ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αποσαφήνιση Λέξεων με Βάση τα Google 5-grams

Πολυξένη Π. Κατσιούλη

Επιβλέπων: Θεόδωρος Καλαμπούκης, Καθηγητής ΟΠΑ

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2008

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αποσαφήνιση Λέξεων με Βάση τα Google 5-grams

Πολυξένη Π. Κατσιούλη

Α.Μ.: ΕΥ0609

ΕΠΙΒΛΕΠΩΝ:

Θεόδωρος Καλαμπούκης, Καθηγητής ΟΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Θεόδωρος Καλαμπούκης, Καθηγητής ΟΠΑ

Ίων Ανδρουτσόπουλος, Επίκουρος Καθηγητής ΟΠΑ

ΠΕΡΙΛΗΨΗ

Ο στόχος της ανάκτησης πληροφοριών (Information Retrieval, IR) είναι να βρεθούν και να ανακτηθούν έγγραφα σχετικά με ένα ερώτημα, όπου τα έγγραφα και το ερώτημα είναι στην ίδια γλώσσα. Με τις προόδους της έρευνας και της τεχνολογίας ο στόχος επεκτάθηκε πέρα από τα γλωσσικά όρια για να περιλάβει την ανάκτηση κειμένων σε διαφορετικές από το ερώτημα γλώσσες, μια διαδικασία η οποία είναι γνωστή ως δια-γλωσσική ανάκτηση πληροφοριών (Cross Language Information Retrieval, CLIR).

Η δια-γλωσική ανάκτηση πληροφοριών όπου τα ερωτήματα των χρηστών και τα έγγραφα είναι γραμμένα σε διαφορετικές γλώσσες αποτελεί ένα από τα σημαντικότερα ερευνητικά πεδία της ανάκτησης πληροφοριών. Για να είναι αποδοτικό ένα CLIR σύστημα είναι απαραίτητο να γίνει σωστή μετάφραση των ερωτημάτων του χρήστη.

Στην παρούσα εργασία περιγράφεται και υλοποιείται μια μέθοδος με την οποία γίνεται αποσαφήνιση της έννοιας των λέξεων ώστε να επιλεγεί η πιο κατάλληλη μετάφραση. Τα ερωτήματα υποβάλλονται στην ελληνική γλώσσα ενώ χρησιμοποιείται μια λίστα με τις αγγλικές μεταφράσεις λημμάτων. Η μέθοδος βασίζεται σε ένα σύνολο δεδομένων της Google -τα λεγόμενα Google 5-grams- το οποίο περιέχει αγγλικά πεντάγραμμα (ακολουθίες πέντε λέξεων) καθώς και τις συχνότητες εμφάνισής τους. Η αποσαφήνιση της έννοιας των λέξεων και η επιλογή της κατάλληλης μετάφρασης γίνεται με τη βοήθεια πιθανοκρατικών αλγορίθμών όπως ο αλγόριθμος Naïve Bayes και τα Γλωσσολογικά Μοντέλα. Η μέθοδος αυτή δοκιμάστηκε και αξιολογήθηκε με ένα σύνολο από ιατρικά ελληνικά ερωτήματα.

Θεματική Περιοχή: δια-γλωσσική ανάκτηση πληροφοριών, αυτόματη μετάφραση

Λέξεις Κλειδιά: επεξεργασία φυσικής γλώσσας, ανάκτηση πληροφοριών, αποσαφήνιση της έννοιας των λέξεων

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ **3**
ΠΕΡΙΕΧΟΜΕΝΑ **4**
ΛΙΣΤΑ ΕΙΚΟΝΩΝ **6**
ΛΙΣΤΑ ΠΙΝΑΚΩΝ **6**
ΠΡΟΛΟΓΟΣ **7**
ΚΕΦΑΛΑΙΟ 1 ..... **9**
ΕΙΣΑΓΩΓΗ **9**
1.4 Αυτόματη Μετάφραση 11
1.3 Περιγραφή Εργασίας 11
ΚΕΦΑΛΑΙΟ 2 ..... **13**
ΕΠΙΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ **13**
2.1 Διγλωσσική Ανάκτηση Πληροφοριών με χρήση Διαδικτυακών Καταλόγων 13
2.1.1 Εξαγωγή χαρακτηριστικών γνωρισμάτων για κάθε κατηγορία 13
2.1.2 Διαδικασία ανάκτησης 15
2.2 Αποσαφήνιση της έννοιας μιας λέξης με γλωσσολογική ανάλυση 16
2.2.1 Γλωσσολογικό Μοντέλο 17
2.2.2 Στατιστικό Μοντέλο 18
2.3 Μέθοδοι αποσαφήνισης μεταφράσεων 20
2.3.1 Μέθοδος αποσαφήνισης μεταφράσεων βασισμένη στο Web 21
2.3.2 Μέθοδος αποσαφήνισης μεταφράσεων βασισμένη σε ένα σώμα κειμένου 22
2.4 Δια-γλωσσική Ανάκτηση Πληροφοριών με χρήση Οντολογιών 23
ΚΕΦΑΛΑΙΟ 3 ..... **27**
ΠΙΘΑΝΟΚΡΑΤΙΚΕΣ ΜΕΘΟΔΟΙ ΑΡΣΗΣ ΑΜΦΙΣΗΜΙΑΣ
ΤΗΣ ΕΝΝΟΙΑΣ ΤΩΝ ΛΕΞΕΩΝ **27**
3.1 Ο Αλγόριθμος Naïve Bayes 28
4.2 Γλωσσολογικά Μοντέλα (Language Models) 30
4.2.1 N-gram Μοντέλα 30
ΚΕΦΑΛΑΙΟ 4 ..... **33**
ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΔΟΛΟΓΙΑΣ **33**
4.1 Google n-grams 33
4.2 Μηχανή Αναζήτησης Lucene 34

4.2.1 Δημιουργία ευρετηρίου με τη χρήση του Lucene 35
4.2.1.1 Κλάση IndexWriter 36
4.2.1.2 Κλάση Directory 36
4.2.1.3 Κλάση Analyzer 36
4.2.1.4 κλάση Document 37
4.2.1.5 Κλάση Field 37
4.2.1.6 Ερευτηριοποίηση των Google 5-grams 38
4.2.2 Διαδικασία Αναζήτησης με τη Χρήση του Lucene 39
4.2.2.1 Κλάση IndexSearcher 39
4.2.2.2 Κλάση Query 39
4.2.2.3 Κλάση QueryParser 39
4.2.2.4 Κλάση Hits 40
4.2.2.5 Αναζήτηση Δεδομένων από τα Google 5-grams 40
4.3 Επισκόπηση Μεθοδολογίας 40
ΚΕΦΑΛΑΙΟ 5 ..... **43**
ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ **43**
5.1 MEDLINE Ερωτήματα 43
5.2 Λεξικό 43
5.4 Έμμεση Αξιολόγηση 45
5.4.1 Μέση Ακρίβεια (Average Precision) 46
5.4.2 Βάση Δεδομένων OSHUMED 47
5.4.3 Αποτελέσματα Έμμεσης Αξιολόγησης 47
ΚΕΦΑΛΑΙΟ 6 ..... **49**
ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ **49**
ΠΑΡΑΡΤΗΜΑ Α **51**
ΠΑΡΑΡΤΗΜΑ Β **55**
ΑΚΡΩΝΥΜΙΑ **56**
ΑΝΑΦΟΡΕΣ **57**

ΛΙΣΤΑ ΕΙΚΟΝΩΝ

Εικόνα 1. Στόχος της εργασίας: Παράδειγμα 10
Εικόνα 2. Σύνοψη CLIR μεθόδου με Web directories 14
Εικόνα 3. Διαδικασία ανάκτησης κειμένων 15
Εικόνα 4. Μετάφραση του ερωτήματος 16
Εικόνα 5. Concept 'SHIP' 25
Εικόνα 6. Αλγόριθμος Naïve Bayes 29
Εικόνα 7. Μέγεθος των 5-grams 33
Εικόνα 8. Δείγμα από τα 4-grams 34
Εικόνα 9. Ενσωμάτωση Lucene σε εφαρμογές 35
Εικόνα 10. Δημιουργία ευρετηρίου και προσθήκη των 5-grams σε αυτό 38
Εικόνα 11. Αναζήτηση δεδομένων από το ευρετήριο 40
Εικόνα 12. Επισκόπηση της μεθοδολογίας 41
Εικόνα 13. Διαδικασία μετάφραης ενός ερωτήματος 42
Εικόνα 14. Μέση ακρίβεια ανακτηθένων κειμένων (term frequency) 47
Εικόνα 15. Μέση ακρίβεια ανακτηθέντων κειμένων (document frequency) 48

ΛΙΣΤΑ ΠΙΝΑΚΩΝ

Πίνακας 1. Σύνοψη των τύπων πεδίων και των χαρακτηριστικών τους 38
Πίνακας 2. Στατιστικά στοιχεία ερωτημάτων 44
Πίνακας 3. Αποτελέσματα πειραμάτων 45

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών Επιστήμη των Υπολογιστών του τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών. Το αντικείμενο μελέτης της εργασίας αυτής είναι η υλοποίηση και αξιολόγηση μιας μεθόδου αυτόματης μετάφρασης ερωτημάτων κατά τη δια-γλωσσική ανάκτηση πληροφοριών. Η μέθοδος αυτή βασίζεται σε ένα σύνολο δεδομένων που δημοσίευσε η Google το 2006 για να επιλέξει την κατάλληλη μετάφραση των λέξεων του ερωτήματος. Η επιλογή της κατάλληλης μετάφρασης των λέξεων γίνεται με τη βοήθεια πιθανοκρατικών αλγορίθμων οι οποίοι βασίζονται στο πλαίσιο στο οποίο ανήκουν οι αμφίσημες λέξεις.

Στο σημείο αυτό, θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στον καθηγητή του Οικονομικού Πανεπιστημίου Αθηνών κ. Θεόδωρο Καλαμπούκη για την πολύ σημαντική καθοδήγησή του καθ' όλη τη διάρκεια εκπόνησης της εργασίας. Θα ήθελα επίσης να ευχαριστήσω τον επίκουρο καθηγητή του Οικονομικού Πανεπιστημίου Αθηνών κ. Ίων Ανδρουτσόπουλο για τις πολύτιμες επισημάνσεις και παρατηρήσεις του κατά τη διάρκεια της παρουσίασης της εργασίας. Τέλος, ευχαριστώ την οικογένειά μου και τους καλούς μου φίλους για τη συμπαράσταση και τη στήριξη που μου προσέφεραν καθ' όλη τη διάρκεια της φοίτησής μου στο συγκεκριμένο Πρόγραμμα Μεταπτυχιακών Σπουδών.

Αθήνα, Ιούλιος 2008

Πολυξένη Π. Κατσιούλη

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

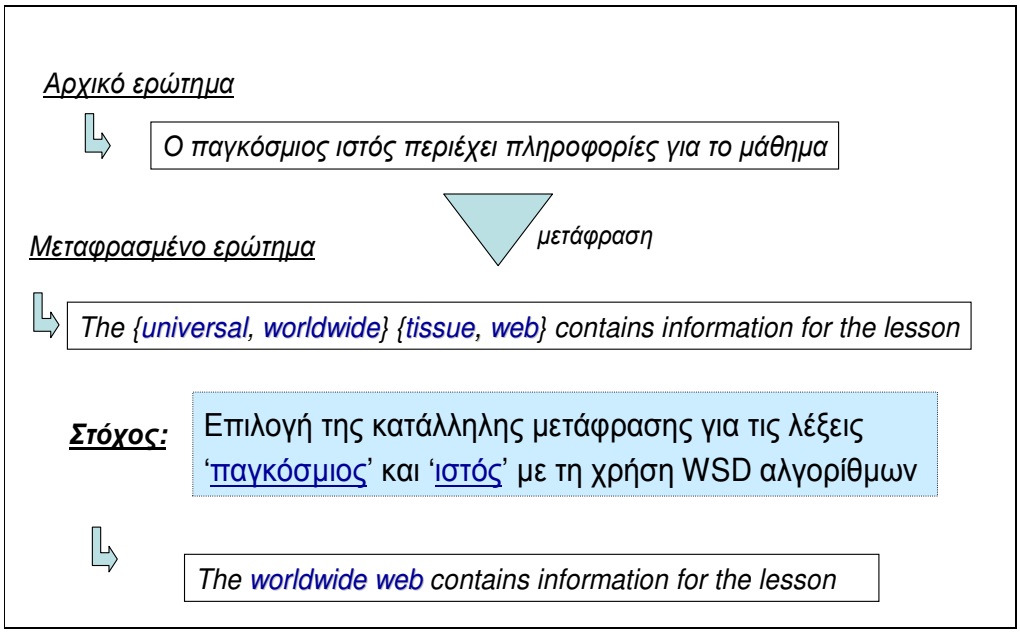
Με την ανάπτυξη του Παγκοσμίου Ιστού (World Wide Web, WWW), όλο και περισσότερες γλώσσες χρησιμοποιούνται στα έγγραφα του Ιστού, και για αυτό είναι τώρα πολύ ευκολότερη η πρόσβαση στα έγγραφα που γράφονται σε διαφορετικές γλώσσες από τη μητρική γλώσσα του κάθε χρήστη. Εντούτοις, οι υπάρχουσες μηχανές αναζήτησης υποστηρίζουν μόνο την ανάκτηση εγγράφων που είναι γραμμένα στην ίδια γλώσσα με το ερώτημα, και έτσι δεν υπάρχει κάποιος αποδοτικός τρόπος για τους χρήστες να ανακτήσουν έγγραφα γραμμένα σε διαφορετικές από το ερώτημα γλώσσες.

Επίσης, υπάρχουν περιπτώσεις, ανάλογα με τις ανάγκες του χρήστη, όπου οι πληροφορίες γράφονται σε μια γλώσσα διαφορετική από τη μητρική γλώσσα του χρήστη. Για να ικανοποιηθούν οι ανάγκες των χρηστών σε ένα χαρακτηριστικό μονόγλωσσο σύστημα ανάκτησης πληροφοριών, οι χρήστες πρέπει να μεταφράσουν μόνοι τους τα ερωτήματα με τη βοήθεια ενός λεξικού. Αυτή η μέθοδος είναι όχι μόνο δύσκολη για το χρήστη, αλλά τα αποτέλεσμα της μετάφρασης είναι μερικές φορές λανθασμένα, ειδικά όταν ο χρήστης δεν είναι εξοικειωμένος με τη γλώσσα.

Στην παρούσα εργασία περιγράφεται και αξιολογείται μια μέθοδος η οποία στοχεύει στην εύρεση των σωστών μεταφράσεων των ερωτημάτων των χρηστών. Η προτεινόμενη μέθοδος εφαρμόζει μια διαδικασία αποσαφήνισης της έννοιας των λέξεων προκειμένου να επιλεγούν οι κατάλληλες μεταφράσεις αυτών. Η μέθοδος χρησιμοποιεί ένα σύνολο δεδομένων το οποίο δεν αναφέρεται σε κάποιο συγκεκριμένο πεδίο, δεν είναι δηλαδή domain specific. Αυτό σημαίνει ότι η παρούσα μέθοδος μπορεί να εφαρμοστεί για τη μετάφραση ερωτημάτων από κάθε πεδίο. Η αποσαφήνιση της έννοιας των λέξεων βασίζεται σε πιθανοκρατικές μεθόδους, ενώ η αξιολόγηση της μεθόδου έδειξε ότι τα αποτελέσματα είναι αρκετά ενθαρρυντικά.

Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να ανακτήσουμε έγγραφα γραμμένα στην αγγλική γλώσσα για το ερώτημα «Ο παγκόσμιος ιστός περιέχει πληροφορίες για το μάθημα». Με τη βοήθεια ενός ελληνικο-αγγλικού λεξικού μεταφράζουμε το ερώτημα

στην αγγλική γλώσσα. Όπως φαίνεται και στην Εικόνα 1, το λεξικό έδωσε δύο πιθανές μεταφράσεις για τις λέξεις 'παγκόσμιος' και 'ιστός'. Στόχος, της παρούσας εργασίας είναι να επιλέξει την κατάλληλη μετάφραση των αμφίσημων λέξεων με τη χρήση αλγορίθμων αποσαφήνισης της έννοιας μιας λέξης και των Google 5-grams τα οποία περιγράφονται αναλυτικά στα Κεφάλαια 3 και 4 αντίστοιχα.



Εικόνα 1. Στόχος της εργασίας: Παράδειγμα

1.1 Δια-γλωσσική Ανάκτηση Πληροφοριών

Η διαγλωσσική ανάκτηση πληροφοριών (CLIR) είναι μια εφαρμογή που αναπτύσσεται με γρήγορο ρυθμό, χάρι στην όλο και περισσότερο ανάπτυξη του Παγκόσμιου Ιστού και στη δημιουργία δικτυατών τόπων σε διάφορες γλώσσες. Ο στόχος της δια-γλωσσικής ανάκτησης γλωσσικών πληροφοριών είναι η εύρεση των πληροφοριών που χρειάζεται ένας χρήστης ακόμα κι αν αυτές είναι γραμμένες σε διαφορετική από τη μητρική γλώσσα του χρήστη. Αυτό επιτυγχάνεται με το σχεδιασμό ενός συστήματος όπου η σχετικότητα ενός ερωτήματος που υποβάλεται σε μια μηχανή αναζήτησης σε μια γλώσσα μπορεί να συγκριθεί με έγγραφα γραμμένα σε μια άλλη γλώσσα.

