Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης
& Υπολογιστων

Στατιστικοί Έλεγχοι στην Επεξεργασία Φυσικής Γλώσσας μέσω Η/Υ

Ανάκτηση Πληροφορίας, Εύρεση Συνεκφερόμενων Λέξεων & Αποσαφήνιση Εννοιών.

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ Τ. ΦΡΑΓΓΟΥ

Πτυχιούχου Πηροφορικής Ε.Κ.Π.Α (1994),
Μαθηματικών Ε.Κ.Π.Α (1985),
& Μεταπτυχιακού Ηλ.Αυτοματισμού Ε.Κ.Π.Α (1997)

Αθήνα, Οκτώβριος 2005

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων & Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης & Υπολογιστων

Στατιστικοί Έλεγχοι στην Επεξεργασία Φυσικής Γλώσσας μέσω Η/Υ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ Τ. ΦΡΑΓΓΟΥ

Συμβουλευτική Επιτροπή: Γιάννης Μαϊστρος (Επιβλέπων)
Στάθης Ζάχος
Εμμανουήλ Σκορδαλάκης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή

Γ. Μαϊστρος Επ. Καθηγητής Ε.Μ.Π.	Σ. Ζάχος Καθηγητής Ε.Μ.Π.	Ι. Βασιλείου Καθηγητής Ε.Μ.Π.
Α-Γ. Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Ν. Κοζύρης Επ. Καθηγητής Ε.Μ.Π.	Ν. Παπασπύρου Λέκτορας Ε.Μ.Π.
	Σ. Μαρκαντωνάτου Επ. καθηγήτρια Παν. Αθηνών	

Αθήνα, Οκτώβριος 2005

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΠΡΟΛΟΓΟΣ

Η ερώτηση:

> Μπορεί μια ενιαία στατιστική μεθοδολογία να απαντήσει σε προβλήματα επεξεργασίας φυσικής γλώσσας πού εμφανίζουν μια ομοιότητα ως προς το στόχο, ο οποίος είναι η επιλογή μεταξύ ανταγωνιζόμενων οντοτήτων: Για παράδειγμα, ανταγωνιζόμενα έγγραφα στην ανάκτηση πληροφορίας (information retrieval), ανταγωνιζόμενες έννοιες στην αποσαφήνιση της έννοιας μιας λέξης (Word Sense Diasambiguation), η ανταγωνιζόμενα ζευγάρια λέξεων (Collocations). Η παρούσα εργασία προσπαθεί να απαντήσει σε αυτό το ερώτημα.

Η στατιστική εχει καταδειχθεί σάν ο κλάδος της μαθηματικής επιστήμης που έχει χρησιμοποιηθεί με την μεγαλύτερη επιτυχία στην επεξεργασία φυσικής γλώσσας (Natural Language Processing: NLP). Τα συστήματα για αναζήτηση πληροφορίας (Information Retrieval: IR), αποσαφήνιση της έννοιας μιας λέξης (Word Sense Disambiguation: WSD), ο σχηματισμός Collocations και η κατηγοριοποίηση κειμένου (Text Categorization) είναι κλάδοι της επεξεργασίας φυσικής γλώσσας που εφαρμόζουν ευρύτατα στατιστικές μεθόδους. Ο σκοπός αυτής της διατριβής είναι να παρουσιάσει την εφαρμογή μιας ενιαίας στατιστικής μεθοδολογίας για την ανάπτυξη συστημάτων για τούς παραπάνω τομείς έρευνας. Συγκεκριμένα, την εύρεση λέξεων που σχηματίζουν collocation (συνεκφερόμενες λέξεις), την αναζήτηση κειμενικής πληροφορίας με βάση το ερώτημα ενός χρήστη και την αποσαφήνιση της έννοιας μιας λέξης από τα συμφραζόμενά της.

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Γιάννη Μαίστρο, για τις πολλές παρατηρήσεις και προτάσεις του καθώς και για την συνεχή του υποστήριξη κατά την διάρκεια αυτής της έρευνας.

Τον καθηγητή του TEI Αθήνας κ. Χρήστο Σκουρλά πού εξέφρασε το ενδιαφέρον του για την δουλειά μου και μοιράστηκε μαζί μου τις γνώσεις του ιδιαίτερα στην περιοχή του Information Retieval.

Την ομάδα ανάπτυξης του Lemur software από το Carnegie Mellon University για την πολύτιμη βοήθειά και υποστήριξη σε θέματα λογισμικού στην περιοχή του Information Retieval.

Τον συνάδελφό μου και καθηγητή στην μέση εκπαίδευση Δρ. Αναστάσιο Κουτσούκο για τις πολύτιμες συμβουλές του και την συνεχή ενθάρυνσή του.

Φυσικά και είμαι ευγνώμων στην Μαρία και γυιό μου Παντελή για την υπομονή τους και την αγάπη τους πρός το πρόσωπό μου.

Κωνσταντίνος Τ. Φράγγος
Αθήνα, Οκτώβριος 2005

ΠΕΡΙΛΗΨΗ

Ο σκοπός αυτής της διατριβής είναι να παρουσιάσει την εφαρμογή μιας ενιαίας στατιστικής μεθοδολογίας για την ανάπτυξη συστημάτων επεξεργασίας φυσικής γλώσσας. Πιό συγκεκριμένα, συστήματα για την εύρεση λέξεων που σχηματίζουν collocation, την αναζήτηση κειμενικής πληροφορίας με βάση το ερώτημα ενός χρήστη και την αποσαφήνιση της έννοιας μιας λέξης από τα συμφραζόμενά της.
Η μέθοδος αυτή βασιζεται στην χρήση των στατιστικών ελέγχων ″καλού ταιριάσματος″ (Goodness of Fit Statistical Tests : GOF).
Είναι απλή στην ουσία της και στηρίζεται στην ικανότητα της στατιστικής επιστήμης για ποσοτική εκτίμηση του επιπέδου σημαντικότητας μιας ″υπόθεσης″ ταιριάσματος μεταξύ ανταγωνιζόμενων οντοτήτων.

Αρχικά, παρουσιάζουμε μια εισαγωγή των στατιστικών μοντέλων που χρησιμοποιούνται στην επεξεργασία φυσικής γλάσσας καθώς επίσης και των μέτρων αποτίμησης της αποδοτικότητας των συστημάτων αυτών. Ακολουθεί η εφαρμογή των στατιστικών ελέγχων στην ανάκτηση πληροφορίας (information retrieval). Μέσα στο πλαίσιο των στατιστικών ελέγχων ″καλού ταιριάσματος″ (*GOF Statistical Tests*) παρουσιάζουμε ένα σύστημα για αναζήτηση κειμενικής πληροφορίας απο ″δεξαμενές″ εγγράφων (document repositories) με βάση το ερώτημα ενός χρήστη. Στην συνέχεια, παρουσιάζουμε στατιστικές μεθόδους για την ″ανακάλυψη″ λέξεων μέσα σε Ελληνικά κείμενα οι οποίες σχηματίζουν collocations και θεμελιώνουμε ένα τρόπο εφαρμογής των στατιστικών ελέγχων στην περιοχή αυτή. Τέλος εφαρμόζουμε τους στατιστικούς ελέγχους στην περιοχή της αποσαφήνισης της έννοιας μιας λέξης (word sense disambiguation). Ένα στατιστικό σύστημα αναπτύσεται για την αποσαφήνιση της έννοιας μια λέξης από τα συμφραζόμενά της κάνοντας χρήση του ηλεκτρονικού λεξικού WordNet σαν λεξικολογική πηγή.

Τα συμπεράσματα που προκύπτουν μετά απο αποτίμηση των μεθόδων πού αναπτύξαμε πάνω σε πειραματικά δεδομένα ελέγχου, είναι οτι τα στατιστικά αυτά συστήματα αποδεικνύονται ″εύρωστα″ και ικανά να δώσουν αποτελέσματα καλύτερα από αυτά των κλασσικών μεθόδων πού χρησιμοποιούνται στα αντίστοιχα επιστημονικά πεδία.

ABSTRACT

Statistical methods have been successfully applied in the area of natural language processing. The aim of this work is to apply a unified statistical method for natural language processing tasks which involve competing entities which try to better fit in a given target linguistic framework. For example, competing words which try to form collocations with a target word, competing documents which try to appear more similar to a given query in information retrieval, competing senses of a target word within a linguistic context, etc.
We start our work with an introduction to the statistical models used in natural language processing systems, as well as to the evaluation measures used to evaluate the performance of the various systems. An application of the statistical tests in the area of information retrieval is followed. Within the goodness of fit statistical test's framework, we develop a system for finding relative documents to a given query from document repositories. We proceed with the development of some statistical methods for finding collocations. At this point, we introduce a framework to apply the goodness-of-fit statistical tests as a method to push forward pairs of words which form collocations. Finally, in the area of the word sense disambiguation we apply the statistical tests to find the correct sense of a target word given its linguistic context. We evaluated all the above systems following the standard evaluation criteria used in the international literature. It is proven that the systems developed within the goodness of fit statistical test's framework are robust systems, performing well and in most of the cases above the baseline of some of the well established in their own area systems.

ΚΩΝΣΤΑΝΤΙΝΟΣ Τ. ΦΡΑΓΓΟΣ
Πτυχιούχος Μαθηματικών & Πληροφορικής του Πανεπιστημίου Αθηνών

Κατάλογος Συντμήσεων

ΕΦΓ	:	Επεξεργασία Φυσικής Γλώσσας
ΑΛ	:	Αποσαφήνιση Λέξης
ΑΠ	:	Αναζήτηση Πληροφορίας
ΣΛ	:	Συνεκφερόμενες Λέξεις
GOF	:	$Goodness\ of\ fit$

Κεφάλαιο 1

Εισαγωγή

Η στατιστική είναι ο κλάδος της μαθηματικής επιστήμης που έχει χρησιμοποιηθεί ευρύτατα στην Επεξεργασία Φυσικής Γλώσσας (ΕΦΓ). Η αλματώδη εξέλιξη της πληροφορικής τα τελευταία χρόνια και η διαθεσιμότητα μεγάλου όγκου κειμένων σε ψηφιακή μορφή (corpora), δημιούργησαν τις συνθήκες για την αναγέννηση των ποσοτικών μεθόδων στην (ΕΦΓ) και σαν συνέπεια την ευρύτατη εφαρμογή και χρήση των στατιστικών μεθόδων στούς διάφορους κλάδους της επεξεργασίας φυσικής γλώσσας.

Η ανάκτηση πληροφορίας (Information Retrieval) ως υποκλάδος της επεξεργασίας φυσικής γλώσσας ασχολείται με την ανάπτυξη αλγορίθμων και μοντέλων για την αναζήτηση πληροφορίας από διάφορες συλλογές κειμένων. Αν και παραδοσιακά η ανάκτηση πληροφορίας ασχολείται με το κείμενο, η αναζήτηση και άλλων μορφών πληροφορίας όπως εικόνας και βίντεο αρχίζει σήμερα να αποκτά ολοένα και περισσότερο ενδιαφέρον. Με την αναγέννηση των ποσοτικών μεθόδων επεξεργασίας φυσικής γλώσσας, οι στατιστικες μέθοδοι έγιναν η κυρίαρχη προσέγγιση ανάπτυξης συστημάτων για ανάκτηση πληροφορίας.

Οι στατιστικές μέθοδοι θεωρούνται ως το αποκλειστικό εργαλείο για την ανάπτυξη συστημάτων για την αποσαφήνιση λεκτικής σημασίας (word sense disambiguation), κατηγοριοποίηση κειμένου (text classification), εύρεση collocations κλπ, τα οποία αναγνωρίζονται σαν υπολογιστικά πολύπλοκα προβλήματα στην επεξεργασία φυσικής γλώσσας και η επίλυσή τους αναμένεται να επηρεάσει καταλυτικά την εξέλιξη του κλάδου της υπολογιστικής γλωσσολογίας (computational linguistic).

Στις επόμενες ενότητες αυτού του κεφαλαίου θα δώσουμε πρώτα μια σύντομη εισα-

γωγή των στατιστικών μεθόδων που χρησιμοποιούνται στα συστήματα επεξεργασίας και ανάκτησης πληροφορίας, καθώς και των μέτρων και δεδομένων που χρησιμοποιούνται για την αποτίμηση της αποδοτικότητας αυτών των συστημάτων. Οι μέθοδοι αυτές προτάσονται εδώ γιατί εφαρμόζονται γενικότερα σε όλα τα στατιστικά συστήματα επεξεργασίας φυσικής γλώσσας.

1.1 Στατιστικά Μοντέλα

Η έρευνα στα στατιστικά συστήματα επεξεργασίας φυσικής γλώσσας ασχολείται με την ανάπτυξη αλγορίθμων και συστημάτων για την αναπαράσταση, αποθήκευση, οργάνωση, επεξεργασία και προσπέλαση των στοιχείων της πληροφορίας. Οι πρώτες προσπάθειες για αναπαράσταση και ανάκτηση πληροφορίας ξεκίνησαν με τα συστήματα αναζήτησης πληροφορίας. Αν και παραδοσιακά ο κλάδος ασχολείτο μόνο με την αναζήτηση κειμένων και την εύρεση εγγράφων, σήμερα, με την πρόοδο της τεχνολογίας και την ανάπτυξη των πολυμέσων υπάρχει έντονο ενδιαφέρον και για άλλες μορφές πληροφορίας όπως εικόνα, ήχος και βίντεο.

Η αναπαράσταση της κειμενικής πληροφρίας σε υπολογίσιμη μορφή παίζει καθοριστικό ρόλο στην ανάπτυξη συστημάτων επεξεργασίας φυσικής γλώσσας. Παρουσιάζουμε παρακάτω τούς τρόπους αναπαράστασης πού κατά καιρούς προτάθηκαν από τα διάφορα συστήματα αναπαράστασης και ανάκτησης πληροφορίας.

Σε ένα τυπικό σύστημα αναζήτησης πληροφορίας, ο χρήστης έχει μια ανάγκη για πληροφορία η οποία μεταφράζεται σε κάποιο ερώτημα (query), συνήθως μια σειρά από λέξεις, και το σύστημα λαμβάνοντας αυτό το ερώτημα απαντά με μια βαθμολογημένη λίστα από σχετικά έγγραφα. Ο μηχανισμός εύρεσης των σχετικών εγγράφων περιγράφεται διά μέσου των μοντέλων αναζήτησης (retrieval models) στα οποία τα έγγραφα παριστάνονται σαν ένα σύνολο από αντιπροσωπευτικές λέξεις κλειδιά (keywords) πού καλούνται index terms. Το έργο του συστήματος αναζήτησης πληροφορίας είναι να αναπαραστήσει κάθε έγγραφο σαν ένα σύνολο από λέξεις κλειδιά, ή αλλοιώς index terms, να αποφασίσει για την σημαντικότητα κάθε index term στην αναπαράσταση του περιεχομένου του εγγράφου, και να αποφασίσει χρησιμοποιώντας έναν αλγόριθμο ταιριάσματος για το πόσο σχετικό είναι το έγγραφο στην ερώτηση του χρήστη.

Ανάλογα με την φύση της διαδικασίας αναπαράστασης ενός εγγράφου σαν σύνολο από index terms πού ο κάθε όρος έχει ένα σχετικό βάρος σημαντικότητας (term weighting scheme), μπορούμε να κατατάξουμε τα πιο σημαντικά μοντέλα αναπαράστασης που προτάθηκαν στίς εξής κύριες κατηγορίες: Δυαδικά μοντέλα (boolean models), διανυσματικά μοντέλα (vector models) και πιθανοτικά μοντέλα (probabilistic models). Το δυαδικό μοντέλο είναι το πιο απλό μοντέλο το οποίο βασίζεται στην θεωρία συνόλων και την boolean άλγεβρα, ενώ τα άλλα δύο θεωρούνται μοντέλα που χρησιμοποιούν στατιστικές μεθόδους. Τα δυαδικά συστήματα είναι καλύτερα γνωστά και σάν μοντέλα του "ακριβούς" ταιριάσματος, διότι επιστρέφουν έγγραφα τα οποία ακριβώς ικανοποιούν μια δομημένη έκφραση (boolean queries). Παρά την απλότητά του, το δυαδικό μοντέλο εφαρμοζόμενο στην αναζήτηση πληροφορίας υποφέρει από αρκετά μειονεκτήματα μεταξύ των οποίων είναι, η στρατηγική αναζήτησης η οποία βασίζεται σε μια δυαδική απόφαση (relevant, non-relevant) χωρίς καμμία διαβάθμιση της σχετικότητας, καθώς επίσης και από την δυσκολία που υπάρχει να εκφρασθεί ένα ερώτημα σε boolean έκφραση από τον χρήστη. Χρησιμοποιούνται όμως ακόμα σε εμπορικά πληροφοριακά συστήματα.

Για το ξεπέρασμα των παραπάνω δυσκολιών αναπτύχθηκαν τα στατιστικά μοντέλα.

Είναι πολύ γνωστό ότι τα συστήματα τα οποία υιοθετούν γενικά ένα μη δυαδικό term weighting scheme εμφανίζουν καλύτερες αποδόσεις στην αναζήτηση πληροφορίας απ΄ ό,τι τα δυαδικά συστήματα. Το διανυσματικό μοντέλο, [1], [2], είναι το πρώτο μοντέλο στην αναζήτηση πληροφορίας το οποίο αναγνωρίζει μεν τους περιορισμούς ενός ακριβούς ταιριάσματος μεταξύ ερωτήματος και εγγράφου, αλλά προτείνει ένα index term weighting σχήμα το οποίο επιτρέπει μερικό ταίριασμα. Σύμφωνα με το διανυσματικό μοντέλο, κάθε όρος k_i σε ένα έγγραφο d_j χαρακτηρίζεται με ένα θετικό μη μηδενικό πραγματικό αριθμό που καλείται βάρος (weight) και εκφράζει την σημαντικότητα τού όρου στον προσδιορισμό της σημασιολογίας του εγγράφου. Επί πλέον και οι λέξεις κλειδιά στο ερώτημα χαρακτηρίζονται και αυτοί με ένα βάρος. Εάν αναπαραστήσουμε ένα έγγραφο d_j σαν ένα διάνυσμα $(w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ και το ερώτημα q σαν ένα διάνυσμα $(w_{1,q}, w_{2,q}, \ldots, w_{t,q})$, όπου t είναι ο συνολικός αριθμός των όρων στο σύστημα, τότε μπορούμε να χρησιμοποιήσουμε το συνημίτονο (cosine) της γωνίας μεταξύ των δύο διανυσμάτων σαν ένα μέτρο της ομοιότητας μεταξύ του ερωτήματος

και του εγγράφου.

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{1.1}$$

Επιστρέφουμε τώρα στο weighting σχήμα του διανυσματικού μοντέλου. Κάποιος θα μπορούσε να μετρήσει τον αριθμό των εμφανίσεων ενός όρου στο έγγραφο (term frequency) και τον αριθμό των εγγράφων στα οποία ο όρος αυτός εμφανίζεται (document frequency) , και να συνδυάσει αυτά σε ένα μοναδικό βάρος που θα χαρακτηρίζει τον όρο ως εξής:

$$w_{i,j} = \begin{cases} a(1 + log(tf_{i,j}))log\frac{N}{df_i} & tf_{i,j} \geq 0 \\ 0 & tf_{i,j} = 0 \end{cases} \tag{1.2}$$

Όπου, $tf_{i,j}$ είναι η συχνότητα του όρου στο έγγραφο, df_i ο αριθμός των εγγράφων στα οποία ο όρος εμφανίζεται και N ο αριθμός των εγγράφων στην συλλογή. Ο όρος N/df_i συχνά αποκαλείται και αντίστροφη συχνότητα εγγράφου (inverse document frequency) ή (idf-weighting). Η συχνότητα του όρου στο έγγραφο και η αντίστροφη συχνότητα εγγράφου έχουν αναχθεί λογαριθμικά.

Ο παραπάνω τύπος είναι ένα παράδειγμα μιας μεγαλύτερης οικογένειας weighting σχημάτων γνωστών σαν $tf.idf$ term weighting σχήματα. Αυτά τα σχήματα είναι απλά, εύκολα στην κατανόηση και έχουν αποδειχθεί "εύρωστα" στην πράξη σε μια ευρεία περιοχή εφαρμογών.

Μια εναλλακτική προσέγγιση στο $tf.idf$ term weighting σχήμα είναι η χρήση της θεωρίας των πιθανοτήτων για την ανάπτυξη ενός μοντέλου που να περιγράφει την κατανομή των διαφόρων όρων στο έγγραφο και με την χρήση αυτής της κατανομής να χαρακτηρίζεται η σημαντικότητα του όρου μέσα στο σύστημα αναζήτησης. Υπάρχουν πολλά μοντέλα στην αναζήτηση πληροφορίας που βασίζονται στην θεωρία πιθανοτήτων [3], [4], [5], [6], [7].

Πρόσφατα μια νέα προσέγγιση στη μοντελοποίηση γλώσσας (language modelling) έχει προταθεί σαν μια εναλλακτική λύση στα παραδοσιακά διανυσματικά και τα άλλα πιθανοτικά μοντέλα. Έχει εφαρμοσθεί με επιτυχία στα συστήματα ανζήτησης πληροφορίας [8], [9], [10], [11]. Ένα στατιστικό μοντέλο γλώσσας είναι ένας πιθανοτικός μηχανισμός για την μελέτη και παραγωγή κειμένου πού η καταγωγή του ανάγεται στην

εποχή του Shannon [12], ο οποίος διατύπωσε την πολύ γνωστή θεωρία του στον τομέα των επικοινωνιών (source channel perspective). Ο Shannon μελέτησε πόσο καλά τα απλά ν-γράμματα μοντέλα (n-gram models) μπορούν να προβλέψουν φυσικό κείμενο.

Αν και για αρκετά χρόνια πρίν υπήρχε μεγάλο ενδιαφέρον για τις μεθόδους που βασίζονται στα μοντέλα γλώσσας για πρόβλεψη και αναπαραγωγή κειμένου σε μια μεγάλη πικοιλία από εφαρμογές στο πεδίο της επεξεργασίας φυσικής γλώσσας, στην αναζήτηση πληροφορίας οι ιδέες της μοντελοποίησης γλώσσας χρησιμοποιήθηκαν μάλλον πρός στην αντίθετη κατεύθυνση. Με τα κλασικά πιθανοτικά μοντέλα [3], [5], [13], [14], χρησιμοποιώντας ένα μοντέλο για την απόδοση μιας πιθανότητας σε μια λέξη (unigram language model), υπάρχει η ανάγκη να κατανείμουμε μια μάζα πιθανότητας (probability mass) πάνω σε ένα τεράστιο χώρο των πιθανών τιμών για κάθε όρο. Αυτή η πιθανότητα είναι δύσκολο να ελεγχθεί γιατί η μόνη ένδειξη στις πιο πολλές περιπτώσεις είναι μόνο οι όροι του ερωτήματος και η κατασκευή ενός ακριβούς μοντέλου κατανομής είναι αδύνατη. Παρά όμως αυτή την δυσκολία, αυτά τα μοντέλα έχουν εφαρμοσθεί σε πολλά πεδία της αναζήτησης πληροφορίας με σημαντική επιτυχία.

Για να ξεπερασθεί αυτό το πρόβλημα, το 1998 οι Ponte και Croft [8], χρησιμοποιώντας μια smoothed εκδοχή του unigram language μοντέλου πρότειναν μια μέθοδο για να αποδώσουν μια τιμή πιθανοφάνειας (likelihood score) από το έγγραφο στο ερώτημα. Αυτή η μέθοδος είναι γνωστή και σαν προσέγγιση του μοντέλου γλώσσας και η οποία μπορεί να ερμηνευθεί ως εξής. Ένα μοντέλο γλώσσας θεωρείται σαν ένα θορυβώδες κανάλι ″noisy channel″ ή, ″translation model″ το οποίο απεικονίζει τα έγγραφα στα ερωτήματα.

Ένα σημαντικό θέμα που αφορά την απόδοση ενός συστήματος στην αναζήτηση πληροφορίας είναι η επαύξηση του ερωτήματος (query expansion) με νέους όρους, ειδικά για πολύ σύντομα (ολίγων λέξεων) μικρά ερωτήματα. Μια δημοφιλής τεχνική πού χρησιμοποιείται στα περισσότερα συστήματα είναι η στρατηγική του relevance feedback. Σε ένα τυπικό σενάριο αυτής της στρατηγικής, ο χρήστης πρώτα σχηματοποιεί και μετά υποβάλλει στο σύστημα ένα ερώτημα το οποίο αναπαριστά την πληροφοριακή του ανάγκη. Έπειτα το σύστημα επιστρέφει κάποια σχετικά έγγραφα και ζητείται από τον χρήστη να αξιολογήσει την σειρά των εγγράφων της επιστρεφόμε-

νης λίστας. Με βάση αυτή την αξιολόγηση το σύστημα ξεκινά νέα αναζήτηση (query reformulation) [2], [3]. Το μειονέκτημα είναι οτι ο χρήστης ενώ ψάχνει για πληροφορία είναι επιφορτισμένος με το επί πλέον καθήκον της ανάλυσης των εγγράφων που έχουν βρεθεί από το σύστημα σε μια πρωταρχική αναζήτηση.

Σαν μια ενναλλακτική περίπτωση αυτής της κατάστασης είναι το γνωστό pseudo-relevance feedback. σε αυτή την μέθοδο αντί να βασιζόμαστε στον χρήστη να βαθμολογήσει ας υποθέσουμε τα k πιο σχετικά έγγραφα, το σύστημα απλά υποθέτει σαν πιο σχετικά έγγραφα τα πρώτα k πιό υψηλά βαθμολογημένα έγγραφα (top ranked) σε ένα αρχικό δοκιμαστικό τρέξιμο του αλγορίθμου, και έπειτα χρησιμοποιεί αυτά τα έγγραφα για να επαυξήσει το ερώτημα του χρήστη.

Στα διανυσματικά μοντέλα (tf-idf schemas), μια δημοφιλής κλασσική τεχνική για την ενσωμάτωση relevance feedback είναι η επαναληπτική μέθοδος μεταβολής του αρχικού ερωτήματος από τα ανατροφοδοτούμενα στο σύστημα σχετικά έγγραφα σύμφωνα με την standard Rochio formula [15].

$$q_m = \alpha q + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} d_j - \frac{\gamma}{|D_n|} \sum_{\forall d_j \in D_n} d_j \tag{1.3}$$

Όπου q_m το διάνυσμα του ερωτήματος που τροποποιείται (modified query vector) , D_r το σύνολο των σχετικών εγγράφων όπως καθορίστηκαν από την διαδικασία ανατροφοδότησης (feedback) , D_n το σύνολο των μη-σχετικών (non-relevant) εγγράφων μεταξύ των ευρεθέντων εγγράφων από το αρχικό τρέξιμο, $|D_r|$, $|D_n|$ ο αριθμός των εγγράφων στα σύνολα D_r, D_n αντίστοιχα και α, β, γ παράμετροι ρύθμισης. Στην αρχική τυποποίηση της μεθόδου, ο Rochio χρησιμοποίησε μια απλοποιημένη εκδοχή της εξίσωσης 1.3 θέτοντας $\beta = 1$. Εάν $\gamma = 0$, έχουμε μια απλοποιημένη στρατηγική θετικής ανατροφοδότησης διότι δεν ενσωματώνουμε πληροφορία από τα μη σχετικά έγγραφα.

1.2 Μέτρα Αποτίμησης

Σε αυτή την ενότητα θα περιγράψουμε τα μέτρα που θα χρησιμοποιήσουμε για την αποτίμηση της αποδοτικότητας (evaluation) των συστημάτων Ανάκτησης Πληροφορίας.

Κάνουμε τις εξής υποθέσεις: θεωρούμε οτι η αποτίμηση της αποδοτικότητας ενός

συστήματος γίνεται πάνω σε μια συλλογή απο έγγραφα που χρησιμοποιείται σαν αναφορά για τον έλεγχο της αποτίμησης Η συλλογή αποτελείται εκτός απο τα έγγραφα και από ένα σύνολο παραδειγμάτων-ερωτημάτων (queries) τα οποία το καθένα εκφράζει μια πληροφοριακή ανάγκη πού πρέπει να ικανοποιηθεί με μια σειρά απο επιστρεφόμενα σχετικά έγγραφα από την συλλογή.

Δοθείσης μιας στρατηγικής αναζήτησης S, (retrieval strategy) το μέτρο της εκτίμησης (evaluation measure) αυτής της στρατηγικής θα πρέπει να αντανακλά την ομοιότητα μεταξύ τού συνόλου των εγγράφων πού βρέθηκαν από την retrieval στρατηγική S και του συνόλου των σχετικών εγγράφων απο την συλλογή, τα οποία παρέχονται από τους ειδικούς (specialists) σαν σχετικά.

Τα μέτρα εκτίμησης της αποδοτικότητας των information retrieval συστημάτων περιφέρονται γύρω απο δύο έννοιες του *precision* και του *recall* . Δίνουμε παρακάτω τους ορισμούς για αυτές τις δύο έννοιες.

Έστω ότι έχουμε ένα αίτημα για πληροφορία ή όπως αλλοιώς λέμε ένα query q και έστω R το σύνολο των σχετικών εγγράφων με αυτό το query έτσι όπως αυτά παρέχονται από τους ειδικούς αξιολογητές. Ας υποθέσουμε τώρα ότι μια retrieval στρατηγική S (η οποία ελέγχεται ως πρός την αποδοτικότητά της), επεξεργάζεται την πληροφοριακή ανάγκη q και παράγει σαν απάντηση ένα σύνολο από έγγραφα A. Εστω $|A|$ ο αριθμός αυτών των αντικειμένων στο σύνολο A. Επί πλέον, έστω $|Ra|$ ο αριθμός των εγγράφων στην τομή (intersection) των συνόλων R και A.

Τα μέτρα αποτίμησης (evaluation measures) *recall* και *precision* ορίζονται ως ακολούθως:

- **Recall** είναι το κλάσμα από τα σχετικά documents (R) τα οποία ευρέθησαν, δηλαδή
$$Recall = \frac{|Ra|}{|R|} \tag{1.4}$$
- **Precision** είναι είναι το κλάσμα από τα ευρεθέντα έγγραφα (A) τα οποία είναι σχετικά, δηλαδή
$$Precision = \frac{|Ra|}{|A|} \tag{1.5}$$

Τα μέτρα *Recall* και *Precision* όπως ορίστηκαν παραπάνω, προυποθέτουν ότι όλα τα στο σύνολο απαντήσεων A έχουν εξετασθεί (ειδωθεί) με την μία, για να υπολογισθούν

τα μέτρα *Recall* και *Precision*. Αντί για αυτό όμως, στην πράξη τα έγγραφα στο σύνολο A πρώτα ταξινομούνται σύμφωνα μέ ένα βαθμό σχετικότητας (δηλαδή δημιουργείται μια βαθμολογημένη λίστα (ranking)). Ο χρήστης έπειτα εξετάζει αυτή την βαθμολογημένη λίστα ξεκινώντας από το πιο πάνω έγγραφο. Σε αυτή την κατάσταση, τα μέτρα *Recall* και *Precision* μεταβάλλονται καθώς ο χρήστης προχωρά την εξέτασή του στο σύνολο απαντήσεων A. Επομένως, συνηθίζεται σαν καλύτερο μέτρο αποτίμησης να αποδίδουμε την γραφική παράσταση του *Precision* ως πρός το *Recall*. Δηλαδή καθώς προχωράμε την εξέταση της επιστρεφόμενης βαθμολογημένης λίστας για κάθε τιμή (level) *recall* που υπολογίζουμε σε ένα σημείο της λίστας υπολογίζουμε την αντίστοιχη τιμή για το *Precision*. Η precision versus recall καμπύλη συνήθως βασίζεται σε 11 standard recall επίπεδα τα οποία είναι (0%, 10%, ..., 100%) και είναι γνωστή ως *non interpolated precision at 11 recall points*.

Τα παραπάνω μέτρα τα υπολογίσαμε για ένα μόνον Ερώτημα (query). Για να υπολογίσουμε την αποδοτικότητα στο retrieval ενός συστήματος πάνω σε ένα σύνολο από Ερωτήματα, βρίσκουμε το μέσο *precision* σε κάθε *recall* επίπεδο ώς ακολούθως:

$$\overline{P(r)} = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q} \tag{1.6}$$

Όπου $\overline{P(r)}$ είναι το μέσο (average) precision στο recall επίπεδο r, N_q είναι ο αροθμός των Ερωτημάτων που χρησιμοποιήθηκαν, και $P_i(r)$ είναι είναι το precision στο recal επίπεδο r για το i-th Ερώτημα.

Η μέση Ακρίβεια (average precision) και η non interpolated precision στα 11 recall σημεία είναι τα μέτρα πού θα χρησιμοποιήσουμε στην παρούσα εργασία για την αποτίμηση της αποδοτικότητας των information retrieval συστημάτων.

