ΨΗΦΙΑΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

Διδάσκων: κ. Σαράντος Καπιδάκης

Εργασία: Πολυγλωσσικότητα και ψηφιακές βιβλιοθήκες: Παρουσίαση της εργασίας "Multilingual Information Retrieval Based on Document Alignment Techniques" των Martin Braschler και Peter Schäuble

Ματθαίος Στρατής

Εθνική Βιβλιοθήκη της Ελλάδος
Τμήμα Καταλόγων
matstrat@tellas.gr

Τίτλος:	**Πολυγλωσσικότητα και ψηφιακές βιβλιοθήκες: Παρουσίαση της εργασίας "Multilingual Information Retrieval Based on Document Alignment Techniques" των Martin Braschler και Peter Schäuble**
Περιγραφή:	Εργασία για το μάθημα Ψηφιακές Βιβλιοθήκες στα πλαίσια του *Προγράμματος Μεταπτυχιακών Σπουδών στην Επιστήμη της Πληροφορίας - Διοίκηση & Οργάνωση Βιβλιοθηκών με έμφαση στις Νέες Τεχνολογίες της Πληροφορίας,* για το χειμερινό εξάμηνο του ακαδημαϊκού έτους 2004.
Θέματα / Λέξεις-κλειδιά:	Πολυγλωσσικές Ψηφιακές Βιβλιοθήκες, Δια-γλωσσική Ανάκτηση Πληροφοριών, CLIR, Ευθυγραμμίσεις, Alignments, Συνάφεια
Δημιουργός:	Ματθαίος Στρατής. Τμήμα Καταλόγων Εθνικής Βιβλιοθήκης της Ελλάδος
Ημερομηνία δημιουργίας:	16-02-2004
Χρόνος έκδοσης:	2004
Χώρα έκδοσης:	GR
Γλώσσα κειμένου:	gre

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΓΕΝΙΚΑ

Η πρόσφατη ταχεία ανάπτυξη των ψηφιακών βιβλιοθηκών, δημιούργησε νέες σημαντικές προκλήσεις όσον αφορά την οργάνωση του περιεχομένου τους, τη διαχείρισή τους και τις διεπαφές με τον χρήστη.

Η ανάπτυξη πολυγλωσσικών ψηφιακών βιβλιοθηκών, δηλαδή ψηφιακών βιβλιοθηκών που περιέχουν τεκμήρια σε περισσότερες από μια γλώσσες, δημιούργησε την ανάγκη για την όσο το δυνατόν μεγαλύτερη ανάκτηση πληροφοριών σε περισσότερες από μία γλώσσες. Βασική επιδίωξη στη συγκεκριμένη περίπτωση είναι ο χρήστης να μπορεί να εντοπίζει και να ανακτά την πληροφορία που τον ενδιαφέρει, ανεξάρτητα από τη γλώσσα στην οποία αυτή είναι γραμμένη καθώς και ανεξάρτητα από τη γλώσσα που χρησιμοποιεί ο ίδιος κατά τη σύνταξη του query.

Η συνολική αυτή δραστηριότητα ονομάζεται δια-γλωσσική ανάκτηση πληροφοριών (Cross-language information retrieval – CLIR)

Η αναγκαιότητα γι αυτό είναι μάλλον προφανής: Αφ' ενός, αυτός που αναζητά την πληροφορία πρέπει να έχει πρόσβαση σε όσο το δυνατόν περισσότερες πηγές, χωρίς τον φραγμό της γλώσσας, αφ' ετέρου και ο παροχέας της πληροφορίας, πρέπει να έχει τη δυνατότητα να κάνει τις εργασίες του και τις ιδέες του διαθέσιμες σε όλους στην δική του προτιμώμενη γλώσσα, γνωρίζοντας οτι αυτό δεν θα περιορίσει την πρόσβαση σ' αυτές.

Οι πολυγλωσσικές ψηφιακές βιβλιοθήκες εξ' άλλου είναι πολύ συνηθισμένες , για παράδειγμα ,σε χώρες στις οποίες δεν χρησιμοποιείται μόνο μια εθνική γλώσσα, σε χώρες οπού και η εθνική και η αγγλική γλώσσα χρησιμοποιούνται από κοινού για επιστημονική ή τεχνική τεκμηρίωση, σε πανευρωπαϊκά ινστιτούτα, σε πολυεθνικές εταιρίες κλπ.

Έτσι, το ζητούμενο σε μια πολυγλωσσική ψηφιακή βιβλιοθήκη, προκειμένου να επιτευχθεί μια επιτυχής δια-γλωσσική ανάκτηση πληροφοριών, είναι η ανάπτυξη μεθόδων και εργαλείων που θα ταιριάζουν επιτυχώς τα queries με τα τεκμήρια, παρακάμπτωντας το φράγμα της γλώσσας.

Υπάρχουν δύο διαφορετικές προσεγγίσεις που μπορούν να ακολουθηθούν κατά την εκτέλεση μιας δια-γλωσσικής ανάκτησης πληροφοριών:

α) Τα τεκμήρια-στόχοι να μεταφράζονται στη γλώσσα αναζήτησης

β) Τα queries μιας αναζήτησης να μεταφράζονται στη γλώσσα των αντίστοιχων τεκμηρίων.

1η ΠΡΟΣΕΓΓΙΣΗ

Η τακτική του να μεταφράζονται τα τεκμήρια-στόχοι στη γλώσσα της αναζήτησης, θα ήταν πολύ βολική για τον χρήστη. Στην πράξη όμως, αποδεικνύεται μη ρεαλιστική γιατί η μετάφραση χιλιάδων τεκμηρίων είναι μια δραστηριότητα δαπανηρή και ταυτόχρονα πολύ απαιτητική όσον αφορά τους πόρους που είναι

απαραίτητο να χρησιμοποιηθούν. Επιπρόσθετα, η μετάφραση θα πρέπει να παργματοποιείται για κάθε πιθανό τεκμήριο και γλώσσα–στόχο ξεχωριστά.

Η αυτόματη μετάφραση (MT) από αντίστοιχα μεταφραστικά προγράμματα δεν μπορεί να προσφέρει ικανοποιητικά αποτελέσματα στην περίπτωση αυτη. Τα προγράμματα αυτά όπως έχει αποδειχθεί, απέχουν πολύ από το να δώσουν αποτελέσματα υψηλής ποιότητας και εκτός από την υψηλή δαπανηρότητα της όλης διαδικασίας όπως είπαμε και πριν, απαιτείται η εκτέλεση πολλών δραστηριοτήτων οι οποίες κρίνονται περιττές εξεταζόμενες υπό το πρίσμα μιας ξεκάθαρης δια-γλωσσικής ανάκτησης πληροφοριών, όπως για παράδειγμα η επιτυχής διαχείριση της σειράς των λέξεων.

Εξ' άλλου, όπως έχει αποδειχθεί, ένα μικρό μόνο ποσοστό των τεκμηρίων μιας συλλογής είναι γενικού ενδιαφέροντος. Είναι λοιπόν αρκετό, να μεταφράζονται μόνο τα τεκμήρια που στην πράξη αποδεικνύεται οτι ενδιαφέρουν τον χρήστη που εκτελεί την αναζήτηση.

2η ΠΡΟΣΕΓΓΙΣΗ

Η προσέγγιση της μετάφρασης των queries, ακολουθεί με τη σειρα της δύο τεχνικές:

1. Τεχνικές knowledge-based (Βασισμένες στη γνώση)

Στην περίπτωση αυτή, κατά την εκτέλεση μια δια-γλωσσικής ανάκτησης πληροφοριών, εφαρμόζονται πολύγλωσσα λεξικά, θησαυροί ή γενικής χρήσης οντολογίες.

1.1. Χρήση λεξικών

Η χρήση λεξικών ακολουθήθηκε κατα τις πρώτες προσπάθειες του ταιριάσματος του query με τα τεκμήρια . Όπως αποδείχθηκε, η μέθοδος αυτή όπου κάθε φράση στο query, αντικαθίσταται από μια λίστα όλων των πιθανών μεταφράσεων, αντιπροσωπεύει μια αποδεκτή αρχική προσέγγιση, παρ' οτι τέτοιες σχετικά απλές μέθοδοι, είναι πολύ λιγότερο αποτελεσματικές συγκρινόμενες με μια μονόγλωσση αναζήτηση. Συγκεκριμένα, η μετάφραση του query με χρήση αυτόματων μηχαναγνώσιμων λεξικών (Automatic Machine Readable Dictionary – MRD), αποδείχθηκε να υστερεί σε απόδοση κατα 40-60%, σε σχέση με μια μονόγλωσση αναζήτηση. Αυτό συμβαίνει κυρίως για τους εξής λόγους: Τα γενικής χρήσης λεξικά δεν περιέχουν συνήθως εξειδικευμένο λεξιλόγιο, υπάρχουν αρκετές «νοθευμένες» μεταφράσεις και συνήθως αποτυγχάνουν να μεταφράσουν όρους αποτελούμενους από σύνθετες λέξεις.

