processing in order to provide the final dataset. Finally, the aforementioned chapter contains the analysis of the annotated corpus that constitutes the testing data of the system.

Section 4 presents our approach. In this section we go deeper in our method and we explain the main idea of the system that we have implemented.

Last but not least, Section 5 contains the evaluation results of the system compared with a naive baseline and the related work, while the Section 6 contains a conclusion over the thesis and proposes future work as well.

Chapter 2

Background

2.1 Feature engineering

Feature engineering is defined as the process of using domain knowledge in order to create features that make machine learning algorithms work efficiently. Feature engineering is an essential process for the machine learning applications. In order to extract features from raw information which is stored in database, it is necessary to understand the properties of the task and how they might interact with the strengths and limitations of the employed machine learning models. Each of the engineered features should be constructed with respect to enhance predictive performance. Moreover, they should satisfy two properties; should be intuitively explained and always possible to be computed.

2.2 Linguistic Feature Extraction

2.2.1 N-grams

N-gram is a contiguous sequence of terms from a given sequence of text or speech. The terms can be phonemes, syllables, letters, words or base pairs according to the application. *N-grams* of texts are extensively used in text mining and natural language processing tasks. The integer N define the length of the sequence.

2.2.2 Text Feature Extraction of Term Weighting in Corpus: Metric TF-IDF

Term weighting is an important aspect of Information Retrieval systems. Terms are words, phrases or any other indexing units used to identify the contents of a text. Since different terms have different level of importance in a text, an important indicator, which is called *(term weight)*, is associated with each term. Three main components that affect the importance of a term in a text are the *term frequency factor (tf)*, the *inverse document frequency factor (idf)* and the *normalization factor.* More specifically:

- *term frequency factor (tf):* Long documents usually use the same terms repeatedly. As a result, the term frequency factors may be large for long documents, increasing the average contribution of its terms towards the query - document similarity.

- *inverse document frequency factor (idf):* Long documents also have numerous different terms. This increases the number of matches between a query and a long document, increasing the query - document similarity and the chances of retrieval of long documents in preference over short documents. Moreover, the high rate of appearance of a term doesn't imply that this term is directly related to the topic of the specific document. The word with the highest occurrence rate may be an auxiliary verb, like the Greek verb *eimai* ('to be'). For this reason,the *inverse document frequency (idf)* is used, which is based on counting the number of documents in the collection being searched which contain the term in question. The intuition is that a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents.

- *Normalization factor:* is a way of penalizing the term weights for a document in accordance with its length. Normalization factor retain the number of featured terms, normalizing the weights and ensuring that all their values are between 0 and 1. The use of a logarithmic function in the *tf-idf* equation constitutes the normalization factor.

As a consequence, in order to calculate the weight of a term of the domain, we use the formula:

$$\begin{aligned}\text{tf-idf}(t,d) &= \text{tf}(t,d)\ \text{idf}(t,d)\\ &= \frac{\text{freq}(t,d)}{|T_d|}\log\frac{|D|}{|D_t|}\end{aligned}$$

where freq(t, d) is the frequency of term t in domain d, T_d is the set of terms appearing in domain d, D is the set of domains, and D_t is the set of domains where t appears. At this point we should mention that we adapted this method by treating all texts of a domain as a single “document”.

2.3 Classification

2.3.1 Maximum TF-IDF Classier (MTC)

The *Maximum Likelihood Classifier* is one of the most popular methods of classification in text/term classification, in which a text/term with the maximum likelihood is classified into the corresponding class. The likelihood is defined as the probability of a text/term belonging to a class. Instead of probabilities, we adopt a different approximation considering that likelihood is defined as the TF-IDF value of a term belonging in domain d.

We use *Maximum TF-IDF Classifier (MTC)* in order to determine probabilistically the domain of a term. Specifically, we have already estimated the TF-IDF value of terms for each domain. Each term is classified in the domain where it appears with the highest TF-IDF value. More formally, given a term t and the set d_t of all the domains where t appears:

$$\text{MTC}(t, d_t) = \text{argmax}_{d \in d_t} \text{tf-idf}(t, d)$$

2.4 Related Work

The recent linguistic studies have focused not only on metaphor detection but and on metaphor interpretation. Apart from the first efforts to detect a metaphorical phrase like the MET* system of Fass [4], which based on Wilks theory [5] that a metaphor represents a violation of selectional restrictions in a certain context and the CorMet system, which had been presented by Mason [6] and is based on corpora of different domains for verbs which may be used with similar complements, the last decade there is a variety of systems of detection and interpretation metaphorical phrases.

Recent systems are striving to be less demanding in the required linguistic resources and rely on more statistical approaches to semantics. The TroFi system [7], for example, assumes a user-provided set of seed sentences and detects metaphors by computing the similarity between a sentence and all of the seed sentences. Specifically, TroFi system is a sentence clustering approach in order to recognize metaphorical phrases. The certain approach started from a set of seed sentences which had been annotated by humans. The system computed the rate of similarity between the sentence with the word to be disambiguated and all of the seed sentences. Other systems rely on semantic hierarchies: Krishnakumaran and Zhu [8] for instance predict metaphorical phrases at the sentence level using the hyponymy relation in WordNet. Also, Shutova [9] interprets metaphorical phrases as a paraphrasing task. So, for each metaphorical expression there is a literal paraphrase which is obtained by applying a probabilistic model in order to rank all the possible paraphrases of the certain metaphorical phrase at the certain context.

Most recent developments in the field, ie Klebanov [10], delivered a more effective supervised word-level classification ML model to discern between metaphorical and non metaphorical words of a text body. The model employs variables engineered from a concreteness database as well as reweighted training data points.

Another novel study on computational semantics regard innovations in *compositional distributional semantics (CDS)* models, as per Gutiérrez [11]. Therein, CDS models are used for the first time in the context of metaphor detection. An innovative approach, that being providing metaphors for the CDS models to train on as linear representations of input variables, resulted in improved performance with respect to semantic representation performance. The mentioned engineered variables aided to achieve a satisfying metaphor detection performance of 0.82 F-score.

A significant concern of the field under investigation, is found to be the necessity of extended annotations by human agents and in addition the respective evaluation being significantly constrained. These issues are both addressed in Shutova [12], where weakly supervised and unsupervised methods are used - based on considerably limited and in some cases even with no annotation at all - to infer on complex metaphors based on quantifiable observations from distribution characteristics of large text data under a concept. Extensive experimentation on three different language settings, revealed a scalable and adaptable character of the employed models that delivered significant performance under limited supervision.

Closer to our setting and methodology is the work by Schulder and Hovy [3] on detecting metaphors using a purely statistical approach to word semantics. Their research proposal is based on the hypothesis that novel metaphoric language is *unusual* in a given context. So, this unusualness of words in a certain context would be an indicator of metaphoricity in the concrete text. In order to calculate whether a term is typical of its context, they use statistical metrics to identify words *commonly used in* and *characteristic of* a domain as opposed to words commonly used across all domains. They extract domain-specific document collections using term searching. The query terms are a set of seed terms that are considered typical for a domain. The evaluation of their results is based on manually chosen seed terms or the terms with the highest relevance for document search, generating a single governance domain. Throughout the certain procedure, Schulder and Hovy found out that term relevance may be more useful when data is sparse.

Chapter 3

Data Acquisition

In order to investigate our research proposal we needed a corpus of articles from newspapers as we thought that if a certain document domains term appears at a different domains texts it has a metaphorical usage. As we had decided that using a corpus will help our research but there is not any corpus suitable for our purposes, we decided to design our own corpus. In order to build a corpus there are a number of factors which have to be taken into account such as size, balance and representativeness.

Trying to create a corpus suitable for our research we asked from a majority of national and regional newspapers the permission in order to incorporate their articles at our corpus. However, only ten regional newspapers permitted us to use their articles. These newspapers are: “Anagnostis”, “Avgi”, “Lefkaditika Nea”, “Machitis”, “Methorios”, “Samiakos Tipos”, “Tharros ton Vioton”, “Thraki”, “Foni tou Nestou” and “To Vima tis Egialias”.

Most of the selected newspapers are of general interest, but some of them have publish and local news. The selection of newspapers has resulted in a corpus with broad topical coverage containing relatively homogeneous data, despite the fact that it is harvested from the web.

3.1 Crawling Data

Having the permission of the editors to use the articles from their newspapers for research we used the command GNU *wget* at the command line of Linux in order to download the files from the web. Due to this command, we can use the total amount of data which may have a site like images, html files etc. At the html files we can find the pure text which we retrieve for our purposes. Using the command *wget* we can have access at the articles which had been published until the certain time. In some special cases, there were some restrictions so we were able to download just a portion from the data. This fact depends on the policy of each web site.

Before starting the processing of the collected material, it was necessary to throw useless data away and to keep only html files. As useless data were considered to be folders, and articles which either appeared more than once or they didn't include the whole article but just a part of it. Then, the html files, which included the articles, had been named with a separate for each article specific name with a specific pattern. This pattern included the name of the newspaper, where each article was published, and an id name of each file.

Therein after, the articles of the corpus were downloaded from the web as HTML files and cleaned into plain text using the Boilerpipe library [13]. The boilerpipe library provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a web page. The library already provides specific strategies for common tasks (for example: article extraction) and may also be easily extended for individual problem settings. Extracting context is very fast (milliseconds), just needs the input document (no global or site-level information is required) and is usually quite accurate. When we give an html file at the boilerpipe, it returns the body of the code. At the certain case, the body of the code is the pure text of the article, which may have and useless information. So, we have to remove any useless information in order to keep only the pure text.

The result of the html files parsing was text files that include the following information:

1. The link of the newspaper

2. The category in which had been classified the article from the editor

3. The date of the publication

4. The title of the article

5. The pure text of the article

The mentioned contents of the text files appeared at several points of the html code of each file. As a result, we had to detect and extract these contents for each newspaper separately. In some cases, a part of this information, such as the category of the text or the date of publication, was not available. In this case, we consider this part as unavailable.

Table 3.1 presents the total number of crawled articles for each newspaper separately. Until now, the newspapers "Foni tou Nestou" and "To Vima tis Egialias" follow the print-based model. As a result, there are not available on-line editions.

Table 3.1: Newspapers and Articles

Name of Newspaper	Crawled articles
Anagnostis	2723
Avgi	38255
Lefkaditika Nea	5305
Machitis	11893
Methorios	1961
Samiakos Tipos	2551
Tharros ton Vioton	1179
Thraki	11365
To Vima tis Egialias	0
Foni tou Nestou	0
Total number of articles	75232

The corpus of the newspapers and the crawler as well, are available on-line (http://metaphor.iit.demokritos.gr/)

3.2 Article Classification

3.2.1 What is a domain?

The universe of documents is characterized by a wide variety of typologies, which correspond to different layouts and logical structures. A domain can be defined as a group of documents which can be clustered with respect to the subject, for instance journals, papers, business letters are different domains. Each domain can also be characterized by some features which can make effective the domain classification. Such features can be concerned with the vocabulary of the text, its terminology and the common themes, which may appear at the texts of the same domain. To sum up, as domain we can consider a wider subject field which includes a variety of texts with similar topics, terminology and vocabulary.

3.2.2 Document domain

The amount of available electronic data is increasing rapidly. So, for the purposes of the Text Mining, it is necessary to manage that data. The text classification is sometimes a hard task but it is very useful in order to extract useful information about texts with relevant features. For this purpose, it is necessary to be primarily determined the subject or the topic of the document. This happens as writing on a certain subject, a certain set of typical words tend to be used. In other words, texts of similar topics will use a similar set of topical words, while texts without similar topics will use far fewer of these words.

Apart from the vocabulary, some specific terms are more likely to appear for a certain topic than for other topics. Simultaneously, other terms are far more generic and will appear in almost every topic regardless of how similar the topics may be. We should have in mind that the same term may be present in both domains.

