"ΛΟΓΟΠΛΟΗΓΗΣΗ"
Μαΐος 1999
Τεύχος 5

Επιστημονικός Υπεύθυνος:
Καθηγητής Γιώργος Καραγιάννης

Υπεύθυνη Έκδοσης:
Δρ. Ιωάννα Μαλαγαρδή

Συνεργάτες:
Αναστάσιος Πατρικάκος
Δέσποινα Σκούταρη
Αθανασία Φούρλα

Γραφίστας:
Άρτεμις Γλάρου

Διεύθυνση:
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 & Επιδαύρου
151 25 Παράδεισος Αμαρουσίου
τηλ.: 6800959 • fax: 6854270
e-mail: ioanna@ilsp.gr
http:// www.ilsp.gr

Την ευθύνη των κειμένων έχουν οι συγγραφείς.

Η χρηματοδότηση της έκδοσης αυτής έγινε από το πρόγραμμα EUROMAP (LE) το οποίο χρηματοδοτήθηκε από την DG XIII της Ευρωπαϊκής Επιτροπής.

Η "Λογοπλοήγηση" διανέμεται δωρεάν.

"LogoNavigation"
May 1999
Issue 5

Scientific Director:
Professor George Carayannis

Edition Executive:
Dr. Ioanna Malagardi

Collaborators:
Anastasios Patrikakos
Despina Scutari
Athanassia Fourla

Graphics Designer:
Artemis Glarou

Address:
Institute for Language and Speech Processing
Artemidos 6 & Epidavrou Str.
151 25 Marousi
Athens, Greece
tel: 301- 6800959 • fax: 301-6854270
E-mail: ioanna@ilsp.gr
http://www.ilsp.gr

The authors are responsible for text content.

Funding for this issue was carried out by the EUROMAP (LE) project which is funded by DG XIII of the European Commission.

"LogoNavigation" is distributed free of charge.

Πίνακας Περιεχομένων / *Table of Contents*

Εισαγωγικό Σημείωμα/*Introductory Note*

Το πέμπτο τεύχος της Λογοπλοήγησης περιέχει τις εξής θεματικές ενότητες:

Η πρώτη θεματική ενότητα περιλαμβάνει τις ομιλίες που δόθηκαν κατά τη διάρκεια των εγκαινίων των νέων κτηριακών εγκαταστάσεων του ΙΕΛ. Τα εγκαίνια του Ινστιτούτου πραγματοποιήθηκαν στις 6 Απριλίου 1998. Κατά την διάρκεια της εκδήλωσης λειτούργησε έκθεση προϊόντων και δραστηριοτήτων του Ινστιτούτου. Τα εγκαίνια τελέσθηκαν από την Υπουργό Ανάπτυξης κυρία Βάσω Παπανδρέου. Στην εκδήλωση παρέστη ο Γενικός Γραμματέας Έρευνας και Τεχνολογίας Καθηγητής Εμμανουήλ Φραγκούλης. Την εκδήλωση τίμησαν με την παρουσία τους πλήθος παραγόντων του επιστημονικού και επιχειρηματικού κόσμου, καθώς επίσης και στελέχη Δημοσίων Υπηρεσιών που ασχολούνται με την Έρευνα και Τεχνολογία.

Η δεύτερη θεματική ενότητα περιλαμβάνει γενικές πληροφορίες και δελτίο τύπου για το Κοινοτικό έργο EUROMAP. Το έργο αυτό δρα καταλυτικά στον Ελληνικό χώρο όπως και στις άλλες Ευρωπαϊκές χώρες προσπαθώντας να ευαισθητοποιήσει τον βιομηχανικό χώρο, αλλά και τους υποψήφιους χρήστες της Γλωσσικής Τεχνολογίας.

Η τρίτη θεματική ενότητα περιλαμβάνει τις περιλήψεις των εισηγήσεων της ημερίδας με θέμα «Ανάκτηση και Εξαγωγή Πληροφορίας» που πραγματοποιήθηκε στο Πανεπιστήμιο Πατρών στις 11 Απριλίου 1998 στο πλαίσιο των δραστηριοτήτων του έργου «Ανθρώπινο Δίκτυο Γλωσσικής Τεχνολογίας». Την διοργάνωση της ημερίδας είχαν αναλάβει το Ινστιτούτο Επεξεργασίας του Λόγου σε συνεργασία με το Εργαστήριο Ενσύρματης Τηλεπικοινωνίας του Πανεπιστημίου Πατρών.

Η τέταρτη θεματική ενότητα περιέχει θέματα σχετικά με τη δημιουργία γλωσσαρίου όρων Γλωσσικής Τεχνολογίας και Πληροφορικής.

Τέλος η πέμπτη θεματική ενότητα περιέχει ειδήσεις σχετικές με την Γλωσική Τεχνολογία και την Πληροφορική.

I. Ομιλίες που δόθηκαν κατά τη διάρκεια των Εγκαινίων των Νέων Κτηριακών Εγκαταστάσεων του ΙΕΛ /*Speeches given during the Inauguration Ceremony of the New Offices of ILSP*

Ομιλία της Υπουργού Ανάπτυξης κας Βάσως Παπανδρέου στα Εγκαίνια του ΙΕΛ

Είμαι ιδιαίτερα ευτυχής που μου δίνεται η ευκαιρία να εγκαινιάσω τις νέες κτηριακές εγκαταστάσεις του Ινστιτούτου Επεξεργασίας του Λόγου (ΙΕΛ) που είναι ένας από τους εποπτευόμενους από το Υπουργείο Ανάπτυξης οργανισμούς. Γνωρίζω το ερευνητικό και αναπτυξιακό έργο του Ινστιτούτου και χαίρομαι που σήμερα θα μου δοθεί η ευκαιρία να έχω μία περισσότερο άμεση αντίληψη.

Στα συμβούλια των Υπουργών της Ευρωπαϊκής Ένωσης συζητάμε συχνά τις νέες τεχνολογίες και είμαι πεπεισμένη ότι η Γλωσσική Τεχνολογία έχει πολλές καινοτομικές δυνατότητες. Στο Πέμπτο Πρόγραμμα Πλαίσιο, το οποίο τώρα ετοιμάζουμε, η Γλωσσική Τεχνολογία έχει ενταχθεί στον άξονα του προγράμματος που προωθεί την δημιουργία της φιλικής κοινωνίας των πληροφοριών.

Σύμφωνα με το ανωτέρω πρόγραμμα η εργασία τα επόμενα χρόνια θα επικεντρωθεί σε προχωρημένες τεχνολογίες του λόγου που θα επιτρέψουν οικονομικά αποτελεσματικές ανταλλαγές μεταξύ γλωσσών και πολιτισμών, φυσικές διεπαφές σε ψηφιακές υπηρεσίες και αμεσότερη σύνδεση και χρήση των βάσεων δεδομένων με περιεχόμενο πολυμέσων.

Μία βασική προτεραιότητα του Πέμπτου Προγράμματος Πλαισίου είναι η προσθήκη της πολυγλωσσικότητας (multilinguality) στα πληροφοριακά συστήματα σε όλα τα επίπεδα του κύκλου της πληροφορίας.

Φαίνεται ότι τελειώνει η εποχή που η πολυγλωσσία μπορούσε να χαρακτηρισθεί σαν εμπόδιο στην συνεννόηση των Ευρωπαϊκών λαών ή σαν εμπόδιο στο εμπόριο. Αυτό συμβαίνει τόσο χάρη στις λύσεις που εξευρίσκει η Γλωσσική Τεχνολογία, όσο και φυσικά χάρη στην Γλωσσική εκπαίδευση που συντελεί στο να μα-

θαίνουμε πολλές γλώσσες στην Ευρωπαϊκή Ήπειρο.

Η συμμετοχή μας στην Ευρωπαϊκή Ένωση είχε για την Ελλάδα ένα επιπλέον θετικό στοιχείο. Εξασφάλισε στην γλώσσα μας ίσες ευκαιρίες με τις άλλες Ευρωπαϊκές γλώσσες στο Ευρωπαϊκό Πολυγλωσσικό Περιβάλλον.

Ιδιαίτερα στα θέματα της τεχνολογίας με τις δυνατότητες συμμετοχής στα Ευρωπαϊκά Προγράμματα, ανοίχτηκε στην Ελληνική γλώσσα ο δρόμος προσπέλασης και επικοινωνίας στα δίκτυα πληροφοριών και έτσι μία επιπλέον δυνατότητα επιβίωσης και διάδοσης, εφ' όσον βέβαια ενταχθεί εγκαίρως στις γλώσσες που δημιουργούν υποδομή από πλευράς πληροφορικής.

Αυτή η υποδομή συμβάλλει στην ισότιμη χρήση των ολιγότερο ομιλουμένων γλωσσών με τις περισσότερο ομιλούμενες γλώσσες. Περιμένουμε λοιπόν από το ΙΕΛ την δημιουργία αυτής της υποδομής σε συνεργασία με τους άλλους φορείς που έχουν τεχνογνωσία στην περιοχή.

Δεν θα πρέπει να παραβλέψουμε το γεγονός ότι η γλώσσα μας είναι μία από τις πιο ανθεκτικές στο χρόνο γλώσσες του πλανήτη με καταγραμμένα έπη πριν από τρεις χιλιάδες χρόνια, η οποία, παρά τις συντακτικές και μορφολογικές αλλαγές που έχουν μεσολαβήσει, παραμένει κοντά στις πρωτογενείς μορφές της. Εκτός από τα έπη στην Ελληνική γλώσσα γράφτηκαν οι πρώτες τραγωδίες, οι πρώτες κωμωδίες, οι πρώτες φιλοσοφικές θεωρίες, οι πρώτοι νόμοι και τα πρώτα μαθηματικά. Στην γλώσσα μας μεταφράστηκε η Παλαιά Διαθήκη από τους 70 σοφούς και γράφτηκαν τα Ευαγγέλια. Υπήρξε η επίσημη γλώσσα μιας αυτοκρατορίας που έζησε 1.000 χρόνια και ομιλείτο σε όλη την λεκάνη της Μεσογείου. Τροφοδότησε με ιδέες την Αναγέννηση και με ελπίδες το κρυφό σχολειό. Τροφοδοτεί ακόμη με χιλιάδες όρους την σύγχρονη επιστήμη και εμφανίζει τον μεγαλύτερο βαθμό διείσδυσης σε άλλες γλώσσες, ιδιαίτερα με λέξεις υψηλού νοήματος.

Σήμερα η Ελληνική γλώσσα ανήκει στην κατηγορία των λιγότερο ομιλουμένων γλωσσών αλλά εξακολουθεί να έχει δύναμη έκφρασης και να αποτελεί πηγή έμπνευσης για βραβευμένους ποιητές και πεζογράφους. Φαίνεται ότι μετά από μια περίοδο αναζήτησης μεταξύ παλαιότερων και νεοτέρων μορφών η Ελληνική γλώσσα τείνει τώρα να ισορροπήσει σε ένα σημείο που θα της επιτρέψει να κατακτήσει νέους στόχους δημιουργίας.

Το Υπουργείο Ανάπτυξης έχει μια συγκεκριμένη πολιτική για την Ελληνική γλώσσα.

Επιδίωξή μας είναι η Ελληνική να είναι μεταξύ των γλωσσών που θα «μιλιώνται» στα δίκτυα πληροφοριών που οικοδομούνται στο πλαίσιο της κοινωνίας των πληροφοριών του μέλλοντος, «να μεταφράζεται», «να αναγνωρίζεται», να μπορεί να χρησιμοποιείται στην ανάκτηση δεδομένων από πολυγλωσσικές βάσεις δεδομένων. Για τους λόγους αυτούς υποστηρίζουμε την δημιουργία εργαλείων λογισμικού για την αποτελεσματική παρουσία της Ελληνικής στον χώρο της επικοινωνίας ανθρώπου-μηχανής.

Αντίστοιχα επιχειρήματα ισχύουν για την γλωσσική εκπαίδευση. Επειδή η διδασκαλία θα χρησιμοποιεί όλο και περισσότερο τα εποπτικά μέσα, ιδιαίτερα αυτά που αξιοποιούν την δυναμική των πολυμέσων, υποστηρίζουμε τις σύγχρονες τάσεις της γλωσσικής διδασκαλίας βασισμένης στον ηλεκτρονικό υπολογιστή και τούτο γιατί πιστεύουμε ότι αφενός μεν προσφέρουν μεγαλύτερες δυνατότητες διάδοσης της γλώσσας μας, ιδιαίτερα αν ο σχεδιασμός είναι γύρω από την αυτοδιδασκαλία, και αφετέρου εξασφαλίζουν καλύτερες δυνατότητες εμπέδωσης μιας γλώσσας.

Η πολιτική του Υπουργείου Ανάπτυξης είχε την τελευταία δεκαετία στα θέματα της Γλωσσικής Τεχνολογίας τους εξής άξονες:

1. Συμμετοχή στα μεταφραστικά προγράμματα EUROTRA και SYSTRAN της Ευρωπαϊκής Ένωσης.
2. Δημιουργία κρίσιμης μάζας επιστημόνων που εργάζονται ερευνητικά σε θέματα Επεξεργασίας Φυσικής Γλώσσας και Τεχνολογίας Φωνής.
3. Ενθάρρυνση της συμμετοχής Ελληνικών Εργαστηρίων στα διεθνή προγράμματα γλώσσας και φωνής με χορήγηση από το Υπουργείο Ανάπτυξης της λεγομένης εθνικής συμμετοχής.
4. Υλοποίηση Εθνικών Προγραμμάτων από την

Γενική Γραμματεία Έρευνας και Τεχνολογίας χρηματοδοτουμένων από τους προϋπολογισμούς των Διαρθρωτικών Ταμείων της Κοινότητας και το Ελληνικό Υπουργείο Εθνικής Οικονομίας.

Υπήρξαν ήδη δύο προγράμματα υποδομής:
α) Το πρόγραμμα ΛΟΓΟΣ (LOGOS)
β) Το πρόγραμμα ΔΙΑΛΟΓΟΣ (DIALOGOS)

Μαθαίνω ότι στα προγράμματα αυτά υπήρξε ευρεία συμμετοχή από πανεπιστήμια και βιομηχανικούς φορείς και ότι η αξιολόγησή τους έδειξε ότι υπήρξαν ιδιαίτερα επιτυχημένα. Προκηρύσσουμε αυτές τις ημέρες το τρίτο Εθνικό Πρόγραμμα σε Γλωσσική Τεχνολογία με μία σειρά από στόχους.

Ένας στόχος είναι η ανάπτυξη πρωτοβουλιών, ιδιαίτερης σημασίας για τον Ελληνικό χώρο, η ενίσχυση της υπάρχουσας τεχνογνωσίας και η ενθάρρυνση της ενσωμάτωσης της καινοτομίας σε τελικά προϊόντα. Με αυτόν τον τρόπο αφενός μεν η Ελληνική γλώσσα θα έχει επιτυχημένη παρουσία στην κοινωνία των πληροφοριών, αφετέρου δε, θα δημιουργηθούν προϊόντα διεθνούς απήχησης σε ηλεκτρονική μορφή.

Ένας άλλος στόχος είναι η προετοιμασία της Ελληνικής συμμετοχής στο 5ο Πρόγραμμα Πλαίσιο της Ευρωπαϊκής Ένωσης, το οποίο περιέχει, όπως σας είπα, μεταξύ των βασικών σκοπών του την δημιουργία μιας φιλικής στον χρήστη κοινωνίας των πληροφοριών. Η Γλωσσική Τεχνολογία μπορεί να συμβάλλει ουσιαστικά σε έναν τέτοιο στόχο.

Το υπό προκήρυξη Εθνικό Πρόγραμμα έχει τους εξής θεματικούς τομείς:
1. Ηλεκτρονική Λεξικογραφία
2. Ορολογικά Ηλεκτρονικά Λεξικά & Συλλογή Κειμένων Εντάσεως Όρων
3. Μεταφραστικά Εργαλεία και Συστήματα Μηχανικής Μετάφρασης
4. Τεχνολογία Φωνής
5. Γλωσσική Τεχνολογία στην Επικοινωνία Ανθρώπου-Μηχανής
6. Γλωσσική Εκπαίδευση

Θα ήθελα να σημειώσω ότι η ΓΓΕΤ έχει προσπαθήσει να κάνει εισαγωγή δράσεων με πολιτιστική και κοινωνική διάσταση. Στα θέματα αυτά πρέπει να έχουμε ιδιαίτερη ευαισθησία.

Οι Γλωσσικές Τεχνολογίες έχουν μεγάλο δυναμικό και μπορούν να βελτιώσουν τόσο την ανθρώπινη επικοινωνία όσο και τον διάλογο με τις μηχανές. Επίσης θα ανοίξουν ορίζοντες νέων αγορών και θα βοηθήσουν στην διατήρηση της Ευρωπαϊκής πολιτιστικής και γλωσσικής ποικιλίας.

Κυρίες και Κύριοι

Χρειαζόμαστε μία εθνική στρατηγική για την επιβίωση και τη διάδοση της Ελληνικής γλώσσας. Πιθανόν η πολιτική αυτή να είναι σύνθετη, να θέλει πολλή σκέψη αλλά πρέπει να υπάρξει σύντομα. Οι οργανισμοί που δουλεύουν για την Ελληνική γλώσσα και τον πολιτισμό πρέπει να συντονισθούν, να συνεργασθούν και να προσφέρουν στην πολιτική ηγεσία ένα σχέδιο.

Με αυτές λοιπόν τις σκέψεις αλλά και με αυτές τις προσδοκίες ας ευχηθούμε στο Ινστιτούτο Επεξεργασίας του Λόγου να επιτύχει στην αποστολή του.

Ευχαριστώ

Ομιλία του Δ/ντή του ΙΕΛ Καθηγητή Γιώργου Καραγιάννη στα Εγκαίνια του Ινστιτούτου

Κυρία Υπουργέ,

Θέλω να σας ευχαριστήσω θερμά εκ μέρους του Επιστημονικού Συμβουλίου του ΙΕΛ που μας κάνετε την τιμή να τελέσετε τα εγκαίνια των νέων μας εγκαταστάσεων.

Κύριε Γενικέ, Κύριοι συνάδελφοι, Κυρίες & Κύριοι,

Ευχαριστούμε που είστε μαζί μας σ' αυτήν την τελετή που για το Ινστιτούτο Επεξεργασίας του Λόγου είναι ορόσημο, μια και πέρασαν 5 χρόνια λειτουργίας του και αποτελεί για το Επιστημονικό Συμβούλιο, τον Διευθυντή και το προσωπικό του Ινστιτούτου ευκαιρία παρουσίασης του έργου που έχει γίνει μέχρι σήμερα. Η Γλωσσική Τεχνολογία έχει καθιερωθεί πρόσφατα σαν κλάδος της πληροφορικής που αρχίζει και ωριμάζει πια και μπορεί να οδηγήσει σε βιομηχανικές

διαδικασίες και προϊόντα. Άλλωστε εμφανίστηκαν και κάποιες συγκεκριμένες ανάγκες και μέσα από την πορεία προς τις νέες κοινωνίες προς τις οποίες βαδίζουμε, την κοινωνία των πληροφοριών και την κοινωνία της δια βίου μάθησης.

Η Γλωσσική Τεχνολογία μπορεί να βοηθήσει σε καλύτερη και φιλικότερη επικοινωνία με την μηχανή. Είναι γνωστές οι εφαρμογές της αναγνώρισης φωνής και της παραγωγής συνθετικής ομιλίας ή των εφαρμογών κατανόησης κειμένου. Η Γλωσσική Τεχνολογία μπορεί επίσης να βοηθήσει σε καλύτερη συνεννόηση μεταξύ των ανθρώπων μειώνοντας τα γλωσσικά λάθη σε μονογλωσσικό περιβάλλον καθώς και με τις δυνατότητες μηχανικής μετάφρασης που προσφέρει σε πολυγλωσσικό περιβάλλον. Ακόμη η τεχνολογία αυτή βοηθάει στον σχεδιασμό δυναμικών εργονομιών σε σύνθετες εκδοτικές κατασκευές όπως είναι ένα λεξικό ή μία εγκυκλοπαίδεια.

Είναι σημαντικό να αναφέρει κανείς ορισμένες ιδιαιτερότητες της γλωσσικής τεχνολογίας:

α) Δεν μπορεί να είναι εισαγόμενη τεχνολογία όπως τόσες άλλες. Δεν είναι τόσο πιθανό να ενδιαφέρει από οικονομικής απόψεως μία μεγάλη εταιρεία του εξωτερικού η ανάπτυξη τεχνολογίας για τα Ελληνικά όταν υπάρχουν τόσες άλλες γλώσσες που ομιλούμενες από μεγαλύτερους πληθυσμούς δίδουν μεγαλύτερες ελπίδες για ικανοποιητικότερες αγορές. Από την άλλη πλευρά δεν μπορεί η τεχνολογία της Ελληνικής να αναπτυχθεί εξ' ίσου καλά εκτός Ελλάδος κυρίως για λόγους τεχνογνωσίας, δηλαδή δεν υπάρχει εκτός Ελλάδος η αναγκαία τεχνογνωσία για τα Ελληνικά. Η γλωσσική τεχνολογία εξαρτάται από την γλώσσα και πρέπει να αναπτυχθεί για κάθε γλώσσα χωριστά.

β) Απαιτεί πολύχρονη εκπαίδευση των τεχνολόγων που την υποστηρίζουν κυρίως λόγω της πολυπλοκότητας της και της διακλαδικής της φύσεως.

γ) Απαιτεί στενή συνεργασία μηχανικών πληροφορικής και ειδικών με βασικές γνώσεις στις ανθρωπιστικές επιστήμες εργαζομένων από κοινού και έχοντας αναπτύξει κοινή γλώσσα.

δ) Απαιτεί πολύχρονη εργασία ρουτίνας για την ανάπτυξη δεδομένου ότι έχει ανάγκη την εισαγωγή στον Η/Υ μεγάλου όγκου γλωσσικών πόρων (linguistic resources) είτε υπό μορφή δομημένων κειμένων, λεξικών, κανόνων γραμματικής ή φωνητικών λογατόμων κατάλληλα κομμένων και ραμμένων.

Για τους ανωτέρω λόγους η ανάπτυξη της Γλωσσικής Τεχνολογίας συνδέεται με την δημιουργία παράδοσης και μακρόπνοης ενασχόλησης.

Η ανάπτυξη της Γλωσσικής Τεχνολογίας μπορεί να επιδράσει σε επίπεδο τεχνολογικό και οικονομικό. Από την τεχνολογική σκοπιά: Πρόκειται για μία δύσκολη τεχνολογία στην αιχμή της σύγχρονης πληροφορικής. Επιστρατεύει επιστημονικές γνώσεις από πολλούς επιστημονικούς τομείς: Επεξεργασία Φυσικής Γλώσσας, Επεξεργασία Σημάτων, Αναγνώριση Προτύπων, Τεχνητή Νοημοσύνη κ.λπ. Επομένως πρόκειται για έναν σύγχρονο κλάδο με διεπιστημονική υπόσταση. Η χρησιμοποίηση μεθόδων από τους ανωτέρω κλάδους υπήρξε στο ΙΕΛ ιδιαίτερα γόνιμος στόχος. Οι οικονομικές επιδράσεις της Γλωσσικής Τεχνολογίας είναι δύο ειδών: α) Δημιουργία δραστηριότητας και αγοράς με μεγάλο αριθμό νέων προϊόντων, β) Επιδράσεις στην αύξηση της παραγωγικότητας στο περιβάλλον του σύγχρονου γραφείου.

Η προσπάθεια στο ΙΕΛ έγινε αυτά τα 5 χρόνια σε έξι διαφορετικά μέτωπα που οδήγησαν στην οργάνωση σε έξη τμήματα, τα πέντε είναι καθαρά επιστημονικά, το έκτο το τμήμα συνδέσμου, όπως το λέμε, ασχολείται με τις υπηρεσίες που προσφέρει το ΙΕΛ και με διάφορα οριζόντια θέματα.

Το κάθε επιστημονικό τμήμα έχει τριπλή δραστηριότητα. Το πρώτο σκέλος είναι η ανάπτυξη, είτε πρόκειται για ανάπτυξη υποδομής είτε για ανάπτυξη προϊόντος. Το δεύτερο σκέλος είναι η έρευνα για να υποστηριχθεί καλύτερα η ανάπτυξη με καινοτομικές συνιστώσες. Το τρίτο σκέλος αφορά την συγκέντρωση γλωσσικών πόρων. Οι γλωσσικοί πόροι είναι το άλφα και το ωμέγα της ανάπτυξης της Γλωσσικής Τεχνολογίας. Για αυτόν τον ανωτέρω λόγο η οργάνωσή μας περιλαμβάνει συντήρηση διαφόρων βάσεων δεδομένων που εμπλουτίζουμε συνεχώς.

Στην δουλειά αυτή μεγάλη σημασία έχει και η διασφάλισή της ποιότητας. Είναι ένα θέμα που μας βασανίζει πάρα πολύ. Προσπαθούμε να μην δώσουμε

προς τα έξω κάτι μέτριο. Ειδικά για το γλωσσικό υλικό έχουμε κάποιες δυσκολίες εξαιτίας της πολυτυπίας που υπάρχει στην Ελληνική γλώσσα την οποία προσπαθούμε να σεβαστούμε όσο είναι δυνατό. Οι συνεχείς αλλαγές στην ορθογραφία που παρατηρούνται π.χ. με την έκδοση ενός νέου λεξικού μας κουράζουν αλλά προσπαθούμε να τις αντιμετωπίσουμε. Τα πέντε επιστημονικά τμήματα είναι:

- Τμήμα Ηλεκτρονικής Λεξικογραφίας
- Τμήμα Γλωσσικών Εφαρμογών Γραφείου
- Τμήμα Εκπαιδευτικής Τεχνολογίας
- Τμήμα Τεχνολογίας Φωνής
- Τμήμα Μηχανικής Μετάφρασης

Γίνεται προσπάθεια συνεργασίας των τμημάτων αυτών όποτε υπάρχει συμπληρωματικότητα στο πλαίσιο κάποιας εφαρμογής.

Ήθελα να αναφερθώ στην δουλειά των τμημάτων μέσα από τα προϊόντα που θα δείτε στην έκθεση που σε λίγο θα εγκαινιάσει η Υπουργός Ανάπτυξης. Το πιο γνωστό προϊόν μας είναι η «λογομάθεια». Ξεκίνησε να κατασκευάζεται εδώ και 5 χρόνια και έχει δώσει ήδη 2 ενδιάμεσα προϊόντα. Το τελικό προϊόν το οποίο θα κυκλοφορήσει αυτόν τον μήνα είναι ένα περιβάλλον εργασίας, χρήσιμο για πολλά χρόνια στην διάρκεια της παιδικής ζωής. Ένας ολόκληρος κόσμος. Υπάρχουν 2 σχολές στο εκπαιδευτικό λογισμικό. Η Σχολή ελευθέρας πλοήγησης και η σχολή του εκπαιδευτικού λογισμικού που συνδέεται με το βιβλίο. Η «λογομάθεια» είναι ένα λογισμικό ελεύθερης πλοήγησης. Ήταν η πρώτη κίνηση του ΙΕΛ μόλις δημιουργήθηκε και έγινε στην λογική της ενίσχυσης της γλώσσας μας για την σωστή εκμάθηση από τα Ελληνόπουλα.

Η επιτυχία στην «λογομάθεια» είναι ότι μπόρεσε να συνδυάσει την γλωσσική με την πολιτιστική εκπαίδευση μέσω της ιδέας του ηλεκτρονικού βραβείου. Έτσι πολλοί ζητούν την «λογομάθεια» περισσότερο για τα βραβεία της παρά για τα γλωσσικά μαθήματα. Όμως για να δημιουργηθούν οι συλλογές των βραβείων και να εξουσιοδοτηθεί ο χρήστης να τα βλέπει πρέπει οπωσδήποτε να περάσει από τα μαθήματα.

Τα γλωσσικά μαθήματα περιέχουν διδασκαλία, το λεγόμενο διδακτικό μέρος, και γλωσσικές ασκήσεις εκ των οποίων πολλές είναι σχεδιασμένες με παιχνιώδη τρόπο. Η ανάγκη της πληρότητας της ύλης καλύπτεται χάρη στα ηλεκτρονικά βιβλία στα οποία μπορεί να ανατρέξει κανείς άμεσα από οπουδήποτε χωρίς να αναγκασθεί να τα ξεφυλλίζει όπως συμβαίνει με τα συμβατικά βιβλία. Έτσι υπάρχουν ηλεκτρονικά βιβλία γραμματικής, συντακτικού, ορθογραφίας και λεξιλογίου που γράφτηκαν εξ' αρχής για την «λογομάθεια», ενώ τα ηλεκτρονικά βραβεία περιλαμβάνουν πίνακες ζωγραφικής από Έλληνες ζωγράφους, κομμάτια κλασικής μουσικής, κομμάτια Ελληνικής ποίησης, ένα σήριαλ εμπνευσμένο από την αργοναυτική εκστρατεία με βάση τα κείμενα του κ. Γεραλή, παραδοσιακά παιδικά τραγούδια, τα παραδοσιακά μας μουσικά όργανα, και τα Ελληνικά νησιά. Κερδίζοντας ένα νησί ο μαθητής στην συνέχεια της επιτυχημένης επίλυσης αριθμού ασκήσεων μπορεί να περιηγηθεί το νησί αυτό βλέποντας κάποια αξιοθέατα, μαθαίνοντας περισσότερες πληροφορίες για το νησί ή ακούγοντας τοπική μουσική.

Τα πολυμέσα μπορούν να προσφέρουν πρωτότυπες λύσεις που σαγηνεύουν τα παιδιά και αυξάνουν τον βαθμό απορρόφησης της γνώσης με δημιουργία ισχυρών εντυπώσεων. Επίσης δίνουν την δυνατότητα καλύτερης και παραστατικότερης ανάλυσης των περισσότερο πολύπλοκων εννοιών.

Γενικά, η επίδραση των διαλογικών πολυμέσων στην εκπαίδευση μπορεί να συνοψισθεί στα εξής α) αύξηση της μαθησιακής τάσης των μαθητών και αύξηση του ενδιαφέροντος για ενεργό εμπλοκή β) ιδιαίτερη χρησιμότητα για μαθητές με μικρή ικανότητα συγκέντρωσης γ) μετακίνηση της μέσης στάθμης μίας τάξης προς τα άνω.