Πολύ σημαντικό ρόλο σε ένα CLIR σύστημα παίζει η αποσαφήνιση της έννοιας με την οποία χρησιμοποιείται η κάθε λέξη, μια διαδικασία γνωστή ως Word Sense Disambiguation (WSD). Στην πράξη σε ένα CLIR σύστημα οι πιθανές έννοιες μιας λέξης αντιπροσωπεύονται από τις πιθανές μεταφράσεις αυτής, οπότε η διαδικασία του WSD

στοχεύει στην επιλογή της πιο κατάλληλης μετάφρασης μιας λέξης που έχει πολλές δυνατές μεταφράσεις.

1.4 Αυτόματη Μετάφραση

Η αυτόματη μετάφραση (Machine Translation, MT) [1] είναι ένα πεδίο της υπολογιστικής γλωσσολογίας (computational linguistics) που ερευνά τη χρήση λογισμικού υπολογιστών για να μεταφράσει ένα κείμενο από μια φυσική γλώσσα σε μια άλλη. Στο βασικό της επίπεδο, η αυτόματη μετάφραση αντικαθιστά τις λέξεις της μιας γλώσσας με τις αντίστοιχες μεταφράσεις στην γλώσσα προορισμού. Η χρήση τεχνικών βασιμένων σε μεγάλα σώματα κειμένου (corpus) επιτρέπει τον καλύτερο χειρισμό κάποιων ιδιαιτεροτήτων της εκάστοτε γλώσσας όπως τη μετάφραση ιδιωματισμών και φράσεων.

Η αυτοματοποίηση της μετάφρασης με τη χρήση υπολογιστών απασχολεί τους ερευνητές για πολλά χρόνια. Τις τελευταίες δεκαετίες έχουν αναπτυχθεί προγράμματα υπολογιστών που στοχεύουν στη μετάφραση κειμένων από μια γλώσσα σε μια άλλη. Η απόδοση όμως των προγραμμάτων αυτών δεν είναι πολύ καλή. Δεν υπάρχει δηλαδή, καμιά αυτόματη μηχανή που να δέχεται ως είσοδο ένα κείμενο σε οποιαδήποτε γλώσσα και να παράγει μια τέλεια μετάφραση του κειμένου αυτού σε κάποια άλλη γλώσσα χωρίς την ανθρώπινη παρέμβαση ή βοήθεια.

Στην πράξη, έχει παρατηρηθεί ότι ο στόχος της αυτόματης μετάφρασης δεν είναι η μετάφραση κειμένων με υψηλό λογοτεχνικό και πολιτιστικό περιεχόμενο. Η μεγάλη πλειοψηφία των επαγγελματιών μεταφραστών απασχολείται για να ικανοποιήσει την τεράστια και αυξανόμενη ζήτηση για τις μεταφράσεις επιστημονικών και τεχνικών εγγράφων, τις εμπορικές και επιχειρησιακές συναλλαγές, τα διοικητικά υπομνήματα, τη νομική τεκμηρίωση, τα εγχειρίδια, τα γεωργικά και ιατρικά βιβλία, τα βιομηχανικά διπλώματα ευρεσιτεχνίας, τα φυλλάδια δημοσιότητας, τις εκθέσεις εφημερίδων, κ.λπ.

1.3 Περιγραφή Εργασίας

Η παρούσα εργασία εστιάζεται στην περιγραφή της μεθοδολογίας με την οποία επιτυγχάνεται η επιλογή της πιο κατάλληλης μετάφρασης ενός ερωτήματος από μια φυσική γλώσσα σε μια άλλη με απώτερο στόχο την ανάκτηση δεδομένων από ένα CLIR σύστημα. Η οργάνωση της εργασίας έχει ως εξής:

Στο Κεφάλαιο 2 περιγράφονται μερικές από τις προσεγγίσεις που έχουν προταθεί στη βιβλιογραφία για την επίλυση του προβλήματος της αυτόματης μετάφρασης.

Στο Κεφάλαιο 3 παρουσιάζονται διεξοδικά οι μέθοδοι που χρησιμοποιήθηκαν στην παρούσα εργασία για την αποσαφήνιση της έννοιας μιας λέξης με σκοπό να επιλεγεί η πιο κατάλληλη μετάφραση των αμφίσημων λέξεων. Συγκεκριμένα, περιγράφεται αναλυτικά ο αλγόριθμος Naïve Bayes καθώς και τα Γλωσσολογικά Μοντέλα (Language Models).

Στο κεφάλαιο 4, περιγράφονται αναλυτικά τα 5-grams του Google πάνω στα οποία βασίστηκε η συγκεκριμένη εργασία καθώς και ο τρόπος με τον οποίο έγινε δυνατή η εξαγωγή πληροφοριών από αυτά. Στο κεφάλαιο αυτό, περιγράφονται επίσης, όλες οι φάσεις τις προτεινόμενης μεθόδου για την αυτόματη μετάφραση των ερωτημάτων.

Το Κεφάλαιο 5 περιλαμβάνει την αξιολόγηση της μεθόδου. Συγκεκριμένα, περιγράφονται τα δεδομένα που χρησιμοποιήθηκαν στην αξιολόγηση καθώς και τα αποτελέσματα που προέκυψαν από τα διάφορα πειράματα.

Η παρούσα εργασία ολοκληρώνεται με το Κεφάλαιο 6 στο οποίο δίνονται τα συμπεράσματα της όλης μελέτης καθώς και κάποια «ανοιχτά» για μελέτη θέματα που αφορούν στο συγκεκριμένο εξεταζόμενο πεδίο έρευνας.

ΚΕΦΑΛΑΙΟ 2

ΕΠΙΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η αυτόματη μετάφραση ερωτημάτων είναι ένα πεδίο έρευνας της ανάκτησης πληροφοριών το οποίο εξελίσσεται διαρκώς τα τελευταία χρόνια. Στο κεφάλαιο αυτό γίνεται μια επισκόπηση της βιβλιογραφίας και περιγράφονται μέθοδοι που αποσαφηνίζουν τις έννοιες των λέξεων κατά τη μετάφραση ερωτημάτων με σκοπό την ανάκτηση κειμένων από ένα CLIR σύστημα. Οι μέθοδοι που περιγράφονται στις επόμενες ενότητες του κεφαλαίου χρησιμοποιούν διαφορετικούς τρόπους για να αντιμετωπίσουν το πρόβλημα της αυτόματης μετάφρασης. Συγκριμένα, περιγράφονται μέθοδοι βασισμένες σε Web καταλόγους αλλά και γενικότερα στο Web (Web-based methods), σε σύνολα κειμένων (corpus-based methods), σε γλωσσολογική και συντακτική ανάλυση καθώς και σε οντολογίες.

2.1 Διγλωσσική Ανάκτηση Πληροφοριών με χρήση Διαδικτυακών Καταλόγων

Στο [2] προτείνεται μια CLIR μέθοδο που χρησιμοποιεί Διαδικτυακούς καταλόγους (Web directories) οι οποίοι είναι διαθέσιμοι σε πολλές γλώσσες (π.χ., το Yahoo [3]). Η μέθοδος αυτή χρησιμοποιεί μία έκδοση του Web καταλόγου στην ίδια γλώσσα με τα ερωτήματα (source language) και μία έκδοση του Web καταλόγου στη γλώσσα των ανακτηθέντων κειμένων (target language).

Στις ενότητες που ακολουθούν περιγράφονται οι φάσεις της προτεινόμενης μεθόδου, σύνοψη της οποίας παρουσιάζεται στην Εικόνα 2.

2.1.1 Εξαγωγή χαρακτηριστικών γνωρισμάτων για κάθε κατηγορία

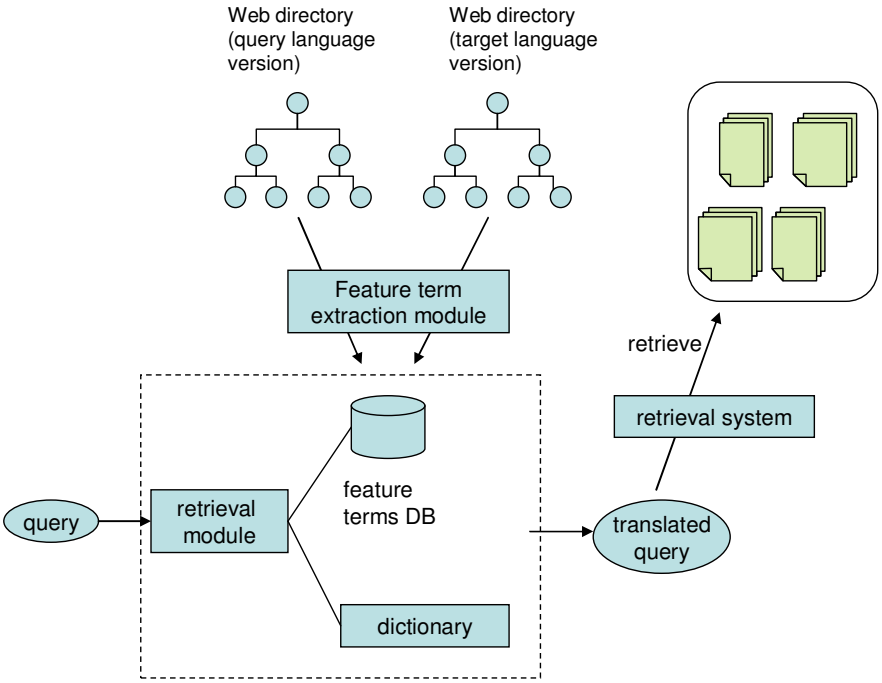
Κατά τη φάση αυτή γίνεται εξαγωγή χαρακτηριστικών (feature terms) γνωρισμάτων για κάθε κατηγορία από του Web καταλόγους. Η διαδικασία αυτή γίνεται ως εξής:

1. Το σύστημα εξάγει όρους από όλες τις εκδόσεις των Web καταλόγων για κάθε κατηγορία.
2. Το σύστημα υπολογίζει τα βάρη των εξαγόμενων όρων.
3. Οι n όροι με το μεγαλύτερο βάρος θεωρούνται ως χαρακτηριστικά γνωρίσματα για κάθε κατηγορία.

Τα βάρη των feature terms υπολογίζονται με βάση τη στατιστική μετρική *tf-icf* (*term frequency – inverse category frequency*) η οποία αποτελεί μια παραλλαγή του *tf-idf* (*term frequency – inverse document frequency*). Το *tf-icf* υπολογίζεται με βάση τον ακόλουθο τύπο:

$$tf \cdot icf(t_i, c) = \frac{f(t_i)}{N_c} + \log\frac{N}{n_i} + 1$$

όπου, t_i είναι ο όρος που εμφανίζεται στην κατηγορία c, $f(t_i)$ είναι η συχνότητα εμφάνισης του όρου t_i, N_c ο συνολικός αριθμός των όρων στην κατηγορία c, n_i είναι ο αριθμός των κατηγοριών που περιέχουν τον όρο t_i, και N είναι ο συνολικός αριθμός των κατηγοριών στον Web κατάλογο.

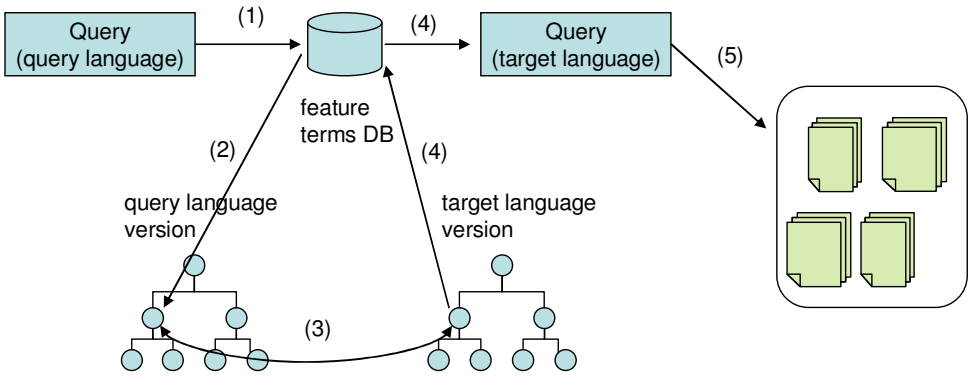


Εικόνα 2. Σύνοψη CLIR μεθόδου με Web directories

2.1.2 Διαδικασία ανάκτησης

Στη φάση αυτή πραγματοποιείται η διαδικασία της ανάκτησης κειμένων σχετικών με το ερώτημα του χρήστη. Η φάση αυτή αποτελείται από τα ακόλουθα βήματα (Εικόνα 3):

1. Για κάθε κατηγορία του Web καταλόγου στην ίδια γλώσσα με τα ερωτήματα υπολογίζεται η σχέση (relevance) του ερωτήματος και των χαρακτηριστικών γνωρισμάτων της εκάστοτε κατηγορίας.
2. Επιλέγεται η πιο σχετική με το ερώτημα κατηγορία.
3. Επιλέγεται η κατηγορία από τον Web κατάλογο στη γλώσσα προορισμού (target language) που είναι πιο σχετική με την κατηγορία που επιλέχτηκε στο προηγούμενο βήμα.
4. Μετάφραση του ερωτήματος με τη βοήθεια των χαρακτηριστικών γνωρισμάτων της αντίστοιχης κατηγορίας.
5. Ανάκτηση κειμένων με βάση το μεταφρασμένο ερώτημα.



Εικόνα 3. Διαδικασία ανάκτησης κειμένων

Στο σύστημα που προτείνεται στο [2] τα ερωτήματα αποτελούνται από λέξεις κλειδιά και όχι από προτάσεις. Ένα ερώτημα αναπαρίσταται από ένα διάνυσμα ως εξής:

$$\vec{q} = (q_1, q_2, \ldots, q_n)$$

όπου q_k είναι το βάρος της k-οστής λέξης-κλειδί του ερωτήματος. Όλα τα βάρη q_k του ερωτήματος έχουν τιμή 1.

Η σχέση ανάμεσα στο ερώτημα και σε κάθε κατηγορία υπολογίζεται πολλαπλασιάζοντας το εσωτερικό γινόμενο του ερωτήματος και του συνόλου των feature terms με τη γωνία που σχηματίζουν τα δύο διανύσματα:

$$rel(q,c) = \vec{q} \cdot \vec{c}\, \frac{\vec{q} \cdot \vec{c}}{|\vec{q}| \cdot |\vec{c}|}$$

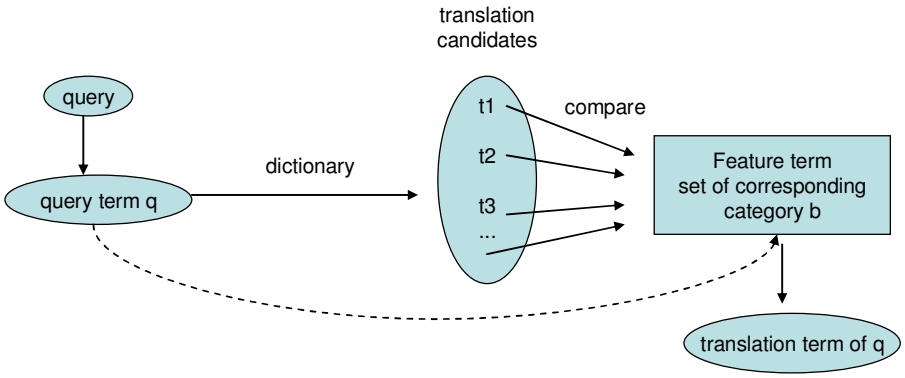
όπου το διάνυσμα c αντιπροσωπεύει κάθε κατηγορία και ορίζεται ως εξής:

$$\vec{c} = (w_1, w_2, \ldots, w_n)$$

όπου w_k είναι το βάρος της k-οστής λέξης-κλειδί στο σύνολο των feature terms της κατηγορίας c.

Ως σχετική με το ερώτημα επιλέγεται εκείνη η κατηγορία της οποίας η τιμή του rel ξεπερνά ένα προκαθορισμένο όριο (threshold).

Στην Εικόνα 4 παρουσιάζεται ο τρόπος με τον οποίο γίνεται η μετάφραση ενός ερωτήματος. Αρχικά, εξάγονται από ένα λεξικό οι πιθανές μεταφράσεις κάθε όρου του ερωτήματος. Στη συνέχεια, το σύστημα εξετάζει αν κάποια από τις μεταφράσεις περιέχεται στο σύνολο των feature terms της αντίστοιχης κατηγορίας. Αν αυτό συμβαίνει, το σύστημα ελέγχει το βάρος της υποψήφιας μετάφρασης στο σύνολο των feature terms. Τέλος, η μετάφραση με το μεγαλύτερο βάρος στο σύνολο των feature terms επιλέγεται ως η πιο κατάλληλη για τον αντίστοιχο όρο.