3.2.3 Mapping domains

The International Press Telecommunications Council (IPTC) [1] creates and maintains sets of concepts in order to be assigned as metadata values to news objects like texts. This allows for a consistent coding of news metadata across news providers and over the course of time. The set of concepts is organized into a hierarchical taxonomy and it is a set of terms to express a facet of news content. Facets could be for example the subject, the genre or even the urgency. A taxonomy could be a flat list of terms or a hierarchical structure.

The news metadata are grouped into categories basing on content, Media Topics, genres and world regions. Especially, the genre indicates a nature, journalistic or intellectual characteristic of an item.

[1] www.iptc.org/site/Home/

According to the IPTC, there are 17 top level topics [1], into which media articles are classified. These topics are:

- Arts, Culture & Entertainment
- Crime, Law & Justice
- Disaster & Accident
- Economy, Business & Finance
- Education
- Environment
- Health
- Human Interest
- Labour
- Life Style & Leisure
- Politics
- Religion & Belief
- Science & Technology
- Society
- Sport
- Conflicts, War & Peace
- Weather

[1] http://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

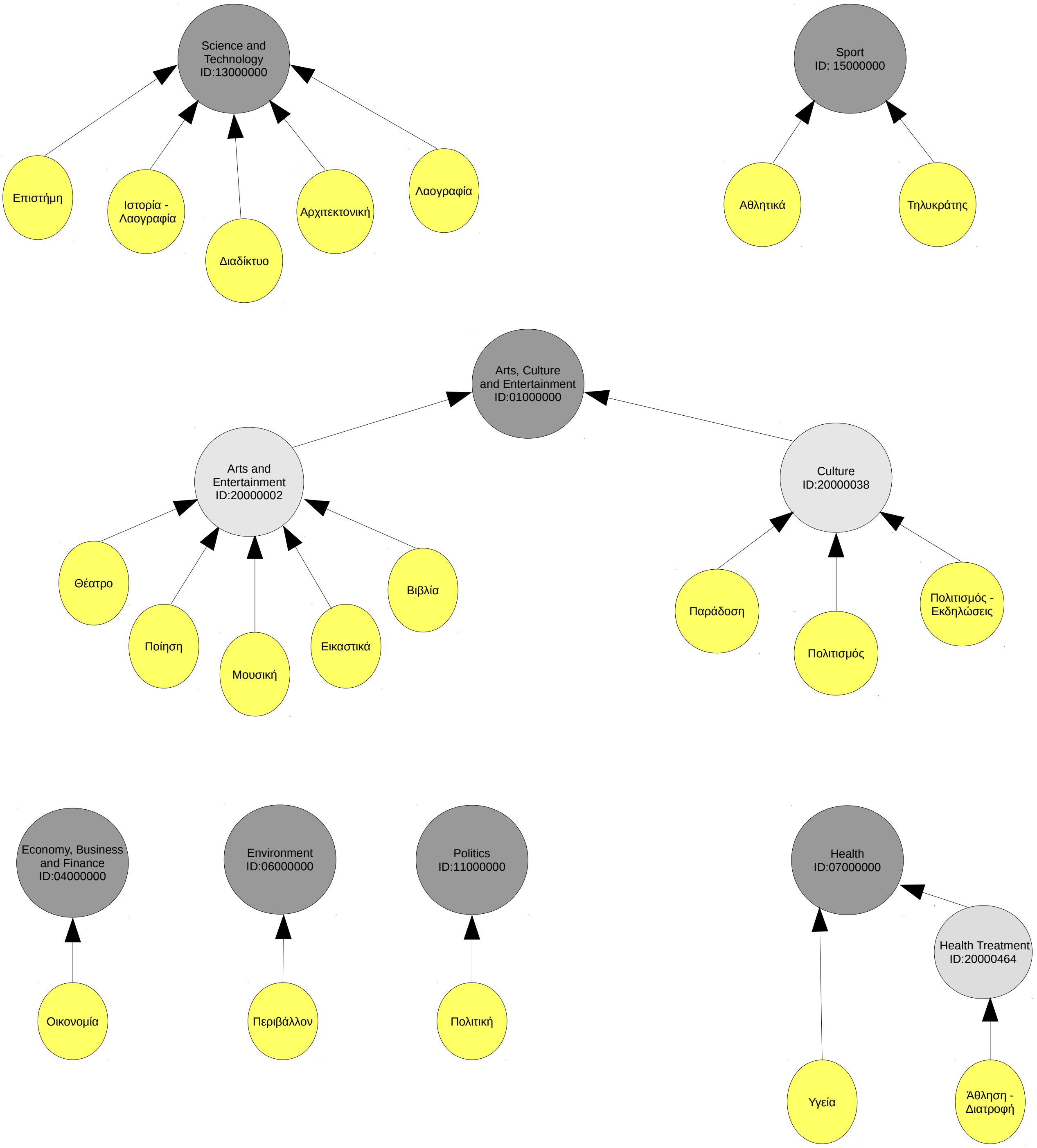
Bearing in mind the previous distinction after collecting the articles, we tried to classify them to taxonomies according to the protype of the IPTC. We focused on seven broad and discrete topics, such as: “Arts, Culture & Entertainment”, “Economy, Business & Finance”, “Environment”, “Health”, “Politics”, “Science & Technology and “Sport”. The aforementioned topics selected because are very common among newspapers and mainly, because it’s easier to distinguish and separate them among the other 17 topics.

Considering the topic of editor as required information we excluded an article from the corpus, whether the topic was unavailable. We have incorporated the certain articles at the category “Unavailable” which had been created in order to incorporate files lacking of available topic. Specifically, we excluded all the articles of the newspaper “Machitis” because all of them were published unsorted without any selected domain from editor. In the rest of newspapers, the number of articles with “Unavailable” domain was negligible.

Having collect the entire corpus from all newspapers, we had to extract the final dataset according to the available topic. There were many articles with topic different from the seven selected topic, such as “Weather”, “Human Interest”. Moreover, newspapers with mainly local news (*e.g.* “Methorios”, “Samiakos Tupos”, etc) use a variety of very specific topics that are mentioned in a person, local event, etc and they are not possible to be classified in one of the mentioned topics. As a results, these articles excluded from our corpus, as non relevant. Also, we have excluded short articles (less than 200 characters). Specifically, there are many articles that contain only photographs or tables and the text is few characters. Thus, we have incorporated the certain articles at the category “Other”, which had been created in order to incorporate files lacking of completeness of the above typical characteristics.

In order to have a robust classification, we mapped the sections of the newspapers to domains from the top level of the relevant taxonomy of the IPTC. In order to categorize all editors’ domain the existing domain is classified into the seven selected domains. The figure 3.1 consist of the categorization of the available domains.

Figure 3.1: Classification of editor's topic.



Finally, the seven selected broad topics were not discrete or available in each newspaper. Especially, for the newspapers cover regional news without national interesting and relevant taxonomy according to the IPTC. Thus, we built our corpus choosing the three newspapers that contain a suitable and appropriate number of available articles: “Avgi”, “Lefkaditika Nea” and “Thraki”. Table 3.2 lists the seven domains and the number of articles in each.

Table 3.2: Distribution of articles in topics.

Topic	Number	Percentage
Art, Culture & Entertainment	2224	11.8%
Economy,Business & Finance	4170	22.2%
Environment	846	4.5%
Health	930	4.9%
Politics	8965	47.7%
Science & Technology	181	1.0%
Sport	1494	7.9%

3.3 Data Processing

Data acquisition doesn’t include only the data collection but also the data processing, namely the manipulation of articles in order to produce meaningful information. Purpose of the data processing is the identification and the evaluation of useful components that are necessary for the implementation of the thesis.

Initially, we read the articles of the corpus word by word without taking into account terms which may be extremely common words and appear to be of little value. These words are called stopwords (Section 3.3.1).

Moreover, we dropped tokens that consist of single alphanumeric characters and several symbols, dropped stress marks and other diacritics. Then, we converted each word to upper case without the Greek punctuation marks to avoid confusion. The next stage focuses on stemming (Section 3.3.2).

3.3.1 Stopwords

Words in a document or in a corpus that are frequently occurring but meaningless in terms of Information Retrieval (IR) are called stopwords. It is repeatedly claimed that stopwords do not contribute towards the context or information of the documents and they should be removed during indexing as well as before querying by an IR system. However, the use of a single fixed stopword list across different document collections could be detrimental to the retrieval effectiveness. Nevertheless, a relatively small set of stopwords is an important part of the total number of words of a text.

For instance, words like *kai* ('and') and *tou* ('him') may appear at every sentence. So, the certain words are very common in Greek language and thanks to this fact cannot be indicators of non-literal use of language. These words are said to have a very low discrimination value when it comes to IR and they are known not only as stopwords but and as noise words and negative dictionary. To sum up, the amount of information carried by these words is negligible. Consequently, it is usually worthwhile to ignore all stopword terms when indexing the documents and processing the queries.

3.3.2 Stemming

In linguistic morphology and IR, stemming is the process for reducing inflected or sometimes derived words to their stem, base or root form generally a written word form. The stem need not be identical to the morphological root of the word. It is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

It is fact that each natural language appears its typical, characteristic features. For this reason, there isn't a generic rule-based algorithm that could apply the same stemming rules for all the languages. Especially, in the case of the Greek language, there is a variety of stemming methods for Greek texts. However, the certain methods are parts of more extended work about morphological analysis or information retrieval from various texts and can't be consider as rule-based stemmers, some of them have just included a set of rules. For this reason, it was necessary to create a more effective Greek stemmer, that could effectively remove specific suffixes of a given word.

Given that in Greek language there are nouns forms that resemble morphologically verb forms for instance, *oi luseis* ('the solutions) is a noun that resembles the verb *luseis* ('you will solve'), it was necessary our method to be based on a Greek Part of Speech

Tagger [14]. As a consequence, before the stemming of a word, we had already found out its part of speech.

After detecting the part of speech of each word, it would be easier for the striping algorithm to remove the suffix of the given word. In order to produce better stems, we should take into account some limitations. Initially, we use only capital letters in order to solve the crucial problem of the appearance of Greek tone-mark in several syllables on the stem. Moreover, we concentrate on suffixes and we have not considered prefix removal in this research. Also, as the Greek language is rich in derivative words, we decided to concentrate only on inflectional endings.

As we have already mentioned, the first step of our procedure is the detection of the part of speech of each word. Generally, a word may belong either to the nominal part of speech or to the verbal. Especially, as verbal part of speech may be considered every verb of the active or passive voice either at the singular number or at the plural number. It is necessary to bear in mind that we took into account irregular verbs like the Greek auxiliary verb "eimai" and some archaic types, which are left in modern Greek language within the context of some typical expressions.

The nominal part of speech includes apart not only the nouns but the adjectives, the participles, the pronouns, the preposition and the adverbs. Initially, we formed a closed set with the certain words that are not declined in Greek language. This set includes some proper names, the letters of the Greek alphabet, the absolute figures, adverbs, prepositions and loan words from other languages, which remain indeclinable and ignored/disregarded from stripping algorithm. Then, we observed that the adjectives and the participles may appear the same suffixes in Greek language, so we formed rules that can handle both of them. Especially, in the case of the adjectives, we created rules that remove the correct suffix even if the adjective is either at the comparative or at the superlative degree. Respectively, we manipulated the comparative and the superlative degree of the adverbs.