Κεφάλαιο 2

Εφαρμογή των Στατιστικών Ελέγχων στην Ανάκτηση Πληροφορίας

Στα περισσότερα πιθανοτικά μοντέλα που χρησιμοποιούμε στις στατιστικές προσεγγίσεις για την Αναζήτηση Πληροφορίας (information retrieval), ενδιαφερόμαστε στο να εκτιμήσουμε πόσο ″καλά″ το μοντέλο του εγγράφου (document model) ″ταιριάζει″ στην πληροφοριακή ανάγκη του χρήστη (query model). Από την άλλη πλευρά στην στατιστική, υπάρχουν καλά θεμελιωμένες τεχνικές για την εκτίμηση του κατά πόσο ένα μοντέλο ″ταιριάζει″ με κάποιο άλλο μοντέλο. Οι στατιστικοί έλεγχοι καλού ″ταιριασματος″ (goodness of fit statistical tests) είναι πολύ γνωστές μέθοδοι για την εκτίμηση της υπόθεσης πού υπόκειται σε ένα σύνολο δεδομένων (data set).

Την υπόθεση ότι οι όροι ενός ερωτήματος (query) κατανέμονται τυχαία στα διάφορα έγγραφα της συλλογής, δηλαδή ακολουθούν μια κατανομή που υπαγορέυεται από τους κανόνες της ″τυχαιότητας″, μπορούμε να την εκτιμήσουμε ποσοτικά με την βοήθεια των στατιστικών ελέγχων καλού ″ταιριασματος″.

Στη βασική θέση της διατριβής αναπτύσσουμε μια τεχνική για Αναζήτηση Πληροφορίας της οποίας η τεχνική της για την βαθμολόγηση των εγγράφων στηρίζεται στον X-τετράγωνο έλεγχο Καλού Ταιριάσματος (X^2 square Goodness of Fit statistical test). Για λόγους συντομίας από εδώ και στο εξής θα χρησιμοποιούμε τον συμβολισμό $X^2 - GOF$ έλεγχος. Η τεχνική αυτή εκτός του ότι αποδεικνύεται ιδιαίτερα αποδοτική, είναι και ευέλικτη ώστε να μπορεί να προσαρμοσθεί και σε διαφορετικά προβλήματα, εκεί όπου υπεισέρχεται η έννοια της εκτίμησης του ″ταιριάσματος″ (fitting),

όπως πχ, η αποσαφήνιση της σωστής έννοιας μιας λέξης στα συμφραζόμενά της (Word Sense Disambiguation) .

Για να υλοποιήσουμε την μέθοδό μας κάνουμε μια υπόθεση για τα δεδομένα, γνωστή και σαν μηδενική υπόθεση ″null hypothesis″ . Σύμφωνα με αυτή θεωρούμε ότι δεν υπάρχει κάποια ιδιαίτερη σχέση ή δεσμός (association) μεταξύ του ερωτήματος (query) και ενός συγκεκριμένου εγγράφου, εκτός από το ότι οι όροι του ερωτήματος μπορεί να εμφανισθούν σε αυτό το έγγραφο από ″τύχη″ και μόνο.

Για να εκτιμήσουμε αυτή την υπόθεση εκτελούμε ένα X-τετράγωνο έλεγχο καλού ″ταιριάσματος″ και υπολογίζουμε την αντίστοιχη chi-square τιμή. Ο μαθηματικός τύπος αναζήτησης βασίζεται στην βαθμολόγηση όλων των εγγράφων της συλλογής σύμφωνα με αυτές τις υπολογιζόμενες X-τετράγωνο τιμές.

Η αποδοτικότητα της μεθόδου αυτής ελέγθηκε και εκτιμήθηκε πάνω σε μεγάλες συλλογές εγγράφων από τα πειραματικά TREC δεδομένα για έλεγχο συστημάτων και αποδεικνύεται πολύ αποτελεσματική. Η απόδοσή της βρίσκεται σταθερά πάνω απο τη γραμμή απόδοσης των κλασσικών tf-idf σχημάτων. Επί πλέον η μέθοδος προσφέρει την δυνατότητα να μοντελοποιήσουμε με διάφορους τρόπους τα ερωτήματα και τα έγγραφα, κάνοντας διαφορετικές υποθέσεις και εκτιμώντας ή υπολογίζοντας οποιαδήποτε ιδιαίτερη θεωρητική κατανομή που να ταιριάζει καλύτερα στην ιδιομορφία που ενδεχομένως παρουσιάζουν τα δεδομένα, και να εκμεταλλευτούμε στην συνέχεια την δύνατότητα για εκτίμηση της σημαντικότητας αυτής της υπόθεσης, που μας προσφέρουν οι GOF στατιστικοί έλεγχοι.

2.1 Εισαγωγή στα Στατιστικά Μοντέλα Γλώσσας

Όπως αναφέραμε και στην πρώτη ενότητα η Αναζήτηση Πληροφορίας είναι η περιοχή του κλάδου της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) όπου η στατιστική έχει εφαρμοσθεί με πολύ μεγάλη επιτυχία.

Δύο στατιστικά μοντέλα για βαθμολόγηση εγγράφων (document ranking) πού αναπτύχθηκαν στις αρχές της δεκαετίας του 70 και παραμένουν ακόμα και σήμερα σε χρήση είναι: Το μοντέλο διανυσματικού χώρου (vector space model) που προτάθηκε

από τον Salton [2] και το πιθανοτικό μοντέλο των Robertson και Sparck-Jones [3].

Για 30 και πλέον χρόνια το στατιστικό γλωσσικό μοντέλο υπήρξε το όχημα για την στατιστική επεξεργασία αναγνώρισης λόγου (statistical speech recognition) και οι περισσότερες από τις στατιστικές τεχνικές που πρωτοεφαρμόσθηκαν σε συστήματα αναγνώρισης λόγου, όπως το μοντέλο του θορυβώδους καναλιού του Shannon (noisy channel model), τα N-γράμματα μοντέλα (n-gram models) καθώς και τα ″Κρυμμένα″ Μαρκοβιανά Μοντέλα (hidden Markov models) χρησιμοποιούνται ακόμα και σήμερα σε μεγάλο μέρος εφαρμογών στην στατιστική επεξεργασία φυσικής γλώσσας.

Μόνο όμως πρόσφατα στην περιοχή της Αναζήτησης Πληροφορίας ανακαινήθηκε ζωηρό ενδιαφέρον για τα στατιστικά μοντέλα γλώσσας. Το στατιστικό γλωσσικό μοντέλο είναι ένας πιθανοτικός μηχανισμός ο οποίος μπορεί να αναπαράγει κείμενο αποδίδοντας στην εμφάνιση των λέξεων του κειμένου κάποια πιθανότητα. Αυτό το μοντέλο όπως είπαμε έχει τις καταβολές του στη εποχή του Shannon [12] ο οποίος στην προσπάθειά του να μοντελοποιήσει την παραγωγή κειμένου φυσικής γλώσσας, πρότεινε το μοντέλο πηγής (source channel perspective) το οποίο εξέταζε το πόσο καλά τα απλά N-γράμματα μοντέλα μπορούσαν να προβλέψουν την εμφάνιση φυσικού κειμένου.

Είναι απορίας άξιο πώς, ενώ για πολλά χρόνια υπήρξε ζωηρό ενδιαφέρον στην εφαρμογή της στατιστικής μοντελοποίησης της γλώσσας για πρόβλεψη και παραγωγή κειμένου σε μεγάλο πλήθος από εφαρμογές στην επεξεργασία φυσικής γλώσσας, στην Αναζήτηση Πληροφορίας μόνο πρόσφατα ανακαινήθηκε τέτοιο ενδιαφέρον. Η εξήγηση μάλλον βρίσκεται στο γεγονός ότι οι ιδέες του language modelling χρησιμοποιήθηκαν πρός την αντίθετη κατεύθυνση.

Στα κλασικά πιθανοτικά μοντέλα των Robertson και Sparck Jones [3], το OKAPI σύστημα [5], τα πολύ γνωστά naive-Bayesian networks [13] και το Inquery σύστημα [14], γίνεται η χρήση στατιστικών μοντέλων γλώσσας αλλά είναι ανάγκη σε αυτά τα μοντέλα να κατανεμηθεί ελεύθερα μια μάζα πιθανότητας πάνω σε ένα χώρο (space) με ένα τεράστιο αριθμό από πιθανές εκβάσεις γεγονότων (events). Αυτή η πιθανότητα είναι εξαιρετικά δύσκολο να ελεγχθεί στην περιοχή της Αναζήτησης Πληροφορίας γιατί η μόνη ένδειξη που έχουμε τις περισσότερες φορές είναι οι λίγες λέξεις του ερω-

τήματος του χρήστη, κατά συνέπεια είναι σχεδόν αδύνατη η κατασκευή ενός ακριβούς πιθανοτικού μοντέλου. Όμως παρά την παραπάνω δυσκολία, αυτά τα μοντέλα έχουν εφαρμοσθεί και μάλιστα με επιτυχία σε πολλές εφαρμογές για Αναζήτηση Πληροφορίας.

Στα προηγούμενα δύο χρόνια έγιναν πολλές δημοσιεύσεις στις οποίες τα στατιστικά μοντέλα γλώσσας χρησιμοποιούνται στον υπολογισμό της εκτίμησης της σχετικότητας ενός εγγράφου με το ερώτημα. Ο Miller και οι υπόλοιποι [48] [11], χρησιμοποιούν ″Κρυμμένα″ Μαρκοβιανά Μοντέλα για την βαθμολόγηση των εγγράφων. Δεν χρησιμοποιούν μονο μοντέλα μιας λέξης αλλα επεκτείνονται στην χρήση συνδυασμών δύο διαδοχικών λέξεων (bigrams) για να μοντελοποιήσουν φράσεις με δύο λέξεις, επί πλέον δε κάνουν χρήση και μιας μεθόδου για επαύξηση των όρων του ερωτήματος με ανάδραση (feedback).

Οι Berger και Lafferty [10] ανέπτυξαν ένα μοντέλο στατιστικής μετάφρασης της γλώσσας για να χρησιμοποιήσουν επί πλέον όρους όπως συνώνυμα και σχετιζόμενες λέξεις.

Στην δική μας προσέγγιση χρησιμοποιούμε με ένα διαφορετικό τρόπο τα στατιστικά μοντέλα γλώσσας. Για να βαθμολογήσουμε τα διάφορα έγγραφα με δεδομένο ένα ερώτημα, βασιζόμαστε στούς X-τεράγωνο ελέγχους οι οποίοι κάνουν μια εκτίμηση κατά πόσο ένα στατιστικό μοντέλο περιγράφει ″καλά″ τα δεδομένα. Πιο συγκεκριμένα, θεωρώντας την μηδενική υπόθεση (null hupothesis) ότι οι διάφοροι όροι κατανέμονται τυχαία στα διάφορα έγγραφα, για να βαθμολογήσουμε ένα έγγραφο d ως προς ένα ερώτημα q μετράμε την συχνότητα εμφάνισης των όρων του ερωτήματος μέσα στο έγγραφο και την ελέγχουμε με αυτή που προκύπτει από την μηδενική υπόθεση. Εάν η διαφορά μεταξύ των συχνοτήτων που παρατηρήθηκαν (observed) και των αναμενόμενων συχνοτήτων (expected) είναι μεγάλη, τότε μπορούμε να θεωρήσουμε ότι το συγκεκριμένο έγγραφο ενδεχόμενα να είναι ″πολωμένο″ (biased) έναντι του ερωτήματος και να του δώσουμε επομένως ένα μεγαλύτερο βαθμό συνάφειας με το ερώτημα.

Εκτελώντας ένα X-τετράγωνο στατιστικό έλεγχο μπορούμε να εκτιμήσουμε την εγκυρότητα της υπόθεσης σχετικά με την τυχαιότητα εμφάνισης των όρων του ερωτήματος στα διάφορα έγγραφα της συλλογής και να τα βαθμολογήσουμε επομένως

σύμφωνα με τις προκύπτουσες αντίστοιχες X-τετράγωνο τιμές.

Στις επόμενες ενότητες αυτού του κεφαλαίου δίνουμε μια σύντομη περιγραφή των X-τετράγωνο στατιστικών έλεγχων "καλού ταιριάσματος", συνεχίζουμε με την περιγραφή της προτεινόμενης μεθόδου αναζήτησης εγγράφων και τελειώνουμε με τα πειραματικά αποτελέσματα της εκτίμησης της αποδοτικότητας της μεθόδου.

2.2 Στατιστικοί Έλεγχοι "Καλού Ταιριάσματος"

Ο σκοπός των έλεγχων καλού ταιριάσματος στην στατιστική είναι να επαληθεύσει ή όχι την υπόθεση ότι τα πειραματικά δεδομένα παράγονται από μια τυχαία μεταβλητή της οποίας η κατανομή είναι πολύ καλά γνωστή. Αυτό είναι ένα πολύ σημαντικό πρόβλημα τόσο στην θεωρητική όσο και πειραματική επιστημονική ανάλυση. Ο στατιστικός πρέπει να αποφασίσει εάν η πειραματική και θεωρητική κατανομή ακολουθούν τον ίδιο νόμο, εάν δηλαδή παράγονται από τον ίδιο φυσικό μηχανισμό. Με άλλα λόγια το πρόβλημα ανάγεται στην επιλογή μιας από τις δύο ενναλλακτικές υποθέσεις: Της μηδενικής υπόθεσης (null hypothesis) H_0, η οποία θεωρεί ότι το δείγμα ακολουθεί την υποκείμενη θεωρητική κατανομή και της ενναλακτικής υπόθεσης (alternative hypothesis) H_1, η οποία θεωρεί ότι αυτό δεν συμβαίνει. Ενας έλεγχος θεωρείται ισχυρός εάν η πιθανότητα της αποδοχής της H_0 είναι χαμηλή όταν η H_0 είναι λάθος.

Ο πιο σημαντικός αλλά και το πιο γνωστός στατιστικός έλεγχος είναι ο X^2-τετράγωνο έλεγχος που προτάθηκε από τον Pearson (Pearson's Chi-squared test) [33]. Προτάθηκε βασικά για την μελέτη διακριτών κατανομών αλλά μπορεί να φανεί χρήσιμος και για συνεχείς κατανομές, εάν τα δεδομένα ομαδοποιηθούν κατάλληλα σε κλάσεις.

Για τον υπολογισμό του X^2-τετράγωνο ελέγχου η στατιστική που χρησιμποιείται είναι η εξής:

$$X^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i \tag{2.1}$$

Όπου O_i ειναι η παρατηρηθείσα συχνότητα για την κλάση ή την κατηγορία i και E_i είναι η αναμενόμενη συχνότητα (expected frequency) που προκύπτει από την

υποτιθέμενη θεωρητική κατανομή πού δεχόμαστε οτι ισχύει για τα δεδομένα μας. Η αναμενόμενη συχνότητα υπολογίζεται απο τον τύπο $E_i = N(F(Y_u) - F(Y_l))$ όπου F είναι η αθροιστική συνάρτηση κατανομής για την κατανομή η οποία ελέγχεται, Y_u είναι το άνω όριο της κλάσης ι, Y_l είναι το κάτω όριο της κλάσης ι, και N το μέγεθος του δείγματος.

Η στατιστική ελέγχου της εξίσωσης 2.1 ακολουθεί κατά προσέγγιση, την X-τετράγωνο κατανομή με $(k - c)$ βαθμούς ελευθερίας όπου k είναι ο αριθμός των κλάσεων και c είναι ο αριθμός των εκτιμώμενων παραμέτρων για την κατανομή σύν ένα. Για παράδειγμα, εάν έχουμε μια κανονική κατανομή (normal distribution) 2 παραμέτρων (m, s), τότε $c = 3$.

Χρησιμοποιώντας κάποιο στατιστικό πακέτο λογισμικού ή πίνακες της X-τετράγωνο κατανομής υπολογίζουμε την πιθανότητα p για την υπολογιζόμενη X-τετράγωνο τιμή από την εξίσωση 2.1 και απορρίπτουμε την μηδενική υπόθεση H_0 εάν το p είναι πολύ μικρό (τυπικά κάτω από ένα επίπεδο σημαντικότητας a) ή διαφορετικά διατηρούμε την H_0. Ο έλεγχος είναι ευαίσθητος στην εκλογή του αριθμού των κλάσεων και για να μπορεί να εφαρμοσθεί θα πρέπει οι ″μετρήσεις″ για τις θεωρητικές συχνότητες να μην είναι μικρότερες του 5.

2.3 Μέθοδος Αναζήτησης Πληροφορίας με την Χρήση Στατιστικού Ελέγχου

Η ουσία της προτεινόμενης μεθόδου είναι να συγκρίνει τις συχνότητες των όρων του συγκεκριμένου ερωτήματος πού παρατηρήθηκαν στο έγγραφο με τις συχνότητες τις αναμενόμενες από καθαρή τύχη. Η σύγκριση αυτή με την βοήθεια του X-τετράγωνο ελέγχου για στατιστική σημαντικότητα μπορεί να ποσοτικοποιήσει μια διαφορά (discrepancy) , η οποία τελικά να χρησιμοποιηθεί σαν κριτήριο βαθμολόγησης της συνάφειας του εγγράφου με το ερώτημα.

Η μηδενική υπόθεση H_0 είναι, όπως αναφέραμε και πιο πρίν, η παραδοχή ότι οι όροι του ερωτήματος κατανέμονται από τύχη στα διάφορα έγγραφα της συλλογής. Αυτό αποτελεί την θεωρητική παραδοχή για την κατανομή των όρων (Theoretical assumption about the distribution) και η στον αντίποδα υπόθεση H_1 είναι ότι οι όροι δεν κατανέμονται συμπτωματικά δηλαδή δεν ακολουθούν τούς κανόνες της τυχαιότητας (chance

distribution.) Απορρίπτοντας την μηδενική υπόθεση έχουμε μια ένδειξη συνάφειας (relatedness) ή πόλωσης (bias) πού υπάρχει μεταξύ του ερωτήματος και του εγγράφου.

Η μηδενική υπόθεση απορρίπτεται όταν η X^2 τιμή που υπολογίζεται απο την εξίσωση 2.1 είναι μεγαλύτερη από την τιμή που λαμβάνουμε από την X-τετράγωνο κατανομή για ένα επίπεδο σημαντικότητας α, δηλαδή όταν ισχύει $X^2 > X^2_{(\alpha,k-c)}$, όπου $X^2_{(\alpha,k-c)}$ είναι η X-τετράγωνο επι τοις εκατό συνάρτηση σημείων (chi-square percent point function) με $k-c$ βαθμούς ελευθερίας και επίπεδο σημαντικότητας α. Δηλαδή όσο μεγαλύτερη είναι η υπολογιζόμενη X^2 τιμή τόσο ισχυρότερη είναι η ένδειξη για να απορρίψουμε την μηδενική υπόθεση και επομένως να έχουμε μια συσχέτιση (relatedness) μεταξύ ερωτήματος και εγγράφου. Επομένως όσον αφορά την τεχνική μας για την βαθμολόγηση εγγράφων, για την μέτρηση δηλαδή της συνάφειας θα μπορούσαμε να χρησιμοποιήσουμε αυτή καθε εαυτή την X^2 τιμή. Τα έγγραφα δηλαδή με τις μεγαλύτερες αντίστοιχες X^2 τιμές θα βαθμολογηθούν και θα καταταχθούν στις πρώτες θέσεις της επιστρεφόμενης βαθμολογημένης λίστας με τα σχετικά για το ερώτημα έγγραφα.

Επειδή για τούς σκοπούς της Αναζήτησης Πληροφορίας, δεν ενδιαφερόμαστε να απορρίψουμε την μηδενική υπόθεση σε ένα επίπεδο σημαντικότητας, δεν ενδιαφερόμαστε για τον αριθμό των κλάσεων στα οποία θα ομαδοποιηθούν τα δεδομένα μας, ούτε ενδιαφερόμαστε να πετύχουμε μια τέτοια ομαδοποίηση ώστε να ικανοποιήσουμε τον ελάχιστο αριθμό 5 των αναμενόμενων συχνοτήτων. Για τούς παραπάνω λόγους δεν υπολογίζουμε την $X^2_{(\alpha,k-c)}$ τιμή από την X-τετράγωνο κατανομή. Η υπολογιζόμενη X^2 τιμή από την εξίσωση 2.1 είναι αρκετή για τούς σκοπούς του information retrieval.

Επειδή εδώ τα δεδομένα μας είναι λέξεις κειμένου, υποθέτουμε ότι υπάρχουν τόσες κλάσεις όσος ακριβώς και ο αριθμός των λέξεων στο ερώτημα, και προσπαθούμε να υπολογίσουμε για κάθε όρο του ερωτήματος την αναμενόμενη και την παρατηρηθείσα συχνότητα. Εάν δηλώσουμε την συχνότητα του όρου w στο έγγραφο d σαν $c(w;d)$, την συχνότητα του όρου στην συλλογή C σαν $c(w;C)$, την συνολική συχνότητα όλων των όρων στο document και στην συλλογή σαν $\sum_w c(w;d)$ και $\sum_w c(w;C)$ αντίστοι-

χα, θα έχουμε για τις αναμενόμενες και παρατηρηθείσες συχνότητες αντίστοιχα τούς ακόλουθους τύπους.

$$E_w = \sum_{w'} c(w'; d) \frac{c(w; C)}{\sum_{w''} c(w''; C)} \tag{2.2}$$

$$O_w = c(w; d) \tag{2.3}$$

Συμβολίζοντας την συχνότητα εμφάνισης ενός όρου του ερωτήματος k_i μέσα σε ένα έγγραφο d με $tf_i = c(k_i; d)$, την συχνότητα του όρου στην συλλογή $\mathcal{C}$ με $ctf_i = c(k_i; \mathcal{C})$, την συνολική συχνότητα όλων των όρων στο έγγραφο και στην συλλογή με $D = \sum_{k_i} c(k_i; d)$ και $C = \sum_{k_i} c(k_i; \mathcal{C})$ αντίστοιχα, τότε οι αναμενόμενες και οι παρατηρηρηθείσες συχνότητες θα δίνονται απο τους ακόλουθους τύπους:

$$E_{k_i} = D \frac{ctf_i}{C} \tag{2.4}$$

$$O_{k_i} = tf_i \tag{2.5}$$

Ο τύπος της προτεινόμενης μεθόδου για την βαθμολόγηση της σχετικότητας των εγγράφων, η εξίσωση 2.6, προκύπτει από την αντικατάσταση των εξισώσεων 2.4 και 2.5 στην 2.1.

$$S(q, d) = \sum_i \frac{(tf_i C - ctf_i D)^2}{ctf_i CD} \tag{2.6}$$

το άθροισμα αναφέρεται πάνω σε όλους τους όρους του ερωτήματος q.

Η διαφορά O_{k_i}-E_{k_i} μπορεί να θεωρηθεί σαν μια βαθμολογία αναζήτησης που λαμβάνεται από κάθε όρο του ερωτήματος δεδομένου του εγγράφου d. Η βαθμολογία αυτή υψώνεται στο τετράγωνο και έπειτα κανονικοποιείται διαιρούμενη με την αναμενόμενη τιμή E_{k_i} για αυτό τον όρο του ερωτήματος. Αυτές οι κανονικοποιημένες διαφορές έπειτα αθροίζονται για να λάβουμε την τελική βαθμολογία αναζήτησης για το έγγραφο d.

Το κύριο πλεονέκτημα αυτής της μεθόδου είναι ότι είναι ένας απλός μη παραμετρικός τρόπος για την βαθμολογία των εγγράφων στην συλλογή. Σε άλλες παραμετρικές

μεθόδους, όπως στην $KL-Divergence$ για παράδειγμα το παραγόμενο μοντέλο αναζήτησης χρειάζεται να υποστηριχθεί από διάφορες άλλες μεθοδολογίες όπως είναι η εκμάθηση των παραμέτρων με εκπαίδευση πάνω σε πραγματικά δεδομένα ($training\ data$).

Στα επόμενα θα ελέγξουμε την αποδοτικότητα της προτεινόμενης X^2 μεθόδου και θα συγκρίνουμε τα αποτελέσματα με τα αποτελέσματα ενός απο τα κλασικά συστήματα στην Αναζήτηση Πληροφορίας, το $tf-idf$ σχήμα, καθώς και με την $KL-Divergence$ που θεωρείται από τις πιο αποδοτικές μεθόδους. Περιγράφουμε πρακάτω τα $tf-idf$ και $KL-Divergence$ συστήματα πού θα χρησιμοποιήσουμε για σύγκριση με την προτεινόμενη X^2 μέθοδο.

2.4 Τα $TFIDF$ Συστήματα Αναζήτησης και η $KL-Divergence$ σαν Βαση Σύγκρισης

2.4.1 $TFIDF$ σχήματα και $OKAPI$ Τύπος Αναζήτησης Πληροφορίας

Τα $tf-idf$ συστήματα Αναζήτησης Πληροφορίας όπως αναφέραμε και στο κεφάλαιο 1 αποκαλούνται και μοντέλα διανυσματικού χώρου και προτάθηκαν για πρώτη φορά από τον Salton το 1971 [2]. Σύμφωνα με αυτό το μοντέλο κάθε όρος k_i σε ένα έγγραφο d_j συνδέεται με ένα θετικό βάρος w_{ij} το οποίο εκφράζει το πόσο σημαντικός είναι ο όρος για τον καθορισμό της σημασιολογίας του εγγράφου και επομένως της σπουδαιότητάς του στο σύστημα αναζήτησης. Επίσης και στούς όρους του ερωτήματος αποδίδεται ένα βάρος. Εάν αναπαραστήσουμε το έγγραφο d_j σαν το διάνυσμα $(w_{1,j}, w_{2,j}, ..., w_{t,j})$ και το ερώτημα q σαν το διάνυσμα $(w_{1,q}, w_{2,q}, ..., w_{t,q})$, όπου t είναι ο συνολικός αριθμός των όρων στο σύστημα, τότε η σχετικότητα του ερωτήματος και του εγγράφου μπορεί να μετρηθεί από την εξίσωση 2.7.

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2.7}$$

Για το βάρος του όρου i στο έγγραφο j χρησιμοποιείται ένας συνδυασμός της συχνότητας του όρου στο έγγραφο και στην συλλογή υπό λογαριθμική κλίμακα (βλέπε κεφάλαιο 1).

$$w_{i,j} = \begin{cases} a(1 + log(tf_{i,j}))log\frac{N}{df_i} & \text{εάν } tf_{i,j} \geq 0 \\ 0 & \text{εάν } tf_{i,j} = 0 \end{cases} \tag{2.8}$$

Όπου, $tf_{i,j}$ είναι η συχνότητα του όρου i στο έγγραφο j (term frequency) , df_i είναι ο αριθμός των εγγράφων στα οποία περιέχεται ο όρος i (document frequency) και N ο αριθμός των εγγράφων στην συλλογή.

Για τις ανάγκες της εκτίμησης της αποδοτικότητας και για να αυξήσουμε την αποδοτικότητα του $tf - idf$ συστήματος, να γίνει δηλαδή πιο ανταγωνιστικό, χρησιμοποιήσαμε μια ελαφρά παραλλαγή του βάρους συγκριτικά με αυτό που δίνεται απο την εξίσωση 2.8, τον OKAPI TF τύπο γνωστό και ως BM25 τύπο για το βέλτιστο ταίριασμα (best matching OKAPI retrieval formula) [49]. Ενώ ο Okapi TF τύπος σχεδιάστηκε για να χρησιμοποιηθεί με το OKAPI πιθανοτικό σύστημα, έχει αποδειχθεί ότι και όταν χρησιμοποιείται με το διανυσματικό μοντέλο δίνει καλύτερα αποτελέσματα Αναζήτησης [66].

Αυτός ο τύπος Αναζήτησης δίνεται από την εξίσωση 2.9.

$$w = ln\frac{N - df + 0.5}{df + 0.5} \cdot \frac{(k_1 + 1)tf}{(k_1 * ((1 - b) + b\frac{dl}{avdl})) + tf} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf} \tag{2.9}$$

όπου tf είναι η συχνότητα του όρου στο έγγραφο, qtf η συχνότητα του όρου στο ερώτημα, N ο συνολικός αριθμός των εγγράφων στην συλλογή, df ο αριθμός των εγγράφων που περιέχουν τον όρο, dl το μήκος του εγγράφου (σε $bytes$), $avdl$ το μέσο μήκος του εγγράφου και τέλος, k_1 μεταξύ 1.0-2.0, b (συνήθως 0.75), k_3 μεταξύ 0-1000 είναι σταθερές.

2.4.2 *KL-Divergence*

Η $KL - Divergence$ [40] είναι μια ιδιαίτερα αποδοτική μέθοδος αναζήτησης πληροφορίας η οποία επεκτείνει την προσέγγιση των μοντέλων γλώσσας στην περιοχή της αναζήτησης πληροφορίας. Η βασική ιδέα έγγειται στην εκτίμηση ενός μοντέλου γλώσσας για το ερώτημα και ενός μοντέλου γλώσσας για το έγγραφο και έπειτα να τα συγκρίνει με την $Kullback - Leibler\ divergence$. Δοθέντων δύο μοντέλων γλώσσας

$p(x)$ και $q(x)$ τότε η $Kullback-Leibler\ divergence$ ή η σχετική εντροπία μεταξύ δύο συναρτήσεων πυκνότητας πιθανότητας δίνεται από τον τύπο 2.10

$$D(p||q) = \sum_x p(x)\frac{p(x)}{q(x)} \tag{2.10}$$

Αν και δεν είναι μια πραγματική απόσταση (δεν είναι συμμετρική και δεν ισχύει η τριγωνική ανισότητα θεωρείται ένα καλό μέτρο σύγκρισης μεταξύ δύο κατανομών πιθανοτήτων. Στην $KL-divergence$, υποθέτουμε ότι το ερώτημα q αναπαράγεται από ένα γεννητικό μοντέλο $p(q|\theta_q)$ και το έγγραφο d από το $p(w|\theta_d)$, όπου θ_q και θ_d δηλώνουν τις παραμέτρους του $unigram$ μοντέλου γλώσσας. Η σχετικότητα του εγγράφου d σε σχέση με το ερώτημα q μπορεί να μετρηθεί από την ακόλουθη αρνητική $KL-divergence$ συνάρτηση 2.11:

$$-D(\theta_q||\theta_d) = \sum_w p(w|\theta_q)\log p(w|\theta_d) + (-\sum_w p(w|\theta_q)\log p(w|\theta_q)) \tag{2.11}$$

Ο δεύτερος από τα δεξιά όρος στον τύπο είναι μια σταθερά εξαρτώμενη από το ερώτημα, η καλύτερα από την εντροπία του μοντέλου του ερωτήματος θ_q, και μπορεί να αγνοηθεί γιατι δεν επηρεάζει την βαθμολόγηση της σχετικότητας των εγγράφων. Στον ίδιο τύπο η σχετικότητα του εγγράφου d σε σχέση με το ερώτημα q εξαρτάται από την εκτίμηση του μοντέλου του ερωτήματος $p(w|\theta_q)$ και του μοντέλου γλώσσας $p(w|\theta_d)$.

Στην παρούσα διατριβή, για την εκτίμηση του μοντέλου γλώσσας $p(w|\theta_d)$ χρησιμοποιούμε μια τεχνική παρεμβολής η οποία κάνει μια γραμμική παρεμβολή της εκτίμησης της μέγιστης πιθανοφάνειας $p_{ml}(w|d) = \frac{c(w;d)}{\sum_{w'} c(w';d)}$ σε σχέση με το μοντέλο συλλογής $\frac{c(w;C)+1}{V+\sum_{w'} c(w';C)}$, χρησιμοποιώντας μια παράμετρο λ για να ελέγξουμε την επίδραση του κάθε μοντέλου [39]. Όπου $c(w;d)$ αναπαριστά την συχνότητα του όρου w στο έγγραφο d, $\sum_{w'} c(w';d)$ τη συνολική συχνότητα των όρων στο έγγραφο d και V είναι το εκτιμώμενο μέγεθος του λεξικού της συλλογής. Το τελικό $smoothing$ μοντέλο συνοψίζεται στην εξίσωση 2.12.

$$p(w|d) = \begin{cases} (1-\lambda)p_{ml}(w|d) + \lambda p(w|C) & \text{εάν η λέξη } w \text{ υπάρχει στο έγγραφο } d \\ \lambda\frac{c(w;C)+1}{V+\sum_{w'} c(w';C)} & \text{διαφορετικά} \end{cases} \tag{2.12}$$

Η απλούστερη μέθοδος για τον υπολογισμό του μοντέλου του ερωτήματος είναι να χρησιμοποιήσει κανείς τον εκτιμητή μέγιστης πιθανοφάνειας $p_{ml}(w|\theta_q) = c(w,q)/|q|$ για τις λέξεις στο κείμενο του ερωτήματος. Ένας πιο ενδιαφέρων σημαντικός τρόπος για τον υπολογισμό του μοντέλου του ερωτήματος στην $KL-Divergence$ μέθοδο είναι να εκμεταλευτεί κανείς έγγραφα από ανατροφοδότηση ($feedback$), χρησιμοποιώντας μια παρεμβολή του εκτιμητή μέγιστης πιθανοφάνειας $p_{ml}(w|\theta_q)$ μέ ένα μοντέλο ανατροφοδότησης $p(w|\theta_F)$ που εκτιμάται από ανατροφοδοτύμενα έγγραφα.