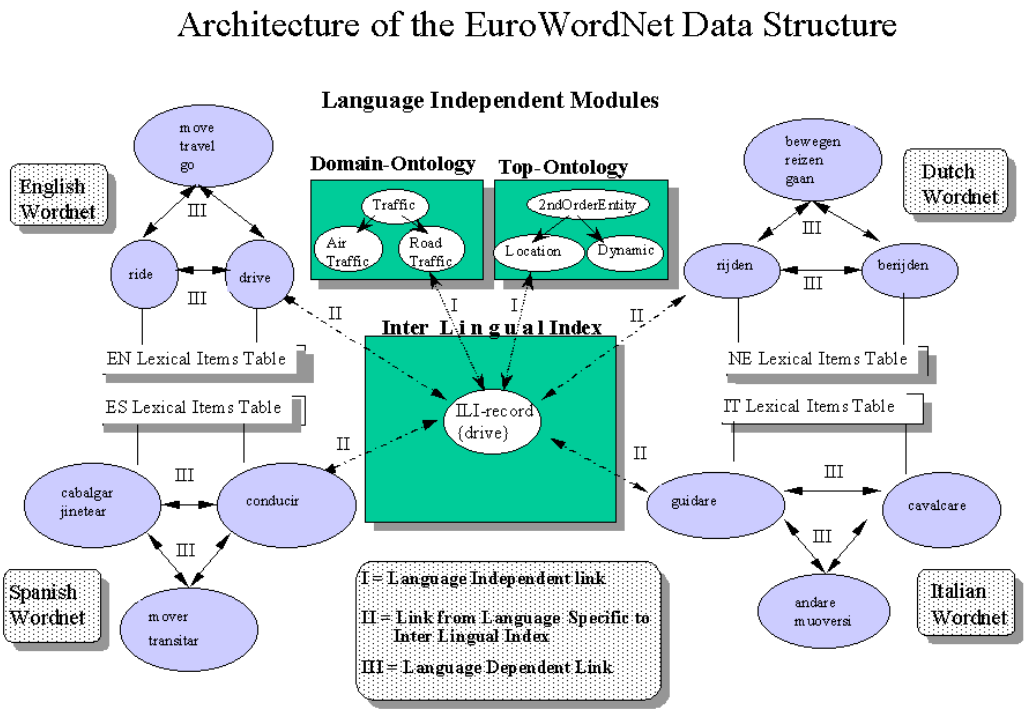
1.2. Χρήση θησαυρών

Οι πιο γνωστές προσεγγίσεις δια-γλωσσικής ανάκτησης πληροφοριών, βασίζονται στη χρήση θησαυρού. Ο θησαυρός είναι μια οντολογία που ειδικεύεται στην οργάνωση ορολογίας. Ένας πολύγλωσσος θησαυρός, οργανώνει την ορολογία για περισσότερες από μία γλώσσες. Παρ' ότι η χρήση πολύγλωσσων θησαυρών φαίνεται να δίνει καλά αποτελέσματα σε μια δια-γλωσσική ανάκτηση πληροφοριών, η δημιουργία και η συντήρηση ενός τέτοιου θησαυρού είναι δαπανηρή και απαιτείται αρκετή εκπαίδευση για την ιδανική χρήση.

1.3. Χρήση οντολογιών

Ένα χαρακτηριστικό παράδειγμα πολύγλωσσης οντολογίας είναι το project Euro WordNet (http://www.illc.uva.nl/EuroWordNet/), η κατασκευή του

οποίου ξεκίνησε το 1996 και ολοκληρώθηκε το 1999. Το Euro Wordnet είναι στην ουσία μια πολυγλωσσική βάση δεδομένων, η οποία για κάθε ξεχωριστή γλώσσα που περιέχει χρησιμοποιεί τα λεγόμενα Wordnets. Για την κατασκευή των Wordnets ακολουθήθηκε η φιλοσοφία του Princeton Wordnet, το οποίο είναι το αμερικανικό Wordnet για την αγγλική γλώσσα. Τα Wordnets είναι επι μέρους οντολογίες, μέσα στις οποίες, τα ουσιαστικά, ρήματα, επίθετα και επιρρήματα κάθε γλώσσας, οργανώνονται σε σετ συνωνύμων, τα λεγόμενα Synsets, τα οποία συνδέονται μεταξύ τους, μέσω των βασικών σημασιολογικών τους σχέσεων. Κάθε Wordnet αντιπροσωπεύει ένα εσωτερικό σύστημα λεξικοποιήσεων για κάθε γλώσσα. Στη συνέχεια τα Wordnets συνδέονται με ένα Δια-γλωσσικό Ευρετήριο (Inter-Lingual Index – ILI), μέσω του οποίου, οι διάφορες γλώσσες διασυνδέονται μεταξύ τους, δίνοντας έτσι τη δυνατότητα μετάβασης από λέξεις μιας γλώσσας, στις αντίστοιχες λέξεις μιας άλλης. Το ευρετήριο αυτό, παρέχει επίσης πρόσβαση σε μια κοινή Κορυφαία Οντολογία (Top Ontology), που αποτελείται από 63 σημασιολογικές διακρίσεις. Αυτή η Κορυφαία Οντολογία παρέχει ένα κοινό σημασιολογικό πλαίσιο για όλες τις γλώσσες που περιέχονται, ενώ οι ξεχωριστές ιδιότητες κάθε γλώσσας διατηρούνται στα αντίστοιχα Wordnets. Η όλη βάση δεδομένων, μπορεί να χρησιμοποιηθεί μαζί με άλλες για δια-γλωσσική ανάκτηση πληροφοριών. Μέχρι στιγμής οι γλώσσες που περιέχονται είναι η αγγλική, γαλλική, γερμανική, ολλανδική, ιταλική, τσεχική και εσθονική). Η αρχιτεκτονική του Euro Wordnet παρουσιάζεται στο ακόλουθο σχήμα:



2. Τεχνικές corpus-based (Βασισμένες στη συλλογή)

Οι corpus-based τεχνικές, βασίζονται στην ίδια τη συλλογή. Χρησιμοποιούν στατιστικά στοιχεία σχετικά με τη χρήση των όρων, προκειμένου να εξάγουν συμπεράσματα όσον αφορά τη σχέση των όρων. Σύμφωνα με την προσέγγιση αυτή, μεγάλες συλλογές κειμένων αναλύονται και απ' την ανάλυση αυτή εξάγονται αυτόματα οι πληροφορίες που χρειάζονται προκειμένου να κατασκευαστούν μεταφραστικές τεχνικές συγκεκριμένης εφαρμογής για την συγκεκριμένη συλλογή. Οι συλλογές που αναλύονται, μπορεί να περιέχουν παράλληλες (μεταφραστικά ισοδύναμες) ή συγκρίσιμες (με σχετικό περιεχόμενο) ομάδες τεκμηρίων. Κατά τον πειραματισμό με τη χρήση τεκμηρίων μιας συλλογής, χρησιμοποιήθηκαν διανυσματικές τεχνικές και τεχνικές πιθανοτήτων.

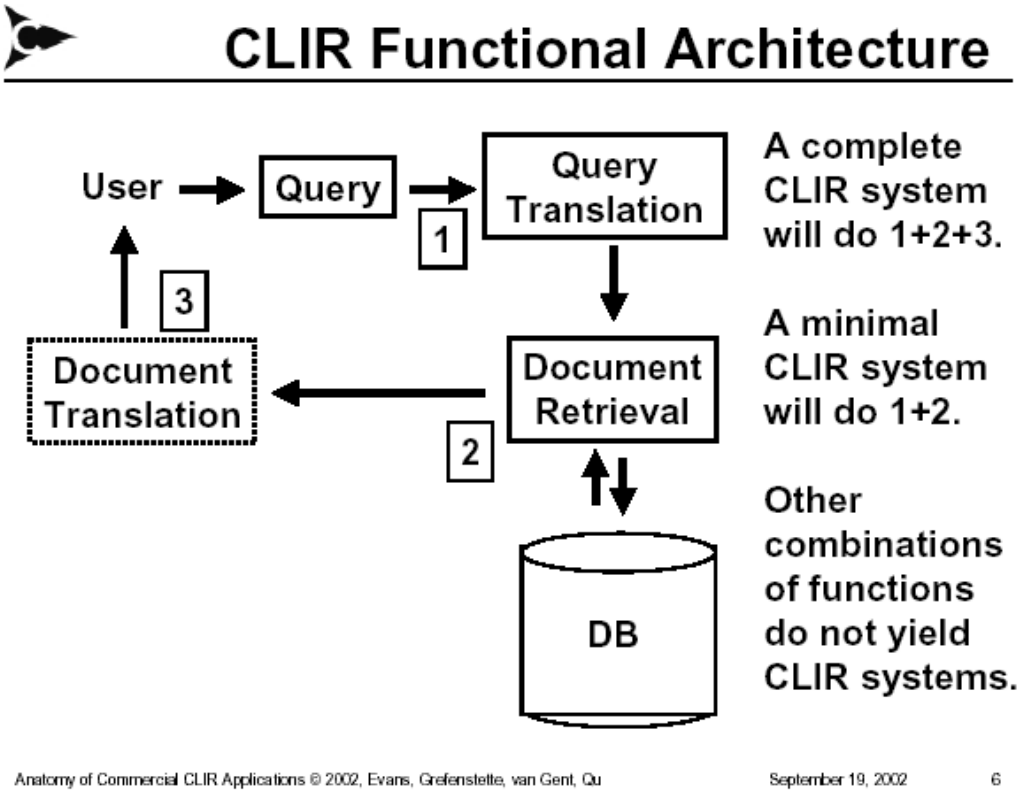
Τα πρώτα τεστ με παράλληλα τεκμήρια, αφορούσαν σε στατιστικές μεθόδους για την εξαγωγή δεδομένων σχετικά με την ισοδυναμία πολύγλωσσων όρων, τα οποία θα μπορούσαν να χρησιμοποιηθούν σαν δεδομένα εισόδου για κάποιο σύστημα αυτόματης μετάφρασης. Η πιο πρόσφατη τεχνική, έχει σαν στόχο την εξαγωγή γλωσσικά ανεξάρτητων όρων καθώς και αναπαραστάσεων τεκμηρίων από παράλληλες ομάδες. Η τεχνική αυτή ονομάζεται Latent Semantic Indexing (LSI) και χρησιμοποιεί μαθηματικές διανυσματικές τεχνικές. Και στη συγκεκριμένη περίπτωση πάντως, το πρόβλημα της σωστής μετάφρασης των όρων παραμένει υπαρκτό. Εξ' αιτίας αυτού, το ενδιαφέρον έχει στραφεί κυρίως στη χρήση συγκρίσιμων ομάδων τεκμηρίων. Μια συλλογή συγκρίσιμων τεκμηρίων είναι εκείνη στην οποία τα τεκμήρια ταξινομούνται και ευθυγραμμίζονται με βάση τη σχετικότητα των θεμάτων που διαπραγματεύονται και όχι με βάση του αν είναι μεταφραστικά ισοδύναμα.

