ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΑΡΧΕΙΟΝΟΜΙΑΣ ΒΙΒΛΙΟΘΗΚΟΝΟΜΙΑΣ
Πρόγραμμα Μεταπτυχιακών Σπουδών στην Επιστήμη της Πληροφορίας
"Διοίκηση & Οργάνωση Βιβλιοθηκών με έμφαση στις Νέες Τεχνολογίες της Πληροφορίας"

ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

Διδάσκων: Καθηγητής Σαράντος Καπιδάκης

Εργασία: **Αυτοματοποιημένη κατηγοριοποίηση κειμένου σε πολυγλωσσικό περιβάλλον Ψηφιακής Βιβλιοθήκης**
The PEKING project

Γεράσιμος Τουρκογιάννης
gtourkog@cc.uoa.gr

Χειμερινό Εξάμηνο: Φεβρουάριος 2004

Περιεχόμενα

Ενότητα Α΄

Εισαγωγή:Αυτόματη κατηγοριοποίηση κειμένων (Automated Text Categorization ATC)

Η αυτοματοποιημένη κατηγοριοποίηση κειμένων έχει γίνει ο στόχος της δημιουργίας λογισμικών εργαλείων που μπορούν να ταξινομήσουν κείμενο ή υπερκείμενα έγγραφα κάτω από προκαθορισμένες κατηγορίες ή θεματικού / ταξινομικούς κώδικες. Έχει παρατηρηθεί ενδιαφέρον για την αναγκαιότητα της αυτοματοποιημένης κατηγοριοποίησης κειμένων λόγω της διαθεσιμότητας μεγάλου αριθμού εγγράφων και εγγράφων σε ψηφιακή μορφή προκειμένου να οργανωθούν και να χρησιμοποιηθούν και εύκολα να ανακτηθούν. Η κατεξοχήν προσέγγιση σήμερα είναι η δημιουργία αυτόματων ταξινομητών κειμένου με την εκμάθηση των χαρακτήρων διαφόρων κατηγοριών μέσα από ένα σύνολο προταξινομημένων εγγράφων. Προηγμένης τεχνολογίας μέθοδοι αναπτύσσοντας την ικανότητα μηχανών για μάθηση (ML) έχουν πρόσφατα εφαρμοστεί γι αυτό το σκοπό που οδηγούν σε εξελιγμένα συστήματα αποτελεσματικότητας σε τρόπο ώστε η ATC να τοποθετείται στο σταυροδρόμι της ανάκτησης πληροφοριών και της ικανότητας μηχανών για μάθηση (ML). Αυτή η πρακτική έχει ενθαρρύνει την εφαρμογή των τεχνικών ATC σε νέους τομείς τής πληροφόρησης όπως είναι η κατηγοριοποίηση ιστοσελίδας μέσα σε καταλόγους ιεραρχικά καθώς επίσης και η κατηγοριοποίηση εγγράφων σε φωνητική μορφή. Έτσι η προοδευτική υιοθέτηση της αυτόματης ή ημιαυτόματης (διαδραστικής) ταξινόμησης από τα διάφορα σχήματα εκεί όπου η χειρωνακτική εργασία / επεξεργασία ήταν κανόνας. Αναμένεται ότι αυτή η πρόοδος στην ATC θα είναι επικερδής και θα δοκιμασθεί η αξιοπιστία αυτών των τεχνολογιών που εφαρμόζονται στην ATC.

Ενότητα Β΄

Αυτόματη πολύγλωσση ευρετηρίαση και ταξινόμηση κειμένων σε ψηφιακές βιβλιοθήκες (CROSS-LINGUAL TEXT CATEGORIZATION)

Τα περισσότερα από τα σημερινά επιστημονικά δημοσιευμένα και τεχνικά άρθρα γράφονται στα αγγλικά. Επομένως, ο αριθμός της συλλογής των εγγράφων στην αγγλική γώσσα που συλλέγονται από τις εμπορικές

εταιρίες που διαχειρίζονται τράπεζες πληροφοριών που είναι και βιβλιογραφικοί παραγωγοί βάσεων δεδομένων ή διαχειρίζονται από βιβλιοθήκες και οι εκδότες αυξάνεται γρήγορα.

Παρόλα αυτά, θα υπάρχουν ακόμα διάφορα έγγραφα που διατίθενται μόνο στη μητρική γλώσσα του συντάκτη, όπως στη γερμανική, γαλλική ισπανική, ιταλική κλπ.

Μια μέθοδος για να διευκολύνει την πρόσβαση σ΄αυτές τις πληροφορίες με αξιόπιστη ανάκληση και ακρίβεια παρέχονται με τις διαδικασίες της «έξυπνης» ευρετηρίασης και ταξινόμησης. Το σύστημα /πρόγραμμα / έργο PEKING προσφέρει τα διάφορα εργαλεία για μονόγλωσση καθώς επίσης και για πολύγλωσση ευρετηρίαση και ταξινόμηση λαμβάνοντας το πλεονέκτημα των περίπλοκων τεχνολογιών επεξεργασίας γλωσσών από ήδη υπάρχοντα ειδικά γλωσσικά βοηθήματα / πηγές όπως οι θησαυροί, τα ταξινομικά σχήματα, τα ειδικά λεξιλόγια κλπ.

Ορισμοί

- η αυτόματη ευρετηρίαση είναι η απλή προέλευση των λέξεων κλειδιών από ένα κείμενο και η παροχή πρόσβασης σε όλες εκείνες τις λέξεις.
- τα πιο σύνθετα αυτόματα συστήματα ευρετηρίασης προσπαθούν να επιλέξουν τους ελεγχόμενους όρους λεξιλογίου (θησαυρός) βασισμένους στους όρους του κειμένου.
- Η αυτόματη ταξινόμηση προσπαθεί να ομαδοποιήσει αυτόματα παρόμοια κείμενα χρησιμοποιώντας είτε : 1. μια πλήρως αυτόματη μέθοδο clustering 2. ένα καθιερωμένο σχήμα ταξινόμησης και ένα σύνολο κειμένων που είναι ήδη ευρετηριασμένα από το σχήμα.
- η αυτοματοποιημένη κατηγοριοποίηση κειμένων είναι η διαδικασία της δημιουργίας εργαλείων λογισμικού ικανών να ταξινομούν τα κείμενα ή τα υπερκείμενα (hypertexts) κάτω από προκαθορισμένες κατηγορίες ή θεματικούς κώδικες
- Clustering είναι η διαδικασία της ομαδοποίησης κειμένων βασισμένων στην ομοιότητα των λέξεων ή των εννοιών των τεκμηρίων όπως ερμηνεύεται από μια αναλυτική μηχανή. Αυτές οι μηχανές χρησιμοποιούν σύνθετους αλγορίθμους όπως Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing), Latent Semantic Analysis, Bayesian statistical analysis και άλλους