Consequently, stemming improves results because the presence of different word forms for the same term makes training harder, and this is more pronounced in morphologically rich languages such as Greek. Although there is a variety of stemmers, the unique morphological system of each language doesn't allow the creation of a global rule-based algorithm which would be able to find out the stem of each word. Especially, in some languages with a rich morphological system, like Greek, it is even more difficult to find the word stem by reducing the suffix from inflected or derived words. It is useful to mention that a wide variety of suffixes exist in the Greek morphological system, some of them may appear in different parts of speech. For this reason, it is necessary to point out the part of speech of the certain word before trying to find out the root of the concrete word. Our stemmer is available on-line.[1]

3.4 Annotations

According to our research proposal the document's domain is different from the domain of metaphors which appear at the certain document. In order to investigate this case, we have constructed a secondary corpus with texts from a range of domains. The texts have been selected from the initial corpus which includes the articles of the newspapers. Initially, the annotators, who were Greek native speakers with expertise in linguistics, had to detect the domain of the document according to the IPTC, as well as the domain of each paragraph of the document. Then, they had to annotate the metaphors of all the texts. For each metaphor they should annotate its type (Section 1.1.3) and they would select its domain. Specifically, we had already found 18 thematical domains. The metaphors can be classified into them. For instance, the metaphorical expression, *oi ekloges sinthetoun to skiniko* ('the elections create the circumstances'), is typical example of the second type of metaphors. The metaphor's domain is "Arts, culture & Entertainment" while the document's domain is "Politics". Finally, the annotators had to detect and annotate the delexical phrases.

The previous paragraph contains the description of the whole annotation schema. In the context of this work, we focused only in the annotations that concern the metaphor spans, regardless of the metaphor type or domain.

The annotators were looking for expressions that at the concrete context had a non literal meaning although the certain expressions on their own or the concrete expressions at another context could have a literal meaning. We should have in mind that may be appeared expressions which have a non literal meaning and they will belong to the same domain with the whole text. Moreover, an expression with a non literal meaning may be appeared at the certain context while the same expression at another context will have a

[1] Please see https://bitbucket.org/dataengineering/greek-stemmer

typically literal meaning. In a few words, we could say that the procedure of metaphor annotations focuses on five steps:

1. Reading of the whole text in order to get a general understanding of it.

2. Determining the lexical units.

3. Establishing the contextual meaning of each lexical unit.

4. Determine if it has another more basic (concrete, body-related, precise, historically older) meaning than the contextual meaning. This basic meaning is not necessarily the most frequent meaning.

5. If the contextual meaning contrasts with the basic meaning and it can be understood in comparison with it then the lexical unit should be marked as metaphorical.

Metaphor is a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them [1]. In order to find the meaning of a term, the annotators look up the certain word in the dictionary of Greek Language Manolis Triadafyllidis [15]. They should have in mind that the basic meaning of a word is more concrete as what it evokes is easier to imagine, see, hear, feel, smell and taste. Moreover, the literal meaning is usually related to bodily action, is more precise and concrete, is historically older and isn't necessarily the most frequent. A lexical unit has to be annotated as metaphorical if its contextual meaning (the meaning observed in a given context) contrasts with its basic meaning, according to the dictionary's definitions, and in order to understand the contextual meaning the annotator has to know the basic meaning of the certain term [2]. On the contrary, a term is literal if its contextual meaning is similar with the basic definition of the certain term or if it is concrete, precise, or bodily-related enough.

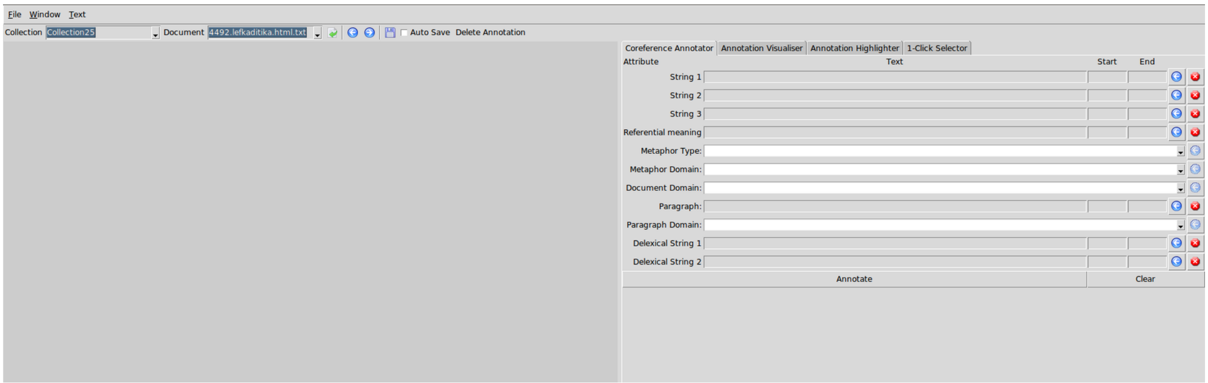
[1] http://www.merriam-webster.com/dictionary/metaphor

[2] It would be useful to be mentioned that the annotator doesn't take into account the historical aspect of the words.

3.4.1 Ellogon

In order to recognize if a phrase has a literal or a non literal meaning at a certain text, we used the tool Ellogon [16]. Ellogon is a multi-lingual, cross-platform, general-purpose language engineering environment, developed in order to aid both researchers in the field of computational linguistics, as well as companies producing and delivering language engineering systems. The tool's documentation is available on the website [1].

Regarding metaphor annotations, an annotation schema was implemented prior assigning the annotation procedure to the annotators. The annotation schema constitutes of an Ellogon module that simulates an annotation procedure. Thus, we constructed a metaphor detection annotation schema, adapting it in the tool. Figure 3.2 presents Ellogon annotation tool environment. The right side of the screenshot contains the available attributes of the metaphor annotation (String, Metaphor domain, etc).



Collection Collection25 Document 4492.lefkaditika.html.txt

□ Auto Save

Figure 3.2: Screenshot for the environment of tool

Aiming to ensure that the annotation environment is understandable and user-friendly, a few volunteers offered themselves to evaluate the annotation schema usability, providing us with important feedback. Moreover, in order to facilitate the annotators adaptation to the tool, appropriate training and support was provided.

The annotation task was defined by extensive, written guidelines provided to the annotators (first stage) in order to study and grasp the metaphor annotations manual. After a short period to acquaint themselves with the guidelines, annotators submitted

[1] http://www.ellogon.org/

their questions (second stage), concerning the comprehension of the metaphor annotations. Following, we trained the annotators over the Ellogon tool (third stage). Moreover, we demonstrated the tool to the annotators, as well as all the features of the tool that required for the metaphor annotation. We also carried out and demonstrated an annotation example for a sample article. Finally, all questions relevant with the tool and its technicalities were covered. Having concluded that latter stage, the annotators did not have the opportunity to make additional questions, unless these concerned issues relevant to the tool.

3.4.2 First Annotation Part

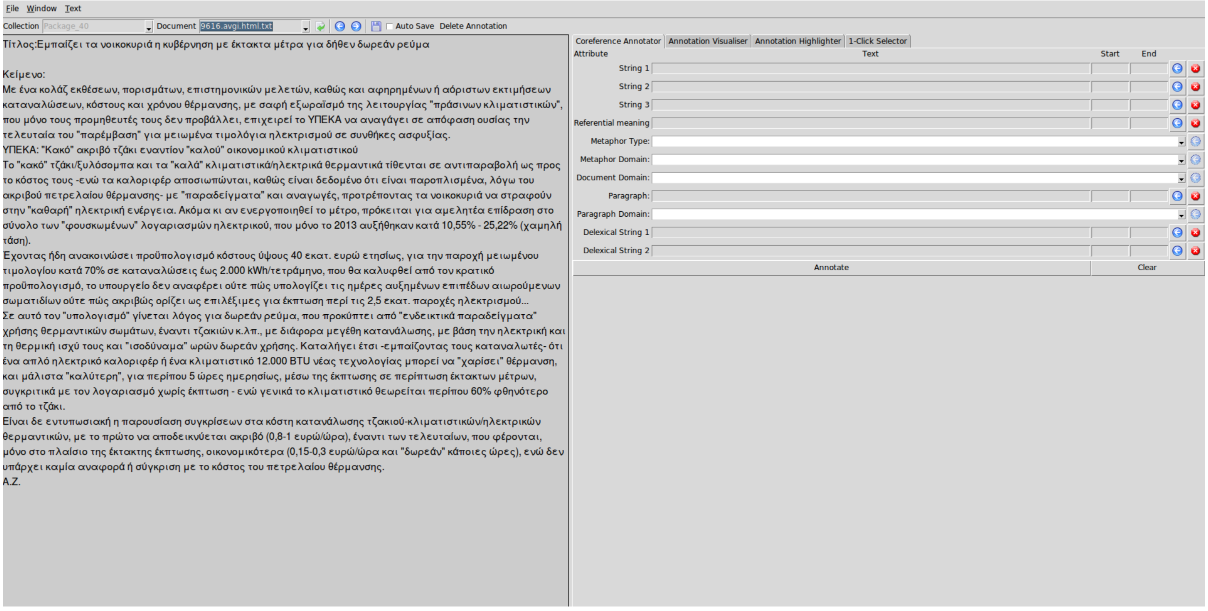
The manual annotation was carried out by two initial annotators. The procedure was organized so that each annotator had to process the same set of articles. Each annotator had to process two (2) articles per day, neither more nor least. Initially, ten (10) sample articles were provided to the annotators in order to evaluate the procedure as well as the robustness of the given guidelines. These files were then excluded from the test corpus and were used only for our own feedback. After that, we gradually provided the annotators with ten (10) more articles that constituted the testing annotated corpus.

Figure 3.3 presents the pure non annotated article, while the figure 3.4 contains a completed annotated article.

Subsequently, we assigned to a third expert annotator to resolve inconsistencies among the initial annotators, in order to create the golden corpus. The expert annotator followed the same annotation procedure. The final annotation corpus contained ten (10) articles. Table 3.3 presents some statistics over the first golden test corpus.

Table 3.3: Annotated articles of golden corpus (Part 1).

Domain	**Articles**	**All words**	**Content words**	**Metaphors**
Art, Culture & Entertainment	1	567	287	11
Economy, Business & Finance	3	2533	1320	111
Health	1	321	158	14
Politics	5	4409	2306	249
All articles	10	7830	4071	385



Collection Package_40 Document 9616.avgi.html.txt

Τίτλος:Εμπαίζει τα νοικοκυριά η κυβέρνηση με έκτακτα μέτρα για δήθεν δωρεάν ρεύμα