Εικόνα 4. Μετάφραση του ερωτήματος

2.2 Αποσαφήνιση της έννοιας μιας λέξης με γλωσσολογική ανάλυση

Στο άρθρο [4] περιγράφεται μια μέθοδος αποσαφήνισης της έννοιας μια λέξης σε μια γλώσσα με τη χρήση στατιστικών δεδομένων από ένα σώμα κειμένου (corpus)

γραμμένο σε διαφορετική γλώσσα. Ο αλγόριθμος αυτός προσδιορίζει συντακτικές σχέσεις ανάμεσα στις λέξεις της αρχικής γλώσσας και αντιστοιχεί τις εναλλακτικές μεταφράσεις των σχέσεων αυτών στην γλώσσα προορισμού με τη χρήση ενός δίγλωσσου λεξικού. Στη συνέχεια, η επιλογή των κατάλληλων μεταφράσεων γίνεται με τη βοήθεια στατιστικών μετρικών.

Η μέθοδος που προτείνεται στο [4] κατασκευάζει ένα γλωσσολογικό (linguistic) και ένα στατιστικό μοντέλο προκειμένου να επιλέξει τη σωστή μετάφραση ενός ερωτήματος. Στις επόμενες παραγράφους περιγράφεται ο τρόπος με τον οποίο δημιουργούνται τα δύο αυτά μοντέλα.

2.2.1 Γλωσσολογικό Μοντέλο

Η μέθοδος χρησιμοποιεί ένα λεξικό με το οποίο όλες οι λέξεις του ερωτήματος μεταφράζονται στη γλώσσα προορισμού. Η βασική έννοια της μεθόδου είναι η *συντακτική πλειάδα* (*syntactic tuple*) η οποία ορίζει μια συντακτική σχέση ανάμεσα σε δύο ή περισσότερες λέξεις. Αποτελείται από το όνομα της συντακτικής σχέσης ακολουθούμενο από μια σειρά από λέξεις που ικανοποιούν τη σχέση αυτή. Για παράδειγμα, (subj-verb: man walk) είναι μια συντακτική πλειάδα η οποία εμφανίζεται στην πρόταση “The man walked home”. Η εξαγωγή των συντακτικών σχέσεων γίνεται με τη βοήθεια ενός parser συντακτικής ανάλυσης.

Οι σχέσεις οι οποίες είναι χρήσιμες για την αποσαφήνιση της έννοιας των λέξεων είναι οι εξείς:

- Σχέσεις ανάμεσα σε ένα ρήμα και στο υποκείμενό του, στους προσδιορισμούς του, στα άμεσα και έμμεσα αντικείμενά του καθώς και στους εμπρόθετους προσδιορισμούς του.
- Σχέσεις ανάμεσα σε ένα ουσιαστικό και στα συμπληρώματά του (complements), συμπεριλαμβανομένων των επιθέτων και των τροποιποιημένων ουσιαστικών (modifying nouns) στα σύνθετα ουσιαστικά.
- Σχέσεις ανάμεσα στα επίθετα ή τα επιρρήματα και τους τροποποιητές τους (modifiers).

Το πρώτο βήμα της μεθόδου είναι η εύρεση όλων των συντακτικών πλειάδων που περιέχουν αμφίσημες λέξεις. Για παράδειγμα, ας υποθέσουμε και πάλι ότι θέλουμε να μεταφράσουμε το ερώτημα “Ο παγκόσμιος ιστός περιέχει πληροφορίες για το μάθημα”.

Η μετάφραση του ερωτήματος αυτού είναι: “The universal {tissue, web, mast} contains information for the lesson”. Στο ερώτημα αυτό υπάρχει μία αμφίσημη λέξη, η λέξη ‘ιστός’ με τρεις πιθανές έννοιες-μεταφράσεις, ‘tissue’, ‘web’ και ‘mast’. Μερικές από τις συντακτικές πλειάδες του ερωτήματος αυτού είναι οι ακόλουθες:

1. (adj-noun: παγκόσμιος ιστός)
2. (subj-verb: ιστός περιέχει)
3. (verb-obj: περιέχει πληροφορίες)

Από τις πλειάδες αυτές μόνο οι 1 και 2 περιέχουν την αμφίσημη λέξη ‘ιστός’.

Στη συνέχεια, οι πλειάδες αυτές αντιστοιχίζονται σε πλειάδες στη γλώσσα προορισμού. Οι πλειάδες 1 και 2 έχουν τρεις πιθανές αντιστοιχίσεις επειδή περιέχουν την αμφίσημη λέξη, ενώ η πλειάδα 3 έχει μόνο μια. Οι αντιστοιχίσεις των παραπάνω πλειάδων στη γλώσσα προορισμού είναι οι εξείς:

1. (adj-noun: universal tissue)
   (adj-noun: universal web)
   (adj-noun: universal mast)
2. (sub-verb: tissue contain)
   (sub-verb: web contain)
   (sub-verb: mast contain)
3. (verb-obj: contain information)

Η δημιουργία του γλωσσολογικού μοντέλου ολοκληρώνεται με τον υπολογισμό της συχνότητας εμφάνισης κάθε πλειάδας στο corpus της γλώσσας προορισμού.

2.2.2 Στατιστικό Μοντέλο

Το δεύτερο μέρος της μεθόδου συνίσταται στη δημιουργία ενός στατιστικού μοντέλου για την επιλογή της κατάλληλης μετάφρασης για κάθε αμφίσημη λέξη. Η επιλογή αυτή βασίζεται στις συχνότητες εμφάνισης των συντακτικών πλειάδων στο corpus της γλώσσας προορισμού όπως αυτές υπολογίστηκαν κατά την ολοκλήρωση της δημιουργίας του γλωσσολογικού μοντέλου.

Αρχικά θεωρείται η περίπτωση όπου μία μόνο συντακτική πλειάδα περιέχει μια αμφίσημη λέξη. Έστω ότι με T συμβολίζεται η αρχική πλειάδα και με T_1, T_2,...,T_k οι k αντιστοιχίσεις της πλειάδας αυτής στη γλώσσα προορισμού. Έστω επίσης ότι οι

συχνότητες εμφάνισης των πλειάδων αυτών στο corpus της γλώσσας προορισμού είναι $n_1, n_2,\ldots, n_k$, όπου $n_1 \geq n_2 \geq \ldots \geq n_k$.

Αφού ο στόχος είναι η επιλογή για την πλειάδα T μια από τις πλειάδες T_i, μπορεί να θεωρηθεί η T ως μια τυχαία διακριτή μεταβλητή με πολυωνυμική κατανομή, με πιθανές τιμές τις $T_1, T_2,...,T_k$. Έστω p_i η πιθανότητα η πλειάδα T_i να είναι η σωστή μετάφραση της T. Η τιμή της p_i υπολογίζεται με βάση τον ακόλουθο τύπο:

$$p_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$$

Πρέπει τώρα να προσδιοριστεί το κριτήριο για την επιλογή της πιο κατάλληλης συντακτικής πλειάδας T στη γλώσσα προορισμού. Η πιο λογική υπόθεση είναι να επιλεγεί η πλειάδα με τη μεγαλύτερη πιθανότητα, δηλαδή η πλειάδα T_1 (η πλειάδα με τη μεγαλύτερη συχνότητα εμφάνισης). Για να θεωρηθεί το κριτήριο αυτό ως αξιόπιστο θα πρέπει η διαφορά της πιθανότητας της πλειάδας T_1 από τις πιθανότητες των άλλων πλειάδων να είναι σημαντική. Για παράδειγμα, αν $\hat{p}_1$=0,51 και $\hat{p}_2$=0,49 η πιθανότητα επιτυχούς επιλογής της πλειάδας T_1 είναι 0,5. Για να αποτραπεί το σύστημα να κάνει τέτοιου είδους επιλογές πρέπει οι πιθανότητες p_i να ικανοποιούν ορισμένες συνθήκες. Μια πιθανή συνθήκη είναι η τιμή της $\hat{p}_1$ να είναι μεγαλύτερη από ένα συγκεκριμένο όριο (threshold). Παρ'ολο που η μέθοδος αυτή ικανοποιεί το πιθανολογικό μοντέλο είναι ευάλωτη στο θόρυβο των δεδομένων. Ο θόρυβος στα δεδομένα εισάγεται εξαιτίας των σφαλμάτων που ενδεχομένως να συμβαίνουν κατά τη διάρκεια της αυτόματης συλλογής των στατιστικών δεδομένων. Προκειμένου να ληφθούν υπόψι τα προβλήματα του θορύβου χρησιμοποιείται ένα άλλο κριτήριο το οποίο λέγεται odds ratio. Η T_1 επιλέγεται ως η καταλληλότερη πλειάδα αν οι λόγοι

$$\frac{\hat{p}_1}{\hat{p}_2}, \frac{\hat{p}_1}{\hat{p}_3}, \ldots, \frac{\hat{p}_1}{\hat{p}_k}$$

ξεπερνούν ένα προκαθορισμένο όριο. Το κριτήριο αυτό είναι λιγότερο ευαίσθητο στο θόρυβο αφού εξαρτάται μόνο από τις δύο μεγαλύτερες μετρήσεις.

Η χρήση του πιθανολογικού μοντέλου εισάγει μερικές παραδοχές στη δομή των αντίτοιχων γλωσσολογικών δεδομένων. Οι παραδοχές αυτές είναι οι ακόλουθες:

- Οι πλειάδες T_i είναι αμοιβαία διαχωρισμένες (mutually disjoint.)

- Η εμφάνιση μιας πλειάδας T στην αρχική γλώσσα μπορεί πράγματι να μεταφραστεί σε μία από τις πλειάδες T_1, T_2, ..., T_k στη γλώσσα προορισμού.
- Κάθε εμφάνιση της πλειάδας T_i στη γλώσσα προορισμού μπορεί να είναι η μετάφραση μόνο της πλειάδας T στην αρχική γλώσσα.

Στη συνέχεια εξετάζεται η περίπτωση στην οποία διάφορες αμφίσημες λέξεις εμφανίζονται σε διάφορες συντακτικές πλειάδες. Δεδομένου ότι οι διαφορετικές σχέσεις μπορούν να οδηγήσουν σε διαφορετικές μεταφράσεις των λέξεων οι συγγραφείς του [4] έχουν υιοθετήσει μια στρατηγική για την επιλογή της κατάλληλης μετάφρασης των λέξεων σε μια πρόταση. Η στρατηγική αυτή αποτελείται από τα ακόλουθα βήματα:

1. Υπολόγισε για κάθε πλειάδα στην αρχική γλώσσα την ποσότητα $B \approx \ln \frac{p_1}{p_2}$. Αν η μεγαλύτερη τιμή του B είναι μικρότερη από το threshold τότε η διαδικασία τερματίζεται.
2. Έστω T η πλειάδα στην αρχική γλώσσα για την οποία η τιμή του B είναι μέγιστη. Επέλεξε ως μετάφραση των αμφίσημων λέξεων της πλειάδας T τις αντίστοιχες λέξεις της πλειάδας T_1 (η πλειάδα με τη μεγαλύτερη συχνότητα εμφάνισης). Αφαίρεσε την T από τη λίστα των πλειάδων της αρχικής γλώσσας.
3. Διέδωσε τον περιορισμό: αφίρεσαι τις πλειάδες στη γλώσσα προορισμού που είναι ασυνεπείς με την επιλογή αυτή. Αν κάποιες από τις πλειάδες στην αρχική γλώσσα γίνουν ασαφείς αφαιρούνται από τη λίστα των πλειάδων της αρχικής γλώσσας.
4. Επενέλαβε τη διαδικασία για τις υπόλοιπες πλειάδες μέχρι να αποσαφηνιστούν όλες οι αμφίσημες λέξεις, ή μέχρι η μέγιστη τιμή του B να γίνει μικρότερη από το threshold.

2.3 Μέθοδοι αποσαφήνισης μεταφράσεων

Στο άρθρο [5] οι Qu, Grefenstette και Evans παρουσιάζουν δύο μεθόδους αποσαφήνισης της έννοιας μια λέξης (translation disambiguation methods) ώστε να επιλεγεί η κατάλληλη μετάφραση αυτής. Οι μέθοδοι αυτοί βασίζονται στις παρακάτω παρατηρήσεις:

- Η σωστή μετάφραση μιας λέξης ενός ερωτήματος δεν είναι αμφίσημη αν ληφθεί υπόψι το πλαίσιο (context) (δηλαδή, οι υπόλοιπες λέξεις του ερωτήματος) στο οποίο βρίσκεται.
- Το Web και τα μεγάλα σώματα κειμένου μπορούν να χρησιμοποιηθούν για την αποσαφήνιση της μετάφρασης μιας λέξης.

Ας υποθέσουμε ότι ένα ερώτημα που αποτελείται από τις λέξεις $s_1,s_2,\ldots,s_5$ σε μία source language έχει τις ακόλουθες μεταφράσεις στη target language.

s_1	s_2	s_3	s_4	s_5
t_{11}	t_{21}	t_{31}	t_{41}	t_{51}
t_{12}	t_{22}	t_{32}		t_{52}
t_{13}		t_{33}		
		t_{34}		

Ο όρος s_1 έχει 3 πιθανές μεταφράσεις: t_{11}, t_{12} και t_{13}. Ένα context για το t_{11} μπορεί να κατασκευαστεί ως μια από τις πιθανές ακολουθίες συμπεριλαμβανομένων και των άλλων μεταφράσεων στην target language. Για παράδειγμα, κάθε μία από τις παρακάτω ακολουθίες συνιστά ένα context για κάθε μετάφραση αναφορικά με τις μεταφράσεις των υπολοίπων λέξεων. Το σύνολο αυτό των ακολουθιών ονομάζεται χώρος μεταφράσεων *(translation space)*.

<t_{11}, t_{21}, t_{31}, t_{41}, t_{51}>
<t_{11}, t_{21}, t_{32}, t_{41}, t_{51}>
<t_{11}, t_{21}, t_{33}, t_{41}, t_{51}>
<t_{11}, t_{21}, t_{34}, t_{41}, t_{51}>
…
<t_{13}, t_{22}, t_{34}, t_{41}, t_{52}>

2.3.1 Μέθοδος αποσαφήνισης μεταφράσεων βασισμένη στο Web

Ο ένας από τους δύο αλγορίθμους που προτείνονται στο [5] για την αποσαφήνιση της μετάφρασης μιας λέξης βασίζεται στο Web. Συγκεκριμένα τα βήματα του αλγορίθμου είναι τα ακόλουθα:

1. *Εξαγωγή όλων των πιθανών μεταφράσεων των λέξεων ενός ερωτήματος.*
2. *Δημιουργία όλων των πιθανών συνδυασμών των μεταφράσεων δηλαδή του translation space.*

3. *Εισαγωγή κάθε ακολουθίας σε μια δικτυακή πύλη (Web portal), π.χ., AltaVista.*
4. *Ορισμός του coherence score ως τον αριθμό των σελίδων που επιστρέφονται για κάθε ακολουθία μεταφράσεων.*
5. *Επιλογή της μετάφρασης για κάθε λέξη με το μεγαλύτερο coherence score.*

Δεδομένου ότι η σειρά των λέξεων δεν διατηρείται από μια γλώσσα σε άλλη, η ερώτηση που στέλνεται στο Web χρησιμοποιεί ένα χειριστή (operator) που επιβάλλει την παρουσία της μεταφρασμένης ακολουθίας αλλά όχι τη διατήρηση της σειράς των λέξεων. Η προηγμένη αναζήτηση της μηχανής αναζήτησης AltaVista [6] υποστηρίζει τον τελεστή NEAR, που εξασφαλίζει ότι οι λέξεις εμφανίζονται σε μικρή απόσταση μεταξύ τους, με οποιαδήποτε διάταξη. Ο τελεστής αυτός χρησιμοποιείται για τον υπολογισμό του coherence score κάθε ακολουθίας. Δεδομένου ότι οι μηχανές αναζήτης δεν αποκόπτουν τις καταλήξεις των λέξεων, το ερώτημα που στέλνεται στη μηχανή αναζήτησης επεκτείνεται ώστε να περιλαμβάνει παράγωγα κάθε λέξης. Για παράδειγμα, αν μια πιθανή μετάφραση ενός ερωτήματος είναι “big black dogs” το ερώτημα που στέλνεται στη μηχανή αναζήτησης είναι:

(big ***OR*** *bigger* ***OR*** *biggest)* ***NEAR*** *(black* ***OR*** *blacks* ***OR*** *blacked* ***OR*** *blacking* ***OR*** *blackest* ***OR*** *blacking)* ***NEAR*** *(dog* ***OR*** *dogs)*

2.3.2 Μέθοδος αποσαφήνισης μεταφράσεων βασισμένη σε ένα σώμα κειμένου

Δύο εναλλακτικές μέθοδοι επιλογής της κατάλληλης μετάφρασης ενός ερωτήματος που προτείνεται στο άρθρο [5] βασίζονται σε ένα σύνολο από κείμενα (corpus) και σε στατιστικές μετρικές για τη μέτρηση του coherence score.

