ΣΤΡΑΤΙΩΤΙΚΗ ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ
Τμήμα Στρατιωτικών Επιστημών

Ε Λ Λ Η Ν Ι Κ Η Δ Η Μ Ο Κ Ρ Α Τ Ι Α
ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΑΚΑΔΗΜΑΪΚΟΥ ΕΤΟΥΣ 2018-20
ΕΦΑΡΜΟΣΜΕΝΗ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ & ΑΝΑΛΥΣΗ
(ΠΔ 97 /2015/ΦΕΚ 163Α'/20.08.2014)

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

ΔΕΝΤΡΑ ΑΠΕΙΚΟΝΙΣΗΣ ΕΞΑΡΤΗΣΕΩΝ

Διατριβή που υπεβλήθη για την μερική ικανοποίηση των απαιτήσεων για την απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης

ΥΠΟ:

ΓΕΩΡΓΙΟΥ ΙΩΑΝΝΙΔΗ

Α.Μ.: 2018018018

ΙΟΥΛΙΟΣ 2021

Μεταπτυχιακή Διατριβή του Γεώργιου Ιωαννίδη εγκρίνεται:

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

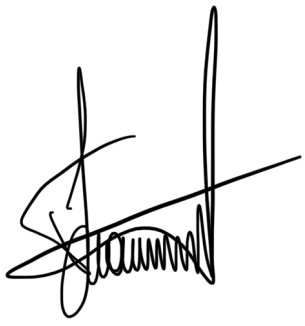
Καθηγητής Δάρας Νικόλαος (Επιβλέπων),……..

Καθηγητής Καϊμακάμης Γεώργιος,……..

Καθηγητής Μουστάκης Βασίλειος,……

Vasilis Moustakis
Digitally signed by Vasilis Moustakis
Date: 2021.07.20 15:42:43 +03'00'

ΣΕΛΙΔΑ ΣΚΟΠΙΜΑ ΚΕΝΗ



Έτος 2021

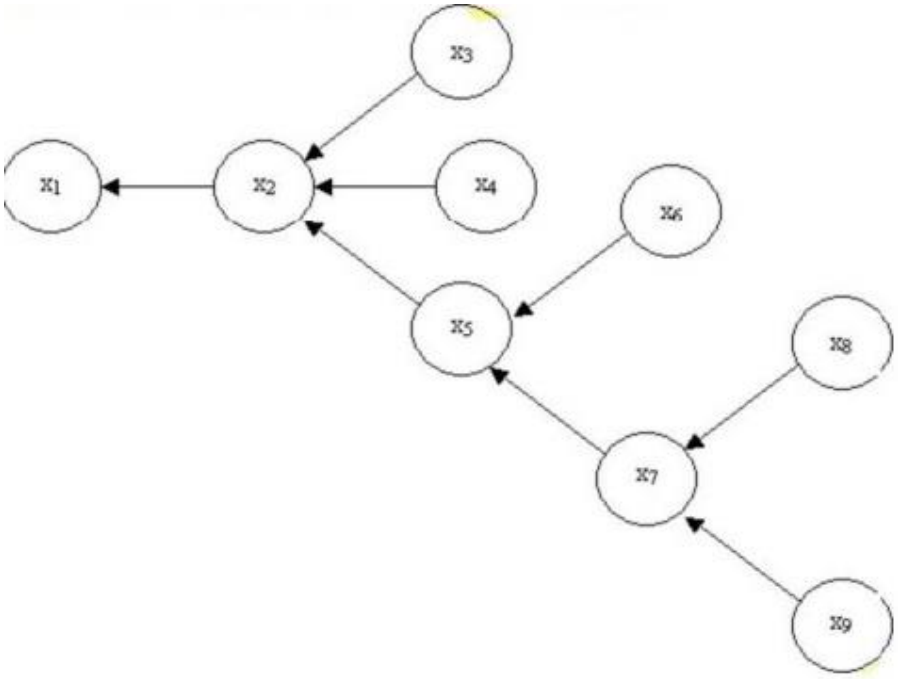
ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή

Η παρούσα διατριβή αναλύει θεωρητικά τα δέντρα εξάρτησης ως κατανομές πιθανότητας και την χρήση τους στην αναγνώριση μοτίβων σε εικόνες. Επίσης εξετάζονται πιθανές εφαρμογές των δέντρων εξάρτησης ως χρήσιμων συντακτικών πληροφοριών σε προβλήματα όπως αυτό της εξεύρεσης απαντήσεων σε ερωτήσεις (Q/A), αναπτύσσοντας έναν αλγόριθμο. Πλην της εφαρμογής τους στο πεδίο των ερωτήσεων/απαντήσεων, θα μελετηθεί η χρησιμότητα τους στην Μηχανική Μετάφραση και ιδίως στην συντακτική στατιστική μηχανική μετάφραση. Στο πεδίο της στατιστικής μηχανικής μετάφρασης θα μελετηθούν μοντελοποιήσεις των δέντρων εξάρτησης σε γλώσσες που παρουσιάζουν ισομορφισμούς και σε περιπτώσεις στις οποίες δεν υπάρχουν. Τέλος θα συμπεριληφθεί και ένα κομμάτι ανάλυσης-δημιουργίας μιας μοντελοποίησης διασύνδεσης ενός δέντρου εξάρτησης με ένα σημασιολογικό δέντρο εξάρτησης.

2. Δέντρα εξάρτησης

Ένα δέντρο εξάρτησης χρησιμοποιείται για την απεικόνιση μιας διακλαδωμένης εξάρτησης ώστε να προσεγγιστούν διάφορες κατανομές πιθανοτήτων. Ένα δέντρο εξάρτησης υποδεικνύει την εξάρτηση στοιχείων από άλλα στοιχεία.



Σχήμα 1: Δέντρο εξάρτησης,Toussaint,1978.

Για παράδειγμα στο σχήμα 1 φαίνεται η σχέση εξάρτησης ενός στοιχείου $\boldsymbol{x_k}$, με τα υπόλοιπα. Το στοιχείο $\boldsymbol{x_2}$ είναι εξαρτώμενο από το στοιχείο $\boldsymbol{x_1}$, και τα στοιχεία $\boldsymbol{x_3}$, $\boldsymbol{x_4}$ και $\boldsymbol{x_5}$ είναι εξαρτώμενα από το $\boldsymbol{x_2}$. Όπως και στις αλυσίδες Makrov, πρέπει να ληφθεί υπ' όψιν η κοινή κατανομή πιθανότητας $\boldsymbol{P(X)}$ όπου το $\boldsymbol{X}$ είναι διάνυσμα των **N** τυχαίων μεταβλητών $\boldsymbol{X = (X_1, X_2, \dots, X_N)}$**.** Για να ορίσουμε την κατανομή πιθανοτήτων για μια συγκεκριμένη διάταξη του $\boldsymbol{X}$, μπορεί να χρησιμοποιηθεί μια προσέγγιση, η οποία να είναι το γινόμενο κατώτερων τάξεων κατανομών πιθανοτήτων. Αυτό ισχύει, γιατί το γινόμενο των προσεγγίσεων είναι επίσης μια έγκυρη κατανομή πιθανοτήτων.

Ας υποθέσουμε, ότι έχουμε ένα ήδη ορισμένο δέντρο εξάρτησης, για κάθε τάξη αντικειμένου που προσπαθούμε να ορίσουμε. Αυτή η μέθοδος με τα δέντρα εξάρτησης,

θα βοηθήσει στην ταξινόμηση των εικόνων σε ήδη ορισμένες τάξεις[1]. Θεωρούμε μια επιτρεπόμενη κατανομή πιθανοτήτων ως προσέγγιση του παρακάτω τύπου:

$$\boldsymbol{P_t(X) = \prod_{i=1}^{n} P(X_{mi} \mid X_{mj(i)})}$$

$$\boldsymbol{0 \le j(i) < i}$$

$\boldsymbol{M = (m_1, m_2, \dots m_n)}$ είναι ένας άγνωστος συνδυασμός ακεραίων, όπου τουλάχιστον μία από τις μεταβλητές απεικονίζεται. Η συνάρτηση $\boldsymbol{j(i)}$ είναι το δέντρο εξάρτησης της προσέγγισης και αναπαριστά την απεικόνιση, η οποία χρειάζεται για να οριστεί το δέντρο εξάρτησης.

Ας πούμε ότι το $\boldsymbol{xi}$ αναπαριστά ένα σημείο στο επίπεδο. Ας υποθέσουμε τώρα δύο μεταβλητές $\boldsymbol{x_q}$ και $\boldsymbol{x_p}$ έτσι ώστε η συνάρτηση απεικόνισης τους να δίνει το αποτέλεσμα $\boldsymbol{p = j(q)}$. Επομένως, η συνάρτηση εξάρτησης μας έχει πει ότι το $\boldsymbol{x_q}$ είναι εξαρτώμενο από το $\boldsymbol{x_p}$. Αυτό ορίζεται στο γράφημα του δέντρου συνάρτησης από ένα βέλος που δείχνει το $\boldsymbol{x_p}$από το $\boldsymbol{x_q}$. Αν $\boldsymbol{j(q) = 0}$ δεν θα υπάρχει κανένα βέλος να δείχνει προς την αντίθετη κατεύθυνση της μεταβλητής $\boldsymbol{x_q}$.

2.1 Βέλτιστη προσέγγιση

Για τον καθορισμό των εξαρτήσεων, μια περαιτέρω διερεύνηση της προσεγγιστικής κατανομής πρέπει να επιχειρηθεί. Έστω η κατανομή πιθανότητας για n διακριτές μεταβλητές όπως $\boldsymbol{X = (X_1, X_2, \dots, X_N)}$, η κοινή πληροφορία μεταξύ της προσέγγισης $\boldsymbol{P_a(X)}$ της κατανομής, με αυτήν της πραγματικής κατανομής $\boldsymbol{P(X)}$ μπορεί να οριστεί ως:

$$\boldsymbol{I(P, P_a) = \sum_x P(X) \log \frac{P(X)}{P_a(X)}}$$

με ιδιότητα $\boldsymbol{I(P, P_a) \ge 0}$α

[1] Στο παρόν κεφάλαιο επιχειρείται ο ορισμός των δέντρων εξάρτησης και η μαθηματική θεωρία που τα συνοδεύει, η οποία βασίζεται στην θεωρία πιθανοτήτων. Οι εφαρμογές των δέντρων εξάρτησης, είναι ποικίλες όπως θα διαπιστωθεί από το σύνολο της εργασίας (μετάφραση, Q/A κ.λπ.). Το παρόν κομμάτι αναφέρεται στην χρήση του συγκείμενου για την αναγνώριση μοτίβων σε εικόνες. Υπάρχουν αρκετοί τρόποι προσέγγισης του προβλήματος του συγκειμένου στην αναγνώριση μοτίβων, με χρήση της μαθηματικής θεωρίας, τέτοιοι είναι οι αλυσίδες Makrov, που και αυτές βασίζονται στην θεωρία πιθανοτήτων, οι καμπύλες χώρου Hilbert κ.ά. Τα δέντρα εξάρτησης στην αναγνώριση μοτίβων σε εικόνες, χρησιμοποιούνται για να βρεθούν οι εξαρτήσεις μεταξύ των pixel που αποτελούν τα μοτίβα, και η πιθανότητες των διατάξεων.

Οι κοινές πληροφορίες μεταξύ δύο κατανομών πιθανοτήτων ορίζονται ως το μέτρο της εγγύτητας μεταξύ της κατανομής $\boldsymbol{P}$ χρησιμοποιώντας την κατά προσέγγιση κατανομή $\boldsymbol{P_a}$. Επιπλέον, το $\boldsymbol{I(P, P_a)}$ μπορεί να θεωρηθεί ως η διαφορά στο περιεχόμενο των πληροφοριών μεταξύ των δύο κατανομών πιθανότητας. Το μέτρο είναι πάντα θετικό εάν οι δύο κατανομές είναι διαφορετικές και μηδέν εάν είναι οι ίδιες. Θα χρησιμοποιήσουμε αυτό το μέτρο για να καθορίσουμε τη προσέγγιση μιας κατανομής πιθανοτήτων ως δέντρο εξάρτησης.

Γνωρίζουμε ότι αν επρόκειτο να κάνουμε μια εξαντλητική αναζήτηση για ένα σύνολο μεταβλητών $\boldsymbol{N}$, τότε υπάρχει η πιθανότητα $\boldsymbol{N^{N-2}}$με $\boldsymbol{N}$ κορυφές. Ως εκ τούτου, για έναν μεγάλο αριθμό $\boldsymbol{N}$, αυτό θα πάρει ένα τεράστιο χρονικό διάστημα, γι' αυτό είναι καλύτερο να προσεγγίσουμε την κατανομή πιθανότητας. Ως εκ τούτου, για μια κατανομή πιθανότητας $\boldsymbol{N - ιοστής}$ $\boldsymbol{σειράς}$ $\boldsymbol{P}$, θέλουμε να βρούμε την κατανομή του δέντρου εξάρτησης $\boldsymbol{P_a}$. Αρχικά, θα πρέπει να εξετάσουμε τους ακόλουθους δύο ορισμούς, προκειμένου να μας βοηθήσουν με το πρόβλημα βελτιστοποίησης:

Ορισμός 1:

Οι αμοιβαίες πληροφορίες μεταξύ δύο τυχαίων μεταβλητών $\boldsymbol{x_i}$ και $\boldsymbol{x_j}$ ορίζονται ως εξής:

$$\boldsymbol{I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{(x_i, x_j)}{P(x_i) P(x_j)}}$$

Ορισμός 2:

Αυτό που θα θέλαμε να καθορίσουμε είναι η εξάρτηση μέγιστου βάρους, όπου κάθε κλαδί έχει το μέγιστο δυνατό βάρος που μπορεί να έχει. Αυτό σημαίνει ότι οι αμοιβαίες πληροφορίες μεταξύ x_i και x_j είναι ένα μέγιστο. Η κατανομή πιθανότητας για ένα δέντρο εξάρτησης $P_t(X)$ είναι μια βέλτιστη προσέγγιση του $P(X)$ εάν και μόνο εάν το δέντρο εξάρτησής του είναι του μέγιστου βάρους.

Αν επεκτείνουμε:

$$\boldsymbol{I(P, P_a) = \sum_{x} P(X) \log \frac{P(X)}{P_a(X)}}$$

χρησιμοποιώντας $\boldsymbol{P_t(X)}$ και $\boldsymbol{P(X)}$ ως τις λειτουργίες που προσπαθούμε να βρούμε τις αμοιβαίες πληροφορίες τους, μπορούμε να επεκτείνουμε τα ακόλουθα:

$$\boldsymbol{I(P,P_t) = -\sum_{x} P(X) \sum_{i=1}^{n} \log P(x_i, x_{j(i)}) + \sum_{x} P(X) \log P(X)}$$

$$\boldsymbol{= -\sum_{i=1, j(i)\neq 0}^{n} \log \frac{P(x_i \mid x_{j(i)})}{P(x_i)P(x_{j(i)})} - \sum_{x} P(X) \sum_{i=1}^{n} \log P(x_i) + \sum_{x} P(X) \log P(X)}$$

(*)

Μπορούμε επίσης να ορίσουμε το $\boldsymbol{H(X)}$ ως το ακόλουθο:

$$\boldsymbol{H(X) = -P(X) \log P(x_i) = -\sum_{x_i} P(x_i) \log P(x_i)}$$

$$\boldsymbol{\sum_{x} P(X) \log \frac{P(x_i, x_{j(i)})}{P(x_i)P(x_{j(i)})} = \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)}) \log \frac{P(x_i, x_{j(i)})}{P(x_i)P(x_{j(i)})} = I(x_i, x_{j(i)})}$$

Τώρα, αν αντικαταστήσουμε στη σχέση (*) παίρνουμε τα ακόλουθα για να καθορίσουμε τη λειτουργία αμοιβαίας πληροφόρησης μας:

$$\boldsymbol{I(P,P_t) = -\sum_{i=1}^{n} I(x_i, x_{j(i)}) + \sum_{i=1}^{n} H(x_i) - H(X)}$$

$\boldsymbol{H(x_i)}$ και $\boldsymbol{H(X)}$ είναι και οι δύο ανεξάρτητες από το δέντρο εξάρτησης, και $\boldsymbol{I(P,P_t)}$ είναι μη αρνητικό, επομένως εάν είμαστε σε θέση να ελαχιστοποιήσουμε το μέτρο εγγύτητας, ή τη λειτουργία αμοιβαίας πληροφορίας, μπορούμε να μεγιστοποιήσουμε το βάρος του κλαδιού $\boldsymbol{I(x_i, x_{j(i)})}$.