Επομένως από τις τελευταίες τάξεις του δημοτικού σχολείου μπορεί και πρέπει να ισχυροποιηθεί η μάθηση της Ελληνικής γλώσσας με χρήση των Η/Υ. Ο συνδυασμός της διδασκαλίας με το παιχνίδι και την ηλεκτρονική επιβράβευση μπορεί να οδηγήσει σε μεγαλύτερη αποδοχή από τους μαθητές.

Μετά την «λογομάθεια», άρχισε ο σχεδιασμός της «φιλογλωσσίας» που έχει εντελώς διαφορετική τεχνοτροπία από την «λογομάθεια». Η «φιλογλωσσία» απευθύνεται σε αρχάριους ξένους που θέλουν να μάθουν Ελληνικά. Έχει επικοινωνιακή δομή, χρησιμοποιούνται διάλογοι από την καθημερινή ζωή και ο μαθη-

τής έχει τον πλήρη έλεγχο των διαλόγων δυνάμενος να τους ακούσει και να τους ξανακούσει, αλλά και να αντικαταστήσει έναν ήρωα με τον εαυτό του όταν νομίσει ότι έχει φθάσει σε ικανό βαθμό κατάκτησης των γλωσσικών τύπων. Η «φιλογλωσσία» περιέχει πλήθος μοντέρνων εργαλείων Γλωσσικής Τεχνολογίας που βοηθούν τον μαθητή. Είναι το προϊόν του ΙΕΛ που ολοκληρώνει τις περισσότερες σύγχρονες τεχνολογίες. Η «φιλογλωσσία» είναι ένα δυναμικό περιβάλλον το πρώτο στο είδος τους που περιέχει ανεξάντλητους πόρους γλωσσικής εκπαίδευσης. Έχει την δυνατότητα να προφέρει και να κλίνει οποιαδήποτε λέξη που κλίνεται με συνθετική φωνή, να γράφει με βάση το φωνητικό αλφάβητο οποιαδήποτε λέξη κ.ο.κ. Η «φιλογλωσσία» θα είναι στον Ευρωπαϊκό στίβο το προϊόν γλωσσικής εκπαίδευσης που θα ολοκληρώνει τις πιο προηγμένες τεχνολογίες γι' αυτό αξίζει να το παρατηρήσετε κατά την διάρκεια των επιδείξεων.

Ένα άλλο προϊόν του ΙΕΛ που δοκιμάζεται τώρα σε κάποια σχολεία είναι η «λογονόστηση» για την διδασκαλία της Ελληνικής ως δεύτερης μητρικής γλώσσας. Το προϊόν αυτό απευθύνεται σε παιδιά παλιννοστούντων από τις περιοχές του Πόντου. Έχει σαν γλώσσα υποστήριξης την ρωσική, δηλαδή το παιδί αν δεν καταλαβαίνει κάτι μπορεί να λάβει κάποιες επεξηγήσεις από τον Η/Υ προφορικές ή/και γραπτές στα ρωσικά. Περιλαμβάνει κυρίως μαθήματα ενισχυτικής διδασκαλίας με κείμενα και διαλόγους που έχουν να κάνουν με το περιβάλλον, την αγωγή υγείας, την κυκλοφοριακή αγωγή και τον πολιτισμό. Έτσι συνδυάζεται η γλωσσική εκπαίδευση με θέματα αγωγής. Το λογισμικό περιέχει και δίγλωσσο Ελληνορωσικό λεξικό.

Το Τμήμα Εκπαιδευτικής Τεχνολογίας που κατασκευάζει όλα αυτά τα προϊόντα, κατασκευάζει και την «λεξιπαιδεία», ένα πολύγλωσσο λεξικό πολυμέσων για παιδιά δημοτικού. Με βάση το Ελληνικό λήμμα το παιδί μπορεί να έχει προσπέλαση στον ορισμό, σε παραδείγματα χρήσης, σε σχόλια, σε φωτογραφίες ή σε κομμάτια video, αλλά και στην απόδοση μίας λέξης σε διάφορες Ευρωπαϊκές γλώσσες τόσο σε γραπτή όσο και σε προφορική μορφή.

Πέραν των εκπαιδευτικών προϊόντων το τμήμα εκπαιδευτικής τεχνολογίας έχει δημιουργήσει (σε συνεργασία με το τμήμα ηλεκτρονικής λεξικογραφίας), ένα σύστημα διόρθωσης ορθογραφικών και συντακτικών λαθών που λειτουργεί κάτω από περιβάλλον «word» και βασίζεται στο μορφολογικό λεξικό του Ινστιτούτου. Το μορφολογικό αυτό λεξικό περιέχει κωδικοποιημένα 62.000 λήμματα της νέας ελληνικής με κατάλληλους κανόνες παραγωγής ώστε να είναι δυνατή η γένεση και η αναγνώριση περίπου 1.500.000 λεκτικών τύπων. Το λεξικό αυτό είναι ένα μεγάλο έργο υποδομής που έγινε τα τελευταία χρόνια και μπορεί να στηρίξει πέραν του ορθογραφικού διορθωτή και άλλα προϊόντα του Ινστιτούτου από συστήματα κειμενικής ανάλυσης μέχρι συστήματα ανάκτησης πληροφορίας και σύνθεσης φωνής μια και περιέχει την βασική πληροφορία για την παραγωγή τόσο μεγάλου αριθμού λεκτικών τύπων καθώς και την πληροφορία για το τι μέρος του λόγου είναι η κάθε λέξη.

Πέραν από το μορφολογικό λεξικό στου οποίου τον σχεδιασμό μετείχε και συνεχώς εμπλουτίζει με γλωσσικό υλικό, το τμήμα Ηλεκτρονικής Λεξικογραφίας δημιούργησε ένα υπολογιστικό λεξικό της ελληνικής των 20.000 λημμάτων με βάση τα πρότυπα που επεξεργάστηκε το Κοινοτικό πρόγραμμα PAROLE. Το λεξικό αυτό είναι ιδιαίτερα χρήσιμη υποδομή και μπορεί να υποστηρίξει διάφορα προϊόντα γλωσσικής τεχνολογίας από πλευράς ελληνικής γλώσσας.

Το Τμήμα Ηλεκτρονικής Λεξικογραφίας είναι το κατεξοχήν τμήμα του Ινστιτούτου που εργάζεται για θέματα δημιουργίας γλωσσικών πόρων. Εκτός των λεξικών συγκεντρώνει και κείμενα αντιπροσωπευτικά του τρόπου που γράφεται η γλώσσα μας στην εποχή μας. Έτσι αυτήν την στιγμή υπάρχουν δύο μεγάλα δείγματα χρήσης της γλώσσας μας σε ηλεκτρονική μορφή το σώμα κειμένων ΤΥΠΟΣ με κείμενα κυρίως από εφημερίδες και περιοδικά και το σώμα κειμένων ΟΡΟΣΗΜΟ με κείμενα από την χρήση της γλώσσας μας σε περιβάλλοντα εντάσεως όρων. Θα δείτε στην έκθεση την περιπέτεια ενός κειμένου με διάφορα στάδια επεξεργασίας.

Τα σώματα αυτά κειμένων θα αρχίσουν να διατίθενται σύντομα τόσο μέσω ειδικευμένων CD's όσο και μέσω του διαδικτύου σε εκείνους τους ερευνητές που ενδιαφέρονται να μελετήσουν την γλώσσα μας. Έχουν κατασκευασθεί πλήθος εργαλείων που βοηθούν στην ανάκτηση και στην γλωσσική έρευνα. Η σύγχρονη άποψη είναι ότι τα προϊόντα Γλωσσικής Τεχνολογίας πρέπει να βασίζονται στην γλώσσα όπως γράφεται

και μιλιέται και για τον λόγο αυτό τα σώματα κειμένων που συλλέγονται από το ΙΕΛ είναι πολύτιμα γιατί αποτελούν πραγματώσεις του σύγχρονου Ελληνικού λόγου. Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά όσους συνέβαλλαν στις συλλογές αυτές κειμένων και ιδιαίτερα τις εφημερίδες ΒΗΜΑ & ΕΛΕΥΘΕΡΟΤΥΠΙΑ καθώς και τους πολυάριθμους εκδοτικούς οίκους, και τους ερευνητές και καθηγητές που μας προσέφεραν τις σημειώσεις τους για να ενισχύσουν τις συλλογές εντάσεως όρων στο πλαίσιο του σώματος κειμένων ΟΡΟΣΗΜΟ.

Το τμήμα Τεχνολογίας Φωνής αναπτύσσει το σύστημα σύνθεσης φωνής με το όνομα “εκφωνητής”. Το σύστημα αυτό δίνει την δυνατότητα στον Η/Υ να εκφωνήσει κάτω από το παραθυρικό περιβάλλον οποιοδήποτε κείμενο που ευρίσκεται σε ηλεκτρονική μορφή. Δουλεύουμε πολλά χρόνια στην σύνθεση φωνής και προσπαθούμε συνεχώς να βελτιώσουμε την ποιότητα. Έχουμε χρησιμοποιήσει δύο διαφορετικές τεχνολογίες μέχρι στιγμής, στις οποίες υπάρχουν εργαστηριακά πρότυπα.

Θα ακούσατε τον συνθέτη μας να σας μιλάει όταν λίγο αργότερα επισκεφτείτε την έκθεσή μας. Πρόσφατα άρχισε να δημιουργείται ένα περιβάλλον στον προσωπικό υπολογιστή που συμπαρίσταται σε άτομα με προβλήματα όρασης εκφωνώντας σε κάθε στιγμή την κατάσταση του Η/Υ. Έτσι ο χρήστης που δεν μπορεί να παρακολουθήσει την εξέλιξη των διαδικασιών από την οθόνη βοηθιέται μέσω της συνθετικής ομιλίας.

Το Τμήμα Φωνής αναπτύσσει επίσης τεχνολογία ανάλυσης σημάτων στον χώρο του χρόνου και της συχνότητας και στον τομέα αυτό θα δείτε συγκεκριμένα προϊόντα μας στην έκθεση. Επίσης θα δείτε μία πλατφόρμα εξαγωγής θορύβου από φωνή (αποθορυβοποίησης), που μπορεί να είναι μία ιδιαίτερα χρήσιμη τεχνολογία. Στο ΙΕΛ διαθέτουμε όλες τις τεχνολογίες συμπίεσης φωνής. Στην αναγνώριση φωνής επίσης έχουμε κάνει πολλά βήματα. Διαθέτουμε και τις δύο τεχνολογίες που χρησιμοποιούνται σήμερα στην αναγνώριση φωνής δηλαδή τα κρυμμένα Μαρκοβιανά μοντέλα και την δυναμική χρονική αναμόρφωση και υπάρχουν στο ΙΕΛ διάφορα μικρά πειραματικά συστήματα αναγνώρισης φωνής. Δυστυχώς μέχρι σήμερα δεν μπόρεσε το Ινστιτούτο να κερδίσει κατάλληλο πρόγραμμα για να κατασκευάσει κάποιο μεγάλο σύστημα αναγνώρισης συνεχούς λόγου. Στον τομέα αυτό η εκμάθηση της μηχανής είναι το κλειδί και στοιχίζει ακριβά.

Το τμήμα Μηχανικής Μετάφρασης που ήταν εξαιρετικά δραστήριο παλαιότερα έχει περιορίσει την δραστηριότητα του γιατί επίσης δεν υπήρξε τα τελευταία χρόνια κάποιο πρόγραμμα. Όπως είναι γνωστό ήταν παλαιότερα ισχυρά δραστηριοποιημένο και ανέπτυξε τις υπολογιστικές γραμματικές του EUROTRA, ενός μεγάλου ευρωπαϊκού προγράμματος για την μετάφραση των Ευρωπαϊκών γλωσσών που έληξε το 1993. Τώρα το τμήμα ασχολείται πιο πολύ με την συντήρηση αυτών των γραμματικών. Πρόκειται για πολύτιμα κομμάτια λογισμικού που δεν θέλουμε να εγκαταλείψουμε εν/όψει των προθέσεων αναβίωσης των δραστηριοτήτων έρευνας στην μετάφραση εκ μέρους της ΕΕ στο 5ο Πρόγραμμα Πλαίσιο.

Πολύ σημαντικές επίσης είναι οι δραστηριότητες του τμήματος Γλωσσικών Εφαρμογών Γραφείου. Το τμήμα αυτό εργάζεται αναπτύσσοντας τεχνολογίες που θα είναι χρήσιμες στο περιβάλλον γραφείου του μέλλοντος που θα είναι κατ’ ανάγκη πολυγλωσσικό και δικτυωμένο. Το γραφείο αυτό χρειάζεται λογισμικό χειρισμού πολυγλωσσικών κειμένων, λογισμικό ανάκτησης και εξαγωγής πληροφορίας από κείμενα καθώς και εργαλεία για την υποβοήθηση του μεταφραστικού έργου. Το τμήμα έχει κατασκευάσει μία από τις πρώτες μεταφραστικές μνήμες διεθνώς. Ήταν μία μεγάλη επιτυχία και πρόκειται να οδηγήσει άμεσα σε προϊόν με διεθνή αγορά. Οι μεταφραστικές μνήμες είναι περιβάλλοντα που αξιοποιούν τις ποιοτικές μεταφράσεις ενός μεταφραστή ή ενός μεταφραστικού γραφείου και χάρη σε ειδικά εργαλεία δίνουν την δυνατότητα κινητοποίησης έτοιμων μεταφράσεων για υποστήριξη του έργου του μεταφραστή όταν θέλει να μεταφράσει κάτι καινούργιο. Ειδικά στην περίπτωση που τα κείμενα προς μετάφραση περιέχουν στερεότυπα, όπως είναι η περίπτωση των διοικητικών γλωσσών σαν αυτή που χρησιμοποιεί η ΕΕ, μία μεταφραστική μνήμη διαθέτει την δυνατότητα μετάφρασης των κειμένων και μπορεί να θεωρηθεί σαν βασική τεχνολογία για πρόχειρη μετάφραση. Ας μην ξεχνάμε ότι στην μετάφραση διοικητικών γλωσσών η ποιότητα είναι ταυτόσημη με την πιστότητα για καλύτερη συνεννόηση. Εκτός από την μεταφραστική μνήμη οι ερευνητές του τμήματος Γλωσσικών Εφαρμογών

Γραφείου θα σας δείξουν ένα σύστημα αυτόματης εξαγωγής πολυλεκτικών όρων από ένα κείμενο καθώς και ένα σύστημα ανάκτησης πληροφορίας από βάσεις δεδομένων που επιτρέπει λάθη πληκτρολόγησης μέχρι δύο ανά λέξη.

Η ανάπτυξη των προϊόντων μας ακολουθεί ένα κύκλο με συγκεκριμένο χρονοδιάγραμμα. Πολλά θα είναι διαθέσιμα άμεσα, άλλα λίγο αργότερα μέσα στο 1998 που είναι για το ΙΕΛ η χρονιά των προϊόντων του όπως το 1997 ήταν η χρονιά της κτηριακής υποδομής.

Στο Τμήμα Συνδέσμου εντάσσονται όπως σας είπα οι οριζόντιες δραστηριότητες του ΙΕΛ. Εκτός από το ότι υποστηρίζει τα άλλα τμήματα αναζητώντας την επίκαιρη πληροφορία που τα αφορά, ιδιαίτερα σε θέματα διαγωνισμών και προγραμμάτων, υποστηρίζει εκδοτικά το Ινστιτούτο, με την εκτύπωση και διανομή των κειμένων εργασίας, της «Λογοπλοήγησης», του ειδικού ενημερωτικού δελτίου που εκδίδει το ΙΕΛ σε θέματα Γλωσσικής Τεχνολογίας, και το οποίο φιλοδοξεί να εξελιχθεί σε ευρωπαϊκό επιστημονικό περιοδικό στον χώρο αυτό. Επίσης πάντα στα εκδοτικά θέματα το τμήμα ετοιμάζει ένα ειδικό τόμο σε θέματα μηχανικής μετάφρασης. Το Τμήμα Συνδέσμου επίσης συντηρεί την web σελίδα του ΙΕΛ και οργανώνει πλήθος ημερίδων 2- 4 ετησίως στα θέματα της Γλωσσικής Τεχνολογίας για ευαισθητοποίηση και πληροφόρηση.

Συμμετέχει επίσης στην χαρτογράφηση του ευρωπαϊκού χώρου σε θέματα Γλωσσικής Τεχνολογίας και συνεργάζεται με την Κοινότητα και το Εθνικό Κέντρο Τεκμηρίωσης για την διάδοση της κοινοτικής πληροφορίας στον ελληνικό χώρο. Στο Τμήμα Συνδέσμου λειτουργεί το γραφείο EUROMAT που προσφέρει δωρεάν υπηρεσίες πρόχειρης μετάφρασης σε όλα τα Ελληνικά Υπουργεία όταν υπάρχει ανάγκη ταχύτατης και άμεσης μετάφρασης κειμένων κυρίως από τα Αγγλικά στα Ελληνικά. Χρησιμοποιείται το σύστημα μετάφρασης της κοινότητας στο οποίο έχει ενταχθεί η Ελληνική γλώσσα σύμφωνα με την πολιτική της ΓΓΕΤ από το 1989. Οι μεταφράσεις που γίνονται για τα Ελληνικά υπουργεία ανατροφοδοτούν το σύστημα το οποίο βελτιώνεται συνεχώς με τον τρόπο αυτό. Η ανατροφοδότηση του συστήματος γίνεται χάρη στην εργασία της ομάδας του EUROMAT και την συνεργασία με την ομάδα του Λουξεμβούργου.

Το Παράρτημα Ξάνθης του ΙΕΛ ιδρύθηκε και λειτουργεί από τις αρχές του 1996 με απόφαση του Επιστημονικού Συμβουλίου του Ινστιτούτου για να εξυπηρετήσει τους παρακάτω σκοπούς:

- Υποστήριξη στα ακριτικά μονοθέσια σχολεία της περιοχής που επιθυμούν την εισαγωγή εκπαιδευτικού λογισμικού και των δικτυακών εφαρμογών στην εκπαιδευτική διαδικασία.

- Παραγωγή εκπαιδευτικών λογισμικών με γλωσσική και πολιτισμική διάσταση στην Β. Ελλάδα σε συνεργασία με το Πανεπιστήμιο Θράκης και άλλους ενδιαφερόμενους οργανισμούς.

- Κατασκευή ηλεκτρονικών λεξικών από και προς τις Βαλκανικές γλώσσες.

- Κατασκευή υπολογιστικών γραμματικών χρήσιμων στη μετάφραση από και προς τις Βαλκανικές γλώσσες και της συγκριτικής ανάλυσης των γλωσσών αυτών σε σχέση με την Ελληνική για θέματα μετάφρασης.

- Μελέτη των αναγκών των Βαλκανικών χωρών σε Γλωσσική Τεχνολογία και οργάνωση ημερίδων ευαισθητοποίησης στις γειτονικές χώρες ως προς τη χρησιμότητα της τεχνολογίας αυτής.

Κάποιοι από τους στόχους αυτούς έχουν αρχίσει να υλοποιούνται.

Κυρία Υπουργέ, Κύριε Γενικέ,

το ΙΕΛ είναι ένας οργανισμός καλά οργανωμένος που ευτύχησε χάρη στην υποστήριξή σας και αυτήν της ΕΕ να έχει στην διάθεσή του αυτές τις σύγχρονες εγκαταστάσεις επιλύοντας κατά τον καλύτερο τρόπο ένα χρόνιο και οξύ πρόβλημα στέγασης. Θέλουμε να σας ευχαριστήσουμε θερμά για την τόσο καλή λύση που δώσατε. Μας επέτρεψε να οργανωθούμε καλύτερα και να αυξήσουμε την παραγωγικότητά μας. Το ΙΕΛ αποτελεί έναν μηχανισμό στα χέρια του Υπουργείο Ανάπτυξης που μπορεί να συμβάλλει στην ανάπτυξη στον χώρο της πληροφορικής προσφέροντας πρωτότυπες και καινοτομικές λύσεις με προστιθέμενη αξία που μπορεί να σχεδιάσει σε συνεργασία με τον ιδιωτικό τομέα και τους λοιπούς δη-

μόσιους οργανισμούς.

Δράττομαι της ευκαιρίας να ευχαριστήσω εκ μέρους του Επιστημονικού Συμβουλίου του ΙΕΛ όλους και όλες, όσοι συμπαραστάθηκαν στο έργο του ΙΕΛ. Κατά κύριο λόγο τα μέλη της πολιτικής ηγεσίας της Γενικής Γραμματείας και τους υπευθύνους τόσο της Γενικής Γραμματείας Έρευνας & Τεχνολογίας όσο και της Ευρωπαϊκής Ένωσης που σε δύσκολες στιγμές μας βοήθησαν.

Θα ήθελα να ευχαριστήσω θερμά επίσης τους στενούς συνεργάτες μου αναγνωρίζοντας την προσπάθειά τους για να πετύχει αυτός ο οργανισμός. Ιδιαίτερα θα ήθελα να αναφερθώ στους υπευθύνους των τμημάτων του ΙΕΛ κ.κ. Μ. Γαβριηλίδου, Γ. Βολτή, Σ. Φούρλα, Σ. Πιπερίδη, Σ. Μπακαμίδη, Γ. Σταϊνχάουερ καθώς και τον υπεύθυνο του παραρτήματος του ΙΕΛ Ξάνθης κ. Ν. Χατζηγεωργίου, που από το πρωί μέχρι αργά το βράδυ δεν φείδονται κόπου για να δημιουργηθούν τα προϊόντα του ΙΕΛ. Θα ήθελα επίσης να ευχαριστήσω θερμά τα μέλη του Επιστημονικού Συμβουλίου του Ινστιτούτου κ.κ. Μπαμπινιώτη, Κοπιδάκη, Στρίντζη και Σπυράκη για την τόσο εποικοδομητική συμμετοχή τους στην διοίκηση του Ινστιτούτου και για τις καλοπροαίρετες συμβουλές τους στα διάφορα θέματα του ΙΕΛ.

Θέλω να σας παρακαλέσω να μας ακολουθήσετε στο σημερινό πρόγραμμά μας. Θα ανεβούμε στον ημιώροφο όπου η κ. Παπανδρέου θα κόψει συμβολικά την κορδέλα εγκαινιάζοντας τις νέες εγκαταστάσεις και συγχρόνως την έκθεση μας όπου θα μπορέσετε να δείτε τα προϊόντα μας.

ΙΙ. Το Κοινοτικό Έργο "Euromap"

Τί είναι το EUROMAP

Το EUROMAP είναι η πρώτη έρευνα μεγάλης κλίμακας που έχει διεξαχθεί για τις ανάγκες της αγοράς για προϊόντα Γλωσσικής Τεχνολογίας στην Ευρώπη. Με ευκαιρία την έρευνα αυτή μελετήθηκαν οι πολιτικές, οι ερευνητικές και αναπτυξιακές δραστηριότητες, οι προμηθευτές και οι δυνητικοί χρήστες πληροφοριακών και επικοινωνιακών συστημάτων με ενσωματωμένη Γλωσσική Τεχνολογία στις 15 χώρες της Ευρωπαϊκής Ένωσης μαζί με την Νορβηγία και την Ισλανδία.

Χρησιμοποιώντας τα αναλυτικά αποτελέσματα από τους διάφορους ερευνητικούς στόχους, το EUROMAP, όπως υποδηλώνει και το όνομά του, σχεδίασε έναν χάρτη ρεαλιστικών δυνατοτήτων αγοράς για τις εφαρμογές της Γλωσσικής Τεχνολογίας στα επόμενα χρόνια. Η πληροφορία αυτού του είδους έχει βοηθήσει στον σχεδιασμό της θεματολογίας της ευρωπαϊκής έρευνας στην Γλωσσική Τεχνολογία, ειδικότερα για το Πέμπτο Πρόγραμμα Πλαίσιο που είναι υπεύθυνο για κοινά ερευνητικά και αναπτυξιακά (Ε&Α) προγράμματα στην Ευρώπη. Θα προσφέρει επίσης στις εθνικές αρχές μια εικόνα των δυνατοτήτων και απαιτήσεών τους σε αυτόν τον τομέα.

Το EUROMAP ξεκίνησε και χρηματοδοτείται κυρίως από την 13η Γενική Διεύθυνση της Ευρωπαϊκής Επιτροπής. Η έρευνα έχει διεξαχθεί από ένα δίκτυο οργανισμών γνωστών ως Εθνικά Εστιακά Σημεία, τα οποία λειτουργούν σε κάθε χώρα (το ΙΕΛ αποτελεί το σημείο αυτό στην Ελλάδα). Σκοπός τους ήταν να αξιολογήσουν τις δυνατότητες της τοπικής αγοράς σχετικά με τη Γλωσσική Τεχνολογία και να διατυπώσουν σχετικές συστάσεις βάσει των ευρημάτων τους σε εθνικό επίπεδο.

Η συνέχιση του προγράμματος έχει πρωταρχικό στόχο την ευρύτερη και συστηματικότερη διάδοση των αποτελεσμάτων του EUROMAP, καθώς και την γενικότερη ενημέρωση για τις δυνατότητες της Γλωσσικής Τεχνολογίας και τις προβλεπόμενες στο Πέμπτο Πρόγραμμα-Πλαίσιο σχετικές δράσεις.

Στο πλαίσιο της δεύτερης φάσης του προγράμματος

(1998-99) έχει ξεκινήσει ένας αριθμός δραστηριοτήτων διάχυσης πληροφοριών (ημερίδες, ενημερωτικά φυλλάδια, γραφείο ενημέρωσης, δελτία τύπου, συμμετοχή σε εκθέσεις κ.λπ.), με στόχο την δημιουργία αποτελεσματικότερου διαλόγου και συνεργασίας ανάμεσα στην ερευνητική κοινότητα, τους προμηθευτές και τους χρήστες τόσο μέσα σε κάθε χώρα όσο και μεταξύ των χωρών. Ελπίζουμε ότι οι αναφορές "΄Ερευνα για την Εθνική Πολιτική στην Γλωσσική Τεχνολογία" και "Εθνική ΄Εκθεση για την Γλωσσική Τεχνολογία" που εξέδωσε το ΙΕΛ στο πλαίσιο του EUROMAP για την Ελλάδα θα συμβάλουν στην καλύτερη κατανόηση των δυνατοτήτων ανάπτυξης της αγοράς Γλωσσικής Τεχνολογίας.

Με την οργάνωση ενημερωτικών δραστηριοτήτων το EUROMAP θα προωθήσει την συμμετοχή των εθνικών φορέων σε ευρωπαϊκά προγράμματα (Πέμπτο Πρόγραμμα-Πλαίσιο) και θα ενισχύσει τις διακρατικές δραστηριότητες με σκοπό την καλύτερη συνεργασία ανάμεσα σε ερευνητές, εταιρείες ανάπτυξης και πιθανούς χρήστες στον χώρο της Γλωσσικής Τεχνολογίας.

Για την αποτελεσματική ολοκλήρωση του έργου αυτού δημιουργήθηκε ένα νέο γραφείο πληροφόρησης για την Γλωσσική Τεχνολογία, τις εφαρμογές της καθώς και για τις τεχνικές που συμβάλλουν με δυνατότητες φυσικής γλώσσας στις σημερινές τεχνολογίες πληροφορικής και επικοινωνιών. Το γραφείο αυτό αποτελεί κέντρο ενημέρωσης για όλους τους φορείς που ενεργοποιούνται στον τομέα της προηγμένης τεχνολογίας ηλεκτρονικών υπολογιστών στην Ελλάδα. Προωθεί επίσης την ευρύτερη ενημέρωση σχετικά με τις δυνατότητες της αγοράς στο χώρο της Ευρωπαϊκής Γλωσσικής Τεχνολογίας γενικότερα.

Το γραφείο ενημέρωσης του EUROMAP προσφέρει στους ερευνητές, τους προμηθευτές και τους δυνητικούς χρήστες της Γλωσσικής Τεχνολογίας ένα πρωτότυπο κέντρο εθνικής και ευρωπαϊκής ενημέρωσης σχετικά με οργανισμούς, προϊόντα, προγράμματα και γεγονότα σ'αυτόν τον αναπτυσσόμενο τομέα.

Μέσω του γραφείου αυτού διοχετεύονται οι πιο σύγχρονες πληροφορίες σχετικά με υπάρχοντα εθνικά προγράμματα, προϊόντα, προμηθευτές, χρήστες και εθνικούς φορείς που έχουν δραστηριότητες στον τομέα της Γλωσσικής Τεχνολογίας. Το γραφείο πληροφόρησης αποτελεί εθνικό κρίκο για όλες τις ενέργειες πληροφόρησης από πλευράς της Ευρωπαϊκής Επιτροπής σε θέματα Γλωσσικής Τεχνολογίας. Κυρίως, λειτουργεί ως πηγή πληροφόρησης για την γραμμή δράσης "Τεχνολογία Ανθρώπινης Γλώσσας" του προγράμματος Τεχνολογίας της Κοινωνίας των Πληροφοριών στο Πέμπτο Ευρωπαϊκό Πρόγραμμα Πλαίσιο για την ΄Ερευνα και την Τεχνολογία στην Ευρώπη, το οποίο χρηματοδοτείται από την Ευρωπαϊκή Επιτροπή. Από την υπηρεσία αυτή θα επωφεληθούν ερευνητικά και αναπτυξιακά κέντρα, μικρές και μεγάλες επιχειρήσεις και άλλοι οργανισμοί οι οποίοι θέλουν να αναπτύξουν συνεργασίες με σκοπό την υποβολή προτάσεων για χρηματοδότηση στο πλαίσιο αυτού του προγράμματος.

Όσοι ενδιαφέρονται για περισσότερες πληροφορίες σχετικά με το πώς η Γλωσσική Τεχνολογία μπορεί να τους διευκολύνει στο χώρο εργασίας τους ή στον ελεύθερο χρόνο τους μπορούν να απευθύνονται στην υπεύθυνη πληροφόρησης στην Ελλάδα, η οποία θα τους παρέχει τις συγκεκριμένες πληροφορίες σχετικά με αυτήν την πρωτότυπη διάσταση της μελλοντικής κοινωνίας των πληροφορίων στην Ελλάδα.