Και πάλι όπως και με τη μέθοδο των παράλληλων συλλογών, αποδεικνύεται οτι η μέθοδος αυτή είναι εξαρτώμενη από την εφαρμογή. Δεν είναι ακόμα πολύ σαφές πόσο καλά μπορεί να εφαρμοστεί στην αναζήτηση μέσα σε μια μεγάλη ετερογενή συλλογή.

Κάνοντας πάντως μια συνολική αποτίμηση, διαπιστώνουμε οτι κάθε ξεχωριστή μέθοδος παρουσιάζει περιορισμούς και έχει ιδιαίτερες απαιτήσεις. Οι υπάρχοντες πόροι όπως τα ηλεκτρονικά δίγλωσσα λεξικά, είναι συνήθως ανεπαρκείς. Η κατασκευή συγκεκριμένων εργαλείων όπως οι θησαυροί ή οι οντολογίες, γενικά δεν είναι επαναχρησιμοποιήσιμοι. Μια νέα πολύγλωσση εφαρμογή θα απαιτήσει τη δημιουργία νέων εργαλείων ή την προσαρμογή των παλαιότερων. Πρέπει επίσης να τονίσουμε οτι πολλά από τα συστήματα και οι μέθοδοι που αναφέρθηκαν, επικεντρώνονται κυρίως σε ζεύγη γλωσσών και όχι σε πολλές γλώσσες ταυτόχρονα. Αυτό συμβαίνει γιατί η κατάσταση περιπλέκεται πολύ περισσότερο όταν επιχειρούμε αναζήτηση μέσα σε πολλές γλώσσες.

Η παρούσα τάση είναι ο πειραματισμός με συνδυασμό πολλών μεθόδων μαζί, π.χ λεξικών, θησαυρών, συγκρίσιμων ομάδων καθώς και διεπαφή με τον χρήστη.

Αρχιτεκτονική λειτουργίας μιας CLIR



Πολύ σημαντικό ζήτημα στην κατασκευή μιας πολυγλωσσικής ψηφιακής βιβλιοθήκης, είναι επίσης και η αναγνώριση, χειρισμός και εμφάνιση των διαφόρων γλωσσών που περιέχονται σ' αυτήν. Οι βασικές απαιτήσεις μιας πολύγλωσσης εφαρμογής είναι να:

1. Υποστηρίζει τα character sets και τις κωδικοποιήσεις που χρησιμοποιούνται για την αναπαράσταση της πληροφορίας που διαχειρίζεται.
2. Παρουσιάζει τα δεδομένα με τρόπο που να έχουν νόημα
3. Χειρίζεται τα πολυγλωσσικά δεδομένα εσωτερικά.

Προκειμένου να επιτευχθεί αφ' ενός η «διεθνικότητα» (internationalization) μιας ψηφιακής βιβλιοθήκης, δηλαδή η δυνατότητα πρόσβασης και χρήσης της ανεξάρτητα από τοπικούς ή γλωσσικούς φραγμούς, όσο και η «τοπικότητα» της (localization) αφ' ετέρου, δηλαδή η προσαρμογή της σε τοπικές ιδιαιτερότητες, είναι απαραίτητη η χρήση των οδηγιών και κανόνων που αφορούν τόσο στο πρωτόκολλο HTTP, όσο και στη γλώσσα HTML. Ειδικότερα, για την κωδικοποίηση των χαρακτήρων, η τελευταία έκδοση της HTML (HTML 4.01), χρησιμοποιεί το σετ χαρακτήρων *Universal Character Set (UCS)* το οποίο περιγράφεται με το ISO 10646. Το σετ αυτό περιλαμβάνει χιλιάδες χαρακτήρες και ουσιαστικά απεικονίζει όλες τις γλώσσες του κόσμου, είναι δε ισοδύναμο χαρακτήρα-προς-χαρακτήρα με το Unicode. Εκτός αυτού, ο

διαχειριστής της πληροφορίας, είναι απαραίτητο να γνωρίζει και τη συγκεκριμένη κωδικοποίηση χαρακτήρων (Character encoding) η οποία έχει χρησιμοποιηθεί για τη μετατροπή των χαρακτήρων ενός τεκμηρίου σε μια ακολουθία από bytes. Η πληροφορία αυτή δίνεται μέσω μιας παραμέτρου στο πεδίο επικεφαλίδας "Content-Type" της HTML. Η παράμετρος αυτή έχει το όνομα "charset". Έτσι, π.χ. για να δείξουμε οτι κάποιο τεκμήριο χρησιμοποιεί το σετ χαρακτήρων greek7, η επικεφαλίδα θα πρέπει να περιέχει τη γραμμή:

Content-type: text/html; csISO88Greek7

Παρ' όλα αυτά, υπάρχει πάντα η πιθανότητα κάποιος χαρακτήρας για κάποια γλώσσα να μην απεικονίζεται με το ISO 10646 εξ' αιτίας της έλλειψης κατάλληλων fonts. Σ' αυτές τις περιπτώσεις η σχετική εφαρμογή πρέπει να εφοδιάζεται με τις κατάλληλες οδηγίες, προκειμένου να επιτευχθεί η localization μιας ψηφιακής βιβλιοθήκης.

Με βάση όσα αναφέραμε πιο πάνω, είναι προφανές οτι μια πολυγλωσσική ψηφιακή βιβλιοθήκη προκειμένου να είναι λειτουργική και φιλική προς τον χρήστη, πρέπει να διαθέτει το ανάλογο πολύγλωσσο interface. Είναι λοιπόν απαραίτητο:

1. Όλα τα επιμέρους interfaces να εμφανίζονται σε κάθε προτιμώμενη γλώσσα
2. Όλα τα μηνύματα να εμφανίζονται σε κάθε προτιμώμενη γλώσσα
3. Όλα τα στοιχεία των επιμέρους πινάκων να εμφανίζονται σε κάθε προτιμώμενη γλώσσα.

Το ιδανικό θα ήταν, το interface να εμφανίζεται σε όλες τις γλώσσες που αντιστοιχούν στα περιεχόμενα τεκμήρια. Κάτι τέτοιο είναι βέβαια δύσκολο να εφαρμοστεί σε μεγάλες ετερογενείς ψηφιακές βιβλιοθήκες, καλό όμως είναι να προσπαθούμε να εμφανίζουμε το interface σε όσες περισσότερες γλώσσες είναι δυνατόν.

Ανακεφαλαιώνοντας όλα όσα είπαμε μέχρι τώρα μπορούμε να δώσουμε έναν διαφορετικό ορισμό μιας πολυγλωσσικής ψηφιακής βιβλιοθήκης:

«Μια πολυγλωσσική ψηφιακή βιβλιοθήκη, είναι μια ψηφιακή βιβλιοθήκη, όλες οι λειτουργίες της οποίας εφαρμόζονται ταυτόχρονα σε όσες γλώσσες είναι επιθυμητό και της οποίας οι λειτουργίες αναζήτησης και ανάκτησης είναι ανεξάρτητες από τη γλώσσα».[1]

[1] Pavani, Ana M. B. , *A model of Multilingual Digital Libray,* Ci. Inf., Brasília, v. 30, n. 3, p. 73-81, set./dez. 2001

2. ΠΑΡΟΥΣΙΑΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ*: "Multilingual Information Retrieval Based on Document Alignment Techniques"*

Η συγκεκριμένη εργασία των Braschler και Schäuble, παρουσιάστηκε στο European Conference for Digital Libraries του 1998.