Φυσικά, δεν θέλουμε να ψάξουμε κάθε πιθανότητα βάρους των κλαδιών προκειμένου να καταλάβουμε ποιο είναι το βέλτιστο δέντρο. Ως εκ τούτου, πρέπει να χρησιμοποιήσουμε μια διαδικασία βελτιστοποίησης.

2.2 Διαδικασία βελτιστοποίησης

Το πρόβλημα είναι να κατασκευαστεί ένα δέντρο εξάρτησης με το μέγιστο βάρος για το κλαδί του κάθε δέντρου.

1. Πρώτα γίνεται ο εντοπισμός του δείγματος $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n})$ από την εικόνα ώστε να ταξινομηθεί.
2. Έπειτα πραγματοποιείται ο υπολογισμός των μέτρων αμοιβαίας πληροφόρησης του κάθε ζεύγους $\boldsymbol{I}(\boldsymbol{x_i}, \boldsymbol{x_{j(i)}})$ ώστε να αξιολογηθούν.
3. Μετά πρέπει να βρεθούν τα μέγιστα μέτρα της αμοιβαίας πληροφορίας από τα ζεύγη ώστε να γίνουν οι πρώτοι κόμβοι των δέντρων εξάρτησης. Στη συνέχεια προστίθενται κλαδιά, αφού τα μέγιστα μέτρα αμοιβαίας πληροφόρησης είναι γνωστά. Το πρόβλημα είναι να γίνει το δέντρο εξάρτησης μέγιστου συνολικού βάρους κλαδιών.
   - διάταξη των κλαδιών ανάλογα με το μειούμενο βάρος τους, το βάρος είναι το βάρος είναι το μέτρο αμοιβαίων πληροφοριών ζευγών που παρουσιάζονται παραπάνω.
   - για κάποια κλαδιά b_iγια να είναι το βάρος του λιγότερο ή ίσο με το βάρος κάποιων κλαδιών $\boldsymbol{b_j}$, ισχύει $\boldsymbol{i < j}$
   - οι μοναδικές λύσεις είναι δυνατές εάν όλα τα βάρη των κλαδιών είναι διαφορετικά και οι ξεχωριστές λύσεις είναι δυνατές εάν υπάρχουν κάποια βάρη που είναι ίδια.

2.2.1 Παράδειγμα

Έστω η ακόλουθη κατανομή πιθανοτήτων για τέσσερα στοιχεία:

Οι αριθμοί εδώ είναι αρκετά άσχετοι, αλλά αυτό που είναι πιο σημαντικό, είναι να κατανοηθεί η διαδικασία που απαιτείται για τη δημιουργία ενός δέντρου εξάρτησης.

Μια δυαδική κατανομή πιθανοτήτων

$x_1x_2x_3x_4$	$P(x_1x_2x_3x_4)$	$P(x_1)\,P(x_2)\,P(x_3)\,P(x_4)$
0000	0.100	0.046
0001	0.100	0.046
0010	0.050	0.056

0011	0.050	0.056
0100	0.000	0.056
0101	0.000	0.056
0110	0.100	0.068
0111	0.050	0.068
1000	0.050	0.056
1001	0.100	0.056
1010	0.000	0.068
1011	0.000	0.068
1100	0.050	0.068
1101	0.050	0.068
1110	0.150	0.083
1111	0.150	0.083

Διάγραμμα 1: Mallar.

Από τα παρακάτω στοιχεία υπολογίζουμε τα ζεύγη αμοιβαίων πληροφοριών:

$$I(x_1, x_2) = 0.079$$

$$I(x_1, x_3) = 0.00005$$

$$I(x_1, x_4) = 0.0051$$

$$I(x_2, x_3) = 0.189$$

$$I(x_2, x_4) = 0.0051$$

$$I(x_3, x_4) = 0.0051$$

Κινούμενη εικόνα 1: Απεικόνιση της κατασκευής ενός δέντρου εξάρτησης (πατήστε διπλό κλικ) (*Mallar, McGill University).*

3. Εφαρμογή των δέντρων εξάρτησης στην απάντηση ερωτήσεων

Η απάντηση ερωτήσεων (Q/A) ανοιχτού τομέα σε φυσικές γλώσσες αποτελεί μια δύσκολη πρόκληση στην επεξεργασία φυσικών γλωσσών και έχει λάβει ιδιαίτερη προσοχή τα τελευταία χρόνια (Voorhees, 2000, 2001, 2002). Στον διαγωνισμό ερώτησης-απάντησης, κατά την διάρκεια του Συνέδριου Ανάκτησης Κειμένου (TREC)[2], δόθηκε το εξής ερώτημα ανοιχτής φόρμας, «Ποιο ήταν το μεγαλύτερο πλήθος που παρακολούθησε ποτέ τον Michael Jordan;» (Voorhees, 2002). Το σύστημα μπορεί να έχει πρόσβαση σε μια μεγάλη συλλογή άρθρων εφημερίδων, για να βρει την ακριβή απάντηση, π.χ. «62.046», μαζί με μια σύντομη πρόταση που να υποστηρίζει την απάντηση.

Οι συνολικές διεργασίες για να απαντηθεί ακόμα και ένα ερώτημα απλού τύπου όπως το παραπάνω είναι αρκετά περίπλοκες. Μια πλήρης Q/A, απαιτεί την δυνατότητα:

[2] Το Συνέδριο Ανάκτησης Κειμένου (TREC) είναι μια σειρά workshop που εστιάζουν σε ομάδες διαφορετικών ερευνητικών περιοχών ανάκτησης πληροφοριών (IR). Σκοπός του είναι να υποστηρίξει και να ενθαρρύνει την έρευνα στην κοινότητα ανάκτησης πληροφοριών παρέχοντας την απαραίτητη υποδομή για την ευρεία αξιολόγηση των μεθοδολογιών ανάκτησης κειμένου, για να αυξήσει την ταχύτητα της μεταφοράς τεχνολογίας από το εργαστήριο στην αγορά.

1. Ανάλυσης ερωτήσεων, προκειμένου να προσδιοριστεί το περιεχόμενο της ερώτησης (Li και Roth, 2002).
2. Ανάκτησης πιθανών υποψηφίων απαντήσεων από την δεδομένη συλλογή άρθρων
3. Καθορισμός της τελικής απάντησης στην ερώτηση.

Το 3 αφορά μόνο το τελευταίο στάδιο. Δηλαδή, υποθέτουμε ότι έχει ήδη δοθεί ένα σύνολο πιθανών απαντήσεων και στοχεύουμε στην επιλογή της κατάλληλης απάντησης.

Ένα πρόβλημα που προκύπτει, σχετίζεται με την αξιολόγηση της απόστασης μεταξύ μιας ερώτησης και καθεμιάς από τις υποψήφιες απαντήσεις. Η υποψήφια απάντηση με τη χαμηλότερη απόσταση από την ερώτηση επιλέγεται ως η τελική απάντηση. Η απλή τεχνική του bag-of-word[3] δεν αποδίδει σε αυτήν την περίπτωση, όπως φαίνεται στο ακόλουθο παράδειγμα που έχει ληφθεί από το (Harabagiu και Moldovan, 2001).

Ποιο είναι το γρηγορότερο αυτοκίνητο στον κόσμο;

Οι υποψήφιες απαντήσεις είναι:

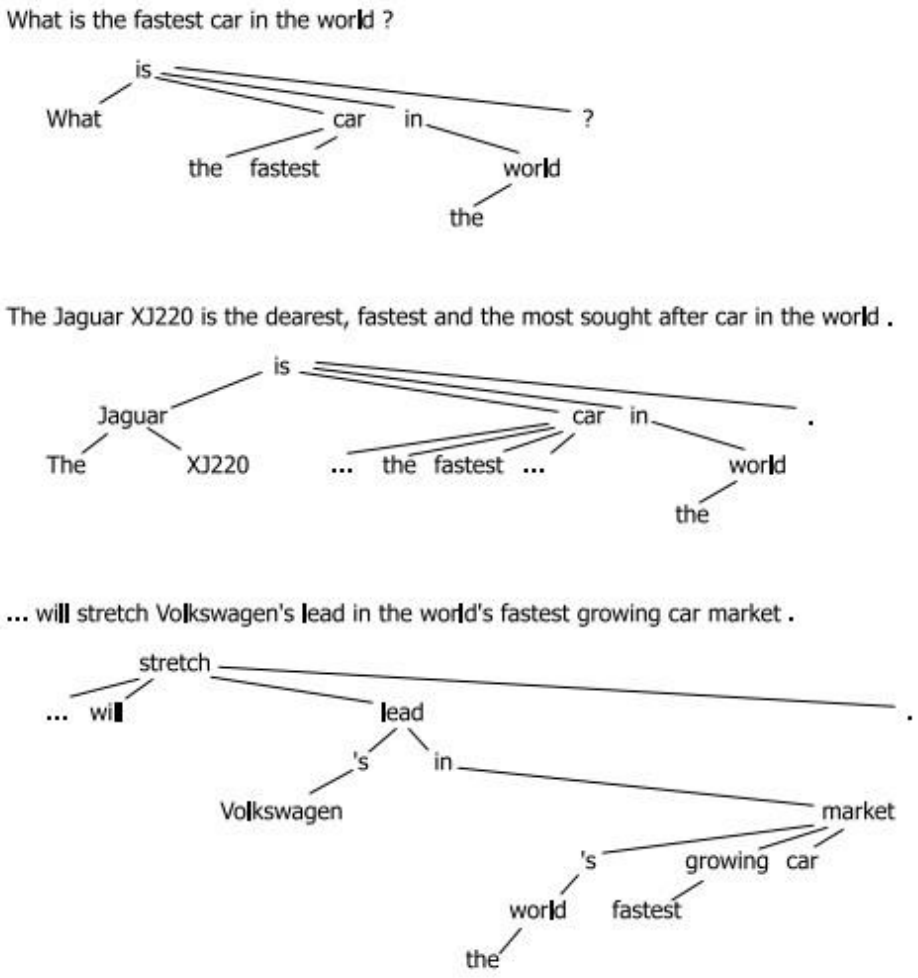
1. Η Jaguar XJ220 είναι το πιο ακριβό (415.000 λίρες), ταχύτερο (217 μίλια/ώρα) και το πιο περιζήτητο αυτοκίνητο στον κόσμο.
2. … θα επεκτείνει το προβάδισμα της Volkswagen στην ταχύτερα αναπτυσσόμενη αγορά οχημάτων στον κόσμο.

Χωρίς βαθιά ανάλυση των προτάσεων, δεν θα ήταν κατανοητό ότι το «γρηγορότερο» στη δεύτερη υποψήφια απάντηση δεν τροποποιεί (modify) το αυτοκίνητο όπως στο πρώτο και η προσέγγιση του «bag-of-word» θα αποτύχει. Επομένως, για να απαντήσουμε, αντί να καθορίσουμε την απόσταση από την αρχική ερώτηση, αναπαριστούμε πρώτα την ερώτηση και ύστερα την απάντηση, υπό την μορφή δέντρου εξάρτησης. Στη συνέχεια, ορίζουμε ένα μέτρο απόστασης μεταξύ των δέντρων εξάρτησης, λαμβάνοντας υπόψη τη δομή τους και ορισμένες σημασιολογικές ιδιότητες που συνάγουμε. Το σχήμα 2 δείχνει τα δέντρα εξάρτησης της ερώτησης και τις υποψήφιες απαντήσεις του προηγούμενου παραδείγματος. Αυτή η αναλυτική απεικόνιση μέσων δέντρων μας επιτρέπει να ταιριάξουμε καλύτερα ερώτηση με απάντηση.

Η τεχνική της αντιστοίχισης δέντρων εξάρτησης, έχει πρόσφατα προσελκύσει το ενδιαφέρον στην κοινότητα επεξεργασίας φυσικών γλωσσών στο πλαίσιο της μηχανικής μετάφρασης (Eisner, 2003; Gildea, 2003; Ding, 2003), αλλά όχι μέχρι στιγμής στην Q/A. επίσης, η αλγοριθμική προσέγγιση που παρουσιάσαμε είναι διαφορετική

[3] Το μοντέλο «bag-of-word» είναι μια απλοποιημένη αναπαράσταση που χρησιμοποιείται στην επεξεργασία των φυσικών γλωσσών και στην ανάκτηση πληροφοριών (IR). Σε αυτό το μοντέλο, ένα κείμενο (όπως μια πρόταση ή ένα έγγραφο) αντιπροσωπεύεται ως σάκος (πολλαπλό σύνολο) λέξεων του, αγνοώντας τη γραμματική και ακόμη και τη σειρά λέξεων, αλλά διατηρώντας την πολλαπλότητα.

από εκείνη που χρησιμοποιείται στη μηχανική μετάφραση. Η προσέγγισή μας χρησιμοποιεί την απόσταση επεξεργασίας με τον κατά προσέγγιση αλγόριθμο αντιστοίχισης δέντρων για τη μέτρηση της απόστασης μεταξύ των δέντρων (Zhang και Shasha, 1989).



Σχήμα 2: Ένα παράδειγμα δέντρων εξάρτησης για μια ερώτηση και τις υποψήφιες απαντήσεις. Για λόγους κατανόησης παραλείπονται ορισμένα τμήματα του δέντρου που δεν έχουν σημασία. (*Punyakanok,Roth & Yih,2004)*

Εξετάζεται η προσέγγιση με τα δέντρα εξάρτησης, στα ερωτήματα της TREC - 2002 Q/A. Η σύγκριση μεταξύ της απόδοσης της προσέγγισής των δέντρων εξάρτησης και της «bag-of-word» δείχνει το πλεονέκτημα της μεθόδου των δέντρων εξάρτησης.

3.1 Αντιστοίχηση δέντρων εξάρτησης στην απάντηση ερωτήσεων

Το ζητούμενο είναι η εύρεση της καλύτερης πρότασης, που περιέχει την απάντηση σε κάθε δοσμένη ερώτηση. Για να το κάνουμε αυτό, χρειαζόμαστε κάποιο μηχανισμό που μπορεί να μετρήσει πόσο συγκλίνει η υποψήφια απάντηση στην ερώτηση. Αυτό μας επιτρέπει να επιλέξουμε την τελική απάντηση που να ταιριάζει περισσότερο στην ερώτηση.

Για να επιτευχθεί αυτό, πρέπει να αναλυθεί το πρόβλημα σε δύο επίπεδα. Αρχικά, χρειάζεται μια απεικόνιση των προτάσεων που να επιτρέπει την σύλληψη χρήσιμων πληροφοριών προκειμένου να διενεργηθεί η διαδικασία αντιστοίχισης. Δεύτερον, χρειαζόμαστε μια αποτελεσματική διαδικασία αντιστοίχισης για να επεξεργαστούμε την επιλεγμένη απεικόνιση.

Στο πρώτο επίπεδο, η απεικόνιση πρέπει να συλλαμβάνει τόσο τη συντακτική όσο και τη σημασιολογική δομή μιας πρότασης. Για να συλλάβουμε την συντακτική δομή, αντιπροσωπεύουμε ερωτήσεις και απαντήσεις με τα δέντρα εξάρτησης που μας επιτρέπουν να δούμε ξεκάθαρα τις συντακτικές σχέσεις μεταξύ λέξεων στις προτάσεις. Η χρήση δέντρων μας επιτρέπει επίσης να ενσωματώσουμε με ευελιξία άλλες πληροφορίες, συμπεριλαμβανομένης της σημασιολογικής δομής. Επιτρέποντας σε κάθε κόμβο του δέντρου να περιέχει περισσότερα από την επιφανειακή μορφή της αντίστοιχης λέξης, μπορούμε να προσθέσουμε σημασιολογικές πληροφορίες π.χ. σε ποιον τύπο ονομαστικών οντοτήτων ανήκει η λέξη, συνώνυμα των λέξεων ή άλλες σχετικές λέξεις στον κόμβο. Επιπλέον, κάθε κόμβος μπορεί να γενικευτεί ώστε να περιέχει περισσότερες γλωσσικές μονάδες από μια λέξη (π.χ μια φράση ή μια ονομαζόμενη οντότητα).