EUROMAP Γραφείο Ενημέρωσης για την Γλωσσική Τεχνολογία:
Υπεύθυνη: Δέσποινα Σκούταρη
Τηλ:6800952, 6, 9
Fax: 6856794
email:euromap@ilsp.gr
http://www.euromap.gr

III. Περιλήψεις εισηγήσεων ημερίδας με θέμα «Ανάκτηση και Εξαγωγή Πληροφορίας» / *Workshop on "Information Retrieval and Information Extraction" - 11 Απριλίου 1998*

1. Information Retrieval from modern Greek text collections using SMART

Professor Theodor Kalamboukis and Simos Nikolaidis
Department of Informatics
Athens University of Economics & Business
76 Patission St., 104 34 Athens, Hellas
E-mail: tzk@aueb.

Abstract

We present results of retrieval from two collections in modern Greek using SMART. Experiment with different stemming algorithms reveal that further research is needed for sub-clustering the indexing terms after stemming. This is due to the fact that modern Greek is a rich language both in inflectional and derivational suffixes.

Ανάκτηση Πληροφοριών από Συλλογές Ελληνικών Κειμένων με το SMART

Καθηγητής Θεόδωρος Καλαμπούκης και
Σίμος Νικολαίδης
Τμήμα Πληροφορικής
Οικονομικό Πανεπιστήμιο Αθηνών
Πατησίων 76, 104 34 Αθήνα

1. Το σύστημα SMART

Το σύστημα ανάκτησης πληροφοριών SMART (Storage Management and Retrieval) αναπτύχθηκε στο πανεπιστήμιο Cornell από την δεκαετία του 60 και βασίζεται στο μοντέλο διανυσματικού χώρου, που οφείλεται στον G. Salton [1,2,3]. Βασικός σκοπός του συστήματος είναι να παράσχει ένα πλαίσιο για έρευνα στο πεδίο της ανάκτησης πληροφοριών.

Το SMART έχει τα πλεονεκτήματα και μειονεκτήματα που έχει κάθε ακαδημαϊκό ερευνητικό λογισμικό. Εχει σχεδιαστεί να είναι πολύ ευέλικτο αλλά δεν παρέχει την βέλτιστη λύση για κάθε συγκεκριμένη διαδικασία. Ο κώδικας είναι σχετικά απλός και εκτελείται στα περισσότερα UNIX συστήματα με λίγες μετατροπές. Η παρούσα έκδοση του SMART v.11 είναι περισσότερο ευέλικτη και μπορεί να τροποποιηθεί ευκολότερα, αλλά διατηρεί το βασικό μειονέκτημα του συστήματος, την έλλειψη φιλικής διεπαφής προς τον μη ειδικό χρήστη.
Αν και ο κύριος στόχος του SMART είναι η έρευνα, το σύστημα απευθύνεται εκτός από τους ερευνητές, στους διαχειριστές βάσεων δεδομένων αλλά και στους απλούς χρήστες.

Τα βασικά χαρακτηριστικά του συστήματος είναι: **Το μέγεθος** (350 πηγαία αρχεία), **η απλότητα**, (παρέχει ομοιόμορφη προσπέλαση στα αρχεία δεδομένων του UNIX), **η διαλογικότητα, η ευελιξία** (είναι πλήρως παραμετρικό), και **η ταχύτητα** (για συλλογές της τάξης των 50 MB η διαδικασία κατασκευής απλών ευρετηρίων (indexing) μπορεί να γίνει με 1 – 5 MB/min).

Οι απαιτήσεις του συστήματος σε δίσκους εξαρτάται από την συλλογή. Τα ευρετήρια μιας συλλογής καταλαμβάνουν περίπου το 40% του χώρου της αρχικής συλλογής των κειμένων.

Ο χρήστης έχει τη δυνατότητα να παρακολουθεί (debugging) τις τιμές κάποιων παραμέτρων, κατά την διάρκεια της εκτέλεσης του προγράμματος. Τέλος ο χρήστης έχει τη δυνατότητα να επιλέξει κάποια από τα ανακτηθέντα κείμενα και με τους όρους αυτών των κειμένων να εμπλουτίσει το αρχικό ερώτημα προκειμένου να ανακτήσει περισσότερα σχετικά κείμενα (Relevance Feedback).

Παρακάτω δίνεται μια σύντομη γενική περιγραφή των μεθόδων προσπέλασης και ανάκτησης με το SMART.

2. Τύποι πληροφορίας και επίπεδα του SMART

Τα κείμενα μιας συλλογής περιέχουν αρκετά είδη πληροφοριών όπως π.χ. ημερομηνίες, ονόματα συγγραφέων, αποστολέα ή παραλήπτη (αν πρόκειται για ηλεκτρονικά μηνύματα), κατάταξη του κειμένου σε μια ιεραρχική λίστα κ.α. Για κάθε ένα από τα παραπάνω είδη πληροφορίας υπάρχει η δυνατότητα διαφορετικής αντιμετώπισης. Αυτό οδηγεί στον ορισμό τύ-

πων ομαδοποίησης των πληροφοριών *(ctypes, classification types)*. Κάθε όρος (concept) του κειμένου ανήκει σε ένα ctype.

Η κατασκευή των ευρετηρίων γίνεται αυτόματα και για κάθε κείμενο της συλλογής δημιουργείται ένα διάνυσμα (υπογραφή) που περιέχει τους όρους του κειμένου. Η αναπαράσταση αυτή αποτελείται από μια λίστα του τύπου
(concept, ctype of concept, weight).

Ο χρήστης υποβάλλει ένα ερώτημα σε φυσική γλώσσα, το οποίο αντικαθίσταται από την παράστασή του, που αποτελείται από τους όρους και τα βάρη τους. Για κάθε κείμενο υπολογίζεται το μέτρο της ομοιότητας (similarity measure) του κειμένου με το ερώτημα και επιστρέφονται τα κείμενα της απάντησης ταξινομημένα σε φθίνουσα σειρά ως προς το μέτρο αυτό.

Όλες οι διαδικασίες του SMART εντάσσονται σε τέσσερα λογικά επίπεδα. Τα επίπεδα αυτά από το υψηλότερο προς το χαμηλότερο είναι:

- *Επίπεδο αίτησης (ερωτήματος) χρήστη (user request level).*
- *Επίπεδο υλοποίησης διεργασιών (Task implementation).*
- *Επίπεδο προσπέλασης σε αντικείμενα (object access level).*
- *Επίπεδο προσπέλασης στη βάση δεδομένων (database access level).*

Ο χρήστης υποβάλλει ένα αίτημα στο υψηλότερο επίπεδο. Το σύστημα αποφασίζει ποιες διεργασίες χρειάζεται να εκτελεστούν, ώστε να ικανοποιηθεί το ερώτημα του χρήστη. Για παράδειγμα στη περίπτωση μιας λειτουργίας ανάκτησης θα εκτελεστούν διεργασίες ανάλυσης (indexing) του ερωτήματος, ανάκτησης των κειμένων, παρουσίασης των ανακτηθέντων κειμένων στο χρήστη και πιθανόν διεργασίες διεύρυνσης του ερωτήματος με νέους όρους (query expansion).

Οι διαδικασίες στο δεύτερο επίπεδο, είναι υπεύθυνες για την εκτέλεση των παραπάνω διεργασιών. Στο επίπεδο αυτό είναι δυνατό να προσπελάσουμε στα δεδομένα της συλλογής μέσω συσχεσιακών αρχείων-αντικειμένων *(relational file objects)*.

Το τρίτο επίπεδο αποτελείται από τις διαδικασίες υπεύθυνες για προσπέλαση στα συσχεσιακά αρχεία-αντικείμενα και τις διαδικασίες ανάγνωσης των αρχείων προδιαγραφών (specification file). Οι διαδικασίες προσπέλασης σε συσχεσιακά αρχεία χρησιμοποιούν διαδικασίες του UNIX για προσπέλαση σε βάσεις δεδομένων.

Το τέταρτο επίπεδο είναι το χαμηλότερο λογικό επίπεδο του SMART και οι διαδικασίες του είναι άγνωστες στα δύο πρώτα επίπεδα. Οι διαδικασίες του επιπέδου αυτού είναι υπεύθυνες για τις μεθόδους αποθήκευσης των συσχεσιακών αρχείων και είναι συνδεδεμένες με τις διαδικασίες του τρίτου επιπέδου.

Τα αρχεία προδιαγραφών αποτελούν τη καρδιά της ευελιξίας του SMART. Όλες οι παράμετροι συμπεριλαμβανομένων και των διαδικασιών που θα εκτελεστούν περιέχονται στα αρχεία αυτά. Τέτοιες παράμετροι καθορίζουν για παράδειγμα τους τύπους (ctype) των όρων, την απόδοση βαρών στους όρους, τη μορφή των αποτελεσμάτων και αξιολόγησή τους και διαδικασίες αποκοπής καταλήξεων ή αφαίρεσης τετριμμένων λέξεων.

Στο specification_file ο διαχειριστής του συστήματος μπορεί να δώσει τιμές στις παραμέτρους ή μπορεί να μεταβάλλει τις προκαθορισμένες τιμές, ώστε αυτές να ανταποκρίνονται στις ιδιαιτερότητες της συλλογής.

3. Ανάκτηση Ελληνικών κειμένων με το SMART

Στο σύστημα SMART, προστέθηκαν ή τροποποιήθηκαν διαδικασίες, στο επίπεδο υλοποίησης διεργασιών, έτσι ώστε να είναι δυνατόν να εκτελεστούν και να αξιολογηθούν διαδικασίες ανάκτησης ελληνικών κειμένων, όπως για παράδειγμα προγράμματα αναγνώρισης ελληνικών χαρακτήρων, αφαίρεσης τετριμμένων λέξεων (stop words), αποκοπής καταλήξεων (stemming) και τροποποιήθηκαν όπου χρειάστηκε οι διαδικασίες συντακτικής ανάλυσης (parsing) και εκτύπωσης. Τέλος μικρές αλλαγές έγιναν σε ορισμένες διαδικασίες του επιπέδου προσπέλασης σε αντικείμενα (object access level), όπως για παράδειγμα στην προσπέλαση του δημιουργούμενου λεξικού (dictionary).

Για να γίνει οποιαδήποτε έρευνα αξιολόγησης αλλά

και βελτίωσης της ανάκτησης κειμένων στην ελληνική γλώσσα είναι απαραίτητη η δημιουργία πειραματικών συλλογών (test collections). Εδώ χρησιμοποιήθηκαν δύο πειραματικές συλλογές:

Η πρώτη συλλογή (COMPUTERS) αποτελείται από 624 κείμενα μεγέθους από 0,4 έως 15 KB από βιβλία πληροφορικής των καθηγητών του Ο.Π.Α. Θ. Καλαμπούκη και Ι. Κάβουρα που ήταν διαθέσιμα σε ηλεκτρονική μορφή. Η βάση έχει μέγεθος 3MB.

Η δεύτερη συλλογή (EXPRESS) αποτελείται από 1200 κείμενα μεγέθους από 0,3 – 6,5 KB της οικονομικής εφημερίδας ΕΞΠΡΕΣ, για τη χρονική περίοδο από τον Ιούνιο του 1997 ως τον Φεβρουάριο του 1998 και αφορούν ειδήσεις σε θέματα οικονομίας, πολιτικής και επενδύσεων. Το μέγεθος της βάσης προς το παρόν είναι 1,6 MB, αλλά συνεχίζει να εμπλουτίζεται με νέα κείμενα και σε πλήρη ανάπτυξη θα περιλαμβάνει περίπου 4000 κείμενα με συνολικό μέγεθος περίπου 6 MB.

Για την πρώτη συλλογή και για την αξιολόγηση διαφορετικών μεθόδων ανάκτησης και διαφορετικών αλγορίθμων αποκοπής καταλήξεων (stemming), εκτελέστηκαν διαδικασίες ανάκτησης για 29 προκαθορισμένα ερωτήματα για τα οποία ήταν γνωστά τα σχετικά τους κείμενα. Τα αποτελέσματα αξιολογήθηκαν με βάση την απόκριση (recall) και την ακρίβεια (precision) του συστήματος.

Για την δεύτερη συλλογή υπήρχαν 13 προκαθορισμένα ερωτήματα με αντίστοιχα σχετικά κείμενα για κάθε ερώτημα.

4. Αποκοπή καταλήξεων

Η ελληνική γλώσσα είναι πλούσια σε κλητικές και παραγωγικές καταλήξεις συγκρινόμενη με την αγγλική. (Έχουμε δέκα μέρη του λόγου, τα άρθρο, ουσιαστικό και επίθετο έχουν τρία γένη, κάθε γένος έχει δύο αριθμούς, και κάθε αριθμός έχει τέσσερις πτώσεις, στα μοντέρνα ελληνικά δεν υπάρχει η δοτική εκτός από ορισμένες στερεότυπες λέξεις της καθαρεύουσας και τέλος τα ρήματα έχουν χρόνους και πρόσωπο).

Ο αλγόριθμος αποκοπής των καταλήξεων, που χρησιμοποιήθηκε [4] λειτουργεί σε δύο στάδια: στο πρώτο στάδιο αποκόπτεται η κλητική κατάληξη της λέξης η οποία καθορίζει και την γραμματολογική της κατηγορία. Στο δεύτερο στάδιο αφαιρούνται οι παραγωγικές καταλήξεις σύμφωνα με την γραμματολογική κατηγορία.

Ο αλγόριθμος σταματά όταν η λέξη (ρίζα) είναι πολύ μικρή (πλήθος γραμμάτων <3).

Ο αλγόριθμος περιλαμβάνει μόνο 65 τύπους καταλήξεων συνολικά.

Στο παρακάτω πίνακα παρουσιάζονται τα αποτελέσματα της ανάκτησης από τη βάση COMPUTERS, ως προς το μέτρο της απόκρισης και ακρίβειας για 11 σημεία. Η απόδοση της ανάκτησης συγκρίνεται για τρεις περιπτώσεις: χωρίς την αποκοπή καταλήξεων (no stem), με αποκοπή μόνο των κλητικών καταλήξεων (stem_infl), και τέλος με τη χρησιμοποίηση του παραπάνω περιγραφέντος αλγορίθμου (tzk-stem).

Recall	no stem	stem_infl	tzk-stem
0,00	0,8989	0,9444	0,9368
0,10	0,8874	0,9358	0,9195
0,20	0,8360	0,9031	0,8913
0,30	0,7094	0,8631	0,8602
0,40	0,6261	0,7910	0,7859
0,50	0,5721	0,7678	0,7721
0,60	0,3778	0,6819	0,6822
0,70	0,3212	0,5613	0,6044
0,80	0,2525	0,4454	0,4748
0,90	0,1670	0,3770	0,4184
1,00	0,1597	0,3510	0,4081
Μέση ακρίβεια για όλα τα σημεία	0,5280	0,6929	0,7049
Μέση ακρίβεια για 3 ενδιάμεσα σημεία (0,20 - 0,50 - 0,80)	0,5535	0,7054	0,7127

5. Προβλήματα Ανάκτησης Ελληνικών Κειμένων - Συμπεράσματα

Από την επεξεργασία ελληνικών συλλογών φάνηκε ότι ορισμένα προβλήματα που παρουσιάζονται στην ελληνική γλώσσα επηρεάζουν την επίδοση της ανάκτησης. Για παράδειγμα πολλές λέξεις στη νεοελληνική συναντώνται με διαφορετική ορθογραφία, η χρήση

της δημοτικής γλώσσας δεν είναι αυστηρά με αποτέλεσμα να συναντώνται πολλές φορές και τύποι της καθαρεύουσας. Επίσης η χρήση λατινικών χαρακτήρων, στη περίπτωση που είναι κοινοί με το ελληνικό αλφάβητο, επηρεάζει την ανάκτηση.

Η ελληνική γλώσσα είναι πλούσια σε κλητικές και παραγωγικές καταλήξεις, κατά συνέπεια χρησιμοποιώντας ένα κατάλληλο αλγόριθμο αποκοπής καταλήξεων δημιουργούνται μεγάλες ομάδες όρων με κοινή ρίζα, γεγονός που οδηγεί στην βελτίωση της απόκρισης, αλλά σε μείωση της ακρίβειας.

Φαίνεται λοιπόν επιτακτική η ανάγκη έρευνας για περαιτέρω επεξεργασία των ομάδων των λέξεων με κοινή ρίζα, για την δημιουργία υποομάδων με βάση τις ομοιότητες των όρων σε σχέση με τη συλλογή των κειμένων.

6. Αναφορές

1. SMART Staff,
   User's manual for the SMART Information Retrieval System, Technical Report 71-95,
   Revices April 1974,
   Cornell University, 1974.

2. Salton G., McGill M.,
   Introduction to Modern Information Retrieval,
   McGraw-Hill, New York, 1983.

3. Buckley C.,
   Implementation of the SMART Information Retrieval System, Technical Report 85-686,
   Cornell University 1985.

4. Kalamboukis T.,
   Suffix stripping with modern Greek,
   Program automated library and information systems.
   Vol 29 ,ASLIB, July 1995.

2. Information and Knowledge Extraction from Medical Texts

Dr. Ioanna Malagardi and Professor John Kontos
Artificial Intelligence Laboratory
Department of Informatics
Athens University of Economics and Business
76 Patission St., 104 34 Athens, Hellas
e-mail: ioanna@ilsp.gr
jpk@aueb.gr

Abstract

In the present paper we study the application of information and knowledge extraction techniques to medical texts. Linguistic and computational problems faced with such texts are reviewed in the present paper. In particular we study noun phrase analysis, causal sentence recognition and analysis, sentence pair localization with common elements and knowledge extraction from such pairs by causal and other forms of reasoning. The processing and extraction of information and knowledge is based on microcosm or domain knowledge.

Εξαγωγή Πληροφορίας και Γνώσης από Ιατρικά Κείμενα

Δρ. Ιωάννα Μαλαγαρδή και Καθηγητής Ιωάννης Κόντος
Εργαστήριο Τεχνητής Νοημοσύνης Τμήμα Πληροφορικής
Τμήμα Πληροφορικής
Οικονομικό Πανεπιστήμιο Αθηνών
Πατησίων76, 104 34 Αθήνα

1. Εισαγωγή

Η έρευνα και η εφαρμογή στην ιατρική απαιτεί την αξιοποίηση μεγάλου όγκου πληροφοριών και γνώσης. Ένα μεγάλο μέρος του υλικού αυτού είναι καταγεγραμμένο σε μορφή κειμένου φυσικής γλώσσας. Πολλά από αυτά τα κείμενα είναι πλέον διαθέσιμα σε ηλεκτρονική μορφή και επεξεργάσιμα από υπολογιστή όπως στο διαδίκτυο. Τα κείμενα αυτά αφορούν ιατρικές επιστημονικές εργασίες σε διάφορους επί μέρους τομείς. Η αυτόματη εξαγωγή πληροφορίας από κείμενα αυτού του τύπου πρέπει να στηριχτεί σε μεγάλες βάσεις γλωσσολογικών γνώσεων και ειδικών γνώσεων για τον κάθε τομέα ή μικρόκοσμο [1]. Η

έρευνα πάνω στο συγκεκριμένο θέμα βρίσκεται διεθνώς σε αρχικά στάδια. Στην Ελλάδα οι πρώτες ερευνητικές προσπάθειες έγιναν στο πλαίσιο του έργου Holist του Οικονομικού Πανεπιστημίου που ξεκίνησε το 1990 και συνεχίζονται από ερευνητική ομάδα της οποίας ορισμένα νεότερα αποτελέσματα παρουσιάζονται εδώ. Η μεθοδολογία που εφαρμόζεται είχε αρχικά προταθεί από τον Ι. Κόντο όπως αναφέρεται στις εργασίες [4] και [5].

Στην παρούσα εργασία αφού γίνει μία ανασκόπηση του θέματος και της μέχρι τώρα προόδου θα επικεντρωθούμε στην παρουσίαση γλωσσολογικών και υπολογιστικών προβλημάτων που αφορούν την εξαγωγή πληροφορίας από ιατρικά κείμενα με τη μελέτη των οποίων έχουμε ασχοληθεί. Τα προβλήματα αυτά είναι τα εξής:

- Η ανάλυση και αναγνώριση ονοματικών φράσεων
- Η ανάλυση και αναγνώριση αιτιακών προτάσεων
- Η εξαγωγή ζευγών αιτιακών προτάσεων με κοινά στοιχεία
- Η εξαγωγή αιτιακής γνώσης και συμπερασμός

2. Μορφή των προτάσεων των ιατρικών κειμένων

Η βασική μορφή των προτάσεων στα ιατρικά κείμενα είναι αυτή των προτάσεων που εκφράζουν τη σχέση **αιτία → αποτέλεσμα**. Τα κείμενα αυτά εκτός από αυτή τη βασική δομή περιέχουν και προτάσεις με άλλες μορφές, οι οποίες είναι φορείς άλλων γνώσεων πέραν των αιτιακών. Παραδείγματα τέτοιων περιπτώσεων από το κείμενο "Aspirin" του Weissmann [2] είναι:

α) ιστορικές
By the the early 1970s no useful hypothesis had yet explained how salicylates exerted their various effects.
Μέχρι τις αρχές της δεκαετίας του 1970 καμμία χρήσιμη υπόθεση δεν είχε ακόμη εξηγήσει πως τα σαλικιλικά προξενούν αποτελέσματα.

The first satisfactory mechanism for the action of aspirin was proposed in 1971 by John R. Vane and his colleagues at the Royal College of Surgeons in London.
Ο πρώτος ικανοποιητικός μηχανισμός για τη δράση της ασπιρίνης προτάθηκε το 1971 από τον John A. Vane και τους συνεργάτες του στο Βασιλικό Κολλέγιο Χειρουργών στο Λονδίνο.

β) ερωτηματικές
And why did some patients develop the nasal polyps accompanied by sniffles and wheezes known as aspirin hypersensitivity?
Και γιατί κάποιοι ασθενείς αναπτύσσουν το ρινικό πολύποδα συνοδευόμενο από φτάρνισμα και από συριγμό γνωστό ως υπερευαισθησία στην ασπιρίνη?

γ) γνωστικές
Vane had been impressed by the fact that many forms of tissue injury are followed by the release of prostaglandins.
Ο Vane εντυπωσιάστηκε από το γεγονός ότι πολλές μορφές τραυματισμού ιστών ακολουθούνται από την παραγωγή προσταγλανδίνης.

Vane and his colleagues argued not only that prostaglandins were produced at sites of inflammation but also that they could, alone or in concert with other mediators, provoke the cardinal signs of inflammation.
Ο Vane και οι συνεργάτες του επιχειρηματολόγησαν ότι όχι μόνο παρήχθησαν προσταγλανδίνες στο σημείο της φλεγμονής αλλά επίσης ότι μπορούσαν, μόνες τους ή σε συνδιασμό με άλλους παράγοντες, να προκαλέσουν τα οξέα σημεία της φλεγμονής.

δ) πειραματογενείς
Vane then used radioactively labeled arachidonic acid to demonstrate that aspirin and related drugs inhibited the synthesis of prostaglandins E2 and F2a.
Ο Vane τότε χρησιμοποίησε ραδιενεργά σημασμένο αραχιδονικό οξύ για να αποδείξει ότι η ασπιρίνη και τα σχετιζόμενα φάρμακα εμποδίζουν την σύνθεση των προσταγλανδινών Ε2 και F2a.

ε) περιγραφικές
NSAIDs (Nonsteroidal Anti-Inflammatory Drugs) are planar, anionic molecules that have an affinity for lipid environments such as the lipid bilayers of plasma membranes.
Τα NSAIDs (Μη- Στεροειδή Αντιφλεγμονώδη Φάρμακα) είναι ανιόντα μόρια που έχουν συνάφεια με λιπιδικά περιβάλλοντα όπως τα λιπιδικά διπλά στρώματα των μεμβρανών του κυτταροπλάσματος.

3. Γλωσσολογικά προβλήματα στα ιατρικά κείμενα

Στη συνέχεια θα ασχοληθούμε με γλωσσολογικά προβλήματα που αφορούν στις προτάσεις που εκφράζουν αιτιακές σχέσεις. Η επεξεργασία αυτή μπορεί να αποβλέπει σε διάφορες εφαρμογές όπως είναι οι εξής:

1. Η εξαγωγή αιτιακής γνώσης από κείμενα. Δεδομένης κάποιας παρατήρησης επί ενός ασθενούς αναζητούνται οι πιθανές αιτίες που αναφέρονται σε ένα σώμα κειμένων.

2. Η εξαγωγή γνώσης με συμπερασμό η οποία δεν αναφέρεται ρητά στο κείμενο αλλά προκύπτει από αυτό. Εδώ πρόκειται για την ανεύρεση μη αναφερόμενων αιτιακών σχέσεων σε ένα κείμενο βάσει άλλων που αναφέρονται και βάσει γενικής προϋπάρχουσας γνώσης που είναι αποθηκευμένη σε υπολογιστή.

3. Η μηχανική μετάφραση προτάσεων με αιτιακή γνώση από μία φυσική γλώσσα στην άλλη, η οποία μπορεί να είναι χρήσιμη σε περιπτώσεις όπου με μία έστω και κατά προσέγγιση μετάφραση μπορεί να ενημερωθεί ένας ιατρός για αιτιακές σχέσεις που περιέχονται σε μία εργασία που είναι γραμμένη σε μία γλώσσα που δεν γνωρίζει.

Η ολοκλήρωση μιας τέτοιας επεξεργασίας προτάσεων που εκφράζουν αιτιακές σχέσεις προϋποθέτει τουλάχιστον τα εξής στάδια:

i. Αναγνώριση λέξεων και ανάκληση των σχετικών πληροφοριών από το λεξικό.
ii. Συντακτική αναγνώριση της φραστικής δομής των προτάσεων.
iii. Ανάλυση των επί μέρους φράσεων σε συνδιασμό με προϋπάρχουσα γνώση και πραγματολογικές πληροφορίες.

Για να καταστεί δυνατή η εκτέλεση των σταδίων αυτών είναι απαραίτητη η αντιμετώπιση των γλωσσολογικών προβλημάτων που παρουσιάζονται [3]. Τα γλωσσολογικά προβλήματα επισημάνθηκαν με την ανάλυση ορισμένων φραστικών δομών, όπως είναι οι:

- Οι Ονοματικές Φράσεις (ΟΦ)
- Οι Προθετικές Φράσεις (ΠΦ)

Οι τύποι των Ονοματικών Φράσεων που αφορούν σε μέρη του σώματος εντοπίστηκαν σε ένα σώμα αγγλικών κειμένων και είναι οι εξής:

- Απλές Ονοματικές Φράσεις που συνίστανται από άρθρα, ονόματα και επίθετα.
- Σύνθετες Ονοματικές Φράσεις που συνίστανται από απλές Ονοματικές Φράσεις συνοδευόμενες από προθέσεις, συνδέσμους, αναφορικές αντωνυμίες και επιρρήματα.

Ο ρόλος των προθέσεων είναι αρκετά σημαντικός. Από τα κείμενα επιλέξαμε δομές αποτελούμενες από απλές Ονοματικές Φράσεις συνοδευόμενες από Προθετικές Φράσεις οι οποίες περιέχουν οντότητες (entities) που δηλώνουν μέρη του σώματος. Βάσει της επιλογής αυτής παρατηρήσαμε σχετικά με τις προθέσεις ότι στις περισσότερες προτάσεις που περιέχουν μια Ονοματική Φράση (οντότητα ή μέρος του σώματος) εμφανίζονται προθέσεις όπως οι **"of"**, **"in"**, **"from"** και **"on".** Η πρόθεση **"of"** είναι αυτή η οποία εμφανίζεται με τη μεγαλύτερη συχνότητα.

1. Πρόθεση "of"
- Η πρόθεση "of" σχεδόν σε όλες τις περιπτώσεις συνοδεύει μια ΟΦ η οποία περιέχει μέρη του σώματος.
- Η ΟΦ που δηλώνει ένα μέρος του σώματος σε ορισμένες περιπτώσεις συνοδεύεται από επίθετο και βρίσκεται πάντα δεξιά της πρόθεσης "of".
- Αριστερά της πρόθεσης "of" υπάρχουν άλλες οντότητες (ΟΦ) οι οποίες δεν είναι συνήθως μέρη του σώματος, αλλά άφορούν σε άλλες οντότητες και ιδιαίτερα σε οντότητες που υποδηλώνουν κάποια διαδιακασία.

2. Πρόθεση "in"
- Η πρόθεση "in" δεν εμφανίζεται με τόση συχνότητα όσο η "of" και συνοδεύει οντότητες, οι οποίες συνοδεύονται συνήθως από ένα άρθρο (Det) και ένα επίθετο (Adj) και υποδηλώνουν μέρη του σώματος.

3. Πρόθεση "from"
- Η πρόθεση "from" εμφανίζεται σπάνια και σε διαφόρων ειδών προτάσεις.

4. Πρόθεση "on"
- Η λιγότερο συχνά εμφανιζόμενη πρόθεση είναι η "on".

Εκτός από τις παρατηρήσεις αυτές προχωρήσαμε στην επισήμανση των σημασιακών σχέσεων που αντιστοιχούν στην κάθε πρόθεση, φυσικά βάσει του σώματος των κειμένων που χρησιμοποιήσαμε. Οι σχέσεις μεταξύ των όρων που μελετήθηκαν είναι:

- *αιτία (cause)*
- *διαδικασία και οντότητα (process- entity)*
- *ταξινόμηση (classification)*
- *ποιότητα (quality)*
- *ποσοδεικτικότητα (quantification)*
- *ικανότητα πρόκλησης (capability of cause)*
- *τμηματική σχέση (part relations)*
- *χρόνος αναφοράς (time of reference)*
- *περίσταση (circumstances)*
- *ιδιότητα δυναμική (quality dynamic)*
- *θέση (position)*
- *ποσότητα (quantity)*
- *προέλευση (origin)*

4. Υλοποίηση της ανάλυσης και της αναγνώρισης ονοματικών φράσεων

Η εξαγωγή πληροφορίας απαιτεί κατ' αρχήν την αναγνώριση ονοματικών φράσεων που αναφέρονται σε οντότητες οι οποίες εμπλέκονται στις αιτιακές σχέσεις που αποτελούν τα βασικά στοιχεία γνώσης τα οποία θέλουμε να εξαγάγουμε.