Παρουσιάζει μια μέθοδο πολυγλωσσικής ανάκτησης πληροφοριών, κατά την οποία, ο χρήστης συνθέτει το query του στην προτιμώμενη από αυτόν γλώσσα, προκειμένου να ανακτήσει σχετιζόμενα τεκμήρια σε οποιαδήποτε από τις γλώσσες που περιέχονται σε μια πολυγλωσσική ψηφιακή συλλογή. Είναι μια επέκταση της κλασσικής δια-γλωσσικής ανάκτησης πληροφοριών (CLIR), κατά την οποία, ο χρήστης μπορεί μεν να ανακτήσει τεκμήρια τα οποία είναι γραμμένα σε διαφορετική γλώσσα απ' αυτή του query, αλλά όπως είπαμε πιο πάνω, οι περισσότερες εφαρμογές της αφορούν κυρίως ζεύγη γλωσσών. Η τεχνική που ακολουθείται εδώ επιχειρεί να πετύχει ανάκτηση τεκμηρίων, γραμμένων σε περισσότερες από δύο γλώσσες.

Είναι corpus-based (βασισμένη στη συλλογή) και χρησιμοποιεί τις λεγόμενες ευθυγραμμίσεις τεκμηρίων (document alignments). Η διαδικασία ευθυγράμμισης, οργανώνει τεκμήρια με σχετικό περιεχόμενο σε ζεύγη, παράγοντας μια ενιαία «χαρτογράφηση» (mapping) των σχετιζόμενων μεταξύ τους τεκμηρίων από διαφορετικές συλλογές.

Στη συγκεκριμένη περίπτωση, χρησιμοποιήθηκαν οι συλλογές δύο ειδησεογραφικών πρακτορείων, του Associated Press (AP), με κείμενα στην αγγλική γλώσσα και του SDA (Ελβετικού Πρακτορείου Ειδήσεων), με κείμενα στη γερμανική και γαλλική γλώσσα.. Η συλλογή των πρακτορείων αυτών περιλαμβάνει κείμενα ειδήσεων.

Ένα παράδειγμα ζεύγους ευθυγραμμισμένων τεκμηρίων, το οποίο προέκυψε από τις ευθυγραμμίσεις γερμανικών και γαλλικών κειμένων του SDA είναι το ακόλουθο:

Example of an alignment pair

Condor-Maschine bei Izmir abgestürzt: Mutmasslich 16 Tote. (Condor plane crashed near Izmir: probably 16 dead)
Un avion ouest-allemand s'écrase près d'Izmir: 16 morts. (A Western German plane crashes near Izmir: 16 dead)

Όπως είναι προφανές, τα δύο αυτά κείμενα, καλύπτουν το ίδιο γεγονός.

Οι ξεχωριστές συλλογές σε διαφορετικές γλώσσες, μαζί με το mapping που προκύπτει απ' τις ευθυγραμμίσεις των σχετιζόμενων τεκμηρίων, συγκροτούν μια

πολυγλωσσική συγκρίσιμη συλλογή.Η συλλογή αυτή μας δίνει τη δυνατότητα να κατασκευάσουμε μια δομή δεδομένων για τη μετάφραση του query. Όπως αναφέραμε νωρίτερα, μια τέτοιου είδους συλλογή περιέχει σχετικά ως προς το περιεχόμενό τους τεκμήρια, ενώ μια παράλληλη συλλογή περιέχει μεταφραστικά ισοδύναμα τεκμήρια. Γενικά, οι παράλληλες συλλογές είναι πολύ σπάνιες, δηλαδή είναι δύσκολο να βρεθούν ξεχωριστές συλλογές που να περιέχουν ακριβώς τα ίδια τεκμήρια σε διαφορετικές όμως γλώσσες, ενώ οι συγκρίσιμες συλλογές, με τεκμήρια σχετικού περιεχομένου σε διαφορετικές γλώσσες, είναι φυσικά πιο διαθέσιμες.

Πολλές σχετικές προσπάθειες πάνω στην ευθυγράμμιση των τεκμηρίων έχουν εκπονηθεί στο παρελθόν. Όμως οι περισσότερες βασίζονταν σε παράλληλες συλλογές και οι ευθυγραμμίσεις γίνονταν σε επίπεδο γλωσσικής πρότασης ή ακόμα και σε επίπεδο λέξης. Έχουν επίσης χρησιμοποιηθεί και Θησαυροί Συσχετίσεων, οι οποίοι βασίζονταν σε συσχέτιση όρου-προς-όρο. Η τεχνική που εξετάζουμε εδώ, βασίζεται στη σύγκριση τεκμηρίου-προς-τεκμήριο.

1. Ευθυγράμμιση τεκμηρίων

Χρήση δεικτών

Οι ευθυγραμμίσεις των τεκμηρίων παράγονται με τη χρήση «δεικτών» προκειμένου να εντοπιστούν οι σχέσεις μεταξύ ζευγών τεκμηρίων από τις συλλογές που αναζητούνται. Ενδείξεις συνάφειας μεταξύ των τεκμηρίων μπορεί να είναι οι ακόλουθες:

- Τα τεκμήρια περιέχουν κοινά κύρια ονόματα (Η ορθογραφία των ονομάτων σε παρόμοιες γλώσσες είναι συνήθως σταθερή)
- Τα τεκμήρια περιέχουν κοινούς αριθμούς (Οι αριθμοί σε μεγάλο βαθμό δεν εξαρτώνται απ' τη γλώσσα)
- Αν στα τεκμήρια έχουν αποδοθεί συμβατοί ταξινομητές (classifiers), αυτοί μπορούν να χρησιμοποιηθούν
- Η ίδια ιστορία ή είδηση συνήθως δημοσιεύεται σε κοντινές ημερομηνίες από τα ειδησεογραφικά πρακτορεία. Κατά συνέπεια, οι ημερομηνίες μπορούν να χρησιμοποιηθούν σαν δείκτες
- Λέξεις που περιέχονται και στα δύο τεκμήρια μπορούν να χρησιμοποιηθούν σαν ένδειξη συνάφειας. Ειδικά γι αυτό, μπορεί να χρησιμοποιηθεί λεξικό για τη μετάφραση των όρων από γλώσσα σε γλώσσα.

Όπως είναι φανερό, μόνο το τελευταίο είδος δεικτών απαιτεί τη χρήση κάποιου γλωσσικού εργαλείου, στη συγκεκριμένη περίπτωση του λεξικού.

Η βασική ιδέα στη διαδικασία ευθυγράμμισης των τεκμηρίων, είναι να χρησιμοποιούνται τα κείμενα της πρώτης συλλογής σαν queries, τα οποία θα «τρέξουν» ως προς τα κείμενα της δεύτερης συλλογής, έχοντας σαν αποτέλεσμα την ανάκτηση των πιο συναφών αντίστοιχών τους. Αυτά τα ζεύγη των συναφών τεκμηρίων, δίνουν ένα mapping μεταξύ των δύο συλλογών.

2. Παραγωγή των ευθυγραμμίσεων

Παραγωγή ευθυγραμμίσεων έγινε αφ' ενός για τη συλλογή αγγλικών κειμένων του AP και γερμανικών κειμένων του SDA (English AP - German

SDA)και αφ' ετέρου για τη συλλογή γερμανικών κειμένων του SDA και γαλλικών κειμένων του SDA (German SDA – French SDA). Παρ' όλο που τα γερμανικά και γαλλικά κείμενα του SDA δεν είναι ακριβείς μεταφράσεις ίδιων κειμένων, η θεματική τους επικάλυψη, είναι αρκετά μεγάλη, κάνοντας έτσι ευκολότερη την ευθυγράμμιση. Επιπλέον, στα κείμενα αυτά του SDA, έχουν αποδοθεί classifiers με το χέρι, οι οποίοι είναι συμβατοί μεταξύ τους. Αυτό απλοποιεί πολύ τη διαδικασία ευθυγράμμισης και επειδή οι δύο αυτές συλλογές έχουν αρκετές ομοιότητες, η ευθυγράμμιση μπορεί να γίνει χωρίς καμμία χρήση γλωσσικών εργαλείων όπως θα δούμε αναλυτικά πιο κάτω.

Από την άλλη πλευρά οι συλλογές των αγγλικών κειμένων του AP και των γερμανικών κειμένων του SDA, είναι πολύ διαφορετικές, πράγμα που σημαίνει ότι στη συγκεκριμένη περίπτωση, κάποιου είδους γλωσσικό εργαλείο είναι απαραίτητο να χρησιμοποιηθεί.

I. Ευθυγράμμιση συλλογών AP – German SDA

Για την ευθυγράμμιση των κειμένων αυτών των συλλογών, χρησιμοποιήθηκαν οι όροι μετρίας συχνότητας. Είναι μια πρακτική αντίστοιχη με τη χρήση των "stopwords" με τη διαφορά ότι εδώ παραλείπονται όχι μόνο οι όροι υψηλής συχνότητας (όπως συμβαίνει με τα stopwords), αλλά και οι όροι χαμηλής συχνότητας. Έτσι το σύνολο των όρων που παραλείπονται τελικά είναι πολύ μεγαλύτερο. Όπως έδειξε η πράξη οι όροι υψηλής συχνότητας που απαντώνται στο 10% του συνόλου των τεκμηρίων, δεν βοηθούν στη διάκριση σχετικών ή μη τεκμηρίων. Οι όροι χαμηλής συχνότητας πάλι, που απαντώνται μόνο σε ένα ή δύο τεκμήρια συνολικά, είναι πιθανό να δώσουν καθαρά τυχαίες ομοιότητες.