Με την χρήση της κατάλληλης αναπαράστασης, αυτό που απομένει είναι να βρεθεί η αντιστοίχιση μεταξύ των κόμβων στην ερώτηση και στην πιθανή απάντηση. Για να διενεργηθεί αυτό, χρησιμοποιείται η κατά προσέγγιση αντιστοίχιση δέντρων που εξηγείται στην επόμενη ενότητα. Υποθέτουμε για κάθε ερώτηση q_1 τις υποψήφιες απαντήσεις, $\boldsymbol{A_1} = \{\boldsymbol{a_1}, \boldsymbol{a_2}, .. \boldsymbol{a_{n_i}}\}$ κάθε μια από τις οποίες είναι μια πρόταση. Έχουμε την τελική απάντηση για τη q_1 ,

$$\boldsymbol{a_i} = \underset{\boldsymbol{a} \in \boldsymbol{A_i}}{\arg\, \boldsymbol{min}} \boldsymbol{DR}(\boldsymbol{q_i}, \boldsymbol{a}),$$

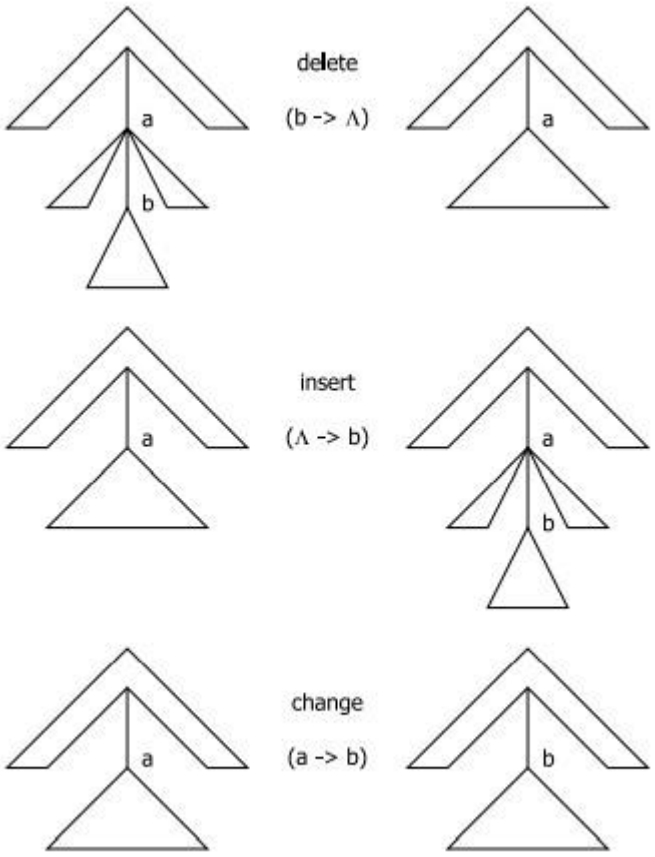
όπου DR δίνει την ελάχιστη κατά προσέγγιση αντιστοίχιση δέντρων.

3.2 Επεξεργασία αποστάσεων και αντιστοίχιση κατάλληλου δέντρου εξάρτησης

Η κατά προσέγγιση αντιστοίχιση δέντρων χρησιμοποιείται (Zhang και Shasha, 1989) για να αποφασιστεί η ομοιότητα μεταξύ των δοσμένων ζευγαριών των δέντρων. Αρχικά, εισάγεται η επεξεργασία απόστασης (Tai, 1979), που είναι το μέτρο απόστασης που χρησιμοποιείται ως κριτήριο αντιστοίχισης. Στη συνέχεια, εξηγείται με ακρίβεια πώς αυτό το μέτρο εφαρμόζεται στο πρόβλημα αντιστοίχισης των δέντρων.

Γίνεται χρήση εδώ, του τυπικού ορισμού που παρουσιάστηκε από (Tai, 1979) και (Zhang και Shasha, 1989). Θεωρούνται, διατεταγμένα επισημασμένα (labeled) δέντρα στα οποία κάθε κόμβος έχει επισημανθεί από κάποια πληροφορία και η σειρά από αριστερά προς τα δεξιά είναι σήμανση της σημαντικότητας. Η επεξεργασία απόστασης μετρά το κόστος μιας ακολουθίας λειτουργιών που μετατρέπει ένα διατεταγμένο δέντρο με ετικέτα σε άλλο. Οι λειτουργίες περιλαμβάνουν τη διαγραφή ενός κόμβου, την εισαγωγή ενός κόμβου και την αλλαγή ενός κόμβου. Το σχήμα 2 απεικονίζει την επίδραση αυτών των εργασιών. Πιο συγκεκριμένα, όταν ένας κόμβος n έχει διαγραφεί, «τα παιδιά του» θα συνδεθούν με το μητρικό του n. Η εισαγωγή είναι το αντίστροφο της διαγραφής. Η αλλαγή ενός κόμβου, είναι η αλλαγή της ετικέτας του. Κάθε λειτουργία ενέχει ένα σχετιζόμενο κόστος. Το κόστος μιας ακολουθίας λειτουργιών είναι το άθροισμα των κοστών κάθε λειτουργίας. Ζητούμενο αποτελεί η εξεύρεση μιας ακολουθίας ελάχιστου κόστους που επεξεργάζεται ένα δέντρο και το μετατρέπει σε άλλο.

Τυπικά μια λειτουργία αντιπροσωπεύεται από ένα ζεύγος $(\boldsymbol{a}, \boldsymbol{b})$ όπου το a αντιπροσωπεύει τον κόμβο προς επεξεργασία και το b το αποτέλεσμα της. Χρησιμοποιούμε $(\boldsymbol{a}, \boldsymbol{\Lambda})$ και $(\boldsymbol{\Lambda}, \boldsymbol{b})$ για να αντιπροσωπεύσουμε τη λειτουργία διαγραφής και εισαγωγής αντίστοιχα. Κάθε λειτουργία $(\boldsymbol{a}, \boldsymbol{b}) \neq (\boldsymbol{\Lambda}, \boldsymbol{\Lambda})$ σχετίζεται με το μη αρνητικό κόστος $\boldsymbol{\gamma}(\boldsymbol{a} \rightarrow \boldsymbol{b})$. Το κόστος μιας ακολουθίας λειτουργιών $\boldsymbol{S} = \langle \boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_k} \rangle$ είναι $\boldsymbol{\gamma}(\boldsymbol{S}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{\gamma}(\boldsymbol{s_i})$.



Σχήμα 3: Το αποτέλεσμα της διαγραφής, της εισαγωγής και της αλλαγής. (*Punyakanok,Roth & Yih,2004).*

Δεδομένου ενός δέντρου $\boldsymbol{T}$, συμβολίζουμε με $\boldsymbol{s(T)}$ το δέντρο που προκύπτει από την εφαρμογή της λειτουργίας $\boldsymbol{s}$ στο $\boldsymbol{T}$, και $\boldsymbol{s(T) = s_k(s_{k-1}(\dots(s_1(T))\dots))}$. Δεδομένου δύο δέντρων $\boldsymbol{T_1}$ και $\boldsymbol{T_2}$, πρέπει να βρεθεί το:

$$\boldsymbol{\delta(T_1, T_2) = min\{\gamma(S) | S(T_1) = T_2\}}$$

Εάν το κόστος ικανοποιεί την τριγωνική ανισότητα, δηλαδή $\boldsymbol{\gamma(a \to c) \leq \gamma(a \to b) + \gamma(b \to c) \forall a, b, c}$ τότε βάσει του (Tai, 1979) το ελάχιστο κόστος $\delta(T_1, T_2)$ είναι το ελάχιστο κόστος χαρτογράφησης. Η χαρτογράφηση $\boldsymbol{M}$ από $\boldsymbol{T_1}$ έως $\boldsymbol{T_2}$ είναι ένα σύνολο ακέραιων ζευγών που ικανοποιούν τις ακόλουθες ιδιότητες. Έστω $\boldsymbol{T[i]}$ να αντιπροσωπεύει τον $\boldsymbol{i - οστό}$ κόμβο του δέντρου $\boldsymbol{T}$ σε οποιαδήποτε δεδομένη σειρά, με N_1 και N_2 να είναι οι αριθμοί των κόμβων σε $\boldsymbol{T_1}$και$\boldsymbol{T_2}$,αντίστοιχα.

1. Για οποιοδήποτε ζεύγος $\boldsymbol{(i, j) \epsilon M}$, $\boldsymbol{1 \leq i \leq N_1}$ και $\boldsymbol{1 \leq j \leq N_2}$.

2. Για οποιαδήποτε ζεύγη $\boldsymbol{(i_1, j_1)}$ και $\boldsymbol{(i_2, j_2) \in M}$,

a. $\boldsymbol{i_1 = i_2}$ εάν και μόνο εάν $\boldsymbol{j_1 = j_2}$,

b. Το $\boldsymbol{T_1[i_1]}$ βρίσκεται στα αριστερά του $\boldsymbol{T_1[i_2]}$ εάν και μόνο εάν το $\boldsymbol{T_2[j_1]}$βρίσκεται στα αριστερά του $\boldsymbol{T_2[j_2]}$,
c. $\boldsymbol{T_1[i_1]}$ είναι πρόγονος του $\boldsymbol{T_1[i_2]}$ εάν και μόνο εάν το $\boldsymbol{T_2[j_1]}$ είναι πρόγονος του $\boldsymbol{T_2[j_2]}$.

Το κόστος μιας χαρτογράφησης M είναι

$$\boldsymbol{\gamma(M) = \sum_{(i,j)\epsilon M} \gamma(T_1[i] \rightarrow T_2[j]) + \sum_{(i,j)\epsilon I} \gamma(T_1[i] \rightarrow \Lambda) + \sum_{(i,j)\epsilon J} \gamma(\Lambda \rightarrow T_2[j])}$$

όπου $\boldsymbol{I}$ είναι το σύνολο του ευρετηρίου των κόμβων στο $\boldsymbol{T_1}$ που δεν αντιστοιχίζεται από το $\boldsymbol{M}$ και το $\boldsymbol{J}$ είναι αυτό των κόμβων στο $\boldsymbol{T_2}$.

Γενικά η απόσταση επεξεργασίας μπορεί να χρησιμοποιηθεί για να αποφασιστεί η ομοιότητα κάθε ζεύγους δέντρων. Ωστόσο, κατά την αντιστοίχιση ερωτήσεων και προτάσεων απαντήσεων στον τομέα απαντήσεων-ερωτήσεων, η ακριβής απάντηση σε μια ερώτηση μπορεί να βρίσκεται μόνο σε μια φράση της πρότασης. Επομένως, η αντιστοίχιση της ερώτησης με ολόκληρη την υποψήφια πρόταση-απάντηση μπορεί να οδηγήσει σε λανθασμένη αντιστοίχιση, παρόλο που η πρόταση περιέχει τη σωστή απάντηση. Η κατά προσέγγιση αντιστοίχιση δέντρων μας επιτρέπει να αντιστοιχίσουμε την ερώτηση σε αποκλειστικά μέρη της πρότασης και όχι σε ολόκληρη. Συγκεκριμένα, δεν υπάρχει πρόσθετο κόστος εάν διαγραφούν ορισμένα δευτερεύοντα δέντρα της απάντησης.

Έστω ότι $\boldsymbol{T_1}$ και $\boldsymbol{T_2}$ είναι δύο δέντρα προς αντιστοίχιση. Ένα δάσος $\boldsymbol{S}$ ενός δέντρου $\boldsymbol{T}$ είναι ένα σύνολο υποδένδρων στο $\boldsymbol{T}$ έτσι ώστε όλα τα υποδέντρα στο $\boldsymbol{S}$ να διαχωρίζονται και το $\boldsymbol{T \setminus S}$ είναι το νέο δέντρο που προκύπτει από την κοπή όλων των υποδέντρων στο $\boldsymbol{S}$ από το $\boldsymbol{T}$. Έστω ότι το $\boldsymbol{S}$ ($\boldsymbol{T}$) αντιπροσωπεύει το σύνολο όλων των πιθανών δασών του T. Το κατά προσέγγιση δένδρο που ταιριάζει μεταξύ T_1και T_2 είναι το:

$$\boldsymbol{DR(T_1, T_2) = \min_{S\epsilon S(T_2)} \delta(T_1, T_2 \setminus S)}$$

Στην εργασία των (Zhang και Shasha, 1989) παρέχεται ένας αποτελεσματικός αλγόριθμος δυναμικού προγραμματισμού για τον υπολογισμό της κατά προσέγγιση αντιστοίχισης δέντρων.

Αξίζει να σημειωθεί εδώ, ότι παρόλο που οι λειτουργίες κόστους που χρησιμοποιούμε στο παρακάτω πείραμα δεν ικανοποιούν την τριγωνική ιδιότητα, αυτό δεν επηρεάζει τις υποκείμενες θεωρίες του αλγορίθμου. Η ιδιότητα απαιτείται μόνο στην

απόδειξη της σχέσης μεταξύ της ελάχιστης απόστασης λειτουργίας επεξεργασίας και της χαρτογράφησης ελάχιστου κόστους.

3.3 Πείραμα

Το πείραμα συντελέστηκε με τις 500 ερωτήσεις που δόθηκαν στον διαγωνισμό TREC-2002 Q/A . Υπήρχαν 46 ερωτήσεις που δεν είχαν σωστή απάντηση. Οι σωστές απαντήσεις για κάθε ερώτηση, αν υπήρχαν, δόθηκαν μαζί με τις απαντήσεις όλων των συμμετεχόντων μετά την ολοκλήρωση του διαγωνισμού. Κατασκευάστηκε λοιπόν μια ομάδα υποψήφιων απαντήσεων, για κάθε ερώτηση από όλες τις πιθανές απαντήσεις των συμμετεχόντων (σωστές ή μη). Οι συμπερίληψη όλων των απαντήσεων, έκανε το πρόβλημα πιο δύσκολο για τον αλγόριθμο. Κανονικά, μια διαδικασία επιλογής απαντήσεων αξιολογείται με βάση το υποψήφιο σύνολο σωστών απαντήσεων που δημιουργήθηκε από τη σωστή απάντηση και την έξοδο από μια μηχανή αναζήτησης πληροφοριών. Ωστόσο, η ομάδα υποψηφίων περιείχε και τις εσφαλμένες απαντήσεις που δόθηκαν από άλλα συστήματα.

Λόγω της διαφορετικότητας της δομής των προτάσεων σε σχέση με την ερώτηση, οι ερωτήσεις μετασχηματίστηκαν χρησιμοποιώντας απλούς κανόνες ευρετικής[4]. Σε αυτόν τον μετασχηματισμό, η ερωτηματικές αντωνυμίες (π.χ. *τι, πότε* ή *πού)* αντικαταστάθηκαν με ένα ειδικό διακριτικό *ANS*. Ακολουθεί ένα παράδειγμα αυτού του μετασχηματισμού.

Πού είναι ο πύργος του διαβόλου ;

Ο πύργος του διαβόλου βρίσκεται σε *ANS*

Κάθε πρόταση πέρασε μια αρχική επεξεργασία όπου προστέθηκε μια ετικέτα από ένα πρόγραμμα SNoW. Στη συνέχεια, η αυτόματη πλήρης συντακτική ανάλυση (Collins, 1997) εκτελέστηκε για την παραγωγή των δέντρων. Δεδομένου ότι αυτός ο αναλυτής εξάγει επίσης την κύρια λέξη κάθε συστατικού, θα μπορούσαμε να μετατρέψουμε άμεσα τα δέντρα ανάλυσης στο αντίστοιχο δέντρο εξάρτησης, παίρνοντας απλώς τη κύρια λέξη ως γονική. Επιπλέον, εξαγάγαμε πληροφορίες ονομαστικής οντότητας με το αναγνωριστικό ονομαστικής οντότητας που χρησιμοποιήθηκε στο (Roth et.al., 2001). Τέλος, για κάθε ερώτηση, χρησιμοποιήθηκε ένας ταξινομητής ερωτήσεων

[4] «Ευρετική» είναι οποιαδήποτε προσέγγιση για την επίλυση προβλημάτων, που χρησιμοποιεί μια πρακτική μέθοδο που δεν είναι εγγυημένη ότι είναι βέλτιστη, τέλεια ή λογική, αλλά είναι ωστόσο επαρκής για την επίτευξη ενός άμεσου, βραχυπρόθεσμου στόχου ή προσέγγισης. Όταν η εξεύρεση βέλτιστης λύσης είναι αδύνατη ή ανέφικτη, μπορούν να χρησιμοποιηθούν ευρετικές μέθοδοι για να επιταχυνθεί η διαδικασία εύρεσης μιας ικανοποιητικής λύσης.

(Li και Roth, 2002), που προέβλεπε τον τύπο των απαντήσεων που αναμενόταν από την ερώτηση.

Μετά την εύρεση της απάντησης, επιστράφηκε το αναγνωριστικό εγγράφου που περιείχε την απάντηση. Μετρήσαμε ως σωστό, αν το αναγνωριστικό εγγράφου που επιστράφηκε, ταιριάζει με αυτό της σωστής απάντησης.