Περίληψη

Αυτό το άρθρο ασχολείται με το πρόβλημα της διαγλωσσικής κατηγοριοποίησης κειμένου που εντοπίζεται όταν έγγραφα από διαφορετικές γλώσσες πρέπει να ταξινομηθούν σύμφωνα με το ίδιο ταξινομικό δέντρο. Περιγράφονται εδώ πρακτικές και οικονομικές λύσεις για διαγλωσσική κατηγοριοποίηση εγγράφων που είτε υπάρχει ικανός αριθμός παραδειγμάτων για κάθε καινούργια γλώσσα είτε για κάποιες γλώσσες δεν υπάρχουν καθόλου παραδείγματα. Παρουσιάζονται τα αποτελέσματα έρευνας της δίγλωσσης ταξινόμησης του ILO corpus (σώμα κειμένων) με έγγραφα στα αγγλικά και ισπανικά εφαρμόζοντας δίγλωσση εκπαίδευση , μετάφραση ορολογίας , και μετάφραση κατ' επιλογή.

Εισαγωγή

Η κατηγοριοποίηση κειμένου είναι σημαντική αλλά συνήθως ένα μάλλον «αναξιοποίητο» κομμάτι του τομέα της διαχείρισης των δεδομένων ή πιο γενικά του τομέα διαχείρισης της γνώσης (ΚΜ). Χρησιμοποιείται σε πολλά ιδρύματα, οργανισμούς και επιχειρήσεις παροχής πληροφοριών άλλοτε με τη μορφή της ιεραρχικής απλής ταξινόμησης ή σαν σύνθετη ταξινόμηση χρησιμοποιώντας καθόλου ή περισσότερες λέξεις κλειδιά του κειμένου με σκοπό τη σύνθετη αλλά και την απλή αναζήτηση.

Οι αυτοματοποιημένες τεχνικές κατηγοριοποίησης που βασίζονται σε χειρονακτικά δομημένα προφίλ κατηγοριών έχουν δείξει ότι μπορεί να επιταχυνθεί υψηλή ακρίβεια αλλά το κόστος της διατήρησης αυτού του μοντέλου είναι αρκετά υψηλό. Τα αυτοματοποιημένα συστήματα κατηγοριοποίησης που βασίζονται σε εποπτευόμενη εκμάθηση μπορούν να φτάσουν σε έναν παρόμοιο βαθμό ακρίβειας ώστε η αυτοματοποιημένη κατά το ήμισυ ταξινόμηση των μονόγλωσσων εγγράφων έχει γίνει συνήθης πρακτική. Τώρα το ερώτημα που προκύπτει είναι πως θα αντιμετωπιστούν με επιτυχία συλλογές εγγράφων που είναι σε περισσότερες από μία γλώσσες και οι οποίες θα πρέπει να ταξινομηθούν σύμφωνα με το ίδιο ταξινομικό δέντρο.

Αυτό το άρθρο περιγράφει τις τεχνικές διαγλωσσικής ταξινόμησης που αναπτύχθηκαν με το σύστημα PEKING και παρουσιάζει τα αποτελέσματα που προκύπτουν ταξινομώντας το ILO corpus χρησιμοποιώντας την αυτόματη ταξινόμηση με το LCS (Γλωσσολογικό Ταξινομικό Σχήμα).

Στα επόμενα δύο κομμάτια σχετίζουμε την έρευνά μας με προηγούμενες έρευνες στην αναζήτηση στην διαγλωσσική αναζήτηση πληροφορίας και περιγράφουμε την πιλοτική προσέγγιση στην εφαρμογή του ILO corpus .

Στη συνέχεια προσδιορίζουμε ένα υπόβαθρο για την μονογλωσσική ταξινόμηση του ILO σώμα κειμένου χρησιμοποιώντας διαφορετικούς αλγορίθμους ταξινόμησης. Στη συνέχεια προτείνονται τρεις διαφορετικές λύσεις για την διαγλωσσική ταξινόμηση υποδεικνύοντας θεαματικά πιο σύντομες (και φυσικά πιο φτηνές) διαδικασίες μετάφρασης. Τέλος περιγράφουμε τους βασικούς μας πειραματισμούς στη διαγλωσσική ταξινόμηση και συγκρίνουμε τα αποτελέσματα.

Προηγούμενη έρευνα

Όταν ξεκινήσαμε την έρευνά μας, αναφέρουν οι συγγραφείς του άρθρου, δεν μπορέσαμε να βρούμε δημοσιεύσεις που να αφορούν στο πεδίο της διαγλωσσικής κατηγοριοποίησης κειμένου όπως αυτή. Από την άλλη μεριά υπάρχει πλούσια βιβλιογραφία που σχετίζεται με το πρόβλημα της αναζήτησης διαγλωσσικής πληροφορίας, (CLIR, Cross-lingual Information Retrieval). Και τα δύο CLIR και CLTC, Cross-lingual Text Categorization βασίζονται σε έναν υπολογισμό των ομοιοτήτων μεταξύ των κειμένων συγκρίνοντας τα κείμενα με τη βοήθεια ερωτήσεων και σχεδιάγραμμα τάξεων. Η πιο σημαντική διαφορά ανάμεσα τους είναι ότι το CLIR βασίζεται σε ερωτήσεις που περιλαμβάνουν μόνο λίγες λέξεις ενώ στο CLTC κάθε τάξη καθορίζεται από ένα αναλυτικό προφίλ (το οποίο μπορεί να αντιμετωπιστεί ως μία συλλογή εγγράφων με μετρημένη βαρύτητα ανάκτησης). Αναπτύσσοντας τεχνικές για το CLTC θα πρέπει να έχουμε υπόψη ό,τι μάθαμε από το CLIR.