Τα επεξεργασμένα τεκμήρια του AP μετατρέπονται σε queries και μεταφράζονται στη γλώσσα-στόχο (στη συγκεκριμένη περίπτωση στα γερμανικά). Όπως είπαμε πιο πριν, λόγω της μεγάλης ανομοιότητας των συλλογών, η χρήση κάποιου λεξικού είναι απαραίτητη. Χρησιμοποιήθηκε μια wordlist (απλοποιημένη μορφή λεξικού), η οποία συντέθηκε από δωρεάν πηγές στο internet, προκειμένου ν' αποφευχθεί η απόκτηση κάποιου ακριβού λεξικού. Οι πηγές αυτές δίνουν πολύ απλοποιημένες λίστες μεταφράσεων, χωρίς να δίνουν ιδιαίτερες πληροφορίες σχετικά με τα μέρη του λόγου, τη συχνότητα των μεταφράσεων, κλπ. Παρ' ότι στις wordlists δεν αποφεύγονται ορθογραφικά και μεταφραστικά λάθη, η πράξη έδειξε οτι η όλη διαδικασία ευθυγράμμισης είναι αρκετά ανεκτική σε τέτοιου είδους λάθη, χωρίς να μειώνεται η αποτελεσματικότητά της.

Για την επίτευξη καλύτερων αποτελεσμάτων, κρίθηκε απαραίτητη η χρήση δύο ακόμα τεχνικών: Του thresholding και της κανονικοποίησης της ημερομηνίας.

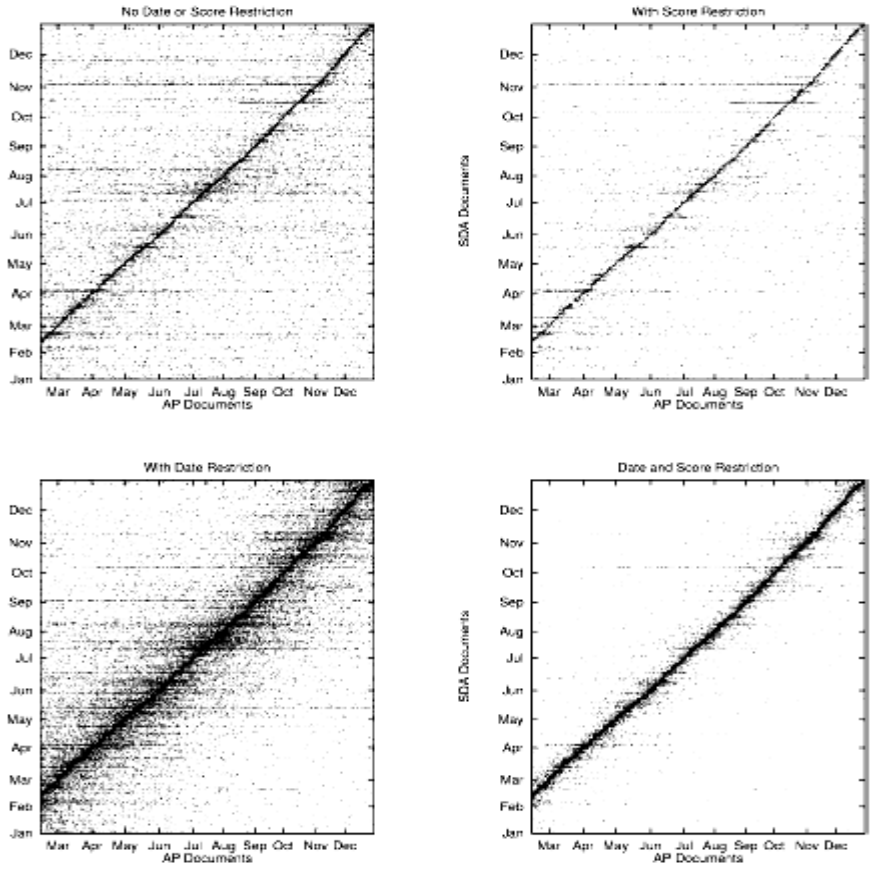
Η τεχνική του thresholding (Τεχνική κατώτερου ορίου ή τεχνική χρήσης κατωφλίου), χρησιμοποείται για να αντιμετωπιστεί η διαφορετικότητα των συλλογών και για να τεθεί ένα φίλτρο στα αποτελέσματα. Επειδή οι συλλογές αυτές είναι διαφορετικές, για αρκετά τεκμήρια,δεν υπάρχουν ικανοποιητικά αντίστοιχά τους, άρα δεν πρέπει να παράγεται κανένα ζεύγος ευθυγραμμισμένων τεκμηρίων. Έτσι, για παράδειγμα, μπορούμε να πούμε ότι αν για ένα query, δεν βρεθούν τουλάχιστον πέντε ευθυγραμμισμένα ζεύγη, θεωρούμε οτι δεν έχει βρεθεί κανένα. (Το κατώτερο όριο σ' αυτό το παράδειγμα είναι ο αριθμός 5). Ζητούμε δηλαδή έναν ελάχιστο αριθμό ευθυγραμμίσεων. Στην πρακτική του thresholding, για τον καθορισμό του αρμόζοντος κατώτατου ορίου λαμβάνεται υπ' όψιν και το μήκος του query. Επειδή εδώ το βάρος δίνεται στη χρήση ενός

κατώτατου ορίου και δεν τίθεται ένα ανώτατο όριο όσον αφορά τα αποτελέσματα, η πρακτική του thresholding παραμένει ένα ζήτημα που απαιτεί περαιτέρω εξέταση και από μόνη της δεν αποτελεί την ιδανική λύση.

Η κανονικοποίηση της ημερομηνίας (date normalization), χρησιμοποιείται επίσης για το φιλτράρισμα των αποτελεσμάτων. Επειδή τα ειδησεογραφικά πρακτορεία, δημοσιεύουν ειδήσεις για σχετικά γεγονότα σε κοντινές ημερομηνίες, η πιθανότητα ενός καλού ταιριάσματος είναι μεγαλύτερη, αν οι ημερομηνίες δημοσίευσης δύο πιθανών σχετικών τεκμηρίων είναι κοντινές. Η τεχνική αυτή αυξάνει την πιθανότητα σωστών ευθυγραμμίσεων, διαιρώντας τον αριθμό των ακατέργαστων αποτελεσμάτων με τον λογάριθμο της απόστασης σε ημέρες μεταξύ δύο κειμένων.

II. Απεικόνιση των ευθυγραμμίσεων για τις συλλογές AP – German SDA

Οι ευθυγραμμίσεις, καθώς και η επίδραση του thresholding και της κανονικοποίησης της ημερομηνίας επάνω τους φαίνονται στους παρακάτω πίνακες: Τα ζεύγη των ευθυγραμμισμένων τεκμηρίων εμφανίζονται σαν κουκίδες στο επίπεδο που καθορίζεται από δύο χρονικούς άξονες. Ο οριζόντιος άξονας αντιπροσωπεύει όλα τα τεκμήρια του AP για το 1988 με ημερολογιακή σειρά, ενώ ο κάθετος άξονας αντιπροσωπεύει τα αντίστοιχα τεκμήρια του SDA επίσης για το 1988.



Το πάνω αριστερά γράφημα απεικονίζει, τις ευθυγραμμίσεις χωρίς την εφαρμογή του thresholding και της κανονικοποίησης ημερομηνίας. Όπως είναι αναμενόμενο, πολλές κουκίδες βρίσκονται μέσα στη διαγώνιο, καθώς οι ημερομηνίες δημοσίευσης των τεκμηρίων που απαρτίζουν το αντίστοιχο ζεύγος , είναι κοντινές. Όμως, πάρα πολλές κουκίδες, βρίσκονται διασπαρμένες μέσα στο επίπεδο. Το πάνω δεξιά γράφημα, απεικονίζει τα αποτελέσματα με την εφαρμογή του thresholding. Εδώ δεν έχει εφαρμοστεί η κανονικοποίηση ημερομηνίας. Παρατηρούμε ότι με την εφαρμογή του thresholding, το οποίο περιορίζει τα αποτελέσματα και φιλτράρει δραστικά τα 2/3 των ζευγών, η διαγώνιος παρουσιάζεται και πάλι πολύ καθαρά. Και τώρα οι κουκίδες που βρίσκονται εκτός διαγωνίου είναι αρκετές και αρκετά διασκορπισμένες, σε αραιά όμως διαστήματα. Το γράφημα κάτω αριστερά απεικονίζει το αποτέλεσμα με την εφαρμογή της κανονικοποίησης ημερομηνίας, χωρίς την εφαρμογή thresholding. Εδώ βλέπουμε ότι οι κουκίδες έχουν συγκεντρωθεί αρκετά κοντά στη διαγώνιο. Το τελευταίο γράφημα απεικονίζει τα αποτελέσματα με την εφαρμογή thresholding και κανονικοποίησης ημερομηνίας. Η διαγώνιος εμφανίζεται με αρκετό πάχος, ενώ οι κουκίδες αραιώνουν όσο απομακρύνονται απ' αυτήν. Αυτό είναι και το επιθυμητό αποτέλεσμα.