Για τον έλεγχο της ορθότητας των παραπάνω περιγραφών υλοποιήθηκε μία γραμματική συντακτικής αναγνώρισης ονοματικών φράσεων σε Turbo Prolog. Η πειραματική γραμματική χρησιμοποιεί την κατηγοριοποίηση 300 λέξεων. Για την αναγνώριση Ονοματικών Φράσεων του τύπου που περιγράψαμε χρησιμοποιούνται μέχρι τώρα 30 κανόνες. Ενδεικτικά αναφέρουμε το εξής παράδειγμα:

Παίρνουμε δύο φράσεις οι οποίες έχουν την ίδια συντακτική δομή, δηλώνουν όμως δύο διαφορετικές σχέσεις.

- The synthesis of prostaglandin
- The viscosity of membranes

Παρατηρούμε ότι ενώ οι δύο φράσεις έχουν την ίδια συντακτική δομή δηλώνουν δύο διαφορετικές σχέσεις. Στην πρώτη "The synthesis of prostaglandin" πρόκειται για τη σχέση process-entity (διαδικασία-οντότητα) η οποία δηλώνεται από την πρόθεση "of", ενώ στη δεύτερη "The viscosity of membranes" πρόκειται για τη σχέση quality-entity (ποιότητα-οντότητα) η οποία πάλι δηλώνεται από την πρόθεση "of". Στα κείμενα που μελετήσαμε διαπιστώσαμε ότι υπάρχει μία μεγάλη ποικιλία ΟΦ με διαφορετικές συντακτικές και σημασιολογικές μορφές.

5. Η ανάλυση και αναγνώριση αιτιακών προτάσεων

Η εκφορά αιτιακής γνώσης σε ιατρικά κείμενα στηρίζεται στη δήλωση αιτιακών σχέσεων. Μια αιτιακή σχέση ορίζεται ως ένα ζεύγος αποτελούμενο από το "προηγούμενο" (αιτία) και το "επόμενο" (αποτέλεσμα). Στη φυσική γλώσσα οι αιτιακές σχέσεις μπορεί να εκφραστούν με μια ποικιλία γλωσσικών μορφών. Τα προηγούμενα και τα επόμενα εκφράζονται ως δύο συνδεδεμένες προτάσεις ή φράσεις. Το σύστημα που υλοποιήθηκε με τη μέθοδο ARISTA [6] καλύπτει τις εξής γλωσσικές μορφές για την εκφορά της γνώσης:

- Προτάσεις ενεργητικής φωνής του τύπου " ΟΦ Ρ ΟΦ".
- Προτάσεις παθητικής φωνής του τύπου "ΟΦ Ρ από ΟΦ".

Το σύστημα αναγνωρίζει τους εξής τύπους ονοματικών φράσεων:

- Ένα όνομα οντότητας.
- Ένα άρθρο ή ποσοδείκτη ακολουθούμενο από όνομα οντότητας.
- Ένα όνομα οντότητας ακολουθούμενο από μια ΠΦ (Προθετική Φράση).
- Ένα όνομα διαδικασίας ακολουθούμενο από μια απλή ΠΦ.
- Ένα όνομα διαδικασίας ακολουθούμενο από μια σύνθετη ΠΦ.

Οι τύποι των Προθετικών Φράσεων που προβλέπονται είναι:

- Πρόθεση ακολουθούμενη από όνομα οντότητας.
- Πρόθεση ακολουθούμενη από όνομα διαδικασίας.
- Σύνθετη ΠΦ.

Οι αιτιακές σχέσεις που εκφράζονται από προτάσεις του κειμένου αναγνωρίζονται από ένα κατηγόρημα πέντε ορισμάτων που ονομάζεται "caused by". Τα ορίσματα του κατηγορήματος αυτού είναι:

- Διαδικασία- αποτέλεσμα που περιέχεται στο "επόμενο".
- Η οντότητα την οποία αφορά το αποτέλεσμα.
- Η διαδικασία- αιτία του "προηγούμενου".
- Η οντότητα την οποία αφορά η αιτία.
- Η κατεύθυνση της αιτιακής σχέσης.

6. Εξαγωγή ζευγών αιτιακών προτάσεων με κοινά στοιχεία

Με βάση τα παραπάνω αναλύθηκε το κείμενο "Aspirin" [2] και εντοπίστηκαν αυτομάτως 90 αιτιακές προτάσεις και αναλύθηκαν οι ονοματικές τους φράσεις, χρησιμοποιώντας τα εργαλεία που είχαν ήδη αναπτυχθεί για μικρότερα κείμενα με τις ανάλογες προσαρμογές στο θέμα του κειμένου.

Μετά την αναγνώριση μίας πρότασης με αιτιακή γνώση συγκρίνεται αυτομάτως η πρόταση αυτή με βάση τις ονοματικές φράσεις τις οποίες περιέχει με όλες τις άλλες αιτιακές προτάσεις του κειμένου. Εντοπίζεται έτσι το ζεύγος που τυχόν θα έχει κάποια ονοματική φράση που περιέχει την ίδια οντότητα [7].

Για το συγκεκριμένο κείμενο βρέθηκαν και παρουσιάζονται ενδεικτικά τα εξής ζεύγη:

aspirinlike drugs blocked ***prostaglandin*** *synthesis*
ασπιρινοειδή φάρμακα εμποδίζουν τη σύνθεση προσταγλανδίνης
high doses of stable **prostaglandins** inhibit inflammation in animals with arthritis
υψηλές δόσεις ευσταθούς προσταγλανδίνης αποτρέπουν τον ερεθισμό σε ζώα με αρθρίτιδα

Με έντονα στοιχεία χαρακτηρίζεται η κοινή οντότητα του ζεύγους και με υπογράμμιση η ονοματική φράση που περιέχει την κοινή οντότητα σε κάθε πρόταση. Πιθανό συμπέρασμα από τον συδυασμό των δύο προτάσεων είναι:

aspirinlike drugs **cause** inflammation in animals with arthritis
*ασπιρινοειδή φάρμακα **προκαλούν** τον ερεθισμό σε ζώα με αρθρίτιδα*

Το παραπάνω συμπέρασμα προκύπτει με την υπόθεση ότι ο συνδυασμός block (εμποδίζω) και inhibit (αποτρέπω) αντιστοιχεί νοηματικά στο ρήμα cause (προκαλώ) ή εναλλακτικά does not inhibit (δεν αποτρέπω). Επισημαίνουμε με το παραδειγμα αυτό πόσο δύσκολο είναι να εξαχθεί ένας ορθός συμπερασμός. Θα πρέπει για την εξαγωγή ορθών συμπερασμών να υπάρχει η γνώση της μακροδομής του κειμένου και της κατηγορίας του κειμένου. Σε κείμενο επιθεώρησης π.χ. θα υπάρχουν πολλά αντιφατικά στοιχεία, ενώ σε ένα κείμενο μελέτης που παρουσιάζει συγκεκριμένη διαπίστωση δεν θα υπάρχουν πιθανώς αντιφατικά στοιχεία. Γενικώς το θέμα τέτοιου είδους συμπερασμών είναι πολύπλοκο και αποτελεί αντικείμενο έρευνας σε εξέλιξη.

Θα αναφερθούμε και σε ένα δεύτερο ζεύγος αιτιακών προτάσεων πιθανό για εξαγωγή συμπερασμού:

aspirinlike drugs blocked **prostaglandin** synthesis
ασπιρινοειδή φάρμακα εμποδίζουν τη σύνθεση προσταγλανδίνης
prostaglandins induce vasodilation
Οι προσταγλανδίνες προκαλούν φλεβοδιαστολή

Πιθανό συμπέρασμα από τον συδυασμό των δύο προτάσεων είναι:

aspirinlike drugs **inhibit** vasodilation
ασπιρινοειδή φάρμακα **εμποδίζουν** τη φλεβοδιαστολή

Το παραπάνω συμπέρασμα φαίνεται λογικό. Η δυνατότητα να εξαχθούν λογικά συμπεράσματα εξαρτάται σε μεγάλο βαθμό από την καταλληλότητα του κειμένου.

Ζεύγη όπως τα παραπάνω μπορούν να αποτελέσουν κρίκους αλυσίδας συμπερασμού. Η εκτέλεση του αιτιακού συμπερασμού με τον υπολογιστή για κείμενα αυτού του μεγέθους προϋποθέτει και άλλα ισχυρότερα εργαλεία που βρίσκονται σε εξέλιξη.

7. Εξαγωγή αιτιακής γνώσης και συμπερασμός

Μέχρι σήμερα έχουν αναπτυχθεί ορισμένα εργαλεία για την εξαγωγή αιτιακής γνώσης και την αξιοποίησή

της για συμπερασμό από κείμενα όπως αυτό της Ασπιρίνης [8]. Τα εργαλεία αυτά δεν είναι κατάλληλα για μεγαλύτερα κείμενα λόγω της χρονοβόρας διαδικασίας που απαιτείται για την εφαρμογή τους. Για τον λόγο αυτόν βρίσκεται σε εξέλιξη η δημιουργία ευφυέστερων εργαλείων για τον συμπερασμό για μεγάλα κείμενα.

Παράδειγμα Αιτιακής Αλυσίδας και Συμπερασμού

Many forms of tissue	*οντότης*
Injury	*διαδικασία*
are followed by	*αιτιακός σύνδεσμος*
the release	*διαδικασία*
of **prostaglandins**	*οντότης*
prostaglandin	*οντότης*
provokes	*αιτιακός σύνδεσμος*
inflammation	*διαδικασία*
(of tissue?)	*ελλείπουσα οντότης*

Συμπέρασμα: Injury causes inflammation

8. Βιβλιογραφία

[1]. Pazienza, M. T. *Information Extraction*. LNAI Tutorial. Springer, 1997.

[2]. Weissmann, G. Aspirin. *Scientific American*.1990.

[3]. Μαλαγαρδή Ι. Γλωσσολογικά Προβλήματα Ιατρικών Κειμένων. *Ανακοίνωση στην Ημερίδα για το Ερευνητικό Έργο HOLIST/Stride. Οικονομικό Πανεπιστήμιο Αθηνών*. ΕΙΕ, Αθήνα, 1994.

[4]. Kontos J. 1980: "Syntax-Directed Processing of Texts with Action Semantics". *Cybernetica,* 23, 2, 157-175.

[5]. Kontos J. 1983: " Syntax-Directed Fact Retrieval from Texts with a Micro-Computer". Proc. MELECON 83 , Athens.

[6]. Kontos, J. ARISTA: Knowledge Engineering with Scientific Texts. *Information and Software Technology*, Vol. 34, No 9, pp. 611-616, 1992.

[7]. Κόντος Ι. *Τεχνητή Νοημοσύνη και Λογομηχανική* (Επεξεργασία Λόγου). (Αθήνα: Εκδόσεις Ε. Μπένου) 1996.

[8]. Μαλαγαρδή Ι. Προσδιορισμός με Υπολογιστή της Υπονοούμενης Σχέσης μεταξύ των Συστατικών Ονοματικών Φράσεων σε Υπογλώσσες. *17η Συνάντηση Εργασίας ΑΠΘ*, Θεσσαλονίκη, 1996.

3. Question Answering for Information and Knowledge Extraction

Professor John Kontos
Artificial Intelligence Laboratory
Department of Informatics
Athens University of Economics and Business
76 Patission St., 104 34 Athens, Hellas
E-mail: jpk@aueb.gr

Abstract

In this paper we present the implementation and application of a number of Question Answering Systems for the processing of questions expressed in Greek that refer to the extraction of information and knowledge. The implementation is based on the programming language Prolog. The functions performed by these systems include the processing of questions that consist of sentences with relative clauses and noun-noun combinations. The processing is based on the Knowledge Bases that are used for the resolution of ambiguities of the questions. Artificial Intelligence methods are used extensively in these systems.

Επεξεργασία Ερωτήσεων για Εξαγωγή Πληροφορίας και Γνώσης

Καθηγητής Ιωάννης Κόντος
Εργαστήριο Τεχνητής Νοημοσύνης Τμήμα Πληροφορικής
Τμήμα Πληροφορικής
Οικονομικό Πανεπιστήμιο Αθηνών
Πατησίων76, 104 34 Αθήνα

1. Εισαγωγή

Στην εργασία αυτή παρουσιάζονται η υλοποίηση και η εφαρμογή συστημάτων Επεξεργασίας Ερωτήσεων με Υπολογιστή εκφρασμένων στην Ελληνική που αφορούν στην αναζήτηση και εξαγωγή πληροφοριών και γνώσεων. Η υλοποίηση έχει γίνει με την γλώσσα προγραμματισμού Prolog. Οι λειτουργίες που εκτελούνται περιλαμβάνουν επεξεργασία ερωτήσεων που περιέχουν αναφορικές προτάσεις και συνδυασμούς ονομάτων και απαντώνται από βάσεις γνώσεων ή κειμένων. Η επεξεργασία υποστηρίζεται από βάσεις γνώσεων που χρησιμοποιούνται για την αποσαφήνιση

αμφισημιών των ερωτήσεων.

Ένα σύστημα επεξεργασίας ερωτήσεων σε φυσική γλώσσα προσφέρει την δυνατότητα εξαγωγής πληροφοριών από βάσεις γνώσεων ή κειμένων. Οι βάσεις γνώσεων ή κειμένων μπορεί να περιέχουν περιγραφές ή γνώσεις. Η απάντηση των ερωτήσεων μπορεί να απαιτεί συμπερασμό καθώς και τη χρήση βάσης γνώσης του σχετικού μικρόκοσμου. Στη συνέχεια θα περιγραφούν τα εξής συστήματα επεξεργασίας ελληνικών ερωτήσεων που έχουμε αναπτύξει.

- Σύστημα με ενδιάμεση γλώσσα και περιγραφές εικόνων
- Σύστημα με ενδιάμεση γλώσσα και συμπερασμό
- Σύστημα με ενδιάμεση γλώσσα την SQL
- Σύστημα απάντησης ερωτήσεων χωρίς ενδιάμεση γλώσσα

2. Συστήματα με Ενδιάμεση Διαδικαστική Γλώσσα

Στα περισσότερα από τα σημερινά συστήματα διεπαφών φυσικής γλώσσας, η ερώτηση μετατρέπεται αρχικά σε μία ενδιάμεση λογική ερώτηση, που εκφράζεται σε τυπική γλώσσα παράστασης. Η λογική αυτή παράσταση στην συνέχεια μετατρέπεται σε μία έκφραση της γλώσσας ερωτήσεων της βάσης δεδομένων και δίνεται στο σύστημα διαχείρισης της βάσης για να εκτελεσθεί. Σε ένα σύστημα αυτού του τύπου το πρώτο τμήμα επιτελεί τη λεκτική ανάλυση των ερωτήσεων και την ανάκτηση των λεκτικών χαρακτηριστικών και το δεύτερο τμήμα μεταφράζει αυτόματα τις ερωτήσεις σε προγράμματα γραμμένα στην ενδιάμεση διαδικαστική τυπική γλώσσα του συστήματος. Το τρίτο τμήμα δέχεται το πρόγραμμα που παράγεται στο δεύτερο τμήμα και το εκτελεί. Η εκτέλεση αυτή μπορεί να αφορά μία από τις δύο βασικές λειτουργίες. Η πρώτη λειτουργία είναι η ενημέρωση μίας βάσης δεδομένων και η δεύτερη είναι η άντληση πληροφοριών από τη βάση δεδομένων και η παραγωγή απάντησης στην υποβληθείσα ερώτηση. Από τα πρώτα υλοποιημένα συστήματα αυτού του τύπου είναι τα DELFI και SHRDLU που δημοσιεύθηκαν για πρώτη φορά στις εργασίες (Kontos J., 1970) και (Winograd T., 1972) αντίστοιχα. Νεώτερη έκδοση του DELFI περιγράφεται στην εργασία (Kontos J., 1971). Η αρχιτεκτονική ενός συστήματος που χρησιμοποιεί ενδιάμεση παράσταση έχει την εξής δομή: Η είσοδος φυσικής γλώσσας, αρχικά αναλύεται συντακτικά από τον μεριστή, ο οποίος παράγει ένα συντακτικό δένδρο. Στην συνέχεια ο σημασιολογικός αναλυτής μετατρέπει το συντακτικό δένδρο σε ενδιάμεση λογική παράσταση. Στα νεότερα συστήματα που έχουμε ερευνήσει αποφεύγουμε κατά το δυνατόν τις ενδιάμεσες παραστάσεις των προτάσεων φυσικής γλώσσας.

3. Σύστημα με Ενδιάμεση Γλώσσα και Περιγραφές Εικόνων

Στη συνέχεια περιγράφεται το σύστημα DELFI που είναι σύστημα απαντήσεως ερωτήσεων προς μία βάση που περιέχει περιγραφές εικόνων. Στην εφαρμογή αυτήν το σύστημα μπορεί να απαντήσει σε πολύπλοκες ερωτήσεις που αφορούν ένα σύνολο αντικειμένων που συνδέονται μεταξύ τους με σχέσεις χώρου. Η είσοδος μπορεί να είναι είτε δηλωτικές προτάσεις είτε ερωτήσεις. Οι δηλωτικές προτάσεις περιέχουν πληροφορίες για την ενημέρωση της βάσης ενώ οι ερωτήσεις χρησιμοποιούνται για την εξαγωγή πληροφοριών από τη βάση.

Η βάση περιέχει περιγραφές αντικειμένων καθένα των οποίων μπορεί να είναι κύκλος ή τετράπλευρο που έχουν διχοτομηθεί οριζόντια με έναν άξονα. Το σύνολο των αντικειμένων αυτών αποτελεί τον μικρόκοσμο της εφαρμογής. Ιστορικά οι έννοιες αυτές έχουν εμφανιστεί σε φιλοσοφικά κείμενα που ασχολούνται με τη γλώσσα όπως (Αριστοτέλης, 4ος π.Χ.), (Wittgenstein L., 1953). Στον μικρόκοσμο αυτό κάθε αντικείμενο μπορεί να βρίσκεται πάνω ή κάτω από ένα άλλο αντικείμενο ανάλογα με τη σχετική θέση του άξονά του, να βρίσκεται εντός ή εκτός σε σχέση με ένα άλλο αντικείμενο ανάλογα με τη σχετική θέση του περιγράμματός του, να έχει περίγραμμα συνεχές ή διακεκομμένο και άξονα συνεχή ή διακεκομμένο. Τα αντικείμενα διακρίνονται με αριθμούς. Στη συνέχεια δίνουμε παραδείγματα περιγραφής των σχέσεων του αντικειμένου 4 με τα άλλα αντικείμενα.

Παράδειγμα 1: Το αντικείμενο 4 είναι κάτω από τα αντικείμενα 2 και 3 και πάνω από τα αντικείμενα 1 και 5 και εντός των αντικειμένων 3 και 1.

Πρέπει να σημειωθεί ότι οι έννοιες "πάνω" και "κάτω"

έχουν οριστεί για το συγκεκριμένο μικρόκοσμο με τρόπο που μπορεί να διαφέρει από τον τρόπο αντίληψης ενός ανθρώπου ώστε να επιδειχθεί η δυνατότητα αυθαιρεσίας στον ορισμό του μικρόκοσμου. Έτσι φαίνεται πώς η έννοια μιας λέξης μπορεί να εξαρτάται από τους νόμους του μικρόκοσμου. Εάν θέλαμε ο ορισμός του μικρόκοσμου να συμπίπτει με τη συνήθη αντίληψη ενός ανθρώπου τότε θα ορίζαμε π.χ. τη σχέση "πάνω" ως εξής: Το "αντικείμενο α" είναι πάνω από το "αντικείμενο β" εφόσον ο άξονας του "α" είναι πάνω από τον άξονα του "β" και το "α" δεν είναι εντός του "β".

Στην περίπτωση του διαμορφωμένου αυτού μικρόκοσμου το Παράδειγμα 1 μεταγράφεται ως εξής:

Παράδειγμα 2: Το αντικείμενο 4 είναι κάτω από το αντικείμενο 3 και πάνω από το αντικείμενο 5.

Η λειτουργία του συστήματος με την υπόθεση ότι στον αναλυόμενο μικρόκοσμο ισχύουν οι σημασιολογικοί νόμοι ορισμού των σχέσεων που προϋποθέτει το Παράδειγμα 1. Για την προετοιμασία του συστήματος πρέπει να προηγηθεί μία φάση ενημέρωσης του συστήματος, που απαιτεί την τροφοδότησή του με πληροφορίες σχετικές με τα αντικείμενα του μικρόκοσμου. Μετά τη συμπλήρωση της φάσης της ενημέρωσης μπορούμε να υποβάλουμε στο σύστημα ερωτήσεις και να λάβουμε απαντήσεις.

4. Σύστημα με Ενδιάμεση Γλώσσα και Συμπερασμό

Ένα παράδειγμα απάντησης ερωτήσεων με συμπερασμό που εφαρμόστηκε η δεύτερη έκδοση του DELFI αφορά την απάντηση ερωτήσεων σχετικά με αεροπορικά δρομολόγια (Kontos J., 1971). Η βάση δεδομένων περιέχει στοιχεία πτήσεων όπως παρουσιάζονται παρακάτω.

Η ΠΤΗΣΗ-1 ΕΙΝΑΙ ΜΕΛΟΣ ΤΩΝ ΠΤΗΣΕΩΝ
Η ΠΤΗΣΗ-2 ΕΙΝΑΙ ΜΕΛΟΣ ΤΩΝ ΠΤΗΣΕΩΝ
Η ΠΤΗΣΗ-3 ΕΙΝΑΙ ΜΕΛΟΣ ΤΩΝ ΠΤΗΣΕΩΝ
Η ΠΤΗΣΗ-1 ΑΝΑΧΩΡΕΙ ΑΠΟ ΑΘΗΝΑ ΚΑΙ ΠΕΤΑΕΙ ΓΙΑ ΡΩΜΗ
Η ΠΤΗΣΗ-2 ΑΝΑΧΩΡΕΙ ΑΠΟ ΡΩΜΗ ΚΑΙ ΠΕΤΑΕΙ ΓΙΑ ΠΑΡΙΣΙ
Η ΠΤΗΣΗ-3 ΑΝΑΧΩΡΕΙ ΑΠΟ ΡΩΜΗ ΚΑΙ ΠΕΤΑΕΙ ΓΙΑ ΛΟΝΔΙΝΟ
Η ΠΤΗΣΗ-1 ΕΧΕΙ ΧΡΟΝΟ ΑΝΑΧΩΡΗΣΗΣ 9 ΚΑΙ ΕΧΕΙ ΧΡΟΝΟ ΑΦΙΞΗΣ 11
Η ΠΤΗΣΗ-2 ΕΧΕΙ ΧΡΟΝΟ ΑΝΑΧΩΡΗΣΗΣ 13 ΚΑΙ ΕΧΕΙ ΧΡΟΝΟ ΑΦΙΞΗΣ 14
Η ΠΤΗΣΗ-3 ΕΧΕΙ ΧΡΟΝΟ ΑΝΑΧΩΡΗΣΗΣ 10 ΚΑΙ ΕΧΕΙ ΧΡΟΝΟ ΑΦΙΞΗΣ 12

Η λειτουργία του συστήματος μπορεί να θεωρηθεί ως λειτουργία ενός εμπείρου συστήματος. Οι κανόνες συμπερασμού του συστήματος αποτελούν τη "βάση γνώσης" του έμπειρου συστήματος.

Οι κανόνες συμπερασμού που χρησιμοποιούνται στην εφαρμογή αυτή είναι οι εξής:

ΑΝ Π-1 ΕΙΝΑΙ ΜΕΛΟΣ ΤΩΝ ΠΤΗΣΕΩΝ ΚΑΙ Π-2 ΕΙΝΑΙ ΜΕΛΟΣ ΤΩΝ ΠΤΗΣΕΩΝ ΚΑΙ Π-1 ΠΕΤΑΕΙ ΓΙΑ Τ-1 ΚΑΙ Π-2 ΑΝΑΧΩΡΕΙ ΑΠΟ Τ-1 ΤΟΤΕ Π-2 ΑΚΟΛΟΥΘΕΙ Π-1.

ΑΝ Π-1 ΑΚΟΛΟΥΘΕΙ Π-2 ΚΑΙ Π-1 ΕΧΕΙ ΧΡΟΝΟ ΑΝΑΧΩΡΗΣΗΣ ΜΕΓΑΛΥΤΕΡΟ ΚΑΤΑ 2 ΑΠΟ ΤΟ ΧΡΟΝΟ ΑΦΙΞΗΣ ΤΟΥ Π-2 ΤΟΤΕ Π-1 ΣΥΝΔΕΕΤΑΙ ΜΕ Π-2.

ΑΝ Π-1 ΣΥΝΔΕΕΤΑΙ ΜΕ Π-2 ΚΑΙ Π-2 ΑΝΑΧΩΡΕΙ ΑΠΟ Τ-1 ΚΑΙ Π-1 ΠΕΤΑΕΙ ΓΙΑ Τ-2 ΤΟΤΕ Τ-2 ΕΙΝΑΙ ΠΡΟΣΠΕΛΑΣΙΜΟΣ ΑΠΟ Τ-1.

Στην ερώτηση: "Είναι το Παρίσι προσπελάσιμο από Αθήνα ;" Η απάντηση που δίνει το σύστημα είναι καταφατική. Στην ερώτηση: "Είναι το Λονδίνο προσπελάσιμο από Αθήνα ;" Η απάντηση που δίνει το σύστημα είναι αρνητική.

5. Σύστημα με Ενδιάμεση Γλώσσα την SQL

Στο τμήμα αυτό παρουσιάζεται ένα σύστημα που στηρίζεται στην παραγωγή προτάσεων στη γλώσσα SQL από ερωτήσεις διατυπωμένες σε φυσική γλώσσα. Το σύστημα χρησιμεύει για τη δημιουργία φιλικών διεπαφών του χρήστη για σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων (RDBMS's) τα οποία δέχονται ερωτήσεις σε γλώσσα SQL. Στο σύστημα που περιγράφεται εδώ η επεξεργασία της φυσικής γλώσσας έχει γίνει με γλώσσα Prolog. Η αναπαράσταση της βάσης δεδομένων γίνεται με ένα σύστημα διαχείρισης βάσεων δεδομένων, το οποίο δέχεται ερωτήσεις SQL. Οι βάσεις δεδομένων μπορεί να έχουν μπορεί να

έχουν προκύψει από την εξαγωγή πληροφορίας από κείμενα με τη μέθοδο των "templates".

Τα ρήματα αποτελούν σημαντικό στοιχείο των ερωτήσεων καθόσον προσδιορίζουν τη σχέση μεταξύ των αντικειμένων της βάσης για τα οποία ο χρήστης αναζητεί πληροφορίες. Υπάρχουν όμως περιπτώσεις στις οποίες αυτό δεν είναι δυνατόν εξαιτίας των ελλείψεων που παρουσιάζονται στις ερωτήσεις σε φυσική γλώσσα με αποτέλεσμα το σύστημα να κινδυνεύει να δώσει λάθος αποτέλεσμα. Ως παραδείγματα αναφέρονται οι εξής ερωτήσεις:

- ΠΟΙΕΣ ΕΙΝΑΙ ΟΙ ΠΡΩΤΕΥΟΥΣΕΣ ΤΩΝ ΧΩΡΩΝ ΤΗΣ ΑΣΙΑΣ;
- ΠΟΙΕΣ ΕΙΝΑΙ ΟΙ ΠΡΩΤΕΥΟΥΣΕΣ ΤΩΝ ΧΩΡΩΝ ΤΗΣ ΕΕ;

Ενώ οι δύο αυτές ερωτήσεις είναι σχεδόν όμοιες, η μοναδική διαφορά που παρουσιάζουν μεταξύ τους έγκειται σε μία μόνο λέξη, δηλαδή η πρώτη πρόταση έχει "ΑΣΙΑ" (όνομα ηπείρου) εκεί που η δεύτερη πρόταση έχει "ΕΕ" (όνομα οργανισμού). Η διαφορά αυτή επηρεάζει δραστικά τη μορφή του παραγόμενου προγράμματος σε SQL. Στις δύο αυτές ερωτήσεις αφενός μεν ελλείπουν οι λέξεις ΗΠΕΙΡΟΣ και ΟΡΓΑΝΙΣΜΟΣ αφετέρου δε δεν υπάρχει ρήμα που να οδηγεί το σύστημα για την απάντηση σε αυτές τις ερωτήσεις. Επειδή η ΑΣΙΑ είναι ΗΠΕΙΡΟΣ η εντολή SELECT του παραγόμενου προγράμματος σε SQL πρέπει να αναφέρεται στον πίνακα countries, ενώ επειδή η λέξη ΕΕ είναι ΟΡΓΑΝΙΣΜΟΣ η αντίστοιχη παραγόμενη εντολή SELECT πρέπει να αναφέρεται στον πίνακα orgbase. Για τον λόγο αυτόν χρησιμοποιείται κατά την ανάλυση και κατανόηση της ερώτησης ένα άλλο κατηγόρημα, το kind(s, s), που περιέχει γνώση του κόσμου.

Ως παραδείγματα αναφέρονται οι παρακάτω δηλώσεις γνώσης:

```
kind("ΑΣΙΑ","ΗΠΕΙΡΟΣ").kind("ΕΥΡΩΠΗ","ΗΠΕΙΡΟΣ").
kind("ΑΜΕΡΙΚΗ", "ΗΠΕΙΡΟΣ").
kind("ΕΕ","ΟΡΓΑΝΙΣΜΟΣ").kind("ΝΑΤΟ","ΟΡΓΑΝΙΣΜΟΣ").
kind("ΟΗΕ", "ΟΡΓΑΝΙΣΜΟΣ").
```

Με τέτοιες δηλώσεις το σύστημα είναι σε θέση να προσδιορίσει την ορθή μετάφραση των δύο ερωτήσεων του προηγούμενου παραδείγματος επιλέγοντας και τον κατάλληλο πίνακα. Η μετάφραση μιας ερώτησης παρουσιάζεται στον παρακάτω πίνακα:

ΠΟΙΕΣ ΧΩΡΕΣ ΕΞΑΓΟΥΝ ΚΡΑΣΙ ΚΑΙ	<S> → <QW><NP>
ΒΡΙΣΚΟΝΤΑΙ ΣΤΗΝ ΕΥΡΩΠΗ;	<QW> → ΠΟΙΕΣ¦ΠΟΣΕΣ
SELECT DISTINCT (expbase.country)	<NP> → <ENT><VP1>ΚΑΙ<VP2>
FROM expbase, countries	<ENT> → ΧΩΡΕΣ¦...
WHERE expbase.country=countries.country	<VP1> → <TV><PRODUCT>
AND expbase.product='ΚΡΑΣΙ'	<TV> → ΕΞΑΓΟΥΝ¦ΕΧΟΥΝ
AND countries.continent='ΕΥΡΩΠΗ' ;	<PRODUCT> → ΚΡΑΣΙ ¦ΛΑΔΙ ...
	<VP2> → <IV><PREP><CONT>
	<IV> → ΑΝΗΚΟΥΝ¦ΒΡΙΣΚΟΝΤΑΙ
	<PREP> → ΣΤΗΝ ¦ΣΤΑ ...
	<CONT> → ΕΥΡΩΠΗ¦ΑΜΕΡΙΚΗ...