Ανάκτηση διαγλωσσικής πληροφορίας (CLIR)

Το CLIR ασχολείται με το πρόβλημα της διαμόρφωσης ερωτήσεων εκ μέρους του χρήστη σε μία γλώσσα με σκοπό την αναζήτηση εγγράφων σε διάφορες άλλες γλώσσες. Δύο προσεγγίσεις έχουν αναπτυχθεί:

1. Τα συστήματα που βασίζονται στην μετάφραση ή μεταφράζουν ερωτήσεις στη γλώσσα ή γλώσσες του εγγράφου ή μεταφράζουν τα έγγραφα στη γλώσσα των ερωτήσεων.
2. Τα «ενδιάμεσα» συστήματα παρουσίασης μεταφέρουν και τις ερωτήσεις και τα έγγραφα σε μία ανεξάρτητη γλωσσική παρουσίαση την οποία φτιάχνουν σε μορφή θησαυρού με οντολογική παρουσίαση ή βασισμένη σε ένα μοντέλο γλωσσικής ανεξαρτησίας με διαστήματα διανυσμάτων (vector space model).

Είναι πολύ σημαντικό να σημειώσουμε ότι όλες οι τρέχουσες / σύγχρονες προσεγγίσεις έχουν κάποια προβλήματα. Η μετάφραση των

εγγράφων σε μια δεδομένη γλώσσα για ένα μεγάλο αριθμό εγγράφων είναι μια μάλλον δαπανηρή προσέγγιση ιδιαίτερα σε χρόνο. Χρησιμοποιώντας την ταξινομία των θησαυρών ή των οντολογιών απαιτείται η δυνατότητα παράλληλης ή συγκριτικής σωμάτων κειμένων και το ίδιο απαιτείται και για τις διαγλωσσικές τεχνικές διανύσματος.

Για να συγκεντρώσεις και να επεξεργαστείς αυτό το υλικό είναι πολύ χρονοβόρο άλλα κυρίως αυτές οι τεχνικές βασίζονται σε στατιστικές προσεγγίσεις που μειώνουν την ακρίβεια όταν δεν υπάρχει επαρκές υλικό.

Η λιγότερο ακριβή προσέγγιση είναι η μετάφραση των ερωτήσεων. Η πιο διαδεδομένη τεχνική μετάφρασης των ερωτήσεων είναι αυτή που πρώτα αναγνωρίζει τις λέξεις του περιεχομένου (απλές λέξεις ή σύνθετες) και μετά προσδίδει σ΄ αυτές όλες τις πιθανές μεταφράσεις. Αυτές οι μεταφράσεις μπορούν να χρησιμοποιηθούν σε απλές μηχανές αναζήτησης μειώνοντας το κόστος ανάπτυξης.

Υπάρχουν έρευνες όπου ερευνάται η επίδραση της ποιότητας της πηγής που μεταφράζεται. Επιπλέον συγκρίνονται τα αποτελέσματα πριν και μετά την μετάφραση, δηλαδή ένα ερώτημα το οποίο αποτελείται από έναν αριθμό λέξεων εμπλουτίζεται πριν ή μετά την μετάφραση ή και τα δύο, χρησιμοποιώντας συναφείς λέξεις από λεξικό ή από κάποιο σώμα κειμένων (corpus).

Αυτή η τεχνική του εμπλουτισμού αποτελεί κάποιας μορφής «ψευτο συναφούς ανταπόκρισης» (pseudo relevance feedback). Σύμφωνα με αυτή την τεχνική χρησιμοποιείται είτε το αρχικό ερώτημα ή η μεταφρασμένη του μορφή και ανακτώνται κείμενα στην ίδια γλώσσα με αυτήν (την γλώσσα της μετάφρασης) και τα 25 πρώτα έγγραφα θεωρούνται σαν επιτυχή παραδείγματα. Από αυτά τα κείμενα φτιάχτηκε μία νέα ομάδα από 60 βασικά ερωτήματα τα οποία εμπεριείχαν και τους αρχικούς όρους. Αυτό προσμετράται ως ένας συνδυασμός της ανάπτυξης των ερωτήσεων και αναθεώρησης της βαρύτητας των όρων.

Η επίπτωση της μείωσης της ποιότητας των γλωσσολογικών πηγών γίνεται σταδιακά. Για το λόγο αυτό είναι αναμενόμενο ότι πρέπει επίσης να γίνεται σταδιακή βελτίωση στις πηγές. Οι αδυναμίες των πηγών μετάφρασης μπορεί να αντισταθμιστεί από την ανάπτυξη των ερωτήσεων.

Σε μία άλλη έρευνα η χρήση του γλωσσολογικού μοντέλου στο IR αναπτύχθηκε σε δίγλωσσο CLIR. Στα μονόγλωσσα IR αυτό το συναφές μοντέλο επιλέγεται παίρνοντας μία ομάδα εγγράφων σχετική με τις ερωτήσεις. Στο CLIR χρειαζόμαστε ένα συναφές μοντέλο τόσο για την αρχική γλώσσα όσο και για την γλώσσα που στοχεύουμε να αναζητήσουμε. Για να επιτύχουμε αποτελέσματα στη γλώσσα του στόχου μας μπορούμε να χρησιμοποιήσουμε ένα παράλληλο σώμα κειμένων ή ένα δίγλωσσο λεξικό δίνοντας πιθανές μεταφράσεις. Σε αυτή την έρευνα υποστηρίζεται έντονα η προσέγγιση του μοντέλου της γλώσσας (language modeling approach) στο IR , παρόλο που χρησιμοποιούνται απλοποιημένα μοντέλα γλώσσας (unigram). Χρησιμοποιώντας εμπλουτισμένες μεταφράσεις (linguistically motivated terms) το αποτέλεσμα θα είναι ακόμα καλύτερο.

Διαγλωσσική κατηγοριοποίηση κειμένου (CLTC)

Ή το CLTC ή η διαγλωσσική ταξινόμηση είναι ένα νέο θέμα έρευνας στο οποίο δεν υπάρχει προηγούμενη βιβλιογραφία. Ωστόσο υπάρχει ένα πρόβλημα το οποίο σταδιακά διογκώνεται στα τμήματα τεκμηρίωσης στους πολυεθνείς ή διεθνείς οργανισμούς οι οποίοι βασίζονται στην αυτόματη ταξινόμηση εγγράφων. Επίσης το πρόβλημα είναι προφανές και στις μηχανές αναζήτησης στο διαδίκτυο, οι οποίες βασίζονται στην ιεραρχική ταξινόμηση των σελίδων του διαδικτύου για να μειώσουν την πολυπλοκότητα της αναζήτησης και να επιτύχουν όσο το δυνατό μεγαλύτερη ακρίβεια. Πώς θα πρέπει να συνδυάσουν την ιεραρχία με την ταξινόμηση των διαφόρων γλωσσών;

Θα πρέπει να καθορίσουμε δύο πρακτικές του CLTC:

Πολύγλωσση εκπαίδευση μηχανής (poly-lingual training)
Ένας ταξινομητής (μία μηχανή ταξινόμησης) έχει εκπαιδευτεί σε χαρακτηρισμένα έγγραφα που είναι γραμμένα σε διαφορετικές γλώσσες (ή η πιθανή χρησιμοποίηση διαφορετικών γλωσσών μέσα στο ίδιο έγγραφο).