Οι παρατηρήσεις αυτές οδηγούν σε μια πιθανή βελτιστοποίηση: Καθώς τα περισσότερα ζεύγη που παράγονται βρίσκονται σε μια στενή λωρίδα γύρω από τη διαγώνιο, η ευθυγράμμιση μπορεί να υπολογιστεί με τη χρήση ενός κυλιόμενου «Παραθύρου ημερομηνίας», το οποίο περιορίζει δραστικά την έκταση του πεδίου έρευνας. Αντί να αναζητούμε για ένα ταίριασμα στο σύνολο των 3 ετών των τεκμηρίων του SDA για κάθε τεκμήριο του AP, μπορούμε να λάβουμε υπ' όψη, μόνο τα τεκμήρια μέσα σ' ένα παράθυρο ημερομηνίας με έκταση λίγων ημερών. Οι αναζητήσεις πραγματοποιούνται πάνω σε μια σειρά αλληλοκαλυπτόμενων μικρών υποσυλλογών , καθώς το παράθυρο ημερομηνίας μετακινείται μέσα στη συλλογή. Τα πειράματα έδειξαν ότι ένα παράθυρο έκτασης 15 ημερών είναι μια αρκετά καλή επιλογή. Για πιο σχετικές συλλογές, η έκταση του παραθύρου, μπορεί να είναι μικρότερη.

Τα ευθυγραμμισμένα ζεύγη που παράγονται είναι μη αναστρέψιμα. Αυτό σημαίνει ότι κάποια τεκμήρια του AP μπορεί να ευθυγραμμιστούν με το ίδιο τεκμήριο του SDA, αλλά όλα τα τεκμήρια του SDA δεν ανήκουν απαραίτητα σε κάποιο ευθυγραμμισμένο ζεύγος. Αν χρειαζόμαστε ευθυγραμμίσεις με την αντίθετη κατεύθυνση, οι ρόλοι των τεκμηρίων του AP και του SDA αντιστρέφονται: Τωρα τα τεκμήρια του SDA μετατρέπονται σε queries, μεταφράζονται στα αγγλικά και τρέχουν πάνω στη συλλογή του AP.

III. Ευθυγράμμιση συλλογών French SDA – German SDA

Όπως είπαμε πιο πάνω, οι δύο αυτές συλλογές είναι αρκετά παραπλήσιες. Στα κείμενα του SDA έχουν αποδοθεί classifiers με το χέρι, οι οποίοι δεν εξαρτώνται από τη γλώσσα. Η ταξινόμηση είναι αρκετά χονδροειδής και περιλαμβάνει πάνω από 300 διαφορετικούς classifiers (κυρίως κωδικούς χωρών). Παρ' όλα αυτά όμως είναι βοηθητική.

Στη συγκεκριμένη περίπτωση, δεν χρησιμοποιήθηκαν wordlists, καθώς τέτοιες δεν ήταν δυνατόν να βρεθούν για γαλλικά-γερμανικά και έτσι οι ευθυγραμμίσεις έγιναν χωρίς τη χρήση οποιουδήποτε λεξικού. Αυτό δε θα ήταν

δυνατόν αν οι συλλογές ήταν πολύ διαφορετικές μεταξύ τους. Δείχνει όμως, ότι η παραγωγή ευθυγραμμίσεων είναι δυνατή, χωρίς τη χρήση γλωσσικών εργαλείων, υπό την προϋπόθεση οι συλλογές να έχουν αρκετές ομοιότητες.

Πέραν της χρήσης classifiers, η ευθυγράμμιση SDA - SDA, χρησιμοποιεί σαν δείκτες, τα κύρια ονόματα και τους αριθμούς. Έτσι, κάποιο τεκμήριο ανταποκρίνεται σ' ένα query, αν έχει κοινό τουλάχιστον έναν classifier και οποιοδήποτε κύριο όνομα ή αριθμό.

Η υψηλή ομοιότητα των δύο αυτών συλλογών, επιτρέπει τη χρήση μικρότερου παραθύρου ημερομηνίας, απ' ότι στην περίπτωση AP - German SDA. Τα πειράματα έδωσαν τα καλύτερα αποτελέσματα για παράθυρα έκτασης μίας ημέρας μόνο. Σ' αυτή την περίπτωση το thresholding δεν επέφερε ιδιαίτερες βελτιώσεις, πιθανώς γιατί φιλτραρίστηκαν τόσο κακά όσο και καλά ζεύγη. Αυτό δείχνει και πάλι την ανάγκη για μια καλύτερη στρατηγική χρήσης του.

3. Αξιολόγηση των ευθυγραμμίσεων

Η αξιολόγηση των ευθυγραμμίσεων, μπορεί να αποβλέπει σε διαφορετικούς βασικούς στόχους. Από τη μία, κάποιος μπορεί να αποτιμήσει πόσα από τα παραγόμενα ζεύγη αποτελούνται από τεκμήρια τα οποία πραγματικά σχετίζονται μεταξύ τους (οι αποκαλούμενες «καλές ευθυγραμίσεις»). Αυτή είναι μια γενική άποψη, κατά την οποία, η ποιότητα των ευθυγραμμίσεων κρίνεται χωρίς να λαμβάνεται υπ' όψη κάποια συγκεκριμένη εφαρμογή. Από την άλλη, κάποιος μπορεί να θέλει να εκτιμήσει την απόδοση κάποιας συγκεκριμένης εφαρμογής που χρησιμοποιεί ευθυγραμμίσεις. Αυτό το δεύτερο σκεπτικό μπορεί να έχει το μειονέκτημα της απώλειας της γενικότητας, ο κίνδυνος όμως της επικέντρωσης σε καθαρά θεωρητικό επίπεδο, χωρίς στην ουσία να υπάρχει κανένα όφελος στην πράξη, αποφεύγεται.

Για την αξιολόγηση των ευθυγραμμίσεων, ανεξάρτητα από την εφαρμογή, εξετάζεται η ποιότητα ενός δείγματος από ζεύγη. Η πρακτική αυτή είναι παραπλήσια με την αποτίμηση σχετικότητας για την αξιολόγηση των ανακτώμενων αποτελεσματων σε μια απλή διαδικασία ανάκτησης πληροφοριών (π.χ. σε μια μονόγλωσση βάση δεδομένων), υπάρχουν όμως οι εξής σημαντικές διαφορές:

- Δεν είναι απολύτως ξεκάθαρο πώς μπορεί να κριθεί η ποιότητα ενός ζεύγους. Ο ορισμός της διαδικασίας ευθυγράμμισης σαν «το να εντοπίζεται η πιο παραπλήσια αντιστοιχία στην άλλη συλλογή», σημαίνει ότι κάποιος ανθρώπινος κριτής θα πρέπει να διαβάσει *ολόκληρη τη συλλογή* για να σιγουρευτεί ότι δεν υπάρχει κάποιο πιο συναφές κείμενο, πράγμα καθαρά μη πρακτικό.
- Είναι επίσης δύσκολο να ξεκαθαριστεί πώς θα κριθεί το ποσοστό συνάφειας μεταξύ δύο ευθυγραμμισμένων τεκμηρίων. Όταν κάνουμε αποτίμηση σχετικότητας για τα τεκμήρια που ανακτώνται σε μια απλή διαδικασία ανάκτησης, τα αποτελέσματα συγκρίνονται με query σχετικά μικρού μήκους και ο ανθρώπινος κριτής μπορεί να κάνει εύκολα τη σύγκριση. Στη συγκεκριμένη περίπτωση της αποτίμησης των ευθυγραμμίσεων, το query είναι ένα ολόκληρο τεκμήριο και η σύγκριση είναι δυσκολότερη. Για να

διευκολυνθούν τα πράγματα, εφαρμόζεται μια κλίμακα πέντε κατηγοριών (βλ. πίνακες)

- Ο ανθρώπινος κριτής πρέπει να διαβάζει δύο τεκμήρια για κάθε αποτίμηση σχετικότητας αντί για ένα (όπως συμβαίνει στην αποτίμηση απλών διαδικασιών ανάκτησης). Αυτό συμβαίνει, γιατί το query είναι διαφορετικό για κάθε ευθυγραμμισμένο ζεύγος.