Γενικά, ενσωματώνοντας έγγραφα με ανατροφοδότηση είναι μια δημοφιλής τεχνική για να επεκτείνουμε ($expanding$) τους όρους το ερωτήματος στην Αναζήτηση Πληροφορίας, αλλά η εφαρμογή αυτής της τεχνικής ξεφεύγει από τους σκοπούς της παρούσης διατριβής.

2.5 Εκτίμηση του X^2 Συστήματος Αναζήτησης Πληροφορίας

Στα παραδοσιακά συστήματα αναζήτησης πληροφορίας, τα έγγραφα στην συλλογή παραμένουν στατικά ενώ τα ερωτήματα (queries) εισέρχονται στο σύστημα και το σύστημα απαντά με μια βαθμολογημένη λίστα από σχετικά έγγραφα. Αυτό είναι γνωστό σαν *ad hoc* retrieval πρόβλημα πάνω στο οποίο θα ελέγξουμε την αποδοτικότητα της προτεινόμενης X^2-GOF μεθόδου Αναζήτησης Πληροφορίας. Σε αυτή την ενότητα θα περιγράψουμε πρώτα τα δεδομένα πάνω στα οποία θα γίνει ο έλεγχος της αποτίμησης της προτεινόμενης μεθόδου, και έπειτα τα πειραματικά αποτελέσματα καθώς και την σύγκριση με δύο δημοφιλείς μεθόδους Αναζήτησης, τις $tf-idf$ $OKAPI$ μέθοδο και την KL-Divergence .

Περιγράφουμε στην επόμενη ενότητα τα $TREC$ δεδομένα.

2.5.1 Περιγραφή των *TREC* Δεδομένων για Έλεγχο Αποτίμησης

Μια συλλογή αναφοράς πού χρησιμοποιείται επί χρόνια για την αποτίμηση της αποδοτικότητας (evaluation) των συστημάτων Αναζήτησης Πληροφορίας (information retrieval systems), είναι η TIPSTER/TREC collection [44], η οποία λόγω του μεγάλου της μεγέθους και των πολυάριθμων πειραμάτων που έχουν γίνει με αυτή, θεωρείται σήμερα σαν στάνταρτ στην περιοχή της Αναζήτησης Πληροφορίας .

Το 1990, υπό την καθοδήγηση της Domna Harman μιας διευθύντριας στο National Institute of Standards and Technology (NIST) στο Maryland, ξεκίνησε μια προσπάθεια που σκοπό είχε να προωθήσει (promotion) την ιδέα της διοργάνωσης σε ετήσια βάση ενός διαγωνισμού, υπό το όνομα TREC από το Text Retrieval Conference, για information retrieval συστήματα πού θα διαγωνίζονται πάνω σε συλλογές με εκατομμύρια έγγραφα.

Ετσι λοιπόν ξεκίνησε μια σειρά από διασκέψεις (TREC conferences) με συνδιοργανωτές το National Institute of Standards (NIST), το Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) σαν μέρος του TIPSTER Text Program. Σε κάθε τέτοια conference η οποία φέρει το όνομα TREC conference και το έτος που πραγματοποιείται (πχ, TREC conference 2003) σε κάθε διαγωνιζόμενο στο retrieval σύστημα του δίδεται η ίδια test collection η οποία αποτελείται από περίπου 2 gigabytes από κείμενο (πάνω από ένα εκατομμύριο έγγραφα) και τους δίνεται και μια σειρά από topics (queries) τα οποία εκφράζουν πληροφοριακές ανάγκες για retrieval.

Επειδή οι συλλογές αυτές δημιουργήθηκαν υπό το χρηματοδοτούμενο από το DARPA TIPSTER πρόγραμμα αναφαφέρονται και σαν TIPSTER ή TIPSTER/TREC test collection.

Η TREC συλλογή αυξάνεται σταθερά από χρόνο σε χρόνο και σήμερα διατίθενται επί αγορά σε έξι CD-ROM's πού το καθένα περιέχει χονδρικά περίπου 1 gigabyte συμπιεσμένο κείμενο. Αναφέρουμε παρακάτω μερικές από τις πηγές από τις οποίες προέρχονται τα κείμενα αυτά:

- WSJ → Wall Street Journal
- AP → Associated Press (news wire)

- ZIFF → Computer Selects (articles), Ziff-Davis
- FR → Federal Register
- DOE → Us DOE Publications (abstracts)
- FBIS → Foreign Broadcast Information Service

Τα έγγραφα σε όλες τις υποσυλλογές έχουν επισημειωθεί (tagged) με την γλώσσα SGML για εύκολο parsing. Οι πιο μεγάλες δομές, όπως ένα πεδίο που αναφέρεται στον αριθμό εγγράφου (καθορίζεται με <DOCNO>) και ένα πεδίο για το κείμενο του εγγράφου (καθορίζεται με <TEXT>) είναι κοινές για όλα τα έγγραφα.

Ενα παράδειγμα TREC εγγράφου είναι το παρακάτω (πιο συγκεκριμένα ένα τμήμα απο το κείμενό του για λόγους οικονομίας) παρμένο από την συλλογή *EFILES* Disk 3.

<DOC>
<DOCNO> CR93E-2713 < /DOCNO>
<DOCID> E08OC4-299 < /DOCID>
<CENTER><PRE> [*Page: E2194*] *< /PRE>< /CENTER>*
<DATE>19941008< /DATE>
<FLD003>E< /FLD003>
<TEXT>
<TTL>IN RECOGNITION OF U.S. REP. HELEN DELICH BENTLEY – HON. FRANK R. WOLF (Extension of Remarks - October 08, 1994)< /TTL>
<CENTER>HON. FRANK R. WOLF< /CENTER>
<CENTER>OF VIRGINIA< /CENTER>
<CENTER>in the House of Representatives< /CENTER>
<CENTER>Friday,October 7, 1994< /CENTER>
Mr. WOLF. Mr. Speaker, it is a pleasure for me to join with my colleagues in recognizing the outstanding service to Congress and our nation of my friend and colleague, Helen Delich Bentley. For the last 10 years, Helen has been the beloved representative for Maryland's 2nd District, which stretches all the way from southeastern Baltimore County, near the Port of Baltimore, to the northern borders of Maryland, near the Delaware and Pennsylvania borders. Needles to say, her constituency is

diverse and varied; nevertheless Helen has earned the respect and admiration of her constituents, who have re-elected her the last three times impressively with over 65 percent of the vote.
<FLD001>E02194< /FLD001>
<FLD002>WOLF< /FLD002>
< /TEXT>
< /DOC>

Η TREC συλλογή συμπεριλαμβάνει ένα σύνολο από ερωτήματα (queries), που δεν είναι τίποτα άλλο από παραδείγματα αιτημάτων για κάποια πληροφοριακή ανάγκη, και που με αυτά μπορεί να ελεγχθεί ένας νέος ranking αλγόριθμος ως πρός την αποδοτικότητά του στο retrieval. Στην TREC ορολογία ένα αίτημα για πληροφορία (query) ονομάζεται *topic*. Ένα παράδειγμα *topic* είναι το παρακάτω:

<top>
<num> Number: 301
<title> International Organized Crime
<desc> Description: Identify organizations that participate in international criminal activity, the activity, and, if possible, collaborating organizations and the countries involved. <narr> Narrative: A relevant document must as a minimum identify the organization and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organization(s) involved would not be relevant. < /top>

Στα topics περιλαμβάνονται συνήθως ο τιτλος "Title" που αποτελείται από μια έως τρεις λέξεις - κλειδιά που χαρακτηρίζουν την πληροφοριακή ανάγκη, η περιγραφή "Description" που είναι μια περιγραφή που αποτελείται από μια έως δύο προτάσεις και η αφήγηση "Narrative" πού είναι πιο αναλυτική περιγραφή που συνήθως περιλαμβάνει και χαρακτηριστικά παραδείγματα πού ικανοποιούν την συγκεκριμένη πληροφοριακή ανάγκη.

Για κάθε ένα *topic* δημιουργείται και ένα σύνολο από σχετικά έγγραφα (relevant

documents) τα οποία χρησιμοποιούνται σαν αναφορά για την εκτίμηση της αποδοτικότητας των συστημάτων στο retrieval. Το σύνολο αυτό λαμβάνεται από μια δεξαμενή από έγγραφα, η οποία δεξαμενή δημιουργείται από τα K πιο υψηλά βαθμολογημένα έγγραφα (συνήθως $K = 100$) πού παράγονται από τα διάφορα συμμετέχοντα διαγωνιζόμενα συστήματα. Τα έγγραφα της δεξαμενής επιδεικνύονται κατόπιν σε ανθρώπους εκτιμητές πού έχουν τελικά τον τελευταίο λόγο για την σχετικότητα η όχι ενός εγγράφου.

2.5.2 Σύγκριση με τα $tf - idf$ σχήματα - $OKAPI$ μέθοδος

Οπως είναι γνωστό η αποδοτικότητα ενός συγκεκριμένου συστήματος αναζήτησης πληροφορίας εξαρτάται απο τα δεδομένα που έχουν επιλεγεί για έλεγχο. Επίσης μπορεί να μεταβάλλεται από συλλογή σε συλλογή. Για να έχουμε μια καλύτερη εκτίμηση της δυνατοτήτας για Αναζήτηση του X^2 αλγορίθμου μας, επιλέξαμε να γίνει ο έλεγχος σε τρεις μεγάλες υποσυλλογές απο την συλλογή δεδομένων TREC για έλεγχο, [44]. Οι συλλογές αυτές είναι οι ακόλουθες:

- FBIS συλλογή εγγράφων από τον δίσκο 5, μέγεθος αρχείου περίπου 490 MB.
- EFILES συλλογή εγγράφων από τον δίσκο 4, μέγεθος αρχείου περίπου 50 MB.
- LATIMES συλλογή εγγράφων από τον δίσκο 5, μέγεθος αρχείου περίπου 45 MB.

Στόν πίνακα 2.1 εμφανίζονται τα στατιστικά στοιχεία που αφορούν τις συλλογές, όπως ο αριθμός των εγγράφων στην συλλογή, ο αριθμός των όρων, ο αριθμός των διακριτών λέξεων στο λεξικό της συλλογής, το μέγιστο μήκος του εγγράφου και τέλος το μέσο μήκος του εγγράφου.

Ως ερωτήματα χρησιμοποιήσαμε τα θέματα 351-340 (topics 351-340), τα οποία χρησιμοποιήθηκαν και στο συνέδριο TREC 7 για συστήματα αναζήτησης πληροφορίας. Όπως αναμένεται ο X^2 έλεγχος ″καλού ταιριάσματος″ θα πρέπει να αποδίδει καλύτερα με ερωτήματα πού αποτελούνται από περισσότερους όρους. Για να ελέγξουμε αυτό το ενδεχόμενο, αλλά και τη δυνατότητα της προτεινόμενης μεθόδου στα

	fbis	efiles	latimes
Num of documents	130,471	11,358	12,423
Num of terms	403,672	83,416	98,789
Num of words	71,780,110	7,910,399	6,707,435
Max document length	143,651	221,130	26,072
Average Doc length	550.16	696.46	539.92

Πίνακας 2.1: *Στατιστικά στοιχεία των 3 συλλογών που χρησιμοποιήθηκαν για έλεγχο*

σύντομα ερωτήματα (μικρός αριθμός λέξεων), εκτελέσαμε ξεχωριστά δύο πειράματα αποτίμησης: Ένα πείραμα χρησιμοποιώντας μόνο τους τίτλους από τα ερωτήματα 351-400 του συνεδρίου TREC 7 και ένα άλλο πείραμα με μια μεγαλύτερη έκδοση των παραπάνω ερωτημάτων. Η τελευταία έκδοση των ερωτημάτων θεωρείται ότι είναι πιο κοντά στα ερωτήματα των χρηστών που υποβάλλονται σε ένα πραγματικό σύστημα.

Για όλα τα πειράματα που εκτελέσαμε δεν εφαρμόσαμε καμμιά προεπεξεργασία στα κείμενα, όπως πχ, tokenization, stemming, ούτε εφαρμόσαμε καμμία λίστα αποκλεισμού συχνών λέξεων (stopword list), όπως άρθρων, συνδέσμων, επιρρημάτων κλπ. Αντίθετα λάβαμε υπ'οψιν όλες ανεξαίρετως τις λέξεις όλων των εγγράφων στην συλλογή.

Και στα δύο διαγωνιζόμενα συστήματα, δηλαδή το X^2GOF και το $tf - idf$, χρησιμοποιήσαμε σαν μέτρο σύγκρισης την μέση ακρίβεια χωρίς παρεμβολή (Average Non Interpolation Precision). Την ακρίβεια αυτή την υπολογίσαμε από τα 1000 έγγραφα με την πιο υψηλή βαθμολογία (top ranked documents) στην επιστρεφόμενη βαθμολογημένη λίστα εγγράφων από τους αλγορίθμους. Αυτό το μέτρο θεωρείται σαν ένα πολύ καλό μέτρο σύγκρισης διότι αντανακλά την συνολική ακρίβεια βαθμολόγησης (overall ranking accuracy) του κάθε αλγορίθμου. Επίσης για τον κάθε αλγόριθμο δίνουμε αριθμητικά τα αποτελέσματα αλλά και γραφήματα για την precision/recall αποδοτικότητα.

Για την αποτίμηση του πρώτου πειράματος, χρησιμοποιήσαμε τον ″Τίτλο″ και την ″Περιγραφή″ από τα ερωτήματα 301-350. (″Titles+Description″), δηλαδή την εκδοχή που περιλαμβάνει τις περισσότερες λέξεις. Στον πίνακα 2.2 παρουσιάζεται η μέση ακρίβεια (average precision) και για τα δύο διαγωνιζόμενα συστήματα, ενώ στον πίνακα 2.3

εμφανίζονται τα αποτελέσματα για την απόδοση precision/recall . Είναι προφανές από την σύγκριση των παραπάνω αποτελεσμάτων, ότι ο X^2 έλεγχος ταιριάσματος αποδίδει πολύ καλύτερα από ότι η $tf - idf$.

Όσον αφορά το δεύτερο πείραμα αποτίμησης, αν και χρησιμοιήσαμε μόνο ″Τίτλους″, δηλαδή την σύντομη έκδοση των 301-350 ερωτημάτων, η X^2 μέθοδος αναζήτησης αποδίδει και εδώ καλύτερα απο το $tf - idf$ σύστημα αναζήτησης.

Τα αποτελέσματα για το δεύτερο πείραμα αποτίμησης φαίνονται στους πίνακες 2.4 και 2.5 για την μέση ακρίβεια (average precision) και την non-interpolaed precision στα 11 *Recall* σημεία αντίστοιχα.

Για καλύτερη οπτική εμφάνιση της αποδοτικότητας δίνουμε γραφήματα και για τα δύο πειράματα πάνω στις 3 συλλογές ελέγχου. Αυτά φαίνονται στις εικόνες 2.1, 2.2, 2.3, 2.4, 2.5, 2.6. Οι 3 πρώτες αφορούν το πρώτο πείραμα με τα ερωτήματα των πολλών λέξεων για τις συλλογές FBIS, EFILES, LATIMES αντίστοιχα και οι 3 τελευταίες εικόνες παρουσιάζουν το δεύτερο πείραμα με τα σύντομα ερωτήματα για τις ίδιες συλλογές.

Θα μπορούσαμε να εκτελέσουμε και ένα τρίτο πείραμα χρησιμοποιώντας το πεδίο ″Narrative″ που υπάρχει στα ερωτήματα. Το ″Narrative″ είναι ένα επιπλέον πεδίο που δίνει μια σύντομη αφήγηση της πληροφοριακής ανάγκης του χρήστη αποτελούμενο από 20-30 λέξεις περίπου. Είναι εύλογο να αναμένουμε σε μια τέτοια περίπτωση την συντριπτική υπεροχή του συστήματος X^2 ελέγχου. Κάτι τέτοιο όμως δεν θεωρούμε ότι ανταποκρίνεται στις συνθήκες μιας πραγματικής εφαρμογής. Σε μια πραγματική εφαρμογή ο χρήστης συνήθως υποβάλλει ένα ερώτημα που αποτελείται απο μερικές μόνο λέξεις.

2.5.3 Σύγκριση με την $KL - Divergence$ μέθοδο στην $TREC$ συλλογή

Για την καλύτερη αποτίμηση των δυνατοτήτων της X^2 τεχνικής στην Αναζήτηση Πληροφορίας εκτελέσαμε ένα μεγαλύτερο πείραμα πάνω σε ολόκληρη την $TREC$ συλλογή

	fbis	efiles	latimes
X^2GOF	0.157	0.1673	0.1951
$tf-idf$	0.068	0.1499	0.1535

Table 2.2: *Αποτελέσματα αποτίμησης Μέση Ακρίβεια (Average Precision) των μεθόδων X^2GOF και $tf-idf$ για τις 3 συλλογές TREC (Ερωτήματα 301-350 "Title + Description" έκδοση)*

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
X^2GOF	0.403	0.272	0.213	0.200	0.175	0.158	0.145	0.114	0.090	0.061	0.048	fbis
$tfidf$	0.241	0.133	0.112	0.087	0.072	0.061	0.052	0.042	0.034	0.027	0.017	
X^2GOF	0.319	0.272	0.260	0.223	0.191	0.173	0.118	0.102	0.093	0.069	0.065	efiles
$tfidf$	0.298	0.294	0.251	0.198	0.177	0.159	0.096	0.089	0.078	0.059	0.052	
X^2GOF	0.296	0.287	0.274	0.245	0.163	0.162	0.158	0.152	0.152	0.151	0.151	latimes
$tfidf$	0.235	0.235	0.235	0.180	0.146	0.146	0.120	0.112	0.112	0.111	0.111	

Table 2.3: *Average non interpolated precision at 11 recall points για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf-idf$ (Ερωτήματα 301-350 "Title + Description" έκδοση)*

	fbis	efiles	latimes
X^2GOF	0.2077	0.1777	0.2025
$tf-idf$	0.1864	0.1715	0.1645

Table 2.4: *Αποτελέσματα Αποτίμησης Μέση Ακρίβεια (Average Precision) των μεθόδων X^2GOF και $tf-idf$ για τις 3 συλλογές TREC (Ερωτήματα 301-350 "Title" έκδοση)*

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
X^2GOF	0.450	0.313	0.279	0.250	0.222	0.209	0.187	0.164	0.151	0.120	0.103	fbis
$tfidf$	0.340	0.291	0.256	0.209	0.195	0.185	0.168	0.147	0.134	0.119	0.096	
X^2GOF	0.315	0.277	0.277	0.248	0.207	0.186	0.129	0.116	0.107	0.090	0.085	efiles
$tfidf$	0.307	0.271	0.268	0.243	0.221	0.195	0.126	0.113	0.101	0.092	0.085	
X^2GOF	0.314	0.313	0.286	0.251	0.167	0.165	0.160	0.156	0.155	0.154	0.153	latimes
$tfidf$	0.244	0.244	0.244	0.199	0.154	0.152	0.132	0.125	0.124	0.122	0.122	

Table 2.5: *Average non interpolated precision στα 11 σημεία recall, για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf-idf$ (Ερωτήματα 301-350 "Title" έκδοση)*

και συγκρίναμε τις αποδοτικότητες και των τριών συστημάτων Αναζήτησης, της X^2 τεχνικής, της $KL-Divergence$, και της $OKAPI$ μεθόδου.

Για το πείραμα αυτό χρησιμοποιήσαμε ολόκληρη την συλλογή των $TREC$ δεδομένων από τα $CD's$ 4 και 5. Αυτά τα δεδομένα περιέχουν 556,000 έγγραφα (γύρω στα 2.1 *Gbytes* δεδομένα) από τις *Congressional Record*, *Federal Register*, *Financial Times*, *Foreign Broadcast Information Service* και *LA Times* συλλογές. Σαν

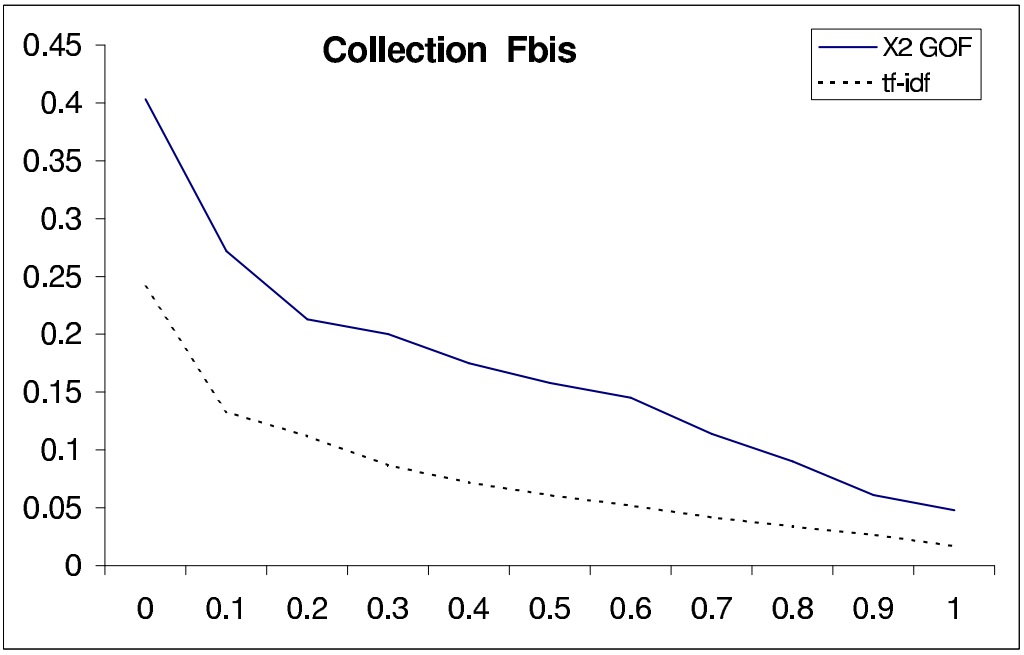


Figure 2.1: *Συλλογή FBIS. Average non interpolated precision στα 11 σημεία recall, για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf - idf$ (Ερωτήματα 301-350 "Title + Description" έκδοση)*

Πίνακας 2.6: *Στατιστικά της TREC συλλογής*

CD's 4 και 5 χωρίς να συμπεριλαμβάνεται το *CR*	
Αριθμός Εγγράφων	528,155
Αριθμός Όρων	254,333,060
Αριθμός Μοναδικών Όρων	731,442
Μέσο Μήκος Εγγράφου	481

ερωτήματα χρησιμοποιήσαμε τα θέματα (*topics*) των $TREC-7$ και $TREC-8$ διαγωνισμών για *adhoc retrieval* (50 *topics* ο κάθε διαγωνισμός). Από την παραπάνω συλλογή παραλείψαμε τα έγγραφα από την υποσυλλογή *Congressional Record*, διότι αυτά αφαιρέθηκαν από τούς $TREC-7$ και $TREC-8$ διαγωνισμούς.

Τα στατιστικά στοιχεία της συλλογής δίνονται από τον πίνακα 2.6

Στα θέματα (*topics*) του $TREC-7$ και $TREC-8$ σε καθένα από τα 50 θέματα χρησιμοποιήσαμε εδώ μόνο το πεδίο "τίτλος" γιατί πιστεύουμε ότι είναι πιο κοντά σε ένα πραγματικό ερώτημα.

Ελέγξαμε την προτεινόμενη $\chi^2 - GOF$ μέθοδο και την συγκρίναμε με την $KL-Divergence$ μέθοδο, καθώς και με την $OKAPI\ BM25$ μέθοδο πάνω σε ολόκληρη

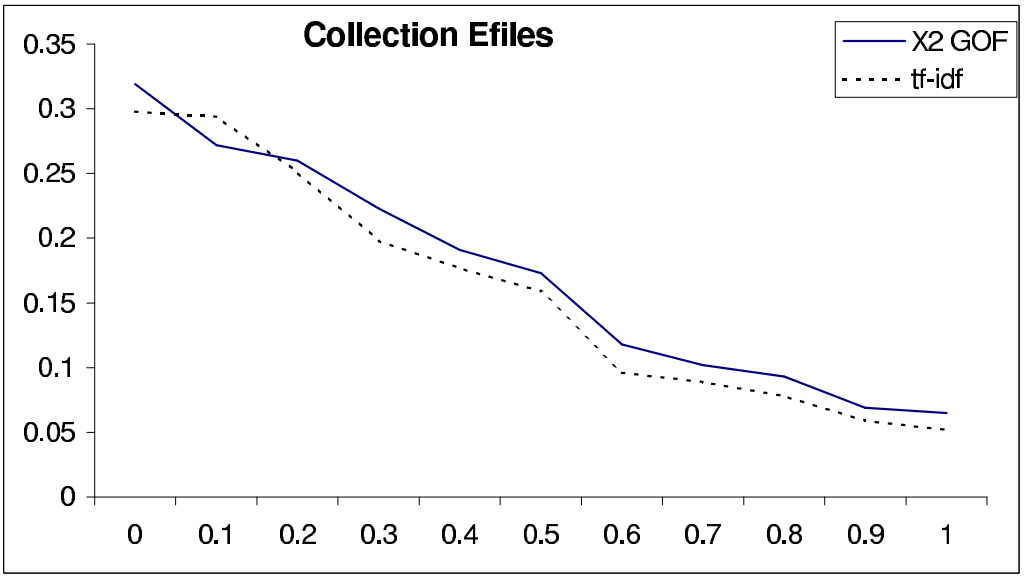


Figure 2.2: *Συλλογή EFILES. Average non interpolated precision στα 11 σημεία Recall για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf-idf$ (Ερωτήματα 301-350"Title+ Description" έκδοση)*

την $TREC$ συλλογή. Τα αποτελέσματα παρουσιάζονται στους πίνακες 2.7 και 2.8 για τα θέματα $TREC-7$ και $TREC-8$ αντίστοιχα.

Αν και στην προτεινόμενη $\chi^2 - GOF$ μέθοδο χρησιμοποιούνται μόνο καθαρές συχνότητες, η μέθοδος ξεπερνά σε απόδοση σταθερά την $OKAPI\ BM25$ μέθοδο, ωστόσο και στις δύο περιπτώσεις $TREC-7$ και $TREC-8$ η $KL-Divergence$ έχει την καλύτερη απόδοση. Η μέθοδος αυτή όμως έχει το μειονέκτημα ότι είναι παραμετρική και χρειάζεται την εκτίμηση των παραμέτρων πάνω σε ολόκληρη την συλλογή.

Η βασική πληροφορία που χρησιμοποιείται στην προτεινόμενη $\chi^2 - GOF$ μέθοδο για την βαθμολόγηση ενός εγγράφου d είναι η συχνότητα tf_i του όρου k_i στο έγγραφο d, η συχνότητα ctf_i του όρου στην συλλογή $\mathcal{C}$, η συνολική συχνότητα D όλων των όρων στο έγγραφο και η συνολική συχνότητα C όλων των όρων στην συλλογή. Είναι σημαντικό να παρατηρήσουμε την διαφορά της προτεινόμενης $\chi^2 - GOF$ μεθόδου και της κλασσικής $OKAPI$ από το μοντέλο των διανυσματικών χώρων. Στην κλασσική οικογένεια των τόσο γνωστών $tf - idf$ σχημάτων ο τύπος για την βαθμολόγηση ενός εγγράφου είναι στενά συνδεδεμένος με την χρήση της λεγόμενης συχνότητας εγγράφου ($document\ frequency$), η οποία εκφράζει τον αριθμό των εγγράφων της συλλογής στα οποία ο όρος k_i εμφανίζεται και μπορεί να διερμηνευθεί σαν ένα μέτρο της πληροφορίας που παρέχει ο όρος. Τα πειραματικά δεδομένα δείχνουν ότι η απουσία

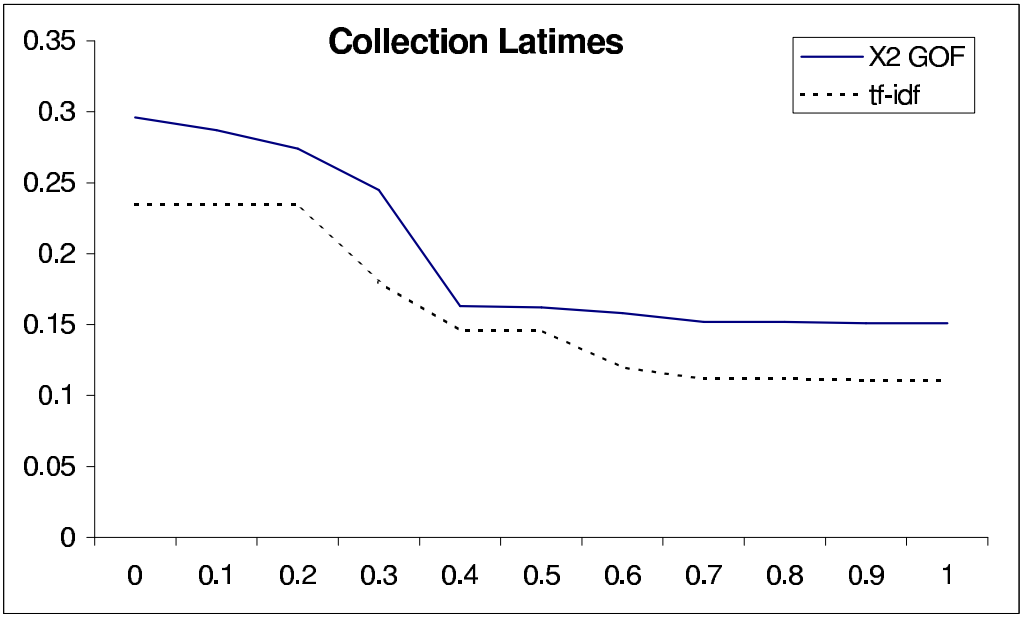


Figure 2.3: *Συλλογή LATIMES. Average non interpolated precision στα 11 σημεία Recall, για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf - idf$ (Ερωτήματα 301-350 "Title + Description" έκδοση)*

αυτής της ποσότητας από την προτεινόμενη μέθοδο δεν επηρεάζει καθόλου την αποδοτικότητά της. Η απλότητα είναι ένα από τα κύρια πλεονεκτήματα της προτεινόμενης $\chi^2 - GOF$ μεθόδου. Χρησιμοποιώντας μόνο καθαρές συχνότητες η υπολογιζόμενη $\chi^2 - GOF$ τιμή βελτώνει την αποδοτικότητα και επιτρέπει την ανεύρεση εγγράφων τα οποία προσεγγίζουν τις συνθήκες του ερωτήματος. Επι πλέον, η μέθοδος μας επιτρέπει να αποφασίσουμε εάν υπάρχει μια στατιστικά σημαντική σχέση μεταξύ του ερωτήματος και του εγγράφου. Εάν η υπολογιζόμενη τιμή είναι μεγάλη συμπεραίνουμε ότι υπάρχει μεταξύ του ερωτήματος και του εγγράφου και η τιμή αυτή επαρκεί για να ταξινομήσει τα έγγραφα σύμφωνα με το βαθμό ομοιότητάς τους ως πρός το ερώτημα.

Οι αποδόσεις των συγκρινόμενων αλγορίθμων σε αυτά τα πειράματα φαίνεται να είναι διαφορετικές. Για να εξετάσουμε πιό τυπικά το βαθμό στον οποίο τα αποτελέσματα αυτά μας παρέχουν μια πειστική απόδειξη για το ότι οι αποδόσεις είναι διαφορετικές, θα μπορούσαμε να εκτελέσουμε ένα έλεγχο $paired\ t - test$. Ο έλεγχος $paired\ t - test$ χρησιμοποιείται για να προσδιορίσουμε εάν οι μέσοι δύο δειγμάτων από δύο πληθυσμούς είναι ίσοι.