Ο ένας από τους τρόπους επιλογής της κατάλληλης μετάφρασης βασίζεται στην αμοιβαία πληροφορία των λέξεων του ερωτήματος. Συγκεκριμένα, τα βήματα της μεθόδου αυτής είναι τα ακόλουθα:

1. *Εξαγωγή όλων των πιθανών μεταφράσεων των λέξεων ενός ερωτήματος.*
2. *Δημιουργία όλων των πιθανών συνδυασμών των μεταφράσεων δηλαδή του translation space.*
3. *Υπολογισμός της αμοιβαίας πληροφορίας (mutual information, MI) για όλα τα ζεύγη των όρων για κάθε δυνατή ακολουθία μεταφράσεων του translation space.*

4. *Άθροισμα των τιμών της αμοιβαίας πληροφορίας. Το άθροισμα αυτό καλείται coherence score της ακολουθίας.*
5. *Επιλογή της μετάφρασης με το μεγαλύτερο coherence score.*

Η αμοιβαία πληροφορία δύο όρων t_1 και t_2 ορίζεται ως εξής:

$$MI(t_1, t_2) \approx \log \frac{A \times N}{(A + C) \times (A + B)}$$

όπου,

- A: η συχνότητα συνεμφάνισης των όρων t_1 και t_2,
- B: η συχνότητα εμφάνισης του όρου t_1 χωρίς τον t_2,
- C: η συχνότητα εμφάνισης του όρου t_2 χωρίς τον t_1, και
- N: ο συνολικός αριθμός των κειμένων.

Ο δεύτερος εναλλακτικός αλγόριθμος μετάφρασης περιλαμβάνει τα ακόλουθα βήματα:

1. *Εξαγωγή όλων των πιθανών μεταφράσεων των λέξεων ενός ερωτήματος.*
2. *Δημιουργία όλων των πιθανών συνδυασμών των μεταφράσεων δηλαδή του translation space*
3. *Αναζήτηση για κάθε δυνατή ακολουθία μεταφράσεων σχετικών κειμένων από το corpus.*
4. *Υπολογισμός του coherence score ως το άθροισμα του βαθμού ομοιότητας (similarity score) των N πρώτων ανακτηθέντων κειμένων για κάθε ακολουθία μεταφράσεων.*
5. *Επιλογή της ακολουθίας με το μεγελύτερο coherence score.*

2.4 Δια-γλωσσική Ανάκτηση Πληροφοριών με χρήση Οντολογιών

Τα περισσότερα συστήματα δια-γλωσσικής ανάκτησης πληροφοριών βασίζονται σε λεξικά προκειμένουν να αποσαφηνίσουν την έννοια μιας αμφίσημης λέξης και να επιλέξουν την κατάλληλη μετάφραση αυτής. Οι συγγραφείς στο [7] περιγράφουν ένα CLIR σύστημα στο οποίο η μετάφραση του αρχικού ερωτήματος βασίζεται σε μια οντολογία.

Η οντολογία είναι μια έννοια που χρησιμοποιήθηκε για πρώτη φορά από τους αρχαίους Έλληνες φιλοσόφους στην προσπάθειά τους να απαντήσουν σε κάποια φιλοσοφικά ερωτήματα σχετικά με την ουσία και την ύπαρξη κάποιων πραγμάτων και εννοιών. Με τον όρο οντολογία εννοούμε την ακριβή περιγραφή πραγμάτων και εννοιών καθώς και των σχέσεων που υπάρχουν ανάμεσα τους. Ο πιο γνωστός ορισμός για την οντολογία, στην επιστήμη των υπολογιστών, πάνω στον οποίο στηρίχτηκαν και άλλοι ορισμοί, δόθηκε από τον Gruber [8] και είναι ο ακόλουθος:

- An ontology is an explicit specification of a conceptualization

Κάθε έννοια της οντολογίας (concept) περιέχει ένα σύνολο από χαρακτηριστικά τα οποία επιτρέπουν στο χρήστη να αποσαφηνίσει το concept αυτό. Επίσης, κάθε concept συνδέτεται με άλλα concepts μέσω ενός συνόλου από καθορισμένες σχέσεις. Για παράδειγμα, στην Εικόνα 5 φαίνονται τα χαρακτηριστικά του concept 'SHIP' καθώς και οι αντίστοιχες μεταφράσεις στην Ισπανική και Κινεζική γλώσσα.

Η διαδικασία με την οποία γίνεται η ανάκτηση πληροφοριών από την Αγγλική γλώσσα στην Ισπανική είναι η εξής:

1. *Εισαγωγή ενός ερωτήματος από το χρήστη στην Αγγλική γλώσσα.*
2. *Παραγωγή της λίστας με τα concepts.*
3. *Επιλογή του κατάλληλου concept από τη λίστα.*
4. *Όλες οι Ισπανικές λέξεις που συνδέονται με το concept θεωρούνται ως Ισπανικά ερωτήματα.*
5. *Τα ανακτηθέντα έγγραφα παρουσιάζονται στο χρήστη.*

Το σύστημα αυτό έχει το βασικό μειονέκτημα ότι δεν αποσαφηνίζει μόνο του την έννοια κάθε λέξης αλλά αφήνει στο χρήστη να επιλέξει με ποια έννοια χρησιμοποιείται η εκάστοτε λέξη του ερωτήματος.

Εικόνα 5. Concept 'SHIP'

ΚΕΦΑΛΑΙΟ 3

ΠΙΘΑΝΟΚΡΑΤΙΚΕΣ ΜΕΘΟΔΟΙ ΑΡΣΗΣ ΑΜΦΙΣΗΜΙΑΣ ΤΗΣ ΕΝΝΟΙΑΣ ΤΩΝ ΛΕΞΕΩΝ

Με τον όρο αποσαφήνιση της έννοιας των λέξεων (Word Sense Disambiguation, WSD) εννοούμε τη διαδικασία με την οποία προσδιορίζεται η έννοια με την οποία χρησιμοποιείται μια λέξη μέσα σε μια πρόταση. Η αυτόματη αποσαφήνιση των εννοιών μιας λέξης αποτελεί σημείο ενδιαφέροντος από τη δεκαετία του 1950. Το WSD [9] είναι μια «ενδιάμεση διαδικασία» η οποία δεν είναι αυτοσκοπός αλλά συνήθως είναι απαραίτητη σε εφαρμογές επεξεργασίας φυσικής γλώσσας (Natural Language Processing, NLP) [10] όπως:

- αυτόματη μετάφραση: επιλογή της κατάλληλης μετάφρασης μιας λέξης. Για παράδειγμα η λέξη '*ιστός*' ανάλογα με το πλαίσιο (context) στο οποίο βρίσκεται μεφράζεται ως '*web*', '*tissue*', ή '*mast*'.
- ανάκτηση πληροφοριών: όταν ψάχνουμε για έγγραφα με συγκεκριμένες λέξεις-κλειδιά είναι επιθυμητό να αποκλείουμε έγγραφα στα οποία οι λέξεις χρησιμοποιούνται με μια ακατάλληλη έννοια.
- γραμματική ανάλυση (grammatical analysis): η αποσαφήνιση της έννοιας μιας λέξης είναι χρήσιμη για να βρούμε το μέρος του λόγου κάθε λέξης σε μια πρόταση, μια διαδικασία γνωστή ως Part-Of-Speech tagging.
- επεξεργασία κειμένου (text processing): η αποσαφήνιση της έννοιας μιας λέξης χρησιμοποιείται όταν γίνονται ορθογραφικές διορθώσεις σε ένα κείμενο.

Στο κεφάλαιο αυτό περιγράφονται δύο πιθανοκρατικές μέθοδοι που χρησιμοποιήθηκαν για την άρση της αμφισημίας της έννοιας των λέξεων με σκοπό την επιλογή της κατάλληλης μετάφρασης αυτών. Συγκεκριμένα, στην ενότητα 3.1 περιγράφεται ο

αλγόριθμος Naïve Bayes και στην ενότητα 3.2 περιγράφονται τα Γλωσσολογικά Μοντέλα (Language Models).

3.1 Ο Αλγόριθμος Naïve Bayes

Ο Naïve Bayes [11, 23] ανήκει στην κατηγορία των αλγορίθμων εποπτευόμενης μάθησης (supervised learning). Οι αλγόριθμοι της κατηγορίας αυτής χρησιμοποιούν ένα σώμα κειμένου ως training set. Στην παρούσα εργασία τον ρόλο του training set παίζουν τα Google 5-grams τα οποία περιγράφονται στην ενότητα 4.1.

Η ιδέα του Naïve Bayes αλγορίθμου είναι ότι εξετάζει τις λέξεις που ανήκουν στο ίδιο context με τη αμφίσημη λέξη, τη λέξη δηλαδή της οποίας την έννοια προσπαθεί να αποσαφηνίσει.

Στόχος του Naïve Bayes είναι η επιλογή της κατάλληλης μετάφρασης t_k μιας αμφίσημης λέξης w σε ένα δοσμένο context c. Έστω ότι $v_1,v_2,\ldots,v_j$ είναι οι λέξεις που ανήκουν στο context c. Συμβολίζουμε επίσης με $t_1,t_2,\ldots,t_k$ όλες τις πιθανές μεταφράσεις της λέξης w. Ο Naïve Bayes εφαρμόζει τον ακόλουθο κανόνα απόφασης (decision rule).

$$\textit{Bayes decision rule} \rightarrow \textit{Επέλεξε τη μετάφραση } t' \textit{ αν } P(t' \mid c) > P(t_k \mid c) \textit{ για } t_k \neq t'$$

Ο Bayes κανόνας απόφασης είναι βέλτιστος γιατί ελαχιστοποιεί την πιθανότητα σφάλματος εφόσον σε κάθε περίπτωση επιλέγει την έννοια με τη μεγαλύτερη υπό συνθήκη πιθανότητα.

Επειδή δεν γνωρίζουμε την πιθανότητα $P(t_k|c)$ μπορούμε να την υπολογίσουμε χρησιμοποιώντας την ακόλουθη σχέση:

$$P(t_k \mid c) = \frac{P(c \mid t_k)}{P(c)} P(t_k)$$

$P(t_k)$ είναι η prior πιθανότητα της μετάφρασης t_k, η πιθανότητα δηλαδή να συναντήσουμε τη λέξη tk αν δεν ξέρουμε τίποτα για το context. Η πιθανότητα $P(t_k)$ πολλαπλασιάζεται με τον παράγοντα $P(c|t_k)/P(c)$ στον οποίο περικλείεται η γνώση που έχουμε για το context.

Αν αυτό που θέλουμε είναι να επιλέξουμε την κατάλληλη μετάφραση, μπορούμε να απλοποιήσουμε τη διαδικασία της αποσαφήνισης αφαιρώντας τον όρο P(c), ο οποίος είναι σταθερός για όλες τις μεταφράσεις και δεν επηρεάζει το αποτέλεσμα. Μπορούμε επίσης να χρησιμοποιήσουμε τους λογαρίθμους των πιθανοτήτων για να κάνουμε τους υπολογισμούς πιο εύκολους. Έχουμε λοιπόν:

$$t' = \arg\max P(t_k \mid c) \Rightarrow$$
$$t' = \arg\max \frac{P(c \mid t_k)}{P(c)} P(t_k) \Rightarrow$$
$$t' = \arg\max P(c \mid t_k) P(t_k) \Rightarrow$$
$$t' = \arg\max[\log P(c \mid t_k) + \log P(t_k)]$$

Χρησιμοποιώντας την υπόθεση ανεξαρτησίας του Naïve Bayes σύμφωνα με την οποία όλα τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, ισχύει δηλαδή η σχέση

$$P(c \mid t_k) = \prod_{v_j} P(v_j \mid t_k)$$

καταλήγουμε στον ακόλουθο τροποποιημένο κανόνα απόφασης του Naïve Bayes:

Naïve Bayes decision rule → *Επέλεξε τη μετάφραση t′ αν*

$$t' = arg\ \max[\log P(t_k) + \sum_{vj \in c} \log P(v_j \mid t_k)]$$

Οι πιθανότητες $P(t_k)$ και $P(v_j|t_k)$ υπολογίζονται από το training set, δηλαδή από τα Google 5-grams. Στην Εικόνα 6, περιγράφεται η διαδικασία επιλογής της κατάλληλης μετάφρασης μιας λέξης με τη χρήση του αλγορίθμου Naïve Bayes. Ο όρος C(x, y) αναπαριστά τη συχνότητα εμφάνισης των λέξεων x και y στο ίδιο context.

Training phase

For all translations t_k of *w* **do**
  For all words v_j in the vocabulary **do**

$$P(v_j \mid t_k) = \frac{C(v_j, t_k)}{\sum_i C(v_i, t_k)}$$

  End
End
For all translations t_k of *w* **do**

$$P(t_k) = \frac{C(t_k)}{\sum_i C(t_i)}$$

End

Disambiguation phase

For all translations t_k of *w* **do**

$$score(t_k) = \log P(t_k)$$

  For all words v_j in the context window **do**

$$score(t_k) = score(t_k) + \log P(v_j \mid t_k)$$

  End
End
choose $t' = \arg\max_{tk} score(t_k)$

Εικόνα 6. Αλγόριθμος Naïve Bayes

Ας θεωρήσουμε για παράδειγμα το ερώτημα '*Ο παγκόσμιος ιστός περιέχει πληροφορίες για το μάθημα*'. Το ερώτημα αυτό μεταφράζεται ως εξής: '*The universal {tissue, web, mast} contains information for the lesson*'. Στο ερώτημα αυτό η λέξη ιστός είναι αμφίσημη. Η λέξη ιστός έχει τις ακόλουθες πιθανές μεταφράσεις: t_1 = tissue, t_2 = web και t_3 = mast, ενώ οι λέξεις οι οποίες συμμετέχουν στο context είναι: v_1 = universal, v_2 =

contains, v_3 = information και v_4 = lesson (τα άρθρα και οι προθέσεις δε συμμετέχουν στη διαδικασία της αποσαφήνισης για αυτό και αγνοούνται από το context).

4.2 Γλωσσολογικά Μοντέλα

Ένας εναλλακτικός τρόπος αποσαφήνισης της έννοιας των αμφίσημων λέξεων που χρησιμοποιήθηκε στην παρούσα εργασία είναι τα γλωσσολογικά μοντέλα (language models) [12, 23]. Τα γλωσσολογικά μοντέλα χρησιμοποιούνται για να προβλέψουν την εμφάνιση της λέξης w_2 αμέσως μετά τη λέξη w_1.

Η χρήση των μοντέλων αυτών είναι θεμελιώδης στην αναγνώριση ομιλίας (speech recognition) και στην οπτική αναγνώριση χαρακτήρων. Χρησιμοποιείται επίσης για τη διόρθωση ορθογραφίας, την αναγνώριση γραφής, και τη στατιστική αυτόματη μετάφραση (statistical machine translation).

4.2.1 N-gram Μοντέλα

Ο στόχος της πρόβλεψης της επόμενης λέξης σε μια ακολουθία λέξεων μπορεί να θεωρηθεί ως ο υπολογισμός της πιθανότητας $P(w_n \mid w_1, w_2, \ldots, w_{n-1})$.

Σε ένα τέτοιο πιθανολογικό πρόβλημα, χρησιμοποιούμε μια ταξινόμηση των προηγούμενων λέξεων (ιστορία), για να προβλέψουμε την επόμενη λέξη. Έχοντας εξετάσει ένα μεγάλο σώμα κειμένου είναι δυνατόν να γνωρίζουμε ποιες λέξεις τείνουν να ακολουθήσουν άλλες λέξεις.

Εντούτοις, είναι αδύνατο να εξετάζουμε κάθε ιστορία χωριστά: τις περισσότερες φορές συναντάμε μια πρόταση την οποία δεν έχουμε ξανασυναντήσει οπότε δεν υπάρχει κάποια όμοια ιστορία στην οποία θα μπορούσαν να βασιστούν οι προβλέψεις μας. Έτσι, χρειαζόμαστε μια μέθοδο ομαδοποίησης των ιστοριών που είναι παρόμοιες με κάποιο τρόπο από τις οποίες να μπορούμε να εξάγουμε ποιες λέξεις αναμένεται να ακολοθήσουν μια γνωστή ακολουθία λέξεων. Ένας πιθανός τρόπος να πραγματοποιηθεί η ομαδοποίηση αυτή είναι να υποθέσουμε ότι μόνο οι τελευταίες λέξεις στην ακολουθία επηρρεάζουν τη λέξη που ακολουθεί (Markov assumption). Αν κατασκευάσουμε ένα μοντέλο στο οποίο όλες οι ιστορίες που έχουν όμοιες τις τελευταίες n-1 λέξεις τοποθετηθούν στην ίδια κλάση, τότε έχουμε ένα μοντέλο Markov (n-1) τάξης.

Στην εργασία αυτή χρησιμοποιούμε δύο N-gram μοντέλα: bigram και trigram μοντέλο για τα οποία N=2 και N=3 αντίστοιχα. Το bigram μοντέλο χρησιμοποιεί τον ακόλουθο τύπο για τον υπολογισμό της πιθανότητας εμφάνισης μιας ακολουθίας από n λέξεις.

$$P(w_1, w_2, ..., w_n) = P(w_1)P(w_2 \mid w_1) \cdots P(w_n \mid w_{n-1})$$

Με βάση τον τύπο αυτό η εμφάνιση κάθε λέξης εξαρτάται μόνο από την προηγούμενη λέξη.

Στην περίπτωση του trigram μοντέλου η πιθανότητα εμφάνισης μιας ακολουθίας από n λέξεις είναι ίση με:

$$P(w_1, w_2, ..., w_n) = P(w_1)P(w_2 \mid w_1)\prod_{j=3}^{n} P(w_j \mid w_{j-1}w_{j-2})$$

όπου η εμφάνιση κάθε λέξης εξαρτάται από τις δύο προηγούμενες λέξεις. Στους παραπάνω τύπους η πιθανότητα $P(w_1)$ ισούται με τον αριθμό των πενταγράμμων του Google που περιέχουν τη λέξη w_1 προς τον αριθμό όλων των πενταγράμμων.