Πίνακας 1. Κατηγορίες για την αποτίμηση των ευθυγραμμισμένων ζευγών

Class 1	Same story	Two documents cover exactly the same story/event (e.g. the results of the same candidate in the same primary for the US presidential election)
Class 2	Related story	Two documents cover two related events (e.g. two different primaries for the US presidential election)
Class 3	Shared aspect	Two documents address various topics, but at least one of them is shared (e.g. update on US politics of the day, one item is about the upcoming presidential election)
Class 4	Common terminology	Two documents are not really related, but a significant amount of the terminology is shared (e.g. one document on the US, the other on the French presidential election)
Class 5	Unrelated	Two documents have no apparent relation (e.g. one document about the US presidential election, the other about vacation traffic in Germany)

Ένα δείγμα της τάξης του 1% από τα τελικά ζεύγη που προκύπτουν από την ευθυγράμμιση των συλλογών AP – SDA αποτιμάται με βάση αυτές τις πέντε κατηγορίες. Τα αποτελέσματα είναι τα ακόλουθα:

Πίνακας 2. Αποτελέσματα αποτίμησης δείγματος 1% επί του συνόλου

1% sample:	852 out of 85125 pairs	
Class 1:	327	38.38%
Class 2:	174	20.42%
Class 3:	16	1.88%
Class 4:	123	14.44%
Class 5:	212	24.88%

4. Εφαρμογές των ευθυγραμμίσεων για την ανάκτηση πληροφοριών

I. Δια-γλωσσική ανάκτηση πληροφοριών με τη χρήση των ευθυγραμμίσεων

Οι ευθυγραμμίσεις μπορούν να χρησιμοποιηθούν για την κατασκευή ενός συστήματος για πολυγλωσσική ανάκτηση πληροφοριών. Ένα τέτοιο σύστημα είναι συγκριτικά απλό και πολύ οικονομικό καθώς πολύ λίγοι γλωσσικοί πόροι απαιτούνται για την ευθυγράμμιση των τεκμηρίων. Είναι επίσης αρκετά ανεξάρτητο από συγκεκριμένους συνδυασμούς γλωσσών υπό την προϋπόθεση οτι μπορούν να βρεθούν κατάλληλες συλλογές προς ευθυγράμμιση.

Χρησιμοποιήθηκε μια εκπληκτικά απλή στρατηγική που μπορεί να παραχθεί ως εξής: Στην περίπτωση που οι δύο συλλογές στις οποίες θα γίνει η αναζήτηση είναι παράλληλες (η μια δηλαδή με τεκμήρια που είναι ακριβείς μεταφράσεις των τεκμηρίων της άλλης), θα ήταν πιθανό η αναζήτηση να γίνει μόνο στη συλλογή που ανταποκρίνεται στη γλώσσα του query και να επιστραφεί μια λίστα που να έχει παραχθεί από την αντικατάσταση κάθε τεκμηρίου που έχει βρεθεί με το αντίστοιχό του στην άλλη συλλογή. Αυτή βέβαια δεν είναι μια τόσο ενδιαφέρουσα περίπτωση, καθώς σπάνια μια συλλογή υπάρχει και σε μεταφρασμένη μορφή. Η απαίτηση άρα του να βρεθεί μια παράλληλη συλλογή αντικαθίσταται με την απαίτηση να βρεθεί μια συγκρίσιμη συλλογή. Αυτή η απαίτηση συναντάται σε πολλές πραγματικές εφαρμογές. Ένα παράδειγμα είναι η ευθυγράμμιση των συλλογών του AP και του SDA που δίνει τη δυνατότητα σε κάποιον να ψαξει τα κείμενα του SDA με query στα αγγλικά.

Κατ' αρχάς το query εκτελείται στη συλλογή πηγής δίνοντας μια λίστα αποτελεσμάτων όπου τα τεκμήρια κατατάσσονται με σειρά σχετικότητας. Αντί να αντικατασταθούν τα ανευρεθέντα τεκμήρια από τις μεταφράσεις τους, το mapping που προκύπτει σε επίπεδο τεκμηρίου, χρησιμοποιείται για να τα αντικαταστήσει με τα πιο συναφή αντίστοιχά τους από την άλλη συλλογή, αν υπάρχουν τέτοια διαθέσιμα. Αυτό έχει σαν αποτέλεσμα την παραγωγή μιας νέας λίστας που περιέχει τεκμήρια στη γλώσσα-στόχο. Επειδή όμως πολλά από τα τεκμήρια δεν είναι μέρος ενός ευθυγραμμισμένου ζεύγους, με αυτή τη στρατηγική δεν θ' ανακτηθούν ποτέ.

Το πρόβλημα αυτό προσεγγίστηκε με τη χρήση *βρόχου*[*] *ψευδο relevance feedback* μετά το στάδιο της αντικατάστασης (αλλά πριν την πραγματοποίηση της αναζήτησης στη γλώσσα-στόχο). Ένας ικανός αριθμός των τεκμηρίων με την υψηλότερη βαθμολογία σχετικότητας *υποτίθεται* οτι είναι σχετικά και απ' αυτά εξάγονται όροι που θεωρείται οτι τα αντιπροσωπεύουν καλά. Αυτοί οι όροι συνθέτουν ένα query που χρησιμοποιείται για τη νέα αναζήτηση. Επειδή τα τεκμήρια είναι ήδη στη γλώσσα-στόχου, το ίδιο συμβαίνει και με το query που παράγεται. Το feedback πραγματοποιείται πάνω σ' ένα σύνολο από ευθυγραμμισμένα τεκμήρια και

[*] **Βρόχος (Loop):** «Μια ακολουθία εντολών που επαναλαμβάνεται έως ότου ικανοποιηθεί μια προδιαγεγραμμένη συνθήκη, όπως η συμφωνία με ένα στοιχείο δεδομένων ή η ολοκλήρωση μιας απαρίθμησης» (Το Επιστημονικό Λεξικό της Πληροφορικής: Oxford-Κλειδάριθμος, 1998)

χρησιμοποιείται για να παράγει *το ίδιο* τη μετάφραση. Οι νέοι όροι δεν μπορούν να συνδυαστούν με το πρωτότυπο query γιατί οι γλώσσες τους δεν ταιριάζουν. Μόνο όροι που προέρχονται από διαδικασία του feedback σχηματίζουν το νέο query. Αυτή η στρατηγική αποδίδει καλά για συγκεκριμένα queries, αποτυγχάνει όμως αν το αρχικό query δεν ανακτήσει σχετικά τεκμήρια. Αυτό μπορεί να διορθωθεί με τον συνδυασμό αυτής της στρατηγικής με ένα λεξικό Η ίδια wordlist που χρησιμοποιήθηκε για την παραγωγή των ευθυγραμμίσεων, μπορεί να χρησιμοποιηθεί για μια λέξη προς λέξη μετάφραση του query. Αυτές οι μεταφράσεις συνήθως έχουν προβλήματα, ακόμα κι αν χρησιμοποιείται λεξικό υψηλής ποιότητας. Οι μεταφράσεις είναι συχνά διφορούμενες και αν συμπεριληφθούν πολλές λανθασμένες μεταφράσεις, αυτό μπορεί να είναι πολύ επιβλαβές για την ανάκτηση. Άλλο πρόβλημα είναι η απουσία πολλών όρων από το λεξικό επειδή ο όρος του query μπορεί να είναι κλιμένος ή επειδή απουσιάζει ακόμα και η ίδια βασική μορφή του.

Γι αυτό η απλή λέξη-προς-λέξη αντικατάσταση με όλες τις πιθανές μεταφράσεις συνήθως δεν αποδίδει καλά και ταιριάζει λιγότερο για τον υπολογισμό των ευθυγραμμίσεων. Επιπλέον προσπάθειες απαιτούνται, όπως η χρήση μερών του λόγου, αποσαφήνιση εννοιών, λεξικογραφικά σωστή κανονικοποίηση λέξεων κ.λπ. Είναι πάντως εύκολο να συνδυαστούν τέτοιες απλές μεταφράσεις μέσω λεξικού με τη διαδικασία του relevance feedback που περιγράψαμε. Αντί να διαγράφεται το πρωτότυπο query πριν το feedback, αυτό μπορεί να αντικατασταθεί με μια «ψευδο-μετάφραση». Επειδή η βαρύτητα των όρων ξαναϋπολογίζεται στη διαδικασία του feedback, αυτόματα λύνεται και το πρόβλημα της απόδοσης βαρών στους δύο πληροφοριακούς πόρους (λεξικό vs ευθυγραμμίσων). Οι ευθυγραμμίσεις, σαν πρόσθετη πληροφοριών βοηθούν στο να μειωθούν τα προβλήματα στη μετάφραση που αναφέραμε.

Στην περίπτωση που η συλλογή που ερευνάται δεν είναι ευθυγραμμισμένη, δύο ανεξάρτητες ευθυγραμμισμένες μπορούν να χρησιμοποιηθούν για την παραγωγή του query. Οι δύο αυτές συλλογές, χρησιμοποιούνται μόνο για τη μεταφορά του query στη γλώσσα-στόχο ενώ η αναζήτηση πραγματοποιείται στην τρίτη συλλογή.