Κείμενο:
Με ένα κολάζ εκθέσεων, πορισμάτων, επιστημονικών μελετών, καθώς και αφηρημένων ή αόριστων εκτιμήσεων καταναλώσεων, κόστους και χρόνου θέρμανσης, με σαφή εξωραϊσμό της λειτουργίας "πράσινων κλιματιστικών", που μόνο τους προμηθευτές τους δεν προβάλλει, επιχειρεί το ΥΠΕΚΑ να αναγάγει σε απόφαση ουσίας την τελευταία του "παρέμβαση" για μειωμένα τιμολόγια ηλεκτρισμού σε συνθήκες ασφυξίας.
ΥΠΕΚΑ: "Κακό" ακριβό τζάκι εναντίον "καλού" οικονομικού κλιματιστικού
Το "κακό" τζάκι/ξυλόσομπα και τα "καλά" κλιματιστικά/ηλεκτρικά θερμαντικά τίθενται σε αντιπαραβολή ως προς το κόστος τους -ενώ τα καλοριφέρ αποσιωπώνται, καθώς είναι δεδομένο ότι είναι παροπλισμένα, λόγω του ακριβού πετρελαίου θέρμανσης- με "παραδείγματα" και αναγωγές, προτρέποντας τα νοικοκυριά να στραφούν στην "καθαρή" ηλεκτρική ενέργεια. Ακόμα κι αν ενεργοποιηθεί το μέτρο, πρόκειται για αμελητέα επίδραση στο σύνολο των "φουσκωμένων" λογαριασμών ηλεκτρικού, που μόνο το 2013 αυξήθηκαν κατά 10,55% - 25,22% (χαμηλή τάση).
Έχοντας ήδη ανακοινώσει προϋπολογισμό κόστους ύψους 40 εκατ. ευρώ ετησίως, για την παροχή μειωμένου τιμολογίου κατά 70% σε καταναλώσεις έως 2.000 kWh/τετράμηνο, που θα καλυφθεί από τον κρατικό προϋπολογισμό, το υπουργείο δεν αναφέρει ούτε πώς υπολογίζει τις ημέρες αυξημένων επιπέδων αιωρούμενων σωματιδίων ούτε πώς ακριβώς ορίζει ως επιλέξιμες για έκπτωση περί τις 2,5 εκατ. παροχές ηλεκτρισμού...
Σε αυτό τον "υπολογισμό" γίνεται λόγος για δωρεάν ρεύμα, που προκύπτει από "ενδεικτικά παραδείγματα" χρήσης θερμαντικών σωμάτων, έναντι τζακιών κ.λπ., με διάφορα μεγέθη κατανάλωσης, με βάση την ηλεκτρική και τη θερμική ισχύ τους και "ισοδύναμα" ωρών δωρεάν χρήσης. Καταλήγει έτσι -εμπαίζοντας τους καταναλωτές- ότι ένα απλό ηλεκτρικό καλοριφέρ ή ένα κλιματιστικό 12.000 BTU νέας τεχνολογίας μπορεί να "χαρίσει" θέρμανση, και μάλιστα "καλύτερη", για περίπου 5 ώρες ημερησίως, μέσω της έκπτωσης σε περίπτωση έκτακτων μέτρων, συγκριτικά με τον λογαριασμό χωρίς έκπτωση - ενώ γενικά το κλιματιστικό θεωρείται περίπου 60% φθηνότερο από το τζάκι.
Είναι δε εντυπωσιακή η παρουσίαση συγκρίσεων στα κόστη κατανάλωσης τζακιού-κλιματιστικών/ηλεκτρικών θερμαντικών, με το πρώτο να αποδεικνύεται ακριβό (0,8-1 ευρώ/ώρα), έναντι των τελευταίων, που φέρονται, μόνο στο πλαίσιο της έκτακτης έκπτωσης, οικονομικότερα (0,15-0,3 ευρώ/ώρα και "δωρεάν" κάποιες ώρες), ενώ δεν υπάρχει καμία αναφορά ή σύγκριση με το κόστος του πετρελαίου θέρμανσης.
A.Z.

Figure 3.3: Screenshot from a non-annotated article

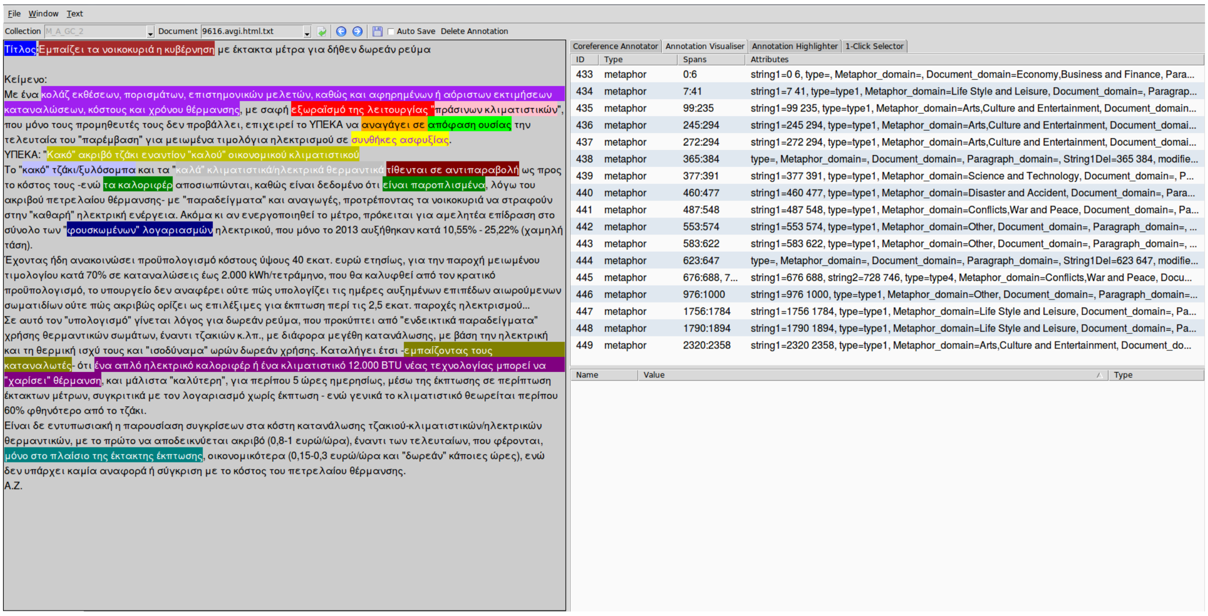
3.4.3 Second Annotation Part

Regardless of the evaluation results over the first golden corpus, we encountered a number of issues through the annotation procedure based on the annotators feedback. Firstly, there were some misunderstandings regarding the guidelines in need of elaboration, that were fully covered. We realized that the annotators faced difficulties defining the metaphors contexts. Specifically, difficulties came up regarding detection of the number of terms that a metaphor is extended to. For instance, the first annotator used to annotate too many contents for a metaphor (too many terms), while the second annotator used to annotate very few contents of metaphors. Despite the fact that the third annotator solved the issues, overall confusion was non trivial. Also, it was very common for the annotators to miss out metaphors that were placed in an already annotated metaphor. Metaphor spans can possibly overlap as in the example below:

(5)	[[Rokanizontas	dramatika]	ton	rolo]	tou	tupografou
	Shaving (carpentry)	dramatically	the	role	the	typographer-GEN

Dramatically gnawing at the typographer's role

where neither "dramatically gnawing" nor "gnawing at the role" are interpreted literally.



Τίτλος:Εμπαίζει τα νοικοκυριά η κυβέρνηση με έκτακτα μέτρα για δήθεν δωρεάν ρεύμα

Κείμενο:
Με ένα κολάζ εκθέσεων, πορισμάτων, επιστημονικών μελετών, καθώς και αφηρημένων ή αόριστων εκτιμήσεων καταναλώσεων, κόστους και χρόνου θέρμανσης, με σαφή εξωραϊσμό της λειτουργίας "πράσινων κλιματιστικών", που μόνο τους προμηθευτές τους δεν προβάλλει, επιχειρεί το ΥΠΕΚΑ να αναγάγει σε απόφαση ουσίας την τελευταία του "παρέμβαση" για μειωμένα τιμολόγια ηλεκτρισμού σε συνθήκες ασφυξίας.
ΥΠΕΚΑ: "Κακό" ακριβό τζάκι εναντίον "καλού" οικονομικού κλιματιστικού
Το "κακό" τζάκι/ξυλόσομπα και τα "καλά" κλιματιστικά/ηλεκτρικά θερμαντικά τίθενται σε αντιπαραβολή ως προς το κόστος τους -ενώ τα καλοριφέρ αποσιωπώνται, καθώς είναι δεδομένο ότι είναι παροπλισμένα, λόγω του ακριβού πετρελαίου θέρμανσης- με "παραδείγματα" και αναγωγές, προτρέποντας τα νοικοκυριά να στραφούν στην "καθαρή" ηλεκτρική ενέργεια. Ακόμα κι αν ενεργοποιηθεί το μέτρο, πρόκειται για αμελητέα επίδραση στο σύνολο των "φουσκωμένων" λογαριασμών ηλεκτρικού, που μόνο το 2013 αυξήθηκαν κατά 10,55% - 25,22% (χαμηλή τάση).
Έχοντας ήδη ανακοινώσει προϋπολογισμό κόστους ύψους 40 εκατ. ευρώ ετησίως, για την παροχή μειωμένου τιμολογίου κατά 70% σε καταναλώσεις έως 2.000 kWh/τετράμηνο, που θα καλυφθεί από τον κρατικό προϋπολογισμό, το υπουργείο δεν αναφέρει ούτε πώς υπολογίζει τις ημέρες αυξημένων επιπέδων αιωρούμενων σωματιδίων ούτε πώς ακριβώς ορίζει ως επιλέξιμες για έκπτωση περί τις 2,5 εκατ. παροχές ηλεκτρισμού...
Σε αυτό τον "υπολογισμό" γίνεται λόγος για δωρεάν ρεύμα, που προκύπτει από "ενδεικτικά παραδείγματα" χρήσης θερμαντικών σωμάτων, έναντι τζακιών κ.λπ., με διάφορα μεγέθη κατανάλωσης, με βάση την ηλεκτρική και τη θερμική ισχύ τους και "ισοδύναμα" ωρών δωρεάν χρήσης. Καταλήγει έτσι -εμπαίζοντας τους καταναλωτές- ότι ένα απλό ηλεκτρικό καλοριφέρ ή ένα κλιματιστικό 12.000 BTU νέας τεχνολογίας μπορεί να "χαρίσει" θέρμανση, και μάλιστα "καλύτερη", για περίπου 5 ώρες ημερησίως, μέσω της έκπτωσης σε περίπτωση έκτακτων μέτρων, συγκριτικά με τον λογαριασμό χωρίς έκπτωση - ενώ γενικά το κλιματιστικό θεωρείται περίπου 60% φθηνότερο από το τζάκι.
Είναι δε εντυπωσιακή η παρουσίαση συγκρίσεων στα κόστη κατανάλωσης τζακιού-κλιματιστικών/ηλεκτρικών θερμαντικών, με το πρώτο να αποδεικνύεται ακριβό (0,8-1 ευρώ/ώρα), έναντι των τελευταίων, που φέρονται, μόνο στο πλαίσιο της έκτακτης έκπτωσης, οικονομικότερα (0,15-0,3 ευρώ/ώρα και "δωρεάν" κάποιες ώρες), ενώ δεν υπάρχει καμία αναφορά ή σύγκριση με το κόστος του πετρελαίου θέρμανσης.
A.Z.

ID	Type	Spans	Attributes
433	metaphor	0:6	string1=0 6, type=, Metaphor_domain=, Document_domain=Economy,Business and Finance, Para...
434	metaphor	7:41	string1=7 41, type=type1, Metaphor_domain=Life Style and Leisure, Document_domain=, Paragrap...
435	metaphor	99:235	string1=99 235, type=type1, Metaphor_domain=Arts,Culture and Entertainment, Document_domain...
436	metaphor	245:294	string1=245 294, type=type1, Metaphor_domain=Arts,Culture and Entertainment, Document_domai...
437	metaphor	272:294	string1=272 294, type=type1, Metaphor_domain=Arts,Culture and Entertainment, Document_domai...
438	metaphor	365:384	type=, Metaphor_domain=, Document_domain=, Paragraph_domain=, String1Del=365 384, modifie...
439	metaphor	377:391	string1=377 391, type=type1, Metaphor_domain=Science and Technology, Document_domain=, P...
440	metaphor	460:477	string1=460 477, type=type1, Metaphor_domain=Disaster and Accident, Document_domain=, Para...
441	metaphor	487:548	string1=487 548, type=type1, Metaphor_domain=Conflicts,War and Peace, Document_domain=, Pa...
442	metaphor	553:574	string1=553 574, type=type1, Metaphor_domain=Other, Document_domain=, Paragraph_domain=, ...
443	metaphor	583:622	string1=583 622, type=type1, Metaphor_domain=Other, Document_domain=, Paragraph_domain=, ...
444	metaphor	623:647	type=, Metaphor_domain=, Document_domain=, Paragraph_domain=, String1Del=623 647, modifie...
445	metaphor	676:688, 7...	string1=676 688, string2=728 746, type=type4, Metaphor_domain=Conflicts,War and Peace, Docu...
446	metaphor	976:1000	string1=976 1000, type=type1, Metaphor_domain=Other, Document_domain=, Paragraph_domain=...
447	metaphor	1756:1784	string1=1756 1784, type=type1, Metaphor_domain=Life Style and Leisure, Document_domain=, Pa...
448	metaphor	1790:1894	string1=1790 1894, type=type1, Metaphor_domain=Life Style and Leisure, Document_domain=, Pa...
449	metaphor	2320:2358	string1=2320 2358, type=type1, Metaphor_domain=Arts,Culture and Entertainment, Document_do...