Διαγλωσσική εκπαίδευση μηχανής (Cross-lingual training)
Χαρακτηρισμένα έγγραφα που είναι διαθέσιμα μόνο σε μία γλώσσα και που θέλουμε να τα ταξινομήσουμε μαζί με έγγραφα που είναι γραμμένα σε άλλη γλώσσα.

Οι περισσότερες περιπτώσεις συναντώνται μεταξύ των δύο αυτών ακραίων μοντέλων. Οι πειραματισμοί μας θα δείξουν ότι τα ακόλουθα αποτελούν ένα πιθανό σενάριο:

Ένας οργανισμός ο οποίος χρησιμοποιεί ήδη ένα αυτοματοποιημένο σύστημα ταξινόμησης θέλει να επεκτείνει αυτό το σύστημα για να ταξινομεί και έγγραφα σε άλλες γλώσσες. Για να διευκολυνθεί αυτή η επέκταση κάποια έγγραφα σε άλλες γλώσσες εισάγονται στο σύστημα είτε σε μη μεταφρασμένη μορφή αλλά χαρακτηρισμένα από τον χρήστη, είτε σε μεταφρασμένη μορφή χωρίς κανένα άλλο χαρακτηριστικό. Με περιορισμένη επέμβαση από το χρήστη φορτώνονται κάποιες εντολές στο σύστημα ώστε τα έγγραφα σε όλες αυτές τις γλώσσες να μπορούν να ταξινομούνται αυτόματα στην πρωταρχική τους μορφή μέσω ενός απλού πολυγλωσσικού ταξινομητή.

Χρησιμοποιώντας έναν αριθμό πειραμάτων θα πειραματιστούμε σε ένα υποθετικό σενάριο:

Πολύγλωσση εκπαίδευση μηχανής: Ταυτόχρονη εκπαίδευση σε χαρακτηρισμένα έγγραφα στις γλώσσες Α και Β μας επιτρέπει να ταξινομήσουμε και τα Α και τα Β έγγραφα με τον ίδιο ταξινομητή.

Διαγλωσσική εκπαίδευση μηχανής: Ένας μονόγλωσσος εκπαιδευμένος ταξινομητής για τη γλώσσα Α μαζί με τη βοήθεια βασικών μεταφρασμένων όρων της γλώσσας Β μας επιτρέπει να ταξινομήσουμε έγγραφα γραμμένα στη γλώσσα Β.

Τι συμπεραίνουμε από CLIR και CLTC

Στην CLTC για να επιτύχουμε μεταφράσεις πρέπει να χρησιμοποιήσουμε παρόμοιες γλωσσικές πηγές όπως και στην CLIR. Επειδή οι πηγές δεν είναι ικανοποιητικές η αντιστάθμιση γίνεται εφαρμόζοντας επέκταση προ και μετά. Στην CLTC το ρόλο των ερωτήσεων (με των οποίων τα έγγραφα συγκρίνονται) παίζουν οι θεματικές ενότητες (class profiles) οι οποίες συνθέτονται από πολλά έγγραφα, αυτό μπορεί να έχει την ίδια επίπτωση σαν μία διεξοδική ανάπτυξη του εγγράφου ή το προφίλ του εγγράφου με μορφολογικές παραλλαγές και συνώνυμα. Στην ουσία το θεματικό προφίλ μπορεί να αντιμετωπιστεί σαν ένα προσεγγιστικό γλωσσικό μοντέλο (approximative language model – unigram) του εγγράφου σε αυτό το συγκεκριμένο θέμα.

Η πειραματική διαδικασία

Όλα τα πειράματα γίνονται με την έκδοση 2.0 του LCS (Γλωσσολογικό Ταξινομικό σχήμα) που αναπτύχθηκε στο έργο PEKING, το οποίο έθεσε σε ισχύ τους αλγόριθμους Winnow και Rocchio. Πρέπει να συγκριθούν αυτοί οι δύο αλγόριθμοι διότι περιμένουμε να μας δείξουν ποιοτικές δια-

φορές στη συμπεριφορά κάποιων εργασιών. Στον Rocchio το θεματικό προφίλ είναι στοιχειωδώς υπολογισμένο σαν centroid, δηλαδή σε ένα προϋπολογισμένο σύνολο από επεξεργασμένα έγγραφα, όπου ο αλγόριθμος Winnow μέσω των ευριστικών μεθόδων υπολογίζει (Support Vector Machines) έναν ιδανικό γραμμικό διαχωριστή στο διάστημα μεταξύ του θετικού και αρνητικού παραδείγματος. Στα πειράματα χρησιμοποιήσαμε την 25/75 ή την 50/50 μέθοδο διαχωρισμού των δεδομένων για την εκμάθηση και τον πειραματισμό με 12-fold ή 16-fold διασταυρούμενης αξιολόγησης.

Ο στόχος μας είναι να συγκρίνουμε την επίδραση των διαφορετικών παρουσιάσεων των δεδομένων και όχι τον υψηλό βαθμό ακρίβειας..

Παρόλο που το ILO corpus είναι μονό-ταξινομημένο (συγκεκριμένα κάθε έγγραφο ανήκει σε μια θεματική τάξη) επιτρέψαμε στο ταξινομητή να μας δώσει 0-3 ταξινομήσεις για κάθε έγγραφο (το οποίο δίνει μία ένδειξη της πιστότητας στην πολύ-ταξινόμηση). Η πολύ-ταξινόμηση (multiclassification) επιτρέπει μεγαλύτερα περιθώρια λάθους και για αυτό έχει κατά κάποιο τρόπο πιο περιορισμένη πιστότητα από την μονό-ταξινόμηση.

Για κάθε παρουσίαση πρώτα αποφασίζουμε την ιδανικότερη ρύθμιση και επιλογή όρων για το σύνολο των εγγράφων. Η ιδανική αξιολόγηση των παραμέτρων εξαρτάται από το σώμα κειμένων (corpus) και από την παρουσίαση του εγγράφου. Ο συνδυασμός έχει καθοριστική επίδραση στη πιστότητα και χωρίς αυτόν τα αποτελέσματα από διαφορετικά πειράματα είναι δύσκολο να συγκριθούν.