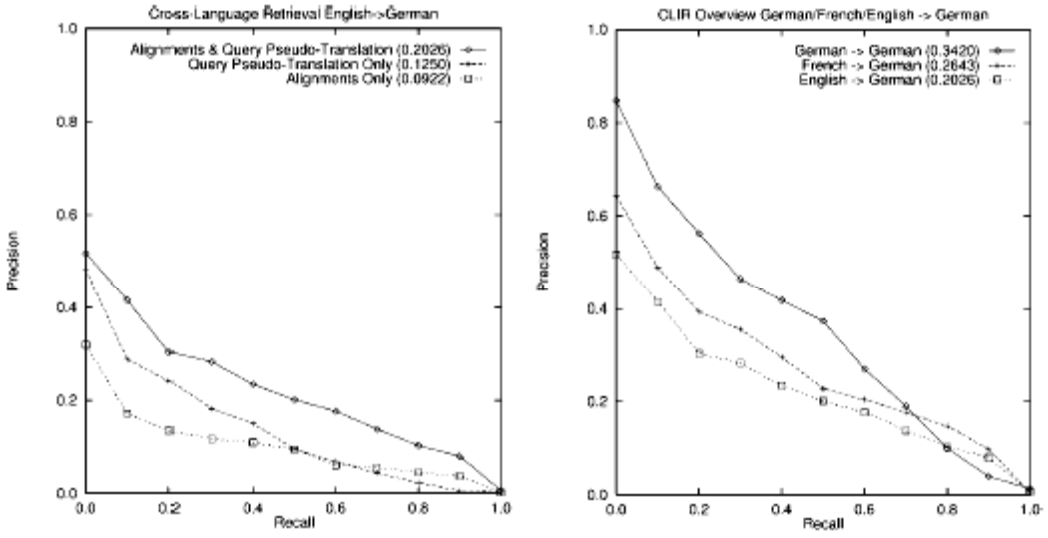
II. Αποτελέσματα στη συλλογή του TREC-6

Το Text REtrieval Conference (TREC) (http://trec.nist.gov) επιχορηγείται από το National Institute of Standards and Technology (NIST) των ΗΠΑ και το Αμερικανικό Υπουργείο Άμυνας. Σκοπός του TREC είναι να υποστηρίξει την έρευνα πάνω στην ανάκτηση πληροφοριών, παρέχοντας τη βάση για ευρεία αξιολόγηση μεθοδολογιών ανάκτησης κειμένων. Για κάθε συνέδριο, το NIST παρέχει ένα test set τεκμηρίων και ερωτήσεων. Οι συμμετέχοντες τρέχουν τα δικά τους συστήματα ανάκτησης πάνω στα δεδομένα και επιστρέφουν στο NIST μια λίστα με τα καλύτερα αποτελέσματα. Το NIST εξετάζει τα αποτελέσματα, κρίνει αν τα ανακτημένα τεκμήρια είναι αυτά που ανταποκρίνονται στα αντίστοιχα ερωτήματα και αξιολογεί τα αποτελέσματα. Κάθε συνέδριο κλείνει με ένα workshop όπου όλοι οι συμμετέχοντες μοιράζονται τις εμπειρίες τους.

Οι στρατηγικές που προαναφέραμε εφαρμόστηκαν πάνω στη βάση εργασίας του TREC-6. Η συλλογή πάνω στην οποία έγινε η αναζήτηση ήταν ένας συνδυασμός των συλλογών των ειδησεογραφικών πρακτορείων German SDA και NZZ (Neue Zurcher Zeitung). Και οι δύο συλλογές είναι στη γερμανική γλώσσα, άρα ο συγκεκριμένος συνδυασμός είναι υπερσύνολο της συλλογής που ευθυγραμμίστηκε με

το AP και το French SDA. Αξιολογήθηκε η σχετικότητα των αποτελεσμάτων για 21 queries.

Τα ακόλουθα γραφήματα παρουσιάζουν μια σύγκριση μεταξύ της χρήσης μόνο των ευθυγραμμίσεων, της χρήσης της ψευδο-μετάφρασης μέσω λεξικού, και τέλος, της χρήσης συνδυασμού των δύο μεθόδων, δηλαδή της ανάκτησης αρχικά με μετάφραση μέσω λεξικού και βελτίωσης της μετάφρασης κατόπιν με τη χρήση των ευθυγραμμίσεων, λαμβάνοντας υπ' όψιν την ακρίβεια/ανάκληση.



Το αριστερό γράφημα συγκρίνει τα αποτελέσματα σε μια αγγλο-γερμανική CLIR. Ο συνδυασμός και των δύο μεθόδων δίνει σαφώς το καλύτερο αποτέλεσμα, βελτιώνοντας τη μέση ακρίβεια κατά 62% σε σχέση με τη χρήση λεξικού μόνο.Το δεξί γράφημα συγκρίνει τις CLIR προερχόμενες από τα αγγλικά και τα γαλλικά ως προς μια μονόγλωσση ανάκτηση στα γερμανικά. Η γαλλο-γερμανική CLIR παράγει ελαφρά καλύτερα αποτελέσματα από την αγγλο-γερμανική. Αυτό συμβαίνει γιατί εδώ η διαδικασία ευθυγράμμισης είναι πολύ απλούστερη και δείχνει ότι όταν πολύ συναφείς συλλογές είναι διαθέσιμες για ευθυγράμμιση, μπορεί να κατασκευαστεί ένα σύστημα χωρίς τη χρήση πόρου με τη μορφή λεξικού.

4. Σύνοψη και συμπεράσματα

Όπως είδαμε, η διαδικασία της ευθυγράμμισης δεν απαιτεί τη χρήση ιδιαίτερων πόρων. Δεν απαιτούνται ούτε ακριβό hardware, ούτε ακριβά γλωσσικά εργαλεία υψηλής ποιότητας. Είναι επίσης εύκολο να προσαρμοστεί σε ένα δυναμικό περιβάλλον όπου νέα τεκμήρια προστίθενται συνεχώς σε μια συλλογή. Σ' αυτή την περίπτωση, χάρη στη χρήση των «παραθύρων ημερομηνίας», οι ευθυγραμμίσεις μπορούν να επεκτείνονται, χωρίς να χρειάζεται να πεταχτούν τα ήδη υπάρχοντα ζεύγη.

Η χρήση των ευθυγραμμίσεων, δίνει εξαιρετικά αποτελέσματα, αποδεικνύοντας την αξία τους για πραγματικές εφαρμογές. Άλλες μελλοντικές

εφαρμογές των ευθυγραμμίσεων είναι επίσης πιθανές, όπως αυτόματη παραγωγή λεξικών, δημιουργία θησαυρών κ.λπ.

Υπάρχει βέβαια ακόμα μεγάλο περιθώριο βελτίωσης στην όλη διαδικασία και από τις μέχρι τώρα ενδείξεις, φαίνεται ότι είναι σημαντικό να εξάγονται τα σωστά τμήματα των τεκμηρίων, φιλτράροντας τους όρους υψηλής και χαμηλής συχνότητας. Αυτή η ιδέα θα μπορούσε να υλοποιηθεί περαιτέρω με τη χρήση των πιο συναφών προτάσεων ανάμεσα σε δύο τεκμήρια, ή με την τεχνική εξαγωγής περιλήψεων από τα τεκμήρια και σύγκρισης των περιλήψεων αυτών. Τα μέχρι στιγμής πειράματα πάντως έδειξαν μια απαγορευτική αύξηση στη δυσκολία των διάφορων υπολογισμών.

Ίσως το πιο ενδιαφέρον θέμα να είναι η ενοποίηση περισσότερων ή καλύτερων γλωσσικών πόρων. Καθώς υπάρχει ελπίδα όλο και περισσότερα γλωσσικά εργαλεία να γίνονται διαθέσιμα, η προοπτική αυτή φαίνεται αρκετά πιθανή. Η διαδικασία ευθυγράμμισης μπορεί να εκμεταλλευτεί τα πλεονεκτήματα αυτής της ενοποίησης, με αποτέλεσμα την περαιτέρω βελτίωση και της ίδιας.

3. ΛΟΓΙΣΜΙΚΑ ΕΦΑΡΜΟΓΗΣ ΤΗΣ CLIR

I. **Cindor** της TextWise (http://www.cindorsearch.com)
II. **TwentyOne** της Irion Technologies (http://www.irion.nl/products/index.html)
III. **Pidgin** της Irion Technologies (http://www.pidgin.nl)
IV. **AnswerWorks** της WexTech (http://www.wextech.com/products.html)
V. **Lirix** της Xerox (http://www.xrce.xerox.com/programs/lirix/)
VI. **Relevancy** της Eurospider (http://www.eurospider.com/en/relevancy/relevancy.htm)

4. ΠΗΓΕΣ

I. **Braschler, M., Schäuble, P**. *Multilingual Information Retrieval Based on Document Alignment Techniques στο* "Lecture Notes in Computer Science: Research and Advanced Technology for Digital Libraries: Second European Conference, ECDL'98, Heraklion, Crete, Greece, September 1998. Proceedings" (http://www.springerlink.com/app/home/contribution.asp?wasp=d0phrmquxmw6bewumma1&referrer=parent&backto=searcharticlesresults,8,13;)

II. **Peters C., Picchi E.** *Across Languages, Across Cultures: issues in multilinguality and digital libraries*, στο D-Lib, May 1997 (http://www.dlib.org/dlib/may97/peters/05peters.html)

III. **Pavani A**. *A Model of Multilingual Digital Library*, στο Ci.Inf., Brasilia, v.30, n.3.,p.73-81, Sep./Dec. 2001 (http://www.ibict.br/cionline/300301/3031001.pdf)

IV. **Oard, Douglas W**. *Serving users in many languages: Cross-Language Information Retrieval for Digital Libraries,* στο D-Lib, December 1997 (http://www.dlib.org/dlib/december97/oard/12oard.html)

V. **Evans, D. A…. [et al]** *Anatomy of Commercial CLIR Applications,* στο CLEF Workshop 2002, Rome, Italy, September 19, 2002 (http://clef.iei.cnr.it:2002/workshop2002/presentations/com-clir.pdf)