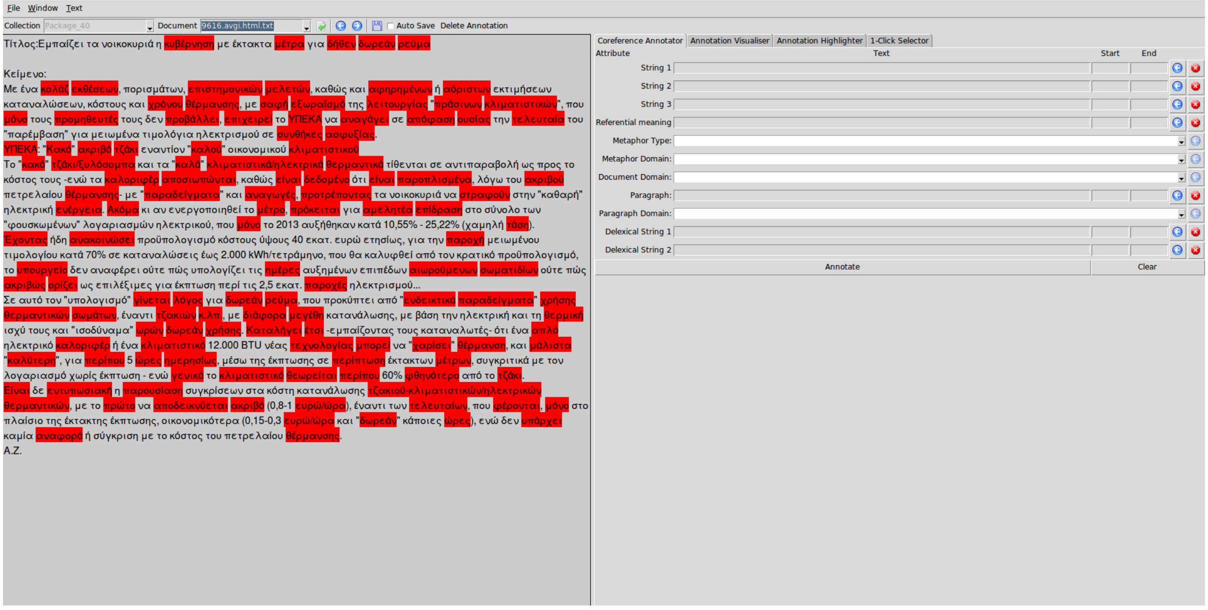
Name	Value	Type

Figure 3.4: Screenshot from an annotated article

Annotation is always a demanding task. Moreover, metaphor detection is a non trivial task for the layman. It requires understanding of the whole article and the establishment of contextual meaning of each lexical unit. As a result, metaphor annotation, besides its difficulty, is additionally very time-consuming.

Focusing on resolving these issues, we needed a second improved annotation task with a larger corpus in order to extract more accurate results. In the context of the second annotation part, firstly, we tried to improve the annotation procedure optimizing the guidelines of the metaphor detection employing the feedback already acquired from the first attempt. Secondly, we were aiming to train the annotators even further, leveraging the latter experience. Last but not least, we tried to reduce the metaphor annotation time and make the task less exhausting for the annotators using pointers in the text. Specifically, at every article have been automatically annotated (details in Section 4.1) certain words as possible metaphors (words with red color, figure 3.5). The annotators had to verify that the certain terms have actually a metaphorical meaning at the concrete context. Also, they have to look for other terms with non literal meaning at the certain context which had not been automatically annotated by the system.

In the second part, the manual annotation was carried out by one initial annotators,



Collection Package_40 Document 9616.avgi.html.txt

Τίτλος:Εμπαίζει τα νοικοκυριά η κυβέρνηση με έκτακτα μέτρα για δήθεν δωρεάν ρεύμα

Κείμενο:

Με ένα κολάζ εκθέσεων, πορισμάτων, επιστημονικών μελετών, καθώς και αφηρημένων ή αόριστων εκτιμήσεων καταναλώσεων, κόστους και χρόνου θέρμανσης, με σαφή εξωραϊσμό της λειτουργίας "πράσινων κλιματιστικών", που μόνο τους προμηθευτές τους δεν προβάλλει, επιχειρεί το ΥΠΕΚΑ να αναγάγει σε απόφαση ουσίας την τελευταία του "παρέμβαση" για μειωμένα τιμολόγια ηλεκτρισμού σε συνθήκες ασφυξίας.

ΥΠΕΚΑ: "Κακό" ακριβό τζάκι εναντίον "καλού" οικονομικού κλιματιστικού

Το "κακό" τζάκι/ξυλόσομπα και τα "καλά" κλιματιστικά/ηλεκτρικά θερμαντικά τίθενται σε αντιπαραβολή ως προς το κόστος τους -ενώ τα καλοριφέρ αποσιωπώνται, καθώς είναι δεδομένο ότι είναι παροπλισμένα, λόγω του ακριβού πετρελαίου θέρμανσης- με "παραδείγματα" και αναγωγές, προτρέποντας τα νοικοκυριά να στραφούν στην "καθαρή" ηλεκτρική ενέργεια. Ακόμα κι αν ενεργοποιηθεί το μέτρο, πρόκειται για αμελητέα επίδραση στο σύνολο των "φουσκωμένων" λογαριασμών ηλεκτρικού, που μόνο το 2013 αυξήθηκαν κατά 10,55% - 25,22% (χαμηλή τάση).

Έχοντας ήδη ανακοινώσει προϋπολογισμό κόστους ύψους 40 εκατ. ευρώ ετησίως, για την παροχή μειωμένου τιμολογίου κατά 70% σε καταναλώσεις έως 2.000 kWh/τετράμηνο, που θα καλυφθεί από τον κρατικό προϋπολογισμό, το υπουργείο δεν αναφέρει ούτε πώς υπολογίζει τις ημέρες αυξημένων επιπέδων αιωρούμενων σωματιδίων ούτε πώς ακριβώς ορίζει ως επιλέξιμες για έκπτωση περί τις 2,5 εκατ. παροχές ηλεκτρισμού...

Σε αυτό τον "υπολογισμό" γίνεται λόγος για δωρεάν ρεύμα, που προκύπτει από "ενδεικτικά παραδείγματα" χρήσης θερμαντικών σωμάτων, έναντι τζακιών κ.λπ., με διάφορα μεγέθη κατανάλωσης, με βάση την ηλεκτρική και τη θερμική ισχύ τους και "ισοδύναμα" ωρών δωρεάν χρήσης. Καταλήγει έτσι -εμπαίζοντας τους καταναλωτές- ότι ένα απλό ηλεκτρικό καλοριφέρ ή ένα κλιματιστικό 12.000 BTU νέας τεχνολογίας μπορεί να "χαρίσει" θέρμανση, και μάλιστα "καλύτερη", για περίπου 5 ώρες ημερησίως, μέσω της έκπτωσης σε περίπτωση έκτακτων μέτρων, συγκριτικά με τον λογαριασμό χωρίς έκπτωση - ενώ γενικά το κλιματιστικό θεωρείται περίπου 60% φθηνότερο από το τζάκι.

Είναι δε εντυπωσιακή η παρουσίαση συγκρίσεων στα κόστη κατανάλωσης τζακιού-κλιματιστικών/ηλεκτρικών θερμαντικών, με το πρώτο να αποδεικνύεται ακριβό (0,8-1 ευρώ/ώρα), έναντι των τελευταίων, που φέρονται, μόνο στο πλαίσιο της έκτακτης έκπτωσης, οικονομικότερα (0,15-0,3 ευρώ/ώρα και "δωρεάν" κάποιες ώρες), ενώ δεν υπάρχει καμία αναφορά ή σύγκριση με το κόστος του πετρελαίου θέρμανσης.

Α.Ζ.

Figure 3.5: Screenshot of a pre-annotated article

with a second expert annotator resolving omissions or pleonasms to create the golden corpus. Similarly, the annotation task follows the same flow and contains 89 articles. Table 3.4 presents some statistics regarding the second golden test corpus.

Concluding, the second metaphor annotation task improves the results (details in Section 5). Also, many of the aforementioned problems had been resolved up to a point. Unfortunately, the annotation time problem remains, although we used pointers in the text. As a result, we strongly believe that the metaphor annotation time for each of the annotators significantly depends on his experience and workflow and not as much in the quality of the guidelines and the training of the annotators.

Table 3.4: Annotated articles of golden corpus (Part 2).

Domain	**Articles**	**All words**	**Content words**	**Metaphors**
Art, Culture & Entertainment	12	9533	4883	552
Economy, Business & Finance	13	9288	4820	565
Environment	13	8102	4058	327
Health	12	4946	2676	217
Politics	13	10029	5113	663
Science & Technology	13	7719	4009	287
Sport	13	6181	2841	331
All articles	89	55798	28400	2942

Chapter 4

Our Approach

This research focuses on the specific area of metaphors in language and the meaning of the words within a context, and how the term weighting can be applied to the unsupervised metaphor detection.

In this work we try to detect the metaphorical phrases which may appear at the articles of newspapers. Our research proposal is based on the hypothesis that the domain of the newspaper's article is different from the domain of the metaphors which may appear at the certain article. For instance, the metaphors which may appear at a politic article will belong to another domain.

The following example contains a paragraph from a Greek politic article.

> “Η ελληνική κυβέρνηση επέλεξε [τον ρόλο του σιωπηλού παρατηρητή] που κατάφερε μέσα σε λίγες μέρες να προκαλέσει [ανυπολόγιστες καταστροφές στην κυπριακή οικονομία] και ταυτόχρονα να αυξήσει τον κίνδυνο για ολόκληρη την ευρωζώνη. Για την Κύπρο [οι εξελίξεις θα είναι δραματικές]. Η απότομη συρρίκνωση του τομέα των χρηματοπιστωτικών......"

If we had read the whole text, it would be easy to ascertain that it is a part from a text which deals with political matters. So, the document domain of this text is “Politics”. The phrase “τον ρόλο του σιωπηλού παρατηρητή” has a non literal usage at the certain context. The domain of the certain metaphor is the “Arts, culture & entertainment”. At the same sentence, the word “ανυπολόγιστες” has a metaphorical meaning with the rest context “καταστροφές στην κυπριακή οικονομία”. The whole phrase belongs to the metaphor domain "Disaster and Accident ". Also, at the same text there is the phrase "οι εξελίξεις θα είναι δραματικές." and belongs to the metaphor domain "Arts, culture & entertainment".

Figure 4.1 is an abstract visualization of the processing stages and relevant implementations that constitute the system's data-flow. In the following sections we will go though the several processes in order to explain the main approach behind them.

4.1 Corpus Preparation

In order to investigate our research proposal we started with compiling a corpus of articles from three Greek newspapers that offer content on-line: "Lefkaditika Nea", "Thraki", and "Avgi". The other available newspapers was either incompatible for our purpose or unavailable for downloading. The articles of the corpus were downloaded from the web and cleaned into plain text using the Boilerpipe library [13]. Section 3.1 contains even more details about the crawling of the data.

Considering the given topic article from editor as required information we extract a corpus from the three mentioned newspapers that constitute a collection of roughly 19,000 newspaper articles. In order to have an initial classification, we mapped the sections of the three newspapers to domains from the top level of the relevant taxonomy of the *IPTC*. Especially, we focused on seven broad and discrete topics. These topics selected because are very common among newspapers and mainly, because it's easier to distinguish and separate them among the other topics. Table 3.2 lists the seven domains and the number of articles in each. More details about the article classification are contained in section 3.2.