ΑΠΑΝΤΗΣΗ = Ελλάδα, Ισπανία, Γαλλία, Ιταλία

Στο αριστερό τμήμα του παραπάνω πίνακα δίνεται η ερώτηση που εισάγεται από τον χρήστη σε φυσική γλώσσα και από κάτω το πρόγραμμα σε γλώσσα SQL που παράγεται ως μετάφραση της ερώτησης. Στο δεξιό τμήμα του πίνακα δίνεται σε μορφή BNF το τμήμα της γραμματικής που ενεργοποιείται για τη συντακτική ανάλυση της ίδιας ερώτησης.

6. Σύστημα Απάντησης Ερωτήσεων χωρίς Ενδιάμεση Γλώσσα

Έχουμε υλοποιήσει και συστήματα που απαντούν ερωτήσεις απευθείας από κείμενα χρησιμοποιώντας την πρωτότυπη μέθοδο ARISTA (Kontos J. 1980,1983,1992,1996). Στο παρακάτω παράδειγμα το κείμενο γνώσης ή "γνωσιακό" κείμενο περιγράφει τη σύνταξη και τη σημασιολογία απλών λογικών εκφράσεων.

Το κείμενο έχει ως εξής :

Μία εντολή σημαίνει δώσε την τιμή της έκφρασης στη μεταβλητή.
Η μορφή μιας εντολής είναι μεταβλητή, σχέση, έκφραση.
p είναι μία μεταβλητή. q είναι μία μεταβλητή. r είναι μία μεταβλητή.
"ισούται" είναι μία σχέση.
Η μορφή μίας έκφρασης είναι μεταβλητή, συνδετικό, μεταβλητή.
Το "και" είναι ένα συνδετικό. Το "ή" είναι ένα συνδετικό.
Το p είναι Ψευδές. Το q είναι Αληθές.

Ένας χρήστης μπορεί να υποβάλλει την ερώτηση:

"Ποια είναι η έννοια του "r ισούται p και q" ;"

Δεδομένων των συνήθων πινάκων αληθείας των λογικών συνδετικών "ή" και "και" και δεδομένων των τιμών αληθείας των p και q που δίνονται στο κείμενο μια εύλογη απάντηση στην παραπάνω ερώτηση ,η οποία και παράγεται αυτομάτως, είναι:
"Η έννοια του "r ισούται p και q" είναι "δώσε την τιμή Ψευδές στο r"

7. Συμπεράσματα

Στην εργασία αυτή παρουσιάστηκαν η υλοποίηση και η εφαρμογή συστημάτων Επεξεργασίας Ερωτήσεων με Υπολογιστή εκφρασμένων στην Ελληνική που αφορούν στην αναζήτηση και εξαγωγή πληροφοριών και γνώσεων. Οι λειτουργίες επεξεργασίας λόγου που εκτελούνται από τα συστήματα αυτά στηρίζονται σε μεθοδολογίες προερχόμενες κυρίως από τον χώρο της Τεχνητής Νοημοσύνης. Η επιλογή της κατάλληλης μεθόδου εξαρτάται από την εκάστοτε εφαρμογή.

8. Βιβλιογραφία

Αριστοτέλης. *Άπαντα* (Αθήνα: Εκδόσεις Κάκτος 1994).

Kontos J. & G. Papakonstantinou 1970: "A Question-Answering System Using Program Generation". *Proceedings of A.C.M. International Computing Symposium*, Bonn Germany.

Kontos J. & A. Kossidas 1971: "On the Question-Answering System DELFI and its Application". *Proceedings of AGARD Symposium on Artificial Intelligence.* Rome, Italy.

Kontos J. 1980: "Syntax-Directed Processing of Texts with Action Semantics". *Cybernetica*, 23, 2, 157-175.

Kontos J. 1983: " Syntax-Directed Fact Retrieval from Texts with a Micro-Computer". Proc. MELECON 83, Athens.

Kontos J. 1992: "ARISTA: Knowledge Engineering with Scientific Texts". *Information and Software Technology*, 34, 611-616.

Κόντος Ι. 1996: *Τεχνητή Νοημοσύνη και Λογομηχανική.* (Αθήνα: Εκδόσεις Ε. Μπένου).

Winograd T. 1972: *Understanding Natural Language.* Academic Press.

Wittgenstein L. 1953: *Philosophical Investigations.* New York: Macmillan.

4. Automatic Term Extraction Based on Pattern Grammars

Byron Georgantopoulos and Stelios Piperidis
Institute for Language and Speech Processing
Epidaurou & Artemidos 6 Paradeisos Amarousiou
151 25 Athens, Hellas
spip@ilsp.gr, byron@ilsp.gr

Abstract

In this paper, we present a method for the automatic extraction of terms from machine-readable text corpora. The method is based on a pattern grammar endowed with regular expressions and feature-structure unification capacity. The text corpus we have used consisted of a software manual by Hewlett-Packard extending to around 90000 wordforms, containing a term index against which the results of the method were evaluated. The method extracted 124 out of 214 manually coded terms, featuring a 58% recall.

Αυτόματη Εξαγωγή Όρων με Χρήση Γραμματικής Προτύπων

Βύρων Γεωργαντόπουλος και Στέλιος Πιπερίδης
Ινστιτούτο Επεξεργασίας του Λόγου
Επιδαύρου & Αρτέμιδος6 Παράδεισος Αμαρουσίου
151 25 Αθήνα

Περίληψη

Στο άρθρο αυτό παρουσιάζονται τα πρώτα αποτελέσματα μιας μεθόδου αυτόματης εξαγωγής όρων από σώματα κειμένων. Η μέθοδος στηρίζεται στην εφαρμογή μιας γραμματικής προτύπων που χρησιμοποιεί το φορμαλισμό ενοποίησης (feature-structure unification) και τελεστές κανονικών εκφράσεων-γραμματικών (regular expressions). Το σώμα κειμένων που χρησιμοποιήθηκε είναι ένα εγχειρίδιο οδηγιών της Hewlett-Packard μεγέθους περίπου 90000 λέξεων που περιελάμβανε έναν κατάλογο όρων έναντι του οποίου αξιολογήθηκαν τα αποτελέσματα της μεθόδου. Η μέθοδος εξήγαγε 124 από τους 214 όρους που είχαν εξαχθεί χειρωνακτικά, παρουσιάζοντας ποσοστό ανάκτησης 58%.

1. Εισαγωγή

Στο άρθρο αυτό παρουσιάζονται τα πρώτα αποτελέσματα μιας μεθόδου αυτόματης εξαγωγής όρων από σώματα κειμένων. Η αυτόματη εξαγωγή όρων αποκτά ιδιαίτερο ενδιαφέρον σήμερα που μεγάλοι όγκοι κειμένων παράγονται πλέον ηλεκτρονικά, γεγονός που οδηγεί στην διατύπωση νέων απαιτήσεων για την διαχείριση και επεξεργασία τους (αυτόματη ταξινόμηση, ανάκτηση πληροφοριών, κλπ). Η εφαρμογή συστημάτων γλωσσικής τεχνολογίας για την ικανοποίηση των αναγκών αυτών απαιτεί την προσαρμογή (customisation) του συστήματος στην θεματική περιοχή, το γνωστικό πεδίο, των προς επεξεργασία κειμένων. Βασικό βήμα στην διαδικασία αυτή αποτελεί η βελτίωση και ο εμπλουτισμός των γλωσσικών πόρων (language resources) με την ενσωμάτωση της κατάλληλης ορολογίας. Η εφαρμογή μεθόδων αυτόματης εξαγωγής όρων προσφέρει μια έγκυρη, γρήγορη και χαμηλού κόστους λύση στην διαδικασία προσαρμογής.

Η εξαγωγή όρων βρίσκει πολλές εφαρμογές στο χώρο της επεξεργασίας φυσικής γλώσσας και ειδικά με τον διαρκώς αυξανόμενο όγκο ηλεκτρονικών κειμένων σήμερα:

- **δεικτοδότηση κειμένων (text indexing)** - οι εξαγόμενοι όροι χρησιμοποιούνται απευθείας στον κατάλογο όρων του κειμένου
- **κατηγοριοποίηση-ταξινόμηση κειμένων (text classification)** - κείμενα με παρόμοιους όρους ταξινομούνται στην ίδια θεματική περιοχή
- **ανάκτηση/εξαγωγή πληροφορίας (information retrieval/extraction)** - ο χρήστης αναζητά κείμενα που τον ενδιαφέρουν με τη μορφή ερωτήσεων αποτελούμενων από όρους-κλειδιά. Από όλα τα διαθέσιμα κείμενα επιστρέφονται μόνο αυτά που περιέχουν αυτούς τους συγκεκριμένους όρους
- **κατασκευή περίληψης (text abstracting / summarisation)** - οι προτάσεις που περιέχουν όρους του κειμένου είναι κατά κανόνα και οι σημαντικότερες προτάσεις, αυτές που υποδηλώνουν σαφέστερα το περιεχόμενό του.
- **παραλληλοποίηση κειμένων (text alignment)** - όροι της μιας γλώσσας αντιστοιχούν συνήθως σε έναν μόνο όρο μιας άλλης γλώσσας

2. Μεθοδολογικές προσεγγίσεις

Ως όρους ενός κειμένου ορίζουμε γενικά τις γλωσσικές πραγματώσεις των εννοιών ενός κειμένου. Δύο είναι οι βασικές μεθοδολογικές τάσεις στην εξαγωγή όρων σήμερα:

1. Με χρήση μιας ειδικά σχεδιασμένης γραμματικής όρων (συνήθως ελεύθερης συμφραζομένων), η οποία εφαρμόζεται σε κείμενα κατάλληλα γραμματικά σχολιασμένα και εξάγει όσες φράσεις αναγνωρίζονται από αυτή τη γραμματική [1] .

2. Με χρήση στατιστικών εργαλείων αντίστοιχων με αυτά που χρησιμοποιούνται για εφαρμογές ανάκτησης πληροφοριών και δεικτοδότησης κειμένων. Στα εργαλεία αυτά περιλαμβάνονται μετρήσεις συχνοτήτων, μετρικές από τη θεωρία πληροφορίας, μετρικές που υπολογίζουν τα συμφραζόμενα των λέξεων κ.α.[2], [8]

Αξίζει να σημειωθούν κάποιες διαφορές ανάμεσα στις δύο αυτές μεθόδους. Η γραμματική όρων περιγράφει τη συντακτική δομή που πρέπει να ικανοποιεί κάθε έγκυρος όρος, χωρίς να αποκλείεται το ενδεχόμενο αυτές οι συντακτικές δομές να ικανοποιούνται και από άλλες ακολουθίες λέξεων που δεν θεωρούνται σωστοί όροι. Αν, για παράδειγμα, ένας από τους κανόνες περιγράφει ότι ένα επίθετο και ένα ουσιαστικό συγκροτούν έναν όρο, η εφαρμογή της γραμματικής στην προηγούμενη πρόταση θα επιστρέψει ως αποτέλεσμα τις φράσεις "συντακτικές δομές", "έγκυρος όρος" και "σωστοί όροι". Για τη θεματική κατηγορία του παρόντος κειμένου, ο πρώτος όρος είναι αποδεκτός, ο δεύτερος αποδεκτός σε ευρύτερο πλαίσιο αλλά ο τρίτος όχι. Η αδυναμία της γραμματικής έγκειται στο ότι εφαρμόζει τους κανόνες της χωρίς διάκριση, περιγράφοντας την ικανή αλλά όχι και αναγκαία συνθήκη για να είναι μια ακολουθία λέξεων όρος. Επιπλέον μπορεί να εντοπίσει μόνο όρους με περισσότερες από μία λέξεις, μιας και μόνο σε αυτούς μπορεί να αποδοθεί συντακτική δομή. Συμπερασματικά, ο απώτερος στόχος μιας γραμματικής όρων είναι ο εντοπισμός σε ένα πρώτο στάδιο "υποψήφιων όρων".

Η στατιστική προσέγγιση στηρίζεται στην υπόθεση ότι οι όροι, ως λέξεις ή φράσεις που είναι χαρακτηριστικές της θεματικής περιοχής του κειμένου, έχουν

την τάση να εμφανίζονται συχνά. Η συχνότητα επιδέχεται δύο διαφορετικές ερμηνείες: (1) συχνότερα από ότι σε ένα κείμενο που δεν ανήκει στη συγκεκριμένη θεματική περιοχή και (2) απλά συχνότερα από τις άλλες λέξεις ή φράσεις του κειμένου. Με βάση αυτή τη συγκριτική αντίληψη, για κάθε φράση υπολογίζεται ένα βάρος που εκφράζει τη σημασία της για το κείμενο, εξαιρώντας τις γραμματικές λέξεις, άρθρα, αντωνυμίες, προθέσεις κλπ. οι οποίες εμφανίζουν αρκετά υψηλή συχνότητα σε οποιοδήποτε κείμενο αλλά δεν θεωρούνται όροι. Οι φράσεις για τις οποίες υπολογίζεται το μεγαλύτερο βάρος παρουσιάζουν την μεγαλύτερη πιθανότητα να είναι οι όροι του κειμένου. Στα χαρακτηριστικά της προσέγγισης αυτής είναι η δυνατότητα εντοπισμού μονολεκτικών όρων. Στα μειονεκτήματά της καταγράφεται η αδυναμία να εξάγει όρους που δεν ικανοποιούν τα στατιστικά κριτήρια, καθώς είναι πιθανό έγκυροι όροι να εμφανίζονται μόνο μία ή γενικά λίγες φορές στο κείμενο. Τέλος, η επιλογή της στατιστικής φόρμουλας επηρεάζει την αποδοτικότητα της προσέγγισης αυτής, με τρόπο ανάλογο με αυτόν που η καλυπτικότητα της γραμματικής επηρεάζει την προηγούμενη προσέγγιση.

Αλλες προσεγγίσεις συνδυάζουν την στατιστική επεξεργασία με την γλωσσολογική μοντελοποίηση [3], [4], [5], [6]. Πρόκειται για υβριδικά συστήματα που αρχικά δημιουργούν μια λίστα υποψήφιων όρων με τη βοήθεια γραμματικών και στη συνέχεια "φιλτράρουν" αυτούς τους όρους με στατιστικά εργαλεία ώστε να απομακρύνουν τους όρους εκείνους που ικανοποιούν μεν τη γραμματική, αλλά δεν είναι χαρακτηριστικοί της θεματικής περιοχής του κειμένου ώστε να αποτελούν έγκυρους όρους.

3. Περιγραφή της μεθόδου

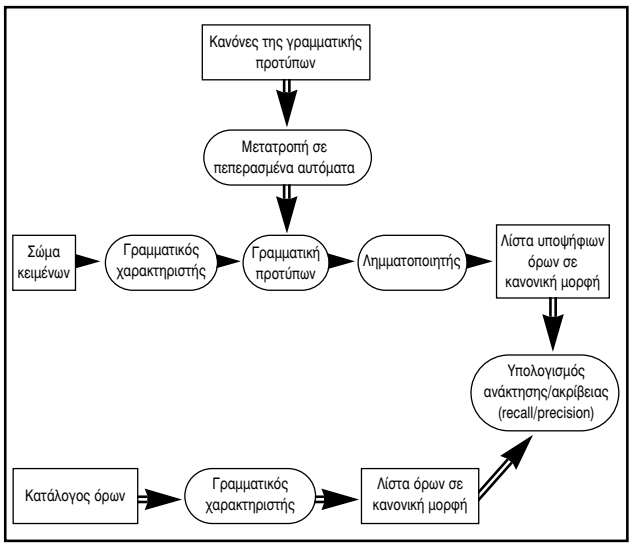
Η μέθοδος που περιγράφεται στο άρθρο αυτό έχει στόχο την εξαγωγή υποψήφιων όρων, η εγκυρότητα των οποίων θα ελεγχεί χειρωνακτικά. Τα βασικά στάδια της μεθόδου συνίστανται σε:

α. γραμματικό χαρακτηρισμό με βάση ένα μορφολογικό λεξικό και ένα σύστημα κανόνων για επίλυση μορφολογικών αμφισημιών

β. συντακτική ανάλυση με βάση μια γραμματική προτύπων

γ. λημματοποίηση με βάση το μορφολογικό λεξικό και την γραμματική κατηγορία που προκύπτει από τον γραμματικό χαρακτηρισμό.

Το διάγραμμα ροής της μεθόδου απεικονίζεται στο παρακάτω σχήμα:



Η **γραμματική** που χρησιμοποιήθηκε για την συντακτική ανάλυση είναι ένα υποσύνολο της γραμματικής προτύπων που παρουσιάστηκε στο [7]. Πρόκειται για μια γραμματική που χρησιμοποιεί το φορμαλισμό ενοποίησης (feature-structure unification) και τελεστές κανονικών εκφράσεων - γραμματικών (regular expressions).

Από την γραμματική του [7] που αριθμούσε 77 κανόνες κωδικοποιήθηκε ένα υποσύνολο που αναγνωρίζει δίλεκτους και τρίλεκτους όρους. Κάθε κανόνας μετατράπηκε σε ένα **πεπερασμένο αυτόματο** (finite-state automaton) ενισχυμένο (1) με δυνατότητες ενοποίησης συντακτικών χαρακτηριστικών και (2) με τελεστές κανονικών εκφράσεων. Τα χαρακτηριστικά αυτά, όπως φαίνεται από το παράδειγμα, μπορεί να είναι η γραμματική κατηγορία (ουσιαστικό, άρθρο, επίρρημα, κλπ.) ή χαρακτηριστικά υποκατηγοριοποίησης όπως γένος, πτώση, αριθμός, έγκλιση, φωνή κλπ. Οι τελεστές κανονικών εκφράσεων περιλαμβάνουν τελεστές όπως προεραιτικότητα, επανάληψη, διάζευξη κλπ.

Το **σώμα κειμένων** που χρησιμοποιήθηκε για την εφαρμογή της μεθόδου είναι ένα εγχειρίδιο οδηγιών

της Hewlett-Packard μεγέθους περίπου 90000 λέξεων. Το κείμενο αυτό επιλέχτηκε επειδή συμπεριλάμβανε έναν **κατάλογο όρων** έναντι του οποίου αξιολογούνται τα αποτελέσματα της μεθόδου. Κατά την αξιολόγηση χρησιμοποιείται η **κανονική μορφή** των όρων στην οποία κάθε λέξη αντικαθίσταται από το λήμμα της.

4. Αποτελέσματα - εκτιμήσεις

Η αξιολόγηση των αποτελεσμάτων βασίστηκε στην σύγκριση των όρων που εξάγει η μέθοδος με τους όρους που απαρτίζουν τον κατάλογο όρων που συνόδευε το κείμενο. Προηγουμένως όλοι οι όροι μετασχηματίστηκαν σε μια κανονικοποιημένη μορφή η οποία περιλαμβάνει μόνο τα λήμματα των λέξεων. Με αυτόν τον τρόπο ταυτίστηκαν όροι που περιείχαν τις ίδιες λέξεις ελάχιστα διαφοροποιημένες, π.χ. στην πτώση. Για παράδειγμα, ο όρος "δείκτης επιλογής" του καταλόγου όρων απαντάται στο κείμενο μόνο ως."δείκτη επιλογής"

Εξαιρώντας τους μονολεκτικούς όρους, το ê_ôάλογος όρων του κειμένου περιείχε συνολικά 214 όρους. Η μέθοδος εξήγαγε 4729 όρους από τους οποίους 124 περιλαμβάνονταν στους 214 σωστούς όρους. Υπολογίστηκαν έτσι:

ποσοστό ανάκτησης (recall) 124/214 = 58%
ποσοστό ακρίβειας (precision) 124/4729= 2,6%.

Το ποσοστό ανάκτησης κρίνεται ικανοποιητικό. Μελέτη των όρων που δεν εντοπίστηκαν έδειξε ότι το 17% από αυτούς περιείχε μη ελληνικές λέξεις, λέξεις που δεν περιέχονταν στο λεξικό του γραμματικού χαρακτηριστή ή λέξεις για τις οποίες ο γραμματικός χαρακτηριστής απέδιδε λανθασμένη γραμματική κατηγορία. Ποσοστό 8,8% ήταν όροι αποτελούμενοι από 4 λέξεις, ενώ η γραμματική περιελάμβανε κανόνες κάλυψης όρων μέχρι 3 λέξεων. Αντίθετα, το ποσοστό ακρίβειας είναι χαμηλό, γεγονός αναμενόμενο που αποδίδεται στην εγγενή ιδιότητα των γραμματικών να παράγουν περισσότερες υποψήφιες φράσεις επειδή οι κανόνες τους είναι γενικοί και παραμένουν πάντα στο συντακτικό επίπεδο.

Η παρούσα γραμματική προτύπων μπορεί να εμπλουτιστεί με επιπλέον χαρακτηριστικά που θα βελτιώσουν την αποδοτικότητά της. Σε αυτά περιλαμβάνονται:

- Η στατιστική επεξεργασία (με μεθόδους όπως: μετρήσεις συχνοτήτων, υπολογισμός βάρους με TFIDF [10], NC-value [6], log-likelihood, mutual information [2]) των όρων που εξάγει η γραμματική ώστε να προκριθούν οι έγκυροι όροι του κειμένου.

- Η κωδικοποίηση σε πεπερασμένο αυτόματο κανόνων που αναγνωρίζουν όρους μεγαλύτερου μήκους.

- Η χρήση μόνο του μέγιστου σε κάλυψη όρου, σε περίπτωση που αυτός εμπεριέχει μικρότερους σε μήκος όρους. Κατ'αυτόν τον τρόπο οι ανακτώμενοι όροι μειώνονται σημαντικά.

- Ο αποκλεισμός των λειτουργικών λέξεων (functional words) από τους όρους κατά τη διαδικασία αξιολόγησης.

- Η χρησιμοποίηση επιπλέον συντακτικής πληροφορίας (όπως η κεφαλή στις ονοματικές φράσεις) ώστε να ταυτίζονται ονοματικές φράσεις με το ίδιο περιεχόμενο αλλά με διαφορετική σειρά λέξεων (π.χ. εταιρίες κατασκευών, κατασκευαστικές εταιρίες).

5. Αναφορές

[1] Bourigault D. (1992).
Surface Grammatical
Analysis for the Extraction of Terminological Noun Phrases.
Proceedings of the 14th International Conference on Computational Linguistics.

[2] Church K. W. and Hunks P. (1990)
Word Association, Norms, Mutual Information, And Lexicography Computational Linguistics, Vol 16, Number 1.

[3] Dagan I. and Church K. W. (1994) Termight: Identifying and Translating Techical Terminology. Proceedings of the EACL 1994.

[4] Daille B., Gaussier E., Lange J. M.,(1994)
Towards automatic extraction of monolingual and bilingual terminology,
Proceedings of COLING 94,
pp 515-521.

[5] Daille B. (1994),
Study and implementation of combined techniques for automatic extraction of Terminology.
in The Balancing Act: Combining Symbolic and Statistical Approaches to Languages,
Workshop at the 32nd Annual Meeting of ACL,
Las Cruces, Nouveau Mexique.

[6] Frantzi, K.T. and Ananiadou, S. (1997)
Automatic term recognition using contextual clues,
Proceedings of Mulsaic 97, IJCAI, Japan

[7] Gavriilidou M, Lambropoulou P.
Report on the Constituent Grammar,
RENOS project,
LREI- 62-048, Athens, 1994

[8] Hatcher A.J. (1960)
An introduction to the analysis of English noun compounds.
In Word, 16, 356-373.

[9] Smadja F. A. and McKeown K. R. (1990)
Automatically Extracting and Representing Collocations For Language Generation,
Proceedings of the 28th annual Meeting of the ACL.

[10] Salton, G. (1989), Automatic text processing: the transformation, analysis, and retrieval of information by computer,
Reading, Mass. Wokingham: Addison-Wesley.

5. Ανάκτηση παραδειγματικών προτάσεων στο πλαίσιο σύγχρονων μεθόδων μετάφρασης

Χρήστος Μαλαβάζος, Στέλιος Πιπερίδης
Ινστιτούτο Επεξεργασίας Λόγου
Αρτέμιδος 6 και Επιδαύρου, Παράδεισος Αμαρουσίου,
151 25 Αθήνα
email : spip, christos@ilsp.gr

Abstract
This paper describes **TR•AID**, a multi-level architecture for a computer-aided translation (CAT) system platform. The system employs different levels of information and processing in an attempt to maximize past translation reuse as well as terminology and style consistency in the translation of specific types of text. Such tools have come in the bibliography under the term Translation Memory (TM) tools.

1. Εισαγωγή

Η πρόταση αξιοποίησης τεχνικών αναγνώρισης προτύπων και μηχανικής μάθησης στην περιοχή της αυτόματης μετάφρασης (Nagao 84) και η επανάκαμψη των στατιστικών μεθόδων στις αρχές της δεκαετίας 90 (Brown et al. 93) έχουν αναζωπυρώσει το ενδιαφέρον της συζήτησης σχετικά με την αρχιτεκτονική αλλά και την σύσταση των σύγχρονων συστημάτων μηχανικής μετάφρασης. Η επεξεργασία παράλληλων δίγλωσσων κειμένων και ιδιαίτερα η στοίχισή τους (alignment) με στόχο την εξαγωγή παραδειγματικών προτάσεων μετάφρασης και σκοπό την επαναχρησιμοποίησή τους στην μετάφραση άλλων, νέων, κειμένων έχει οδηγήσει στην δημιουργία ενός νέου ρεύματος στην μηχανική μετάφραση.

Τα παραδοσιακά συστήματα μετάφρασης, συστήματα βασισμένα σε κανόνες (rule-based machine translation systems) αντιμετωπίζουν προβλήματα τόσο ποιότητας και προσαρμογής σε νέες θεματικές περιοχές όσο και ταχύτητας. Η μετάφραση βάσει παραδειγμάτων (example-based machine translation), συχνά αναφερόμενη ως μετάφραση βάσει μνήμης (memory-based machine translation), προτείνει εναλλακτικούς τρόπους αντιμετώπισης των παραπάνω

προβλημάτων και κυρίως της απόκτησης γνώσης (knowledge acquisition), μειώνοντας το βαθμό αυτοματοποίησης και προτείνοντας μοντέλα ανάκτησης μεταφραστικών παραδειγμάτων στο πλαίσιο της μετάφρασης με τη βοήθεια υπολογιστή.

2. Μεθοδολογία

Η διαδικασία της μετάφρασης ενός κειμένου συχνά χαρακτηρίζεται από τρεις παραμέτρους: επανάληψη, σημαντικές απαιτήσεις σε ποιότητα, συνέπεια και αποτελεσματικότητα. Αυτό ισχύει ειδικά στην μετάφραση διοικητικών και τεχνικών κειμένων, και ακόμα περισσότερο στην περίπτωση νομικών κειμένων (συμβολαίων, κανονισμών, κλπ), όπως επίσης στα εγχειρίδια χρήσης προϊόντων και υπηρεσιών. Η εκτίμηση για τα συγκεκριμένα είδη κειμένων είναι ότι η επανάληψη τμημάτων τους μπορεί σε πολλές περιπτώσεις να υπερβεί το 70%.

Ο σκοπός ενός εργαλείου μεταφραστικής μνήμης είναι να αντιμετωπίσει τα προβλήματα αυτά προσφέροντας ένα υπολογιστικό περιβάλλον το οποίο:

- θα απαλλάσσει τους μεταφραστές από την μετάφραση των επαναλαμβανόμενων τμημάτων, χρησιμοποιώντας ήδη μεταφρασμένα κείμενα
- θα αυξάνει την ποιότητα και την συνέπεια της μετάφρασης, έχοντας την δυνατότητα να ενσωματώσει βοηθητικά μεταφραστικά εργαλεία.

Μια κατάλληλη οργάνωση και αποθήκευση τμημάτων δίγλωσσων κειμένων στη γλώσσα-πηγή (Γ.Π.) και στη γλώσσα-στόχο (Γ.Σ.), καθώς και ένα σύνολο εργαλείων για την ανάκτηση εφαρμόσιμων λύσεων (μεταφράσεων) θα μπορούσαν να αυξήσουν σημαντικά την παραγωγικότητα ενός μεταφραστή, και ταυτόχρονα θα βελτίωναν την ποιότητα και την συνέπεια της μετάφρασης (Freibott 92, Ishida 94).

Η προσέγγιση αυτή βασίζεται σε τέσσερις βασικούς άξονες:

- αυτόματη στοίχιση παράλληλων κειμένων, δηλ. προσδιορισμό μεταφραστικά ισοδύναμων μονάδων (προτάσεων, φράσεων, λέξεων) παραλλήλων κειμένων
- οργάνωση παράλληλων και πολύγλωσσων σωμάτων κειμένων, δηλαδή κειμένων με το ίδιο περιεχόμενο σε διαφορετικές γλώσσες (το ένα μετάφραση του άλλου), έτσι ώστε να καθίσταται δυνατή η αποτελεσματική αποθήκευση και ανάκτηση μεταφραστικών και ορολογικών δεδομένων
- ανάπτυξη έξυπνων τεχνικών ταιριάσματος κειμένων με σκοπό την γρήγορη σύγκριση και ανάκτηση των καταλληλότερων για την μετάφραση προτασιακών παραδειγμάτων
- ανάπτυξη έξυπνων τεχνικών εντοπισμού και μετάφρασης ορολογίας

Στο πλαίσιο της συγκεκριμένης αρχιτεκτονικής έχουν εξετασθεί εναλλακτικές τεχνικές για κάθε υποσύστημα. Η συγκριτική μελέτη των τεχνικών αυτών έχει αναδείξει τις πλέον αποτελεσματικές λύσεις οι οποίες έχουν ολοκληρωθεί στο περιβάλλον του συστήματος **TR•AID (Translation Aid)**.