ΚΕΦΑΛΑΙΟ 4

ΠΕΡΙΓΡΑΦΗ ΜΕΘΟΔΟΛΟΓΙΑΣ

Στο κεφάλαιο αυτό περιγράφεται η μεθοδολογία που υλοποιήθηκε και αξιολογήθηκε στην παρούσα εργασία για την μετάφραση ερωτημάτων. Συγκερκιμένα, περιγράφονται οι τεχνολογίες καθώς και τα δεδομένα εισόδου που χρησιμοποιήθηκαν για την εκτέλεση της παρούσας εργασίας.

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο ο αλγόριθμος Naïve Bayes για να εκτελεστεί απαιτεί ένα σώμα κειμένου (corpus) το οποίο χρησιμοποιεί προκειμένου να υπολογιστούν οι απαιτούμενες πιθανότητες κατά την εκτέλεσή του. Για το σκοπό αυτό χρησιμοποιήθηκαν τα 5-grams του Google τα οποία περιγράφονται αναλυτικά στην ενότητα 4.1. Στην ενότητα 4.2 αναλύεται η μηχανή αναζήτησης Lucene με τη βοήθεια της οποίας έγινε δυνατή η χρήση του συγκεκριμένου συνόλου δεδομένων. Στην ενότητα 4.3 μια συνοπτική αναφορά στο λεξικό που χρησιμοποιείται για τη μετάφραση των λέξεων, ενώ στην ενότητα 4.5 γίνεται μια επισκόπηση ολόκληρης της μεθοδολογίας.

4.1 Google n-grams

Το 2006 η Google Inc. εξέδωσε ένα σύνολο δεδομένων (dataset), τα λεγόμενα Google n-grams [13], το οποίο περιέχει ακολουθίες αγγλικών λέξεων καθώς και τις συχνότητες εμφάνισης τους. Το μήκος των n-grams κυμαίνεται από τα unigrams (μεμονωμένες λέξεις) έως τα πεντάγραμμα (5-grams), τα οποία αποτελούν ακολουθίες πέντε λέξεων.

```
Number of tokens:    1,024,908,267,229
Number of sentences:    95,119,665,584
Number of unigrams:         13,588,391
Number of bigrams:         314,843,401
Number of trigrams:        977,069,902
Number of fourgrams:     1,313,818,354
Number of fivegrams:     1,176,470,663
```

Εικόνα 7. Μέγεθος των 5-grams

Τα n-grams παρήχθησαν από περίπου ένα τρισεκατομμύριο λέξεις από ιστοσελίδες γραμμένες στην Αγγλική γλώσσα. Μόνο τα n-grams με συχνότητα εμφάνισης μεγαλύτερη από 40 προστέθηκαν στο τελικό dataset. Το dataset αυτό αποτελείται από συμπιεσμένα αρχεία κειμένου συνολικού μεγέθους 24GB. Στην Εικόνα 7 φαίνεται ο αριθμός των n-grams διαφόρου μεγέθους που συνιστούν το dataset αυτό, ενώ στην Εικόνα 8 παρουσιάζεται ένα δείγμα από 4-grams. Στην παρούσα εργασία χρησιμοποιήθηκαν μόνο τα 5-grams.

```
serve as the incoming   92
serve as the incubator  99
serve as the independent        794
serve as the index      223
serve as the indication 72
serve as the indicator  120
serve as the indicators 45
serve as the indispensable      111
serve as the indispensible      40
serve as the individual 234
serve as the industrial 52
serve as the industry   607
serve as the info       42
serve as the informal   102
serve as the information        838
serve as the informational      41
serve as the infrastructure     500
serve as the initial    5331
serve as the initiating 125
```

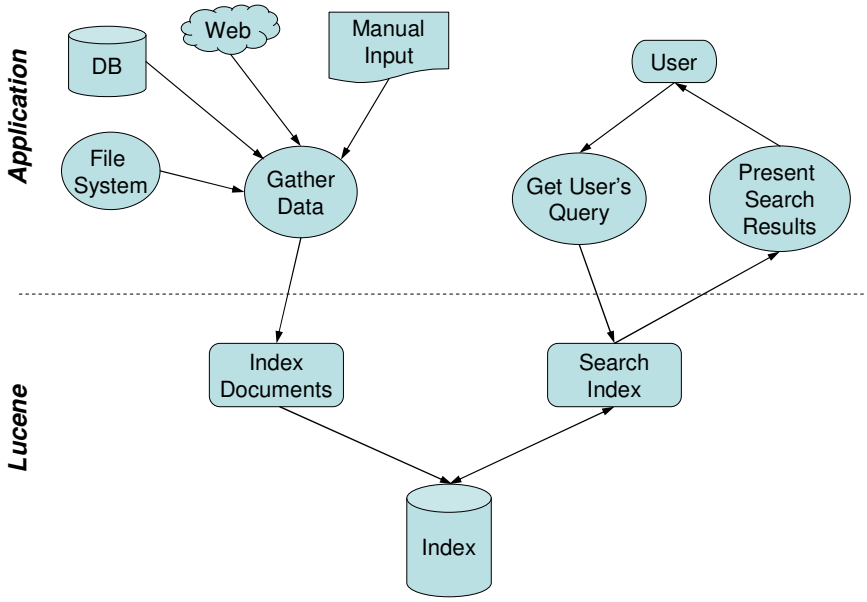
Εικόνα 8. Δείγμα από τα 4-grams

4.2 Μηχανή Αναζήτησης Lucene

Το Lucene [14, 15] είναι μία ανοιχτού κώδικα (open source) βιβλιοθήκη ευρετηριασμού και αναζήτησης κειμένων η οποία υλοποιήθηκε αρχικά σε Java από τον Doug Cutting και υποστηρίζεται από το Apache Software Foundation [16].

Το Lucene δίνει τη δυνατότητα προσθήκης ευρετηρίων και αναζήτησης αυτών σε εφαρμογές. Το Lucene μπορεί να ερευτηριοποιήσει και να καταστήσει δυνατή την αναζήτηση μεγάλου πλήθους δεδομένων αρκεί αυτά να μπορούν να μετατραπούν σε text μορφή. Όπως δείχνει και η Εικόνα 9 στο Lucene δεν παίζει ρόλο η πηγή των δεδομένων, η μορφή τους ή γλώσσα αρκεί να μπορούν να μετατραπούν σε κείμενο. Αυτό σημαίνει ότι το Lucene μπορεί να χρησιμοποιηθεί για την ευρετηριοποίηση και την αναζήτηση δεδομένων αποθηκευμένα στο τοπικό σύστημα αρχείων, σε απομακρυσμένους εξυπηρετητές δικτύων (remote web servers), σε βάσεις δεδομένων,

κ.ά. Επίσης, το Lucene μπορεί να χρησιμοποιηθεί για την ερευτηριοποίηση Microsoft Word αρχείων, HTML ή PDF αρχείων ή οποιασδήποτε άλλης μορφής αρχείων.



Εικόνα 9. Ενσωμάτωση Lucene σε εφαρμογές

Καρδιά όλων των μηχανών αναζήτησης είναι η δημιουργία του ευρετηρίου, ο τρόπος δηλαδή με τον οποίο θα επεξεργαστούν τα δεδομένα προκειμένου να είναι αποδοτική και γρήγορη η αναζήτησή τους. Στις ενότητες που ακολουθούν περιγράφεται ο τρόπος με τον οποίο ευρετηριοποίηθηκαν τα Google 5-grams καθώς και ο τρόπος με τον όποιο γίνεται η αναζήτησή τους με τη βοήθεια του Lucene.

4.2.1 Δημιουργία ευρετηρίου με τη χρήση του Lucene

Οι περισσότερες μηχανές αναζήτησης χρησιμοποιούν τα B-δέντρα για τη δημιουργία ευρετηρίου στα οποία οι αναζητήσεις και οι εισαγωγές κόμβων γίνονται με πολυπλοκότητα O(logn) (όπου n είναι το πλήθος των κόμβων του δέντρου). Το Lucene υιοθετεί μια ελαφρώς διαφορετική προσέγγιση: αντί να διατηρεί ένα μόνο ευρετήριο δημιουργεί πολλαπλά τμήματα αυτού (index segments) και τα συχωνεύει περιοδικά. Για κάθε νέο έγγραφο που ευρετηριοποιείται το Lucene δημιουργεί ένα νέο index segment, ενώ συγχωνεύει τα μικρά τμήματα με μεγαλύτερα κρατώντας έτσι τον αριθμό των index segments μικρό. Με τον τρόπο αυτό οι αναζητήσεις παραμένουν γρήγορες. Κατά τη

συγχώνευση των τμημάτων το Lucene δημιουργεί ένα νέο τμήμα και διαγράφει όλα τα παλιά.

Οι κυριότερες κλάσεις του Lucene που χρησιμοποιούνται για τη δημιουργία του ευρετηρίου είναι οι ακόλουθες:

- IndexWriter
- Directory
- Analyzer
- Document
- Field

Στις ακόλουθες παραγράφους δίνεται μια μικρή περιγραφή των κλάσεων αυτών.

4.2.1.1 Κλάση IndexWriter

Η κλάση IndexWriter αποτελεί την κεντρική συνιστώσα δημιουργίας του ευρετηρίου. Δημιουργεί ένα νέο ευρετήριο και προσθέτει δεδομένα (documents) σε ένα υπάρχον. Χρησιμοποιείται για να γράψει δεδομένα σε ένα ευρετήριο και όχι για να διαβάσει ή να αναζητήσει δεδομένα από αυτό.

4.2.1.2 Κλάση Directory

Η κλάση Directory αναπαριστά την τοποθεσία στην οποία είναι αποθηκευμένο το ευρετήριο του Lucene. Στην παρούσα υλοποίηση το ευρετήριο είναι αποθηκευμένο στο δίσκο και όχι στη μνήμη. Για το λόγο αυτό, για τη δημιουργία ενός αντικειμένου της κλάσης Directory χρησιμοποιήθηκε ένας κατάλογος του τοπικού συστήματος αρχείων. Συγκεκριμένα, όπως φαίνεται και στην Εικόνα 10 χρησιμοποιήθηκε η κλάση FSDirectory η οποία διατηρεί μια λίστα με πραγματικά αρχεία του συστήματος αρχείων.

4.2.1.3 Κλάση Analyzer

Η κλάση Analyzer χρησιμοποιείται για να φιλτράρει τα δεδομένα που ευρετηριοποιούνται. Είναι μια αφηρημένη κλάση (abstract class) άλλα το Lucene περιέχει αρκετές υλοποιήσεις αυτής. Μερικές από αυτές είναι υπεύθυνες για την αφαίρεση των stop words από τα κείμενα που προστίθενται στο ευρετήριο. Ο όρος stop

words αναφέρεται σε λέξεις που χρησιμοποιούνται πολύ συχνά αλλά δε συμμετέχουν στη διάκριση ενός κειμένου από ένα άλλο (π.χ., on, an, the, him, her). Άλλες υλοποιήσεις της κλάσης αυτής μετατρέπουν τα κείμενα σε πεζά γράμματα (lowercase) ώστε να μην υπάρχει διάκριση πεζών-κεφαλαίων κατά τη διάρκεια των αναζητήσεων.

4.2.1.4 κλάση Document

Ένα αντικείμενο της κλάσης Document αναπαριστά ένα σύνολο από πεδία (fields). Ένα Document μπορεί να είναι ένα σύνολο από δεδομένα, μια ιστοσελίδα, ένα μήνυμα ηλεκτρονικού ταχυδρομείου ή ένα αρχείο κειμένου το οποίο θέλουμε να προσθέσουμε στο ευρετήριο ώστε να μπορέσουμε να το ανακτήσουμε.

4.2.1.5 Κλάση Field

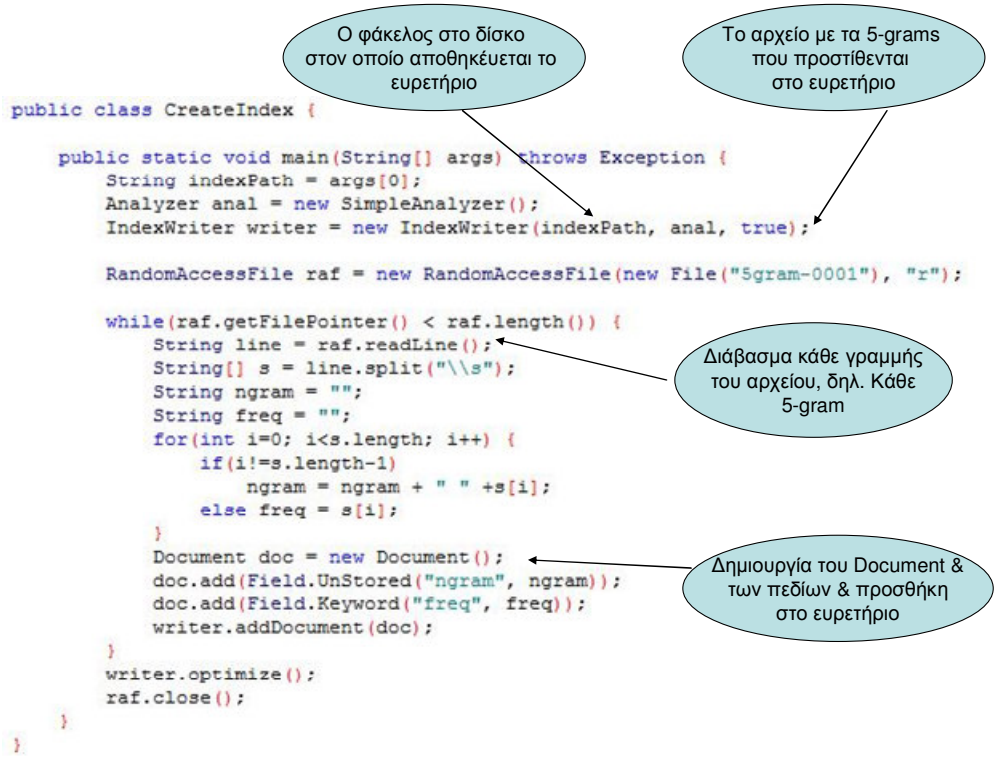
Κάθε αντικείμενο της κλάσης Document περιέχει ένα ή περισσότερα πεδία τα οποία αναπαριστώνται από την κλάση Field. Κάθε πεδίο αντιστοιχεί είτε στα δεδομένα που χρησιμοποιούνται κατά την αναζήτηση ως λέξεις κλειδιά ή στα δεδομένα που ανακτώνται από το ευρετήριο κατά την αναζήτηση. Υπάρχουν τέσσερις τύποι πεδίων που θα μπορούσαν να οριστούν:

- Field.Keyword: Τα δεδομένα του τύπου αυτού αποθηκεύονται, προστίθενται στο ευρετήριο αλλά δεν κατακερματίζονται (tokenization). Ο τύπος αυτός χρησιμοποποιείται κυρίως για τα πεδία που πρέπει να αποθηκευτούν αυτούσια (π.χ., ημερομηνίες).
- Field.Text: Τα δεδομένα αυτού του τύπου αποθηκεύονται, προστίθενται στο ευρετήριο και κατακερματίζονται. Δεν θα πρέπει να χρησιμοποιείται με δεδομένα μεγάλου μεγέθους (π.χ., το πλήρες κείμενο ενός άρθρου), διότι αυτά θα αποθηκεύονται στο ευρετήριο, το οποίο θα γίνει ιδιαίτερα μεγάλο.
- Field.UnStored: Τα δεδομένα του πεδίου αυτού δεν αποθηκεύονται αλλά προστίθενται στο ευρετήριο και κατακερματίζονται. Δεδομένα μεγάλου μεγέθους θα πρέπει να προστίθενται στο ευρετήριο μέσω πεδίων τύπου UnStored.
- Field.UnIndexed: Τα δεδομένα του πεδίου αυτού αποθηκεύονται αλλά δεν προστίθενται στο ευρετήριο ούτε κατακερματίζονται. Το πεδίο αυτό χρησιμοποιείται για δεδομένα που θέλουμε να μας επιστραφούν στις αναζητήσεις μας, αλλά δεν αναζητούμε πάνω σε αυτά τα δεδομένα.

Τα χαρακτηριστικά των παραπάνω τύπων πεδίων συνοψίζονται στον Πίνακα 1.

Πίνακας 1. Σύνοψη των τύπων πεδίων και των χαρακτηριστικών τους

Field Type	Analyzed	Indexed	Stored	Usage
Field.Keyword		x	x	telephones, dates, URLs
Filed.Text	x	x	x	document titles and content
Field.UnStored	x	x		document titles and content
Field,UnIndexed			x	document type (e.g.,PDF, HTML)



Εικόνα 10. Δημιουργία ευρετηρίου και προσθήκη των 5-grams σε αυτό

4.2.1.6 Ερευτηριοποίηση των Google 5-grams

Στην παρούσα εργασία χρησιμοποιήθηκε το Lucene για την ερευτηριοποίηση των Google 5-grams και την ανάκτηση δεδομένων από αυτά. Για κάθε πεντάγραμμο

δημιουργήθηκε ένα αντικείμενο της κλάσης Document με δύο πεδία. Το ένα πεδίο -με όνομα ngram- αποτελεί η ακολουθία των λέξεων και το άλλο -με όνομα freq- αποτελούσε η συχνότητα εμφάνισης του αντίστοιχου πεντάγραμμου. Στην Εικόνα 10, φαίνεται ο πηγαίος κώδικας δημιουργίας του ερευτηρίου και προσθήκης των πενταγράμμων ενός αρχείου σε αυτό.