Ορίσαμε τρεις τύπους λειτουργιών κόστους , δηλαδή, διαγραφή, εισαγωγή και αλλαγή, όπως φαίνεται στο σχήμα 4. Η λίστα λέξεων διακοπής περιείχε μερικές από τις πολύ κοινές λέξεις που δεν θα ήταν πολύ σημαντικές, π.χ. το άρθρο όπως "ένα", "το". Οι μορφές του λήμματος εξήχθησαν χρησιμοποιώντας το WordNet (Miller et. Al., 1990).

Συγκρίθηκε η προσέγγισή των δέντρων εξάρτισης με την απλή μέθοδο του bag-of-words. Σε αυτήν την απλή προσέγγιση, μετρήσαμε την ομοιότητα μεταξύ μιας ερώτησης και μιας υποψήφιας απάντησης με τον αριθμό των κοινών λέξεων, είτε στις επιφανειακές τους μορφές είτε στις μορφές λήψης μεταξύ της ερώτησης και της απάντησης διαιρεμένη με τη διάρκεια αυτής της απάντησης. Η τελική απάντηση ήταν αυτή που παρήγαγε την υψηλότερη ομοιότητα.

1. διαγραφή:
   εάν το a είναι λέξη διακοπής, $\gamma(a \rightarrow \Lambda) = 5$,
   αλλιώς $\gamma(a \rightarrow \Lambda) = 200$.

2. Εισαγωγή
   εάν το a είναι λέξη διακοπής, $\gamma(\Lambda \rightarrow \alpha) = 200$,
   αλλιώς $\gamma(\Lambda \rightarrow \alpha) = 5$.

3. Αλλαγή
   εάν το a είναι *ANS*
   εάν το b ταιριάζει με τον αναμενόμενο τύπο απάντησης,$\gamma(a \rightarrow b) = 5$,
   αλλιώς $\gamma(a \rightarrow b) = 200$

   αλλιώς

   εάν η λέξη a είναι ίδια με τη λέξη b, $\gamma(a \rightarrow b) - 0$
   αλλιώς αν a και b έχουν την ίδια μορφή λήμματος,$\gamma(a \rightarrow b) = 1$, αλλιώς
   $\gamma(a \rightarrow b) = 200$.

Σχήμα 4: Υπόδειγμα συναρτήσεων υπολογισμού κόστους, μετάφραση από: *Punyakanok,Roth & Yih,2004)*

Αξίζει να σημειωθεί ότι η μέθοδος αξιολόγησης που χρησιμοποιήθηκε εδώ είναι διαφορετική από εκείνη στον διαγωνισμό TREC-2002 Q/A. Στο TREC, η απάντηση που δόθηκε από το σύστημα αποτελείται από το κλειδί της απάντησης και το έγγραφο που υποστηρίζει την απάντηση. Η απάντηση θεωρείται σωστή μόνο όταν το κλειδί απάντησης και το υποστηρικτικό έγγραφο είναι σωστά. Επειδή το εν λόγω σύστημά δεν παρέχει το κλειδί απάντησης, η αξιολόγηση του συστήματός βασίζεται μόνο στην εύρεση του σωστού δικαιολογητικού. Ωστόσο, αυτό δεν απλοποιεί σημαντικά το έργο, καθώς το δυσκολότερο μέρος της επιλογής απαντήσεων είναι η εύρεση του σωστού εγγράφου. Το κλειδί απάντησης μπορεί να εξαχθεί αργότερα με ορισμένους κανόνες ευρετικής. Επίσης, στην πράξη, ένας χρήστης που χρησιμοποιεί ένα σύστημα Q/A είναι πολύ πιθανό να απορρίψει το εύρημα, χωρίς ένα σωστό έγγραφο υποστήριξης. Ακόμα κι αν το σύστημα δεν παρέχει ένα σωστό κλειδί απάντησης, ο χρήστης μπορεί εύκολα να το βρει με δεδομένο ένα σωστό υποστηρικτικό έγγραφο.

Στον Πίνακα 1 φαίνεται μια σύγκριση μεταξύ των δύο μεθόδων και η αποτελεσματικότητα της χρήσης των δέντρων εξάρτησης.

		Correct	
Method	#	%	%(454)
Tree Matching	183	36.60	40.31
Bag-of-word	131	26.20	28.85

Πίνακας 1: Η σύγκριση της απόδοσης της προσεγγιστικής μεθόδου της αντιστοίχισης των δέντρων και της μεθόδου Bag-of-words. Η τελευταία στήλη δείχνει το ποσοστό μόνο για τις 454 ερωτήσεις που έχουν απαντήσεις. (Punyakanok,Roth & Yih,2004).

4. Δέντρα δομής εξάρτησης στην συντακτική μηχανική μετάφραση

Η Μηχανική Μετάφραση (ΜΤ) αποτελεί ένα ενδιαφέρον και άλυτο πρόβλημα. Η «Μηχανική Μετάφραση» ασχολείται με τη μετάφραση μιας πρότασης από μια γλώσσα (πηγή) σε μια πρόταση σε άλλη γλώσσα (στόχο), διατηρώντας το νόημα στην ολότητα του. Αυτό απαιτεί από τον υπολογιστή να κωδικοποιήσει το νόημα και των δύο γλωσσών σε μια αναπαράσταση που μπορεί να χρησιμοποιηθεί κατά το χρόνο εκτέλεσης, για τη μετάφραση. Τα προηγούμενα χρόνια το μεταφραστικό πρόβλημα προσεγγίστηκε με τρεις βασικούς τρόπους.

1. Με την απομνημόνευση όλων των προτάσεων της πηγής και της γλώσσας-στόχου, πριν τη στιγμή της επεξεργασίας, μέθοδος που ονομάζεται Μετάφραση Μνήμης (ΤΜ) και μεταφράζοντας κατά το χρόνο εκτέλεσης με μια απλή διαδικασία αναζήτησης-αντιστοίχισης (Hutchins και Somers, 1992).
2. Χρησιμοποιώντας πλήρεις γνώσεις της γλώσσας πηγής για ανάλυση και παραγωγή της μετάφρασης σύμφωνα με τη σύνταξη της γλώσσας-στόχου (Nirenburg et al., 1992)
3. Μια πιο σημασιολογικά υποκινούμενη και συγχρονική προσέγγιση της μετάφρασης. Σε αυτή την προσέγγιση η γλώσσα-πηγή μεταφράζεται σε μια ενδιάμεση γλώσσα που ονομάζεται Interlingua, μια αφηρημένη γλωσσικά-ανεξάρτητη αναπαράσταση. Η γλώσσα-στόχος προκύπτει μέσα από την Interlingua (Hutchins και Somers, 1992).

Αν και αυτές οι προσεγγίσεις έχουν μελετηθεί εκτενώς τις τελευταίες δεκαετίες, μια πιο ελπιδοφόρα προσέγγιση που ονομάζεται Στατιστική Μηχανική Μετάφραση (SMT) βασιζόμενη σε στατιστικούς και μαθηματικούς κανόνες έχει ανέλθει στο μεταφραστικό προσκήνιο κατά την τελευταία δεκαετία (Brown et al. 1993; Koehn, Och, και Marcu, 2003).

Η Στατιστική Μηχανική Μετάφραση (SMT) χρησιμοποιεί μεγάλες ποσότητες πηγών/δεδομένων για να «μάθει» μοντέλα μετάφρασης σε υπο-προτασιακό[5] επίπεδο. Τα μοντέλα, αυτά, μπορούν να γενικευτούν/εφαρμοστούν επιτυχώς σε άγνωστα δεδομένα. Η Στατιστική Μηχανική Μετάφραση αντιμετωπίζει το πρόβλημα της

[5] Σαν υπό-προτασιακό αποδίδεται το sub-sentential, δηλαδή την μετάφραση μερών της πρότασης, επί παραδείγματι την απόδοση φράσεων.

μετάφρασης ως ένα θορυβώδες κανάλι[6], όπου το μοντέλο του καναλιού (διαύλου) ονομάζεται συνήθως το «μοντέλο μετάφρασης» και το μοντέλο της πηγής ονομάζεται «μοντέλο γλώσσας». Το μεταφραστικό μοντέλο υπολογίζεται από μια παραγωγική διαδικασία αντιστοίχισης λέξεων. Το γλωσσικό μοντέλο εκτιμάται ως μια n-gram[7] αλληλουχία με αλυσίδες Markov[8]. Υπό αυτή την επεξεργασία το μεταφραστικό πρόβλημα είναι γλωσσικά αγνωστικιστικό και δεν περιλαμβάνει κανενός είδους συντακτικές πληροφορίες, ούτε από την γλώσσα-πηγή, ούτε από την γλώσσα-στόχο. Η μετάφραση που παράγεται, στο πλαίσιο της στατιστικής μηχανικής μετάφρασης, τείνει συνήθως να είναι κατακερματισμένη και ανεξάρτητη του μεταφραστικού πλαισίου.

Με την χρήση τεχνικών εκτίμησης και ευρετικής, μπορεί να γίνει ενσωμάτωση του πλαισίου (Koehn, Och, και Marcu, 2003), πράγμα που έχει βελτιώσει τις μεταφράσεις τα τελευταία χρόνια. Αλλά, η ποιότητα απέχει πολύ από αυτή της ανθρώπινης μετάφρασης. Αυτό, έχει να κάνει κυρίως με το γεγονός, ότι αυτά τα μοντέλα είναι αγνωστικιστικά, χωρίς συντακτικές πληροφορίες και έτσι είναι ανεπαρκώς ενημερωμένα σχετικά με τις συντακτικές διαφορές μεταξύ των γλωσσών. Κάποιες πρόσφατες προσεγγίσεις εξέτασαν την ενσωμάτωση της σύνταξης σε διάφορες φάσεις της μεταφραστικής διαδικασίας με επιτυχία (Yamada και Knight, 2001; Chiang, 2005).

Η σύνταξη μπορεί να είναι οποιασδήποτε μορφής, που κυμαίνεται από προσθήκες βασιζόμενες σε καταγραφή προφορικού λόγου, έως ολόκληρα δέντρα ανάλυσης που μπορούν να συμπεριλάβουν διάφορους γραμματικούς κανόνες.

Παρακάτω, θα εξεταστούν προσεγγίσεις ενσωμάτωσης των δέντρων εξάρτησης στη μεταφραστική διαδικασία. Ειδικότερα, όσα σχετίζονται με τα μοντέλα συντακτικής μετάφρασης και την εκτίμησή τους.

[6] Το θορυβώδες μοντέλο καναλιού είναι ένα πλαίσιο που χρησιμοποιείται σε ορθογραφικούς ελέγχους, ερωτήσεις, αναγνώριση ομιλίας και στη Μηχανική Μετάφραση. Σε αυτό το μοντέλο, ο στόχος είναι να βρούμε την επιδιωκόμενη λέξη μέσω μιας λέξη όπου τα γράμματα έχουν κωδικοποιηθεί με κάποιο τρόπο. Έστω ένα αλφάβητο $\boldsymbol{\Sigma}$ *με* $\boldsymbol{\Sigma}*$ το σύνολο όλων των πεπερασμένων σειρών του $\boldsymbol{\Sigma}$**.** Έστω το $\boldsymbol{D}$ το λεξικό των επιτρεπτών λέξεων να είναι υποσύνολο του $\boldsymbol{\Sigma}*$**,** $\boldsymbol{D} \subseteq \boldsymbol{\Sigma}*$**.** Το θορυβώδες κανάλι είναι ο πίνακας $\boldsymbol{\Gamma_{ws}} = \mathbf{Pr}\,(\boldsymbol{s} \,/\, \boldsymbol{w})$, όπου $\boldsymbol{w} \in \boldsymbol{D}$, είναι η ζητούμενη λέξη και $\boldsymbol{s} \in \boldsymbol{\Sigma}*$, είναι η κωδικοποιημένη λέξη που λάβαμε μέσα από το κανάλι.

[7] Στον τομέα της υπολογιστικής γλωσσολογίας, n-gram ονομάζεται μια συνεχόμενη ακολουθία n αντικειμένων από ένα δεδομένο δείγμα κειμένου ή ομιλίας

[8] Μια αλυσίδα Markov είναι ένα στοχαστικό μοντέλο που περιγράφει μια ακολουθία πιθανών γεγονότων στα οποία η πιθανότητα κάθε συμβάντος εξαρτάται μόνο από γνώση του προηγούμενου συμβάντος.

4.2 Σύνταξη στη στατιστική μηχανική μετάφραση

4.2.1 Στατιστική μηχανική μετάφραση

Τα μοντέλα IBM (Brown et al., 1993) εισήχθησαν για να μοντελοποιήσουν το μεταφραστικό πρόβλημα σε στατιστικό πλαίσιο. Ονομάζονται επίσης και μοντέλα ευθυγράμμισης και κυμαίνονται από το 1 έως το 5 με αυξανόμενη εκλέπτυνση, για να εξηγήσουν τις διάφορες αποκλίσεις που συμβαίνουν μεταξύ των γλωσσών. Αυτό συνέβαλε στη βελτίωση της αντιστοίχισης λέξη προς λέξη που είναι σημαντική για τη συνολική μετάφραση. Ωστόσο, υπήρχαν μερικά προβλήματα, τα μοντέλα μετάφρασης σε επίπεδο λέξης δεν μπορούσαν να αποδώσουν όπως τα μεταφραστικά μοντέλα που συμπεριλάμβαναν πληροφορίες πλαισίου. Αυτό οδήγησε την έρευνα στον τομέα μετάφρασης φράσεων (PBSMT), τομέας ο οποίος απομακρύνεται από τους θεμελιώδεις περιορισμούς των μοντέλων που βασίζονται σε λέξη προς λέξη μετάφραση. Τα μοντέλα μετάφρασης φράσεων, θεωρούν τις φράσεις ως την μικρότερη μονάδα μετάφρασης, διασκελίζοντας έτσι την ανάγκη σύνθεσης, των μοντέλων μετάφρασης λέξη προς λέξη. Μερικές φορές, ακόμη και αν μια φράση μεταφραστεί λέξη προς λέξη με μια τέλεια γραμματική μετάφραση, μπορεί να μην είναι έγκυρη για τη γλώσσα-στόχο. Ακόμη και όταν μια φράση εμφανίζεται ως σύνθεση μετάφρασης λέξη προς λέξη η ενσωμάτωση των πληροφοριών του υπόλοιπου περιβάλλοντος βοηθά (Quirk και Menezes, 2006). Αυτό όχι μόνο επιτρέπει την κατασκευή άπταιστων μεταφράσεων για τα υπο-προτασιακά τμήματα, αλλά και συμπεριλαμβάνει εγγενώς τη δυνατότητα λεκτικής αναδιάταξης εντός της φράσης.

4.2.2 Σύνταξη στη Μηχανική Μετάφραση

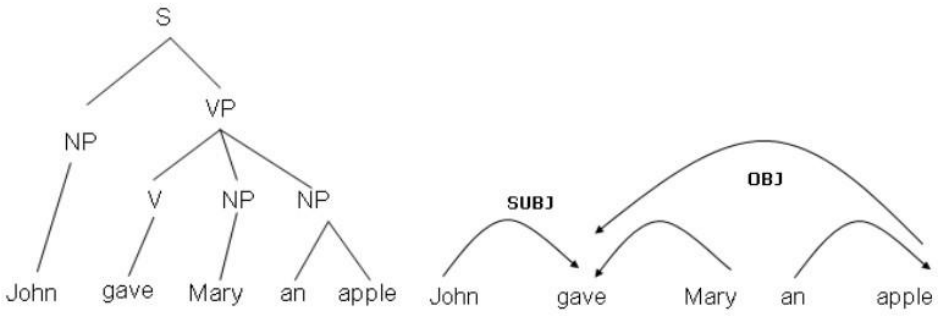
Τα τελευταία χρόνια υπάρχει αυξανόμενο ενδιαφέρον για μεθόδους που ενσωματώνουν συντακτικές πληροφορίες/γνώσεις στην μηχανική μετάφραση. Οι συντακτικοί κανόνες αναδιάταξης μπορούν να χρησιμοποιηθούν ως ένα προηγούμενο βήμα της μηχανικής μετάφρασης φράσεων, για τη μείωση της απόκλισης των λέξεων και της συντακτικής στρέβλωσης μεταξύ των γλωσσών προέλευσης και του στόχου (Xia και McCord, 2004). Υπήρξαν επίσης προσεγγίσεις, για την επαναξιολόγησης της κατάταξης του τελικού n-βέλτιστου καταλόγου των πιθανών μεταφράσεων, που παράγονται από τα τυποποιημένα συστήματα μηχανικών μεταφράσεων. Ορισμένες προσεγγίσεις

έχουν επιχειρήσει να συνδέσουν ακόμα πιο άμεσα σύνταξη και μεταφραστικό μοντέλο. Έχουν επίσης αναπτυχθεί διάφορα ιεραρχικά και συντακτικά μοντέλα, τα οποία εφαρμόζονται κατά την αποκωδικοποίηση (Yamada και Knight, 2001). Πολλές από αυτές τις προσεγγίσεις περιλαμβάνουν την αυτόματη εκμάθηση και εξαγωγή των υποκείμενων συντακτικών κανόνων από τα δεδομένα. Οι υποκείμενοι κανόνες που χρησιμοποιούνται εμπεριέχουν μια ευρεία γκάμα και περιλαμβάνουν απλούς κανόνες όπως ITGs (Wu, 1997), ιεραρχικούς συγχρονικούς κανόνες (Chiang, 2005), σειρές στα πρότυπα δέντρων, σύγχρονα μοντέλα CFG (Xia & McCord, 2004) (Yamada & Knight, 2001).