Στο σχήμα 1 παρουσιάζονται τα εργαλεία και ο τρόπος ολοκλήρωσης τους και επικοινωνίας στο ολοκληρωμένο περιβάλλον TrAID.

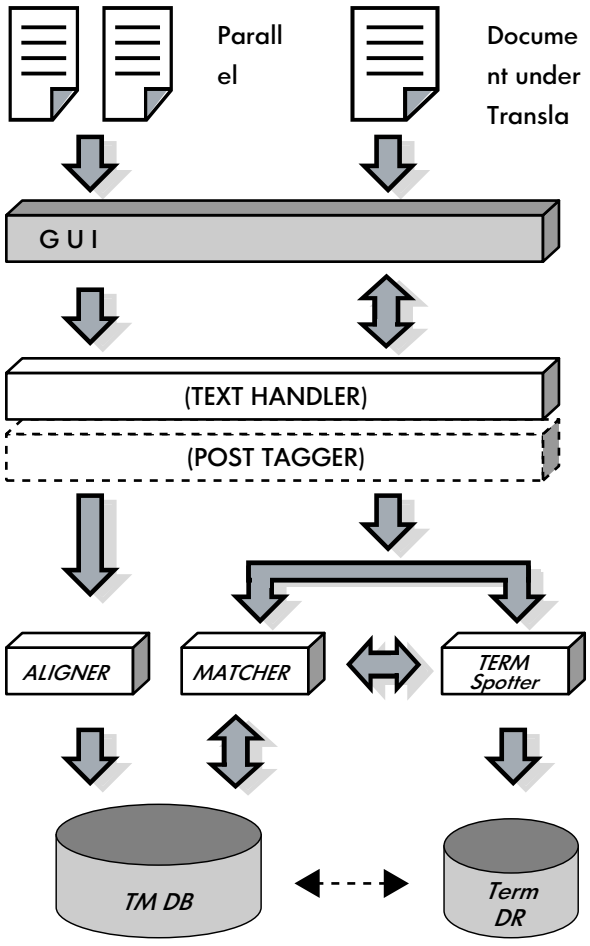


Figure 1: TrAID

3. Συμπεράσματα

Η πραγματική προστιθέμενη αξία των λογισμικών πακέτων μετάφρασης έγκειται στην δυνατότητά τους αφενός να βελτιώνουν την αποτελεσματικότητα της μεταφραστικής διαδικασίας μέσω της μείωσης του κόστους και του χρόνου μετάφρασης και αφετέρου να αξιοποιούν την υπάρχουσα ποιοτική γνώση μετάφρασης σε μια θεματική περιοχή. Με αυτά τα κριτήρια η πλήρως αυτόματη μετάφραση θεωρείται ακόμη μη εφικτή.

Ο στόχος, επομένως, συνίσταται στον βέλτιστο συνδυασμό εργαλείων και γλωσσικών πόρων διαφορετικών επιπέδων εξειδίκευσης και πολυπλοκότητας που θα έχουν την δυνατότητα να επεξεργάζονται και να μεταφράζουν κείμενα σε διαφορετικές γλώσσες και διαφορετικές γνωστικές περιοχές.

ΙΕΛ - Ινστιτούτο Επεξεργασίας του Λόγου

"Ανάκτηση παραδειγματικών προτάσεων στο πλαίσιο σύγχρονων μεθόδων μετάφρασης"

Τμήμα Γλωσσικών Εφαρμογών

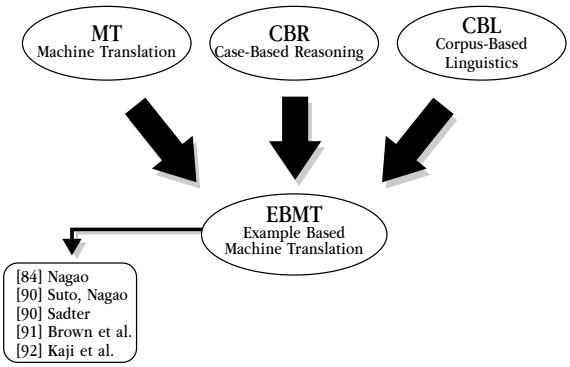
1

Τάσεις ...

- **Information Management**
  Αποδοτική διαχείριση πληροφορίας
- **Corpus Linguistics**
  Αναζωπύρωση του ερευνητικού ενδιαφέροντος στις αρχές της δεκαετίας του '80
- **Machine Assisted Translation**
- **Examplle Based Translation - EBMT**
  Μετάφραση βασισμένη σε παραδείγματα (Nagao, Kaji et al., IBM Team, ...)

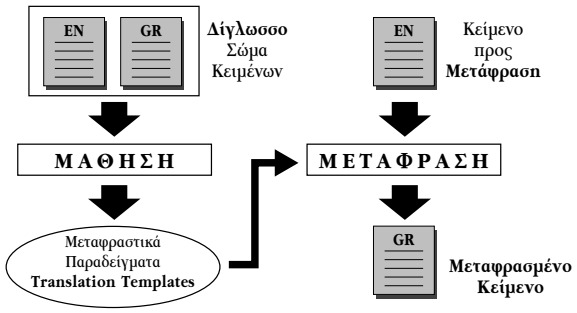
2

Μετάφραση Βασισμένη σε Παραδείγματα



3

Μετάφραση Βασισμένη σε Παραδείγματα



4

RBMT-EBMT Σημεία Σύγκρισης ...

- **RBMT - EBMT**
  (Μετάφραση Βάσει Κανόνων - Μετάφραση Βάσει Παραδειγμάτων)
  - Ποιότητα Μετάφρασης
  - Απαιτούμενη Γλωσσολογική Προεπεξεργασία (μορφολογία, σύνταξη, σημασιολογία)
  - Επεκτασιμότητα - Μάθηση
  - Δυνατότητα υπολογισμού αξιοπιστίας αποτελεσμάτων

5

Στόχοι

Έρευνα εναλλακτικών τρόπων μετάφρασης

⬇

Σχεδιασμός & ανάπτυξη "πιλοτικών" πρωτοτύπων

⬇

Ανάπτυξη Προϊόντος ➡ ***Tr•AID***

6

Μεθοδολογία

- ανάπτυξη (ημι-)αυτόματων μεθόδων εντοπισμού "αντιστοίχων τμημάτων" κειμένου μέσα από "παράλληλα κείμενα". **(alignment)**
- σχεδιασμός βελτίωσης αρχιτεκτονικής και ανάπτυξη μηχανισμών για αποθήκευση των "παραλλήλων τμημάτων" ως μεταφραστικών παραδειγμάτων. **(DB storage)**
- ανάπτυξη μηχανισμού για ανάκτηση καταλλήλων παραδειγμάτων μετάφρασης, από τα ήδη αποθηκευμένα. **(matching & retrieval)**
- Συνσυασμός των ανωτέρω σε ένα ενιαίο περιβάλλον εργασίας με ελάχιστες απαιτήσεις τεχνικών γνώσεων. **(GUI)**

7

Σημεία Έρευνας

- **Παραλληλοποίηση (alignment)**
  Τι θεωρούμε ως "μόριο" (αδιαίρετη-"μεταφράσιμη" μονάδα) κειμένου;
  Παράγραφοι - Προτάσεις - "Φράσεις" - Λέξεις ...
- **Σχεδιασμός ΒΔ Μεταφραστικών Παραδειγμάτων**
  Αναπαράσταση Πληροφορίας
  Ταχύτητα πρόσβασης
- **Μηχανισμός Ταιριάσματος & Ανάκτησης Προτάσεων (Matching & Retrieval)**
  Μετρική ομοιότητας;
  Παραμετρική ανίχνευση

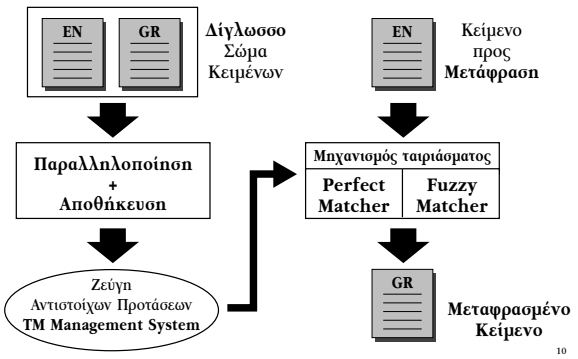
8

Συστατικά Συστήματος

- Σύστημα προ-επεξεργασίας κειμένων (Handler)
  Λεκτική ανάλυση, αναγνώριση προτάσεων, αναγνώριση λεκτικών μονάδων ιδιαίτερης σημασίας για την μετάφραση (αριθμοί, ημερομηνίες, συντμήσεις ...). Τεχνολογία πεπρασμένων αυτομάτων
- Παραλληλοποιητής (Aligner)
  Στατιστικό-αριθμητικό σύστημα "ανεξάρτητο γλωσσών"
  Ολοκληρωμένο περιβάλλον εργασίας
- Μηχανισμός ταιριάσματος Προτάσεων (Matcher)
  Pattern Matching, Παραμετρικός
- Σχεσιακό Μοντέλο Βάσης Μεταφρ. Παραδειγμάτων
  RDBMS
  Μηχανισμοί αποθήκευσης παραδειγμάτων
  "Ανοικτή" αρχιτεκτονική

9

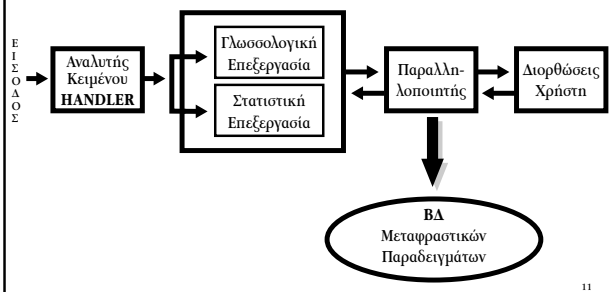
Μετάφραση Βασισμένη σε Παραδείγματα



10

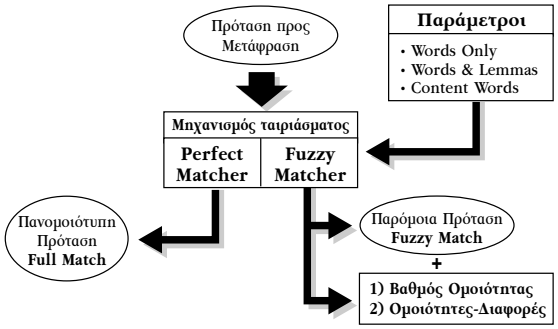
Δημιουργία ΒΔ Μεταφραστικών Παραδειγμάτων

Προ-επεξεργασία Κειμένου ...



11

Μηχανισμός Ταιριάσματος



12

Μερικό Ταίριασμα (Fuzzy Matching)

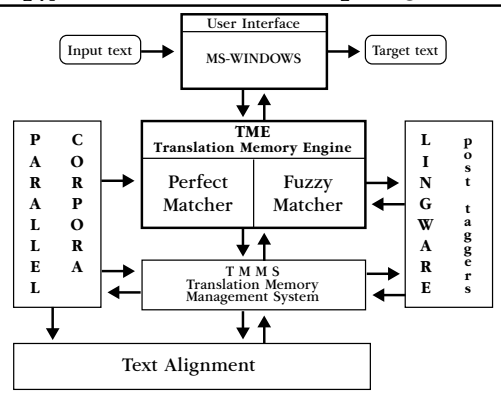
Πιθανοί Συνδυασμοί μεταξύ Τμημάτων Κειμένου ...

IS: Sa Sb	*IS*: Sa Sb **Sc**
SL: Sa Sb **Sc**	*SL*: Sa Sb
IS: Sa Sc	*IS*: Sa **Sb** Sc
SL: Sa **Sb** Sc	*SL*: Sa Sc
IS: Sa Sb **Sc**	*IS*: Sa **Sb Sc**
SL: Sa Sb **Sd**	*SL*: Sa **Sc Sb**

Αύξηση διαφοράς => Μείωση βαθμού Ομοιότητας

13

Αρχιτεκτονική του Συστήματος



14

Ανοικτά θέματα ...

- **Βελτίωση του συστήματος παραλληλοποίησης**
  Μελέτη διαφορετικών προσεγγίσεων
- **Μονάδες μετάφρασης μικρότερες της πρότασης**
- **Μεταφραστική Μνήμη & Ορολογία**

15

Βιβλιογραφία

(Brown et al. 91) P. F Brown, J. C. Lai, R. L. Mercer, *Aligning Sentences in Parallel Corpora*. Proc. of the 29th Annual Meeting of the ACL, pp 169-176, 1991.

(Brown et al. 93) P. F Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, *The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics,* June 1993.

(Catizone et al. 89) R. Catizone, G. Russell, S. Warwick, *Deriving translation data from bilingual texts*, Proc. of the First Lexical Acquisition Workshop, Detroit 1989

(Chanod & Tapanainen 96) J. P. Chanod and P. Tapanainen. *A non-deterministic tokenizer for finite-state parsing*, Proceedings of the ECAI 96 Workshop, 1996.

(Cranias et al. 94) L. Cranias, H. Papageorgiou and S. Piperidis, *A matching technique in Example-Based Machine Translation*, Proc. of Coling, pp 100-105.

(Frakes 84) W. B. Frakes *Term Conflation for Information Retrieval*. Research and Development in Information Retrieval, New York: Cambridge University Press, 1984.

(Freibott 92) G.P. Freibott, *Computer Aided Translation in an Integrated Document Production Process:Tools and Applications,* Translating and the Computer 14, pp 45-66, 1992.

(Furuse & Iida 92) O. Furuse and H. Iida, *Cooperation between Transfer and Analysis in Example-Based Framework.* Proc. Coling, pp 645-651, 1992.

(Gale & Church 91) W. A. Gale and K. W. Church *A Program for Aligning Sentences in Bilingual Corpora*. Proc. of the 29th Annual Meeting of the ACL., pp 177-184, 1991.

(Grefenstette & Tapanainen 94) G. Grefenstette and P. Tapanainen *What is a word, What is a sentence? Problems of tokenization*, COMPLEX 94.

(Ishida 94) R. Ishida, (1994), *Future translation workbenches: some essential requirements,* Aslib Proceedings, vol.46, no. 6, pp 163-170, June 1994.

(Kaji et al. 92) H. Kaji, Y. Kida and Y. Morimoto, *Learning Translation Templates from Bilingual Text.* Proc. Coling., pp 672-678, 1992.

(Kay & Roscheisen 91) M. Kay, M. Roscheisen, *Text-Translation Alignment,* Computational Linguistics Vol. 19, No 1, 1991.

(Nagao 84) M. Nagao, *A framework of a mechanical translation between Japanese and English by analogy principle.* Artificial and Human Intelligence, ed. Elithorn A. and Banerji R., North-Holland, pp 173-180, 1984.

(Ney 84) H. Ney, *The use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition,* IEEE vol. ASSp-32, No 2, 1984.

(Nirenburg et al. 93) S. Nirenburg, C. Domashnev D. J. Grannes. *Two Approaches to Matching in Example-Based Machine Translation*. Proc. of TMI-93, Kyoto, Japan, 1993.

(Palmer & Hearst 94) D. Palmer and M. A. Hearst, *Adaptive sentence boundary disambiguation*, Report No. UCB/CSD 94/797.

(Papageorgiou et al. 94) H. Papageorgiou, L. Cranias and S. Piperidis, *Automatic alignment in parallel corpora*, Proc. of the 32nd Annual Meeting of the ACL, 1994.

(Reynar & Ratnaparkhi 97) J. C. Reynar and A. Ratnaparkhi, *A maximum entropy approach to identifying sentence boundaries*, Computational Linguistics Archive cmp-lg/9704002, 1997.

(Sadler & Vendelmans 90) V. Sadler and R. Vendelmans, *Pilot Implementation of a Bilingual Knowledge Bank.* Proc. of Coling, pp 449-451, 1990.

(Sato 92) S. Sato, CTM: *An Example-Based Translation Aid System*. Proc. of Coling, pp 1259-1263, 1992.

(Sato & Nagao 90) S. Sato and M. Nagao, *Toward Memory-based Translation.* Proc. of Coling, pp 247-252, 1990.

(Simard et al. 92) M. Simard, G. Foster and P. Isabelle, *Using cognates to align sentences in bilingual corpora*, Proc. of TMI, 1992.

(Sumita & Iida 91) E. Sumita and H. Iida, *Experiments and Prospects of Example-based Machine Translation*. Proc. of the 29th Annual Meeting of the Association for Computational Linguistics, pp 185-192, 1991.

(Sumita & Tsutsumi 88) E. Sumita and Y. Tsutsumi, *A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching*. TRL Research Report, Tokyo Research Laboratory, IBM, 1988.

6. Activities of NCSR "Demokritos" in Information Extraction

Dr. Constantine D. Spyropoulos
NCSR "Demokritos", Institute of Informatics & Telecommunications
Tel: 01-6503196,
Fax: 01-6532175,
{costass}@iit.demokritos.gr

Abstract

This article presents the research and development activities of the Institute of Informatics & Telecommunications of NCSR "Demokritos" in Information Extraction. More specifically it describes briefly the work in the information extraction projects ECRAN and GIE in which the Institute is currently involved and presents a relevant Conference Session that the Institute organised recently.

Δραστηριότητες του ΕΚΕΦΕ "Δημόκριτος" στην Εξαγωγή Πληροφορίας από Κείμενα

Δρ. Κων/νος Δ. Σπυρόπουλος
Ε.Κ.Ε.Φ.Ε. "Δημόκριτος",
Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών
Τηλ: 01-6503196,
Fax: 01-6532175,
{costass}@iit.demokritos.gr

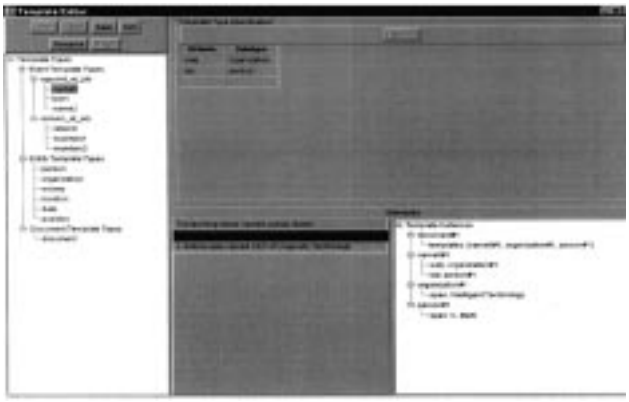
Τα συστήματα εξαγωγής πληροφορίας επεξεργάζονται κειμενικές βάσεις δεδομένων παρουσιάζοντας στους χρήστες συγκεκριμένες πληροφορίες που τους ενδιαφέρουν και όχι μόνο τα σχετικά κείμενα όπως συμβαίνει με τα συστήματα *ανάκτησης* και *φιλτραρίσματος πληροφορίας* (information retrieval and filtering) [Gaizauskas & Wilks, 1997]. Τα υπάρχοντα συστήματα *εξαγωγής πληροφορίας* εξάγουν προκαθορισμένους τύπους πληροφορίας από κείμενα μιας θεματικής περιοχής, γραμμένα σε μια συγκεκριμένη γλώσσα. Για να μπορέσει η τεχνολογία εξαγωγής πληροφορίας να χρησιμοποιηθεί στην πράξη σε βιομηχανικές και εμπορικές εφαρμογές, τα συστήματα εξαγωγής πληροφορίας πρέπει να είναι σε θέση να προσαρμόζονται εύκολα σε νέες θεματικές περιοχές και στα ενδιαφέροντα νέων χρηστών, καθώς επίσης και σε νέες γλώσσες.

Στους τομείς αυτούς εστιάζεται και το ενδιαφέρον του Ινστιτούτου Πληροφορικής & Τηλεπικοινωνιών (Ι.Π.&Τ.) του ΕΚΕΦΕ "Δημόκριτος", μέσα από τη συμμετοχή του σε ερευνητικά προγράμματα. αλλά και από τις πρωτοβουλίες που αναπτύσσει με τη διοργάνωση σχετικών ημερίδων [Karkaletsis & Spyropoulos, 1998]. Συγκεκριμένα το Ι.Π.&Τ. συμμετέχει στα ερευνητικά προγράμματα εξαγωγής πληροφορίας *ECRAN* και *GIE*, τα οποία βρίσκονται σε εξέλιξη, ενώ διοργάνωσε πρόσφατα και Ειδική Συνεδρία στο πλαίσιο του Διεθνούς Συνεδρίου EURISCON'98 με θέμα *"Adaptive and Multilingual Information Extraction - AMIE".*

Το ECRAN είναι ερευνητικό έργο χρηματοδοτούμενο από το πρόγραμμα TELEMATICS - Language Engineering της ΕΕ, με στόχο την ανάπτυξη συστημάτων εξαγωγής πληροφορίας από κειμενικές βάσεις δεδομένων (της Αγγλικής, Ιταλικής και Γαλλικής γλώσσας), τα οποία προσαρμόζονται εύκολα σε νέες θεματικές περιοχές και χρήστες. Ανάδοχος του έργου είναι η εταιρία Thomson-CSF Γαλλίας, ενώ οι υπόλοιποι εταίροι είναι: Universita di Ancona και Universita di Roma Tor Vergata Ιταλίας, Smart Information Services (SIS) GmbH Γερμανίας, Universite de Fribourg Ελβετίας, και University of Sheffield Αγγλίας. Το Ι.Π.&Τ. συμμετείχε, στην 1η φάση του έργου, στην ανάπτυξη ενός ερευνητικού συστήματος μοντελοποίησης χρηστών (User Modeling in Information Extraction - UMIE) [Benaki et al., 1997] [Karkaletsis et al, 1997]. Το UMIE έχει υλοποιηθεί ως εφαρμογή του Διαδικτύου (http://www.iit.demokritos.gr/UMIE) και μπορεί να συνεργάζεται με συστήματα ανάκτησης και εξαγωγής πληροφορίας. Στο συγκεκριμένο έργο το UMIE συνεργάζεται με το σύστημα εξαγωγής πληροφορίας του ECRAN. Οι χρήστες του WWW επικοινωνούν με το σύστημα αυτό μέσω του UMIE, το οποίο συγκεντρώνει αρχικά πληροφορία για προσωπικά στοιχεία του χρήστη (τμήμα εταιρίας στο οποίο εργάζεται) καθώς και πληροφορία για τα ενδιαφέροντά του στις θεματικές περιοχές που καλύπτει το σύστημα. Η πληροφορία που συγκεντρώνεται αποθηκεύεται στη βάση μοντέλων χρηστών. Στη συνέχεια η πληροφορία που υπάρχει στο μοντέλο κάθε χρήστη εμπλουτίζεται ανάλογα με τις επιλογές του κατά την αλληλεπίδρασή του με το UMIE. Η πληροφορία που έχει εξαχθεί από το σύστημα εξαγωγής πληροφορίας δρομολογείται σε κάθε χρήστη ανάλογα με την πληροφορία που υπάρχει στο μοντέλο του. Κατά τη διάρκεια της 2ης φάσης του έργου, το Ι.Π.&Τ. σε συνεργασία με το University of Sheffield ανέλαβε και ανέπτυξε ερευνητικά συστήματα για την προσαρμογή της τεχνολογίας εξαγωγής πληροφορίας σε νέες θεματικές περιοχές, τα οποία αξιολογούνται από τις εταιρείες Thomson και SIS που συμμετέχουν στο έργο. Συγκεκριμένα αναπτύχθηκαν τα ακόλουθα: σύστημα για την δημιουργία προτύπων (patterns) και πλαισίων (templates) για νέες θεματικές περιοχές [Karkaletsis *et al*, 1998b], σύστημα αποσαφήνισης εννοιών λέξεων με χρήση τεχνικών μηχανικής μάθησης και αξιοποίηση του θησαυρού WordNet και του λεξικού LDOCE [Paliouras *et al* 1998b], και σύστημα εκμάθησης κανόνων αναγνώρισης ονοματικών οντοτήτων [Paliouras *et al* 1998a]. Το σύστημα δημιουργίας προτύπων και πλαισίων περιγράφεται στο [Karkaletsis *et al*, 1998b] και αποτελείται από ένα φιλικό περιβάλλον χρήσης για την επιλογή των γεγονότων που ενδιαφέρουν και τον καθορισμό των ρόλων διαφόρων οντοτήτων για κάθε γεγονός. Το τμήμα που είναι αυτοματοποιημένο είναι η διαδικασία προεπιλογής και ταξινόμησης των ρημάτων, σύμφωνα με την πιθανότητά τους να αντιστοιχούν σε σημαντικούς τύπους γεγονότων για μια θεματική περιοχή, καθώς επίσης και η διαδικασία που κατασκευάζονται τα πρότυπα αφού ο χρήστης επιλέξει τα ρήματα που εκφράζουν γεγονότα που τον ενδιαφέρουν. Στόχος της μεθοδολογίας που ακολουθείται στο ECRAN είναι η δημιουργία συντακτικών προτύπων (syntactic patterns) για τα ρήματα, από παραδειγματικά κείμενα μιας θεματικής περιοχής, ο εμπλουτισμός τους με σημασιολογική πληροφορία (π.χ. ότι το υποκείμενο ενός ρήματος είναι πρόσωπο), και η ομαδοποίησή τους ανάλογα με τις συντακτικές και σημασιολογικές ομοιότητές τους [Basili *et al*, 1998], [Karkaletsis *et al,* 1998b].

Το αποτέλεσμα είναι ένα σύνολο από πρότυπα, τα οποία καθοδηγούν στη συνέχεια τη διαδικασία σχεδιασμού της βάσης πλαισίων, πάντα όμως κάτω από την επίβλεψη του χρήστη, ο οποίος καθορίζει την τελική μορφή των πλαισίων (βλ. Σχήμα 1). Τα πρότυπα και τα πλαίσια που έχουν κατασκευαστεί για μια θεματική περιοχή χρησιμοποιούνται στη συνέχεια από το σύστημα εξαγωγής πληροφορίας σε άγνωστα κείμενα της περιοχής αυτής. Το σύστημα εξαγωγής πληροφορίας είναι μια *"μηχανή" ταιριάσματος προτύπων (pattern matching engine)* που εντοπίζει τις φράσεις/προτάσεις από όπου θα εξαχθεί η πληροφορία την

οποία και αποθηκεύει στη βάση πλαισίων (templates base). Τα πρώτα αποτελέσματα από τη χρήση αυτού του συστήματος είναι ιδιαίτερα ενθαρρυντικά.



Σχήμα 1
Εργαλείο δημιουργίας πλαισίων

Το Ι.Π.&Τ. διαθέτει επίσης σημαντική πείρα στην προσαρμογή μεθόδων γλωσσικής τεχνολογίας για την Ελληνική. Η πείρα αυτή αποκτήθηκε με το διακρατικό ερευνητικό πρόγραμμα GIE (Greek Information Extraction) [GIE] [Karkaletsis et al, 1998a], το οποίο υλοποιείται σε συνεργασία με το University of Sheffield, και στα πλαίσια του οποίου προσαρμόστηκαν αρκετά εργαλεία της πλατφόρμας GATE (General Architecture for Text Engineering) [Cunnigham et al, 1996] για την εξαγωγή πληροφορίας από Ελληνικά κείμενα. Συγκεκριμένα:

- Συλλέχθηκαν τρία μεγάλα σώματα Ελληνικών κειμένων, τα οποία ασχολούνται με χρηματιστηριακά νέα, ειδήσεις μετακινήσεως προσωπικού σε εταιρίες και κείμενα από αποφάσεις της Ευρωπαϊκής Ένωσης.

- Υλοποιήθηκε διαχωριστής λέξεων και προτάσεων (tokeniser, sentence splitter).

- Υλοποιήθηκε σύστημα αναγνώρισης μερών του λόγου, με τη χρήση της μεθόδου μηχανικής μάθησης Brill tagger [Brill, 1993].

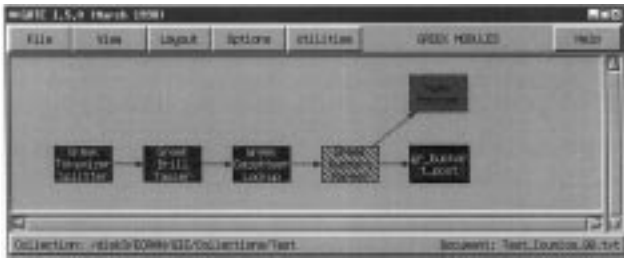
- Υλοποιήθηκε συντακτικός αναλυτής για την αναγνώριση ονοματικών φράσεων [Μανουσοπούλου, 1997].

- Χρησιμοποιήθηκαν τα εργαλεία του Αγγλικού συστήματος VIE [Humphreys et al., 1997] για την αναζήτηση σε ονοματικούς καταλόγους (gazetteer lookup) και την συντακτική ανάλυση με χρήση της γραμματικής ονοματικών οντοτήτων (named-entity bottom up chart parser). Τα εργαλεία αυτά χρησιμοποιήθηκαν με Ελληνικούς ονοματικούς καταλόγους και Ελληνική γραμματική αντίστοιχα.

- Κατασκευάστηκαν περιορισμένου μεγέθους ονοματικοί κατάλογοι της Ελληνικής.

- Κατασκευάστηκαν χειρωνακτικά γραμματικοί κανόνες για την αναγνώριση ονοματικών οντοτήτων σε κείμενα της Ελληνικής.

Το σύστημα αναγνώρισης ονοματικών οντοτήτων GIE, με ορισμένα από τα εργαλεία που αναφέρθηκαν παραπάνω, απεικονίζεται στο Σχήμα 2.