Στην περίπτωσή μας αν θεωρήσουμε τις λαμβανόμενες τιμές μέσης Ακρίβειας στα 11 σημεία της κάθε μεθόδου σαν δείγματα πληθυσμών, τότε η υπολογιζόμενη *paired*

Πίνακας 2.7: *Μέση ακρίβεια AvgPrec και μέση ακρίβεια με παρεμβολή στα 11 σημεία recall για τις χ^2-GOF, OKAPI και $KL-Divergence$ μεθόδους (CD's 4 και 5, Ερωτήματα 351-400 "Title" version)*

$CD's$ 4 και 5, Ερωτήματα 351-400, σχετικά έγγραφα 4674			
	$\chi^2 - GOF$	$OKAPI$	$KL - Divergence$
$AvgPrec$	0.1248	0.097	0.1827
0	0.4994	0.4006	0.6074
0.1	0.2871	0.2094	0.3758
0.2	0.2134	0.1447	0.3058
0.3	0.1381	0.1059	0.2415
0.4	0.0904	0.0787	0.1606
0.5	0.062	0.0521	0.1236
0.6	0.037	0.0400	0.0667
0.7	0.0177	0.0157	0.0507
0.8	0.0147	0.0121	0.0386
0.9	0.0099	0.0090	0.0282
1	0.0015	0.0008	0.0110

Πίνακας 2.8: *Μέση ακρίβεια AvgPrec και μέση ακρίβεια με παρεμβολή στα 11 σημεία recall για τις χ^2-GOF, $OKAPI$ και $KL-Divergence$ μεθόδους ($CD's$ 4 και 5, Ερωτήματα 401-450 "Title" version)*

$CD's$ 4 και 5, Ερωτήματα 401-450, σχετικά έγγραφα 4728			
	$\chi^2 - GOF$	$OKAPI$	$KL - Divergence$
$AvgPrec$	0.1270	0.1112	0.2195
0	0.4254	0.4210	0.6460
0.1	0.2421	0.2207	0.4529
0.2	0.1803	0.1573	0.3240
0.3	0.1472	0.1341	0.2595
0.4	0.1088	0.0867	0.1982
0.5	0.0902	0.0680	0.1697
0.6	0.0708	0.0494	0.1280
0.7	0.0528	0.0317	0.0948
0.8	0.0424	0.0251	0.0727
0.9	0.0275	0.0192	0.0501
1	0.0089	0.0103	0.0180

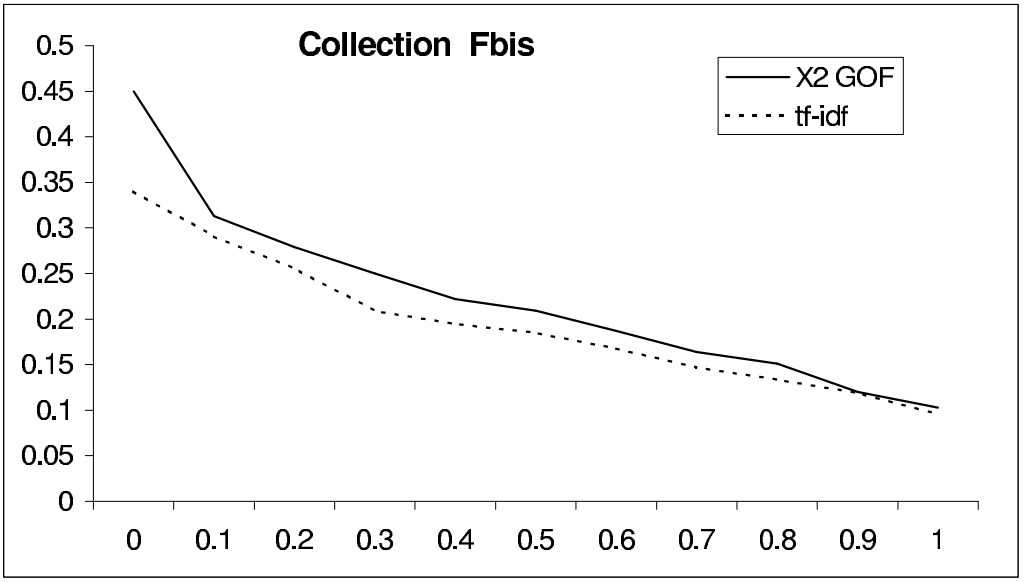


Figure 2.4: *Συλλογή FBIS. Average non interpolated precision στα 11 σημεία Recall για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf-idf$ (Ερωτήματα 301-350 "Title" έκδοση)*

$t-test$ τιμή δίνει την πιθανότητα οι μέσες ακρίβειες να είναι διαφορετικές. Όσο πιο μικρότερη είναι η πιθανότητα τόσο πιο σημαντική είναι η διαφορά μεταξύ των μέσων Ακριβειών.

Εκτελώντας τον έλεγχο *paired* $t-test$ για τα μοντέλα χ^2-GOF και $OKAPI$, τότε επιστρεφόμενη πιθανότητα ($p-value$) για τα θέματα 351-400 είναι 0.0326 και για τα θέματα 401-450 είναι 0.00010608. Επομένως, συμπεραίνουμε οτι οι λαμβανόμενες μέσες ακρίβειες για τα μοντέλα χ^2-GOF και $OKAPI$, είναι διαφορετικές με βεβαιότητα 96.74% και 99.98% για τα θέματα 351-400 και 401-450 αντίστοιχα. Όμοια, συγκρίνοντας τα μοντέλα χ^2-GOF και $KL-divergence$ βρίσκουμε επιστρεφόμενες πιθανότητες 0.0004 για τα θέματα 351-400 και 0.0018 για τα θέματα 401-450.

Όπως αναφέραμε παραπάνω η προτεινόμενη χ^2-GOF μέθοδος Αναζήτησης μας προσφέρει την δυνατότητα να δοκιμάσουμε απλούς εναλλακτικούς τύπους Αναζήτησης, χρησιμοποιώντας διαφορετικές βασικές υποθέσεις για τα δεδομένα. Στην εργασία *divergence from randomness* του *Amati* [67], προτείνεται ένα βασικό μοντέλο τυχαιότητας της κατανομής των όρων στα διάφορα έγγραφα. Σύμφωνα με αυτό οι διαδικασίες κατανομής όρων μπορούν να ορισθούν σαν τυχαίες εκλογές όρων (*random drawings*) από ένα "δοχείο" (*urn*) που περιέχει τους διαθέσιμους όρους.

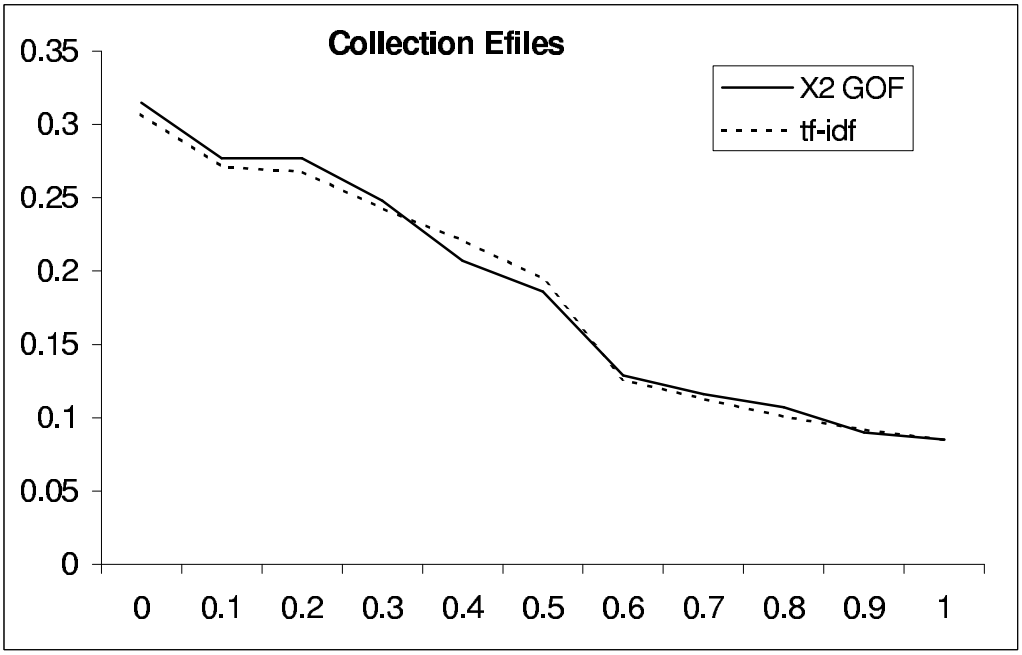


Figure 2.5: *Συλλογή EFILES. Average non interpolated precision στα 11 σημεία Recall, για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf-idf$ (Ερωτήματα 301-350 "Title" έκδοση)*

Ακολουθώντας αυτή την πρόταση, αλλάξαμε το μοντέλο της τυχαιότητας από αυτό της ομοιόμορφης κατανομής στο διωνυμικό μοντέλο (*binomial model*). Σύμφωνα με αυτό το μοντέλο η εμφάνιση ενός μοναδικού όρου i σε ένα έγγραφο d θεωρείται σαν *Bernoulli* διαδικασίαμε με πιθανότητα $p = 1/N$, όπου N ο αριθμός των εγγράφων στην συλλογή. Εάν ctf_i η συχνότητα του όρου i στην συλλογή, τότε μπορούμε να κάνουμε την υπόθεση ότι ο όρος i κατανέμεται πάνω στα N έγγραφα σύμφωνα με το διωνυμικό νόμο. Επομένως η πιθανότητα tf_i εμφανίσεων σε ένα έγγραφο δίνεται από τον τύπο

$$p = B(N, ctf_i, tf_i) = \binom{ctf_i}{tf_i} p^{tf_i} q^{ctf_i - tf_i}, \tag{2.13}$$

όπου $p = 1/N$ και $q = (N-1)/N$.

Ο αναμενόμενος αριθμός των εμφανίσεων ενός όρου σε ένα έγγραφο είναι $\lambda = (1/N)ctf_i$ και είναι σταθερός για όλα τα έγγραφα στην συλλογή. Χρησιμοποιώντας tf_i σαν την παρατηρηθείσα συχνότητα και ctf_i/N σαν την αναμενόμενη συχνότητα, τότε από την εξίσωση 2.1 λαμβάνουμε τον ακόλουθο τύπο Αναζήτησης για την νέα υπόθεση για το

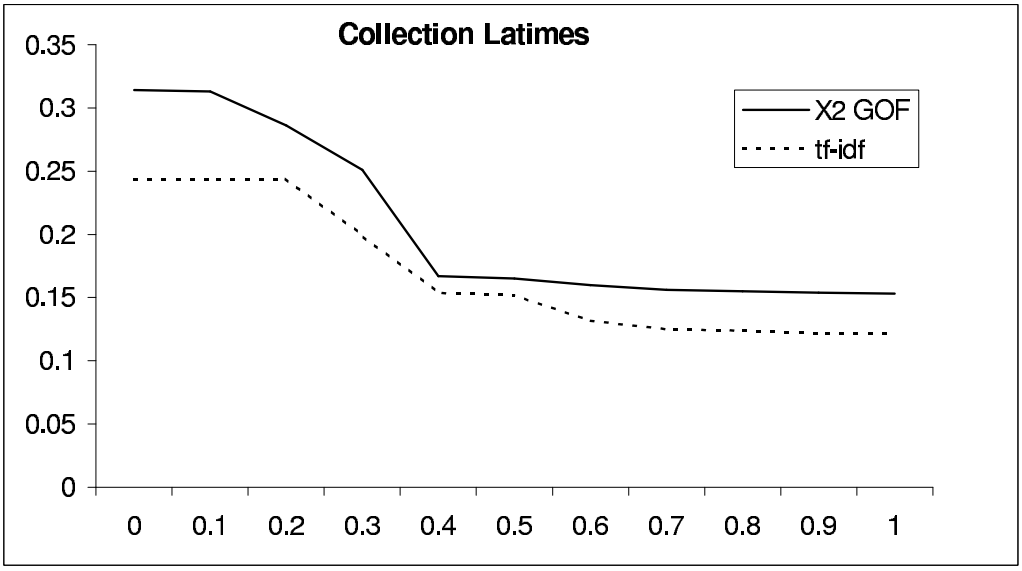


Figure 2.6: *Συλλογή LATIMES. Average non interpolated precision στα 11 σημεία Recall, για τις μεθόδους αναζήτησης πληροφορίας X^2GOF και $tf - idf$ (Ερωτήματα 301-350 "Title" έκδοση)*

μοντέλο τυχαιότητας της κατανομής των όρων

$$S(q,d) = \sum_i \frac{(tf_i N - ctf_i)^2}{N ctf_i}, \tag{2.14}$$

το άθροισμα είναι πάνω σε όλους τους όρους του ερωτήματος q.

Συγκρίναμε την απόδοση αυτού του νέου τύπου Αναζήτησης (διωνυμική κατανομή) με αυτήν του αρχικού τύπου Αναζήτησης (ομοιόμορφη κατανομή), σε ένα δεύτερο πείραμα χρησιμοποιώντας την $fbis$ συλλογή εγγράφων από το CD 5 των $TREC$ δεδομένων.

Τα στατιστικά της $fbis$ συλλογής και τα αποτελέσματα της εκτίμησης της αποδοτικότητας φαίνονται στούς πίνακες 2.9, 2.10 και 2.11.

Από αυτά τα αποτελέσματα συμπεραίνουμε ότι το μοντέλο της ομοιόμορφης κατανομής αποδίδει ελαφρά καλύτερα από το μοντέλο της διωνυμικής κατανομής.

Στατιστικά στοιχεία	*fbis* συλλογή
Αριθμός εγγράφων	130,471
Αριθμός όρων	403,672
Αριθμός λέξεων	71,780,110
Μέγιστο μήκος εγγράφου	143,651
Μέσο μήκος εγγράφου	550.16

Πίνακας 2.9: *Στατισικά στοιχεία της fbis συλλογής εγγράφων από το CD 5 των TREC δεδομένων για έλεγχο*

2.6 Συμπέρασμα

Σε αυτό το κεφάλαιο παρουσιάσαμε μια μέθοδο για την εφαρμογή του X^2 στατιστικού ελέγχου ταιριάσματος στην Αναζήτηση Πληροφορίας. Η εφαρμογή ενός τέτοιου ελέγχου ξεκινάει πάντα με την διατύπωση της μηδενικής υπόθεσης (null hypothesis). Στην συγκεκριμένη περίπτωση διατυπώνουμε τον ισχυρισμό ότι οι όροι ενός ερωτήματος κατανέμονται τυχαία στα διάφορα έγγραφα της συλλογής, δηλαδή από σύμπτωση ακολουθώντας τους νόμους της τυχαιότητας, και ότι δεν υπάρχει καμμία συσχέτιση μεταξύ ερωτήματος και εγγράφου πέραν της συμπτωματικής εμφάνισης των όρων του ερωτήματος στο έγγραφο.

Χρησιμοποιώντας X^2 στατιστικό έλεγχο ταιριάσματος για την εκτίμηση του κατά πόσο τα δεδομένα συμφωνούν στατιστικά με την παραπάνω υπόθεση, ποσοτικοποιούμε μια "διάσταση" (discrepancy) μεταξύ αναμενόμενων και παρατηρηθεισών συχνοτήτων από την υπολογιζόμενη X^2 τιμή. Αυτή η διάσταση χαρακτηρίζει και την συνάφεια μεταξύ ερωτήματος και εγγράφου (relatedness) και χρησιμοποιείται σαν κριτήριο για Αναζήτηση Πληροφορίας.

Η μέθοδος αποδεικνύεται εύρωστη (robust) και αποδοτική αποδίδοντας καλά για πολύ σύντομα ερωτήματα, όσο και για περισσότερο 'φλύαρα' ερωτήματα.

Η εφαρμογή του X^2 στατιστικού ελέγχου ταιριάσματος πού εφαρμόζουμε εδώ έχει το πλεονέκτημα ότι μας επιτρέπει να μοντελοποιήσουμε τα ερωτήματα και τα έγγραφα με διάφορους τρόπους, αποδίδοντας οποιαδήποτε ιδιαίτερη θεωρητική κατανομή για τα δεδομένα και να χρησιμοποιήσουμε έπειτα την δύναμη των στατιστικών ελέγχων ταιριάσματος για να αποτιμήσουμε την κατανομή πού διέπει τα δεδομένα μας.

Κάποιες διαφορετικές υποθέσεις για τα δεδομένα όπως η κανονικότητα (normal-

Πίνακας 2.10: *Μέση Ακρίβεια AvgPrec και μέση ακρίβεια με παρεμβολή στα 11 Recall σημεία για τις μεθόδους χ^2-GOF ομοιόμορφη κατανομή και χ^2-GOF διωνυμική κατανομή (συλλογή εγγράφων fbis, Ερωτήματα 351-400 "Title" version)*

fbis Συλλογή, Ερωτήματα 351-400, Σχετικά έγγραφα 1339		
	$\chi^2 - GOF\ uniform$	$\chi^2 - GOF\ binomial$
AvgPrec	0.1412	0.1409
0	0.4022	0.3548
0.1	0.2624	0.2465
0.2	0.1760	0.1740
0.3	0.1498	0.1392
0.4	0.1350	0.1324
0.5	0.1148	0.1206
0.6	0.0869	0.0959
0.7	0.0748	0.0869
0.8	0.0563	0.0760
0.9	0.0493	0.0644
1	0.0452	0.0592

Πίνακας 2.11: *Μέση Ακρίβεια AvgPrec και μέση ακρίβεια με παρεμβολή στα 11 Recall σημεία για τις μεθόδους χ^2-GOF ομοιόμορφη κατανομή και χ^2-GOF διωνυμική κατανομή (συλλογή εγγράφων fbis, Ερωτήματα 401-450 "Title" version)*

fbis Συλλογή, Ερωτήματα 401-450, Σχετικά έγγραφα 1667		
	$\chi^2 - GOF\ uniform$	$\chi^2 - GOF\ binomial$
AvgPrec	0.1609	0.1557
0	0.3709	0.3707
0.1	0.2592	0.2870
0.2	0.2296	0.2424
0.3	0.2084	0.2049
0.4	0.1796	0.1404
0.5	0.1629	0.1208
0.6	0.1223	0.0971
0.7	0.0821	0.0778
0.8	0.0694	0.0673
0.9	0.0450	0.0567
1	0.0401	0.0480

ity), Weibull κλπ, πιθανόν να είναι πολύ καλές ενναλλακτικές υποθέσεις, αλλά και η δοκιμή και άλλων διαφορετικών στατιστικών ελέγχων όπως Kolmogorov-Smirnov και Anderson-Darling [33].

Ολα αυτά τα θέματα αξίζει να διερευνηθούν σε μια επόμενη εργασία.

Κεφάλαιο 3

Στατιστικές Μέθοδοι για την Εύρεση Συνεκφερόμενων Λέξεων

3.1 Εισαγωγή

Σε αυτό το κεφαλαιο θα περιγράψουμε την εφαρμογή δύο στατιστικών μεθόδων για την εξαγωγή λέξεων που συνεκφέρονται πολύ συχνά μαζί (collocations) μέσα σε κείμενα γραμμένα στην Ελληνική γλώσσα. Η πρώτη μέθοδος είναι η μέθοδος του μέσου και της διασποράς (*Mean* and *Variance* method) η οποία υπολογίζει τις αποστάσεις (distances, offsets) μεταξύ των λέξεων σε ένα corpus και ψάχνει για πρότυπα αποστάσεων με χαμηλή διασπορά (spread). Η δεύτερη μέθοδος βασίζεται στον X-τετράγωνο έλεγχο ($chi - square\ test$), εμφανίζεται περισσότερο ευέλικτη επειδή δεν κάνει την υπόθεση οτι οι λέξεις στο corpus ακολουθούν την κανονική κατανομή. Οι δύο μέθοδοι παράγουν ενδιαφέροντα collocations τα οποία μπορούν να φανούν χρήσιμα σε πολλές εφαρμογές, όπως: computational lexicography, language generation και machine translation.

Τα collocations είναι κοινό χαρακτηριστικό των φυσικών γλωσσών και μπορούν να εμφανισθούν τόσο σε απλό κείμενο φυσικής γλώσσας όσο και σε τεχνικό και επιστημονικό κείμενο. Ενα collocation μπορεί να είναι συνδυασμός λέξεων (ή φράσεων) οι οποίες εμφανίζονται συχνά μαζί στην γλώσσα με ένα τρόπο πού να φαίνεται πολύ φυσικός νοηματικά από τα συμφραζόμενα (context) δηλαδή το πλαίσιο μέσα στο οποίο εμφανίζεται το collocation, παρότι που η απομονωμένη σύνθεση των επί μέρους νοημάτων των λέξεων που συνθέτουν το collocation οδηγεί σε ένα νοηματικό περιεχόμενο άσχετο με το context.

Τα collocations σε φυσικές γλώσσες με πλούσιο κλιτικό σύστημα, όπως η ελληνική γλώσσα μπορούν να εμφανίζονται με δύο τρόπους: με ένα αυστηρό άκαμπτο τρόπο, δηλαδή οι λέξεις πού συνθέτουν το collocation να εμφανίζονται συχνότατα με τόν ίδιο συντακτικό τρόπο, πχ. οι Ελληνικές λέξεις 'Χρηματιστήριο' και 'Αξία' εμφανίζονται ως 'Χρηματιστήριο Αξιών', ενώ μπορεί σε ένα collocation οι λέξεις να εμφανίζονται με ένα πιο 'χαλαρό' τρόπο, πχ. οι Ελληνικές λέθεις 'Στρώνω / στρώνομαι' και 'δουλειά' μπορεί να εμφανισθούν σε διάφορες φράσεις με περισσότερους από ένα τρόπο:
'Στρώνομαι στην δουλειά'
'Η δουλειά μου στρώνει'.

Για τα collocations υπάρχουν πολλοί ορισμοί αφού οι διάφοροι ερευνητές έχουν εστιάσει πάνω σε διάφορα χαρακτηριστικά των collocations. Σύμφωνα με το Firth [55] "Collocations of a given word are statements of the habitual or customary places of the word".

Οι Benson and Morton [50] ορίζουν τα collocations σαν "an arbitrary and recurrent word combination", η λέξη *recurrent* σημαίνει ότι αυτοί οι συνδυασμοί εμφανίζονται συχνά για ένα δεδομένο context.
Η Smadja [64] καθορίζει τέσσερα χαρακτηριστικά των collocations χρήσιμα για τις διάφορες υπολογιστικές εφαρμογές.
α) τα collocations είναι αυθαίρετα, αυτό σημαίνει ότι δεν αντιστοιχούν σε κάποια συντακτική ή σημασιολογική παραλλαγή.
β) Τα collocations είναι domain-dependent , επομένως ο χειρισμός κειμένου σε ένα πεδίο απαιτεί σαφή γνώση, φυσικά της σχετικής ορολογίας αλλά και επι πλέον των domain-dependent collocations.
g) Τα collocations είναι *recurrent* (όπως ορίστηκε παραπάνω).
δ) Τα collocations είναι cohesive lexical clusters, η έννοια του cohesive lexical cluster σημαίνει ότι η εμφάνιση μιας η περισσότερων λέξεων συχνά συνεπάγεται και την εμφάνιση των υπόλοιπων λέξεων του collocation.

Στην εργασία τού Lin [59] τα collocations ορίζονται σαν "a habitual word combi-

nation", ενώ οι Gitsaki et. al [56] τα ορίζουν σαν "a recurrent word combination". Οι Howarth and Nesi [57] δίνουν μια διαφορετική προσέγγιση της χρήσης των collocations απο την σκοπιά του ανθρώπου πού μαθαίνει μια ξένη γλώσσα. Σύμφωνα με τούς Manning and Schutze [60] τα collocations χαρακτηρίζονται από *limited compositionality.* (περιορισμένη συνθετικότητα). Μια έκφραση φυσικής γλώσσας είναι "compositional" εάν η ένοια της έκφρασης μπορεί να προβλεφθεί από την σύνθεση των εννοιών που συνθέτουν το collocation. Για παράδειγμα στην έκφραση 'γερό ποτήρι' ο συνδυασμός των δύο λέξεων έχει μια νέα σημασία, αυτός πού καταναλώνει μεγάλες ποσότητες αλκοόλ, κάτι όμως πού είναι τελείως διαφορετικό από την έννοια πού προκύπτει από την σύνθεση των δύο επι μέρους εννοιών. Τέλος άλλο ένα χαρακτηριστικό των collocations είναι η απυσία έγκυρων συνωνύμων του [60], [59]: για παράδειγμα ενώ οι αγγλικές λέξεις *baggage* and *luggage* είναι συνώνυμες θα μπορούσαμε να γράψουμε μόνο *emotional, historical ή psychological baggage.*

3.2 Η Λογική της Εξαγωγής collocations σε Εφαρμογές ΕΦΓ

Τα collocations είναι σημαντικά για έναν αριθμό εφαρμογών όπως: Natural Language generation, machine translation, text simplification και computational lexicography. Οι Howarth and Nasi [57] ισχυρίζονται οτι οι περισσότερες εκφράσεις φυσικής γλώσσας περιέχουν τουλάχιστον ένα collocation. Επίσης η αυτόματη παραγωγή φυσικής γλώσσας χρειάζεται την σαφή γνώση των σωστών συνδυασμών των λέξεων της γλώσσας.

Το Machine Translation (MT) θεωρείται σαν ένα από τα πιό δύσκολα έργα στην Επεξεργασία Φυσικής Γλώσσας, και στην Τεχητή Νοημοσήνη γενικότερα. Είναι επίσης ένα δύσκολο έργο να μεταφράσουμε collocations απο την μια γλώσσα στην άλλη. Το γεγονός αυτό έχει άμεση επίδραση σε ένα σύστημα μηχανικής μετάφρασης. Σύμφωνα με τον Gitsaki [56] τα collocations διαφέρουν πολύ από την μια γλώσσα στην άλλη. Για παράδειγμα, το *clear road* της Αγγλικής είναι ' ελευθερος δρομος ' (free road) στα Ελληνικά, και όχι ' καθαρός δρόμος '.

Η πληροφορία πού προέρχεται από τα collocations είναι επίσης σημαντική για έργα

όπως η απλοποίηση κειμένου. Αυτό έχει να κάνει με αντικατάσταση δύσκολων λέξεων με πολύ απλούστερες. Χωρίς την γνώση των collocations και των σχετιζόμενων περιορισμών αυτή η αντικατάσταση μπορεί να οδηγήσει σε ένα ακατανόητο κείμενο. Ενα παράδειγμα έργου που αφορά απλοποίηση κειμένου είναι για την Αγγλική γλώσσα το *Practical Simplification of English Text (PSET)* project [51].

Τα collocations είναι επίσης σημαντικά στο Computational Lexicography. Χρησιμοποιούνται για να χαρακτηρίσουν πλήρως τις λεξικές καταχωρήσεις. Σύμφωνα με το Richardson [63] 'Για μια λεπτομερή λεξικογραφική ανάλυση, μόνο τα collocations πού είναι παρόντα σε ένα λεξικό θα παρέχουν πλήρη συνθετικά χαρακτηρηστικά, τα οποία θα μπορούσαν να αποκαλύψουν τις κατ' ευθείαν σημασιολογικές σχέσεις και να συμβάλουν έτσι στον χαρακτηρισμό της λεξικής καταχώρησης '.
O Smith [65] μελέτησε collocations για να ανακαλύψει 'συμβάντα' σχετιζόμενα με πληροφορίες τόπου και χρόνου σε μη δομημένο κείμενο.

Στο παρόν κεφάλαιο της διατριβής, περιγράφουμε δύο στατιστικές μεθόδους εξαγωγής collocations από κείμενα γραμμένα στην νέα Ελληνική γλώσσα. Η πρώτη μέθοδος είναι η μέθοδος του μέσου και της διασποράς (*Mean* and *Variance* method) η οποία υπολογίζει αποστάσεις (distances, offsets) μεταξύ των λέξεων σε ένα corpus και ψάχνει για πρότυπα αποστάσεων με χαμηλή διασπορά Η δεύτερη μέθοδος βασίζεται στο Χ-τετράγωνο έλεγχο. Στην ενότητα 3 εστιάζουμε στις κύριες ιδέες των μεθόδων αυτών. Στην ενότητα 4 περιγράφονται λεπτομερώς οι μέθοδοι πού χρησιμοποιούνται εδώ. Μια μικρή περιγραφή των δεδομένων (data) που χρησιμοποιου΄νται για την αποτίμηση (evaluation) αυτών των μεθόδων καθώς και πειραματικά αποτελέσματα δίνουμε στην ενότητα 5. Τέλος κλείνουμε αυτό το κεφάλαιο με ένα συμπερασματικό σχολιασμό των μεθόδων καθώς και κατευθύνσεις για περαιτέρω συνέχιση της εργασίας πάνω στα collocations.

3.3 Εύρεση collocations με Χρήση Στατιστικών Μεθόδων

Η παραδοσιακή προσέγγιση για την εξαγωγή collocations είναι η λεξικογραφική προσέγγιση. Σύμφωνα με τούς Benson and Morton [50] δεν μπορούμε να χειρισθούμε ξεχωριστά τα συμμετέχοντα μέρη σε ένα collocation (collocates θα τα αποκαλούμε από εδώ και στο εξής). Επομένως το έργο της εξαγωγής των κατάλληλων collocates δεν είναι προβλέψιμο και γενικά τα collocations πρέπει να εξάγονται 'χειρωνακτικά' και έπειτα να παρατίθενται στα λεξικά.

Προσφάτως, οι στατιστικές μέθοδοι έχουν εφαρμοσθεί στην μελέτη των φυσικών γλωσσών καθώς και στην εξαγωγή collocations. Επηρεαζόμενος εν μέρει από την μεγάλη διαθεσιμότητα ηλεκτρονικών κειμένων (corpora) έτοιμων για ανάγνωση από ένα υπολογιστή, ο Choueka [52] δοκίμασε να εξαγάγει αυτόματα από κείμενο collocations χρησιμοποιώντας 'N-γράμματα' (N-grams) από 2 έως 6 λέξεις.
Ένα απλό κριτήριο εξαγωγής collocations από ένα corpus είναι *η συχνότητα της εμφάνισης*. Εάν δύο ή περισσότερες λέξεις εμφανίζονται πολύ συχνά μαζί αυτό είναι μια ένδειξη για collocation. Ατυχώς, η επιλογή των πιο συχνά εμφανιζόμενων *N-grams* δεν οδηγεί πάντοτε σε πολύ ενδιαφέροντα αποτελέσματα. Για παράδειγμα εάν αναζητήσουμε τα συχνότερα *bigrams* σε ένα Αγγλικό κείμενο το αποτέλεσμα θα είναι μια λίστα απο φράσεις του τύπου *of the, in the, to the* κλπ.
Για να ξεπεράσουν αυτό το πρόβλημα οι Justeson and Katz [58] πρότειναν μια 'ευριστική' μέθοδο που βελτιώνει τα προηγούμενα αποτελέσματα. 'Πέρασαν' τις υποψήφιες φράσεις από ένα part-of-speech φίλτρο και διάλεξαν μόνο εκείνα τα *N-grams* τα οποία είναι πιθανά να είναι φράσεις. Μερικά patterns που χρησιμοποιήθηκαν σε αυτό το φίλτρο είναι: *AN, NN, AAN και ANN*. Όπου *A* σημαίνει επίθετο και *N* ουσιαστικό. Αν και η ευριστική αυτή μέθοδος είναι πάρα πολύ απλή, οι συγγραφείς ανέφεραν πολύ σημαντική βελτίωση στα αποτελέσματα.

Η βασιζόμενη στην ' *συχνότητα εμφάνισης* ' μέθοδος δουλεύει πολύ καλά με φράσεις ουσιαστικών. Ωστόσο, πολλά collocations περιέχουν λέξεις με άλλες πιο ευέλικτες συσχετίσεις μεταξύ των. Η μέθοδος *mean and variance* [64] ξεπερνάει αυτό

το πρόβλημα υπολογίζοντας τις αποστάσεις μεταξύ των *collocates* και βρίσκοντας την διασπορά (spread) της κατανομής αυτών των αποστάσεων. Δηλαδή, η μέθοδος υπολογίζει το μέσο και την διασπορά του offset (προσημασμένη απόσταση) μεταξύ δύο λέξεων στο corpus. Αυτή η μέθοδος φαίνεται λογική. Εάν η διασπορά της κατανομής είναι χαμηλή τότε έχουμε μια στενή κορυφούμενη κατανομή των αποστάσων και αυτό είναι μια ένδειξη για collocation. Απο την άλλη πλευρά εάν η διασπορά είναι υψηλή τότε τα offsets κατανέμονται τυχαία, δηλαδή δεν υπάρχει ένδειξη για collocation.

Mutual information είναι ένα μέτρο πού χρησιμοποιείται για την εξαγωγή collocations [53]. Ο ορος *Mutual information* έχει την καταγωγή του από την Θεωρία της Πληροφορίας. Ο όρος *information* έχει την περιορισμένη έννοια ενός συμβάντος το οποίο λαμβάνει χώρα αντιστρόφως ανάλογα με την πιθανότητά του και ορίζεται σαν μια ιδιότητα πού κατέχουν οι τιμές των τυχαίων μεταβλητών. Είναι χονδρικά ένα μέτρο του πόσο πολύ μία λέξη μας πληροφορεί για μία άλλη.