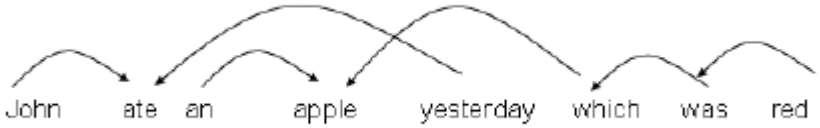
4.3 Δέντρα δομής εξάρτησης

Οι δομές εξάρτησης αντιπροσωπεύουν τις γραμματικές σχέσεις μεταξύ των συστατικών στοιχείων μιας πρότασης. Είναι πιο αφηρημένες σε σύγκριση με τα συντακτικά δέντρα, υπό την έννοια ότι δεν περιορίζουν ή προ-ορίζουν μια συγκεκριμένη σειρά λέξεων. Είναι πιο προσδιορισμένες σε σημασιολογικό επίπεδο και οι σχέσεις μεταξύ των λέξεων είναι πιο σαφείς.

Ένα δέντρο εξάρτησης μιας πρότασης είναι ένα ακυκλικό προσανατολισμένο γράφημα λέξεων με τις λέξεις να είναι οι κόμβοι και τις σχέσεις των λέξεων να είναι τα άκρα. Κάθε λέξη της πρότασης είτε τροποποιεί μια άλλη λέξη, είτε τροποποιείται από μια λέξη. Η ρίζα του δέντρου είναι η μοναδική λέξη που τροποποιεί αλλά δεν τροποποιείται (εξαρτάται) από κάποια άλλη λέξη. Η σχέση μεταξύ δύο λέξεων στο δέντρο μπορεί να δοθεί ως σχέση «γονέα-παιδιού» ή «κύριας-εξαρτώμενης» ή «κυβερνήτη-κυβερνώμενου». Η ειδική σχέση μεταξύ των δύο λέξεων δίνεται ως τίτλος στα άκρα που συνδέουν τους κόμβους. Η κατεύθυνση των βελών είναι συνήθως από την «κύρια προς την εξαρτώμενη», αλλά το αντίθετο είναι επίσης πιθανό, δεδομένης μιας προσυμφωνημένης σημειογραφίας για ολόκληρο το δέντρο. Τυπικά, το δέντρο εξάρτησης ορίζεται ως εξής: Δεδομένης μιας πρότασης $\boldsymbol{S}\{\boldsymbol{w_0}, \dots, \boldsymbol{w_n}\}$, ένα σύνολο από άκρες $\boldsymbol{E}\{\boldsymbol{e_1}, \dots, \boldsymbol{e_n}\}$ ορίζονται έτσι ώστε κάθε e_i να συνδέει δύο λέξεις στην πρόταση και w0 να είναι η ρίζα που συνδέεται με μια λέξη χωρίς να εξαρτάται από άλλη.



Σχήμα 5: Δέντρο δομής ***φράσης*** *και το αντίστοιχο δέντρο δομής* ***εξάρτησης***. (Ambati,2005).



Σχήμα 6: Ένα παράδειγμα ***μη προβολικού*** *δέντρου εξάρτησης* (Ambati,2005).

Στο δέντρο εξάρτησης στο σχήμα 5 τα άκρα παίρνουν την ονομασία τους από την σχέση την οποία περιγράφουν. Ένα δέντρο εξάρτησης με ονομασίες είναι χρήσιμο για την κατανόηση των γραμματικών λειτουργιών των λέξεων και των ρόλων που παίζουν σε μια πρόταση. Για παράδειγμα, ο «Ιωάννης» είναι το υποκείμενο της πρότασης που έχει ρήμα το «έφαγε» και το μήλο είναι το αντικείμενο της ίδιας πρότασης. Οι επισημασμένες εξαρτήσεις (σχέσεις) έχουν αποδειχθεί χρήσιμες στην επεξεργασία των φυσικών γλωσσών, για παράδειγμα στην απάντηση ερωτήσεων και την ανάλυση λόγου. Αν και οι επισημασμένες εξαρτήσεις έχουν χρησιμοποιηθεί για τη βελτίωση της αξιολόγησης της Μηχανικής Μετάφρασης (Owczarzak, van Genabith, and Way, 2007), δεν υπάρχει κάποια εργασία με αντικείμενο αναφοράς την απλή μετάφραση.

Στο σχήμα 5 αν και η δομή εξάρτησης δεν έχει αναπαρασταθεί ως ένα διακλαδισμένο δέντρο που να αντιστοιχεί με ένα δέντρο μετάφρασης ολόκληρων φράσεων, έχει χρησιμοποιηθεί ένας τρόπος αναπαράστασης με γραμμική σειρά που διατηρεί τη θέση των λέξεων. Με αυτή την αναπαράσταση παρατηρείται πιο εύκολα η αλληλεξάρτηση μεταξύ μακρινών σε απόσταση λέξεων. Είναι επίσης εύκολο με αυτό τον τρόπο (σχήμα 5), να εξηγηθεί μία αρχή των δομών εξάρτησης που ονομάζεται προβολικότητα[9]. Όλες οι προτάσεις στη φυσική γλώσσα μπορούν να εξηγηθούν από μια διχοτόμηση των δέντρων εξάρτησης σε προβολικά και μη προβολικά δέντρα. Ένα προβολικό δέντρο είναι αυτό που όταν αναπαριστάται με τις λέξεις σε μια προκαθορισμένη γραμμική σειρά, δεν υπάρχουν διασταυρώσεις μεταξύ των άκρων. Μπορούμε επίσης να πούμε ότι ένα δέντρο είναι προβολικό, αν με άκρο που ξεκινάει από την $\boldsymbol{w_i}$ και φτάνει στη $\boldsymbol{w_j}$, οποιαδήποτε λέξη $\boldsymbol{w_k}$ ανάμεσα, είναι μια άμεση ή έμμεση εξαρτώμενη της $\boldsymbol{w_i}$.

Για την αγγλική γλώσσα, τα περισσότερα από τα δέντρα γραμματικής ανάλυσης είναι προβολικά. Ωστόσο, υπάρχουν ορισμένα παραδείγματα στα οποία ένα μη προβολικό δέντρο είναι προτιμότερο. Έστω η πρόταση, ο Γιάννης έφαγε ένα μήλο χθες που ήταν κόκκινο. Εδώ η αναφορική πρόταση «που ήταν κόκκινο» και το αντικείμενο

[9] Η «προβολικότητα» είναι μια αρχή των δομών των δέντρων με την οποία εντοπίζονται και ορίζονται ασυνέχειες. Μια δομή δέντρου λέγεται ότι είναι προβολική εάν δεν υπάρχουν άκρα εξάρτησης διασταύρωσης ή / και γραμμές προβολής.

που αναφέρεται το «μήλο» διαχωρίζεται με έναν χρονικό τροποποιητή του κύριου ρήματος. Δεν υπάρχει τρόπος να σχεδιαστεί το δέντρο εξάρτησης για αυτήν την πρόταση χωρίς να διασταυρωθούν τα άκρα, αυτό απεικονίζεται στο σχήμα 6. Αν και αυτή η διάκριση είναι πολύ σημαντική για την έρευνα στην ανάλυση των προτάσεων, καθώς οι αλγόριθμοι αλλάζουν με βάση το είδος του δέντρου (McDonald, Cramer, και Pereira, 2005), δεν λαμβάνεται ιδιαίτερα υπ' όψιν στη Μηχανική Μετάφραση.

4.4. Εξάρτηση βασισμένη σε μοντέλα μετάφρασης

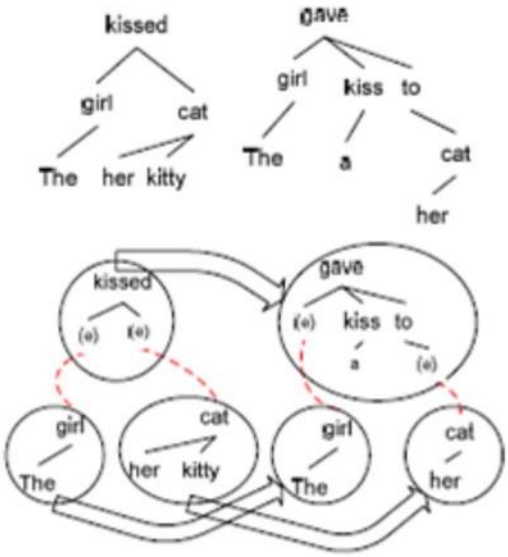
Η Μηχανική Μετάφραση μεταξύ μιας γλώσσας, που πλειοψηφεί από την πλευρά της πηγής και μιας γλώσσας που μειονεκτεί από την πλευρά του στόχου ή το αντίστροφο, είναι τα πιο κοινά σενάρια. Σε τέτοιες περιπτώσεις όπου έχουμε μόνο ένα δέντρο ανάλυσης για τη μία πλευρά του ζεύγους γλωσσών απαιτείται να χρησιμοποιηθεί κάποιου είδους τεχνικής προβολής βασιζόμενη στην «διαίσθηση» της γλώσσας για να πραγματοποιηθεί η μεταφορά του νοήματος.

Έχουμε ως δεδομένα δύο παράλληλα corpus (ένα της γλώσσας-πηγής και ένα της γλώσσας-στόχου), η τεχνική προβολής εφαρμόζεται με τον ακόλουθο τρόπο για να δημιουργηθεί δομή στις προτάσεις. Το κείμενο της γλώσσας-πηγής επισημαίνεται/αναλύεται είτε με το χέρι είτε μέσω μιας αυτοματοποιημένης ανάλυσης εξάρτησης. Οι αντιστοιχίες μεταξύ των λέξεων, στις παράλληλες προτάσεις του κειμένου προσδιορίζονται εκ των προτέρων, μέσω της καλύτερης πιθανής ευθυγράμμισης viterbi ή μέσω των n-καλύτερων ευθυγραμμίσεων. Οι πληροφορίες που προκύπτουν από αυτή την διαδικασία χρησιμοποιούνται στη συνέχεια μαζί με ορισμένες τεχνικές ευρετικής για να γίνει η επισήμανση. Η ποιότητα των αντιστοιχιών, καθορίζει την ακρίβεια της μεταφοράς/μετάφρασης. Δεν υπάρχουν δύο γλώσσες που να είναι εντελώς πανομοιότυπες στη σύνταξη. Είναι βέβαιο ότι θα υπάρξουν αποκλίσεις και ως εκ τούτου η μεταφορά των επισημάνσεων/αναλύσεων δεν μπορεί να είναι εγγυημένη για την παραγωγή μιας απολύτως σωστής σύνταξης από την πλευρά της γλώσσας-στόχου. Αυτό απαιτεί κάποιου είδους περεταίρω επεξεργασίας ανάλογα με τον τύπο της επισήμανσης και το είδος της γλώσσας.

Η υπόθεση της άμεσης αντιστοιχίας (DCA) εισάγεται για πρώτη φορά στο (Hwa et al., 2005). Χρησιμοποιείται αρχικά για την προβολή σχέσεων εξάρτησης. Η υπόθεση αναφέρει ότι ένα συγκεκριμένο είδος σχέσεων εξάρτησης διατηρείται υπό προϋποθέσεις μέσα στην άμεση προβολή. Οι υποθέσεις άμεσης αντιστοιχίας συνήθως προέρχονται από εμπειρικές μελέτες, πάνω σε δύο διαφορετικά γλωσσικά corpus (Fox, 2002). Αυτές οι μελέτες αποτελούν τη βάση πραγμάτευσης των περισσότερων προβλημάτων της σύνταξης προβολής. Ωστόσο, η υπόθεση της άμεσης αντιστοιχίας τείνει,

επίσης, να δημιουργεί πολύ θορυβώδεις[10] επισημάνσεις για τη γλώσσα-στόχο, επειδή είναι πολύ απλή και ντετερμινιστική, δεδομένης της πολυπλοκότητας των πραγματικών γλωσσών. Λόγω αυτού, χρησιμοποιούνται συνήθως πιθανά θεωρητικά μοντέλα, αντί της υπόθεσης της άμεσης αντιστοιχίας για την στιβαρότητα της προβολής (Smith και Eisner, 2006).

Οι τεχνικές προβολής που χρησιμοποιούν την προσέγγιση της άμεσης αντιστοιχίας, έχουν ως αποτέλεσμα τη δημιουργία σύνταξης για τη γλώσσα-στόχο, η οποία είναι ισόμορφη της δομής του δέντρου της γλώσσας προέλευσης. Τα μεταφραστικά μοντέλα που κατασκευάζονται από τέτοια corpus ονομάζονται ισομορφικά μοντέλα μετάφρασης. Αυτός ο ισομορφισμός, μπορεί να μην υπάρχει πάντα μεταξύ δύο γλωσσών. Ακόμη και αν υπάρχει μεγάλος βαθμός ισομορφισμού, είναι συχνά δύσκολο να παρατηρηθεί στα υπό εξέταση corpus-κείμενα, λόγω της ελεύθερης φύσης της μετάφρασης ή του θορύβου στην ποιότητα της μετάφρασης.



Σχήμα 7: μη ισομορφισμός μεταξύ δύο προτάσεων (Ambati,2005).

Για παράδειγμα, η κατασκευή στο σχήμα 7 δείχνει τον μη ισομορφισμό μεταξύ δύο προτάσεων. Τα μοντέλα μετάφρασης που βασίζονται στην εκμάθηση σε ένα τέτοιο σενάριο παρουσιάζουν προβλήματα, καθώς ο εντοπισμός της ευθυγράμμισης μεταξύ

[10] Η έννοια του «θορύβου» σε όλη την έκταση της εργασίας είναι παρμένη από την θεωρία της πληροφορίας και αναφέρεται σε πιθανές αλλοιώσεις που προκύπτουν κατά την μεταφορά ενός μηνύματος.

των μονάδων ενός υποδέντρου είναι εξαιρετικά δύσκολος. Το παράδειγμα του σχήματος 7, κατατάσσεται στα «μοντέλα μετάφρασης που βασίζονται στον μη ισομορφισμό».

4.5. Ισομορφικά μοντέλα μετάφρασης

Σε αυτή την ενότητα παρουσιάζεται η έρευνα για την ενσωμάτωση των δέντρων εξάρτησης στο έργο της μετάφρασης, σε γλωσσικά corpus όπου ο ισομορφισμός θεωρείται δεδομένος.