Σχήμα 2
Το σύστημα Αναγνώρισης Ονοματικών Οντοτήτων GIE

Το Ι.Π.&Τ. ανέλαβε επίσης πρόσφατα τη διοργάνωση της Ειδικής Συνεδρίας με τίτλο *"Towards Adaptive and Multilingual Information Extraction Systems - AMIE"* στο πλαίσιο του 3rd European Robotics, Intelligent Systems & Control Conference (EURISCON'98) (http://www.robotics.ece.ntua.gr/euriscon-softcom98.html), το οποίο διοργανώθηκε στην Αθήνα, στο διάστημα 22-25 Ιουνίου 1998. Στόχος του AMIE ήταν να φέρει σε επαφή Ευρωπαίους ερευνητές στην περιοχή της εξαγωγής πληροφορίας για να συζητήσουν τα ζητήματα αιχμής της προσαρμοστικότητας και πολυγλωσσικότητας. Στη διάρκεια του AMIE παρουσιάστηκαν οι ακόλουθες εργασίες:

- R.Basili, M.T.Pazienza. *Adaptive NLP-driven systems:acquisition of linguistic information for Informaton Extraction purposes*

- I.Blank. *Computer-aided analysis of multilingual*

patent texts.

- L.Dini. *Parallel IE Systems for Multilingual Information Gathering*
- V.DiTomaso, G.D'Angelo. *Information Extraction Techniques for Multilevel Sentence Matching*
- E.Karkaletsis, C.D.Spyropoulos, G.Petasis. *Named Entity Recognition from Greek Texts: the GIE Project*
- J.Kontos, I.Malagardi. *Question Answering and Information Extraction from Texts*
- S.Piperidis, B.Georgantopoulos. *Eliciting Terminological Knowledge for Information Extraction Applications*
- F.Vichot, F.Wolinski, H.C.Ferri, D.Urbani. *Using Information Extraction for Knowledge Entering*
- S. E. Michos, N. Fakotakis, and G. Kokkinakis *Using Functional Style Features to enhance Information Extraction from Greek Texts*

Οι εργασίες αυτές, μαζί και με τις εργασίες των υπολοίπων Sessions του EURISCON'98, θα αποτελέσουν κεφάλαια ενός βιβλίου που θα εκδοθεί σύντομα από τον διοργανωτή του Συνεδρίου Καθ. Σ. Τζαφέστα από την εκδοτική εταιρεία Kluwer, με τίτλο "Advances in intelligent systems: concepts, tools and applications".

Συνοψίζοντας, η τεχνολογία εξαγωγής πληροφορίας αποτελεί σημαντικό τμήμα των δραστηριοτήτων του Ι.Π.&Τ., το οποίο δίνει ιδιαίτερη έμφαση στην ανάπτυξη της τεχνολογίας αυτής για την Ελληνική γλώσσα και στην ανάπτυξη εργαλείων που διευκολύνουν την προσαρμογή της σε νέες θεματικές περιοχές και γλώσσες. Για την επίτευξη των στόχων αυτών το Ι.Π.&Τ. συνεργάζεται με Ευρωπαϊκούς οργανισμούς για την αξιολόγηση της τεχνολογίας εξαγωγής πληροφορίας και με Ελληνικούς οργανισμούς για την ανάπτυξη των απαραίτητων γλωσσικών πόρων για την Ελληνική γλώσσα. Η ομάδα του Ι.Π.&Τ. εκτός από τον υπογράφοντα αποτελείται από 3 μεταδιδακτορικούς νέους ερευνητές, 3 μεταπτυχιακούς υποψήφιους διδάκτορες και 2 τεχνικούς υποστήριξης έρευνας. Εκτός των ανωτέρω έργων, η ομάδα έχει υποβάλει διάφορες ερευνητικές προτάσεις για περαιτέρω χρηματοδότηση, οι οποίες βρίσκονται στο στάδιο της αξιολόγησης, με στόχο την ανάπτυξη ενός ολοκληρωμένου συστήματος εξαγωγής πληροφορίας για Ελληνικά κείμενα.

Βιβλιογραφία

[Benaki et al., 1997a] Benaki, E., Karkaletsis, V. and Spyropoulos, C.D. "Integrating User Modeling into Information Extraction: the UMIE Prototype", in Proceedings of the *6th International Conference on User Modeling (UM97)*, CISM No 383, Springer Wien New York, pp. 55-58, 1997.

[Brill 1993] Brill, E. "A corpus-based approach to language learning". *Doctoral Dissertation*, Univ. of Pennsylvania, 1993.

[Cunnigham et al., 1996] Cunnigham, H., Wilks, Y., Gaizauskas, R., 1996. *Gate – A General Architecture for Text Engineering*. Proceedings of COLING-96, Copenhagen, 1996.

[ECRAN] ECRAN: Extraction of Content: Research at Near-Market, Language Engineering Project (LE-2110). http://www2.echo.lu/langeng/en/le1/ecran/ecran.html

[GIE] GIE: Greek Information Extraction. http://www.iit.demokritos.gr/skel

[Humphreys et al., 1997] Humphreys, K., Gaizauskas, R., Cunningham, H., and Azzam, S. VIE technical Specifications. Department of Computer Science, University of Sheffield, 1997.

[Karkaletsis et al, 1997] Karkaletsis, V., Spyropoulos, C.D., Benaki, E. "Customising Information Extraction Templates according to Users Interests", in Proceedings of *the Workshop "Lexically Driven Information Extraction* - LDIE'97", Frascatti, Rome, July 16, 1997, pp. 23-38.

[Karkaletsis et al, 1998a] Karkaletsis, V., Spyropoulos, C. D. and Petasis, G. "Named-entity recognition

from Greek texts: the GIE Project." *In Advances in intelligent systems: concepts, tools and applications,* ed. S. Tzafestas, Kluwer Academic Publishers, 1998.

[Karkaletsis et al, 1998b] Karkaletsis, V., Androutsopoulos, I., Paliouras, G., Spyropoulos, C.D., Catizone, R., and Stevenson, M. "Domain Modelling and Template Customisation". *ECRAN, Deliverable 3.1.1,* September 1998.

[Karkaletsis & Spyropoulos, 1998] Karkaletsis, V., and Spyropoulos, C.D., 1998. "Towards a User friendly Information Society". In Proceedings of the *Conference "ELSNET in Wonderland"*, Soestenberg, Holland, 25-27 March 1998, pp. 66-72.

[Μανουσοπούλου, 1997] Μανουσοπούλου, Α. Γ., *et al., «Εντοπισμός και σημείωση ονοματικών συνόλων σε κείμενα της Νέας Ελληνικής»,* 6ο Πανελλήνιο Συνέδριο Πληροφορικής, 1997.

[Paliouras et al 1998a] Paliouras, G., Karkaletsis, V. and Spyropoulos, C. "Automated acquisition of named-entity recognition rules", submitted to the Journal of Natural Language Engineering (under review).

[Paliouras et al 1998b] Paliouras, G., Karkaletsis, V. and Spyropoulos, C. "Machine learning for domain-adaptive word-sense disambiguation." *In Proceedings of the Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications,* 1st International Conference on Language Resources and Evaluation, Granada, Spain, May 26, 1998.

7. Information Extraction from Texts

Dr. Vangelis Karkaletsis, Dr. George Paliouras

NCSR "Demokritos",
Institute of Informatics & Telecommunications
Tel: 01-6503197,
Fax: 01-6532175,
{vangelis, paliourg}@iit.demokritos.gr

Abstract

This article presents the existing types of text-based information technology, emphasising on Information Extraction. It describes the processing stages of a an information extraction system and discusses the two major issues in the area: adaptation to new domains and multilingual information extraction. The exploitation of machine learning techniques for domain adaptation and the major approaches in multilingual information extraction are briefly discussed.

Αυτοματοποιημένη Εξαγωγή Πληροφορίας από Κείμενα

Δρ. Βαγγέλης Καρκαλέτσης, Δρ. Γιώργος Παλιούρας
Ε.Κ.Ε.Φ.Ε. "Δημόκριτος",
Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών
Τηλ: 01-6503197,
Fax: 01-6532175,
{vangelis, paliourg}@iit.demokritos.gr

Η συνδυασμένη χρήση νέων τεχνολογιών στις Τηλεπικοινωνίες (δίκτυα υψηλής ταχύτητας και μεγάλου όγκου μετάδοσης πληροφορίας) και στην Πληροφορική (φθηνά μέσα αποθήκευσης) έχουν επιτρέψει τη διάθεση στους χρήστες τεράστιου όγκου κειμένων σε ηλεκτρονική μορφή. Η εποχή της υπερπληροφόρησης με παροχή πληροφορίας σε τόσο μεγάλες ποσότητες απαιτεί την ανάπτυξη τεχνικών για την αποτελεσματική πρόσβαση των χρηστών σε κειμενικές βάσεις δεδομένων (ΚΒΔ). Τρεις διαφορετικές τεχνολογίες υπάρχουν για την πρόσβαση σε ΚΒΔ: *ανάκτηση πληροφορίας* (Information Retrieval), *φιλτράρισμα πληροφορίας* (Information Filtering) και *εξαγωγή πληροφορίας* (Information Extraction). Τα συστήματα που ανήκουν στις δύο πρώτες τεχνολογίες παρέχουν εκείνα τα κείμενα από τις ΚΒΔ που είναι σχετικά είτε με

τις λέξεις-κλειδιά στην ερώτηση του χρήστη (ανάκτηση πληροφορίας) είτε με τα ενδιαφέροντα του χρήστη διατυπωμένα συνήθως σαν κατηγορίες κειμένων (φιλτράρισμα πληροφορίας). Αντίθετα τα συστήματα εξαγωγής πληροφορίας εξάγουν συγκεκριμένα δεδομένα από κείμενα και παρέχουν στον χρήστη μόνο αυτά τα δεδομένα και όχι ολόκληρα τα κείμενα.

Κατά τη διάρκεια της τελευταίας δεκαετίας σημειώθηκε σημαντική πρόοδος στην ανάπτυξη αξιόπιστης τεχνολογίας εξαγωγής πληροφορίας, στην ΕΕ, τις ΗΠΑ και την Ιαπωνία. Αυτή η πρόοδος είναι αποτέλεσμα της αύξησης των διαθέσιμων πόρων (λεξικά αξιοποιήσιμα υπολογιστικά και συλλογές κειμένων σε ηλεκτρονική μορφή), της διάθεσης αυξημένης υπολογιστικής ισχύος σε ταχύτητα και όγκο επεξεργασίας, και της ανάπτυξης τεχνικών από την περιοχή της Γλωσσικής Τεχνολογίας που εφαρμόζονται πλέον στην πράξη. Η πρόοδος στην τεχνολογία εξαγωγής πληροφορίας αποδεικνύεται από τα αποτελέσματα σχετικών Συνεδρίων αξιολόγησης συστημάτων εξαγωγής πληροφορίας *(Message Understanding Conferences - MUCs)* που διοργανώνονται τα τελευταία χρόνια στις ΗΠΑ [DARPA, 1995; DARPA, 1998]. Η μεγάλη συμμετοχή επιχειρήσεων στα συνέδρια αυτά, καταδεικνύει την ιδιαίτερη σημασία που αποδίδουν οι επιχειρήσεις αυτές στην τεχνολογία εξαγωγής πληροφορίας. Η τεχνολογία αυτή εξετάζεται σε πραγματικές εφαρμογές, όπως οι εξαγορές επιχειρήσεων [Jacobs & Rau, 1990; Cowie et al., 1993], τα οικονομικά αποτελέσματα επιχειρήσεων [Vichot et al., 1998; Andersen et al., 1992], τα ιατρικά περιστατικά [Cavazza & Zigenbaum, 1992], καθώς επίσης και αστυνομικά περιστατικά [Schneider, 1998; Gaizauskas, 1992]. Τα συστήματα που παίρνουν μέρος στα Συνέδρια Αξιολόγησης MUC καλούνται να επεξεργαστούν κείμενα (διαφορετικά πεδία εφαρμογών σε κάθε συνέδριο), να αναγνωρίσουν τα σχετικά με το πεδίο κείμενα και να γεμίσουν *βάσεις πλαισίων (templates)* που περιέχουν πεδία με πληροφορίες για τις σημαντικές *οντότητες* και *γεγονότα*. Ενδεικτικά συστήματα εξαγωγής πληροφορίας που συμμετείχαν στα MUC-6 και MUC-7 είναι τα ακόλουθα: PROTEUS του New York University [Yangarber & Grishman, 1998], Alembic της MITRE [Day et al., 1998], LaSIE-II του University of Sheffield [Humphreys et al., 1998], LOLITA του University of Durham [Costantino et al., 1997], CRL/NMSU του New Mexico State University [Cowie, 1995]. Οι θεματικές περιοχές που έχουν εξεταστεί στα Συνέδρια Αξιολόγησης MUC είναι οι εξής: μηνύματα ναυτικού (MUCK, MUCK-II), ειδήσεις για τρομοκρατικές επιθέσεις (MUC-3, MUC-4), επιχειρηματικές ειδήσεις (joint ventures, micro-electronics products) (MUC-5), επιχειρηματικές ειδήσεις (management succession) (MUC-6), ειδήσεις για εκτοξεύσεις πυραύλων (MUC-7).

Ένα τυπικό σύστημα εξαγωγής πληροφορίας [Hobbs, 1993], είναι ουσιαστικά ένα σύστημα επεξεργασίας φυσικής γλώσσας που αποτελείται από τα παρακάτω υπο-συστήματα:

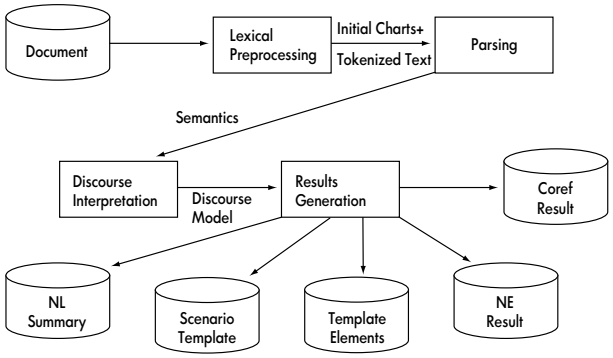
- *Λεξική Ανάλυση:* διαχωρισμός προτάσεων και λέξεων, αναγνώριση μέρους του λόγου και λήμματος κάθε λέξης.

- *Συντακτική Ανάλυση:* προσδιορισμός συντακτικής δομής φράσεων και προτάσεων, αναγνώριση ονοματικών οντοτήτων.

- *Σημασιολογική και Πραγματολογική Ανάλυση:* προσδιορισμός των προτάσεων που αναφέρονται στο γεγονός για το οποίο πρέπει να εξαχθεί πληροφορία, προσδιορισμός του ρόλου των οντοτήτων στο συγκεκριμένο γεγονός, προσδιορισμός σχέσεων αναφοράς μεταξύ των οντοτήτων. Οι εργασίες αυτές συνήθως επιτυγχάνονται με τη χρήση προτύπων (patterns).

- *Γέμισμα της βάσης πλαισίων:* η σημασιολογική αναπαράσταση κάθε γεγονότος και των εμπλεκόμενων οντοτήτων χρησιμοποιείται για να γεμίσει μία βάση πλαισίων (templates base) που περιέχουν πεδία για τα γεγονότα και τους ρόλους των σχετικών οντοτήτων.

Τα υπάρχοντα συστήματα εξαγωγής πληροφορίας δεν ακολουθούν όλα την παραπάνω διαδικασία. Ανάλογα με την εφαρμογή δίνουν λιγότερη ή περισσότερη έμφαση σε κάποια από τα παραπάνω υπο-συστήματα. Ένα χαρακτηριστικό παράδειγμα συστήματος ΕΠ που ακολουθεί όλα τα στάδια της παραπάνω διαδικασίας είναι το σύστημα LaSIE-II του University of Sheffield [Humphreys et al., 1998] (βλ. Σχήμα 1).

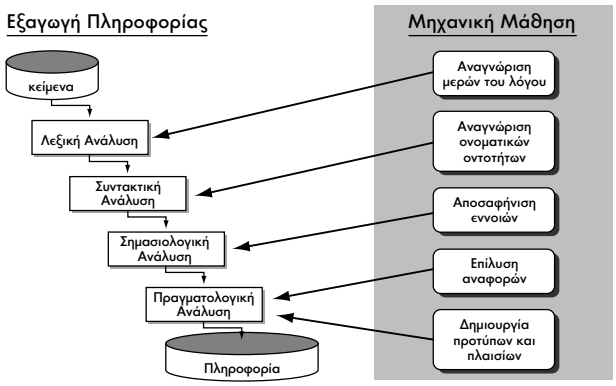


Σχήμα 1
Αρχιτεκτονική του συστήματος εξαγωγής πληροφορίας LaSIE-II

Κάθε κείμενο στο LaSIE-II υφίσταται λεξική ανάλυση (lexical preprocessing), συντακτική ανάλυση (parsing), σημασιολογική και πραγματολογική ανάλυση (discourse interpretation). Το αποτέλεσμα της παραπάνω επεξεργασίας χρησιμοποιείται για την παραγωγή αποτελεσμάτων για 5 διαφορετικές εργασίες, σύμφωνα με τα συνέδρια αξιολόγησης MUC: Αναγνώριση Ονοματικών Οντοτήτων (Named Entity Recognition), Προσδιορισμός Αναφορών Οντοτήτων (Coreference Identification), Εξαγωγή Πληροφορίας για τις Οντότητες (Template Elements Filling), Εξαγωγή Πληροφορίας για τα Γεγονότα στα οποία εμπλέκονται οι Οντότητες (Scenario Template Elements Filling), Παραγωγή περιλήψεων (NL Summary).

Τα υπάρχοντα συστήματα εξαγωγής πληροφορίας εξάγουν προκαθορισμένους τύπους πληροφορίας από κείμενα μιας θεματικής περιοχής, γραμμένα σε μια συγκεκριμένη γλώσσα. Για να μπορέσει η τεχνολογία εξαγωγής πληροφορίας να χρησιμοποιηθεί στην πράξη σε βιομηχανικές και εμπορικές εφαρμογές, τα συστήματα εξαγωγής πληροφορίας πρέπει να είναι σε θέση να προσαρμόζονται εύκολα σε νέες θεματικές περιοχές και στα ενδιαφέροντα νέων χρηστών, καθώς επίσης και σε νέες γλώσσες.

Ο περιορισμός του προβλήματος της εξαγωγής πληροφορίας σε ένα συγκεκριμένο θεματικό πεδίο κάνει το πρόβλημα πρακτικά επιλύσιμο, αλλά ταυτόχρονα δημιουργεί την ανάγκη για συνεχή προσαρμογή του συστήματος σε νέες θεματικές περιοχές. Η προσαρμογή αυτή είναι ιδιαίτερα δαπανηρή, εξαιτίας της πληθώρας γλωσσικών πόρων που χρησιμοποιούνται για την εξαγωγή πληροφορίας από κείμενα. Χαρακτηριστικό δείγμα της σημασίας αυτού του προβλήματος είναι τα χρονικά περιθώρια των συνεδρίων αξιολόγησης MUC, τα οποία στενεύουν συνεχώς, ωθώντας τους συμμετέχοντες στην εξεύρεση μεθόδων, οι οποίες να επιταχύνουν την προσαρμογή των συστημάτων τους. Η προφανής λύση στο πρόβλημα είναι η αυτοματοποίηση της προσαρμογής των διαφόρων τμημάτων του συστήματος εξαγωγής πληροφορίας, στο βαθμό που αυτή είναι δυνατή. Η αυτοματοποίηση αυτή επιτυγχάνεται συνήθως με τη χρήση μεθόδων μηχανικής μάθησης, οι οποίες χρησιμοποιούν κείμενα εκπαίδευσης (training texts) για να προσαρμόσουν τα μεταβλητά τμήματα του συστήματος στη θεματική περιοχή [Cardie, 1997]. Στο Σχήμα 2 απεικονίζονται τα τμήματα ενός συστήματος εξαγωγής πληροφορίας στα οποία μπορούν να χρησιμοποιηθούν μέθοδοι μηχανικής μάθησης.



Σχήμα 2
Χρήση μεθόδων μηχανικής μάθησης στην εξαγωγή πληροφορίας από κείμενα

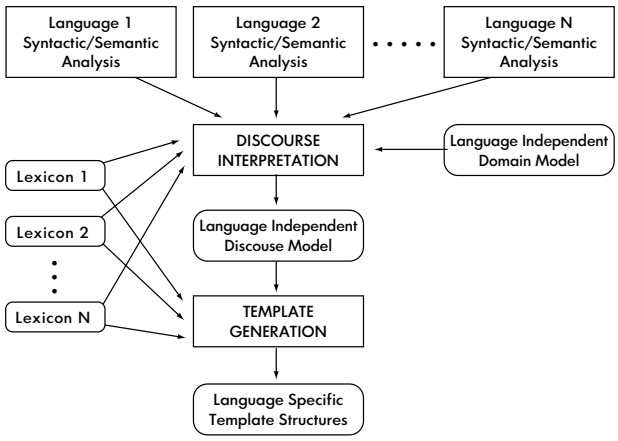
Όσον αφορά το δεύτερο κύριο ζήτημα στην τεχνολογίας εξαγωγής πληροφορίας, δηλ. την προσαρμογή σε περισσότερες από μία γλώσσες, αυτή μπορεί να πραγματοποιηθεί με δύο κύριους τρόπους:

- Σύστημα ΕΠ, που εκτελεί μονογλωσσική ΕΠ, αλλά για περισσότερες από μία γλώσσες, υπάρχει δηλ. διαφορετική έκδοση του συστήματος για κάθε γλώσσα. Σε κάθε τέτοια έκδοση η αρχική (source) και η τελική (extraction) γλώσσα είναι ίδιες.

- Σύστημα ΕΠ που εκτελεί διαγλωσσική (cross-lingual) ΕΠ Στην περίπτωση αυτή η αρχική και η τελική γλώσσα είναι διαφορετικές.

Ένα παράδειγμα συστήματος πολυγλωσσικής ΕΠ

από κείμενα μιας θεματικής περιοχής παρουσιάζεται στο Σχήμα 3.



Σχήμα 3
Αρχιτεκτονική του συστήματος πολυγλωσσικής εξαγωγής πληροφορίας ML-LaSIE

Πρόκειται για το σύστημα ML-LaSIE του Πανεπιστημίου του Sheffield. Κάθε κείμενο αναλύεται λεξικά, συντακτικά και σημασιολογικά από τον αντίστοιχο αναλυτή. Κατά τη διάρκεια της πραγματολογικής ανάλυσης η σημασιολογική αναπαράσταση του κειμένου απεικονίζεται σε ένα γλωσσικά ανεξάρτητο μοντέλο της θεματικής περιοχής. Ανάλογα με τη γλώσσα στην οποία επιθυμούμε να παρουσιαστεί η πληροφορία, χρησιμοποιείται το αντίστοιχο λεξικό για την "μετάφραση" από το γλωσσικά ανεξάρτητο μοντέλο στη βάση πλαισίων και στη συγκεκριμένη γλώσσα.

Συνοψίζοντας, θα πρέπει να σημειώσουμε την ιδιαίτερη σημασία της τεχνολογίας εξαγωγής πληροφορίας στη σημερινή εποχή της υπερπληροφόρησης και την ανάγκη αντιμετώπισης προβλημάτων προσαρμογής της τεχνολογίας αυτής σε νέες θεματικές περιοχές και γλώσσες με στόχο την αξιοποίησή της στην πράξη σε βιομηχανικές και εμπορικές εφαρμογές. Αρκετά από τα προβλήματα προσαρμογής της τεχνολογίας μπορούν να αντιμετωπιστούν με τεχνικές μηχανικής μάθησης.

Βιβλιογραφία

[Andersen et al., 1992] Andersen P.M., Hayes P.J., Huettner A.K., Nirenburg I.B., Schmandt L.M.and Weinstein S.P.. Automatic extraction of facts from press releases to generate news stories. In Proceedings of the Third Conference on Applied Natural Language Processing, pages 170-177. ACL, 1992.

[Cardie, C., 1997] Cardie, C. Empirical Methods in Information Extraction. AI Magazine, 18:4, 65-79, 1997.

[Cavazza & Zweigenbaum, 1992] Cavazza, M., and Zweigenbaum, P. *Extracting implicit information from free text technical reports.* Information Processing and Management, 28(5), 1992.

[Costantino et al. 1997] Costantino, M., Morgan, R.G., and Collingham R.J. Financial Information Extraction Using Pre-Defined and User-Definable Templates in the LOLITA System. In *CIT – Journal of Computing and Information Technology*, 1997.

[Cowie 1995] Cowie, J. "Desription of the CRL/NMSU System Used for MUC-6". In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, San Francisco, Calif.: Morgan Kaufmann.

[Cowie et al., 1993] Cowie J., Wakao T., Jin W., Pustejovsky J. and Waterman S.. The diderot information extraction system. In Proceedings of the First Conference of the Pacific Association for Computational Linguistics (PACLING 93), Vancouver, Canada, 1993.

[DARPA, 1995] Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6),* Morgan Kaufmann, 1995.

[DARPA, 1998] Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, (forthcoming).

[Day et al., 1998] Day, D., Robinson, P., Vilain, M., and Yeh, A. Description of the ALEMBIC system as used for MUC-7. In *[DARPA, 1998]*.

[Gaizauskas et al., 1992] Gaizauskas R., Evans R., Cahill L.J., Richardson J., and Walker J.. Poetic: A system for gathering and disseminating traffic information. In S.G.Ritchie and G.T.Hendrickson,

editors, Conference Preprints of the International Conference on Artificial Intelligence Applications in Transportation Engineering, pages 79-98, San Buenaventura, California, 1992.

[Hobbs, 1993] Hobbs, J.R. The Generic Information Extraction System. In *Proceedings of the 5th Message Understanding Conference (MUC-5)*. Morgan-Kaufman, 1993, 87-91.

[Humphreys et al., 1998] Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. 1998. Description of the LaSIE-II system as used for MUC-7. In [DARPA, 1998].

[Jacobs & Rau, 1990] Jacobs P.S. and Rau L.F.. Scisor: Extracting information from on-line news. Communications of the ACM, 33(11):88-97, 1990.

[Schneider, 1998] Schneider, T. Multilingual Information Processing: the AVENTINUS Project. In *Proceedings of the First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998, pp. 775-782.

[Vichot et al., 1998] Vichot, F., Wolinski, F., Ferri, H.C., Urbani, D. "Using Information Extraction for Knowledge Entering". In *Proceedings of the Adaptive and Multilingual Information Extraction Workshop (AMIE'98)*, Athens, June 1998.

[Yangarber & Grishman, 1998] Yangarber, R., and Grishman, R. 1998. NYU: Description of the Proteus/PET System as Used for MUC-7 ST. In [DARPA, 1998].

8. Προς μια σύγχρονη πλατφόρμα εξαγωγής πληροφορίας

Στέλιος Πιπερίδης, Σωτήρης Μπούτσης
Ινστιτούτο Επεξεργασίας του Λόγου
Αρτέμιδος 6 και Επιδαύρου,
Παράδεισος Αμαρουσίου, 151 25 Αθήνα
e-mail: spip, sboutsis@ilsp.gr

Abstract

In this paper we describe the plan and potential architecture for the development of an information extraction (IE) platform for Greek. The differentiation between retrieval and extraction of information from text is first discussed and the resulting necessary features of an IE platform are then presented. An overview of the text processing chain for monolingual IE applications is given next. Concluding, we discuss the issues pertaining to information extraction in a multilingual context.

Περίληψη

Στην ανακοίνωση αυτή περιγράφεται η σχεδίαση και η πιθανή αρχιτεκτονική μιας πλατφόρμας εξαγωγής πληροφορίας (ΕΠ) από τα ελληνικά κείμενα. Συζητούνται τα σημεία διαφοροποίησης μεταξύ ανάκτησης και εξαγωγής πληροφορίας και παρουσιάζονται τα ποιοτικά χαρακτηριστικά μιας πλατφόρμας ΕΠ. Στη συνέχεια, δίδεται η συνολική εικόνα των σταδίων επεξεργασίας με σκοπό την εξαγωγή πληροφορίας από μονόγλωσσα ελληνικά κείμενα, ενώ, τέλος, συζητούνται τα ζητήματα που προκύπτουν σε εφαρμογές εξαγωγής πληροφορίας σε πολύγλωσσο περιβάλλον.

Towards an Information Extraction Platform for Greek

Objective

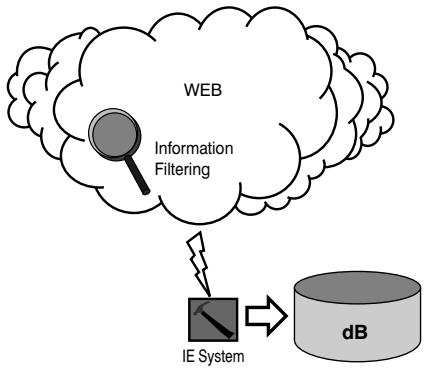
Develop an information extraction platform out of *existing, customised* or *application specific* natural language processing modules

Target Applications

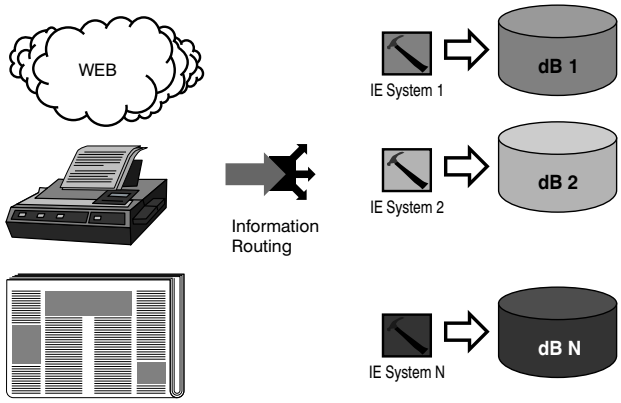
Automatic processing of very large texts in order to :

- filter the vast quantities of incoming information
- on acceptable terms wrt cost/time investments
- with satisfactory precision and recall

Retrieval vs. Extraction



Retrieval vs. Extraction



Implications on System Design & Performance

Robust

Wrt free-text phenomena

Efficient

To deal with the large quantities of incoming text

Easily adaptable

So that the IE system can be extended to cover new domains without prohibitive investments

The Engineering Approach to IE

Finite State Techniques

Pattern rules can be translated to DFSA's by standard techniques. DFSA's allow processing in linear time = O(length of the text)

Partial vs. Complete Analysis

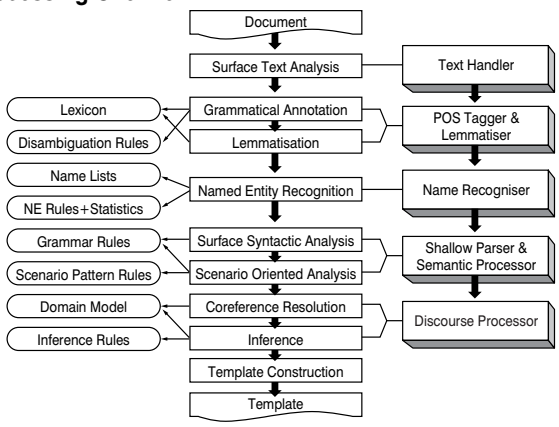
E.g. full parsing can be very expensive and may not apply. Partial parsing can be a good choice, aiming to recover partial syntactic information efficiently from unrestricted text, by sacrificing completeness and depth of analysis

Porting to new domains

Several approaches e.g.