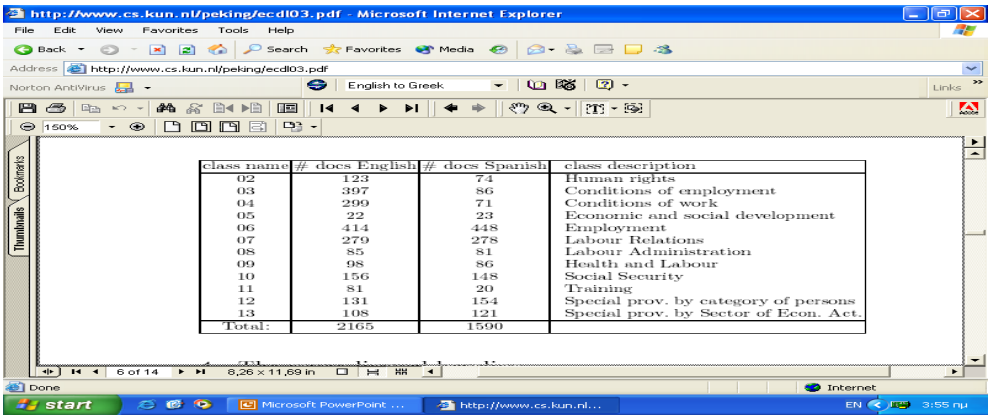
Το σώμα κειμένων ILO

Το ILO corpus είναι μία συλλογή εγγράφων πλήρους κειμένου το κάθε ένα από τα οποία είναι χαρακτηρισμένο με έναν θεματικό όρο (monoclassification) που έχουμε κατεβάσει από την ιστοσελίδα ILOLEX του Διεθνούς Οργανισμού Εργασίας (ILO). Το ILOLEX περιγράφεται ως μία τρίγλωσση βάση δεδομένων. Οι γλώσσες που εμπεριέχονται είναι τα Αγγλικά, τα Ισπανικά και τα Γαλλικά. Από το ILOLEX αποσπάσαμε ένα δίγλωσσο corpus (μόνο Αγγλικά και Ισπανικά) εγγράφων θεματικά επεξεργασμένα για ταξινόμηση. Παρόλο που στην αρχική βάση δεδομένων κάθε έγγραφο έχει και τη μετάφρασή του εμείς για να φτιάξουμε το δικό μας corpus εγγράφων διαλέξαμε ένα χονδρικό αριθμό εγγράφων αποφεύγοντας ομοιότητες σε σχέση με τη γλώσσα, το οποίο σημαίνει ότι διαλέξαμε έγγραφα στα αγγλικά και στα ισπανικά και μερικά μόνο σε μία γλώσσα. Εδώ παραθέτουμε κάποια στατιστικά του ILO corpus:

1. Η αγγλική έκδοση συνίσταται από 2165 έγγραφα. Ο μέσος όρος κάθε εγγράφου είναι 1942 λέξεις. Και το μήκος κάθε εγγράφου ποικίλει από 39 έως 38.646 λέξεις.
2. Η ισπανική έκδοση συνίσταται από 1590 έγγραφα, μήκος εγγράφου από 117 έως 7500 λέξεις και τα περισσότερα έγγραφα έχουν περίπου 2000 λέξεις.

Το σώμα των κειμένων είναι ταξινομημένο θεματικά σε 12 κατηγορίες ταξινόμησης (κάθε έγγραφο ανήκει σε μια μόνο θεματική τάξη και αυτό ονομάζεται μονό-ταξινόμηση).

Ακολουθεί ένας πίνακας με τις κατηγορίες.



class name	# docs English	# docs Spanish	class description
02	123	74	Human rights
03	397	86	Conditions of employment
04	299	71	Conditions of work
05	22	23	Economic and social development
06	414	448	Employment
07	279	278	Labour Relations
08	85	81	Labour Administration
09	98	86	Health and Labour
10	156	148	Social Security
11	81	20	Training
12	131	154	Special prov. by category of persons
13	108	121	Special prov. by Sector of Econ. Act.
Total:	2165	1590	

Η μονόγλωσση βασική γραμμή

Για να τεκμηριώσουμε τις βασικές προϋποθέσεις με τις οποίες θα συγκρίναμε τα αποτελέσματα της διαγλωσσικής ταξινόμησης, πρώτα καταμετρήσαμε την πιστότητα της μονόγλωσσης ταξινόμησης στα Ισπανικά και τα Αγγλικά έγγραφα στο ILO corpus. Επίσης συγκρίναμε τον παραδοσιακό τρόπο παρουσίασης με λέξεις –κλειδιά με εκείνον στο οποίο πολλαπλοί όροι συνδυάζονται σε έναν μονολεκτικό όρο (normalized keywords).

Μονόγλωσσες λέξεις-κλειδιά

Στα αρχικά έγγραφα έχει γίνει κάποια προεργασία όπως : μετατροπή των κεφαλαίων γραμμάτων σε πεζά (de-capitalization), κατάτμηση της πρότασης σε λέξεις και εξάλειψη κάποιων ειδικών χαρακτήρων. Συγκεκριμένα δεν έχει γίνει λημματοποίηση (lemmatization).

Λημματοποιημένες λέξεις-κλειδιά (Lemmatized keywords)

Εδώ γίνεται επεξεργασία των ουσιαστικών και των ρημάτων με τη μορφή του λήμματος (ρίζα κάθε λέξης) στα έγγραφα. Με τον υπολογισμό των αλγορίθμων στις δυο γλώσσες (Αγγλικά και Ισπανικά) στην κατηγοριοποίηση του κειμένου η λημματοποίηση των όρων δεν βελτιώνει την ακρίβεια στην ανάκτηση αν και η λημματοποίηση ενισχύει την ανάκληση των όρων και επηρεάζει αρνητικά την ακρίβεια.