4.5.1 Μετάφραση βασισμένη στην ελάχιστη κάλυψη δέντρων

Ο (Lin 2004) έχει επεξεργαστεί ένα μοντέλο μεταφοράς βασισμένο σε διαδρομές[11] για τη Μηχανική Μετάφραση. Το μοντέλο αναπτύσσεται/μαθαίνει από την ευθυγράμμιση λέξεων μεταξύ παράλληλων corpus κειμένων, όπου η πλευρά της πηγής αποτελείται από δέντρα εξάρτησης. Ο αλγόριθμος εξάγει ένα σύνολο διαδρομών στα δέντρα εξάρτησης της προέλευσης και καθορίζει τις αντίστοιχες μεταφράσεις των διαδρομών χρησιμοποιώντας ευθυγραμμίσεις λέξεων. Το αποτέλεσμα της εκμάθησης είναι ένα σύνολο κανόνων μετάφρασης που δεδομένης μιας συγκεκριμένης διαδρομής της πηγής, παρέχουν το ισοδύναμο τμήμα μετάφρασης στον στόχο. Οι κανόνες εξάγονται όταν όλες οι λέξεις στη διαδρομή έχουν συνδέσεις μετάφρασης. Επίσης, για τις προθέσεις επιτρέπεται να μην είναι ευθυγραμμισμένες. Κάθε κανόνας όχι μόνο κωδικοποιεί τις σχέσεις εξάρτησης για την πλευρά προορισμού (γλώσσα-στόχου), αλλά και τη γραμμική σειρά μεταξύ των κόμβων, αυτός είναι ο ρόλος της διαδρομής. Ο αλγόριθμος εξάγει επίσης δύο είδη διαδρομών για τον κόμβο στο δέντρο προέλευσης. Μια «κύρια διαδρομή» που είναι η ακολουθία λέξεων που ευθυγραμμίζεται με τον κόμβο, και τη «διαδρομή φράσης» που είναι το μέγιστο εύρος διαδρομών όλων των υποδέντρων κάτω από τον κόμβο. Αυτές οι διαδρομές, χρησιμοποιούνται κατά την εξαγωγή κανόνων για την αποφυγή σφαλμάτων. Προκειμένου να γενικευθεί το μεταφραστικό μοντέλο, μερικοί από τους κανόνες που εξάγονται από τις μεταφρασμένες προτάσεις γενικεύονται. Η γενίκευση είναι πολύ περιορισμένη ώστε να μην είναι τεράστιος ο αριθμός των κανόνων που εξάγονται. Επί του παρόντος, είναι μόνο οι τελικοί κόμβοι κάθε τμήματος δέντρου που γενικεύονται. Για να εκχωρηθούν πιθανότητες σε κάθε έναν από τους κανόνες που εξάγονται, υπολογίζεται η $\boldsymbol{P(T_i/S_i)}$ με μια σταθερά εξομάλυνσης για τη μείωση του θορύβου.

[11] Το μοντέλο μεταφοράς βασισμένο σε διαδρομές αποδίδει το «path based transfer model».

Η μετάφραση με ένα μοντέλο που εξάγεται από την παραπάνω διαδικασία, παίρνει μια δεδομένη πρόταση προέλευσης και την αναλύει για να παράξει ένα σύνολο διαδρομών από το δέντρο εξάρτησης της. Στη συνέχεια, εντοπίζει ένα σύνολο κανόνων μεταφοράς/μετάφρασης που «καλύπτουν» ολόκληρο το δέντρο εξάρτησης και παράγουν ένα σύνολο από μέρη του δέντρου, στη γλώσσα στόχου. Ένα σύνολο διαδρομών λέγεται ότι καλύπτει ένα δέντρο εξάρτησης, εάν το σύνολο των κόμβων και οι συνδέσεις τους σε αυτό το σύνολο διαδρομών περιλαμβάνει όλους τους κόμβους και τις συνδέσεις του δέντρου. Η μετάφραση διαβάζεται από αυτό το δέντρο προορισμού. Κύρια πρόκληση εδώ είναι, να είναι δυνατό, να συγχωνευτούν τα θραύσματα των δέντρων που λαμβάνονται, για διαφορετικές διαδρομές σε ένα ενιαίο δέντρο που να έχει την υψηλότερη πιθανότητα. Τα θραύσματα των δέντρων συνδυάζονται για να σχηματίσουν ένα δέντρο $\boldsymbol{T} \ast= \boldsymbol{argmaxP}(\boldsymbol{T_i}/\boldsymbol{S_i})$. Η συγχώνευση γίνεται συνήθως στους κόμβους προορισμού που ευθυγραμμίζονται με τον ίδιο κόμβο προέλευσης και δεν εισάγουν μια λούπα στο δέντρο προορισμού. Όσον αφορά την ταξινόμηση των λέξεων στα θραύσματα των δέντρων σε περίπτωση που οι «κανόνες μεταφοράς» που έρχονται με θραύσματα δέντρων είναι μοναδικά ή από το ίδιο δέντρο δεν υπάρχει πρόβλημα, αλλά αν είναι από διαφορετικά παραδείγματα δέντρων, τότε χρησιμοποιούνται κάποιες σχετικές εκτιμήσεις εγγύτητας.

4.6 Μη ισομορφικά μοντέλα μετάφρασης

Σε αυτή την ενότητα εξετάζονται προσεγγίσεις, όπου τα μοντέλα μετάφρασης εξάρτησης, κατασκευάζονται σε ζεύγη γλωσσών, όπου αμφότερες έχουν κατασκευασμένα δέντρα εξάρτησης. Εξ' αυτού, η υπόθεση του ισομορφισμού δεν ισχύει, αλλά ταυτόχρονα η απουσία της δυσκολεύει τον εντοπισμό των αντιστοιχιών μεταξύ των κόμβων των δέντρων για τις δύο γλώσσες, με σκοπό τη δημιουργία σχετικών μοντέλων μετάφρασης.

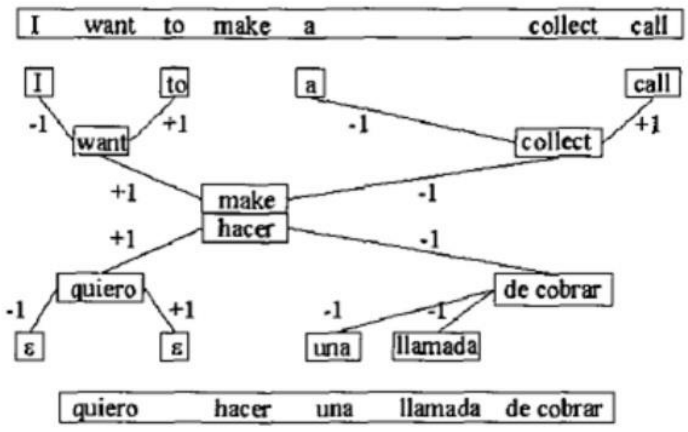
4.6.1 Μετάφραση βασισμένη σε Κύριους μετατροπείς: Alshwai

Οι (Alshawi, Douglas, και Bangalore, 2000) προτείνουν ένα μεταφραστικό μοντέλο εξάρτησης μέσω ενός συνόλου σταθμισμένων κύριων μετατροπέων. Ένας «κύριος μετατροπέας», σε αντίθεση με ένα πεπερασμένο μετατροπέα[12] που καταναλώνει

[12] Ένας «Μετατροπέας Πεπερασμένης Κατάστασης» («FST») είναι μια μηχανή πεπερασμένης κατάστασης με δύο ταινίες μνήμης, ακολουθώντας την ορολογία για τις μηχανές Turing: μια ταινία εισόδου και μια ταινία εξόδου. Ο FST εκτελεί μια αυτοματοποιημένη λειτουργία πεπερασμένης κατάστασης, που αντιστοιχίζει δύο σύνολα συμβόλων. Ένας FST θα διαβάσει ένα σύνολο συμβολοσειρών στην ταινία εισόδου και θα δημιουργήσει ένα σύνολο σχέσεων στην ταινία εξόδου.

εισροές από αριστερά προς τα δεξιά, καταναλώνει εισροές «από τη μέση προς τα έξω». Επομένως, ένας κύριος μετατροπέας εκτός από το γεγονός ότι προσφέρει τα ίδια με έναν κανονικό μετατροπέα πεπερασμένης κατάστασης, περιέχει πληροφορίες θέσης, για τις συμβολοσειρές που εισήχθησαν και πληροφορίες θέσης για τις σχέσεις εξόδου.

Όταν οι κύριοι μετατροπείς εφαρμόζονται στο μεταφραστικό έργο, τους αποκαλούμε "μοντέλα μετατροπής εξάρτησης". Κάθε μία από τις συμβολοσειρές προέλευσης με μια κύρια λέξη και αριστερά και δεξιά εξαρτώμενες λέξεις, μπορεί να χρησιμοποιηθεί από ένα μετατροπέα για την παραγωγή των αντίστοιχων κύριων λέξεων και των λέξεων αριστερά και δεξιά της αντίστοιχης λέξης στη γλώσσα-στόχο. Μια συλλογή τέτοιων μετατροπέων αποσυνθέτει αναδρομικά τις συμβολοσειρές προέλευσης και προορισμού, για να εξηγήσει τη δομή εξάρτησης μεταξύ των δύο γλωσσών. Το μοντέλο παράγει συγχρονισμένα δέντρα εξάρτησης, όπου κάθε τοπικό δέντρο παράγεται από ένα συγκεκριμένο μετατροπέα. Στο σχήμα 8, κάθε ζεύγος δέντρων πηγής και στόχου παράγεται με αυτόν τον τρόπο. Το κόστος μιας παραγωγής σε ένα τέτοιο πλαίσιο είναι το άθροισμα του κόστους του καθενός από τους μετατροπείς, με αυτό τον τρόπο μπορεί να διαλεχτεί το χαμηλότερο δέντρο κόστους, που μπορεί να κατασκευαστεί χρησιμοποιώντας το μοντέλο εξάρτησης. Η εκμάθηση του κόστους ή των βαρών, για κάθε έναν από τους μεμονωμένους μετατροπείς γίνεται σε ένα μη επιμερισμένο σώμα συμβολοσειρών της πηγής και του στόχου. Η εκπαιδευτική προσέγγιση, υπολογίζει πρώτα στατιστικά στοιχεία συνεμφάνισης από τις προτάσεις και αναζητά μια βέλτιστη ιεραρχική στοίχιση χρησιμοποιώντας τα. Η ιεραρχική στοίχιση εκτελείται χρησιμοποιώντας έναν αλγόριθμο δυναμικού προγραμματισμού που βελτιστοποιεί μια συνάρτηση κόστους που περιλαμβάνει, τις πιθανότητες μετάφρασης λέξεων όπως δίνονται από τις συνεμφανίσεις, καθώς και τις σχετικές αποστάσεις μεταξύ της κύριας λέξης και των εξαρτώμενων, τόσο στην προέλευση, όσο και στον προορισμό. Η ευθυγράμμιση αυτή χρησιμοποιείται για την κατασκευή κύριων μετατροπέων που μπορούν να εξηγήσουν τις προτάσεις, με μια τεχνική εκτίμησης μέγιστης πιθανότητας.



Σχήμα 8: Κύριοι μετατροπείς για μοντέλα μετάφρασης εξάρτησης (Ambati,2005).

Ένα μειονέκτημα της προσέγγισης είναι ότι ο αλγόριθμος δεν μαθαίνει απαραίτητα δομές εξάρτησης που έχουν γλωσσικά κίνητρα, αλλά μάλλον εκείνες που προσπαθούν να εξηγήσουν το σύγχρονο φαινόμενο μεταξύ των δύο γλωσσών. Ωστόσο, οι συγγραφείς παρατηρούν ότι στις περισσότερες περιπτώσεις συμβαδίζει με μεμονωμένα δέντρα εξαρτημένης δομής και για τις δύο γλώσσες. Επίσης, δεδομένου ότι ήταν ένα πολύ πρώιμο έργο οι συγγραφείς δεν το συγκρίνουν με τις παραδοσιακές προσεγγίσεις όπως τα IBM Models. Η προσέγγιση επίσης δεν κλιμακώνεται για μεγαλύτερες προτάσεις, και έτσι εφαρμόζουν το σύστημά τους σε μικρές προτάσεις με λιγότερο από 20 λέξεις σε μήκος.

5. Σημασιολογικά Δέντρα Εξάρτησης

Θα εξεταστεί η δυνατότητα δημιουργίας ενός μη καθορισμένου φορμαλισμού, που ονομάζεται «Σημασιολογικό Δέντρο Εξάρτησης», βασιζόμενου σε συναρτησιακές εξαρτήσεις τύπου Skolem μεταξύ των εμπλεκόμενων οντοτήτων/μονάδων που συναποτελούν την πρόταση. Οι μη καθορισμένες δομές στα σημασιολογικά δέντρα εξαρτήσεων βασίζονται σε ένα απλό γράφημα *G* που αντιπροσωπεύει τις σχέσεις κατηγορήματος-ορίσματος[13] που εμφανίζονται στην πρόταση. Οι κόμβοι του ***G*** αναφέρονται είτε σε κατηγορήματα είτε σε μεταβλητές, οι μεταβλητές καλούνται με την σειρά τους «αναφερόμενα λόγου». Κάθε τόξο συνδέει ένα κατηγόρημα με ένα αναφερόμενο λόγου και επισημαίνεται/ονοματίζεται από την θέση του ορίσματος. Κάθε αναφερόμενο λόγου συσχετίζεται επίσης με έναν γενικευμένο ποσοδείκτη[14], σύμφωνα με μια συνάρτηση quant, στον συσχετισμό υπάρχουν περιορισμοί, σύμφωνα με μια συνάρτηση περιορισμού restr.

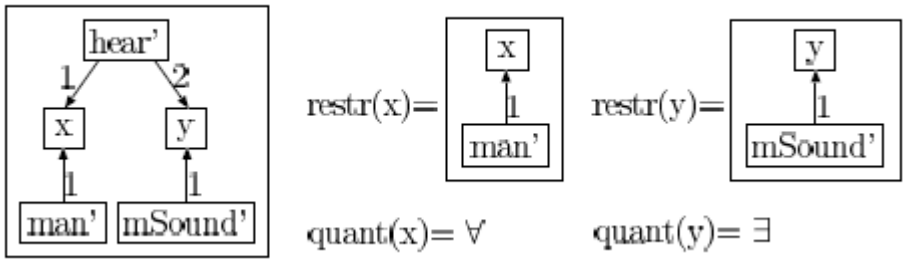
Ενώ η τιμή του quant είναι σαφής, η τιμή του restr περιλαμβάνει τη συντακτική δομή της πρότασης και κατά συνέπεια, τη χαρτογράφηση σύνταξης-σημασιολογίας. Διαισθητικά, η τιμή restr(d), όπου το ***d*** είναι ένα αναφερόμενο λόγου, είναι το υπογράφημα που περιλαμβάνει όλους τους κόμβους που συμβάλλουν στην αναγνώριση των στοιχείων που αναφέρονται. Να σημειωθεί ότι από τυπική πλευρά, το restr(d) είναι απλώς ένα υπογράφημα του G.

Ως ένα πρώτο απλό παράδειγμα, παρουσιάζεται στο σχήμα 9 η πλήρως μη καθορισμένη αναπαράσταση σημασιολογικών δέντρων εξάρτησης για την πρόταση

[13] Στην γλωσσολογία ως «κατηγόρημα» αναφέρεται το μέρος της πρότασης που προσδιορίζει υποκείμενο ή απλή μορφή της ρηματικής φράσης. Το «όρισμα» είναι μια έκφραση που συμπληρώνει τη σημασία του κατηγορήματος.

[14] Οι «Γενικευμένοι Ποσοδείκτες» («GQ») είναι ευρέως αποδεκτοί ως βασικά λογικά στοιχεία που χρειάζονται για να αντιπροσωπεύουν σωστά τη σημασιολογία των προτάσεων των φυσικών γλωσσών.

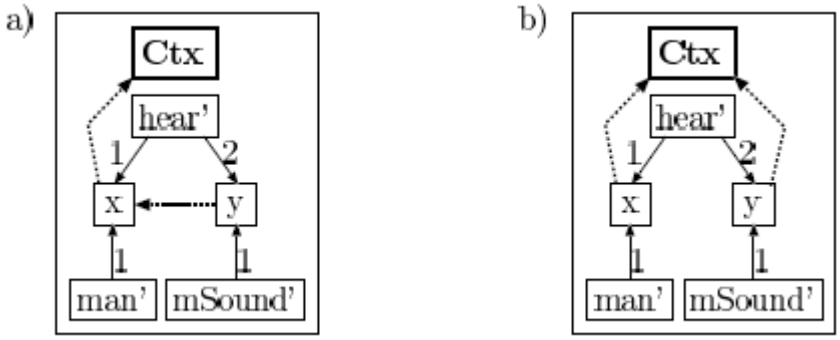
«Every man heard a mysterious sound». Στο σχήμα απεικονίζονται οι σχέσεις κατηγορήματος-ορίσματος της πρότασης, οι εμπλεκόμενοι ποσοδείκτες και οι σχέσεις κατηγορήματος-ορίσματος που απαιτούνται για τον προσδιορισμό του συνόλου των στοιχείων στα οποία κυμαίνονται τα ***x*** και ***y*** αντίστοιχα.



Σχήμα 9: Σημασιολογικό δέντρο εξάρτησης μη καθορισμένης δομής για την πρόταση: Every man heard a mysterious sound. (Robaldo & Lesmo,2006).