About the preprocessing of the data, after tokenization and stopword removal, we dropped tokens that consist of single alphanumeric characters and several symbols, dropped stress marks and other diacritics, and stemmed the data. The preprocessing and the stemming as well are analyzed with more details in Section 3.3.

As we have already mentioned in Section 3.4.3, a number of issues in first annotation part led us to repeat the annotation procedure. In order to simplify the annotation task and make it less exhausting and time consuming we used pointers in the text, annotating automatically terms as non-literal (words with red color, figure 3.5).

The automatic annotation is composed from two parts. The first part concerns the detection of the candidate non-literal terms. In order to capture this terms we used the model that we describe in the next sections. The second part concerns the annotation-coloration of these terms in order to be seen form the annotators. This part is implemented integrating the previous model in an Ellogon components [1]. Thus, we developed an Ellogon component that give us the opportunity to import additional attributes in the existing annotation schema taking advantages of the metaphor detection model. For the annotation schema we have already discussed in Section 3.4.

4.2 Term Weighting

Terms with a great impact receive high scores, while low scores are assigned to terms that are either not frequent in the document or otherwise are too frequent among documents. If also a term doesn't appear in the document, it takes the score 0. We adapt *tf-idf* method treating all text of a domain as a single document. Thus, the feature matrix contains seven (7) features, as the number of domains.

tf-idf values of individual terms (unigrams) cannot capture situations where modifiers or other multi-word constructions radically alter the semantics of a word. In order to capture some of this context, we also compute bigrams, trigrams and 4-grams. The use of larger order n-grams was not possible because of the exponential increase in the number of different n-grams that appear and the consequential sparsity in the corpus and computational intractability.

4.3 Term Classification

In our experiments we used a *Maximum TF-IDF Classifier* that uses term weighting as its only metric. We, then, assume that the *native* or literal-usage domain of a term is the domain where the term has the highest weight. The feature matrix of classifier contains *tf-idf* scores and not likelihood probabilities (like *Maximum Likelihood Classifier*). As a result, we applied a general use of classifiers, although the values of the features matrix are occasionally between 0 and 1.

[1] details in Ellogon's Developers Guide http://www.ellogon.org/index.php/download/all-categories/category/7-ellogon-documentation-manuals

If a term has zero *tf-idf* values for all domains, then this term remains unclassified and is not used in the model of metaphor detection. As a result, we needed an additional feature in order to handle this issue (details in Section 4.3.1).

If a term has almost the same probability to belong in more than one literal domains, that term is classified to all these possible domains. Especially, there are classified terms in domain “Politics” or “Economy, Business and Finance” that are difficult to distinguish in which of two domains truly belong.

Moreover, due to the nature of *tf-idf*, a term with low score might also indicate a term that is common among all domains. To filter out such candidates, we excluded ambiguous classified terms detecting only terms with high impact, namely with high *tf-idf* scores.

In Table 4.1, we can see the top three classified terms in each domain. The first column contains the domain of the terms. The words in second column are the top classified terms for the specific domain, while the last columns contains the *tf-idf* score of the term in each domain. Each *tf-idf* value corresponds to the separately domain of the table. First value displays the *tf-idf* value in the domain “Economy, Business and Finance”, second value displays the value in the domain “Sport”, third value displays the value in the domain “Environment”, etc.

4.3.1 Document Frequency

As already mentioned in section 4.3, if for all domains the terms have zero *tf-idf* values, then they remain unclassified. Trying to improve our system and in order to investigate the influence of the number of classified terms in the model we estimate an alternative method to exploit these unclassified words using *document frequency* feature as a common relevance indicator. We set a rule which classify each term depending on the *term frequency* (calculated as in *tf-idf* above) and *document frequency (df)*. Document frequency of term t in document collection C is defined as the average number of occurrences of t in each document of C:

$$\mathrm{df}(t, C) = \frac{\mathrm{freq}(t, C)}{|C|}$$

After that, we categorize terms with (zero *tf-idf* scores and) low *df* score (threshold determined empirically) using *tf* metric instead of *tf-idf*. More details about the influence of the *df* feature in our model, are listed in the next section.

Table 4.1: Top three terms for each domain.

	Term				**TF-IDF values**			
Domain		1	2	3	4	5	6	7
1	interest rate	0.05196	0	0	0.00796	0	0	0.00049
	surplus	0.02980	0	0	0.01962	0	0	0.00447
	income	0.02374	0.00307	0.00111	0.00486	0.00209	0	0.00207
2	goal	0	0.21569	0	0.00046	0	0	0
	match	0	0.13288	0	0.00011	0.00118	0	0.00131
	cup	0	0.11845	0	0.00008	0.00032	0.00074	0
3	pollution	0.00078	0	0.05543	0.00064	0	0	0.00525
	trash	0.00025	0	0.05040	0.00136	0.00019	0	0.00295
	ecosystem	0.00063	0	0.04495	0.00117	0	0	0.00032
4	memorandum	0.01671	0.00141	0.00353	0.03169	0.00130	0	0.00497
	negotiation	0.01581	0.00044	0.00139	0.02850	0.00091	0	0.00433
	center-left	0.00021	0	0	0.02047	0.00048	0	0
5	actor	0.00004	0.00055	0	0.00178	0.04782	0	0.00059
	direction	0	0.00013	0.00020	0.00012	0.03110	0.00044	0
	sculpture	0	0	0.00202	0.00011	0.02264	0.00371	0
6	higgs	0	0	0	0	0	0.04134	0
	excavation	0.00012	0	0.00020	0.00015	0.01193	0.03351	0
	CERN	0	0	0	0.00039	0	0.02495	0
7	flu	0	0	0	0.00013	0	0	0.18217
	uninsured	0.00469	0.00013	0	0.00206	0	0.00089	0.05248
	pharmacist	0.00034	0.00013	0	0.00053	0.00006	0	0.04458

Legend:

1	Economy,Business and Finance
2	Sport
3	Environment
4	Politics
5	Art, Culture and Entertainment
6	Science and and Technology
7	Health

Term	**Greek stem**
interest rate	epitoki-
surplus	pleonasmat-
income	esod-
goal	gkol
match	mats
cup	kupell-
pollution	rupans-
trash	aporrimmat-
ecosystem	ikosistimay-
memorandum	mnimoni-
negotiation	diapragmateus-
center-left	kentroarister-
actor	ithopoi-
direction	skinothesi-
sculpture	glipt-
higgs	higs
excavation	anaskaf-
CERN	CERN
flu	grip-
uninsured	anasfalist-
pharmacist	farmakopi-

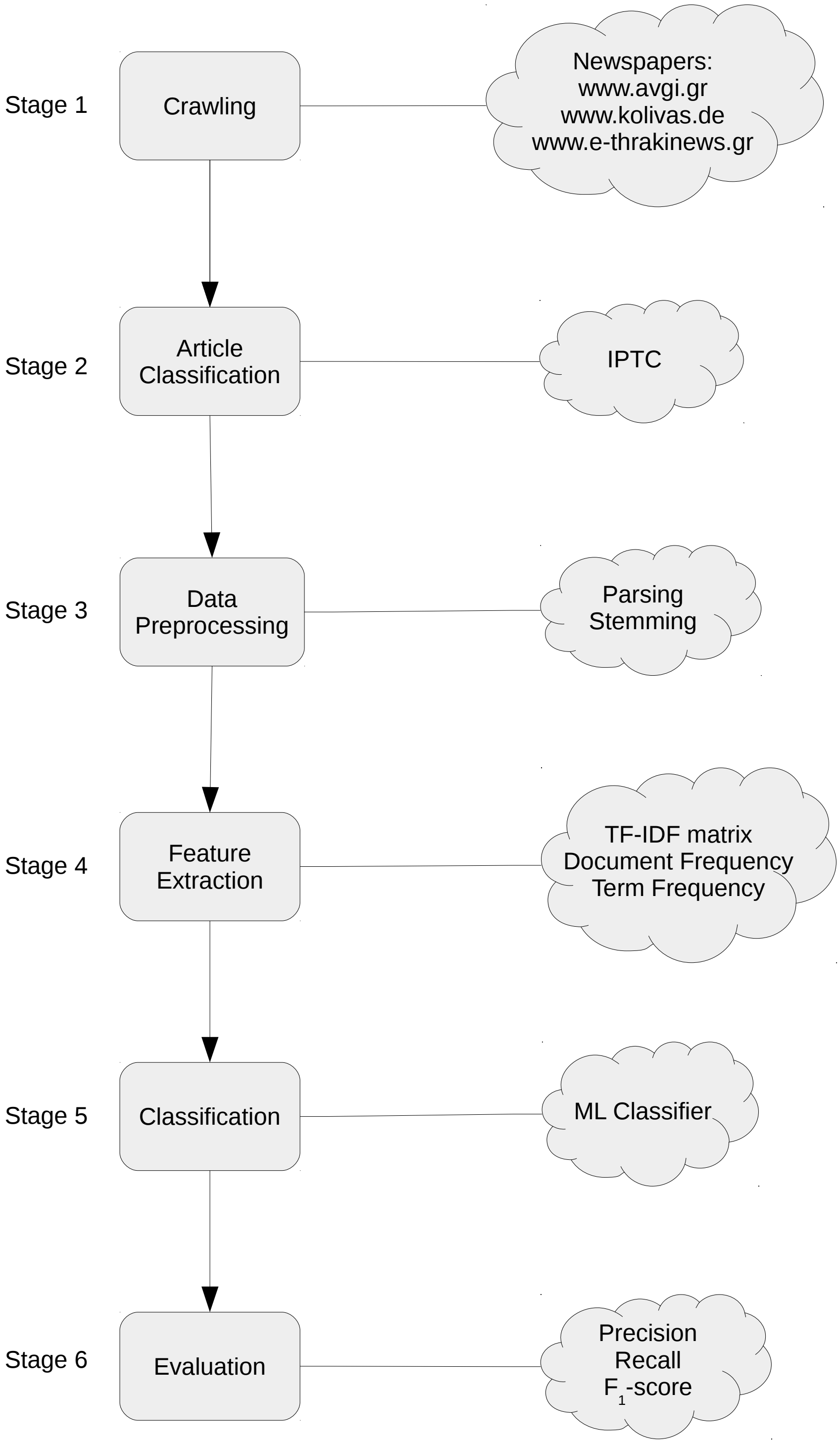


Figure 4.1: Processing stages.

Chapter 5

Results

In order to evaluate our method, we employ the model on the two manually annotated corpora (details in Section 3.4.2 and Section 3.4.3). The second golden corpus constitutes our final test corpus and the reference point, that we evaluate the system. Table 3.3 and Table 3.4 give some statistics over the golden corpora.

Our method yields single-term binary decisions about metaphorical usage. Following standard practice, we define *Precision*, *Recall* as follows:

- Precision is the percentage of positive decisions that were inside at least one span annotated as metaphor.
- Recall is the percentage of spans annotated as metaphors that include at least one positive decision.

F_1-*score* is defined in the usual manner over this Precision and Recall.

5.1 Baseline

In order to have a better intuition about the evaluation of our model, we tried to estimate a baseline over the system. Thus, we implement a naive experiment over the existing corpus.

Initially, we consider that all the words of the vocabulary belong to the most common domain. As we can see from the Table 3.2, "Politics" is the most popular domain. Therefore, all the words of vocabulary are classified in the domain, "Politics". Then, we evaluate our system over this assumption (see Table 5.1).

Table 5.1: Baseline results for the 2nd golden corpus (43,812 classified words).