Θα περιγράψουμε τις κύριες ιδέες πίσω από τις μεθόδους πού θα εφαρμόσουμε εδώ. Την *mean and variance* μέθοδο και την X^2 test (προφέρεται Chi-square test). Θα δώσουμε επίσης ένα ενναλλακτικό τύπο για τον υπολογισμό της X^2 στατιστικής για την περίπτωση της εξαγωγής bigrams από το corpus. Το X^2 test είναι πολύ καλά ορισμένη στατιστική προσέγγιση για την εκτίμηση του εάν ή όχι ένα συμβάν είναι ένα αποτέλεσμα της τύχης, δηλαδή τυχαίο γεγονός. Αυτό είναι γενικότερα ένα από τα κλασικά προβλήματα της στατιστικής και συνήθως διατυπώνεται από την άποψη του hypothesis testing. Στην περίπτωσή μας, θέλουμε να ξέρουμε κατά πόσο δύο λέξεις 'συμβαίνουν' μαζί περισσότερο συχνά από ό,τι στην τύχη. Διατυπώνουμε για τον σκοπό αυτό την μηδενική υπόθεση (null hypothesis) H_0 για ένα δείγμα εμφανίσεων. Η υπόθεση αυτή δηλώνει ότι δεν υπάρχει διασύνδεση των δύο λέξεων πέρα από αυτήν της εμφάνισης μαζί από τύχη. Υπολογίζουμε την πιθανότητα p πού θα είχε το *συμβάν* εάν η H_0 ήταν αληθινή. Εάν η p είναι πολύ χαμηλή κάτω από ένα προκαθορισμένο επίπεδο σημαντικότητας $p < 0.005$ η $p < 0.001$, απορρίπτουμε την H_0, διαφορετικά την δεχόμαστε ώς αληθινή.

Για τον υπολογισμό γενικότερα τέτοιων πιθανοτήτων για την απόρριψη ή μη της

μηδενικής υπόθεσης συνήθως χρησιμοποιούμε την t statistic:

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{3.1}$$

Όπου $\overline{x}$ δηλώνει τον μέσο τού δείγματος, s^2 η διασπορά τού δείγματος, N το μέγεθος του δείγματος και μ ο μέσος της κατανομής x.
Εάν η τιμή της t είναι αρκετά μεγάλη τότε μπορούμε να απορρίψουμε την *μηδενική υπόθεση* H_0. Το πρόβλημα με την *t statistic* είναι ότι υποθέτει δεδομένα πού κατανέμονται ακολουθώντας την κανονική κατανομή, κάτι πού δεν είναι κατανάγκη σωστό για την περίπτωση γλωσσολογικών δεδομένων. Αυτός είναι ο λόγος που επιλέξαμε το X^2 έλεγχο, γιατί δεν υποθέτει κανονικά κατανεμημένα δεδομένα. Ωστόσο, με αυτή την στατιστική έχουν παρατηρηθεί γενικότερα παρενέργειες στην περίπτωση πού έχουμε *αραιά δεδομένα* (sparse data).
Ο Dunning [54] πρότεινε ένα εναλλακτικό έλεγχο, τον likelihood ratios testτο οποίο δουλεύει καλύτερα από τα στατιστικούς ελέγχους tests στην περίπτωση των *αραιών δεδομένων*.

3.4 Μέθοδοι για την Εύρεση collocations

3.4.1 Ο Μέσος και η Διασπορά

Ο Μέσος *(Mean)* είναι ο αριθμητική μέση τιμή των δεδομένων. Εάν έχουμε n παρατηρήσεις $x_1, x_2, ..., x_n$, τότε ο αριθμητικός μέσος δίνεται από τον τύπο:

$$Mean = \frac{x_1 + x_2 + ... + x_n}{n} \tag{3.2}$$

Η *Διακύμανση* των n παρατηρήσεων $x_1, x_2, ..., x_n$ είναι η μέση τετραγωνική απόκλιση αυτών των παρατηρήσεων από τον *Μέσο* τους:

$$Variance = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n - 1} \tag{3.3}$$

Η *Τυπική Απόκλιση* s είναι η τετραγωνική ρίζα της *Διακύμανσης*:

$$s = \sqrt{Variance} \tag{3.4}$$

Ας δούμε ένα παράδειγμα από την Ελληνική γλώσσα. Ας θεωρήσουμε το ρήμα *κτύπησε* και ένα από τα πιο συχνά ορίσματά του την λέξη *πόρτα*. Παραθέτουμε μερικές προτάσεις με τις δύο αυτές λέξεις:

- *Κτύπησε την πόρτα του*
- *Κτύπησε δυνατά την πόρτα του*
- *Κτύπησε την σιδερένια πόρτα του*
- *Κτύπησε την σιδερένια και βαριά πόρτα του*

Ο αριθμός των λέξεων πού εμφανίζονται μεταξύ των δύο λέξεων *'κτύπησε'* και *'πόρτα'* δεν είναι σταθερός, έτσι η απόσταση των δύο λέξεων μεταβάλεται από πρόταση σε πρόταση. Μετρώντας την συχνότητα εμφάνισης των *'κτύπησε'* και *'πόρτα'* σε μια σταθερή απόσταση δεν εξάγουμε κανένα συμπέρασμα.
Για να έχουμε ένα μέτρο της συσχέτισης μεταξύ *'κτύπησε'* και *'πόρτα'* μπορούμε να υπολογίσουμε τον Μέσο και την Διακύμνση των offsets (προσημασμένων αποστάσεων).
Για τις παραπάνω προτάσεις υπολογίζουμε το *mean offset* μεταξύ των λέξεων *'κτύπησε'* και *'πόρτα'*, σύμφωνα με την εξίσωση 3.2:

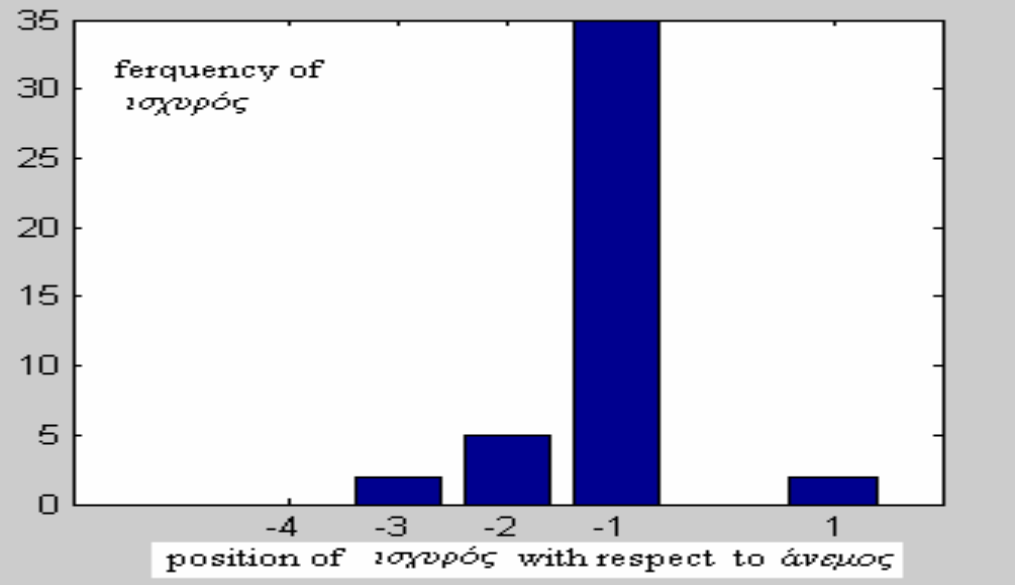
$$Mean = \frac{1}{4}(1 + 2 + 2 + 4) = 2.25 \tag{3.5}$$

Εάν εμφανιζόταν η λέξη *'πόρτα'* πριν την λέξη *'κτύπησε'* τότε θα αποδίδαμε το *offset* με αρνητικό αριθμό.
Η *Διακύμανση* των *offsets* εκτιμά πόσο πολύ το κάθε ξεχωριστό $offset$ αποκλίνει από την μέση τιμή. Εκφράζει την απόκλιση της απόστασης μεταξύ των δύο λέξεων.
Χρησιμοποιώντας την εξίσωση 3.3 υπολογίζουμε την διακύμανση ως ακολούθως:

$$Variance = \frac{1}{3}((1 - 2.25)^2 + (2 - 2.25)^2 + (2 - 2.25)^2 + (4 - 2.25)^2) = 1.58 \tag{3.6}$$

Η *Τυπική Απόκλιση* είναι $\sqrt{1.58} = 1.26$.

Ο *Μέσος* και η *Απόκλιση* μας βοηθά να βρούμε collocations ψάχνοντας για ζευγάρια λέξεων με χαμηλές διακυμάνσεις. Οσο πιο χαμηλή είναι η διακύμανση μεταξύ

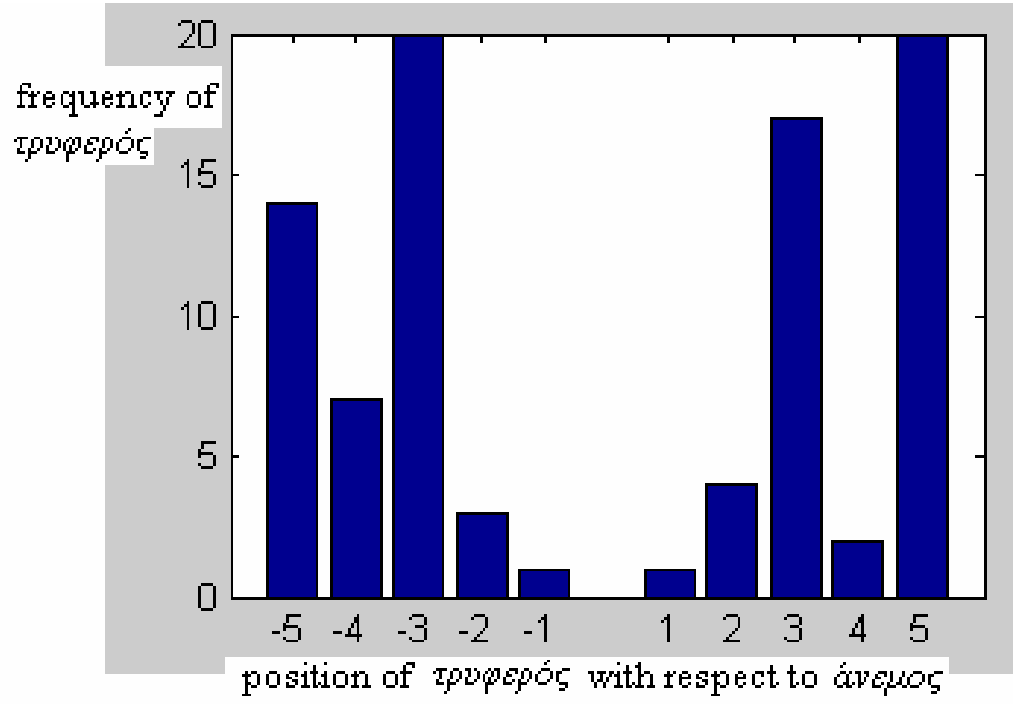


Σχήμα 3.1: *Κατανομή των αποστάσεων της λέξης ισχυρός σε σχέση με την λέξη άνεμος*

των αποστάσεων σε ένα ζευγάρι λέξεων τόσο πιο ισχυρή είναι η ένδειξη ότι αυτό το ζευγάρει μπορεί να σχηματίζει ένα collocation. Μια πολύ χαμηλή τιμή διακύμανσης σημαίνει ότι οι δύο λέξεις εμφανίζονται μαζί με την ίδια περίπου απόσταση. Μπορούμε να εξηγήσουμε εύκολα τα αποτελέσματα εάν θεωρήσουμε την κατανομή της μιας λέξης σε σχέση με την άλλη. Εάν υπάρχει μια στενή οξεία κορυφούμενη κατανομή τότε αυτό είναι μια ένδειξη μια στενής συντακτικής σχέσης μεταξύ των δύο λέξεων. Ας εξηγήσουμε αυτή την περίπτωση με ένα παράδειγμα με μετρήσεις πού έγιναν στο corpus των κειμένων πού είχαμε διαθέσιμο για αποτίμηση. Μετρώντας τις εμφανίσεις του ζευγαριού των λέξεων (*ισχυρός, άνεμος*), υπολογίζουμε τις αποστάσεις της λέξης *ισχυρός* σε σχέση με την λέξη *άνεμος* και προκύπτει η κατανομή του σχήματος 3.1.

Αυτή η κατανομή είναι μια καλή ένδειξη της συντακτικής σχέσης των δύο λέξεων καθ΄ οσον έχουμε μια κορυφούμενη κατανομή με χαμηλό spread. Ενώ αντιθέτως η κατανομή του σχήματος 3.2 που παριστάνει την κατανομή των αποστάσεων της λέξης *τρυφερός* σε σχέση με την *άνεμος* δεν υποδηλώνει καμία τέτοια ένδειξη.

Εως τώρα δεν μιλήσαμε καθόλου για ακραίες τιμές που επηρεάζουν την τιμή του *Μέσου* και της *Διασποράς*. Δυστυχώς, αυτή η απλή μέθοδος μπορεί να οδηγήσει σε αποτυχία στην περίπτωση των πολύ υψηλών συχνοτήτων και των χαμηλών διακυμάνσεων. Ωστόσο μπορούμε να αποφύγουμε τις ακραίες τιμές κάνοντας μια κανονικοποίηση



Σχήμα 3.2: *Κατανομή των αποστάσεων της λέξης τρυφερός σε σχέση με την λέξη άνεμος*

των μετρήσεων σε σχέση με το μέγεθος του δείγματος.

Στο επόμενο κεφάλαιο περιγράφουμε την τεχνική που θα χρησιμοποιήσουμε σε αυτή την εργασία, τον έλεγχο X-τετράγωνο.

3.5 X-τετράγωνο Έλεγχος του Pearson

Το 1900, ο Karl Pearson ανέπτυξε μια στατιστική, την στατιστική X^2 η οποία συγκρίνει τούς παρατηρηθέντες και αναμενόμενους αριθμούς όταν οι δυνατές εκβάσεις ενός πειράματος υποδιαιρούνται σε αμοιβαία αποκλειόμενες κατηγορίες. Ο υπολογισμός της X^2 στατιστικής γίνεται από τον τύπο της εξίσωσης

$$X^2 = \sum \frac{(observed - expected)^2}{expected} \tag{3.7}$$

Όπου το Ελληνικό γράμμα Σ παριστάνει το άθροισμα και υπολογίζεται για τις κατηγορίες όλων των δυνατών εκβάσεων.
Οι *παρατηρηθείσες* και *αναμενόμενες* τιμές μπορούν να εξηγηθούν στο πλαίσιο του hypothesis testing. Εάν έχουμε δεδομένα τα οποία διαιρούνται σε αμοιβαία αποκλειό-

	w_1 = ισχυρός(strong)	$w_1 \neq$ ισχυρός
w_2 = άνδρας (man)	10 Πχ (ισχυρός άνδρας) strong man	1000 Πχ. (σεμνός άνδρας) decent man
$w_2 \neq$ άνδρας	500 Πχ. (ισχυρός άνεμος) strong wind	1,500,000 Πχ. (ασθενής ήχος) weak sound

Σχήμα 3.3: *Πίνακας συνάφειας για το ζευγάρι των λέξεων (ισχυρός, άνδρας)*

μενες κατηγορίες και διατυπώσουμε μια μηδενική υπόθεση για τα δεδομένα, τότε η *αναμενόμενη* τιμή είναι η τιμή για την κάθε κατηγορία εάν η μηδενική υπόθεση είναι αληθινή. Η *παρατηρηθείσα* τιμή είναι η τιμή για την κάθε κατηγορία την οποία μετρούμε από τα δεδομένα του δείγματος.

Ο X^2 έλεγχος είναι ένας αξιοσημείωτα ευέλικτος τρόπος μέτρησης του κατα πόσο τα δεδομένα συμφωνούν με τις υποκείμενες λεπτομέρειες μιας μηδενικής υπόθεσης.

Για να γίνει κατανοητή η εφαρμογή της παραπάνω μεθόδου στην εύρεση collocations δίνουμε ένα παράδειγμα.
Ας υποθέσουμε ότι έχουμε ένα δείγμα από γλωσσολογικά δεδομένα και ενδιαφερόμαστε να εξαγάγουμε collocations από bigrams. Ορίζοντας ένα collocational window μετράμε την συχνότητα εμφάνισης για το ζευγάρι (*ισχυρός, άνδρας*). Υπάρχουν 10 εμφανίσεις του ζευγαριού (*ισχυρός, άνδρας*) μέσα στο corpus, 1000 bigrams όπου η δεύτερη λέξη είναι *άνδρας* αλλά η πρώτη λέξη δεν είναι *ισχυρός*, 500 bigrams όπου η πρώτη λέξη είναι *ισχυρός* αλλά η δεύτερη όχι *άνδρας* και 1.500.000 bigrams πού δεν περιέχουν καμμία απο τις δύο λέξεις στην κατάλληλη θέση δεδομένου του collocational window. Στην περίπτωση αυτή θα ήταν χρήσιμο να χρησιμοποιήσουμε τον πίνακα συνάφειας (contingency table) στον οποίον τα δεδομένα ταξινομούνται όπως φαίνεται στον πίνακα του σχήματος 3.3

Επιπλέον, χρησιμοποιώντας maximum likelihood estimates μπορούμε να υπολογίσουμε τις πιθανότητες του *ισχυρός* και *ανδρας* ως ακολούθως:

$P(\text{ισχυρός}) = 510/1.501.510$

P(ανδρας) = 1010/1.501.510

Η *μηδενική υπόθεση* είναι ότι οι εμφανίσεις του *ισχυρός* και *άνδρας* είναι ανεξάρτητες.

H_0 : P(ισχυρός, ανδρας) = P(ισχυρός) χ P(ανδρας)
=(510/1.501.510)χ(1010/1.501.510) = 1.013 χ 10-5

Επειτα υπολογίζουμε την X^2 τιμή χρησιμοποιώντας την εξίσωση 3.7.
Ψάχνοντας στούς πίνακες της X^2 κατανομής η χρησιμοποιώντας το κατάλληλο λογισμικό στατιστικής βρίσκουμε μια *κρίσμη* τιμή για το επίπεδο σημαντικότητας α (συνήθως $\alpha = 0,05$) και για ένα βαθμό ελευθερίας (η X^2 στατιστική έχει βαθμό ελευθερίας ένα για έναν 2χ2 πίνακα συνάφειας).
Εάν η υπολογιζόμενη X^2 τιμή είναι μεγαλύτερη από την *κρίσμη* τιμή, μπορούμε να απορρίψουμε την μηδενική υπόθεση οτι οι λέξεις *ισχυρός* και *άνδρας* εμφανίζονται ανεξάρτητα. Επομένως για μια μεγάλη υπολογιζόμενη X^2 τιμή έχουμε ισχυρή ένδειξη για το ζευγάρι των λέξεων να σχηματίζει collocation.

Δίνουμε παρακάτω (εξίσωση 3.8) ένα απλούστερο τύπο υπολογισμού που χρησιμοποιούμε στην περίπτωση πού έχουμε ενα 2χ2 πίνακα συνάφειας.

$$X^2 = \frac{N(a_{11}a_{22} - a_{12}a_{21})^2}{(a_{11} + a_{12})(a_{11} + a_{21})(a_{12} + a_{22})(a_{21} + a_{22})} \tag{3.8}$$

όπου α_{ij} είναι οι καταχωρήσεις του 2χ2 πίνακα συνάφειας και N το άθροισμα αυτών των καταχωρήσεων. Υλοποιώντας αυτό τον τύπο προγραμματιστικά πρέπει να φροντίσουμε για την υπερχείλιση μνήμης πού είναι πιθανό να συμβεί όταν διαιρούμε τον αριθμητή με τον παρονομαστή στην εξίσωση 3.8. Για να ξεπεράσουμε αυτή την προβληματική κατάσταση, ειδικά όταν το μέγεθος του corpus είναι πολύ μεγάλο και οι συχνότητες πολύ μικρές παραγοντοποιούμε τον παραπάνω τύπο.

3.6 Πειραματικά Αποτελέσματα

Αρκετά αρχεία κειμένων νέας Ελληνικής γλώσσας ήταν διαθέσιμα από διάφορες πηγές (Internet, electronic databases, κλπ). Μια πρωταρχική μορφολογική διαδικασία part-of-speech tagging σημείωσε το μέρος του λόγου και το λήμμα για κάθε λέξη του σώματος (corpus). Τα αρχεία συνενώθηκαν και δημιουργήθηκε ένα μεγάλο γλωσσολογικό σώμα απαρτιζόμενο από 8.967.432 λέξεις στις οποίες υπήρχε διαθέσιμο το λήμμα. Ατυχώς η προεπεξεργασία μας δεν ήταν ικανή να μας παράσχει τα λήματα για ρήματα κα επιρρήματα. Στο πίνακα 3.1 συνοψίζεται η κατανομή των λημμάτων στο corpus.

Ουσιαστικά	6.739.006
Ρήματα	0
Επίθετα	2.228.426
Επιρρήματα	0
Σύνολο	8.967.432

Πίνακας 3.1: *Κατανομή λημμάτων στο σώμα αποτίμησης*

Τα λήμματα για τα ρήματα και τα επιρρήματα όπως αναφέραμε και προηγουμένως δεν παρέχονται. Παρατηρήστε ότι το συνολοκό άθροισμα των λημμάτων είναι 8.967.432. Τα υπόλοιπα 8.977.083-8.967.432=9.651 λήμματα ανήκουν σε μια κατηγορία πού ο επισημειωτής (tagger) τους δίνει την ετικέτα *RgFwGr* και ανήκουν σε ξένες λέξεις με Ελληνικούς χαρακτήρες που χρησιμοποιούνται όπως προφέρονται.

Τα δέκα συχνότερα ουσιαστικά και επίθετα στο corpus φαίνονται στον πίνακα του σχήματος 3.4

Τα εισαγωγικά εδώ χρησιμοποιούνται για να δηλώσουν ότι αυτές οι λέξεις αποτελούν λήμματα όπως ακριβώς δίνεται απο τον tagger.

3.6.1 Ανάλυση της Διασποράς

Ο μόνος συνδυαμός των διγραμμάτων bigrams που μπορούμε να δοκιμάσουμε είναι 'Επίθετο,Ουσιαστικό' καθώς δεν παρέχονται τα άλλα μέρη του λόγου.
Υπολογίζουμε από το σώμα τις αποστάσεις και την τυπική απόκλιση αυτών των αποστάσεων για όλους τους συνδυασμούς των διγραμμάτων 'Επίθετο, Ουσιαστικό',

	Noun	Frequency	Adjective	Frequency
1	"ελλάδα"	60318	"πολιτική"	25892
2	"νόμος"	33680	"μεγάλη"	15965
3	"κανόνας"	31349	"περισσότερος"	15147
4	"θέση"	31011	"ελληνική"	13508
5	"διεθνός-διεθνώς"	30369	"νέος"	12744
6	"πρόβλημα"	27835	"εθνική"	11360
7	"κυβέρνηση"	26095	"μεγάλος"	10929
8	"χρόνι-χρόνια-χρόνιο"	25580	"όλη"	10680
9	"πρόεδρος"	25302	"οικονομική"	10664
10	"θέμα"	25297	"ελληνικός"	9057

Σχήμα 3.4: *τα δέκα συχνότερα ουσιαστικά και επίθετα*

ορίζοντας ένα collocational παράθυρο των 10 λέξεων, συμπεριλαμβανομένων και των σημείων στίξης. Υπενθυμίζουμε ότι, μια θετική απόσταση $d, (-10 <= d <= 10)$ υποδηλώνει εδώ ότι το ουσιαστικό βρίσκεται d λέξεις δεξιότερα του επιθέτου ενώ μια αρνητική το αντίθετο, ότι δηλαδή το ουσιαστικό βρίσκεται d θέσεις αριστερότερα του επιθέτου. Στους πίνακες 3.2 και 3.3 παρουσιάζονται τα δύο πρώτα διγράμματα που προέκυψαν από τούς υπολογισμούς μας με την πιο υψηλή και πιο χαμηλή αντίστοιχα τυπική απόκλιση.

Στα σχήματα 3.5 και 3.6 σχεδιάζουμε την κατανομή των αποστάσεων για το πρώτο πιο χαμηλής και πιο υψηλής τυπικής απόκλισης αντίστοιχα. Ενώ στους πίνακες 3.7 και 3.8 τα πρώτα δέκα ζευγάρια με τα scores της πιο ξαμηλής και πιο υψηλής τυπικής απόκλισης αντίστοιχα.

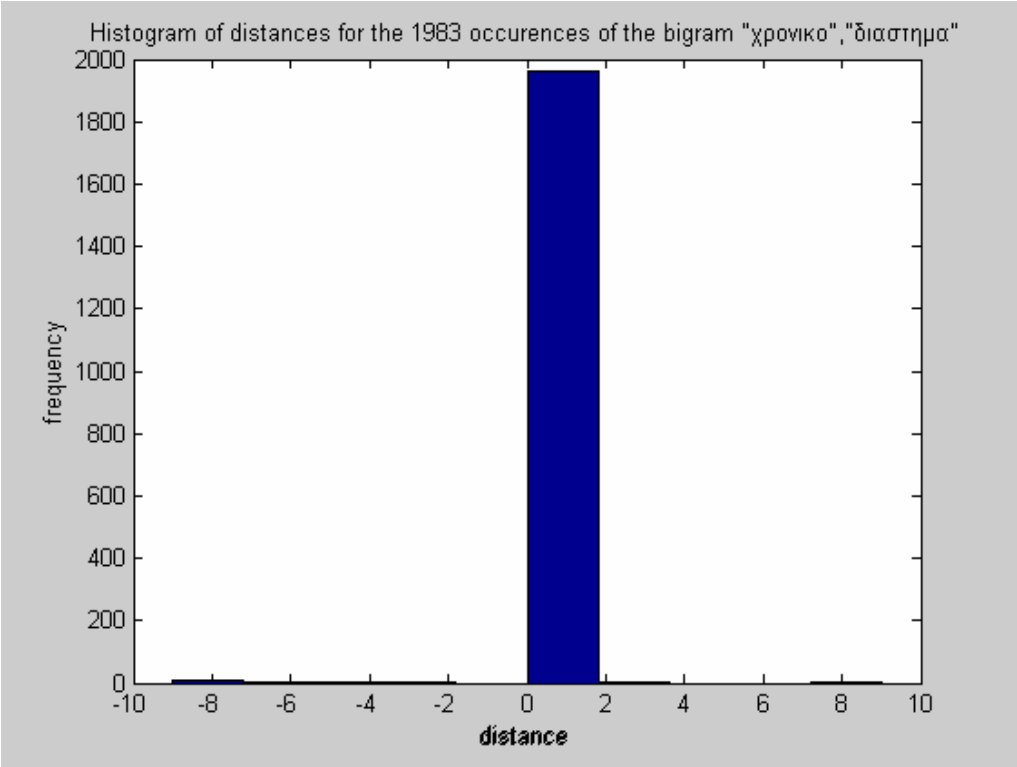
(Επιθ., Ουσιαστικό)	Συχνότητα	Μέσος	Τυπ. Απόκλιση
(*χρονικό, διάστημα*)	1983	0,9561	0,7654
(*κεντρική, σημασία*)	13	1,2308	0,8321

Πίνακας 3.2: *Τα δύο πρώτα διγράμματα με την πιο χαμηλή τυπική απόκλιση*

Σε κάθε καταχώριση εμφανίζονται με την σειρά το ζευγάρι των λέξεων του διγράμματος, η συχνότητα , ο μέσος και η τυπική απόκλιση.
Στις εικόνες 3.5 και 3.6 σχεδιάζουμε τις κατανομές των αποστάσεων για το ζευγάρι με την πιο χαμηλή διακύμανση και την πιο υψηλή αντίστοιχα.

(Επιθ., Ουσιαστικό)	Συχνότητα	Μέσος	Τυπ. Απόκλιση
(*εξωτερικός τρόπος*)	17	1,2353	8,2956
(*εσωτερική, Γιώργος*)	12	0,9167	8,6072

Πίνακας 3.3: *Τα δύο πρώτα δυγράμματα με την πιο χαμηλή τυπική απόκλιση*

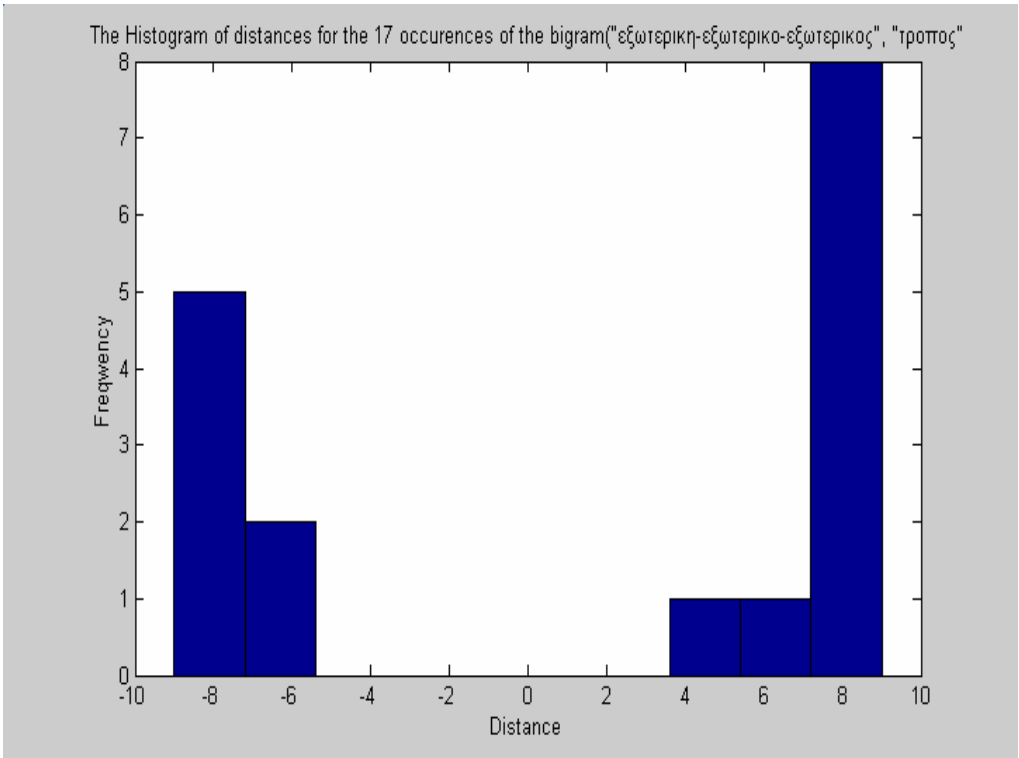


Σχήμα 3.5: *Κατανομή των αποστάσεων για το ζευγάρι με την πιο χαμηλή τυπική απόκλιση (χρονικό, διάστημα)*

Ερμηνεία: Αν ένα δίγραμμα εμφανίζει χαμηλή τυπική απόκλιση έχουμε ισχυρή ένδειξη ότι το δίγραμμα με χαμηλή τυπική απόκλιση και μονοπλευρική υψηλής τιμής κορυφούμενη κατανομή αποτελεί collocation. Οσο πιο στενό είναι το σχήμα και όσο πιο υψηλή είναι η κορυφή τόσο πιο ισχυρή είναι η ένδειξη.