Για την αποσαφήνιση του σχήματος, θα πρέπει να καθοριστούν οι συναρτησιακές εξαρτήσεις μεταξύ των εμπλεκόμενων στοιχείων. Αυτό γίνεται με την εισαγωγή πρόσθετων (διακεκομμένων) τόξων, που ονομάζονται τόξα ***SemDep***, μεταξύ των αναφερόμενων λόγου. Επιπλέον, για να δηλωθεί ότι ένα συγκεκριμένο σύνολο στοιχείων δεν εξαρτάται από τα στοιχεία οποιουδήποτε άλλου συνόλου, το αντίστοιχο αναφερόμενο λόγου συνδέεται, μέσω ενός ***SemDep*** τόξου, με έναν πρόσθετο κόμβο που ονομάζεται **Ctx**. Το **Ctx** αναφέρεται στο πλαίσιο, δηλαδή στο σύνολο των στοιχείων βάσει του οποίου θα αξιολογηθεί η εξάρτηση.

Για παράδειγμα, στο σχήμα 10(a), στα αριστερά, ένα τόξο ***SemDep*** συνδέει ***y*** με **x:** αυτή είναι η ερμηνεία στην οποία, κάθε άτομο εκ του συνόλου όλων των ανδρών έχει ακούσει έναν δυνητικά διαφορετικό ήχο. Αντίθετα, στο σχήμα 10(b), το y συνδέεται με το ***Ctx***. Αυτό σημαίνει ότι ο ήχος πρέπει να είναι ο ίδιος για όλους τους άντρες.



Σχήμα 10: Σημασιολογικά δέντρα εξάρτησης με πλήρως αποσαφηνισμένες δομές για την πρόταση «Every man heard a mysterious sound» (*Robaldo & Lesmo,2006).*

Είναι σημαντικό να σημειωθεί ότι τα τόξα ***SemDep*** αντιπροσωπεύουν μια μεταβατική σχέση και ότι είναι ***ελάχιστα.*** Αυτό σημαίνει ότι:

- Κανένα τόξο δεν εμφανίζεται σε ένα σχήμα που μπορεί να προκύψει λόγω της μεταβατικότητας της σχέσης (έτσι η σημασιολογική εξάρτηση του ***y*** στο ***Ctx*** στο (2.a) δεν αντιπροσωπεύεται ρητά μέσω ενός τόξου).
- Εάν δύο κόμβοι δεν είναι συνδεδεμένοι (είτε άμεσα είτε εξαιτίας της διέλευσης), τότε το ζεύγος αυτών των κόμβων δεν ανήκει στη σχέση. Έτσι, στο σχ. (2.b), δεδομένου ότι το ***x*** δεν είναι συνδεδεμένο στο ***y*** μέσω τόξου ***SemDep*** (και δεν μπορεί να συναχθεί τέτοια σύνδεση μέσω της μεταβατικότητας), θεωρείται ότι τα ***x*** και ***y*** είναι ανεξάρτητα το ένα από το άλλο.

5.1 Κατασκευή σημασιολογικών δέντρων εξάρτησης

Όπως υποδηλώνει το όνομα «σημασιολογικά δέντρα εξάρτησης», οι καλά σχηματισμένες (πλήρως μη καθορισμένες) δομές των σημασιολογικών δέντρων κατασκευάζονται ξεκινώντας από τα δέντρα εξάρτησης που προκύπτουν από τις προτάσεις. Θα εξεταστεί παρακάτω, πώς μπορεί να δημιουργηθεί η διασύνδεση σύνταξης-σημασιολογίας.

5.1.1 Γραμματικές εξαρτήσεις

Η «Γραμματική Εξάρτησης» είναι μια φορμαλιστική προσέγγιση, που επιτρέπει την περιγραφή της σύνταξης των φυσικών γλωσσών, από την άποψη των προσανατολισμένων σχέσεων μεταξύ λέξεων, που ονομάζονται «Σχέσεις Εξάρτησης» ή «Γραμματικές Συναρτήσεις». Συγκεκριμένα, μια ανάλυση γραμματικής εξάρτησης θα πάρει μια πρόταση φυσικής γλώσσας και θα την αναλύσει σε μια ιεραρχική δομή λέξεων που συνδέονται μέσω σχέσεων εξάρτησης.

Όπως προαναφέρθηκε, οι σχέσεις εξάρτησης είναι προσανατολισμένες**.** Επομένως, για κάθε ζεύγος συνδεδεμένων λέξεων, μπορούμε να προσδιορίσουμε μια κύρια (την προέλευση του συνδέσμου) και μια εξαρτώμενη (ο προορισμός του συνδέσμου). Η εξαρτημένη έχει τον ρόλο της «ολοκλήρωσης» της κύριας, δηλαδή να παρέχει μια παραμετροποίηση. Για το λόγο αυτό, στην γραμματική εξάρτησης οι σχέσεις εκφράζονται με όρους ισχύος. Οι σχέσεις εξάρτησης συνήθως επισημαίνονται για να καταστήσουν σαφή τη λειτουργία που παίζει μια εξαρτώμενη λέξη σε σχέση με την κεφαλή. Αν και οι χρησιμοποιούμενες ετικέτες ποικίλλουν από θεωρία σε θεωρία, πολλές γραμματικές λειτουργίες, όπως το υποκείμενο, το αντικείμενο κ.λπ., είναι κοινώς

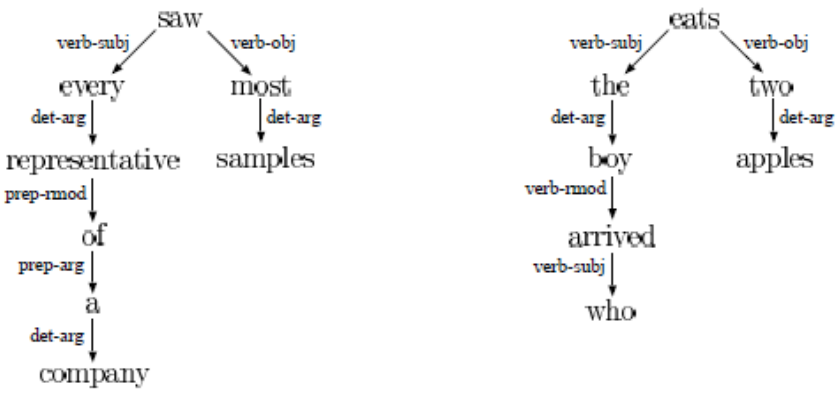
αποδεκτές στη βιβλιογραφία των γραμματικών εξάρτησης. Επίσης, αξίζει να επισημανθεί ότι, σε μια σχέση εξάρτησης, όχι μόνο η άμεσα εξαρτώμενη λέξη, αλλά ολόκληρο το υποδένδρο που είναι εξαρτώμενο από την ρίζα μπορεί να συμβάλει στην «ολοκλήρωση» του νοήματος της κύριας.

Οι σύγχρονες γραμματικές εξάρτησης αποδίδονται στο έργο του Tesniere (Tensiere, 1959). Η δομική σύνταξη του Tensiere αναλύει τις προτάσεις υπό τους όρους της ***λειτουργίας*** που κατέχουν οι λέξεις, που προκαλούν σχέσεις υπόταξης που ονομάζονται ***συνδέσεις.*** Η δομική σύνταξη έχει κωδικοποιηθεί σε ένα υπολογιστικό μοντέλο, το Functional Dependency Grammar (Tapanainen, 1999). Άλλα σημαντικά έργα που ανήκουν στην προσέγγιση της γραμματικής εξάρτησης είναι η Γραμματική του Hudson (Hudson, 1990), μια Γραμματική Εξάρτησης δομημένη σε μια ταξινόμηση «is-a» που επιτρέπει της γλωσσικές γενικεύσεις και επιτρέπει σε κάποιον να ασχολείται με τις προεπιλεγμένες περιπτώσεις-εξαιρέσεις και στην θεωρία νοήματος του κειμένου (Melcuk, 1988, Kahane, 2003).

Μία εκ των πιο πρόσφατων προτάσεων στο τομέα της γραμματικής εξάρτησης είναι η «Εκτατή Γραμματική Εξάρτησης» («Extensible Dependency Grammar») (Debusmann, 2006), ένα δομοστοιχειωτό (modular) πλαίσιο που είναι σε θέση να χειριστεί τις διάφορες πτυχές της φυσικής γλώσσας, κάνοντας χρήση της Θεωρίας νοήματος-κειμένου. Ωστόσο σε αντίθεση με την Θεωρία νοήματος κειμένου, στην εκτατή γραμματική εξάρτησης τα διάφορα επίπεδα περιγραφής, που ονομάζονται διαστάσεις, είναι πιο ολοκληρωμένα. Συγκεκριμένα, ενώ στη Θεωρία νοήματος-κειμένου τα επίπεδα περιγραφής είναι αυστηρά διαδοχικά, δηλαδή οι κανόνες διασύνδεσης ενός επιπέδου με ένα άλλο ορίζονται μόνο για ***παρακείμενα*** επίπεδα, στη εκτατή γραμματική εξάρτησης αυτοί οι κανόνες μπορούν να οριστούν για κάθε ζεύγος διαστάσεων ξεχωριστά.

Οι δομές εξάρτησης, που αναφέραμε σχετικά με την έρευνα για την σημασιολογική γραμματική εξάρτησης χρησιμοποιούνται στο Πανεπιστήμιο του Τορίνο σε διάφορα projects, συμπεριλαμβανομένης της ανάπτυξης μιας τράπεζας δέντρων εξάρτησης για τα ιταλικά (Bosco, 2004). Τα τόξα εξάρτησης επισημαίνονται, σύμφωνα με μια διαδικασία που κωδικοποιεί τις σχέσεις μεταξύ των λέξεων. Το σχήμα βασίζεται σε μια διπλή διάκριση μεταξύ λειτουργικών και μη λειτουργικών εξαρτημένων. Τα μη λειτουργικά εξαρτημένα είναι εξαρτημένα που δεν έχουν τομεακή σημασιολογική εισαγωγή και ταξινομούνται περαιτέρω σε *Aux* (βοηθητικά), *Contin* (συνέχεια, σε ιδιωματισμούς), *Coordinator* (τόξα που σχετίζονται με συνδέσεις), *Separator* (εντάσσονται τα περισσότερα σημεία στίξης), *Visitors* (π.χ. αύξηση δομές), *Interjections* και ορισμένα μόρια χωρίς σημασιολογικό περιεχόμενο. Η κλάση των λειτουργικών εξαρτημένων χωρίζεται σε Επιχειρήματα-*Arguments* και Τροποποιητές-*Modifiers*, που αντιστοιχούν στην τυπική διάκριση μεταξύ των συντακτικά ενεργών και των συντακτικά περιστατικών (Mel'cuk, 2004). Τέλος, οι *Modifiers* μπορεί να

είναι *Rmod* (περιοριστικοί τροποποιητές) ή παραθέσεις-*Appositions* . Αξίζει να σημειωθεί ότι οι ορίζουσες (και οι ποσοδείκτες) είναι οι ρίζες των ονομαστικών υποδέντρων.



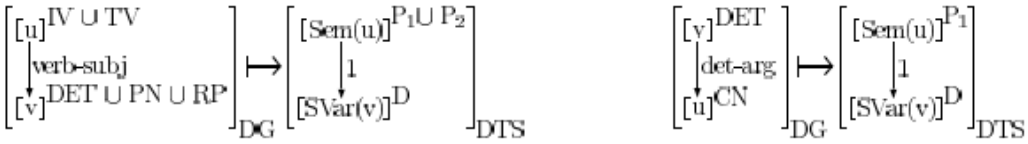
Σχήμα 11: Δέντρα συντακτικής εξάρτησης που σχετίζονται με τις προτάσεις: «Every representative of a company saw most samples» και «The boy who arrived eats two apples».(Robaldo & Lesmo,2006).

5.2 Διασύνδεση σημασιολογικών δέντρων εξάρτησης

Η διασύνδεση (interface) μεταξύ δέντρων εξάρτησης και σημασιολογικών δέντρων εξάρτησης επιτρέπει τη συσχέτιση ενός δέντρου εξάρτησης με ένα πλήρως μη-καθορισμένο καλά σχηματισμένο σημασιολογικό δέντρο εξάρτησης. Είναι σημαντικό να υπενθυμίσουμε ότι αυτή η διασύνδεση είναι συνθετική, δηλαδή ένα δέντρο εξάρτησης συνδέεται μόνο με ένα καλά σχηματισμένο σημασιολογικό δέντρο. Η διεπαφή του δέντρου εξάρτησης με το σημασιολογικό δέντρο εξάρτησης είναι μια γενίκευση της εκτατής γραμματικής εξάρτησης. Η κύρια διαφορά μεταξύ των δύο δέντρου εξάρτησης και σημασιολογικού δέντρου εξάρτησης είναι ότι το πρώτο συσχετίζει δομές με διαφορετικά σύνολα κόμβων ενώ, στο δεύτερο οι διάφορες διαστάσεις μοιράζονται το ίδιο σύνολο κόμβων. Αυτή η επιλογή έχει γίνει για να διαχωριστούν οι κόμβοι στα δέντρα εξάρτησης (λέξεις όπως saw, every, of, κ.λπ.) από τους κόμβους στο γράφημα του σημασιολογικού δέντρου εξάρτησης (κατηγορήματα όπως see, of', John (x), κλπ. ή αναφερόμενα λόγου όπως x, y, z, κ.λπ.). Η χαρτογράφηση από λέξεις στα δέντρα εξάρτησης σε κόμβους στα σημασιολογικά δέντρα εξάρτησης υλοποιείται μέσω των συναρτήσεων ***Sem*** και ***SVar*** ***(οι Lex*** και ***LVar*** υλοποιούν την αντίστροφη χαρτογράφηση).

To Sem σχετίζεται ρήματα, κοινά ουσιαστικά και άλλες λέξεις περιεχομένου στο δέντρο εξάρτησης με ένα κατηγόρημα, ενώ το ***SVar*** συσχετίζει ονόματα και σχετικές αντωνυμίες στο δέντρο εξάρτησης με μια αναφορά λόγου.

Παρακάτω θα εξετασθεί ένα παράδειγμα, μια πολύ απλή περίπτωση γραμματικής εξάρτησης. Το παράδειγμα περιλαμβάνει λέξεις του αγγλικού λεξιλογίου που ταξινομούνται σε επτά μέρη του λόγου: ***IV*** (αμετάβατα ρήματα), ***TV*** (μεταβατικά ρήματα), ***PN*** (κατάλληλα ονόματα), ***CN*** (κοινά ουσιαστικά), ***PREP*** (προθέσεις), ***DET*** (ουσιαστικά), ***PREP*** (προθέσεις), ***DET*** (ορίζουσες), και ***RP*** (σχετικές αντωνυμίες). Οι γραμματικές λειτουργίες είναι υποκείμενο (verb-subj), το αντικείμενο (verb-obj), όρισμα προθέσεως (prep-arg), όρισμα ορίζουσας (det-arg), ο προθετικός τροποποιητής (prep-rmod) και ο λεκτικός τροποποιητής (ρήμα-rmod). Με τα άνωθεν εργαλεία θα εφαρμοστεί η μετάβαση από το δέντρο εξάρτησης στο σημασιολογικό δέντρο εξάρτησης, σε αυτή την απλού τύπου γραμματική. Η χαρτογράφηση ορίζεται από ένα σύνολο εάν-τότε κανόνων, που ονομάζεται $\text{Link}_{\text{DS-DTS}}$. Ένα παράδειγμα τέτοιων κανόνων παρουσιάζεται στο σχήμα 12.

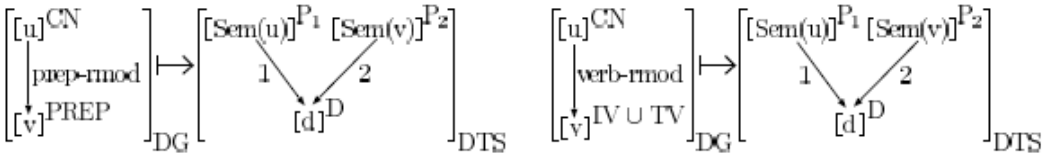


Σχήμα 12: Κανόνες $\text{Link}_{\text{DG-DTS}}$ για ρήματα υποκειμένων και ορίσματα οριζουσών. (Robaldo & Lesmo,2006).