	All PoS	Noun	Adjective	Verb
Precision	0.382	0.473	0.397	0.259
Recall	0.340	0.310	0.300	0.292
$F_{\beta=1}$	0.359	0.375	0.341	0.275

5.2 Evaluation Results

5.2.1 Unigram Model

Initially, we tried to evaluate our model using as terms, single words. Moreover, we tested how metaphor detection interacts with words' features, such as specific PoS (Noun, Verb, Adjective). Specifically, except for the total number of words, we evaluated our model for the three PoS separately. Table 5.2 contains the results over the final golden corpus.

Table 5.2: Evaluation results for the 2nd golden corpus (39,898 classified words).

	All PoS	Noun	Adjective	Verb
Precision	0.492	0.528	0.538	0.426
Recall	0.321	0.178	0.186	0.163
$F_{\beta=1}$	0.388	0.266	0.295	0.251

Comparing results with the baseline (see Table 5.1), there is improvement in Precision but not a significant improvement in the overall system. Moreover, we have to emphasize that the results are achieved using 39,898 classified words, while the baseline results contains 43,812. Also, the performance of the Recall for specific PoS seems slightly disappointing at first, if it is compared with the Recall of All PoS. This fact is explained as each metaphor phrase is very likely to contain one of each PoS.

Comparing results with the first golden corpus (see Table 5.3), there is also a slightly improvement, especially in Precision. Nevertheless, the second and final golden corpus constitutes a more robust and representative corpus containing 89 annotated articles, while the first golden corpus contains only 10 annotated articles.

Table 5.3: Evaluation results for the 1st golden corpus (39,898 classified words).

	All PoS	Noun	Adjective	Verb
Precision	0.442	0.520	0.488	0.321
Recall	0.366	0.171	0.212	0.210
$F_{\beta=1}$	0.400	0.258	0.295	0.254

At this point, let's take a look in the results over the annotated corpus from the initial annotator of the second annotation procedure (see Table 5.4). The evaluation results of the initial annotator appear significant improvement in Precision, while the final golden corpus has better Recall. As we can see from Table 3.4 the golden corpus contains 2,942 annotated metaphors, while the annotated corpus from the annotator contains 4,272 metaphors. As a result, our system detect as metaphorical, literal words that are wrongfully annotated by the annotator, as non-literals. Thus, our model seems to be over-general as we can conclude from the results.

Table 5.4: Evaluation results of the 2nd golden corpus for the initial annotator (39,898 classified words).

	All PoS	Noun	Adjective	Verb
Precision	0.604	0.636	0.646	0.576
Recall	0.285	0.153	0.165	0.131
$F_{\beta=1}$	0.387	0.247	0.262	0.213

A word with low *tf-idf* might also indicate a word that is common among all domains. To filter out such candidates, we excluded the words with low *tf-idf* score. The aforementioned results contains 39,898 classified words. We evaluated our system only with the words which appear to have the strongest impact (see Table 5.5 and Figure 5.1). For 39,898 classified words, we used only the 8,000 words with the highest *tf-idf* score, in order to detect the non-literal phrases. The results are slightly different from the previous experiment. Precision and Recall follow opposite behavior. Fewer classified words result to fewer detected metaphors (demote Recall) and fewer false positives (improve Precision), too. Nevertheless, this experiment achieves the best performance for the Precision.

Table 5.5: Evaluation results for the 8,000 words with the highest TF-IDF value (2nd golden corpus).

	All PoS	Noun	Adjective	Verb
Precision	0.515	**0.546**	0.513	0.422
Recall	0.114	0.073	0.058	0.027
$F_{\beta=1}$	0.187	0.129	0.104	0.052

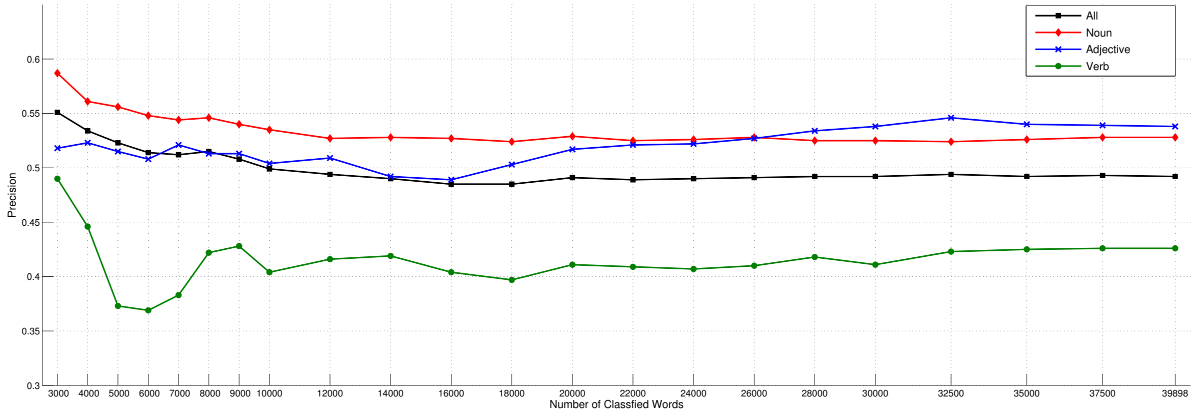


Figure 5.1: Evaluation results of Precision for the words with the highest TF-IDF value (2nd golden corpus).

The vocabulary of the model comprises 43,812 unique words. The 39,898 of these words are classified using the *MTC* classifier. The remaining 3,914 words have zero *tf-idf* and are left unclassified. Trying to improve our system and in order to investigate the influence of the number of classified words in the model we estimated an alternative method to exploit these unclassified words (details in Section 4.3.1).

Table 5.6: Evaluation results (43,812 classified words) using df as additional feature (2nd golden corpus).

	All PoS	Noun	Adjective	Verb
Precision	0.458	0.519	0.493	0.422
Recall	0.788	0.616	0.579	0.669
$F_{\beta=1}$	**0.580**	0.563	0.533	0.518

We categorized words with low *df* score (threshold determined empirically) using *tf* instead of *tf-idf* metric. Of the 3,914 unclassified words because of zero *tf-idf* values, we used *tf* metric to classify the remaining words (see Table 5.6 and Figure 5.2). Although, Precision is slightly worse, increased Recall improves the overall F_1-*score*. The latter model achieves the best performance for F_1-*score* in the final golden corpus.

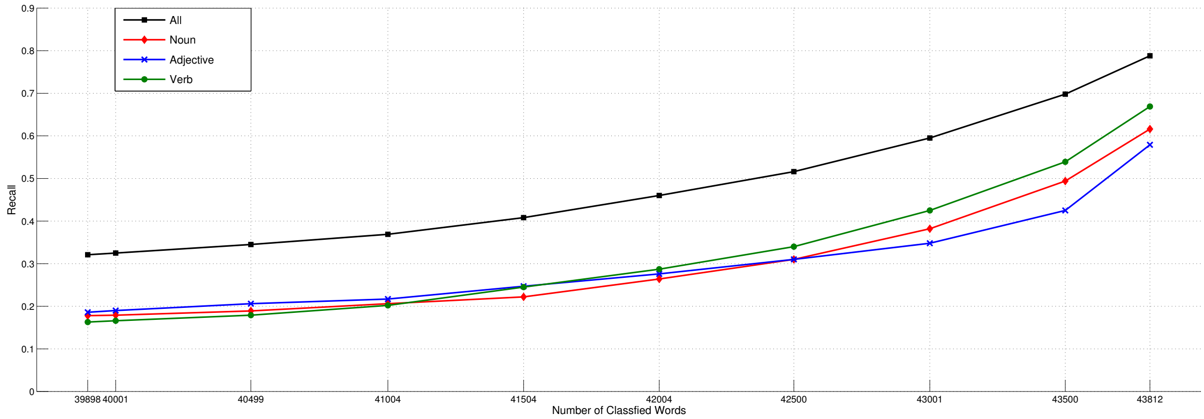


Figure 5.2: Evaluation results of Recall for different number of classified words (2nd golden corpus).

5.2.2 N-gram Model

In the next step we take in account all n-grams up to order 4, and not only unigram terms. Since longer n-grams capture more specific metaphorical contexts, they are preferred over

Table 5.7: Evaluation results (489,934 classified terms).

	Golden Corpus 1	Golden Corpus 2	Annotator
Precision	0.280	0.384	0.507
Recall	0.571	0.475	0.472
$F_{\beta=1}$	0.376	0.425	0.489

more generic, shorter n-grams. For this reason, we apply the longest n-grams first, and if a metaphor is detected the same span of text is not considered for any shorter n-grams.

The metaphor detection for specifics PoS is not possible, because it is difficult to define a specific PoS for a n-gram. Table 5.7 contains the results not only for the final golden corpus, but and for the first golden corpus and for the second annotated corpus from the initial annotator as well.

As in the unigrams, the final golden corpus has achieves slightly better results from the first golden corpus, especially in Precision. Moreover, the results of the annotator confirm the above conclusion that our model seems to be over-general.

The aforementioned results contains 489,934 classified terms. We evaluated our system only with the n-grams which appear to have the strongest impact (see Table 5.8 and Figure 5.3). For 489,934 classified terms, we used only the 6,000 terms with the highest *tf-idf* value in order to detect the non-literal phrases, but the results were not so promising. Recall is too low in this case. Nevertheless, there is improvement for Precision.

Table 5.8: Evaluation results for the 6,000 terms with the highest TF-IDF value.

	Golden Corpus 1	Golden Corpus 2	Annotator
Precision	0.482	0.521	0.640
Recall	0.096	0.058	0.054
$F_{\beta=1}$	0.160	0.105	0.100

Same as before, using *tf* and *df* metrics we categorized the total number of terms in order to investigate the influence of the number of classified terms in the model (see

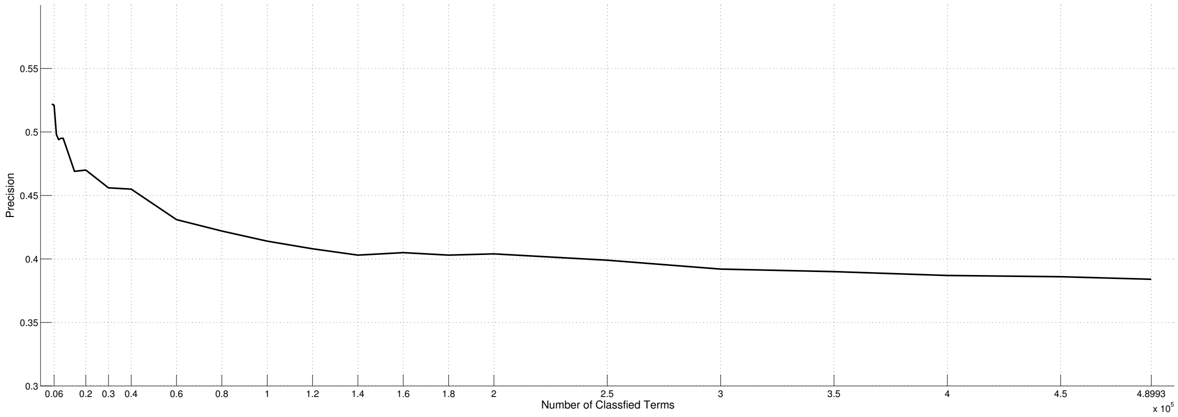


Figure 5.3: Evaluation results of Precision for the terms with the highest TF-IDF value (2nd golden corpus).

Table 5.9 and Figure 5.4). The latter model achieves the best performance for Recall in the final golden corpus.

Table 5.9: Evaluation results (494,764 classified terms) using df as additional feature.

	Golden Corpus 1	Golden Corpus 2	Annotator
Precision	0.370	0.449	0.591
Recall	0.815	**0.801**	0.801
$F_{\beta=1}$	0.509	0.575	0.680

5.2.3 Document Domain in Evaluation Results

Depending on the domain of the article, it's common to notice different writing styles. Specifically, articles with "Politics" or "Economy, Business & Finance" domain usually use a more loose and metaphorical writing style, while science articles use a formal and more literal language.