3.6.2 Αναλυση του Ελέγχου 'Χ τετράγωνο'

Ο έλεγχος 'Χ τετράγωνο' είναι πιο ευέλικτος απ ό,τι η διασπορά η οποία μπορεί να αποτύχει στην περίπτωση των πολύ υψηλών συχνοτήτων. Ο έλεγχος 'Χ τετράγωνο' κάνει μια υπόθεση την *μηδενική υπόθεση* της στατιστικής ανεξαρτησίας μεταξύ των δύο λέξεων που απαρτίζουν το δίγραμμα. Δηλαδή η μηδενική υπόθεση υποθέτει ότι οι



Σχήμα 3.6: *Κατανομή των αποστάσεων για το ζευγάρι με την πιο υψηλή τυπική απόκλιση (εξωτερικός, τρόπος)*

Lemmadj	lemmanou	stdv
"χρονικό"	"διάστημα"	0,7654
"κεντρική"	"σημασία"	0,8321
"ειδικός"	"απάντηση"	1,1875
"μεγάλος"	"βαθμός"	1,1932
"περασμένος"	"κανόνας"	1,3007
"αμερικανική-αμερικανικής"	"κανόνας"	1,3817
"κυριακής"	"ελλάδα"	1,3901
"ανά"	"κόσμος"	1,4151
"οικονομικό"	"παιχνίδι"	1,4434
"εργαζομένα-εργαζομένη-εργαζομένης"	"διεθνός-διεθνώς"	1,4546

Σχήμα 3.7: *Τα πρώτα 10 διγράμματα (επίθετο, όνομα) με την πιο χαμηλή τυπική απόκλιση (χρονικό, διάστημα), (ειδική απάντηση),...)*

Lemmadj	lemmanou	stdv
"ζήτημα"	"πληροφορία-πληροφορίη"	7,854
"χθεσινής"	"συμμετοχή"	7,866
"ανά"	"ιστορία"	7,8671
"μήνα-μήνη"	"χρήση"	7,8758
"ενδεχόμενος"	"θέμα-θέμας"	7,8988
"χθεσινής"	"διεθνός-διεθνώς"	7,9036
"εθνικός"	"μείωση"	7,9174
"συγκεκριμένο"	"ιστορία"	7,9601
"κοινωνική-κοινωνικής"	"λαός"	7,9663
"λίγη"	"διεθνός-διεθνώς"	7,9935

Σχήμα 3.8: *Τα πρώτα 10 δυγράμματα (επίθετο, όνομα) με την πιο υψηλή τυπική απόκλιση (εξωτερικός, τρόπος), (συγκεκριμένη, ιστορία),...)*

δύο λέξεις εμφανίζονται ανεξάρτητες η μία με την άλλη μέσα στο σώμα.
Υπολογίζοντας την τιμή της X^2-στατιστικής απορρίπτουμε την *μηδενική υπόθεση* εάν η τιμή της ξεπερνάει κάποια κρίσιμη τιμή όπως αυτή ορίζεται από την X^2-κατανομή. Για παράδειγμα, εάν κοιτάξουμε ένα στατιστικό πίνακα (η την τιμή που επιστρέφει ένα στατιστικό πακέτο), βρίσκουμε ότι για ένα επίπεδο πιθανότητας $\alpha = 0.05$, (δηλαδή με βεβαιότητα 95%) και με βαθμό ελευθερίας '1' για τα διγράμματα (πίνακας συνάφειας 2χ2), η κρίσιμη τιμή για να απορρίψουμε την *μηδενική υπόθεση* είναι $X^2 = 3.841$.

Φυσικά θα μπορούσαμε να δοκιμάσουμε και ένα μεγαλύτερο επίπεδο σημαντικότητας, ας πούμε 99% ($\alpha = 0.01$) η περισσότερο, αλλά αυτό θα αύξανε περισσότερο την κρίσιμη τιμή.
Η ουσία του ελέγχου είναι να εξετάσει την υπολογιζόμενη X^2 τιμή και να αποφασίσει για την εξάρτηση των δύο λέξεων. Οσο πιο υψηλή είναι η τιμή του X-square τόσο πιο ισχυρή θα είναι και η ένδειξη για collocation.

Πειραματικά Αποτελέσματα: Το σώμα κειμένων που έχουμε στην διάθεσή μας αποτελείται απο 29.539.802 λέξεις. Δοθέντος αυτού του αριθμού και ένός collocational παραθύρου των 10 λέξεων γύρω από ένα στόχο επίθετο, μπορούμε να υπολογίσουμε τον δυνατό ριθμό των διγραμμάτων (επίθετο, ουσιαστικό). Αυτό μπορεί να γίνει από

Adjective	Noun	X2score	a11	a12	a21	a22
"κοινωνικής"	"διάλογος"	3,4057	59	117373	41737	265699004
"κοινωνικής"	"μείωση"	3,3964	10	112994	41786	265703383
"διαφορετικός"	"μέλη"	3,3488	11	116863	43135	265698164
"συγκεκριμένος"	"σημασία"	3,3426	11	111553	45637	265700972
"χρονικό"	"δημόσια"	3,3325	9	112041	41553	265704570
"προοπτική"	"μείωση"	3,2941	11	112993	45169	265700000
"ίδιο"	"παρουσία"	3,1651	11	112471	44161	265701530
"διαφορετικός"	"συμφωνία"	3,1563	11	115063	43135	265699964
"σημερινή"	"συμφωνία"	3,1501	10	115064	42776	265700323
"κυπριακός"	"σημασία"	3,1498	12	111552	47454	265699155

Σχήμα 3.9: *Τα 10 πρώτα δγράμματα με την πιό υψηλή X^2 τιμή*

Adjective	Noun	X2score	a11	a12	a21	a22
"σημερινή"	"στιγμή"	0	16	299972	42770	265515415
"δηλώσης"	"πρόγραμμα"	0	15	280533	121305	265456320
"ουσιαστικά"	"σημείο"	0	18	190638	70182	265597335
"εξωτερική"	"ευρωπαϊκή"	0	251	227377	160849	265469696
"έτοιμος"	"λόγος"	0	21	367683	50235	265440234
"εσωτερική"	"ενδιαφέρον"	0	5	149125	66505	265642538
"μεγάλος"	"άτομο"	0	50	133114	196672	265528337
"ίδιο"	"παιχνίδι"	0	19	211625	44153	265602376
"αργότερα"	"τουρκία"	0	41	299191	108085	265450856
"οικονομικός"	"προσπάθεια"	0	18	239634	55746	265562775

Σχήμα 3.10: *Τα 10 τελευταία δγράμματα με την πιό χαμηλή X^2 τιμή*

τον τύπο:

$$\textit{Total number of bigrams} = (29539802 - 9) * 9 + 36$$

Για το κάθε ένα απο αυτά τα διγράμματα ψάχνουμε στο σώμα και υπολογίζουμε τα α_{ij}, τις καταχωρήσεις δηλαδή του 2χ2 πίνακα συνάφειας 3.8, για να μπορέσουμε τελικά να υπολογίσουμε τη X^2 τιμή. Στους πίνακες 3.9 και 3.10 φαίνονται τα 10 πρώτα διγράμματα με τη πιο υψηλή και τη πιο χαμηλή X^2 τιμή αντίστοιχα. Στον δεύτερο πίνακα η X^2 τιμή δεν είναι ακριβώς μηδέν αλλά προσεγγίζει το μηδέν.

3.7 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε δύο μεθόδους για μηχανική εύρεση collocations για την Ελληνική γλώσσα. Τη μέθοδο ' Μέσου και Διασποράς ' και τον έλεγχο X-τετράγωνο. Στην τελευταία περίπτωση βρήκαμε με τα πειραματικά μας αποτελέσματα ότι είναι δυνατό να δουλέψουμε θαυμάσια με πολύ μεγάλα corpus της Ελληνικής γλώσσας. Σε σχέση με την χρήση και άλλων εργαλείων για Επεξεργασία Φυσικής Γλώσσας, η μέθοδος είναι από μόνη της αυτάρκης, με εξαίρεση την χρήση Λημματοφράφου.

Για να υπολογίσουμε την σημαντικότητα θα μπορούσαμε να χρησιμοποιήσουμε και άλλες πολύ καλά θεμελειωμένες στατιστικές μέθοδοι, όπως mutual information (MI), log likelihood (LL) ratio test, t-test κλπ. Όμως αυτές τις μεθόδους τις απορρίψαμε χάριν της X^2 στατιστικής. Ο λόγος είναι ότι αυτοί οι έλεγχοι έχουν το μειονέκτημα ότι υποθέτουν παραμετρική κατανομή δεδομένων. Αυτό είναι έκδηλα ακατάλληλο όταν υπολογίζουμε συχνότητες λέξεων. Επι πλέον η (MI) συγκρίνει την συνδεδεμένη πιθανότητα $p(w_1, w_2)$ και απαιτεί οι ανεξάρτητες πιθανότητες $p(w_1), p(w_2)$ να συμβαίνουν με οποιονδήποτε τρόπο στο δείγμα, $MI(w1, w2) = log_2 \frac{p(w1,w2)}{p(w1)*p(w2)}$, το οποίο δεν δίνει ρεαλιστική εικόνα στην περίπτωση πολύ χαμηλών συχνοτήτων. Εάν για παράδειγμα μια μη συχνή λέξη έχει συχνότητα 1 σε ένα συγκεκριμένο συνδυασμό, αυτό μπορεί να οδηγήσει σε πολύ υψηλή τιμή του MI και να δεχθούμε μια απόφαση για ένα ισχυρό σύνδεσμο μεταξύ των δύο λέξεων, αν και η συνύπαρξη των δύο λέξεων μπορεί να οφείλεται καθαρά στην τύχη.

Ο έλεγχος X-τετράγωνο είναι ο πιο κοινά χρησιμοποιούμενος έλεγχος για στατιστική σημαντικότητα στην υπολογιστική γλωσσολογία και μπορεί να χρησιμοποιηθεί σε πολλά διαφορετικά contexts.

Τα επόμενα είναι κατευθύνσεις για μελλοντική δουλειά.
Το σύστημα μπορεί να ενσωματώσει γνώση για να βοηθηθεί στην εύρεση των collocations και να βελτιώσει τα αποτελέσματα. Τέτοια γνώση μπορεί να προέλθει από την χρήση λεξικών θησαυρών όπως synonyms, hypernyms, hyponyms, antonyms κλπ. αν μπορούν σε κάποια στιγμή να γίνουν διαθέσιμα για την Ελληνική γλώσσα. Ο Pearsen

[62] έχει δουλέψει με παρόμοιο τρόπο και με επιτυχία στην Αγγλική γλώσσα χρησιμοποιώντας το ηλεκτρονικό λεξικό WordNet.

Χρησιμοποιώντας αυτές τις στατιστικές μεθόδους θα μπορούσαμε ακριβώς να πετύχουμε μια καλή αναπαράσταση της *προτασιακής γνώσης* μας για την Ελληνική γλώσσα. Συνδυάζοντας τέτοιες στατιστικές μεθόδους σε ένα πλαίσιο *Αναπαράστασης* της γνώσης με εννοιακούς γράφους, θα μπορούσαμε να συλλέξουμε πολύτιμη πληροφορία βρίσκοντας εννοιολογικά σχετιζόμενες λέξεις επιτυγχάνοντας έτσι πλουσιότερες γνωσιακές βάσεις.

Κεφάλαιο 4

Στατιστικοί Έλεγχοι για Συστήματα Αποσαφήνισης της Έννοιας μιας Λέξης

Σε αυτό το κεφάλαιο θα περιγράψουμε μία μέθοδο εφαρμογής των στατιστικών ελέγχων για καλό 'ταίριασμα' στην αποσαφήνιση της έννοιας μιας λέξης μέσα στο πλαίσιο όπου εμφανίζεται (word sense diasambiguation). Στη μέθοδο αυτή θα κάνουμε χρήση των σχέσεων μεταξύ των λεξικών καταχωρήσεων του WordNet. Η μέθοδός μας ανήκει στην κατηγορία των "unsupervised" αλγορίθμων για αποσαφήνιση της έννοιας μιας λέξης. Το ″unsupervised″ έχει την έννοια του ότι δεν χρειάζεται προηγούμενη εκπαίδευση πάνω σε αποσαφηνισμένα δεδομένα.

Σύμφωνα με την μέθοδό αυτή εργαζόμαστε ως εξής: επαυξάνουμε το πλαίσιο (context) όπου εμφανίζεται η πρός αποσαφήνιση λέξη, χρησιμοποιώντας τις σχέσεις του WordNet και πιο συγκεκριμένα τα συσχετιζόμενα synsets (έτσι αποκαλούνται οι λεξικές καταχωρήσεις του WordNet πού σχετίζονται με διάφορες λεξικές και σημασιολογικές σχέσεις). Το νέο πλαίσιο που προκύπτει το θεωρούμε σαν ένα στατιστικό δείγμα όπου έπειτα μελετάμε την κατανομή των συσχετιζόμενων synsets. Διατυπώνουμε την ″μηδενική υπόθεση″, σύμφωνα με την οποία για όλες τις έννοιες (senses) της προς αποσαφήνιση λέξης τα συσχετιζόμενα synsets κατανέμονται κανονικά (normally) μέσα στο δείγμα. Για κάθε έννοια της προς αποσαφήνιση λέξης, ποσοτικοποιούμε την απόκλιση από αυτή την υπόθεση με την βοήθεια του X^2 στατιστικού ελέγχου. Η έννοια που εμφανίζει την μεγαλύτερη απόκλιση επιλέγεται σαν η σωστή έννοια της προς

αποσαφήνιση λέξης.

Ο παραπάνω αλγόριθμος αποδεικνύεται πολύ αποδοτικός και εκτιμήθηκε η αποδοτικότητά του πάνω σε πραγματικά δεδομένα πού χρησιμοποιήθηκαν στον english Senseval-2, επίσημο διαγωνισμό για συστήματα αποσαφήνισης λέξεων.

Το σύστημά μας συγκαταλέγεται μεταξύ των πρώτων unsupervised συστημάτων καταλαμβάνοντας την δεύτερη θέση.

4.1 Αποσαφήνιση Λέξης και WordNet

Οι πιο πολλές λέξεις στις φυσικές γλώσσες είναι πολύσημες, δηλαδή έχουν πολλές σημασίες. Το πρόβλημα της απόδοσης της σωστής έννοιας μιας λέξης (target word) μέσα στο πλαίσιο (context) που αποτελείται από τις περιβάλουσες λέξεις, είναι η αποστολή των συστημάτων αποσαφήνισης λέξεων. Αναγνωρίζεται σαν ένα από τα πιο δύσκολα προβλήματα στην επεξεργασία φυσικής γλώσσας. Με το πέρασμα των χρόνων πολλές προσπάθειες έχουν γίνει και πολλά συστήματα έχουν κατασκευαστεί ώστε οι υπολογιστές να αναγνωρίζουν την σωστή έννοια μιας λέξης στο πλαισιό της.

Τα πρώτα συστήματα εβασίζοντο σε σύνολα εμπειρικών κανόνων και αποσαφήνιζαν μόνο ένα μικρό αριθμό από παραδείγματα. Ωστόσο, χρησιμοποιώντας αυτές τις εργασίες σαν αναφορά και με την βοήθεια μικρών λεξικών (όπως ένας απλός κατάλογος του ορισμού των εννοιών) πολλοί αλγόριθμοι παρουσιάστηκαν [16], [17], [18], με την ελπίδα ότι θα μπορούσαν να εφαρμοστούν και σε μεγαλύτερα λεξικά.

Σήμερα η διαθεσιμότητα μεγάλων ηλεκτρονικών λεξικών, όπως είναι το Wordnet [19], δίνει μεγάλη ώθηση στην ανάπτυξη συστημάτων αποσαφήνισης λέξης. Το Wordnet περιέχει μεγάλο αριθμό εννοιών για κάθε λέξη και έχει αυξήσει το ενδιαφέρον για την ανάπτυξη περισσότερο απαιτητικών εφαρμογών για αποσαφήνιση λέξης και γενικότερα συστημάτων Επεξεργασίας Φυσικής Γλώσσας που μπορούν να αποκτήσουν πλεονέκτημα από την διάκριση εννοιών (sense distinction) [20], [21], [22]. Επιπλέον, το γεγονός ότι οι διάφορες έννοιες συνδέονται μεταξύ των δια μέσου σημαντικού αριθμού από σημασιολογικές (semantic) και λεξικές (lexical) σχέσεις, κάνει το WordNet πολύτιμη πηγή για την τυποποίση των δικτύων αναπαράστασης γνώσης, τα οποία είναι πολύ δημοφιλή στούς κόλπους των ερευνητών της υπολογιστικής γλωσσολογίας.

Χρησιμοποιώντας ορισμούς από την ηλεκτρονική λεξικολογική βάση του WordNet, οι Mihalcea και Moldovan [23] συγκέντρωσαν λεξικολογική πληροφορία από το internet για την αυτόματη απόκτηση κειμένων στα οποία έχουν αποσαφηνισθεί και σημειωθεί οι έννοιες των περιεχόμενων λέξεων (sense tagged corpora). Οι Montoyo και Palomar [24] παρουσίασαν μια μέθοδο για αυτόματη αποσαφήνιση των ουσιαστικών. Χρησιμοποίησαν τα ″Specification Marks″ όπως τα ονόμασαν, τα οποία είναι όμοια με τις σημασιολογικές τάξεις (semantic classes) στήν ιεραρχία του WordNet (taxonomy), και οδηγήθηκαν σε βελτιωμένα αποτελέσματα χρησιμοποιώντας ορισμούς. Η εργασία των Banerjee and Pederson [25] παρουσιάζει μια προσαρμογή του αλγορίθμου Lesk [16], χρησιμοποιώντας σημασιολογικές σχέσεις και ορισμούς.

Εκτός απο την χρήση των ορισμών του λεξικού, πολύ δουλειά έχει επίσης πραγματοποιηθεί στην αποσαφήνιση λέξης με την χρήση της σχέσης ιεραρχίας του WordNet (hyponymy/hypernymy relation). Ο Resnick [21] αποσαφήνισε εμφανίσεις ουσιαστικών υπολογίζοντας την σημασιολογική ομοιότητα (semantic similarity) μεταξύ δύο λέξεων διαλέγοντας τον κοινό πρόγονο στην ιεραρχία με το μεγαλύτερο πληροφοριακό περιεχόμενο (the most informative "subsumer"), όπου το πληροφοριακό περιεχόμενο το όρισε σαν συνάρτηση του πλήθους των υπαγόμενων όρων.

Σε μια άλλη προσέγγιση βασιζόμενη και αυτή στην ιεραρχία του WordNet οι Leacock και Chodorow [26] πρότειναν ένα μέτρο για τον υπολογισμό της σημασιολογικής ομοιότητας, υπολογίζοντας το μήκος της διαδρομής μεταξύ των δύο κόμβων της ιεραρχίας. Οι Agirre and Rigau [20] πρότειναν μια μέθοδο βασιζόμενη στην εννοιολογική πυκνότητα (conceptual density) μεταξύ δύο εννοιών στην ιεραρχία και παρουσίασαν και ένα τύπο για την εννοιολογική πυκνότητα για το σκοπό αυτό.

Οι Budanitsky και Graeme [27] παρουσίασαν πειραματικά δεδομένα και συνέκριναν την αποδοτικότητα των παραπάνω μεθόδων σε ένα σύστημα διόρθωσης της ορθογραφίας πραγματικών λέξεων.

4.2 Εισαγωγή

Στο πνεύμα της εφαρμογής των στατιστικών ελέγχων, παρουσιάζουμε μια προσέγγιση για συστήματα αποσαφήνισης λέξης βασιζόμενοι σε μια διαφορετική αντίληψη για την εκτίμηση του μέτρου της σχετικότητας (relatedness) μεταξύ του πλασίου μιας λέξης και των εννοιών της. Αυτό μας οδηγεί στην αναζήτηση ποιοτικών στοιχείων που χαρακτηρίζουν την σχετικότητα, παρά σε ποσοτικά μέτρα σύγκρισης.

Το WordNet συνδέει κάθε λεξική καταχώρηση, η οποία αναπαριστά μια έννοια και αποδίδεται σαν ένα σύνολο από συνώνυμες λέξεις (synset), με άλλες λεξικές καταχωρήσεις δια μέσου των σημασιολογικών συσχετίσεων (semantic relations) δημιουργώντας έτσι για την λεξική καταχώρηση ένα σύνολο από συσχετιζόμενα synsets (related synsets).

Χρησιμοποιώντας αυτές τις σχέσεις, επαυξάνουμε το πλαίσιο (τις περιβάλουσες λέξεις) της προς αποσαφήνιση λέξης με όλα τα συσχετιζόμενα synsets των λέξεων που βρίσκονται στο πλαίσιο. Θεωρώντας αυτό το επαυξημένο πλαίσιο σαν ένα τυχαίο στατιστικό δείγμα μελετάμε την σύνθετη κατανομή των συσχετιζόμενων synsets, του καθενός ξεχωριστά για κάθε μια έννοια, μετρώντας την συχνότητα εμφάνισης στο δείγμα. Επειδή προσδοκούμε η σωστή έννοια να έχει διαφορετική κατανομή των συσχετιζόμενων synsets στο πλαίσιο από ό,τι οι άλλες έννοιες, κάνουμε την υπόθεση, ότι τα συσχετιζόμενα synsets της σωστής έννοιας (correct sense) κατανέμονται λιγότερο κανονικά σε σχέση με τα συσχετιζόμενα synsets των άλλων εννοιών. Ο X^2 έλεγχος καλού "ταίριασματος" [28] εκτιμά ποσοτικά αυτή την υπόθεση συγκρίνοντας τις παρατηρηθείσες συχνότητες στο δείγμα με τις αναμενόμενες. Την ποσοτική εκτίμηση της υπόθεσης που κάναμε την χρησιμοποιούμε σαν κριτήριο για να αποφασίσουμε για την σωστή έννοια της προς αποσαφήνιση λέξης.

Στις επόμενες ενότητες αυτού του κεφαλαίου περιγράφουμε περισσότερο αναλυτικά τις παραπάνω ιδέες. Περιγράφουμε αρχικά τις διάφορες σχέσεις του WordNet , έπειτα παρουσιάζουμε μια μικρή αναφορά στον X^2 έλεγχο και τέλος τον αλγόριθμο αποσαφήνισης καθώς και πειραματικά δεδομένα από την εκτίμηση της αποδοτικότητας του αλγορίθμου πάνω σε πραγματικά δεδομένα.

4.3 Οι σχέσεις του Wordnet

Οι λεξικές καταχωρήσεις στο Wordnet οργανώνονται σε λογικές ομαδοποιήσεις οι οποίες καλούνται synsets. Κάθε synset αποτελείται από μια λίστα από συνώνυμα, δηλαδή λέξεις με την ίδια σημασία τις οποίες μπορούμε να εναλλάξουμε στο ίδιο πλαίσιο (context) χωρίς να αλλοιωθεί η σημασία του πλαισίου. Για παράδειγμα το synset {*administration, governance, establishment, brass, organization, organisation*} αναπαριστά την έννοια του κυβερνητικού σώματος το οποίο διοικεί κάτι.

Το βασικό χαρακτηριστικό το οποίο διαφοροποιεί το WordNet από τα άλλα συμβατικά λεξικά είναι οι σχέσεις του (relations), δείκτες οι οποίοι υποδεικνύουν τις συσχετίσεις μεταξύ των synsets με άλλα synsets. Επί του παρόντος, το WordNet ξεχωρίζει τις σημασιολογικές (semantic) από τις λεξικές (lexical) συσχετίσεις· Οι λεξικές συσχετίσεις ισχύουν μεταξύ των μορφών των λέξεων (word forms) ενώ οι σημασιολογικές συσχετίσεις ισχύουν μεταξύ των εννοιών των λέξεων. Εφόσον μια σημασιολογική συσχέτιση είναι μια σχέση μεταξύ εννοιών και οι έννοιες αναπαρίστανται με synsets, μπορούμε να θεωρήσουμε τις σημασιολογικές συσχετίσεις σαν σχέσεις μεταξύ synsets. Για κάθε synset στο WordNet, αυτοί οι δείκτες συνδέουν το synset με άλλα synsets σχηματίζοντας έτσι μια λίστα από συνδεδεμένα synsets τα οποία τα αποκαλούμε related synsets.

Το WordNet αποθηκεύει πληροφορία για λέξεις οι οποίες ανήκουν στα τέσσερα μέρη του λόγου (parts-of-speech): ουσιαστικά, ρήματα, επίθετα και επιρρήματα. Προθέσεις, σύνδεσμοι και άλλες λειτουργικές συνδετικές λέξεις δεν συμπεριλαμβάνονται στο λεξικό. Εκτός από μονές λέξεις, το WordNet μερικές φορές περιέχει και λέξεις πού συνεκφέρονται (collocations) (πχ ″fountain pen″ , ″take in″) οι οποίες αποτελούνται από δύο η περισσότερες λέξεις αλλά τις χειριζόμαστε από κάθε άποψη σαν μία.

Ο αλγόριθμός μας κάνει χρήση όλων των σχέσεων που μας παρέχει το Wordnet για ουσιαστικά, ρήματα, επίθετα και επιρρήματα, αλλά έχουμε την δυνατότητα να χρησιμοποιήσουμε με την ίδια λογική και μέρος αυτών των σχέσεων προκειμένου να εξασφαλίσουμε καλύτερη αποδοτικότητα του αλγορίθμου. Δίνουμε παρακάτω μια σύντομη περιγραφή για όλες τις διαθέσιμες σχέσεις του WordNet.

4.3.1 Σχέσεις για Ουσιαστικά

Οι ορισμοί των κοινών ουσιαστικών κατά κανόνα περιλαμβάνουν έναν όρο που βρίσκεται στην αμέσως ανώτερη τάξη στην ιεραρχία (superordinate term) και κάποια διακριτικά χαρακτηριστικά που εξειδικεύουν την έννοια που αναπαριστά το ουσιαστικό σε σχέση με την έννοια του superordinate term [19]. Αυτό ακριβώς το γνώρισμα αποτελεί και την βάση για τον τρόπο οργάνωσης των ουσιαστικών σε ιεραρχίες στο WordNet.

Τα ουσιαστικά οργανώνονται σε ιεραρχίες μέσω μιας σχέσης που αποκαλείται *'hypernymy/hyponymy'*, ή *'is-a'*, ή *'is a kind of'*. Εάν αναπαραστήσουμε την σχέση *'is-a'* με ένα δεξί βέλος $\Rightarrow$, μπορούμε να αναπαραστήσουμε την ιεραρχία στο Wordnet σαν μια συνδεδεμένη λίστα, όπου τα βέλη δείχνουν τους διαδοχικούς κόμβους της ιεραρχίας. Ας δούμε ένα πραγματικό παράδειγμα από το WordNet.

{*aid, assistant, help*} $\Rightarrow$ {*resource*} $\Rightarrow$ {*asset, plus*}$\Rightarrow$ {*quality*} $\Rightarrow$ {*attribute*} $\Rightarrow$ {*abstraction*}.

Επί πλέον δείκτες μπορούν να χρησιμοποιηθούν για να υποδηλώσουν και άλλες διαθέσιμες συσχετίσεις από το WordNet που αφορούν τα ουσιαστικά. Αλλες διαθέσιμες σχέσεις είναι οι εξής: *holonymy* και *meronymy* οι οποίες είναι επίσης σημασιολογικές σχέσεις (semantic relations) και συνδέουν δύο synsets εάν η οντότητα η οποία αναφέρεται στο πρώτο synset είναι μέρος της οντότητας του άλλου. Οι παραπάνω σχέσεις είναι συμπληρωματικές με την έννοια ότι εάν το A είναι μέρος του B δηλαδή ένα meronym του B τότε και αντιστρόφως, το B είναι holonym του A.

Μπορούμε να διακρίνουμε τούς εξής τρείς τύπους της *holonymy*: *Member-Of* , *Substance-Of* και *Part-Of* και αντιστρόφως κατά αντιστοιχία τρείς τύπους *meronymy*: *Has-Member*, *Has-Substance* and *Has-Part*.

Ένα ενδιαφέρον χαρακτηριστικό του WordNet είναι οτι οι σημασιολογικές συσχετίσεις για τα ουσιαστικά μπορούν να χρησιμοποιηθούν αναδρομικά και να δώσουν ιεραρχίες. Έτσι μπορούμε να έχουμε hierarhical holonym και hierachical meronym εάν διασχίσουμε την ιεραρχία του Wordnet για το κάθε ένα μέρος χωριστά.

Οι σχέσεις *antonymy* και *attribute* είναι άλλες δύο σχέσεις που ορίζονται για τα ουσιαστικά. Η *antonymy* είναι μια λεξικολογική σχέση (lexical relation) και όχι σημασιολογική, η οποία συνδέει μαζί δύο ουσιαστικά τα οποία είναι αντίθετα το ένα του άλλου και η *attribute* είναι μια σημασιολογική σχέση η οποία συνδέει μαζί ένα ουσιαστικό synset και ένα επίθετο synset. Οι τιμές της *attribute* εκφράζονται με επίθετα και χρησιμοποιούνται για να τροποποιήσουν την έννοια των ουσιαστικών.

4.3.2 Σχέσεις για Ρήματα

Οι σχέσεις *Hyponymy* και *Hypernymy* οι οποίες όπως είπαμε χρησιμοποιούνται μεταξύ ουσιαστικών, χρησιμοποιούνται επίσης και μεταξύ ρημάτων αλλά με λίγο διαφορετικό τρόπο. Εξετάζοντας τα hyponyms ενός ρήματος καθώς και τούς superordinate όρους διαπιστώνουμε ότι όλες αυτές οι συνδεδεμένες λέξεις (lexicalization) περιέχουν πολλά είδη σημασιολογικών λεπτομερειών που καλύπτουν πολλά διαφορετικά σημασιολογικά πεδία [30].

Αυτές οι λεπτομέρειες (elaborations) έχουν συγχωνευθεί σε μια τροπική σχέση, μια σχέση που εκφράζει τρόπο και η οποία καλείται *troponymy* (από την Ελληνική λέξη ″τρόπος″). Αυτή η σχέση μεταξύ δύο ρημάτων μπορεί να εκφρασθεί με τον εξής τρόπο: Το ρήμα synset V_1 είναι ένα hypernym του synset V_2, εάν το V_2 είναι V_1 κατά ένα ιδιαίτερο τρόπο. Το V_1 είναι τότε όπως λέμε troponymy του V_2.

Η *entailment* είναι μια σχέση μεταξύ ρημάτων η οποία μοιάζει με την *meronymy* μεταξύ των ουσιαστικών, αλλά η *meronymy* ταιριάζει καλύτερα για ουσιαστικά παρά για ρήματα. Αυτή η σχέση *entailment* ισχύει μεταξύ δύο ρημάτων εάν το ένα λογικά παράγει το άλλο (entailment). Για παράδειγμα, *snore* entails *sleep*. Ενα πιο συγκεκριμένο (specific) είδος της σχέσης *entailment* είναι η σχέση *cause*. Ισχύει το εξής: Εάν δύο synsets σχετίζονται με την σχέση *cause* τότε σχετίζονται και με την σχέση *entailment*, αλλά όχι αντιστρόφως. Δηλαδή, εάν V_1 causes V_2, τότε επίσης και V_1 entails V_2. Ένα παράδειγμα είναι τα ρήματα (*give* και *have*): Δίνοντας κάτι σε κάποιον προκαλεί (causes) τον αποδέκτη να το έχει, έχοντας κάτι κάποιος δεν σημαίνει ότι του έχει δοθεί.

Αλλη μια σχέση μεταξύ των ρημάτων είναι η *antonymy* η οποία είναι ίδια στην χρήση με την *antonymy* που περιγράψαμε στα ουσιαστικά.

4.3.3 Σχέσεις για Επίθετα και Επιρρήματα

Τα επίθετα στο Wordnet τακτοποιούνται σε ομαδοποιήσεις (clusters), οι οποίες περιέχουν τα synsets κεφαλές (head synsets) και τα δορυφορικά synsets (satellite synsets). Η κεφαλή περιέχει ένα ζευγάρι αντώνυμων επιθέτων που κάθε μέλος του ζευγαριού έχει ένα η περισσότερα δορυφορικά synsets τα οποία αναπαριστούν την ίδια έννοια με το επίθετο της κεφαλής.

Μία άλλη συχνή σχέση η οποία ορίζεται μεταξύ επιθέτων είναι η *similar to*, η οποία είναι μια σημασιολογική σχέση μεταξύ δύο synsets τα οποία είναι επίθετα και παρόμοια στην έννοια. Αυτό είναι κάτι παρόμοιο με τις συνώνυμες λέξεις σε ένα synset (synset words), αλλά όμως όχι ακριβώς το ίδιο. Τα επίθετα που συνδέονται με την σχέση *similar to* είναι μεν όμοια αλλά όχι τόσο όμοια αρκετά, ώστε να τα τοποθετήσουμε μαζί στο ίδιο synset.

Όπως αναφέραμε προηγούμενα, η σχέση *attribute* στην κατηγορία των ουσιαστικών συσχετίζει ένα ουσιαστικό μέ ένα επίθετο. Αυτά τα σχετιζόμενα επίθετα έχουν μια ειδική ονομασία και αποκαλούνται *pertainyms*. Είναι μια ειδική κατηγορία επιθέτων στο Wordnet τα λεγόμενα και σχεσιακά επίθετα (relational adjectives) και δεν ακολουθούν την δομή του ζευγαριού (head, satelite) όπως αναφέραμε προηγουμένως. Μια άλλη τους ιδιότητα είναι ότι δεν έχουν αντίθετα (antonyms): Το synset για ένα pertainym πάρα πολύ συχνά περιέχει μια μόνο λέξη ή ένα (collocation) (ένα συνδυασμό λέξεων) και ένα λεξικολογικό δείκτη (lexical pointer) πρός το ουσιαστικό με το οποίο συνδέεται με την σχέση *pertaining* το συγκεκριμένο επίθετο.