Επεξεργασία φυσικής γλώσσας (NLP) στο κείμενο-πηγή που χαρακτηρίζεται γραμματικώς και λημματοποιείται, πώς;

- Μορφολογική ανάλυση (τύπος του όρου, ρήμα κλπ)
- Λεξική ανάλυση (τι μέρος του λόγου, γραμματικός χαρακτηριστής - tagger)
- Συντακτική ανάλυση
- Φρασεολογική ανάλυση
- Σημασιολογική ανάλυση
- Πραγματολογική ανάλυση

Γλωσσολογικά σύνθετοι / περιφραστικοί όρων (Linguistically motivated terms)

Η χρησιμοποίηση των n-grams (αυτόματη σύνθεση των γραμμάτων μιας λέξης ανά δυο ή τρία από τον αλγόριθμο) αντί για μια λέξη σαν όρο (unigrams) υποστηρίχθηκε για την αυτοματοποιημένη ταξινόμηση κειμένων. Πειράματα σε έρευνα όπου μόνο στατιστικά συναφή n-grams χρησιμοποιήθηκαν δεν έδωσαν καλύτερα αποτελέσματα από τη χρήση μοναδικών λέξεων κλειδιών. Για το πείραμά μας στην CLTC η εξαγωγή των πολύγλωσσων όρων απαιτήθηκε για να μπορέσουμε να βρούμε την καλύτερη μεταφραστική συστοιχία. πχ. Trade union στα Αγγλικά έναντι sindicato στα Ισπανικά.

Στη συνέχεια για τα μονόγλωσσα πειράματα θέλαμε να πειραματιστούμε εάν υπήρχε και μέχρι ποίου σημείου κάποια βελτίωση από τη χρήση γλωσσολογικά σύνθετων / περιφραστικών πολύ-λεξικών όρων σε σχέση με στατιστικά σύνθετους όρους. Για τα Ισπανικά αποσπάσαμε αυτούς του γλωσσολογικά σύνθετους όρους (LMT) χρησιμοποιώντας και ποσοτικές και γλωσσολογικές στρατηγικές:

-Φτιάχτηκε μία αρχική λίστα υποψηφίων όρων χρησιμοποιώντας MILR (Mutual Information and Likelihood Ratio) αντί για το υπάρχον corpus.
-Η αρχική αυτή λίστα των υποψηφίων όρων βελτιώθηκε αντιπαραθέτοντάς την με μία λίστα επιθετικών όρων που ακολουθούσε τη μορφή (ουσιαστικό με ουσιαστικό, ουσιαστικό με επίθετο και ρήμα με πρόθεση και ρήμα)

Αυτή η διαδικασία εξασφάλισε ότι όλες οι Ισπανικές σύνθετες λέξεις ήταν και γλωσσολογικά και στατιστικά ανεπτυγμένες.

Σε αναζήτηση καλύτερων όρων θα χρησιμοποιούμε τον όρο normalized για κείμενο στο οποίο σημαντικές σύνθετες εκφράσεις συνοψίστηκαν σε ένα όρο πχ. Software_engineering .

Ο κατάλογος με τις σύνθετες αγγλικές εκφράσεις δημιουργήθηκε από τις σύνθετες λέξεις που υπάρχουν στη δίγλωσση βάση δεδομένων όπως αυτό προκύπτει από την μετάφραση των Ισπανικών όρων και το LMT.

Για τα αγγλικά το normalization δεν είχε καμία βελτίωση, για τον Rocchio στα Ισπανικά υπήρξε μία μικρή βελτίωση. Για τον Winnow δεν υπήρξε καθόλου βελτίωση.

Ακόμα και όταν χρησιμοποιήθηκαν γλωσσολογικές φράσεις αντί για στατιστικές φράσεις δεν υπήρξε καμία σημαντική βελτίωση στην αυτοματοποιημένη ταξινόμηση.

Διαγλωσσική εκπαίδευση και δοκιμή

Για τη διαγλωσσική κατηγοριοποίηση κειμένου 3 μεταφραστικές στρατηγικές επελέγησαν. Οι 2 πρώτες είναι οικείες με την CLIR:

Μετάφραση εγγράφου : Αν και η μετάφραση του πλήρους κειμένου είναι δυνατή, δεν είναι δημοφιλής στο CLIR γιατί αυτόματη μετάφραση δεν είναι ικανοποιητική και η μη αυτόματη μετάφραση είναι υψηλού κόστους.
Μετάφραση ορολογίας : φτιάχνοντας ορολογία για κάθε τομέα και μεταφράζοντας όλους τους όρους για κάθε τομέα. Αναμένεται ότι αυτό θα εμπεριέχει όλους ή τους περισσότερους όρους που είναι συναφείς με την ταξινόμηση.
Μετάφραση κατ επιλογή : μεταφράζονται μόνο όροι που πραγματικά παρουσιάζονται στην κατηγορία αυτή (οι πιο σημαντικοί όροι)

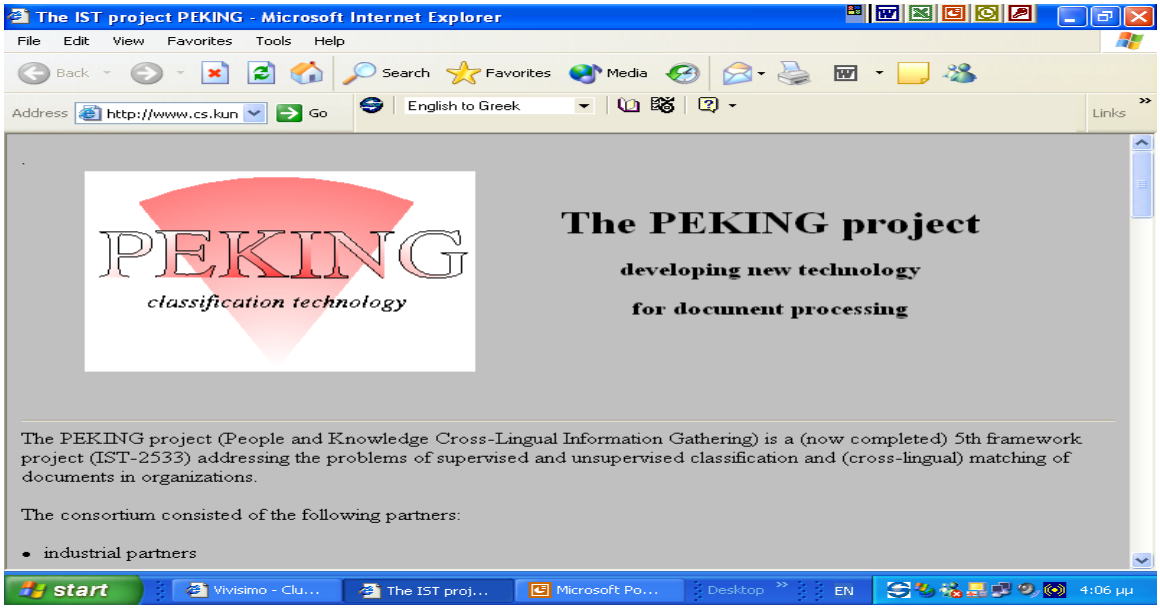
Η μετάφραση όλου του εγγράφου (αυτόματα ή μη) δεν έχει αξιολογηθεί από εμάς επειδή χρειάζεται περισσότερο κόπο από την άλλες μεθόδους αναλογικά και δεν υπόσχεται καλύτερα αποτελέσματα.