In order to investigate the influence of the domain in the method we applied metaphor detection for each domain separately. Specifically, we present domain evaluation results

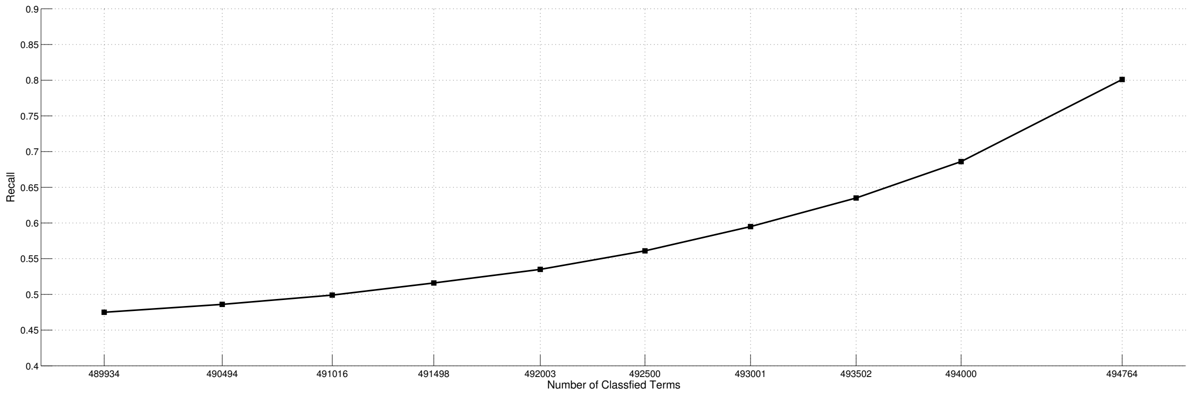


Figure 5.4: Evaluation results of Recall for different number of classified terms (2nd golden corpus).

for 3 of the aforementioned models. The first one concerns the initial unigram model for the final golden corpus (see above Tables 5.2). Table 5.10 contains the evaluation results for each domain. As we can see the domains “Politics” and “Economy, Business & Finance” attain better results compared with articles with domain “Health”, “Science & Technology” and “Sport” confirming the initial notice.

Table 5.10: Evaluation results of the 2nd golden corpus (39,898 classified terms) for each domain.

Domain	Precision	Recall	$F_{\beta=1}$
Art, Culture & Entertainment	0.474	0.349	0.402
Economy, Business & Finance	0.574	0.323	0.414
Environment	0.423	0.336	0.374
Health	0.373	0.290	0.326
Politics	0.603	0.337	0.433
Science & Technology	0.443	0.229	0.302
Sport	0.420	0.320	0.363

Also, Table 5.11 contains the domain evaluation results for our model which achieved the best Recall, namely the n-gram model that use *df* as additional feature (see above Table 5.9). Similarly, domains “Politics” and “Economy, Business & Finance” tend to have the best results among the domains. Furthermore, the domain “Science & Technology” achieves the best performance for Recall, while the domain “Politics” the best $F_{\beta=1}$ in our method.

Table 5.11: Evaluation results (494,764 classified terms) for each domain using df as additional feature (2nd golden corpus).

Domain	Precision	Recall	$F_{\beta=1}$
Art, Culture & Entertainment	0.467	0.809	0.592
Economy, Business & Finance	0.525	0.776	0.626
Environment	0.360	0.813	0.499
Health	0.351	0.797	0.488
Politics	0.539	0.760	**0.631**
Science & Technology	0.397	**0.867**	0.544
Sport	0.409	0.842	0.551

Finally, Table 5.12 presents the domain evaluation results for our model which achieved the best Precision, namely the unigram model that use only the 8,000 terms with the highest *tf-idf* score in order to detect the non-literal phrases (see above Table 5.5). Similarly, domains “Politics” and “Economy, Business & Finance” tend to have the best results among the domains. Moreover, the domain “Politics” achieves the best performance for Precision in our method.

5.3 Evaluation Results and Related Work

Schulder and Hovy [3] performed empirical evaluation using the CRF classifier. Specifically, they employed the CRFsuite [17] providing term relevance, Part-of-Speech and lexicographer sense (WordNet) as generic features. The best performance in terms of $F_{\beta=1}$ (0.373) was achieved using CRF classifier with Precision 0.640 and Recall 0.263. As shown in Table 5.6 our best result outperforms these experiments in terms of $F_{\beta=1}$

Table 5.12: Evaluation results for the 8,000 words with the highest TF-IDF value for each domain (2nd golden corpus).

Domain	Precision	Recall	$F_{\beta=1}$
Art, Culture & Entertainment	0.525	0.132	0.211
Economy, Business & Finance	0.576	0.109	0.184
Environment	0.424	0.116	0.182
Health	0.414	0.119	0.185
Politics	**0.657**	0.123	0.208
Science & Technology	0.491	0.090	0.152
Sport	0.333	0.091	0.142

(0.580) significantly. On the other hand, their learned model achieved Precision that our approach outperforms only in the domain "Politics" (see Table 5.12).

Moreover, their empirical evaluation using their adapted classifier model with seed terms (which is pretty close to our approach) achieved Recall 0.591, Precision 0.245 and $F_{\beta=1}$ 0.356. Nevertheless, the performance of their adapted classifier is lower than our best results in Recall (see Tables 5.6 and 5.9)

5.4 Discussion

There were additional models that we had implemented, but they were not so promising. Another idea we tested was handling terms that are used literally in more than one domains. To test this, we tried keeping the two most probable domains as literal domains of the terms, but this not improve the results at all, indicating that this was not a very fruitful direction to explore with more sophisticated experiments.

In order to investigate the influence of the number of the terms, we trained a model with larger vocabulary. The improvement in Recall and F1 was tiny. As a result, we accept the previous models which retain approximately the same performance with lower complexity and cost.

The n-gram evaluation results concern bigram models (Section 5.2.2). Larger order n-grams applied, too, although the exponential increase in the number of different n-grams that appear and the consequential sparsity in the corpus and computational intractability. Although trigrams and 4-grams were slightly more robust with the regard to parameter changes, there was no the expected improvement over the bigram models. Also, as metaphor processing still is a low resource task for which sufficient dataset is hard to come by, unigrams and bigrams are the most accessible and representative options.

Similarly, as per Schulder and Hovy we performed tests with the CRF classifier using *tf-idf* scores and PoS of each word as features. The results were discouraging. This arises mainly from the nature of the Greek language which is morphologically rich. Specifically, there were a lot of words that were correctly annotated as non-literal in a few sentence, and as literal (correctly too) in many others (see Table 1.2) making difficult to train a robust classifier. Especially, the Recall was very low (less than 0.1), because there were very few words classified as non-literal.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this work we presented a statistical methodology for detecting metaphorical usage of content terms. The main advantage of our methodology is that it only relies on a corpus of documents assigned to broad thematic categories and does not require any other semantic resources. No knowledge of sentence structure information or the metaphor's source domain is required. This gives our method a very wide scope of application across less-resourced languages.

For our experiments we have used a newspaper corpus assuming the *topics* under which articles were posted as such thematic categories. We experimented with the F_1-*score* obtained by our method and found significant variation between the various Parts of Speech. Furthermore, we investigate the influence of *N-grams* in metaphor detection according to the length of N. Also, we figure out specific topics that writing style is prone to use non-literal language compared with other topics with more formal language. Thus, we had the opportunity to study the structure of the Greek articles, to find out words/phrases which can be used as non-literal indicates and to detect common terms which are used to appear at more than one article's domains and so they don't have any value for contribution to differentiation between literal and non-literal speech. In addition, we have also reported improvements in Precision, Recall and F_1-*score* measures over a naive baseline and a relative work.

6.2 Future Work

For future work we plan to revisit our method with more train data of different domains. The vast majority of articles currently belong to “Politics”, “Economy, Business and Finance” and “Art, Culture and Entertainment”. The first two are the hardest domains in that they encompass several themes and might use terms from different domains literally. Domains such as “Enviroment” and “Science and Technology” on the other hand are more distinguishable to our approach, but have considerably fewer articles, especially the latter (Table 3.2). The article collection and pre-processing as more articles becomes available should continue, in the hope that a larger dataset will include enough articles from all domains to allow a more thorough statistical investigation of the differences between domains.

An interesting future work on this approach is to apply our metaphor detection system to different languages. Due to the fact that we presented a purely statistical method without semantic contents, the method could be efficient in others languages too.

As already mentioned, each term is classified in the topic where the term is most characteristic, i.e. where the term’s appearance contributes the most to classifying a document into this topic. Instead of this, we could apply two separate classifiers (instead of one) in order to categorize each term in the two domains with the higher *tf-idf* scores. Then, we could combine the classifiers in order to select each time the classifier that optimize the overall evaluation.

Bibliography

[1] Cobuild, C.: English Grammar. London: Collins CoBUILD (1990) 19

[2] Anastasiadi-Symeonidi, A., Efthimiou, A.: Stereotyped Expressions and Teaching Greek as a Second Language (in Greek). Athens: Patakis Publications (2006) 22

[3] Schulder, M., Hovy, E.: Metaphor detection through term relevance. In: Proceedings of the 2nd Workshop on Metaphor in NLP, 26 June 2014, Baltimore, MD, USA. (2014) 18–26 24, 31, 67

[4] Fass, D.: met*: a method for discriminating metonymy and metaphor by computer. Computational Linguistics **17**(1) (1991) 29

[5] Wilks, Y.: Making preferences more active. Artificial Intelligence **11**(3) (1978) 29

[6] Mason, Z.J.: CorMet: A computational, corpus-based conventional metaphor extraction system. Computational Linguistics **30**(1) (2004) 23–44 30

[7] Birke, J., Sarkar, A.: A clustering approach for the nearly unsupervised recognition of nonliteral language. In: Proceedings of EACL-06, Trento, Italy (2006) 329–336 30

[8] Krishnakumaran, S., Zhu, X.: Hunting elusive metaphors using lexical resources. In: Proceedings of the Workshop on Computational Approaches to Figurative Language, April, 2007, Rochester, New York, Rochester, New York, Association for Computational Linguistics (2007) 13–20 30

[9] Shutova, E.: Metaphor identification as interpretation. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), Atlanta, Georgia, USA, 13-14 June 2013. (2013) 30

[10] Klebanov, B.B., Leong, C.W., Flor, M.: Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In: Proceedings of the 3rd Workshop on Metaphor in NLP, 5 June 2014, Denver, Colorado. (2015) 11–20 30

[11] Gutiérrez, E.D., Shutova, E., Marghetis, T., Bergen, B.: Literal and metaphorical senses in compositional distributional semantic models. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. (2016) 30

[12] Shutova, E., Sun, L., Gutiérrez, E.D., Lichtenstein, P., Narayanan, S.: Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. Association for Computational Linguistics (2016) 30

[13] Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of The Third ACM International Conference on Web Search and Data Mining (WSDM 2010), New York City, NY, USA. (2010) 34, 52

[14] Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., Androutsopoulos, I.: Using Machine Learning Techniques for Part-Of-Speech Tagging in the Greek Language. In Fotiadis, D.I., Nikolopoulos, S.D., eds.: ADVANCES IN INFORMATICS: Proceedings of the 7th Hellenic Conference on Informatics (HCI '99). World Scientific (May 2000) 273–281 http://www.worldscibooks.com/compsci/4320.html. 42

[15] Triadafyllidis, M.: Dictionary of Greek Language. The School of Modern Greek Language of the Aristotle University of Thessaloniki (1998) 44

[16] Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., Spyropoulos, C.D.: Ellogon: A new text engineering platform. In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain. (2002) 45

[17] Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007) 67