Μία άλλη κατηγορία επιθέτων τα αποκαλούμενα "συμμετοχικά" επίθετα (participial adjectives) σχετίζονται με τα ρήματα μέσω της συσχέτισης *participle of* πού είναι λεξικολογικοί δείκτες (lexical pointers) πάνω σε ρήματα από τα οποία θεωρούμε ότι προέρχονται τα συγκεκριμένα επίθετα.

Η *also-see,* άλλη μια σημασιολογική σχέση η οποία συσχετίζει επίθετα με ένα τρόπο όμοιο με αυτόν της αντίστοιχης σχέσης στα ρήματα.

Τέλος για τα επιρρήματα συναντάμε συχνά την σχέση *antonymy* την ίδια ακριβώς που συναντάται στα επίθετα. Επίσης επειδή τα επιρρήματα συχνά προέρχονται από επίθετα, το synset που αντιπροσωπεύει ένα επίρρημα συχνά θα περιέχει έναν λεξικολογικό δείκτη πάνω στο επίθετο από το οποίο παράγεται. Αυτό γίνεται μέσω της σχέσης *pertainym*.

4.4 Η X-τετράγωνο στατιστική και Έλεγχοι Καλού "Ταιριάσματος"

Το 1900, ο Karl Pearson διατύπωσε την X^2-στατιστική στην προσπάθειά του να εκτιμήσει την διαφορά μεταξύ παρατηρηθεισών και αναμενόμενων συχνοτήτων, όταν τα δεδομένα από τα οποία προέρχονται τα δυνατά ενδεχόμενα γεγονότα πού μπορεί να συμβούν (outcomes) υποδιαιρούνται σε αμοιβαία αποκλειόμενες κατηγορίες (mutually exclusive categories) .

Όπως αναφέραμε στο κεφάλαιο 2 οι στατιστικοί έλεγχοι μπορούν να εφαρμοσθούν μέσα στο πλαίσιο του ελέγχου για την "μηδενική υπόθεση" (null hypothesis). Εδώ, αποσαφηνιζουμε περισσότερο την "μηδενική υπόθεση" καθώς και τον τρόπο με τον οποίο θα εφαρμόσουμε τούς στατιστικούς ελέγχους για τις ανάγκες της αποσαφήνισης της έννοιας μιας λέξης.

Εάν έχουμε δεδομένα τα οποία υποδιαιρούνται σε αμοιβαία αποκλειόμενες κατηγορίες και σχηματίσουμε την μηδενική υπόθεση για τα δεδομένα, τότε αυτό πού ορίζουμε αναμενόμενη συχνότητα (expected frequency) είναι η τιμή της κάθε κατηγορίας εάν η "μηδενική υπόθεση" είναι αληθινή. Όσον αφορά την παρατηρηθείσα τιμή (observed value) αυτή είναι η τιμή της κάθε κατηγορίας την οποία παρατηρούμε από τα δεδομένα του δείγματος. Συμπερασματικά έχουμε: η μηδενική υπόθεση είναι η πίστη μας για την "θεωρητική κατανομή" η οποία θεωρούμε ότι διέπει τα δεδομένα μας.

Για ονομαστικές κατηγορίες (nominal data) στις οποίες το πλήθος στοιχείων για κάθε κατηγορία έχει καταγραφεί και πινακοποιηθεί (tabulated), η παρατηρηθείσα τιμή είναι το πραγματικό πλήθος που μετράμε και η αναμενόμενη τιμή είναι το πλήθος το οποίο προβλέπεται από την υποτιθέμενη θεωρητική κατανομή. Ας το δούμε αυτό με

ένα παράδειγμα.

Αν θεωρήσουμε την "μηδενική υπόθεση" ότι σε ένα συγκεκριμένο πληθυσμό μιας πόλης το 60% είναι γυναίκες και το 40% είναι άνδρες, τότε σε ένα δείγμα 100 ανθρώπων από αυτή την πόλη, οι αναμενόμενες τιμές θα είναι 60 άνδρες και 40 γυναίκες σύμφωνα με την μηδενική υπόθεση. Οι παρατηρηθείσες συχνότητες θα είναι η πραγματική μέτρηση για τούς 100 ανθρώπους (ας πούμε, 63 και 47). Ενας πρακτικός κανόνας μάς λέει οτι η X^2 στατιστική είναι έγκυρη όταν οποιαδήποτε αναμενόμενη τιμή δεν είναι μικρότερη από 5. Εάν σε κάποιες κατηγορίες έχουμε μικρές αναμενόμενες τιμές, που αυτό σημαίνει ότι είναι μικρό το μέγεθος του δείγματος, τότε μπορούμε να συνδυάσουμε αυτές τις κατηγορίες σε μια μεγαλύτερη.

Τελειώνοντας μπορούμε να πούμε ότι σε γενικές γραμμές η X^2 στατιστική είναι ένας αξησημείωτα αποτελεσματικός και ακριβής τρόπος για να μετρήσουμε το πόσο καλά συμφωνούν τα δεδομένα με την μηδενική υπόθεση.

Η X^2 στατιστική είναι το άθροισμα των τεραγώνων των διαφορών μεταξύ παρατηρηθεισών και αναμενόμενων συχνοτήτων, με κάθε τεραγωνική διαφορά να διαιρείται με την αναμενόμενη συχνότητα.

$$X^2 = \sum_{i=1}^{k} \frac{(observed_i - expected_i)^2}{expected_i}, \qquad \text{κ ο αριθμός των κατηγοριών.} \tag{4.1}$$

Η παραπάνω στατιστική ακολουθεί την X^2 κατανομή (βλέπε κεφάλαιο 2).

4.4.1 Ελεγχοι Καλού "Ταιριάσματος"

Γενικά οι έλεγχοι καλού ταιριάσματος ελέγχουν την "συμμόρφωση" των παρατηρηθεισών δεδομένων από μια εμπειρική συνάρτηση κατανομής (συνήθως τις πραγματικές μετρήσεις στο δείγμα) με τα αναμενόμενα δεδομένα απο μια υποτιθέμενη συνάρτηση κατανομής. Ο X^2 στατιστικός έλεγχος το κάνει αυτό εφαρμόζοντας τον απλό τρόπο των μετρήσεων στο δείγμα, όμως υπάρχουν και διαφορετικές προσεγγίσεις. Ο έλεγχος Kolmogorov-Smirnov [28] για παράδειγμα υπολογίζει την κάθετη απόσταση (vertical distance) μεταξύ της εμπειρικής και της υποτιθέμενης θεωρητικής κατανομής.

Ανεξάρτητα από την προσέγγιση πού εφαρμόζεται στον έλεγχο της ″συμμόρφωσης″ των εμπειρικών δεδομένων με τα θεωρητικά, το ζητούμενο είναι πρώτον να διατυπωθεί η μηδενική υπόθεση και δεύτερον να προσδιοριστεί η εμπειρική συνάρτηση κατανομής στα πραγματικά δεδομένα.

Θα πρέπει να προσδιορίσουμε μια μονομεταβλητή συνάρτηση κατανομής (univariate distribution function) της οποίας να μπορεί να υπολογίσθεί η αθροιστική συνάρτηση κατανομής πάνω σε κατηγοριοποιημένα δεδομένα (binned data). Δηλαδή, δεδομένα που να μπορούν να κατηγοριοποιηθούν σε κλάσεις, αν και οι έλεγχοι ταιριάσματος μπορούν να εφαρμοσθούν και πάνω σε μη κατηγοριοποιημένα δεδομένα, εάν απλά υπολογίσουμε ένα ιστόγραμμα ή πίνακα συχνοτήτων πρίν την εφαρμογή του ελέγχου.

Για τον υπολογισμό του X^2 στατιστικού ελέγχου που χρησιμοποιούμε εδώ, τα δεδομένα υποδιαιρούνται σε k κατηγορίες, κλάσεις (bins) και η στατιστική ελέγχου υπολογίζεται από την εξίσωση 4.1. Ο έλεγχος όπως έχουμε πεί είναι ευαίσθητος στην εκλογή των κλάσεων διότι απο αυτή την εκλογή εξαρτάται η αναμενόμενη συχνότητα σε κάθε κατηγορία

Μετά την εκλογή των κλάσεων το μόνο που χρειάζεται είναι ο υπολογισμός των αναμενόμενων συχνοτήτων από την υποτιθέμενη θεωρητική κατανομή η οποία θεωρούμε ότι διέπει τα δεδομένα. Εάν F είναι η αθροιστική συνάρτηση κατανομής για την υποτιθέμενη θεωρητική κατανομή, τότε η αναμενόμενη συχνότητα υπολογίζεται από τον ακόλουθο τύπο:

$$Expected_i = N(F(Y_u) - F(Y_l)) \tag{4.2}$$

Όπου η Y_u είναι το πάνω όριο για την κλάση i, Y_l είναι το κάτω όριο για την κλάση i, και N είναι το μέγεθος του δείγματος.

Στην επόμενη ενότητα περιγράφουμε τον τρόπο εφαρμογής του X^2 στατιστικού ελέγχου για την αποσαφήνιση της έννοιας μιας λέξης με την μελέτη των κατανομών των συσχετιζόμενων ″synsets″ του Wordnet.

4.5 Ο Αλγόριθμος Αποσαφήνισης

Όπως αναφέραμε στο προοίμιο αυτού του κεφαλαίου, για την μέθοδο αποσαφήνισης της έννοιας μιας λέξης πού θα αναπτύξουμε εδώ βασιζόμενοι στους στατιστικούς ελέγχους, θα χρησιμοποιήσουμε σχέσεις του WordNet. Αν και είχαμε την δυνατότητα να χρησιμοποιήσουμε πολλούς συνδυασμούς από όλες τις διαθέσιμες σχέσεις που παρέχονται από το λεξικό, αποφασίσαμε να αναπτύξουμε την μέθοδό μας σε 2 εκδόσεις: Την μία χησιμοποιώντας κατανομή των συσχετιζόμενων synsets για όλες τις διαθέσιμες σχέσεις και για όλα τα μέρη του λόγου (part-of-speech), όπως αυτές περιγράφονται στην ενότητα 2 αυτού του κεφαλαίου, και την άλλη, χρησιμοποιώντας μόνο ένα συνδυασμό τριών σχέσεων, τις σχέσεις *antonymy*, *hypernymy* και *hyponymy*. Στην πρώτη περίπτωση θέλαμε να έχουμε μια εκτίμηση της συνολικής συνεισφοράς όλων των σχέσεων του WordNet στην διαδικασία αποσαφήνισης μιας λέξης και στην δεύτερη περίπτωση να αυξήσουμε την αποδοτικότητα του αλγορίθμου μας, επειδή σε μια πρωταρχική αλλά όχι πάντως εξαντλητική δοκιμή των συνδυασμών των σχέσεων ο συγκεκριμένος συνδυασμός παρουσίασε τα καλύτερα αποτελέσματα.

4.5.1 Το Σύνολο των Συσχετιζόμενων Synsets για το Πλαίσιο

Ως γνωστόν για την αποσαφήνιση της έννοιας μιας λέξης η μόνη ένδειξη είναι το πλαίσιο (context) μέσα στο οποίο αυτή εκφέρεται. Όταν λέμε πλαίσιο μιας λέξης εννοούμε τα συμφραζόμενα με τα οποία εμφανίζεται η λέξη σε περιβάλλον φυσικής γλώσσας. Αυτό απαρτίζεται από τις περιβάλλουσες λέξεις της συγκεκριμένης λέξης και συνήθως είναι μία η περισσότερες προτάσεις φυσικού κειμένου.

Η πληροφορία που παρέχεται από το το πλαίσιο της λέξης της οποίας σκοπό έχουμε να αποσαφηνίσουμε τη έννοιά της (target word), αποκαλείται και τοπική πληροφορία (local information). Έχει καταδειχθεί ότι παρέχει σημαντικότατη πληροφοριακή ένδειξη για την σωστή έννοια της προς αποσαφήνιση λέξης.

Για το μήκος του χρησιμοποιούμενου πλαισίου, του αριθμού δηλαδή των λέξεων, οι Leacock και Chodorov [26] σε πειράματα πού έκαναν με έναν τοπικού πλαισίου ταξινομητή (local context classifier) βρήκαν ότι μια βέλτιστη τιμή για το μήκος του πλαισίου είναι ένα παράθυρο από ±6 λέξεις που έχουν λεξικές καταχωρήσεις στο

WordNet γύρω από την προς αποσαφήνιση λέξη.

Η θεματική (topical) πληροφορία μας παρέχει ισχυρή ένδειξη για την σωστή έννοια μιας λέξης, όταν οι έννοιές της δεν σχετίζονται μεταξύ τους η μία με την άλλη. Εάν κάτι τέτοιο ισχύει, ένα μεγάλο παράθυρο θα ήταν περισσότερο αποτελεσματικό για το καθορισμό των διαφόρων εννοιών. Αυτό εξάλλου παρατηρείται και σε αλγορίθμους που χρησιμοποιούν ″δεδομένα για εκπαίδευση″ (training data). Εφόσον οι θεματικές ενδείξεις του πλαισίου (topical contextual clues) για τον καθορισμό της έννοιας εμφανίζονται μέσα στο κείμενο, οι στατιστικές προσεγγίσεις που χρησιμοποιούν ″δεδομένα για εκπαίδευση″ για να καλύψουν το πρόβλημα των ″αραιών δεδομένων″ (sparseness), αυξάνουν το μέγεθος του παραθύρου του πλαισίου.

Ο Gale και οι άλλοι [31] βρήκαν ότι ο Bayesian ταξινομητής τους αποδίδει πολύ πιο καλά σε μεγάλα παράθυρα και προσδιόρισαν ένα βέλτιστο μήκος παραθύρου ± 50 λέξεις γύρω από την προς αποσαφήνιση λέξη.

Στον δικό μας αλγόριθμο θα χρησιμοποιήσουμε μια λίγο διαφορετική προσέγγιση για τον καθορισμό του μήκους του πλαισίου. Δεν θα μετρήσουμε λέξεις γύρω από την προς αποσαφήνιση λέξη αλλά θα εργαστούμε με αυτόνομες γραμματικές προτάσεις. Δηλαδή κάθε λέξη θα αποσαφηνίζεται μέσα στο πλαίσιό της πού θα θεωρείται ότι αποτελείται απ΄ όλες τις άλλες λέξεις πού βρίσκονται στην ίδια πρόταση με την προς αποσαφήνιση λέξη. Κάτι τέτοιο μας απαλλάσσει από αυθαίρετες εκτιμήσεις γύρω από το μήκος του ″παραθύρου″ και αυτή είναι εξάλλου ακριβώς και η μορφή των δεδομένων που θα χρησιμοποιήσουμε για εκτίμηση της αποδοτικότητας του αλγορίθμου, που δίνονται από τον επίσημο οργανισμό SensEval, υπεύθυνο για την διοργάνωση διαγωνισμών για συστήματα Αποσαφήνισης λέξης. Σε αυτά τα δεδομένα (Senseval-2 English lexical sample data) που θα περιγράψουμε παρακάτω, η κάθε λέξη πού ζητείται να αποσαφηνισθεί δίδεται με το πλαίσιό της, το οποίο αποτελείται συνήθως από μία εως τρείς γραμματικές περιόδους.

Χρησιμοποιώντας λοιπόν σαν πλαίσιο της προς αποσαφήνιση λέξης την περίοδο στην οποία αυτή εμφανίζεται, θα δημιουργήσουμε το δείγμα από το οποίο θα προκύψουν οι πρατηρηθείσες συχνότητες που μας χρειάζονται για την εφαρμογή του X^2 στατιστικού ελέγχου. Η δημιουργία γίνεται ως εξής: Για κάθε λέξη του πλαισίου,

συμπεριλαμβανομένης και της προς αποσαφήνιση λέξης, για όλες τις έννοιες και για όλα τα μέρη του λόγου βρίσκουμε από το WordNet τα συσχετιζόμενα synsets για κάθε συγκεκριμένη σχέση και σχηματίζουμε από αυτά το σύνολο των συσχετιζόμενων synsets για το πλαίσιο.

Όπως τονίσαμε παραπάνω, θα εφαρμόσουμε τον αλγόριθμό μας με δύο εκδοχές, την μία χρησιμοποιώντας όλες τις σχέσεις και την άλλη τον ιδιαίτερο εκείνο αποδοτικό συνδυασμό σχέσεων *antonymy*, *hypernymy* και *hyponymy*.

4.5.2 Το Σύνολο των Συσχετιζόμενων Synsets για τις Έννοιες

Όταν μας ζητείται να αποσαφηνίσουμε μια λέξη πού εμφανίζεται σε ένα πλαίσιο μας δίδεται εκτός από την ίδια την λέξη και το μέρος του λόγου (*pos*) (part of speech). Επομένως, για κάθε έννοια της προς αποσαφήνιση λέξης και για το συγκεκριμένο μέρος του λόγου πού μας δίδεται ψάχνουμε στο WordNet τα συσχετιζόμενα synsets για κάθε μια διαθέσιμη σχέση. Με αυτό τον τρόπο δημιουργούμε ξεχωριστά σύνολα από synsets για κάθε μια έννοια της προς αποσαφήνιση λέξης.

Και σε αυτή την περίπτωση, όπως τονίσαμε και για το σύνολο του πλαισίου δημιουργούμε δύο εκδοχές συνόλων από συσχετιζόμενα synsets, την μία φορά χρησιμοποιώντας όλες ανεξαίρετα τις διαθέσιμες από το Wordnet σχέσεις και την άλλη χρησιμοποιώντας τον συγκεκριμένο συνδυασμό των σχέσεων *antonymy*, *hypernymy* και *hyponymy*.

4.5.3 Υλοποίηση του X-τετράγωνο ως Ελέγχου Καλού Ταιριάσματος για Κανονικότητα

Όπως αναφέραμε και στην αρχή, δεν ψάχνουμε για ποσοτικά μέτρα σύγκρισης μεταξύ του πλαισίου και των διαφόρων εννοιών της προς αποσαφήνιση λέξης, όπως κάνουν οι περισσότερες κλασικές μέθοδοι για Αποσαφήνιση Λέξης. Αντιθέτως, επιχειρούμε μια διαφορετική προσέγγιση στο θέμα.

Για κάποια χαρακτηριστικά αναμένουμε ότι η σωστή έννοια (correct sense) θα επιδείξει μια διαφορετική συμπεριφορά απ΄ ό,τι οι άλλες έννοιες. Αυτή τη διαφορετική συμπεριφορά *(non-normal behavior)* προσπαθούμε να την εντοπίσουμε με την βοήθεια

του X-τετράγωνο ελέγχου καλού ταιριάσματος για κανονικότητα και να την χρησιμοποιήσουμε σαν ένδειξη για την σωστή έννοια της πρός αποσαφήνιση λέξης.

Ας χρησιμοποιήσουμε φορμαλισμό για να γίνει κατανοητή η εφαρμογή αυτού του ελέγχου.

Εστω X_i μια τυχαία μεταβλητή (random variable) η οποία μετράει τον αριθμό των εμφανίσεων του *i-οστού* συσχετιζόμενου synset μιας έννοιας μέσα στο δείγμα του πλαισίου στο οποίο εμφανίζεται η προς αποσαφήνιση λέξη. Για την σύνθετη συνάρτηση κατανομής πιθανοτήτων των μεταβλητών, X_i, (composite probability distribution function: pdf) , διατυπώνουμε την μηδενική υπόθεση για κανονικότητα (normality) των συσχετιζόμενων synsets στο δείγμα.

Δηλαδή, ισχυριζόμαστε ότι οι συχνότητες των συσχετιζόμενων synsets για τις διάφορες έννοιες της προς αποσαφήνιση λέξης κατανέμονται κανονικά η προσεγγιστικά κανονικά στο δείγμα του πλαισίου. Αναφέραμε την λέξη προσεγγιστικά γιατί πραγματικά, όπως θα γίνει καταννοητό πιο κάτω, δεν ενδιαφερόμαστε επακριβώς για το πόσο "καλά" η κατανομή για μια έννοια κατανέμεται κανονικά, αλλά για την "κατανεμητική" (distributative) διαφορά μεταξύ των διαφόρων εννοιών.

Για να υλοποιήσουμε ένα X-τετράγωνο έλεγχο για κανονικότητα εργαζόμαστε ως ακολούθως [32]: Για να κάνουμε εφικτό τον υπολογισμό των p-τιμών από την κανονική κατανομή (χρησιμοποιώντας ένα στατιστικό πρόγραμμα ή τους πίνακες κανονικής κατανομής), χρειάζεται να προχωρήσουμε σε τυποποίηση (standardization) των τιμών του τυχαίου δείγματος της μεταβλητής X. Το τυχαίο δείγμα X ολισθαίνεται κατά το εκτιμώμενο στατιστικό της μέσο (mean) και κανονικοποιείται κατά την εκτιμώμενη τυπική της απόκλιση (standard deviation).

$$Z = \frac{X - \mu}{\sigma} \tag{4.3}$$

Για να δημιουργήσουμε πινακοποιημένα δεδομένα (binned data) για την θεωρούμενη κανονική κατανομή, επιλέγουμε τα διαστήματα X_b (bins) με ίσο μήκος. Επίσης για να αποφύγουμε μη επαρκείς αναμενόμενες συχνότητες για ορισμένα synsets πού εμφανίζονται σπάνια χωρίσαμε την ευθεία των πραγματικών αριθμών στα διαστήματα: (-∞ -1.6 -1.2 -0.8 -0.4 0.4 0.8 1.2 1.6 ∞). Οι παρατηρηθείσες συχνότητες προκύπτουν

από τα πινακοποιημένα δεδομένα των τιμών της μεταβλητής Z μέσα σε αυτά τα διαστήματα, ενώ οι αναμενόμενες συχνότητες είναι οι πιθανότητες των διαστημάτων X_b σύμφωνα με την κανονική κατανομή. Αυτές υπολογίζονται από τον ακόλουθο τύπο:

$$Expected_i = \frac{1}{2} N \left[\frac{2}{\sqrt{\pi}} \int_{Xb_i}^{\infty} e^{\frac{-Xb_i^2}{2}} dXb_i \right]' \tag{4.4}$$

Όπου N το μέγεθος του X και X_b τα διαστήματα ελέγχου.

Έχοντας τώρα τις αναμενόμενες και παρατηρηθείσες συχνότητες υπολογίζουμε την X^2 τιμή από την εξίσωση 4.1 και τις αντίστοιχες p-τιμές από την X-τετράγωνο συνάρτηση κατανομής για ένα επίπεδο σημαντικότητας .05 και *"αριθμό διαστημάτων - 3"* βαθμούς ελευθερίας (αφαιρούμε 3 γιατί αποκλείουμε τις ακραίες τιμές)

Οι παραπάνω υπολογισμοί γίνονται για κάθε μια έννοια της προς αποσαφήνιση λέξης ξεχωριστά.

Η έννοια με την μικρότερη p-τιμή επιλέγεται σαν η σωστή έννοια (correct sense). Σε μια πιό ελεύθερη έκφραση θα λέγαμε ότι επιλέγεται η έννοια με την πιο μη κανονική συμπεριφορά (non normal behavior).

Δίνουμε παρακάτω ένα παράδειγμα με πραγματικά δεδομένα.

4.5.4 Παράδειγμα Αποσαφήνισης με την Βοήθεια του Αλγορίθμου μας

Ας εφαρμόσουμε τον αλγόριθμό μας για να αποσαφηνίσουμε μια εμφάνιση, ή στιγμιότυπο (instance) όπως αποκαλείται, του ουσιαστικού *art* μέσα σε ένα πραγματικό πλαίσιο δανεισμένο από τα δεδομένα του Senseval-2 διαγωνισμού. Στο παράδειγμα αυτό για να σχηματίσουμε τα σύνολα των συσχετιζόμενων synsets, τόσο για το πλαίσιο όσο και την καθεμία έννοια ξεχωριστά, θα κάνουμε χρήση όλων των διαθέσιμων σχέσεων του WordNet, και για όλα τα μέρη του λόγου (part-of-speech categories).

Το στιγμιότυπο για την λέξη *art* εμφανίζεται παρακάτω.

<instance id="art.40019" docsrc="bnc_BM9_1279">

<context>

Quickly getting off the bus, he ran to where the plane had impacted and dragged out the injured pilot who was covered in oil moments before the plane caught fire. Alfred Reginald Thomson, R.A., R.P., (1894–1979) — War Artist All branches of the armed services at various times made appointments of official War

Artists, who were commissioned to paint battle scenes or portraits for the armed services. In 1942 one such appointment was made of a deaf artist, Alfred Reginald Thomson, as official War Artist to the Royal Air Force. A.R. Thomson was born in Bangalore, India, in 1894 and was educated at the Royal School for Deaf Children, Margate, in England before he went to study ¡head¿art¡/head¿at the London Art School, Kensington, and exhibited at the Royal Academy from 1920.
</context>
</instance>

Σε αυτό το στιγμιότυπο εμφανίζονται η target λέξη *art*, δηλαδή η λέξη η οποία πρόκειται να αποσαφηνισθεί, η οποία δηλώνεται με το tag <head>, και οι περιβάλλουσες λέξεις οι οποίες συναπαρτίζουν το πλαίσιο της προς αποσαφήνιση λέξης.

Το ουσιαστικό *art* εμφανίζεται στο WordNet με τέσσερεις έννοιες. Από το συνοδευτικό αρχείο για τα στιγμιότυπα των δεδομένων του διαγωνισμού Senseval-2 (αυτό είναι ένα αρχείο που περιέχει τις σωστές απαντήσεις για κάθε στιγμιότυπο), βρίσκουμε ότι η σωστή έννοια εδώ για το παράδειγμά μας είναι η έννοια 2.

Ας δούμε τώρα πώς δουλεύει ο αλγόριθμός μας για να επιλέξει την σωστή έννοια από τις τέσσερεις υποψήφιες έννοιες της λέξης *art*.

Κατά πρώτον ενδιαφερόμαστε να κρατήσουμε (extract) από αυτό το απόσπασμα κειμένου μόνο τις λεγόμενες ″opened-words″, δηλαδή τις λέξεις οι οποίες έχουν μια καταχώριση στο WordNet.

Για κάθε opened-word, για κάθε έννοια αυτής και για κάθε διαθέσιμη σχέση, ψάχνουμε στο WordNet και συλλέγουμε όλα τα συσχετιζόμενα synsets (υπενθυμίζουμε ότι τα synsets είναι σύνολα από συνώνυμες λέξεις που αναπαριστούν μια έννοια). Αυτό είναι και το σύνολο των συσχετιζόμενων synsets για το πλαίσιο. Το προκύπτον σύνολο αποτελείται από ακριβώς 8775 διακριτά synsets.

Το ουσιαστικό *art* έχει τις ακόλουθες τέσσερεις έννοιες στο WordNet:

Sense 1. {*art, fine_art*}

Sense 2. {*art, artistic_creation, artistic_production*}

Sense 3. {*art, artistry, prowess*}

Sense 4. {*artwork, art, graphics, nontextual_matter*}

Συσχετιζόμενα Synsets	Συχνότητα στο Πλαίσιο
slowly, slow, easy, tardily	1
quick, speedy	8
flying, quick, fast	4
deed, feat, effort, exploit	2
acquisition	1
obtainment, obtention	1
catching, contracting	1
appropriation	1
occupation, occupancy, taking_possession, moving in	1
capture, gaining_control, seizure	1
reception, receipt	1
pickup	1
derring	3
Tour_de_force	1
departure, going, going away, leaving	1
Running_away	1
egress, egression, emergence	1
acquiring, getting	1
disposal, disposition	5
rally, rallying	1

Table 4.1: *Ενα απόσπασμα των 20 πρώτων από τα 8775 συσχετιζόμενων synsets και των συχνοτήτων τους όπως δημιουργήθηκαν από το προγραμμά μας για το στιγμιότυπο art.*40019

Δημιουργούμε τώρα για κάθε έννοια και για όλες τις παρεχόμενες σχέσεις από το WordNet τα αντίστοιχα σύνολα από συσχετιζόμενα synsets και μετράμε τις συχνότητες στο σώμα του πλαισίου. Τα αποτελέσματα εμφανίζονται παρακάτω:

(Sense 1) 305 διακριτά synsets οι συχνότητες των οποίων κατανέμονται ως ακολούθως:

{*5,3,3,5,5,3,3,3,18,3,3,7,3,3,3,18,16,3,8,3,3,3,5,3,3,3,16,3,7,47,47,47*}

(Sense 2) 91 διακριτά synsets οι συχνότητες των οποίων κατανέμονται ως ακολούθως:

{*14,3,3,5,3,3,3,3,4,3,3,3,5,24,3,3,3,3*}

(Sense 3) 42 διακριτά synsets οι συχνότητες των οποίων κατανέμονται ως ακολούθως:

{*5,3,3,3,3,3,3,3,3,3,3,3,3,4*}

(Sense 4) 63 διακριτά synsets οι συχνότητες των οποίων κατανέμονται ως ακολούθως:

{*7,3,7,7,3,3,3,3,18,3,3,3*}

Υπολογίζοντας τις p-τιμές για τις παραπάνω κατανομές το πρόγραμμά μας επιστρέφει: $ps_1 = 3.723e{-}011$, $ps_2 = 4.3188e{-}014$, $ps_3 = 1.7569e{-}008$, $ps_4 = 1.3703e{-}005$. Πράγματι, η πιο μικρή p-τιμή είναι αυτή πού αντιστοιχεί στη έννοια 2 της προς αποσαφήνιση λέξης *art*.

Ο αλγόριθμος πολύ σωστά διάλεξε σαν σωστή έννοια την έννοια 2 με την μικρότερη p-τιμή.

Ανακεφαλαιώνοντας τα βασικά βήματα για όλη την διαδικασία αποσαφήνισης (word sense disambiguation procedure) δίνουμε παρακάτω τον αλγόριθμο σε ψευδοκώδικα.

4.5.5 Ο Αλγόριθμος σε Ψευδοκώδικα

```
void Create_Context_SetofRelatedSynsets() {
    for each word wi of the context
        for each part_of_speech of the wi
            for each available relation from WordNet
                Extract_from_WordNet_the_RelatedSynsets();

void disambiguate() {
    Read(target_word, Part_of_speech, context);
```

```
    Create_Context_SetofRelatedSynsets();
    For each sense si of the target word
   {
        Create_Sense_SetodRelatedSynsets();
        Calculate_p-value();
    }
    Select as correct sense that with the minimum p-value
}.
```

4.6 Τα Δεδομένα Αποτίμησης

Εκτιμήσαμε την αποδοτικότητα του αλγορίθμου μας στην Αποσαφήνιση ελέγχοντάς τον επάνω στα δεδομένα [29] πού εδόθησαν για έλεγχο στον επίσημο διαγωνισμό για συστήματα Αποσαφήνισης Λέξης το καλοκαίρι του 2002 και που διοργανώνεται κάθε 2 χρόνια από τον οργανισμό SensEval , ο οποίος έχει συσταθεί για τον σκοπό αυτό.

Αυτά τα δεδομένα ελέγχου είναι ένα πολύ εκτεταμένο σώμα της Αγγλικής γλώσσας το οποίο δημιουργήθηκε παίρνοντας κείμενα από το BNC-2, και το Penn Treebank (συμπεριλαμβάνει κείμενα απο τη Wall Street journal, Brown, IBM manuals, live web σελίδες). Το λεξικό το οποίο παρέχει τον κατάλογο με τις έννοιες για κάθε λέξη (sense inventory) είναι το WordNet version 1.7.1.

Τα δεδομένα ελέγχου, όπως επίσης οι βαθμολογίες τις οποίες πέτυχαν ένας μεγάλος αριθμός από διαγωνιζόμενα συστήματα, αλλά και δεδομένα για εκπαίδευση, είναι όλα διαθέσιμα ελεύθερα από το web site του οργανισμού senseval-2.

Οσον αφορά το σώμα κειμένων επάνω στο οποίο ελέγξαμε την αποδοτικότητα του αλγορίθμου μας αποτελείται από δύο σύνολα δεδομένων: τα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου. Όλες οι λέξεις που ζητείται να αποσαφηνισθούν στα δύο αυτά σύνολα ανήκουν σε μία από τις τρείς κατηγορίες: ουσιαστικό, ρήμα και επίθετο. Όλα τα στιγμιότυπα του σώματος του κειμένου έχουν επανελεγθεί και βρέθηκε με συνέπεια ότι ανήκουν στην σωστή κατηγορία.

Αυτό αποτελεί και μια ευκολία για τους διαγωνιζόμενους, αλλά και για το δικό μας σύστημα, διότι η διαδικασία αποσαφήνισης αποφεύγει το επί πλέον βάρος της

υποσημείωσης του μέρους του λόγου.

Ο αλγόριθμός μας είναι ένας unsupervised αλγόριθμος, με την έννοια ότι δεν χρειάζεται εκπαίδευση (training) πάνω σε κάποιο σώμα εκπαίδευσης *(training* set). Επομένως, χρησιμοποιούμε μόνο το σύνολο των δεδομένων για έλεγχο. Αυτό το σύνολο αποτελείται από 73 εργασίες (tasks) όπου κάθε εργασία αποτελείται από πολλά στιγμιότυπα (instances) της πρός αποσαφήνιση λέξης στόχου μέσα σε αποσπάσματα κειμένου.