Οι δύο κανόνες προσδιορίζουν πως τα υποκείμενα των ρημάτων και τα ορίσματα των οριζουσών ερμηνεύονται: ο κανόνας στα αριστερά δείχνει ότι το υποκείμενο ενός της αμετάβατου ρήματος αντιστοιχεί στο πρώτο μοναδιαίο κατηγόρημα που σχετίζεται με το ρήμα από τη συνάρτηση ***Sem.*** Φαίνεται ότι τα ορίσματα αυτών των κατηγορημάτων είναι οι λόγοι αναφοράς που σχετίζονται με τα εξαρτώμενα του ρήματος μέσω της συνάρτησης **SVar.** Ανάλογοι κανόνες περιορίζουν τη σημασιολογική πραγματοποίηση του άμεσου αντικειμένου και του ορίσματος μιας προθέσεως:

Σχήμα 13: Κανόνες $\text{Link}_{\text{DG-DTS}}$ για ρήματα αντικειμένων και ορίσματα προθέσεων. (Robaldo & Lesmo,2006)

Για την αντιμετώπιση πρόσθετων, είναι απαραίτητο να εισαχθούν κανόνες με ένα επιπλέον επίπεδο πολυπλοκότητας. Οι δύο προαιρετικές σχέσεις εξάρτησης που επιτρέπονται στην απλή εκδοχή της γραμματικής που εξετάζεται, δηλαδή οι προθεσιακοί και οι λεκτικοί τροποποιητές, συνδέουν ένα κοινό ουσιαστικό (που σχετίζεται με ένα κατηγόρημα $\mathbf{P_1}$) με μια πρόταση ή ένα ρήμα (που σχετίζεται με άλλο κατηγόρημα $\mathbf{P_2}$). Τα $\mathbf{P_1}$ και $\mathbf{P_2}$ πρέπει να εφαρμόζονται στο ίδιο αναφερόμενο λόγου. Για παράδειγμα, το «ο άνθρωπος με το καπέλο», υποθέτοντας ότι το «με» σχετίζεται με το κατηγόρημα ***φορά',*** το όρισμα του ***ανθρώπου'*** (η οντότητα που είναι ένας άνθρωπος) και το πρώτο του ***φορά'*** (την οντότητα που φοράει το καπέλο) αναφέρονται στο ίδιο άτομο. Έτσι, εισάγουμε στους κανόνες μια νέα μεταβλητή d, που σηματοδοτεί αυτό το κοινό αναφερόμενο λόγου. Ωστόσο, σε αντίθεση με τις μεταβλητές v και u, στους κανόνες δεν υπάρχει περιορισμός στο αντίστοιχο του ***d*** (δηλαδή στο **LVar (d)**). Αυτοί οι κανόνες φαίνονται στο σχήμα 14.



Σχήμα 14: Κανόνες $Link_{DG\text{-}DTS}$ για προθετικούς και λεκτικούς τροποποιητές (σχετικές αντωνυμίες). (Robaldo & Lesmo,2006).

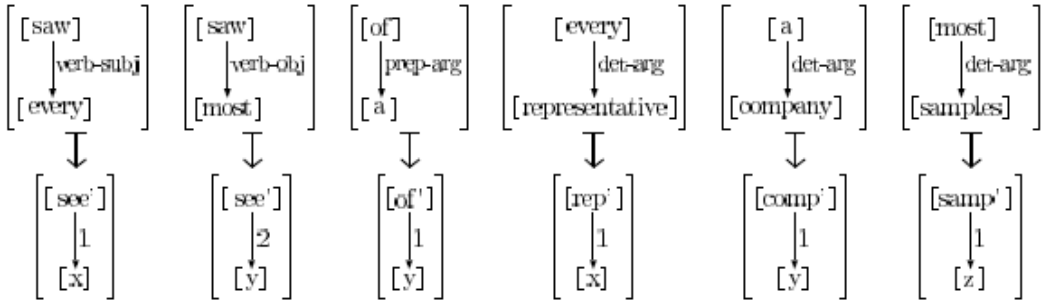
Το τελευταίο συστατικό που απαιτείται για τη δημιουργία της μη καθορισμένης αναπαράστασης του σημασιολογικού δέντρου εξάρτησης είναι ένα κριτήριο για τον ορισμό του restr(x), για κάθε αναφερόμενο λόγου x στο κύριο γράφημα. Το restr (x) είναι απλώς ρυθμισμένο στο να βάζει στο υπογράφημα όλες τις σχέσεις κατηγορήματος-ορίσματος $\boldsymbol{P}\ (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$

- Το $\boldsymbol{P}\ (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ προκύπτει από το υποδένδρο που έχει την ορίζουσα συσχετισμένη με το $\boldsymbol{x}$ ως ρίζα.
- το $\mathbf{x}$ είναι ένα από τα $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$

Αυτό φαίνεται να συμβαδίζει με τις ιδέες που βρίσκονται πίσω από την αρχιτεκτονική ενός δέντρου εξάρτησης. Στην πραγματικότητα, όπως προαναφέρθηκε, σε σχέση εξάρτησης μεταξύ κύριας και εξαρτώμενης, η τελευταία δεν περιορίζει μόνο την έννοια της κύριας, αλλά ολόκληρο το υποδένδρο που έχει το εξαρτώμενο ως ρίζα.

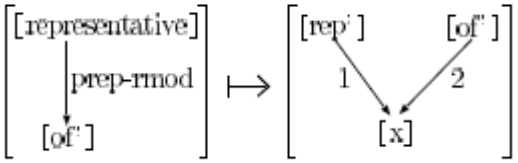
Θα δειχτεί πώς αλληλεπιδρούν οι διαφορετικοί κανόνες στη διαδικασία μεταγωγής. Αν εφαρμοστούν οι κανόνες του δέντρου εξάρτησης στην πρόταση «Every representative of a company saw most samples» που φαίνεται στο σχήμα 11, στα αριστερά. Τα περισσότερα βήματα της μεταγωγής είναι ασήμαντα, δηλαδή η εφαρμογή

των κανόνων στο σχήμα 12 στις υποδομές δέντρου εξάρτησης (κορυφή του σχ. 15), που παράγουν τις υποδομές του σημασιολογικού δέντρου εξάρτηση στο κάτω μέρος του σχήματος.



Σχήμα 15: Ορισμένες αντιστοιχίες συνδέσμων $Link_{DG\text{-}DTS}$ του δέντρο εξάρτησης της πρότασης του σχήματος 11, (Robaldo & Lesmo,2006).

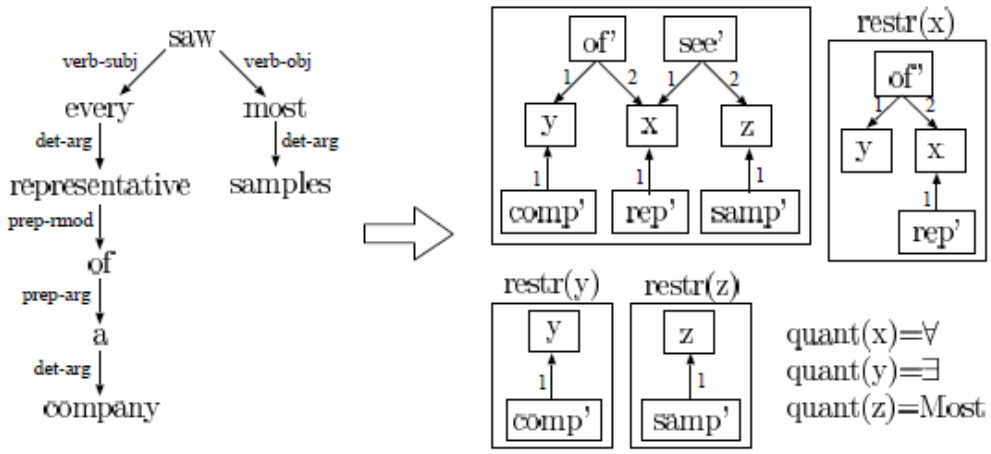
Ένα δύσκολο σημείο αφορά τον προθετικό τροποποιητή. Σε αυτήν την περίπτωση, ο εφαρμοστέος κανόνας είναι ο αριστερός στο σχήμα 14. Η μεταβλητή ***d*** που εμφανίζεται στον κανόνα ενοποιείται με το αναφερόμενο αντικείμενο ***x***. Η εφαρμογή του κανόνα ***prep-rmod*** οδηγεί στη συμπερίληψη του τόξου του ***of'*** όπως φαίνεται στο σχήμα 16.



Σχήμα 16: Αντιστοίχιση $Link_{DG\text{-}DTS}$ για την επικουρική σχέση με το δέντρο εξάρτησης της πρότασης του σχήματος 11. (Robaldo & Lesmo,2006).

Στην περίπτωση του ***prep-rmod*** (όπως για όλους τους τροποποιητές), το σημασιολογικό αποτέλεσμα είναι να προσθέσουμε επιπλέον τόξα σε ήδη υπάρχουσες αναφορές λόγου, δεδομένου ότι οι σχετικές αναφορές λόγου έπρεπε να επαναληφθούν

λόγω της παρουσίας τους σε μια συνδεδεμένη δομή (π.χ. ως συντακτικά ενεργών ρημάτων) και οι τροποποιητές τους προσθέτουν περισσότερες προδιαγραφές. Το τελικό αποτέλεσμα της εφαρμογής της συντακτικής-σημασιολογικής διασύνδεσης του δέντρου εξάρτησης με το σημασιολογικό δέντρο εξάρτησης του σχήματος 11 στα αριστερά φαίνεται στο σχήμα 17.



Σχήμα 17: Εφαρμογή διασύνδεσης του δέντρου εξάρτησης και του σημασιολογικού δέντρου εξάρτησης για την πρόταση του σχήματος 11. (Robaldo & Lesmo,2006)

Βιβλιογραφία

- Alshawi, Hiyan, Shona Douglas, and Srinivas Bangalore. 2000. Learning dependency translation models as collections of finite-state head transducers. Comput. Linguist., 26(1):45–60.
- Ambati Vamshi,2005, Dependency Structure Trees in Syntax Based Machine Translation. Available at: https://www.academia.edu/2807148/Dependency_Structure_Trees_in_Syntax_Based_Machine_Translation
- Bosco, C. 2004. A grammatical relation system for treebank annotation. Ph.D. thesis, University of Turin, Italy.
- Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19(2):263–311.
- Chakravarty Mallar, McGill University, Montreal, Quebec, Course CS-644B, Available at: http://www.bic.mni.mcgill.ca/~mallar/CS-644B / Home.html ?fbclid=IwAR03iIbn9ObQlIL-qEabVC8rugqZlheEsYf3qiFIFAjqRuN-wtuWUXanKYl4
- Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In ACL ’05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 263–270, Morristown, NJ, USA. Association for Computational Linguistics.
- D. Gildea. 2003. Loosely tree-based alignment for machine translation. In Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03), Supporo, Japan.
- D. Roth, G. K. Kao, X. Li, R. Nagarajan, V. Punyakanok, N. Rizzolo, W-T. Yih, C. Ovesdotter, and L. Moran. 2001. Learning components for a question-answering system. In Proceedings of The Tenth Text REtrieval Conference (TREC 2001), Gaithesburg, Maryland.
- Debusmann, R. 2006. Extensible Dependency Grammar: A Modular Grammar Formalism Based On Multigraph Description. Ph.D. thesis, Saarland University, Germany.
- Ding, Yuan and Martha Palmer. 2004. Synchronous dependency insertion grammars: A grammar formalism for syntax based statistical mt. In Geert-Jan M. Kruijff and Denys Duchier, editors, COLING 2004 Recent Advances in Dependency Grammar, pages 90–97, Geneva, Switzerland, August 28. COLING.
- Ding, Yuan and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In ACL ’05: Proceedings of the

43rd Annual Meeting on Association for Computational Linguistics, pages 541–548, Morristown, NJ, USA. Association for Computational Linguistics.

- E. Voorhees. 2000.Overview of the trec-9 question answering track. In The Ninth Text Retrieval Conference (TREC-9), pages 71–80. NIST SP 500-249.
- E. Voorhees. 2001.Overview of the trec 2001 question answering. In The Tenth Text Retrieval Conference (TREC 2001), pages 42–51. NIST SP 500-250.
- E. Voorhees. 2002.Overview of the trec 2002 question answering. In The Eleventh Text Retrieval Conference (TREC 2002). NIST SP 500-251.
- Fox, Heidi J. 2002. Phrasal cohesion and statistical machine translation. In EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pages 304–3111, Morristown, NJ, USA. Association for Computational Linguistics.
- G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990.Wordnet: An on-line lexical database. International Journal of Lexicography, 3(4):235–312.
- Hudson, R. 1990. English word grammar. Oxford and Cambridge: Basil Blackwell
- Hutchins, W. John and Harold L. Somers. 1992. An Introduction to Machine Translation. Academic Press.
- Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. Nat. Lang. Eng., 11(3):311–325, September
- J. Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (companion volume), Supporo, Japan.
- K. Tai. 1979. The tree-to-tree correction problem. Journal of the Association for Computing Machinery, 26(3):422–433.
- K. Zhang and D. Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. SIAM Journal on Computing, 18(6):1245–1262.
- Kahane, S. 2003. The meaning-text theory. In Dependency and Valency: an International Handbook of Contemporary Research. Berlin, De Gruyter.
- Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edomonton, Canada, May 27-June 1.
- Lin, Dekang. 2004. A path-based transfer model for machine translation. In COLING '04: Proceedings of the 20th international conference on Computational Linguistics, page 625, Morristown, NJ, USA. Association for Computational Linguistics.

- Livio Robaldo & Leonardo Lesmo,2006, Dependency Tree Semantics, Foundations of Intelligent Systems, 16th International Symposium, ISMIS 2006, Bari, Italy, September 27-29, 2006, Proceedings M. Collins. 1997. Three generative, lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, pages 16–23, Madrid, Spain.
- McDonald, Ryan, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 91–98, Morristown, NJ, USA. Association for Computational Linguistics.
- Mel'cuk, I. 1988. Dependency syntax: theory and practice. New York: SUNY University Press. Mel'cuk, I. 2004. Actants in semantics and syntax. ii: actants in syntax. Linguistics, (42):247–291.
- Nirenburg, Sergei, Jaime Carbonnell, Masaru Tomita, and Kenneth Goodman. 1992. Machine Translation: A Knowledge-based Approach. Morgan Kaufmann Publishers, Los Altos, CA.
- Owczarzak, Karolina, Josef van Genabith, and Andy Way. 2007. Labelled dependencies in machine translation evaluation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 104–111, Prague, Czech Republic, June. Association for Computational Linguistics.
- Punyakanok Vasin, Roth Dan & Yih Wen-tau, 2004, Mapping Dependencies Trees: An Application to Question Answering.Available at: https://www.academia.edu/2661689/Mapping_dependencies_trees_An_application_to_question_answering.
- Quirk, Chris and Arul Menezes. 2006. Do we need phrases?: challenging the conventional wisdom in statistical machine translation. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 9–16, Morristown, NJ, USA. Association for Computational Linguistics.
- Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: syntactically informed phrasal smt. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 271–279, Morristown, NJ, USA. Association for Computational Linguistics.
- S. Harabagiu and D. Moldovan. 2001.Open-domain textual question answering. In Tuturial of the Second Meeting of the North American Chapter of the Association for Computational Linguistics.
- Smith, David A. and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation, pages 23–30, New York, June.

- Tapanainen, P. 1999. Parsing in two frameworks: finite-state and functional dependency grammar. Ph.D. thesis, University of Helsinki, Finland.
- Tesniere, L. 1959. El´ements de Syntaxe Structurale. Librairie C. Klincksieck, Paris.
- Toussaint Godfried T.,1978, *The use of context in pattern recognition*, V.10, Issue 3, 1978, Pages 189-204. Available at: https://www.sciencedirect.com/science/article/abs/pii/0031320378900274
- Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput. Linguist., 23(3):377–403.
- X. Li and D. Roth. 2002. Learning question classifiers. In COLING 2002, The 19th International Conference on Computational Linguistics, pages 556–562.
- Xia, Fei and Michael McCord. 2004. Improving a statistical machine translation system with automatically learned rewrite patterns. In COLING ’04: Proceedings of the 20th International Conference on Computational Linguistics, page 508, Morristown, NJ, USA. Association for Computational Linguistics.
- Y. Ding, D. Gildea, and M. Palmer. 2003. An algorithm for word-level alignment of parallel dependency trees. In The 9th Machine Translation Summit of International Association of Machine Translation, New Orleans, LA, 9.
- Y. Even-Zohar and D. Roth. 2001. A sequential model for multi-class classification. In Proceedings of 2001 Conference on Empirical methods in Natural Language Processing, Pittsburgh, PA.
- Yamada, Kenji and Kevin Knight. 2001. A syntax-based statistical translation model. In ACL ’01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.