Οι γλωσσολογικές πηγές

Ξέρουμε από τις εφαρμογές του CLIR ότι τα υπάρχοντα αποθέματα λέξεων δεν είναι αρκετά. Για να μεγαλώσουμε το μεταφραστικό φάσμα είναι δυνατόν να αποσπάσουμε ομοειδείς μεταφράσεις από παρεμφερείς μεταφραστικές συλλογές, όμως και οι δυο προσεγγίσεις δείχνουν κάποιες αδυναμίες. Ενώ τα δίγλωσσα λεξικά και οι ιδιωματισμοί μας δίνουν αξιόπιστες πληροφορίες μας προτείνουν περισσότερους από έναν μεταφρασμένους όρους ανά δοθέντος όρου χωρίς δυνατότητα επιλογής.
Οι μεταφραστικές συλλογές που χρησιμοποιούνται για να μεταφράσουν σύγχρονους όρους όπως τεχνικούς όρους, προσφέρουν μεταφράσεις συμφραζομένων, αλλά τα λάθη που παρουσιάζονται στατιστικά αναλύοντας τα κείμενα είναι αξιοσημείωτα. Οι μεταφραστικές πηγές μας δημιουργήθηκαν με την προσέγγιση της παραγωγής ενός σώματος λέξεων που είχαν κριτήριο συχνότητας να εμπεριέχουν ουσιαστικά, επίθετα και ρήματα με μια συχνότητα επανάληψης των 30 εμφανίσεων στο δίγλωσσο σύστημα λέξεων. Τα αποτελέσματα μάς έδωσαν 4462 τύπους λέξεων από το σύνολο των 4.619.681 λέξεων για τα ισπανικά και 5258 από 4.609.670 λέξεων για τα αγγλικά.

Συμπεράσματα

Η CLTC είναι στην πραγματικότητα ευκολότερη από την CLIR για τον ίδιο λόγο που η λημματοποίηση των λέξεων και η επεξεργασία όρου έχουν μικρότερη επίπτωση στην CLTC απ΄ ότι στην CLIR. Δίνοντας έναν τεράστιο αριθμό από επεξεργασμένα έγγραφα ο στατιστικός αλγόριθμος ταξινόμησης λειτουργεί σωστά ακόμα και με την απουσία συνδυασμένων όρων που σημαίνει ότι ισοδυναμεί με το CLTC σύστημα που ωστόσο αποτελεί ανάπτυξη στο σύστημα CLIR. Δεν χρειάζεται μεγάλη επεξεργασία για να επιβεβαιώσουμε ότι όλοι οι γλωσσολογικά συναφείς όροι ή τα συνώνυμά τους είναι συμπτυγμένα γλωσσικά. Αν δυο πανομοιότητες μορφές μιας λέξης παρουσιάζονται αρκετά συχνά ώστε να έχουν επίπτωση στην ταξινόμηση θα έχουν επίσης επίπτωση και σαν ανεξάρτητοι όροι. Έχουμε βρει εφικτές λύσεις για 2 ακραίες περιπτώσεις στην CLTC, μεταξύ των οποίων όλες οι πρακτικές περιπτώσεις μπορούν να τοποθετηθούν. Από την μια πλευρά έχει διαπιστωθεί ότι η πολύγλωσση εκπαίδευση δηλαδή εκπαιδεύοντας μόνο ένα ταξινομητή ώστε να ταξινομεί έγγραφα σε έναν αριθμό γλωσσών είναι η πιο απλή προσέγγιση στην CLTC. Από την άλλη πλευρά όταν για παράδειγμα σε μια καινούργια γλώσσα δεν υπάρχουν χαρακτηρισμένα “εκπαιδευμένα” έγγραφα είναι δυνατόν να χρησιμοποιηθεί μετάφραση ορολογίας.



Το πρόγραμμα PEKING στο διαδίκτυο

Κύριος στόχος του προγράμματος PEKING είναι να αναπτυχθεί ένα σύνολο νέων λειτουργιών για να διευκολύνει και να βελτιώσει την αλληλεπίδραση μεταξύ των συστημάτων διαχείρισης γνώσης (Δ. Γ.) και των χρηστών τους.

Σε ένα σύστημα Δ.Γ. μπορούμε να εξετάσουμε 3 βασικά στοιχεία :

1. Τους **συντελεστές / χρήστες** που παράγουν και χρησιμοποιούν εκείνη τη γνώση.
2. Τον **οργανισμό** στον οποίον οι συντελεστές / χρήστες ανήκουν και όπου παράγεται, χρησιμοποιείται και αποθηκεύεται ένα μεγάλο μέρος της γνώσης.
3. Την καθεαυτό **γνώση**.

Από την οπτική γωνία των συντελεστών / χρηστών διαφορετικές λειτουργίες θα αναπτυχθούν επιτρέποντας στο σύστημα να αναγνωρίσει καλύτερα τους χρήστες τους με τις προτιμήσεις, τα ενδιαφέροντα και τη γνώση. Επιπροσθέτως προβλέπεται ο χρήστης να μπορεί να στείλει τις ερωτήσεις του στο σύστημα στη μητρική του γλώσσα και να είναι σε θέση να λάβει τις πληροφορίες σε οποιαδήποτε άλλη γλώσσα. (διαγλωσσικότητα).

Από τη μεριά του οργανισμού το πρόγραμμα PEKING θα αναπτύξει μια ενότητα των δένδρων στην περιοχή της γνώσης του οργανισμού (δυναμικές οντολογίες).

Από την άποψη της γνώσης το πρόγραμμα PEKING θα αναπτύξει τις λειτουργίες που επιτρέπουν στο σύστημα την αυτόματη και συστηματική κατηγοριοποίηση των κατανεμημένων πληροφοριών.