Κάθε τέτοιο στιγμιότυπο έχει αποσαφηνισθεί προσεκτικά από ειδικούς λεξικογράφους και στην πρός αποσαφήνιση λέξη έχει αποδοθεί η σωστή έννοια από τον κατάλογο του WordNet. Η αποστολή κάθε αλγορίθμου αποσαφήνισης είναι να βρεί αυτή την σωστή έννοια.

Συνοψίζοντας, κάθε στιγμιότυπο αποτελείται από μια πρόταση η οποία περιέχει την προς αποσαφήνιση λέξη και μία έως τρείς άλλες περιβάλλουσες περιόδους που συναποτελούν το πλαίσιο (context). Ένα τυπικό παράδειγμα στιγμιότυπου για το ουσιαστικό της Αγγλικής γλώσσας *art* είναι το ακόλουθο.

<lexelt item="art.n">
<instance id="art.40003" docsrc="bnc_A04_1181">
<context>
Whatever flickerings of potential this young tyro possesses, they cannot cover up the fact that he is a painter with the imagination of a retarded adolescent; no technical mastery; no intuitive feeling for pictorial space; no sensitivity towards, or grasp of, tradition; and a colour sense rather less than that of Congo, the chimpanzee who was taught (among other things) a crude responsiveness to colour harmonies by Desmond Morris in the late 1950s. However, potentially educable as a painter Schnabel may or may not be, his work is just not worthy of serious attention by anyone with a developed taste in this particular art form. Readers need also to be wary of the existence of special markets. The explosive prices for Teddy Bears in the last few years indicate how a market can be created, in this case by a mix of merit and nostalgia. What is clearly a dealers' market is often signalled by the invention of a brand name to group together a variety of material, perhaps rather disparate.

Pop <head>Art</head> is an example.

</context>

<instance>

Η πρός αποσαφήνιση λέξη εμφανίζεται μέσα στο tag <head> και το tag < *instanceid...* > καθορίζει ένα μοναδικό κλειδί (identifier) για το συγκεκριμένο στιγμιότυπο, *"art.40003"* στην συγκεκριμένη περίπτωση. Αυτό το μαναδικό κλειδί αντιστοιχίζεται με το αριθμό της έννοιας του WordNet πού έχει αποδοθεί από τους λεξικογράφους για το συγκεκριμένο στιγμιότυπο και φυλάσσεται στο λεγόμενο αρχείο κλειδιών (key file) το οποίο περιέχει τις σωστές απαντήσεις για κάθε ένα στιγμιότυπο.

Ο σκοπός μας είναι να έχουμε όσες το δυνατόν περισσότερες επιτυχίες όταν συγκρίνουμε τις απαντήσεις του συστήματός μας με αυτές του αρχείου κλειδιών.

4.7 Αποτίμηση της Αποδοτικότητας του Προτεινόμενου Αλγορίθμου

Χωρίσαμε τα στιγμιότυπα των κειμένων του σώματος ελέγχου σε τρείς κατηγορίες για κάθε ένα μέρος του λόγου, δηλαδή μια κατηγορία για ουσιαστικά, μια για ρήματα και μια για επίθετα (επιρρήματα δεν συμπεριλαμβάνονται στα δεδομένα ελέγχου). Για κάθε μέρος του λόγου συγκεντρώσαμε από τα δεδομένα ελέγχου μόνο τα στιγμιότυπα που ανήκουν σε αυτή την συγκεκριμένη κατηγορία, και δημιουργήσαμε με αυτό τον τρόπο τρία ξεχωριστά σώματα ελέγχου. Το γεγονός αυτό μας έδωσε την δυνατότητα να μετρήσουμε ξεχωριστά την αποδοτικότητα του αλγορίθμου μας για κάθε κατηγορία. Για ένα πολύ μικρό αριθμό στιγμιοτύπων για τα οποία δεν παρέχεται ο αριθμός της σωστής έννοιας του WordNet στο αρχείο κλειδιών, αυτά εξαιρέθηκαν από τα δεδομένα ελέγχου. Επίσης εξαιρέσαμε ένα πολύ μικρό αριθμό στιγμιοτύπων στα οποία η πρός αποσαφήνιση λέξη δεν ήταν μοναδική αλλά αποτελούσε collocation (δύο λέξεις μαζί πού τις χειριζόμαστε σαν μια έννοια). Στην περίπτωση αυτή αποδίδεται από τους λεξικογράφους όχι μόνο μια σωστή έννοια από το WordNet, αλλά από μια έννοια πού αντιστοιχεί σε κάθε ένα από τα συστατικά μέρη του collocation.

Για να εκτιμήσουμε την αποδοτικότητα ενός συστήματος αποσαφήνισης λέξης χρη-

σιμοποιούμε μέτρα ανάλογα με αυτά που χρησιμοποιούμε στην αναζήτηση πληροφορίας (information retrieval) και τα οποία ορίσαμε στην εισαγωγή της παρούσης διατριβής. Για την εκτίμηση των συστημάτων στην αναζήτηση πληροφορίας αλλά και γενικότερα ενός στατιστικού μοντέλου επεξεργασίας φυσικής γλώσσας κάνουμε χρήση συνήθως των εννοιών *precision* και *recall*. Αν θεωρήσουμε τα ″στοιχεία″ για τα οποία ένα σύστημα θέλουμε να τα επιλέξει σαν ένα σύνολο στόχο (target set) και το οποίο σύνολο στόχος είναι γενικότερα ένα μέρος από μια μεγαλύτερη συλλογή, τότε το *precision* ορίζεται σαν το μέτρο της αναλογίας των επιλεγμένων στοιχείων στα οποία το σύστημα αποφάσισε ″σωστά″, ενώ το recall ορίζεται σαν το μέτρο της αναλογίας των στοιχείων του συνόλου στόχου τα οποία επέλεξε το σύστημα.

Στην δικιά μας περίπτωση, όλα τα δεδομένα ελέγχου είναι η συλλογή αλλά και ταυτόχρονα το σύνολο στόχος διότι το σύστημά μας θέλουμε να επιλέξει την σωστή έννοια για όλα τα στιγμιότυπα. Επομένως το *precision* και *recall* είναι ίδια.

Στον διαγωνισμό συστημάτων αποσαφήνισης λέξης πού διοργανώθηκε από τον οργανισμό Senseval-2 χρησιμοποιήθηκε σαν μέτρο της αποδοτικότητας το F-measure, το οποίο είναι συνδυασμός του *precision* και *recall* και δίνεται από τον ακόλουθο τύπο:

$$F = \frac{1}{a\frac{1}{p} + (1-\alpha)\frac{1}{R}} \tag{4.5}$$

Όπου P είναι το precision, R είναι το recall και α ένας συντελεστής βάρους ο οποίος καθορίζει την σημαντικότητα πού δίνεται στο precision και στο recall. Αυτός ο τύπος απλοποιείται στο $2PR/(P+R)$ στην περίπτωση που ισοδύναμη σημαντικότητα αποδίδεται και στα δύο ($\alpha = 1/2$).

Στην περίπτωσή μας έχουμε $P = R = F$.

Ο πίνακας 4.2 εμφανίζει τα αποτελέσματα τα οποία λάβαμε και για τα τρία μέρη του λόγου όταν εκτιμήσαμε την αποδοτικότητα του συστήματός μας πάνω στα δεδομένα ελέγχου του διαγωνισμού Senseval-2 χρησιμοποιώντας όλες τις διαθέσιμες σχέσεις από το Wordnet.

Εάν περιοριστούμε σε ένα συγκεκριμένο συνδυασμό των τριών σχέσεων *antonymy*, *hyponymy*, *hypernymy*, όπως αναφέραμε και προηγούμενα, επιτυγχάνουμε μια καλύτερη αποδοτικότητα (0.33) του αλγορίθμου αποσαφήνισης, πίνακας 4.3.

Αποτελέσματα αποτίμησης με χρήση όλων των σχέσεων			
Μέρος του Λόγου	*Precision*	*Recall*	*F-measure*
Ουσιαστικά	0.28	0.28	0.28
Ρήματα	0.11	0.11	0.11
Επίθετα	0.27	0.27	0.27
Σύνολο	0.22	0.22	0.22

Πίνακας 4.2: *Αποτελέσματα Αποτίμησης της μεθόδου μας πάνω στα δεδομένα ελέγχου του διαγωνισμού Senseval-2 , χρησιμοποιώντας όλες τις διαθέσιμες σχέσεις του WordNet*

Αποτελέσματα με χρήση των σχέσεων Antonymy, Hyponymy, Hypernymy			
Μέρος του Λόγου	*Precision*	*Recall*	*F-measure*
Ουσιαστικά	0.37	0.37	0.37
Ρήματα	0.23	0.23	0.23
Επίθετα	0.38	0.38	0.38
Σύνολο	0.33	0.33	0.33

Πίνακας 4.3: *Αποτελέσματα αποτίμησης χρησιμοποιώντας μόνο τις σχέσεις (Antonymy, Hyponymy, Hypernymy)*

Επειδή δεν χρησιμοποιήσαμε καθόλου εκπαίδευση (πάνω στα δεδομένα για εκπαίδευση που παρέχονται από τον διαγωνισμό Senseval-2), το σύστημά μας πρέπει να συγκριθεί μόνο με τα συμμετέχοντα συστήματα πού χρησιμοποίησαν unsupervised τεχνικές. Για να έχουμε μια εκτίμηση της αποδοτικότητας του αλγορίθμου μας σε σύγκριση με τα άλλα συστήματα που συμμετείχαν στον διαγωνισμό αποσαφήνισης λέξεων, παραθέτουμε τον πίνακα 4.4 από το [25]. Σε αυτόν τον πίνακα φαίνονται μαζί οι τιμές για *Recall* και *F-measure* τόσο για τούς συμμετέχοντες στον διαγωνισμό όσο και για το σύστημά μας.

Το σύστημά μας χρησιμοποιώντας μόνο το WordNet σαν λεξικολογική πηγή επιτυγχάνει αποδοτικότητα 0.333 για Recall και F-measure, αποδοτικότητα συγκρίσιμη με την αποδοτικότητα των δύο πρώτων συστημάτων στο διαγωνισμό senseval-2 . Και τα δύο αυτά συστήματα κάνουν χρήση ενός σημαντικού αριθμού από εξωτερικά γλωσσολογικά δεδομένα (corpora) κατά την διαδικασία αποσαφήνισης των λέξεων.

Το UNED-LS-U σύστημα [36] ενσωματώνει πληροφορία από 3,200 βιβλία στα

Όνομα συμμετέχοντα	Recall	F-measure
UNED - LS - U	0.401	0.401
Ο αλγόριθμός μας (AHH)	**0.333**	**0.333**
ITRI - WASPS - WorkBench	0.319	0.412
CL Research - DIMAP	0.293	0.293
IIT 2 (R)	0.244	0.245
IIT 1 (R)	0.239	0.241
IIT 2	0.232	0.232
IIT 1	0.220	0.220
Lesk Implementation	0.183	0.183
Random	0.141	0.141

Πίνακας 4.4: *Αποδοτικότητα των (unsupervised) συστημάτων που συμμετείχαν στον Senseval-2 διαγωνισμό καθώς και του συστήματός μας κατανομής σχέσων του WordNet*

Αγγλικά από το Gutenberg Project για να δημιουργήσει έναν πίνακα σχετικότητας μεταξύ των λέξεων (relevance matrix), όπου η σχετικότητα εξαρτάται από τις αποστάσεις μεταξύ των λέξεων, όπως αυτές εκτιμούνται πάνω στο διαθέσιμο σώμα κειμένων. Για να αποσαφηνίσει τις έννοιες των λέξεων χρησιμοποιείται μια τεχνική βασιζόμενη στην πληροφορία από τον πίνακα σχετικότητας και τα πέντε πρώτα hyponyms από το Wordnet.

Το σύστημα WASPS-Workbench [37] είναι ένα γλωσσολογικό εργαλείο το οποίο ολκληρώνει δύο συστατικά μέρη: λεξικογραφία (lexicography) και αυτόματο σύστημα αποσαφήνισης λέξης. Για την αποσαφήνιση των εννοιών των λέξεων, το Workbench υπολογίζει το ″Word-Sketch″: μία σελίδα από στατιστικώς σημαντικά collocation patterns για αυτή την λέξη υπολογισμένα πάνω στο σώμα κειμένων BNC (British National Corpus). Με βάση αυτά τα πρότυπα (patterns) και χρησιμοποιώντας το Wordnet σαν κατάλογο εννοιών, αποδίδει έννοιες σε συγκεκριμένα πρότυπα. Αυτές οι αποδόσεις χρησιμοποιούνται έπειτα σαν ″σπόροι″ (seeds) για ένα bootstrapping αλγόριθμο, ο οποίος ″συλλαμβάνει″ την σωστή έννοια χρησιμοποιώντας μια λίστα απόφασης από ενδείξεις που συνάγονται από γραμματικά σχετιζόμενα πρότυπα.

4.8 Συμπέρασμα

Σε αυτό το κεφάλαιο παρουσιάσαμε μια μέθοδο για το πώς οι στατιστικοί έλεγχοι "καλού ταιριάσματος" μπορεί να φανούν χρήσιμοι για ένα σύστημα αποσαφήνισης λέξης. Σε όλα σχεδόν τα στατιστικά συστήματα χρησιμοποιείται μια θεωρητική κατανομή για την παραγωγή των αποτελεσμάτων. Η κοινή πρακτική είναι να κάνουμε μια υπόθεση για κάποια συγκεκριμένη κατανομή για τα δεδομένα μας και να χρησιμοποιήσουμε έπειτα κάποιες τυπικές στατιστικές διαδικασίες για να εκτιμήσουμε την θεωρούμενη κατανομή. Αυτές οι τυπικές στατιστικές διαδικασίες είναι οι στατιστικοί έλεγχοι "καλού ταιριάσματος".

Στον αλγόριθμό μας διαφοροποιείται ελαφρά η τεχνική γιατί δεν ενδιαφερόμαστε να βαθμολογήσουμε απόλυτα πόσο καλά "ταιριάζει" η κανονική κατανομή στα δεδομένα μας. Η παραδοχή για την κανονικότητα γίνεται για να δημιουργήσει ένα μοντέλο κατανομής σαν αναφορά (a reference distribution model), σαν μια βάση σύγκρισης για το σκοπό της διάκδρισης (decrimination) μεταξύ των ανταγωνιζόμενων synset κατανομών των διαφόρων εννοιών. Φυσικά θα είχε ενδιαφέρον να δοκιμάσουμε και μια διαφορετική παραδοχή για τα δεδομένα, για παράδειγμα η Weibull κατανομή πιθανόν να ήταν μια καλή επιλογή. Ή, θα μπορούσαμε να δοκιμάσουμε και άλλους ενναλλακτικούς στατιστικούς ελέγχους όπως Kolmogorov-Smirnov και Anderson-Darling ελέγχους [33].

Ένα άλλο χαρακτηριστικό του Wordnet το οποίο μπορεί να είναι χρήσιμο στην διαδικασία αποσαφήνισης λέξης είναι τα glosses [25], [34], [35]. Ο κάθε ορισμός στο Wordnet όπως έχουμε πεί αποτελείται από τις συνώνυμες λέξεις (synsets), το μέρος του ορισμού (defining part) και στην πλειονότητα των περιπτώσεων μερικά παραδείγματα τυπικής χρήσης της έννοιας σε πραγματικά συμφραζόμενα. Μία κατάλληλη μοντελοποίηση της κατανομής των λέξεων πού περιέχονται στα glosses είναι πολύ πιθανόν να είναι καλοί "δείκτες" (indicators) για την έννοια που ορίζουν.

Επίσης μια αποτίμηση της αποδοτικότητας του αλγορίθμου με την χρήση και διαφορετικών συνδυασμών των σχέσεων του Wordnet θα είχε ενδιαφέρον.

Τα παραπάνω αποτελούν κατά την γνώμη μας ενδιαφέροντα χαρακτηριστικά πού ενδεχόμενως επηρεάζουν την αποδοτικότητα και θα πρέπει να αποτελέσουν το αντκείμενο μιας μελλοντικής εργασίας.

Κεφάλαιο 5

Επίλογος

Κλείνοντας την παρούσα διατριβή έχουμε καταλήξει στα ακόλουθα συμπεράσματα.

Οι στατιστικές μέθοδοι είναι κατά γενική ομολογία οι μέθοδοι πού έχουν εφαρμοσθεί με την μεγαλύτερη επιτυχία στην επεξεργασία φυσικής γλώσσας. Παραδείγματα αποτελούν τα στατιστικά συστήματα για τα κλασσικά προβλήματα της εύρεσης collocations, αναζήτηση κειμενικής πληροφορίας και αποσαφήνιση της έννοιας μιας λέξης.

Η θέση μας αυτή σκοπό είχε να παρουσιάσει την εφαρμογή μιας εννιαίας στατιστικής μεθόδου για την επίλυση προβλημάτων στον τομέα της επεξεργασίας φυσικής γλώσσας πού εμφανίζουν μια ομοιότητα ως προς το στόχο, αυτόν της επιλογής μεταξύ ανταγωνιζόμενων οντοτήτων. Η μέθοδος αυτή στηρίχτηκε στην χρήση των στατιστικών ελέγχων (statistical tests) και εφαρμόστηκε σε τρία θεματικά πεδία της επεξεργασίας φυσικής γλώσσας: Εξαγωγή collocations μέσα από ένα μεγάλο σώμα κειμένων, αναζήτηση κειμενικής πληροφορίας από "δεξαμενές" εγγράφων και αποσαφήνιση της έννοιας μιας λέξης από τα συμφραζόμενα της.

Πιό συγκεκριμένα χρησιμοποιήσαμε την μεθοδολογία των στατιστικών ελέγχων "καλού ταιριάσματος" με ένα ειδικά προσαρμοζόμενο τρόπο για την κάθε περίπτωση, ώστε να μπορεί να εφαρμοσθεί και στις τρείς παραπάνω εργασίες.

Πετύχαμε το στόχο μας αναπτύσσοντας συστήματα που επιδεικύουν καλή συμπεριφορά και αποδόσεις καλύτερες σε πολλές περιπτώσεις από αυτές των κλασικών συστημάτων στα αντίστοιχα θεματικά πεδία.

Νομίζουμε ότι τόσο για την βελτίωση αυτών των συστημάτων, όσο και για την εφαρμογή της μεθοδολογίας των στατιστικών ελέγχων "καλού ταιριάσματος" και σε

άλλες περιοχές της επεξεργασίας φυσικής γλώσσας, που εμφανίζονται παρόμοιες στον στόχο, θα άξιζε να ασχοληθούμε συστηματικότερα στο μέλλον.

Αθήνα 2005

Κώστας Τ. Φράγγος

Index

Βιβλιογραφία

[1] G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing. Journal of the ACM, 15(1):8-36, January (1968).

[2] G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice hall Inc., Englewood Cliffs, Nj, (1971).

[3] S. E. Robertson, K. Sparck Jones. Relevance weighting of search terms. Journal of the American Society for Information Sciences, 27(3):129-146, (1976).

[4] W. Croft and D. Harper. Using probabilistic models of document retrieval with out relevance information. J. Do. 35 285-295 (1979)

[5] E. S. Robertson, J. C. Rijsbergen and M. Porter. Probabilistic models of indexing and searching. In information Retrieval Research, S. E. Robertson, C. J. van Rijsbergen, and P. Williams, Eds., Butterworths, Oxford, UK, Chapter 4, 35-36 (1981).

[6] H. Turtle and W. Croft. A comparison of text retrieval models. Comput. J. 35, 3 (June), 279-290 (1992).

[7] S. Wong and Y. Yao. On modeling information retrieval with probabilistic inference. ACM Trans. Inf. Syst. 16, 38-68 (1995).

[8] J. Ponte and B. Croft. A language modeling approach in information retrieval. In proceeding of the 21st 5 ACM SIGIR Conference on Research and Development in Information Retrieval, (Melbourne, Australia), B. Croft, A. Moffat, and C. van Rijsbergen, Eds., ACM, New York, 275-281 (1998).

[9] D. Hiemstra and A. Vries. Relating the new language models of information retrieval to the traditional retrieval models. Res Rep. TR-CTIT-00-09, Center for Telematics and Information technology (2000).

[10] A. Berger and J. Lafferty. "Information Retrieval as statistical Translation". In proceedimgs of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 222-229 (1999).

[11] D. H. Miller, T. Leek and R. Schwartz. A hidden Markov model information retrieval system. In proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval pp. 214-221 (1999).

[12] C. Shannon. A mathematical theory of communication, Bell System Technical Journal, Vol. 27, (1948).

[13] Turtle, H. and Croft, W.. Evaluation of an inference network-based retrieval model. ACM transactions on Information Systems, 9(3):187-222. July (1991).

[14] J. Broglio, J. P. Callan, W. B. Croft and D. W. Nachbar. Document Retrieval and Routing using the INQUERY system. In D. W. Harman, editor, Overview of the Third Retrieval Conference (TREC 3), pages 29-38. NIST Special Publication 500-225, (1995).

[15] J. J. Rochio. Relevance feedback in information Retrieval. InG. Salton, editor, The SMART Retrieval System - Experiments in Automatic Document Processing Prentice Hall Inc., Englewood Cliffs, NJ, (1971).

[16] M. Lesk, Automatic sense disambiguation: How to tell a pine cone from an ice cream cone, in *Proc. of the 1986 SIGDOC Conf., Pages 24-26*, New York. Association of Computing Machinery (1986).

[17] M. Sussna, Word sense disambiguation for free-test indexing using a massive semantic network, in *Proc.2nd Inter. Conf. on Information and Knowledge Management*, Arlington, Virginia, USA (1993).

[18] Y. Wilks, D. Fass, C.-M. Guo, and J. McDonald, a Tractable machine dictionary as a resource for computational semantics, in B. Boguraev and T. Briscoe

(Eds.), *Computational Lexicography for NLP, Chapter 9, pp. 193-228* London: Longman (1989).

[19] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, Introduction to WordNet: An On-line Lexical Database, Five Papers on WordNet *Princeton University* (1993).

[20] E. Agirre, G. Rigau, *Word Sense Disambiguation Using Conceptual Density*, in Proc. 16th Int. Conf. on COLING, Copenhagen (1996).

[21] P. Resnik, WordNet and distributional analysis: A class-based approach to lexical discovery *Statistically-Based Natural-Language-Processing Techniques: Papers from AAAI*(1992).

[22] E. Voorhees, Using WordNet to Disambiguate Word Senses for Text Retrieval *SIGIR* 1993.

[23] R. Mihalcea, D. Moldovan, Automatic Acquisition of Sense tagged Corpora *American Association for Art. Intel.* (1999)

[24] A. Montoyo, M. Palomar, Specification Marks for Word Sense Disambiguation: New Development, A. Gelbukh (Ed.): in *CICLing 2001, LNCS, 182-191* 2001.

[25] S. Banerjee, T. Pedersen, An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, *in Proc. Third Int. Conf. on Intelligent Text Processing and Comput. Ling. (CICLING-02)*,Mexico City, Mexico (2002).

[26] C. Leacock, M. Chodorow, Combining Local Context and WordNet5 Similarity for Word Sense Disambiguation. *Wordnet: An Electronic Lexical Database*, Christiane Fellbaum (1998).

[27] A. Budanitsky, H. Graeme, Semantic distance in WordNet: An experimental, application oriented evaluation of five measures, in *Workshop on the WordNet and Other Lexical Resources, the North American Chapter of the Ass. Comp. Ling.*, Pittsburgh, PA, June 2001.

[28] A. Agresti, *Categorical Data Analysis*, New York: John Wiley & Sons (1990).

[29] http://www.sle.sharp.co.uk/senseval2, 2002.

[30] C. Fellbaum, WordNet, An Electronic Lexical Database, *the MIT Press, Cambridge MA* (1998).

[31] W. Gale, K. W. Church, D. Yarowski, A Method for Disambiguating Word Senses in a Large Corpus, in *Computers and Humanities* 26, 1992

[32] W. T. Eadie, D. Drijard, F. E. James, M. Roos and B. Sadoulet, Statistical Methods in Experimental Physics, *North-Holland, Sec. Reprint*, (1982).

[33] B. R. D'Agostino, and M. A. Stephens, Goodness-of-fit Techniques *eds. New York: Dekker* (1986).

[34] K. Fragos, I. Maistros, C. Skourlas, Word Sense Disambiguation using WORDNET relations in*Proc. 1st Balkan Conf. in Iformatics*, Thessaloniki Greece (2003).

[35] K. Fragos, I. Maistros, C. Skourlas, Using Wordnet Lexical Database and Internet to Disambiguate Word Senses, in*Proc. 9th Panhellenic Conf. in Iformatics*, Thessaloniki Greece, 2003.

[36] D. Fernandez-Amorss, J. Gonzalo and F. Verdejo, the UNED Systems at Senseval-2. *in Proc. 2nd Int. Workshop on Evaluating WSD Systems*, Toulouse France, (2002)

[37] A. Kilgarriff, D. Tugwell, "WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography", *in Proc. workshop "COLLOCATION: Computational Extraction, Analysis and Exploitation"*, pp.32-38. 39th ACL & 10th EACL, Toulouse, July (2001).

[38] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of Royal Statistical Society B, 39:1-38, (1977)

[39] F. Jelinek, R. Mercer. Interpolated estimation of Markov source parameters from sparse data. In Pattern Recognition in Practice, E. S. Gelsema and L. N. Kanal (editors), pages 381-402. North Holland, Amsterdam (1980).

[40] J. Lafferty, C. Zhai. Document language models, query models, and risk minimization for information retrieval. In 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01) (2001).

[41] R. Richardon, A. Smeaton. Using wordnet in a knowledge-based approach to information retrieval. Technical report CA-0395, School of Comp. Science, Dublin City University, (1995).

[42] A. F. Smeaton, C. Berrut. Running TREC 4 experiments: A chronological report of query expansion experiments carried out as part of TREC 4. Technical report CA-2095 School of Comp. Science, Dublin City University, (1995).

[43] C. Stokoe, M. P. Oakes and J. Tait. Word Sense Disambiguation in Information Retrieval Reviseted. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, July (2003).

[44] E. Voorhess and D. Harman, editors. Proceeding of text retrieval Conference (TREC1-9). NIST Special Publications, http://trec.nist.gov/pubs.html (2001).

[45] E. M. Voorhess. Query expansion using lexical semantic relations. In Proceedings of the 7th ACM SIGIR Conference, pages 61-69, (1994).

[46] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In 24th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01) (2001).

[47] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval, Tenth International Conference on Information and Knowledge Management (CIKM 2001), (2001).

[48] (8) D. H. Miller, T. Leek and R. M. Schwartz. BBN at TREC-7: using hidden markov models for information retrieval. In Proceedings of the seventh Text Retrieval Conference, TREC-7, pages 133fffl142. NIST Special Publication 500−242, (1999).

[49] (15) S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In D. K. Harman, editor, the Third Text Retrieval Conference (TREC-3), (1995).

[50] Benson and Morton. The structure of the collocational dictionary. In International Journal of Lexicography 2:1-14, (1989).

[51] Caroll J., G. Minnen, D. Pearse, Y. Canning, S. Delvin and J. Tait.). Simplifying text for language-impaired readers. In Preceedings of the 9th Conference of the European Chapter of the ACL (EACL '99), Bergen, Norway, June (1999).

[52] Y. Choueka, T. Klein and E. Neuwitz. "Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus." Journal for Literary and Linguistic Computing, 4, 34-38. In Information Theory, 36(2), 372-380. Fano, R. (1961). Transmission of Information: A Statistical Theory of Information. MIT Press. Flexner, S., ed. The Random House (1887).

[53] K. Church, and P. Hanks. "Word association norms, mutual information, and lexicography." In Proceedings, 27th Meeting of the ACL, 76–83. Also in Computational Linguistics, 16(1). algorithm." IEEE Transactions on Information Theory, IT-26(1), 15-25. HaUiday, M. A. K., and Hasan, R. (1976). Cohesion in English. Longman (1989).

[54] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, Volume 19, number 1, pp61-74 (1993).

[55] R. J. Firth. A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis, pp 1-32. Oxford: Philological society. Reprinted in F. R. Palmer(ed), Selected papers of J. R. Firth 1952-1959, London: Longman, (1968).

[56] C. Gitsaki, N. Daigaku and R. Taylor. English collocations and their place in the EFL,classroom available at: http//www.hum.nagoya-cu.ac.jp/ taylor/publications/collocations.html (2000).

[57] P. Howarth, and H. Nesi. The teaching of collocations in EAP. Technical report University of Leeds, June (1996).

[58] S. Juteson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Languagr Engineering 1:9-27 (1995).

[59] D. Lin. Extracting collocations from text corpora. In First Workshop on Computational Terminology, Montreal, Canada, August (1998).

[60] C. Manning and H. Schutze.). Foundations of Statistical Natural Language Processing (Fifth Printing 2002). The MIT Press (1999).

[61] G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller. Introduction to WordNet: An On-line Lexical Database. Five Papers on WordNet Princeton University (1993).

[62] D. Pearce. Synonymy in Collocation Extraction. In WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop). pages 41-46. June. Carnegie Mellon University, Pittsburgh (2001).

[63] D. S. Richardson. Determining similarity and inferring relations in a lexical knowledge base [Diss], New York, NY: The City University of New York (1997).

[64] F. Smandja. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177, March (1993).

[65] Searching across language, time, and space: Detecting events with date and place information in unstructured text July 2002. D. A. Smith. In Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries (2002).

[66] Zhai C. Notes on the Lemur TFIDF model. In *School of Computer Science Carnegie Mellon University* 2001.

[67] Amati, G. and Van Rijsbergen. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. In *ACM Transactions on Information Systems* Vol. 20, No.4:357-389, 2002.

Κατάλογος δημοσιεύσεων του συγγραφέα

Λίστα δημοσιεύσεων του συγγραφέα

□

Δημοσιεύσεις (με κριτές)

Disambiguation using WORDNET relations K. Fragos, I. Maistros, C. Skourlas, Word Sense in 1st Balkan Conf. in Iformatics, Thessaloniki Greece 20-22 Oct. (2003).

Using Wordnet Lexical Database and Internet to Disambiguate Word Senses K. Fragos, I. Maistros, C. Skourlas, In 9th Panhellenic Conf. in Iformatics, Thessaloniki Greece, 20-22 Oct. 2003.

Using Conditional Probabilities of Weighted Terms for a Lexicon Based Sense Disambiguation System, K. Fragos, I. Maistros, C. Skourlas, In Proc. Of 3rd WSEAS Conf. in Iformatics, Malta, June 2003.

Discovering Collocations in Modern Greek Language K. Fragos, I. Maistros, C. Skourlas, in Proc. 1st Int. Conf. On Natural Language Understanding and Cognitive Science, Porto, Portugal: 13-14 April 2004.

A X2-Weighted Maximum Entropy Model for Text Classification, K. Fragos, I. Maistros, C. Skourlas, in Proc. 2nd Int. Conf. On Natural Language Understanding

and Cognitive Science, Maimi, Florida: 22-23 May 2005.

(Journals)

A Goodness of Fit Test Approach in Information Retrieval K. Fragos, I. Maistros. To be appeared in the coming issue of the journal of "Information Retrieval".

Distributional Analysis of Related Synsets in Wordnet for a Word Sense Disambiguation Task. K. Fragos, I. Maistro. To be appeared in the coming December issue of the journal of "International Journal of Artificial Intelligence Tools".

Βιογραφικό Σημείωμα

Βιογραφικό σημείωμα του συγγραφέα

□

Κωνσταντινος Φραγγος

Διευθυνση

Διεύθυνση: Δημοκρίτου 7

Πόλη: 34100, Χαλκίδα

Τηλέφωνο: 22210-28559

Email: kfragos@ece.ntua.gr

Προσωπικες πληροφοριες

Γέννος: Άρρεν

Ημερομηνία Γέννησης: 8/5/1962

Τόπος Γέννησης Σκεπαστή Ευβοίας

Σπουδες

9/1981–7/1985: Πτυχίο Μαθηματικών Πανεπιστήμιο Αθηνών

9/1990–9/1993: Πτυχίο Πληροφορικής Πανεπιστήμιο Αθηνών

2/1995–2/1997: Μεταπτυχιακό Ηλεκτρονικού Αυτοματισμού Πανεπιστήμιο Αθηνών

Παρούσα κατάσταση: Υποψήφιος Διδάκτορας στο ΕΜΠ

Ξενες Γλωσσες

Αγγλικά, Γαλλικά

Ερευνητικα Ενδιαφεροντα

Information Retrieval, Word sense disambiguation,
Statistical NLP, Neural Networks, Signal Processing