4.2.2 Διαδικασία Αναζήτησης με τη Χρήση του Lucene

Στην ενότητα αυτή περιγράφονται συνοπτικά οι βασικότερες κλάσεις του Lucene που χρησιμοποιούνται για την ανάκτηση δεδομένων από το ευρετήριο. Οι κλάσεις αυτές είναι οι ακόλουθες:

- IndexSearcher
- Query
- QueryParser
- Hits

4.2.2.1 Κλάση IndexSearcher

Η κλάση IndexSearcher είναι για την αναζήτηση ό,τι είναι η κλάση IndexWriter για την ευρετηριοποίηση. Η κλάση IndexSearcher ανοίγει ένα ευρετήριο για διάβασμα (read-only mode) και προσφέρει ένα σύνολο από μεθόδους αναζήτησης.

4.2.2.2 Κλάση Query

Η τάξη Query είναι μία αφηρημένη τάξη η οποία περιέχει τα κριτήρια αναζήτησης που δημιουργούνται από τον QueryParser. Το Lucene περιέχει αρκετές υλοποιήσεις της κλάσης αυτής μεταξύ των οποίων είναι οι κλάσεις: BooleanQuery, TermQuery, RangeQuery, κ.ά.

4.2.2.3 Κλάση QueryParser

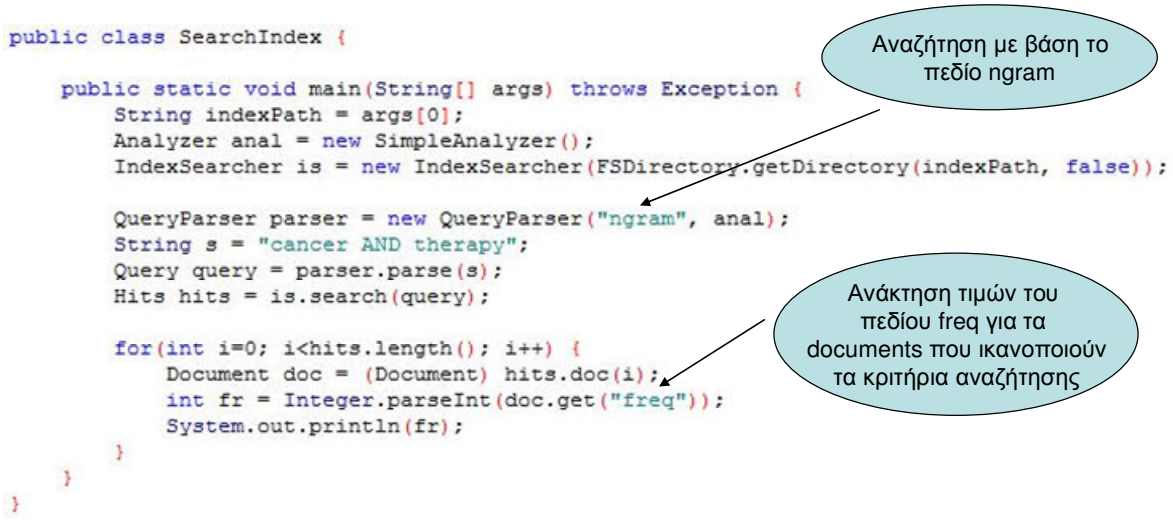
Η τάξη QueryParser χρησιμοποιείται στην κατασκευή ενός parser ο οποίος μπορεί να αναζητήσει δεδομένα σε ένα ευρετήριο. Η κλάση αυτή είναι υπεύθυνη για την ανάλυση του ερωτήματος με το οποίο ζητούνται δεδομένα από το ευρετήριο (βλ. Εικόνα 11).

4.2.2.4 Κλάση Hits

Μία αναζήτηση στο Lucene επιστρέφει ένα αντικείμενο της κλάσης Hits. Ένα αντικείμενο της κλάσης αυτής περιέχει δείκτες προς εκείνα τα Documents τα οποία ικανοποιούν τα κριτήρια της αναζήτησης.

4.2.2.5 Αναζήτηση Δεδομένων από τα Google 5-grams

Στην Εικόνα 11 παρουσιάζεται ο κώδικας με τον οποίο αναζητούνται οι συχνότητες εμφάνισης στο ευρετήριο του ερωτήματος “cancer AND therapy”. Αναζητούνται δηλαδή όλες οι συχνότητες εμφάνισης των πενταγράμμων που περιέχουν τις λέξεις cancer και therapy. Η αναζήτηση γίνεται με βάση το πεδίο ngram ενώ επιστρέφονται οι τιμές του πεδίου freq για εκείνα τα documents που περιέχουν στο πεδίο ngram τις λέξεις cancer και therapy.



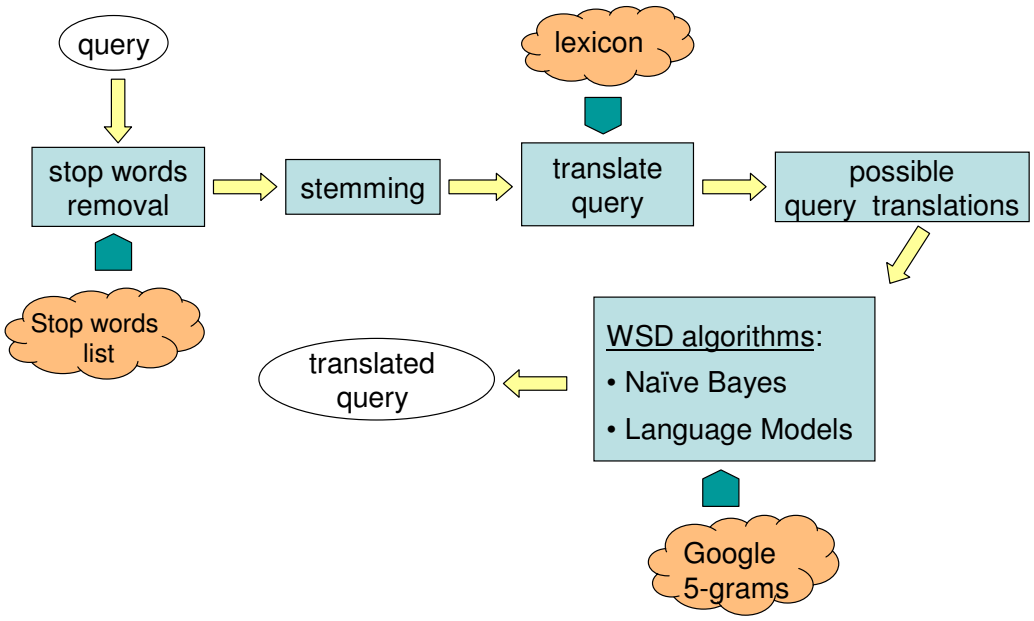

```
public class SearchIndex {

    public static void main(String[] args) throws Exception {
        String indexPath = args[0];
        Analyzer anal = new SimpleAnalyzer();
        IndexSearcher is = new IndexSearcher(FSDirectory.getDirectory(indexPath, false));

        QueryParser parser = new QueryParser("ngram", anal);
        String s = "cancer AND therapy";
        Query query = parser.parse(s);
        Hits hits = is.search(query);

        for(int i=0; i<hits.length(); i++) {
            Document doc = (Document) hits.doc(i);
            int fr = Integer.parseInt(doc.get("freq"));
            System.out.println(fr);
        }
    }
}
```

Εικόνα 11. Αναζήτηση δεδομένων από το ευρετήριο

4.3 Επισκόπηση Μεθοδολογίας

Στην Εικόνα 12 παρουσιάζεται σχηματικά ολόκληρη η διαδικασία της αυτόματης μετάφρασης ενός ερωτήματος από μια αρχική γλώσσα στη γλώσσα προορισμού. Η μέθοδος που προτείνεται στην παρούσα εργασία μεταφράζει ερωτήματα από την ελληνική γλώσσα στην αγγλική. Στις ακόλουθες παραγράφους περιγράφεται ο τρόπος με τον οποίο πραγματοποιείται η μετάφραση των ερωτημάτων.

Αρχικά, δημιουργείται με τη βοήθεια του Lucene ένα ευρετήριο στο οποίο προστίθενται όλα τα 5-grams του Google. Πριν γίνει η ευρετηριοποίηση των πενταγράμμων αφαιρούνται από αυτά –με τη βοήθεια μιας αγγλικής stop words λίστας- όλα τα stop words της Αγγλικής γλώσσας, οι λέξεις δηλαδή που δε συμβάλλουν στην κατανόηση του νοήματος που φέρει κάποιο κείμενο. Τέτοιες λέξεις είναι οι προθέσεις (π.χ., from, of, for), τα βοηθητικά ρήματα (π.χ., is, do, have), οι αντωνυμίες (π.χ., he, she, who), τα άρθρα (π.χ., the, a, an), κτλ. Τα stop words αφαιρούνται επίσης και από τα ερωτήματα με τη βοήθεια μιας ελληνικής stop words λίστας. Ένα ευρετήριο κατασκευάζεται, και πάλι με τη χρήση του Lucene, για το λεξικό με το οποίο μεταφράζονται τα ερωτήματα (βλ. Ενότητα 5.2).



Εικόνα 12. Επισκόπηση της μεθοδολογίας

Στη συνέχεια, τα ερωτήματα χωρίζονται στις επιμέρους φράσεις από τις οποίες αποτελούνται και με τη βοήθεια του λεξικού εξάγονται όλες οι πιθανές μεταφράσεις τους. Τα ερωτήματα δε μεταφράζονται λέξη προς λέξη. Στόχος είναι η εύρεση της μεγαλύτερης ακολουθίας λέξεων που μεταφράζεται. Ο αλγόριθμος με τον οποίο μεταφράζεται ένα ερώτημα περιγράφεται στην Εικόνα 13. Έστω ότι θέλουμε να μεταφράσουμε το ερώτημα Q = “a b c d”. Αρχικά ελέγχεται αν στο λεξικό περιέχεται η μετάφραση ολόκληρου του ερωτήματος Q. Αν αυτό συμβαίνει τότε ο αλγόριθμος επιστρέφει τη μετάφραση και η διαδικασία ολοκληρώνεται. Αν το λεξικό δεν περιέχει ολόκληρη τη μετάφραση του ερωτήματος τότε βρίσκουμε όλες τις υπο-ακολουθίες του αρχικού ερωτήματος με πλήθος λέξεων μια λέξη λιγότερη από το ερώτημα Q. Αν

υπάρχει στο λεξικό η μετάφραση μιας από τις υπο-ακολουθίες τότε αυτή επιστρέφεται και το υπόλοιπο ερώτημα μεταφράζεται λέξη προς λέξη. Με τον τρόπο αυτό είναι δυνατή η μετάφραση των φράσεων και των ιδιωματισμών μιας γλώσσας.

```
1.  Algorithm: Query Translation
2.  Input:
3.     query Q ← "w1 w2 … wn"
4.     L: bilingual lexicon
5.  Output:
6.     translation of Q
7.  int i←1
8.  int n← length of Q
9.  Q′ ← Q
10. do
11.    if (Q′ ∈ L)
12.       T′ ← translation of Q′
13.    find the translation of the rest words w such that w ∈ Q and w ∉ Q′
14.    return the translation of the whole query
15.    else
16.    Q′′ ← {all the possible sub-queries of Q′ of size n-i}
17.    for each q ∈ Q′′
18.       Q′ ← q
19.       i++
20.       goto line 11
21.    end for
22. while (i ≤ n)
```

Εικόνα 13. Διαδικασία μετάφραης ενός ερωτήματος

Πριν γίνει η αναζήτηση των μεταφράσεων στο λεξικό οι λέξεις του ερωτήματος υποβάλλονται σε μια διαδικασία αφαίρεσης των καταλήξεων που ονομάζεται stemming. Για τη διαδικασία αυτή χρησιμοποιείται ένας ελληνικός stemmer, ενώ κρίνεται αναγκαία η εφαρμογή της γιατί στο λεξικό δεν υπάρχουν όλες οι πιθανές πτώσεις ενός των ουσιαστικών και των επιθέτων, ούτε όλα τα πιθανά πρόσωπα και οι χρόνοι των ρημάτων.

Τόσο στον αλγόριθμο Naïve Bayes όσο και στα γλωσσολογικά μοντέλα για να επιλέξουμε την κατάλληλη μετάφραση ενός ερωτήματος πρέπει να υπολογίσουμε τις πιθανότητες συνεμφάνισης των λέξεων του ερωτήματος. Ο υπολογισμός των πιθανοτήτων αυτών μπορεί να γίνει με δύο τρόπους. Ένας από τους τρόπους αυτούς είναι να υπολογίσουμε τη συχνότητα συνεμφάνισης δύο όρων (Term Frequency, TF), το πόσες φορές δηλαδή δύο όρων εμφανίζονται στο ίδιο πεντάγραμμο. Μία εναλλακτική λύση είναι να υπολογίσουμε τον αριθμό των πενταγράμμων που περιέχουν τους όρους που μας ενδιαφέρουν (Document Frequncy, DF).

ΚΕΦΑΛΑΙΟ 5

ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο κεφάλαιο αυτό παρουσιάζεται η αξιολόγηση της μεθόδου, η παρουσίαση των αποτελεσμάτων καθώς και τα συμπεράσματα που εξάγονται από αυτά. Στην αξιολόγηση αυτή χρησιμοποιήθηκαν ερωτήματα στην Ελληνική γλώσσα τα οποία μεταφράστηκαν στην Αγγλική. Συγκεκριμένα, στην ενότητα 5.1 και 5.2 γίνεται μια συνοπτική αναφορά στα ερωτήματα και στο λεξικό αντίστοιχα τα οποία χρησιμοποιήθηκαν κατά την αξιολόγηση. Τα αποτελέσματα και τα συμπεράσματα που προκύπτουν από αυτά παρουσιάζονται στις ενότητες 5.3 και 5.4.

5.1 MEDLINE Ερωτήματα

Η μέθοδος αποσαφήνισης της έννοιας των λέξεων που προτείνεται στην παρούσα εργασία εφαρμόστηκε σε 106 ερωτήματα τα οποία δημιουργήθηκαν από τη MEDLINE [17] και παρουσιάζονται στο Παράρτημα Α. Η MEDLINE είναι μια βάση δεδομένων που καλύπτει τους τομείς της ιατρικής (medicine), της περίθαλψης (nursing), της οδοντιατρικής (dentistry), της κτηνιατρικής, του συστήματος υγειονομικής περίθαλψης, και των προκλινικών επιστημών.

5.2 Λεξικό

Για τη μετάφραση των ερωτημάτων χρησιμοποιήθηκε μια λίστα από 38.044 αγγλικές μεταφράσεις λημμάτων. Για να γίνεται πιο γρήγορα η αναζήτηση των μεταφράσεων αυτών δημιουργήθηκε με τη βοήθεια του Lucene ένα ερευτήριο στο οποίο πρστέθηκαν όλα στα στοιχεία της λίστας αυτής.

Συγκεκριμένα, για κάθε μετάφραση ενός λήμματος δημιουργήθηκε ένα αντικείμενο της κλάσης Document με δύο πεδία. Το ένα πεδίο (με όνομα greek) με βάση το οποίο

γίνονται οι αναζητήσεις περιείχε το ελληνικό λήμμα και το άλλο (με όνομα english) την αγγλική μετάφρασή του.

5.3 Άμεση Αξιολόγηση

Τα 106 ερωτήματα της βάσης δεδομένων MEDLINE που χρησιμοποιήθηκαν για την αξιολόγηση της μεθόδου μεταφράστηκαν με τον τρόπο που περιγράφεται στην Ενότητα 4.3. Στον Πίνακα 2 παρουσιάζονται κάποια στατιστικά στοιχεία που αφορούν στα ερωτήματα αυτά.

Πίνακας 2. Στατιστικά στοιχεία ερωτημάτων

Σύνολο ερωτημάτων	106
Μέσος όρος λέξεων (χωρίς τα stop words)	5,19 ~ 5
Μέσος όρος λέξεων που ανήκουν στο context	2,65 ~ 3
Μέσος όρος αμφίσημων λέξεων	2,47 ~ 2
Σύνολο αμφίσημων λέξεων	262

Η απόδοση της μεθόδου αξιολογήθηκε με δύο τρόπους. Στην ενότητα αυτή περιγράφεται η άμεση αξιολόγηση της μεθόδου όπου υπολογίστηκε η απόδοσή της αναφορικά με τον αριθμό των αμφίσημων λέξεων που αποσαφηνίστηκαν σωστά. Για το σκοπό αυτό χρησιμοποιήθηκαν οι σωστές μεταφράσεις (expert translations) των 106 ιατρικών ερωτημάτων[1]. Για τη μέτρηση της απόδοσης χρσηιμοποιήθηκε η μετρική Precision, η οποία ορίζεται ως ο αριθμός των λέξεων-φράσεων που μεταφράστηκαν σωστά προς τον αριθμό όλων των αμφίσημων λέξεων.