Επομένως θα επιτευχθούν τα παρακάτω πλεονεκτήματα:

1. **Βελτίωση στη δημιουργία γνώσης**. Ως αποτέλεσμα της αυτόματης κατηγοριοποίησης της διαδικασίας εισαγωγής δεδομένων στο σύστημα , από την πλευρά των χρηστών θα είναι γρηγορότερη και αποδοτικότερη.
2. **Βελτίωση στην αναζήτηση και τη διάχυση της γνώσης.** Ως αποτέλεσμα της γνώσης που έχει αποκτήσει το σύστημα από τους χρήστες, η διαγλωσσική και η συνεχής αναπροσαρμογή των δυνατοτήτων του και στα δένδρα στην περιοχή της γνώσης, η αναζήτηση και η διάχυση της γνώσης θα είναι αποδοτικότερες.

Το πρόγραμμα PEKING χρησιμοποιεί αλληλένδετες τεχνικές που επιτυγχάνουν την αυτόματη εξαγωγή και συστηματική κατηγοριοποίηση των διαγλωσσικών πληροφοριών. Αυτή η διαδικασία χρησιμοποιείται και στη Δ.Γ. αλλά και ως ενδιάμεσος για τη διαθέσιμη γνώση.

Η γνώση που ενσωματώνεται σε μια επιχείρηση ή οργανισμό χαρακτηρίζεται σ΄ αυτό το πρόγραμμα από μια ιεραρχική οντότητα που ονομάζεται δένδρο περιοχής της γνώσης (Knowledge Area Tree KAT)

Ενότητα Δ΄

Γλωσσική τεχνολογία στην Ελλάδα : οργανισμοί και έργα

Ελληνικοί οργανισμοί που συμμετέχουν στην Γλωσσική Τεχνολογία με έργα

- Ινστιτούτο Επεξεργασία του Λόγου (οικΟΝΟΜίΑ, METIS, UNL, EuroMAT)
- Γενική Γραμματεία Έρευνας και Τεχνολογίας
- ΕΚΕΦΕ «Δημόκριτος» (Σχηματοποίηση, Ellogon, ΜΙΤΟΣ, Greek Information Extraction GUI)

Επεξεργασία του γραπτού λόγου με τεχνολογίες όπως:

- Γλωσσολογικές μέθοδοι για την Ακρίβεια στην Ανάκτηση Πληροφοριών (IR) και στην Εξαγωγή Πληροφορίας (Information Extraction)
- Υπολογιστική γλωσσολογία (Computational Linguistics)

- Μηχανική Μετάφραση (Machine Translation)
- Γλωσσική Τεχνολογία (Language Technology)
- Ικανότητα μηχανών για εκμάθηση (Machine Learning Methods)
- Διαδικασία λέξεων διανύσματος (word vector processing)

Περιληπτική αναφορά στα ελληνικά έργα:
Α) ΟικΟΝΟΜιΑ : επιφανειακή κατανόηση κειμένων για αποδοτική δεικτοδότηση και εξαγωγή πληροφοριών από οικονομικά δεδομένα. Στάδια επεξεργασίας είναι τα ακόλουθα: γραμματική κατηγοριοποίηση και λήμμα για κάθε λέξη, ονοματικές οντότητες και κατηγοριοποίηση αυτών, ανάλυση της επιφανειακής συντακτική δομής κάθε πρότασης.
Β) Σχηματοποίηση : Στόχος του έργου είναι η δημιουργία ενός ελεγκτή ύφους της Ελληνικής κατάλληλου για συγγραφή πρωτογενών Ελληνικών τεχνικών κειμένων. Συγκεκριμένα το κείμενο υφίσταται γλωσσική επεξεργασία και προστίθενται γλωσσικές υποσημειώσεις που αφορούν τις λεκτικές μονάδες (tokens) τις προτάσεις και τα γραμματικά χαρακτηριστικά των λεκτικών μονάδων.
Γ) το Έλλογον είναι μια πολύγλωσση γενικής χρήσης πλατφόρμα επεξεργασίας κειμένου που χρησιμοποιεί τεχνικές γλωσσολογικής ανάλυσης με τη βοήθεια της υπολογιστικής γλωσσολογίας με σκοπό να βοηθήσει του ερευνητές.
Δ) Το intarget είναι μια προσωποποιημένη υπηρεσία ηλεκτρονικής αποδελτίωσης που απευθύνεται σε ιδιώτες και εταιρικούς συνδρομητές παρέχοντάς τους καθημερινά ημερήσιες δημοσιοποιημένες ειδήσεις. Αυτές οι ειδήσεις ομαδοποιούνται σε κατηγορίες και υποκατηγορίες με βάση τον τίτλο του άρθρου την πηγή και το όνομα του συγγραφέα. Ο πελάτης ορίζει τα θέματα που τον ενδιαφέρουν και για τα οποία θέλει να ενημερώνεται από το σύστημα.
Ε) Ο ΜΙΤΟΣ είναι σύστημα αναζήτησης και εξόρυξης πληροφορίας με εφαρμογή σε χρηματοοικονομικές ειδήσεις. Οι τεχνικές και τα εργαλεία που χρησιμοποιούνται είναι το φιλτράρισμα πληροφορίας, επεξεργασία φυσικής γλώσσας, εξαγωγή πληροφορίας, μοντελοποίηση χρηστών, εξόρυξη δεδομένων.
ΣΤ) GIE, Greek Information Extraction έργο που αντικείμενο του είναι η αυτοματοποιημένη εξαγωγή επιλεγμένων τύπων πληροφορίας από κείμενα της Ελληνικής γλώσσας. Υπάρχουν τέτοια συστήματα αλλά τα κείμενα που χρησιμοποιούνται είναι κυρίως στα αγγλικά. Η μεθοδολογία είναι η ανάπτυξη ενός θεματικού λεξικού, και η εξαγωγή της πληροφορίας με την καθοδήγηση του θεματικού λεξικού με προσημείωση γραμματικής πληροφορίας και συντακτική ανάλυση της συλλογής κειμένων.
Ζ) Το UNL, Universal Networking Language στοχεύει στην κατασκευή ενός πολύγλωσσου συστήματος μηχανικής μετάφρασης για δικτυακές εφαρμογές.

Ηλεκτρονικές Πηγές

- www.htlcentral.org
- www.interpeking.com
- http://www.iit.demokritos.gr/skel/Ellogon/
- http://www.cs.kun.nl/peking/ecdl03.pdf
- http://www.aifb.uni-karlsruhe.de/WBS/aho/pub/lauserhothoecdl03.pdf
- http://www.ilsp.gr/euromap.html
- http://194.219.21.163/index/ie/index.asp
- informationr.net/ir/2-2/paper13.html