Πραγματοποιήθηκαν δύο σύνολα πειραμάτων. Το πρώτο σύνολο πειραμάτων περιλαμβάνει την εκτέλεση των WSD μεθόδων που περιγράφτηκαν στο Κεφάλαιο 3 στους οποίους ο υπολογισμός των πιθανοτήτων γίνεται με τη βοήθεια της συχνότητας των όρων (term frequency), ενώ στο δεύτερο σύνολο ο υπολογισμός των πιθανοτήτων βασίζεται στη συχνότητα των πενταγράμμων (document frequency). Οι τιμές του Precision των μεθόδων σε κάθε διαφορετική εκτέλεση παρουσιάζεται στον Πίνακα 3.

Παρατηρώντας τα αποτελέσματα που συνοψίζονται στον Πίνακα 3 συμπεραίνουνε ότι τα bigram και trigram μοντέλα δεν έχουν μεγάλη διαφορά τόσο με τη χρήση του term frequency όσο και χρησιμοποιώντας την τιμή του document frequency κατά των

[1] Οι μεταφράσεις αυτές πραγματοποιήθηκαν από ειδικό επιστήμονα.

υπολογισμό των πιθανοτήτων. Ο αλγόριθμος Naïve Bayes δίνει μεγαλύτερες τιμές για το Precision και με τους δύο τρόπους υπολογισμού των πιθανοτήτων.

Πίνακας 3. Αποτελέσματα πειραμάτων

document frequency		
Naïve Bayes	Bigram model	Trigram model
68,7%	64,5%	65,64%
term frequency		
Naïve Bayes	Bigram model	Trigram model
72,3%	69,08%	71,3%

Αν και οι τιμές του Precision με τη χρήση του term frequency είναι μεγαλύτερες από τις τιμές που παίρνουμε όταν χρησιμοποιούμε το document frequency το κόστος υπολογισμού των πιθανοτήτων στην πρώτη περίπτωση είναι πολύ μεγάλο και δεν ενδείκνυται η χρησιμοποιήση του term frequency για μια εφαρμογή πραγματικού χρόνου. Και επειδή η ανάκτηση εγγράφων από μια μηχανή αναζήτησης δεδομένου ενός ερωτήματος πρέπει να γίνεται σε πραγματικό χρόνο χωρίς καθυστερήσεις η καλύτερη WSD μέθοδος με βάση τα αποτελέσματα του Πίνακα 3 είναι ο αλγόριθμος Naïve Bayes με τη χρήση του document frequency. Ένας ακόμη λόγος που ξαθιστά τον Naïve Bayes καταλληλότερο είναι το γεγονός ότι εκτελείται πιο γρήγορα από τα γλωσσολογικά μοντέλα, αφού – σε αντίθεση με τα γλωσσολογικά μοντέλα - δε λαμβάνει υπόψι όλες τις πιθανές μεταφράσεις του αρχικού ερωτήματος αλλά στοχεύει στη επιλογή της κατάλληλης μετάφρασης για τις αμφίσημες λέξεις με βάση το context στο οποίο ανήκουν.

5.4 Έμμεση Αξιολόγηση

Εκτός από την άμεση αξιολόγηση της προτεινόμενης μεθόδου που περιγράφτηκε στην ενότητα 5.3, πραγματοποιήθηκε και μια έμμεση αξιολόγηση κατά την οποία το αρχικό ερώτημα και οι μεταφράσεις που προέκυψαν με τη χρήση του Naïve Bayes και των γλωσσολογικών μοντέλων υποβλήθηκαν σε ένα CLIR σύστημα και μετρήθηκε το πόσο σχετικά ήταν τα ανακτηθέντα κείμενα με το εκάστοτε ερώτημα.

Στην έμμεση αξιολόγηση χρησιμοποιήθηκε η μετρική Average Precision, που περιγράφεται στην παράγραφο 5.4.1, για τον υπολογισμό της απόδοσης, ενώ το CLIR σύστημα ανέκτησε δεδομένα από τη βάση OSHUMED (Ενότητα 5.4.2).

5.4.1 Μέση Ακρίβεια (Average Precision)

Στα πειράματα ανάκτησης πληροφοριών, οι δείκτες για τη μέτρηση της αποτελεσματικότητας των συστημάτων ή των μεθόδων είναι σημαντικοί. Η μέση ακρίβεια (Average Precision) [18] είναι ένας δείκτης που χρησιμοποιείται συχνά για την αξιολόγηση της ταξινομημένης (ranked) ανάκτησης εγγράφων σε πειράματα ανάκτησης πληροφοριών.

Έστω ότι N είναι το πλήθος των κειμένων που ανακτώνται από ένα σύστημα ανάκτησης πληροφοριών (εδώ από το CLIR σύστημα) και έστω ότι με h_i συμβολίζουμε το i-στό ανακτηθέν κείμενο. Αντιστοιχίζουμε σε κάθε ένα από τα ανακτηθέντα κείμενα την τιμή 1 ή 0 ανάλογα με το αν το κείμενο είναι σχετικό με το ερώτημα ή όχι αντίστοιχα.

Η ακρίβεια (Precision) των j πρώτων ανακτηθέντων κειμένων είναι ίση με:

$$\mathrm{P}(j) = \sum_{k=1}^{j} rel(k) \Big/ j$$

Η μέση ακρίβεια των N πρώτων ανακτηθέντων κειμένων ορίζεται ως εξής:

$$AveP = \frac{\sum_{r=1}^{N} (P(r) \times rel(r))}{R}$$

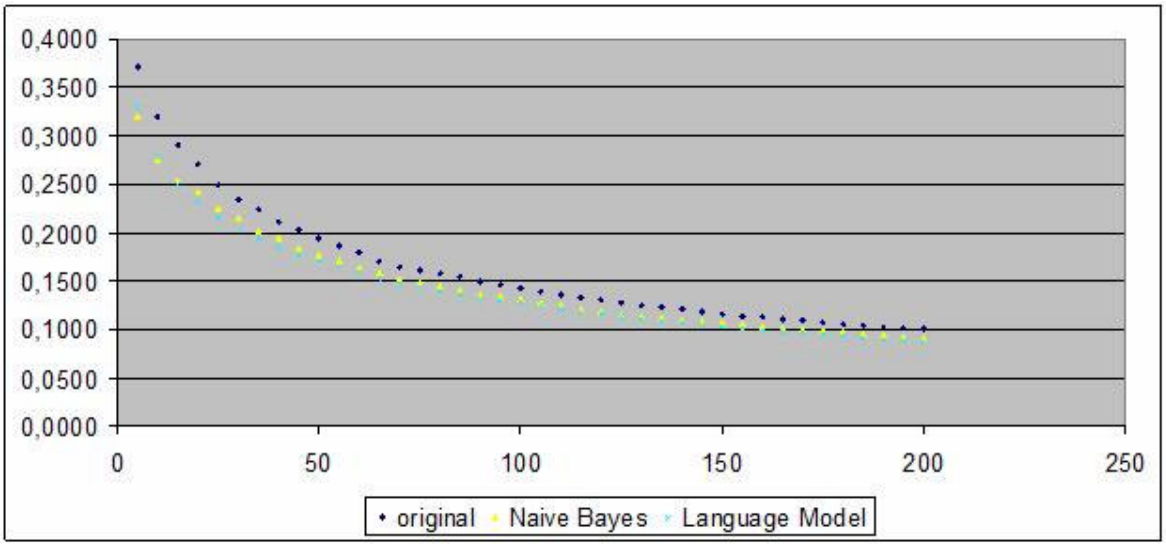
όπου R είναι το πλήθος όλων των σχετικών με το ερώτημα κειμένων στη συλλογή.

Δηλαδή, η μέση ακρίβεια είναι το άθροισμα της ακρίβειας σε κάθε ανακτηθέν κείμενο προς το συνολικό αριθμό των σχετικών εγγράφων στη συλλογή. Η μέση ακρίβεια είναι ένα ιδανικό μέτρο της ποιότητας των μηχανών ανάκτησης. Για να πάρει την τιμή 1, η μηχανή πρέπει να ανακτήσει όλα τα σχετικά έγγραφα (δηλαδή, Recall[2] = 1) και να τα ταξινομήσει τέλεια (δηλ., ακρίβεια P = 1.0).

[2] Το Recall είναι μια μετρική που χρησιμοποιείται στη μέτρηση της απόδοσης των συστημάτων ανάκτησης πληροφοριών και ισούται με τον αριθμό των κειμένων που ανακτήθηκαν και είναι σχετικά με το ερώτημα προς συνολικό αριθμό των σχετικών με το ερώτημα κειμένων της συλλογής.

5.4.2 Βάση Δεδομένων OSHUMED

Η συλλογή OHSUMED [19] περιέχει 348.566 αναφορές, οι οποίες προέρχονται από ένα σύνολο από 270 περιοδικά και αποτελεί υποσύνολο της βάσης MEDLINE. Η συλλογή αυτή περιέχει τις 101 ερωτήσεις που παρήχθησαν από τους πραγματικούς παθολόγους. Κάθε ερώτηση περιέχει μια σύντομη δήλωση για τον ασθενή και την απαιτούμενη πληροφορία. Οι ερωτήσεις είναι γενικά λιτές και περιεκτικές, σε αντίθεση με τις μεγάλες ερωτήσεις που περιέχει το TREC [20]. Στη συλλογή περιέχεται επίσης και ο βαθμός σχετικότητας μεταξύ των ερωτημάτων και των αναφορών. Συγκεκριμένα, ο βαθμός σχετικότητας έχει υπολογιστεί σε δύο επίπεδα: απόλυτα σχετικά (definitely relevant, DR) και απόλυτα ή πιθανόν σχετικά (definitely or possible relevant, D+PR).

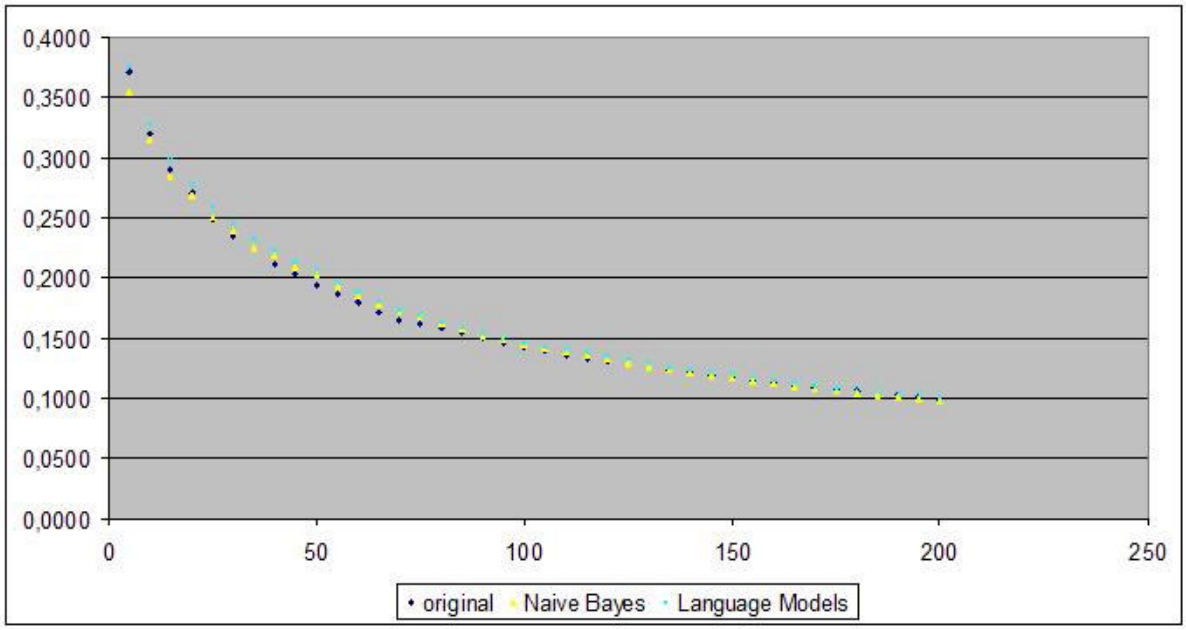


Εικόνα 14. Μέση ακρίβεια ανακτηθέντων κειμένων (term frequency)

5.4.3 Αποτελέσματα Έμμεσης Αξιολόγησης

Τα διαγράμματα των Εικόνων 14 και 15 απεικονίζουν τη μέση ακρίβεια στα k πρώτα ανακτηθέντα κείμενα (οριζόντιος άξονας) από την συλλογή κειμένων OHSUMED με τη χρήση του term frequency και του document frequency αντίστοιχα. Τα μπλε σημεία (original) αντιστοιχούν στη μονογλωσσική ανάκτηση δηλαδή αγγλικά ερωτήματα στην αγγλική βάση. Τα σημεία με κίτρινο και ανοιχτό γαλάζιο χρώμα απεικονίζουν τα αποτελέσματα του Naïve Bayes και του bigram μοντέλου αντίστοιχα.

Παρατηρώτας τα δύο διαγράμματα συμπεραίνουμε ότι η χρήση του term frequency ή του document frequency δεν επηρρεάζει σημαντικά τη μέση ακρίβεια. Ωστόσο, όπως αναφέρθηκε και στην ενότητα 5.3 η χρήση του term frequency δεν ενδείκνυται για ένα σύστημα πραγματικού χρόνου εξαιτίας του υψηλού υπολογιστικού κόστους.



Εικόνα 15. Μέση ακρίβεια ανακτηθέντων κειμένων (document frequency)

ΚΕΦΑΛΑΙΟ 6

ΣΥΜΠΕΡΑΣΜΑΤΑ - ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Ένα από τα κύρια προβλήματα στη βελτίωση της απόδοσης και αποτελεσματκότητας ενός CLIR συστήματος είναι η μείωση της ασάφειας που σχετίζεται με τη μετάφραση του ερωτήματος από την αρχική γλώσσα στη γλώσσα προορισμού. Τα λάθη στις μεταφράσεις οφείλονται κυρίως σε ξένους όρους και στην αποτυχία να μεταφραστούν σωστά οι φράσεις. Επιπρόσθετα, οι πόροι που απαιτούνται για να εξεταστεί το πρόβλημα αυτό απαιτούν συνήθως κάποια χειρωνακτική επεξεργασία και μερικές φορές είναι δύσκολο να αποκτηθούν.

Στην παρούσα εργασία υλοποιήθηκε και αξιολογήθηκε μια μέθοδος αποσαφήνισης της έννοιας των λέξεων ενός ερωτήματος ώστε να επιλεγεί η πιο κατάλληλη μετάφραση ανάλογα με το context στο οποίο ανήκουν οι αμφίσημες λέξεις. Παρά το γεγονός ότι οι πόροι που χρησιμοποιήθηκαν στην προσπάθεια αυτή και συγκεκριμένα τα 5-grams του Google και η λίστα με τις ελληνο-αγγλικές μεταφράσεις λημμάτων περιήχαν αρκετά στοιχεία που έπρεπε να αφαιρεθούν (dummy data) τα πειράματα έδειξαν οτι τα αποτελέσματα είναι πολύ ενθαρρυντικά.

Υπάρχουν αρκετά ανοικτά προς μελέτη θέματα τα οποία θα μπορούσαν να βελτιώσουν ακόμα περισσότερο την απόδοση της όλης διαδικασίας. Ένα από αυτά είναι η χρήση ενός Part of Speech tagger[3] στα μεγάλα ερωτήματα ώστε να περιοριστούν οι πιθανές μεταφράσεις των λέξεων του ερωτήματος.

Η χρήση επίσης ενός συνόλου συνωνύμων για κάθε λέξη του λεξικού θα μπορούσε να βοηθήσει στη μείωση των λέξεων που δε μεταφράζονται εξαιτίας των ελλείψεων του λεξικού που χρησιμοποιείται. Για παράδειγμα, αν θέλουμε να μεταφράσουμε τη λέξη ‘ερυθρό’ και δεν υπάρχει η λέξη αυτή στο λεξικό, γνωρίζοντας ότι οι λέξεις ‘ερυθρό’ και ‘κόκκινο’ είναι συνώνυμες θα μπορούσαμε να αποφύγουμε τη μη μετάφραση της συγκεκριμένης λέξης.

[3] Ένας Part of Speech tagger προσδιορίζει το μέρος του λόγου κάθε λέξης σε μια πρόταση.

Επιπρόσθετα, είναι ενδιαφέρον να μελετηθεί η απόδοση της διαδικασίας με τη χρήση κι άλλων WSD μεθόδων αλλά και με το συνδυασμό του Naïve Bayes και των γλωσσολογικών μοντέλων [21] που μελετήθηκαν στην παρούσα εργασία.

Τέλος, αρκετά σημαντική θα ήταν η αξιολόγηση της μεθόδου με τη χρήση των stop words ώστε να μπορούμε να συμπεράνουμε το αν και πόσο αυτά μπορούν να επηρρεάσουν την απόδοση καθώς και η αντικατάσταση του stemming με αλγόριθμους εύρεσης της ομοιότητας δύο λέξεων (π.χ., ο αλγόριθμος του Levenshtein [22]) ώστε να διαπιστωθεί ποιος από τρόπους αυτούς επηρρεάζει θετικά την απόδοση.

ΠΑΡΑΡΤΗΜΑ Α

Στο Παράρτημα αυτό παρουσιάζονται τα ερωτήματα που χρησιμοποιήθηκαν για την εκτέλεση των πειραμάτων που παρουσιάζονται στο Κεφάλαιο 5.

1. Υπάρχουν ανεπιθύμητες επιδράσεις στα λιπίδια όταν η προγεστερόνη χορηγείται με θεραπεία υποκατάστασης με οιστρογόνα
2. Παθοφυσιολογία και θεραπεία της διάχυτης ενδαγγειακής πήξεως
3. Αντισώματα αντικαρδιολιπίνης και αντιπηκτικό λύκου, παθοφυσιολογία, επιδημιολογία, επιπλοκές
4. Ανασκοπήσεις στην επισκληρίδιο αναισθησία σε υπερήλικες
5. Αποτελεσματικότητα της ετιδρονάτης στη θεραπεία της υπερασβεστιαιμίας σε κακοήθειες
6. Προκαλεί καρκίνο μαστού η θεραπεία υποκατάστασης με οιστρογόνα
7. Ευρήματα στη μαγνητική τομογραφία εγκεφάλου μετά από έκθεση σε τοξικούς παράγοντες
8. Διαγνωστική προσπέλαση της υπέρτασης σε ασθενή με πεταλοειδή νεφρό
9. Τ-λέμφωμα σχετιζόμενο με αυτοάνοσα συμπτώματα
10. Αποτελεσματικότητα της θεραπείας με γάλλιο στην υπερασβεστιαιμία
11. Ανασκόπηση σχετικά με τα έμβολα χοληστερόλης
12. Περιγραφή τραυμάτων που σχετίζονται με θρησκευτικές δραστηριότητες
13. Θεραπεία ανεπάρκειας λακτάσης
14. Πανκυταροπενία στο AIDS, διαγνωστικές εξετάσεις και αιτιολογία
15. Θρομβοκυτάρωση, θεραπεία και διάγνωση
16. Σύνδρομο χρονίας κοπώσεως, αντιμετώπιση και θεραπεία
17. Rh ισοανοσοποίηση, ανασκόπηση άρθρων
18. Ενδοκαρδίτις, διάρκεια αντιμικροβιακής αγωγής
19. Χρήση των β-αναστολέων για θυρεοτοξίκωση κατά την κύηση
20. Συσχέτιση εγκεφαλικής παράλυσης και κατάθλιψης
21. Δευτεροπαθής υπέρταση, πρόσφατη μεθοδολογία για διαγνωστική προσέγγιση
22. Νεώτερες χημειοθεραπείες για εκτεταμένο μεταστατικό καρκίνο μαστού.
23. Αυτόματη ετερόπλευρη γαλακτόρροια, διαφορική διάγνωση και διαγνωστική προσπέλαση
24. Συσχέτιση μεταξύ του Prozac και ηπατικής νόσου
25. Μεμονωμένος υποαλδοστερονισμός, σύνδρομα όπου ο υποαλδοστερονισμός συνυπάρχει με υποκαλιαιμία
26. Παθοφυσιολογία και αιτιολογία του συνδρόμου Stevens-Johnson
27. Δρεπανοκυτταρική αναιμία, θεραπευτικές συμβουλές.
28. Συχνότητα χορήγησης μεβενδαζόλης για την πρόληψη της καχεξίας.
29. Θρομβοπενία στην κύηση, αίτια και αντιμετώπιση.
30. Οξεία σωληναριακή νέκρωση οφειλόμενη σε αμινογλυκοσίδες, σκιαγραφικά, πρόγνωση και θεραπεία.
31. Ανασκόπηση, αντιμετώπιση χρονίου άλγους, χρήση τρικυκλικών αντικαταθλιπτικών

32. Αντιμετώπιση συνδρόμου στέρησης στην κοκαΐνη
33. Πότε πρέπει να γίνεται ενδαρτηρεκτομή καρωτίδων
34. Ανασκόπηση στο σύνδρομο οξείας αναπνευστικής δυσχέρειας των ενηλίκων
35. Προδιαθεσικοί παράγοντες και θεραπεία ηπατοκυτταρικού καρκινώματος.
36. Μπορεί η φαινυτοΐνη ή η φαινοβαρβιτάλη να προκαλέσει αύξηση της γ-γλουταμυλικής τρανσφεράσης ορρού
37. Ινομυαλγία/Συνδετικίτις, διάγνωση και θεραπεία
38. Διαβητική γαστροπάρεση, θεραπεία
39. Ιογενής γαστρεντερίτις, τρέχουσα αντιμετώπιση
40. Αποτελεσματικότερη θεραπεία της κακοήθους αιτιολογίας περικαρδιακής συλλογής στον καρκίνο οισοφάγου.
41. Ασκίτης, διαφορική διάγνωση και διαγνωστική προσπέλαση.
42. Θεραπευτικές επιλογές στον κερατόκωνο.
43. Οσφυαλγία, πληροφορίες για τη διάγνωση και θεραπεία.
44. Μπορεί η ακτινοθεραπεία να προκαλέσει όψιμη εμφάνιση περικαρδιακής συλλογής;
45. Οξεία μεγακαρυοκυτταρική λευχαιμία, θεραπεία και πρόγνωση
46. Μικροσκοπική απώλεια αίματος, ανάγκη για τακτικό έλεγχο
47. Διαφορική διάγνωση επίσχεσης ούρων
48. Ποιες περιφερικές νευροπάθειες σχετίζονται με οιδήματα.
49. Florinef και στεφανιαία νόσος, ενδείξεις
50. Μεμονωμένη συστολική υπέρταση SHEP μελέτη
51. Διαφορική διάγνωση των κυμάτων U
52. Ενδείξεις και επιτυχία της περικαρδιεκτομής και δημιουργίας περικαρδιακού παραθύρου
53. Νεφρίτις λύκου, διάγνωση και αντιμετώπιση.
54. Αναστολείς μετατρεπτικού ενζύμου αγγειοτενσίνης, ανασκόπηση.
55. Πορεία αντιπηκτικής αγωγής με βαρφαρίνη
56. Θεραπεία υποθυρεοειδισμού στη διπολική διαταραχή με ταχεία εναλλαγή
57. Εγκεφαλικό οίδημα δευτεροπαθές σε λοίμωξη, διάγνωση και θεραπεία
58. Διαγνωστική και θεραπευτική προσπέλαση ογκιδίου μαστού
59. Ηπατοχολικές βλάβες σχετιζόμενες με νευροϊνωμάτωση
60. Θεραπεία ενδοκαρδίτιδος με από του στόματος αντιβιοτικά
61. Διασφαγιτιδική σπληνική παράκαμψη, πρόγνωση
62. Εκτίμηση επιπλοκών και αντιμετώπιση της βουλιμίας
63. Θεραπεία της ημικρανίας με β-αναστολείς και αναστολείς ασβεστίου
64. Πρόληψη, παράγοντες κινδύνου, παθοφυσιολογία της υποθερμίας
65. Χρονία φλεγμονώδης απομυελινωτική πολυνευροπάθεια, διαφορική διάγνωση και κριτήρια
66. Επιπλοκές παρατεταμένης χορήγησης προγεστερόνης
67. Παρακολούθηση του διαβήτη στο εξωτερικό ιατρείο, καθιερωμένη αντιμετώπιση του διαβήτη καθώς και νέες τεχνικές αντιμετώπισης
68. Μεσεντέριος αγγειίτις
69. Εκκολπωματίτις, διαφορική διάγνωση και αντιμετώπιση
70. Διαφορική διάγνωση ανεβασμένων επιπέδων αλκαλικής φωσφατάσης και LDH

71. Κυστική ίνωση και νεφρική ανεπάρκεια, επίδραση της επί μακρόν επαναλαμβανόμενης χορήγησης αμινογλυκοσιδών
72. Θυρεοτοξίκωση, διάγνωση και αντιμετώπιση
73. Πυλαία υπέρταση και κιρσοί οισοφάγου, αντιμετώπιση με διασφαγιτιδική ενδοηπατική πυλαιοσυστηματική παράκαμψη
74. Κακοήθες νευροληπτικό σύνδρομο, διαφορική διάγνωση, θεραπεία
75. Καρκινοειδή στο ήπαρ και πάγκρεας, έρευνα, θεραπείες
76. Μετακτινική θυρεοειδίτις, διαφορική διάγνωση, αντιμετώπιση
77. Θερμική εξάντληση, αντιμετώπιση και παθοφυσιολογία
78. Β-αναστολείς και νέγροι με υπέρταση, θεραπευτικές εφαρμογές
79. Επιπλοκές και αντιμετώπιση της ανορεξίας και της βουλιμίας
80. Μάζα επινεφριδίου, διαγνωστική προσπέλαση
81. Ενδοκαρδίτις με αρνητικές καλλιέργειες αίματος, παθογόνοι οργανισμοί, διάγνωση, θεραπεία
82. Άνοια οφειλόμενη σε AIDS, διαγνωστική προσπέλαση
83. Λοιμώξεις σε ασθενείς με μεταμόσχευση νεφρού
84. Χρήσεις θεοφυλλίνης—χρόνιο και οξύ άσθμα
85. Υποτροπιάζουσα κυτταρίτιδα, παράγοντες κινδύνου, αντιμετώπιση, προφύλαξη
86. Κεφαλοσπορίνες και πνευμονικά διηθήματα
87. Σχετίζεται η αυξημένη TSH με σύνδρομο χαμηλής τριιωδοθυρονίνης ορού
88. Καρκίνος πνεύμονα, ακτινοθεραπεία
89. Χειρουργείο έναντι διαδερμικής παροχέτευσης για το απόστημα πνεύμονα
90. Καταμήνιος αναφυλαξία
91. Πλασμαφαίρεση ως θεραπευτική επιλογή για το σύνδρομο Guillain-Barré
92. Σύνδρομο οξείας μεταλοιμώδους πολυνευρίτιδος, ευαισθησία και ειδικότητα της ταχύτητας αγωγής ερεθίσματος
93. Αλλεργική αντίδραση σε βαρφαρίνη, θεραπεία
94. Λοίμωξη ουροποιητικού, κριτήρια για θεραπεία και νοσηλεία
95. Διηθητικές νόσοι λεπτού εντέρου, πληροφορίες σχετικά με λέμφωμα λεπτού εντέρου και νόσου βαρέων α αλύσων
96. Πρωτογενής πρόληψη στον ενήλικα
97. Σιδηροπενική αναιμία, ποια εξέταση είναι η καλύτερη
98. Νόσος Scheuermann, θεραπεία
99. Ορθοσιγμοειδοσκόπηση στην πρόληψη, εάν συνιστώμενη συχνότητα ορθοσιγμοειδοσκόπησης είναι αποτελεσματική και ευαίσθητη στην ανίχνευση του καρκίνου
100. Συσχέτιση νευροληπτικών και περιφερικής νευροπάθειας
101. Οσφυαλγία – ευαισθησία της μαγνητικής τομογραφίας κλπ, συγκρινόμενης με την αξονική τομογραφία οσφυϊκής μοίρας σπονδυλικής στήλης
102. Ποια είναι η καλύτερη αντιμετώπιση του πόνου και της αναπηρίας δευτεροπαθών σε οστεοπόρωση σε πρωτοθεραπευόμενη προχωρημένη νόσο.
103. Διαφορική διάγνωση κολπικής αιμόρροιας απόσυρσης κατά τη διάρκεια αγωγής με οιστρογόνα και προγεστερόνη
104. Χρήση Trental για νευροπάθεια, έχει αποτελέσματα;

105. Ανασκόπηση της αναιμίας χρονίας νόσου
106. HIV και γαστρεντερικό σύστημα, πρόσφατες μελέτες

ΠΑΡΑΡΤΗΜΑ Β

Στο Παράρτημα αυτό παρουσιάζεται ένα απόσπασμα από το XML αρχείο που παράγεται από τη εκτέλεση της προτεινόμενης μεθόδου αποσαφήνισης των εννοιών των αμφίσημων λέξεων και επιλογής της κατάλληλης μετάφρασης. Κάθε στοιχείο που προσδιορίζεται από το tag 'query' περιέχει

- το αρχικό ερώτημα στην ελληνική γλώσσα (tag: gr),
- τη μετάφραση του ερωτήματος αυτού με τις πιθανές μεταφράσεις των αμφίσημων λέξεων μέσα σε άγκιστρα,
- τη μετάφραση του ερωτήματος όπως προκύπτει με την εφαρμογή του αλγορίθμου Naïve Bayes, και
- τη μετάφρση του ερωτήματος όπως προκύπτει από το bigram ή trigram μοντέλο.

```
<query no="2">
  <gr>Παθοφυσιολογία θεραπεία διαχυτης ενδαγγειακής πήξεως</gr>
  <tr>pathophysiology {treatment, therapy} {diffuse, disseminated} intravascular coagulation</tr>
  <nb>pathophysiology therapy diffuse intravascular coagulation</nb>
  <lm>pathophysiology treatment disseminated intravascular coagulation</lm>
</query>
<query no="3">
  <gr>Αντισώματα αντικαρδιολιπίνης αντιπηκτικό λύκου παθοφυσιολογία επιδημιολογία επιπλοκές</gr>
  <tr>antibodies anticardiolipin anticoagulant {lupus, wolf} pathophysiology epidemiology {complication, implication}</tr>
  <nb>antibodies anticardiolipin anticoagulant lupus pathophysiology epidemiology complication</nb>
  <lm>antibodies anticardiolipin anticoagulant lupus pathophysiology epidemiology complication</lm>
</query>
<query no="4">
  <gr>Ανασκοπήσεις επισκληρίδιο αναισθησία υπερήλικες</gr>
  <tr>review {extradural, subdural} {anaesthesia, anesthesia} {very old people, elderly}</tr>
  <nb>review extradural anesthesia elderly</nb>
  <lm>review extradural anaesthesia elderly</lm>
</query>
<query no="5">
  <gr>Αποτελεσματικότητα ετιδρονατης θεραπεία υπερασβεστιαιμίας κακοήθειες</gr>
  <tr>{efficiency, effectiveness} etidronate {treatment, therapy} {hypercalcemia, hypercalcaemia} malignancy</tr>
  <nb>efficiency etidronate treatment hypercalcemia malignancy</nb>
  <lm>effectiveness etidronate therapy hypercalcemia malignancy</lm>
</query>
```

ΑΚΡΩΝΥΜΙΑ

CLIR	Cross Language Information Retrieval
MT	Machine Translation
IR	Information Retrieval
WSD	Word Sense Disambiguation
TF	Term Frequency
DF	Document Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
TF-ICF	Term Frequency – Inverse Category Frequency
WWW	World Wide Web
NLP	Natural Language Processing
DR	Definitely Relevant
D+PR	Definitely or Possible Relevant

ΑΝΑΦΟΡΕΣ

1. Hutchins, W. John; and Harold L. Somers. *An Introduction to Machine Translation*. London: Academic Press, 1992
2. Fuminori Kimura, Akira Maeda, Jun Miyazaki, and Shunsuke Uemura. Query Disambiguation for Cross-Language Information Retrieval Using Web Directories. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI2005)*, pp. 154-159, Tokyo, Japan, Apr. 2005.
3. Yahoo!, www.yahoo.com
4. Ido Dagan and Alon Itai. *Word sense disambiguation using a second language monolingual corpus*. Computational Linguistics, 20(4):563--596, December 1994.
5. Yan Qu, Gregory Grefenstette, David A. Evans: Resolving Translation Ambiguity Using Monolingual Corpora. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002, pp. 223-241
6. Altavista, www.altavista.com
7. Abdelali, A., Cowie, J., Farwell, D., Ogden, W., and Helmreich S., (2003) *Cross-Language Information Retrieval using Ontology*. TALN et multilinguisme, June 2003. Batz-sur-Mer, France
8. TR. Gruber, "A Translation Approach to Portable Ontology Specification", Knowledge Acquisition, 5(2):199-220, 1993
9. Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics, 24*(1), 1-40
10. Eneko Agirre & Philip Edmonds (Eds). Word Sense Disambiguation: Algorithms and Applications, Springer, Vol.33, 2007.
11. Gale, William A., Kenneth W. Church, and David Yarowsky. A method fordisambiguating word senses in a large corpus. *Computers and the Humanities*26:415–439. 1992

12. F. Song and W. B. Croft. *A general language model for information retrieval*. In Proceedings on the 22nd annual international ACM SIGIR conference, pages 279--280, 1999.

13. LDC Catalog, Google n-grams, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

14. Welcome to Lucene, http://lucene.apache.org/

15. Erik Hatcher, and Otis Gospodnetić, Lucene In Action, Manning Publications Co., 2004.

16. The Apache software foundation, http://www.apache.org/

17. MEDLINE, http://www.nlm.nih.gov/databases/databases_medline.html

18. K. Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Nii technical report (nii-2005-014e), NII, 2005Hersh, W., Buckley, C., Leone, T., and Hickman, D. (1994).

19. Oshumed: an interactive retrieval evaluation and new large text collection for research. In Proceedings of SIGIR'94, 17th ACM International Conference on Research and Development in Information Retrieval, pages 192-201

20. Text Retrieval Conference(TREC), http://trec.nist.gov/

21. F. Peng, D. Schuurmans, and S. Wang. Augmenting naive bayes classifiers with statistical language models. Information Retrieval, 7, 317-345, 2004.

22. V. Levenshtein, “Binary codes capable of correcting deletions, insertions and reversals”, Soviet Physics – Doklady 10, 10:707—710, 1966

23. Christopher D. Manning, Hinrich Schutze *Foundations of Statistical Natural Language Processing*, MIT Press (1999).