- **AutoSlog: training on a large annotated corpus,**
- **HASTEN: training from extracted and extensively analysed examples,**
- **NYU: training through user supervised generalisation**

Processing Overview



Handler

Word Boundaries identification

- dates
- numbers
- enumeration lists
- abbreviations
- acronyms
- punctuation
- tokens

Sentence Boundaries identification Tools

- Regular Expression Grammars
- Abbreviation Lists
- A set of filters chained together form the entire segmentation tool:
  split text → isolate punctuation → identify abbreviations, dates, numbers and enumerations → identify sentences

PoS Tagger and Lemmatiser (1)

Morphological analyzer

- lemma
- possible Parts of Speech (ambiguity class lexicon)

E.g.
διατάξεις διάταξη NoCmFePlNm/NoCmFePlAc/NoCmFePlVo
διατάξεις διατάζω VbMnIdXx02SgXxPeAvXx

Training

- use a hand-tagged training corpus if available
- if no such training corpus is available, use the Forward-Backward algorithm
- to estimate the HMM parameters

PoS Tagger and Lemmatiser (2)

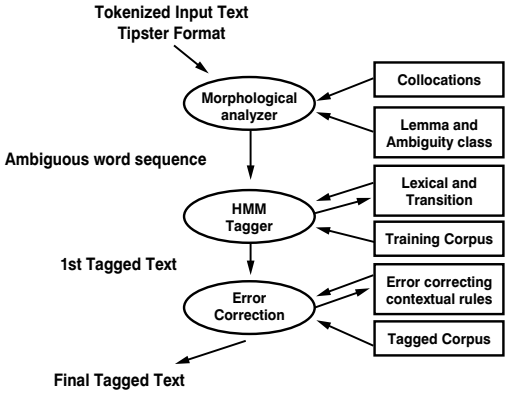
Stochastic Tagging

- probability that a word occurs with a particular tag
- probability of a given sequence of tags
- both tag sequence probabilities and word frequency measurements (HMM)

Rule Based Tagging

- use contextual information to assign tags to unknown or ambiguous words (Brill)
- automatic induction of rules from a manually annotated corpus

PoS Tagger and Lemmatiser (3)



Surface Syntactic Analyser (1)

Standard full-parsers:

- Make the closed world assumption
- Evaluate global parses, not partial parses
- Do all paths search

Full Parsing poses difficult problems because of:

- Incompleteness of grammar
- Long sentences
- "Ungrammatical", noisy input

For a number of LE applications full parsing is very expensive, does not apply.

Partial parsing is a likely solution

aiming to recover partial syntactic information efficiently from unrestricted text, by sacrificing completeness and depth of analysis

Surface Syntactic Analyser (2)

Key Elements of Partial Parsing with Finite State Technology

Grammatical phenomena are described with sets of regular expressions

Grammar rules (regular expressions) are numbered and organised in a cascade e.g.

1. NP -> D? A* N+ ¦ Pron
2. VP -> Md Vb ¦ ...
3. PP -> P NP
4. SV -> NP VP
5. S -> (Adv ¦ PP)? SV NP? (Adv ¦ PP)*

The constituents matched by one rule take part in next ones

Regular expressions compile into a pipeline / cascade of FSA's in a straightforward way

Surface Syntactic Analyser (3)

Serialisation of finite state automata

A4	S				
A4	vg	np	pp		
A4	Επιβάλλεται	np	pp		
A4	Επιβάλλεται	np	με	np	"np"

Vb	n	n_g	n_g	p	art	n	n	n_g	n_g
Επιβάλλεται	φόρος	κύκλου	εργασιών	με	την	ονομασία	"φόρος	προστιθέμενης	αξίας"

Surface Syntactic Analyser (4)

Grammar

17 levels of rules for:

adjectival phrase, noun phrase, verb phrase, prepositional phrase, adverbial phrase, relative clauses and *other subordinate clauses*

following EAGLES (Leech et al 1996).

Performance

Precision and Recall values > 85%

Speed of analysis: ~2000 words / sec

The Multilinguality Aspect in IE

Two possibilities

A system that performs monolingual IE in multiple languages

Monolingual IE: where source and extraction languages are the same

Extraction Language: language of the template fills and/or of summaries generated

A system that does Cross-lingual IE

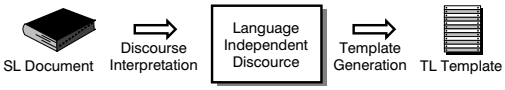
Cross-lingual IE: IE where source language and extraction language differ

Approaches to Cross-lingual IE

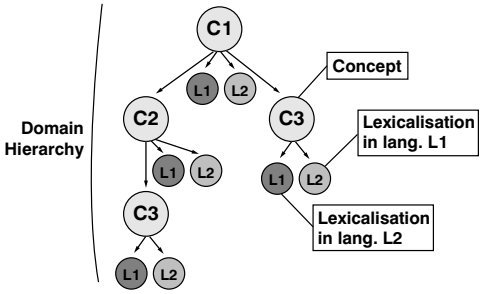
MT system translates text and monolingual IE performs extraction

Monolingual IE performes extraction and MT system translates template fills

Language-dependent front ends map text to language-indep. discourse model



Multilingual Domain Hierarchy



IV. Γλωσσάριο Όρων Γλωσσικής Τεχνολογίας και Πληροφορικής /Language Technology and Informatics Forum

Λάβαμε την ακόλουθη επιστολή εκ μέρους του κ. Νάσου στην οποία παρατίθενται προβληματισμοί σχετικά με το θέμα της απόδοσης ξένων όρων στην Ελληνική.

Μερικές σκέψεις για την απόδοση όρων στα Ελληνικά

Nikolaos Nassos
JMO CC-128/A
European Commission
Rue Wehrer
L-2920 Luxembourg
tel. +352-4301-35153
fax. +352-4301-33066, +352-4301-33829
e-mail: Nassos.NIKOLAOS@DI.cec.be

Εισαγωγή

Στο παρόν κείμενο παραθέτω μερικούς προβληματισμούς σχετικά με το θέμα της απόδοσης ξένων όρων στα Ελληνικά. Αφορμή για το κείμενο έδωσε ένα τεχνικό άρθρο που υποβλήθηκε στο συνέδριο NIT 98. Το άρθρο ήταν γραμμένο στα Ελληνικά αλλά περιείχε πολλούς όρους στην Αγγλική. Οι περισσότεροι ήταν μεταφρασμένοι, κάποιοι άλλοι είχαν μεν εξηγηθεί αλλά στη συνέχεια χρησιμοποιούνταν αμετάφραστοι και τέλος μερικοί δεν αποδίδονταν καθόλου στα Ελληνικά. Σε ορισμένα μάλιστα σημεία το άρθρο έδινε την εντύπωση ότι αποτελούσε γρήγορη μεταφορά ενός αγγλικού πρωτοτύπου ή τουλάχιστον φανέρωνε ότι οι συγγραφείς δυσκολεύονταν πολύ να μιλήσουν για το θέμα τους στα Ελληνικά. Ακόμα και σε σημεία όπου γινόταν χρήση ελληνικών όρων οι συγγραφείς παρέθεταν σε παρένθεση και τον πρωτότυπο. Προσωπικά είχα την εντύπωση ότι μόνον έτσι οι συγγραφείς ήταν σίγουροι πως θα τους καταλάβει ο αναγνώστης του άρθρου.

Στη συνέχεια του κειμένου μου θα δώσω παραδείγματα μερικών πολύ συνηθισμένων όρων, οι περισσότεροι από τους οποίους βρίσκονταν και στο παραπάνω άρθρο, καθώς και αποδόσεις που χρησιμοποιούνται συχνά ή/και προτείνονται από συγκεκριμένους φορείς. Φυσικά δεν έχω εντοπίσει όλους τους φορείς που ασχολούνται με την μετάφραση όρων στην Ελλάδα και στο εξωτερικό και έτσι το κείμενό μου δεν διεκδικεί δάφνες ολοκληρωμένης εργασίας. Θέλω απλώς να δώσω υλικό για προβληματισμό στους συναδέλφους (μηχανικούς και γλωσσολόγους) αλλά και εκτός αυτού, σε όσους ενδιαφέρονται για το θέμα. Θα θεωρήσω ότι το κείμενό μου είναι επιτυχημένο αν ξεκινήσει κάποια συζήτηση με αντικείμενο την απόδοση όρων στα Ελληνικά.

Όροι, απόδοση και σχολιασμός

Διευκρινίζω ότι οι αποδόσεις που αναφέρονται στη συνέχεια δεν είναι δικές μου επινοήσεις ως επί το πλείστον. Πρόκειται για αποδόσεις που έχω συναντήσει σε περιοδικά, βιβλία, λεξικά και αλλού ή που έχω ακούσει από διάφορες πηγές. Στις περιπτώσεις που θυμάμαι την πηγή την παραθέτω.

1. World Wide Web (WWW)

Έχω δει την απόδοση Παγκόσμιο Πλέγμα Πληροφοριών (ΠΠΠ) στο βιβλίο "Το εγχειρίδιο του καλού μπλοφαδόρου για το Internet" των εκδόσεων Δίαυλος αν δεν με απατά η μνήμη μου. Θεωρώ την απόδοση αυτή εξαιρετικά επιτυχημένη. Αποδίδει σωστά το περιεχόμενο του πρωτοτύπου και παρουσιάζει ομοιότητα και στη μορφή, είναι τρεις λέξεις που αρχίζουν μάλιστα με το ίδιο γράμμα. Έτσι, όπως διεθνώς γίνεται λόγος για WWW μπορούμε εμείς να μιλάμε για ΠΠΠ. Επίσης, όπως το World Wide Web για συντομία λέγεται απλά Web, έτσι και το Παγκόσμιο Πλέγμα Πληροφοριών θα μπορεί να λέγεται απλά Πλέγμα.

2. User interface

Πρόκειται για έναν ιδιαίτερα προβληματικό όρο. Από διάφορους (συγγραφείς, περιοδικά) προτείνονται οι εξής αποδόσεις:

διασύνδεση χρήστη
διεπαφή χρήστη

σύζευξη χρήστη
περιβάλλον (εργασίας) χρήστη

Όσον αφορά τις πρώτες αποδόσεις, ίσως θα ήταν καλύτερο να λέμε "διασύνδεση / διεπαφή / σύζευξη με το χρήστη", είναι πάντως σίγουρο ότι η πλειοψηφία των ομιλητών μάλλον δεν είναι ικανοποιημένη με τις αποδόσεις αυτές. Έτσι όλοι ή σχεδόν όλοι εξακολουθούμε να χρησιμοποιούμε το user interface για να συνεννοούμαστε.

Η απόδοση “περιβάλλον (εργασίας) χρήστη” χρησιμοποιήθηκε από τη Microsoft για τις ελληνικές εκδόσεις των προϊόντων της. Κατά τη γνώμη μου είναι η καλύτερη απόδοση.

3. Browser

Πρόκειται για το πρόγραμμα που χρησιμοποιούμε για να δούμε σελίδες του “Πλέγματος” (για να είμαι πιστός στην απόδοση που λίγο πιο πάνω υποστήριξα), όπως είναι το Netscape Navigator ή το Internet Explorer. Επειδή λοιπόν με αυτό το πρόγραμμα βλέπουμε σελίδες και επειδή τις περισσότερες φορές αυτό γίνεται βιαστικά, πρόκειται δηλαδή για ξεφύλλισμα, θα μπορούσαμε να το ονομάσουμε “ξεφυλλιστή” ή, πιο επίσημα, “διαφυλλιστή”.

Τι θα κάνουμε τώρα με το “browsing” την ενέργεια δηλαδή που εκτελούμε με το “διαφυλλιστή”; Προσωπικά θα χρησιμοποιούσα το “ξεφύλλισμα”.

4. Module

Αποδόσεις που έχω συναντήσει:
τμήμα
στοιχείο
ενότητα
μονάδα
δομοστοιχείο

Νομίζω ότι η τελευταία απόδοση είναι η καλύτερη. Προτείνεται από την ΕΛΕΤΟ στην ελληνική έκδοση του λεξικού τηλεπικοινωνιών του J.P Rehahn από τις εκδόσεις Γλώσσημα. Βλέποντας (ή ακούγοντας) κανείς το “δομοστοιχείο” καταλαβαίνει αμέσως ότι πρόκειται για κάποιον τεχνικό όρο με ειδική σημασία. Κατά τη γνώμη μου αυτό είναι απαραίτητο στα τεχνικά κείμενα.

5. On-line (επίθετο)

Άλλος ένας προβληματικός όρος με την έννοια ότι έχει μια ιδιαίτερη σημασία και δεν υπάρχει μια εξίσου ιδιαίτερη απόδοση γενικής αποδοχής.

Η ΕΛΕΤΟ στο προαναφερθέν λεξικό προτείνει “επιγραμμικός”, εμένα προσωπικά η λέξη μου αρέσει. Ίσως γιατί δεν έχει (για μένα) κάποιο άλλο περιεχόμενο και έτσι εύκολα μπορεί να πάρει ένα καινούριο. Τι γίνεται όμως όταν κάποιος δουλεύει “on-line”; Εγώ θα έλεγα ότι δουλεύει “με απευθείας σύνδεση”, μάλλον όμως αυτή η έκφραση δεν καλύτει όλο το περιεχόμενο του “on-line”. Θα μπορούσαμε να χρησιμοποιήσουμε το “επί γραμμής”; Δεν σημαίνει τίποτα, άρα εύκολα μπορεί να αποκτήσει νέο περιεχόμενο. Πιστεύω όμως ότι θα ξενίσει πολλούς χρήστες και δύσκολα θα γίνει αποδεκτό.

Στα προϊόντα της Microsoft γίνεται πάντως λόγος για “άμεση βοήθεια” (on-line help).

6. Script

Πιθανές αποδόσεις

δέσμη ενεργειών (Microsoft)
αρχείο εντολών
σενάριο

Το script είναι όντως ένα αρχείο με εντολές που συνιστούν μια δέσμη ενεργειών! Πάντως σε τεχνικές συζητήσεις όλοι, για να συνεννοούνται, μιλάνε για script ή χαϊδευτικά για “σκριπτάκια”.

7. Back-up

Ως ρήμα: παίρνω αντίγραφα ασφαλείας
Ως ουσιαστικό: αντίγραφο ασφαλείας

Κανονικά θα έπρεπε οι ελληνικοί όροι να έχουν καθιερωθεί αφού είναι συνήθεις και κατανοητοί, οι περισσότεροι χρήστες όμως προτιμούν τους αγγλικούς. Η

ΕΛΕΤΟ προτείνει “εφεδρικοποιώ”.

Τελικές παρατηρήσεις

Κλείνω αυτό το σύντομο κείμενο με κάποιες γενικές παρατηρήσεις. Κατ’ αρχήν πιστεύω ότι στο θέμα της απόδοσης των ξένων όρων στη γλώσσα μας δεν έχουμε δώσει την απαραίτητη σημασία. Μπορεί να έχει γίνει πολλή δουλειά από κάποιους φορείς, όμως οι προτεινόμενοι ελληνικοί όροι δεν έχουν καθιερωθεί, δεν είναι καν γνωστή η ύπαρξή τους πολλές φορές. Στην πράξη λοιπόν, οι περισσότεροι χρησιμοποιούν τους ξένους όρους όχι μόνο γιατί αυτούς συναντάνε συνεχώς στη βιβλιογραφία αλλά, κυρίως, για να διασφαλιστεί η συνεννόηση.

Σε μεγάλο βαθμό η καθιέρωση είναι και θέμα ’βούλησης’ των χρηστών. Κατά τη γνώμη μου, πολλοί χρήστες αδιαφορούν για τις ελληνικές αποδόσεις, είτε επειδή δεν τους αρέσουν ή επειδή δεν τους είναι απαραίτητες. Τους αρκεί δηλαδή που όλοι καταλαβαίνουν τους διεθνείς όρους και η συνεννόηση επιτυγχάνεται. Μετά όμως όλοι έχουμε πρόβλημα όταν πρέπει να κάνουμε μια δημοσίευση ή μια ομιλία με επίσημη γλώσσα τα Ελληνικά. Νομίζω ότι χρειάζεται μεγαλύτερη ευαισθητοποίηση από όλους στον τομέα αυτόν, μεγαλύτερη ενημέρωση από τους φορείς που προτείνουν αποδόσεις καθώς και συνεχής επικοινωνία μεταξύ των ενδιαφερομένων, δηλαδή όλων μας.

Πηγές

Παραθέτω τρεις διευθύνσεις που φιλοξενούν λεξικά ορολογίας.

http://wwli.com/translation/netglos/glossary/glossary.html
NETGLOS, λεξικό όρων Internet.

http://www.di.uoa.gr/~infodict
Πανεπιστήμιο Αθηνών, λεξικό όρων Πληροφορικής.

http://143.233.175.2/infolex
ΤΕΙ Πειραιώς,ηλεκτρονικό λεξικό Πληροφορικής

Στην ανωτέρω επιστολή απάντησε ο Καθηγητής κ. Γ. Καραγιάννης ως εξής:

Αγαπητέ κ. Νάσσο,

Διάβασα με προσοχή το κείμενό σας και επειδή συνήθως κάνω τον αντίλογο στα θέματα της ορολογίας όταν έχω κάποια επιχειρήματα ήθελα να σημειώσετε τα εξής:

1. Στο θέμα του "Web", έχει επικρατήσει η λέξη "Ιστός" και είναι δύσκολο να την αντικαταστήσουμε τώρα πια με τη λέξη "πλέγμα". Άλλωστε η λέξη "ιστός" είναι κοντά και στην αγγλική λέξη.

2. Για την απόδοση του "user interface" αρχίζει να επικρατεί το "διεπαφή χρήστη". Η λέξη "διεπαφή" για το "interface" είναι η περισσότερο δόκιμη και ταιριάζει σε πολλές χρήσεις, τόσο στην πληροφορική όσο και σε άλλες επιστήμες. Αποδίδει ακριβώς το νόημα είτε σε συνεργασία με άλλες λέξεις είτε μόνη της και έχει ιδιαίτερα καλή αισθητική. Μήπως αξίζει τον κόπο να ξεφύγουμε από τις πολλές ποικιλίες στην απόδοση του όρου "interface" στα ελληνικά;

3. Για το "browser" καμμιά από τις δύο προτάσεις δεν ταιριάζει στα ελληνικά. Συζήτησα και τις δύο με γνωστούς γλωσσολόγους και λεξικογράφους και διαφωνούν κάθετα. Ελπίζω στο επόμενο τεύχος μας να έχουμε κάποια νέα πρόταση.

4. Σχετικά με τη λέξη "module" που είναι και η περισσότερο δύσκολη πρέπει να λάβει κανείς υπόψη του και τις λέξεις "modular" και "modularity" που πρέπει επίσης να αποδοθούν. Όλη η προσπάθεια απόδοσης θα ήταν καλό να είναι συνολική και για τους τρεις όρους. Δεν φαίνεται κάποια από τις πέντε αποδόσεις που έχετε συναντήσει να μπορεί να ανταποκριθεί σε αυτήν την ανάγκη. Πρόσφατα από τους ερευνητές του Ιστορικού Λεξικού της Ακαδημίας Αθηνών (κα Μπασέα) προτάθηκε η εξής αντιστοιχία:

   module: συναρμολόγημα
   modular: συναρμολογικός
   modularity: συναρμολογικότητα

Ο όρος αυτός είναι ενδιαφέρων και ίσως καλύπτει τους όρους "σπονδυλωτός" και "αρθρωτός" που κατά καιρούς έχουν προταθεί για να αποδόσουν το "modular".

5. Για το "on-line" χρησιμοποιείται εδώ και πολλά χρόνια στο ΕΜΠ ο όρος "εντός-γραμμής" που είναι περισσότερο κυριολεκτικός από τον όρο "επί-γραμμής". Ταιριάζει επίσης με την αργκώ που θέλει όταν κανείς δεν είναι πλέον εντός γραμμής, να λέει "με πέταξε έξω".

6. Συμφωνώ μαζί σας ότι η καλύτερη απόδοση για το "script" είναι "σκριπτάκι". Φαίνεται να εντάσσεται σχετικά καλά στη γλώσσα μας. Οι άλλοι όροι που προτείνονται είναι εντελώς αδόκιμοι.

7. Σχετικά με το "back-up" νομίζω ότι η ΕΛΕΤΟ έχει δίκιο. Μάλιστα αν το συνδέσουμε με το "copy" μπορούμε να χρησιμοποιήσουμε το δίλεκτο "αντίγραφο εφεδρείας" που μπορεί σιγά-σιγά να γίνει σύμπλοκο.

Με εκτίμηση
Γ. Καραγιάννης

V. Ειδήσεις για τη Γλωσσική Τεχνολογία (news related to Language Technology and Informatics issues)

Συνέδρια / Conferences

3rd European Robotics, Intelligent Systems & Control Conference (EURISCON'98)
Chairman of EURISCON'98,
Prof. Spyros G. Tzafestas.
Athens, Greece, June 22-25, 1998.

Adaptive and Multilingual Information Extraction Systems *(Invited Session)*

Adaptive NLP-driven systems:acquisition of linguistic information for Informaton Extraction purposes
R.Basili, M.T.Pazienza (Italy)

Computer-aided analysis of multilingual patent texts.
I. Blank (Germany)

Parallel IE Systems for Multilingual Information Gathering
L.Dini (Italy)

Information Extraction Techniques for Multilevel Sentence Matching
V.Di Tomaso, G.D'Angelo (Italy)

Named Entity Recognition from Greek Texts: The GIE Project
E.Karkaletsis, C.D.Spyropoulos, G.Petasis (Greece)

Question Answering and Information Extraction from Texts
J.Kontos, I.Malagardi (Greece)

Using Functional Style Features to enhance Information Extraction from Greek Texts
S. E. Michos, N. Fakotakis, and G. Kokkinakis (Greece)

Eliciting Terminological Knowledge for Information

Extraction Applications
S.Piperidis, B.Georgantopoulos (Greece)

Using Information Extraction for Knowledge Entering
F.Vichot, F.Wolinski, H.C.Ferri, D.Urbani (France)

2ο Συνέδριο "Ελληνική Γλώσσα και Ορολογία"

Η Ελληνική Εταιρεία Ορολογίας (ΕΛΕΤΟ), σε συνεργασία με το Πανεπιστήμιο Αθηνών (ΠΑ), το Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ), το Ιόνιο Πανεπιστήμιο (ΙΠ), το Ινστιτούτο Επεξεργασίας Λόγου (ΙΕΛ) και τον Πανελλήνιο Σύλλογο Επαγγελματιών Μεταφραστών (ΠΣΕΜ), διοργανώνουν το δεύτερο συνέδριο για την "Ελληνική Γλώσσα και Ορολογία".

Σκοπός του Συνεδρίου είναι η παρουσίαση της κατάστασης της Ελληνικής Γλώσσας και Ορολογίας στις σημερινές συνθήκες, όπως διαμορφώνεται στο πολύγλωσσο περιβάλλον της Ευρωπαϊκής Ενωσης.

Η θεματολογία του Συνεδρίου είναι:
- Ιστορική θεώρηση
- Γλωσσολογικές αρχές της Ορολογίας και ιδίως της Ελληνικής Ορολογίας (απόδοση ορολογίας, επικοινωνία)
- Ορογραφία - Τεκμηρίωση ορογραφίας
- Λεξικογραφία (ερμηνευτικά, πολύγλωσσα, ειδικά λεξικά)
- Τυποποίηση ορολογίας
- Νέες τεχνολογίες και Ορολογία (τράπεζες και βάσεις ορολογίας, οπτικοί δίσκοι, τεχνικές λειτουργίες λεξικών, μηχανική μετάφραση)
- Διερμηνεία και μετάφραση
- Οργανα Ορολογίας και γλώσσας (συνεργασία, συντονισμός, διαχείριση)
- Διευρωπαϊκό πολύγλωσσο περιβάλλον.

Το Συνέδριο είναι ανοικτό σε όποιον ενδιαφέρεται για την Ελληνική Γλώσσα και Ορολογία, τόσο στην Ελλάδα όσο και στο εξωτερικό. Επίσημες γλώσσες του Συνεδρίου είναι η ελληνική, η αγγλική και η γαλλική γλώσσα.

Το Συνέδριο θα διεξαχθεί στην **Αθήνα** στις **21, 22 και 23 Οκτωβρίου 1999.**

Καλούνται οι εισηγητές που επιθυμούν να παρουσιάσουν εργασία για ανακοίνωση στο συνέδριο, να υποβάλουν περίληψη. Η περίληψη θα πρέπει να είναι σε μια από τις επίσημες γλώσσες του συνεδρίου και δεν θα πρέπει να υπερβαίνει τις 150 λέξεις.

Το φύλλο της περίληψης θα πρέπει να αναφέρει
α) τον τίτλο του Συνεδρίου,
β το όνομα, διεύθυνση και τηλέφωνο/τηλεομοιότυπο του συγγραφέα ή των συγγραφέων.

Τέσσερα αντίγραφα της περίληψης πρέπει να κατατεθούν πριν από τις 26 Μαρτίου 1999 στη Γραμματεία του Συνεδρίου ή να σταλούν ταχυδρομικά στη διεύθυνση:

Ελληνική Εταιρεία Ορολογίας
Σ. Τσάκωνα 5
152 36 ΠΕΝΤΕΛΗ.

Οι συγγραφείς θα ενημερωθούν για την αποδοχή της προτεινόμενης ανακοίνωσης. Θα δοθούν πρόσθετες πληροφορίες για την επεξεργασία των πλήρων εισηγήσεων, οι οποίες θα πρέπει να παραδοθούν πριν από τις **31 Ιουλίου 1999.**

Για πληροφορίες:
Γραμματεία του Συνεδρίου:
κος Ι. Σαριδάκης, τηλ. 212 0113,
κα Α. Παπαναστασίου, τηλ. 202 2466
κος Τ. Ορφανός: τηλ. 611 1020.
Ηλ. Ταχυδρ.: kv121999@dm.ote.gr

Συνέδριο για τη Συνεργασία στον Χώρο της Ορολογίας στην Ευρώπη (Conference on Co-operation in the Field of Terminology in Europe)

17, 18 και 19 Μαΐου 1999
Παρίσι, Γαλλία

Ο κλάδος της ορολογίας άπτεται όλων των πλευρών της επικοινωνιακής διαδικασίας. Τα ορολογικά προϊόντα αποτελούν εργαλεία απαραίτητα για τη μεταφορά της γνώσης.

Έχει ιδιαίτερη σημασία το γεγονός ότι η ορολογία

αναγνωρίζεται ως αυτόνομος επιστημονικός κλάδος και το ότι ιδρύεται μία πανευρωπαϊκή υποδομή για την ανταλλαγή απόψεων, τον σχεδιασμό συγκεκριμένων δράσεων και την υλοποίηση διεθνικών μορφών συνεργασίας μεταξύ ορολόγων και Ευρωπαίων ειδικών, η οποία είναι και ο κύριος στόχος του "Συνεδρίου για τη συνεργασία στον χώρο της ορολογίας στην Ευρώπη".

Το συνέδριο αυτό διοργανώνεται με την πρωτοβουλία της Ευρωπαϊκής Ένωσης Ορολογίας (European Association for Terminology / EAFT) και σε συνεργασία με τις εξής εθνικές ενώσεις ορολογίας:

- AETER (Asociación Española de Terminología),
- ΕΛΕΤΟ (Ελληνική Εταιρεία Ορολογίας),
- BriTerm (Association Britannique de Terminologie),
- DTT (Deutscher Terminologie-Tag),
- TermRom-Bucarest (Association Roumaine de Terminologie),
- TermRom-Moldova (Association Moldave de Terminologie),
- DANTERM (Association Danoise de Terminologie),
- Termip (Association Portugaise de Terminologie),
- Ass.I.Term (Association Italienne de Terminologie),
- NL-Term (Association Néerlandaise de Terminologie), και
- ProTLS (Association des Professionnels du Traitement des Langages Spécialisés).

Στις 6 Δεκεμβρίου 1998, τα μέλη του Διοικητικού Συμβουλίου της EAFT, και οι αντιπρόσωποι, ή οι Πρόεδροι, των ενώσεων που προαναφέρθηκαν, συναντήθηκαν στο Παρίσι για να προετοιμάσουν τη διοργάνωση του Συνεδρίου. Στη διάρκεια της συνάντησης, συζήτησαν τους στόχους του συνεδρίου και τα αναμενόμενα αποτελέσματα.

Με στόχο την επίτευξη των καλύτερων δυνατών αποτελεσμάτων από το Συνέδριο, έχει ξεκινήσει έρευνα μεταξύ των μελών των ευρωπαϊκών εθνικών ενώσεων ορολογίας. Η ανάλυση της έρευνας αυτής θα καθορίσει και τα θέματα που θα συζητηθούν στις ειδικές θεματικές συνεδρίες. Στις συνεδρίες αυτές, ειδικοί θα παρουσιάσουν τα συγκεκριμένα προβλήματα που αφορούν στις δραστηριότητες των ορολόγων. Συγχρόνως, θα οργανωθούν επιδείξεις ορολογικών εργαλείων, σχετικών κόμβων στον Ιστό (websites), κτλ.

Το Συνέδριο θα πραγματοποιηθεί στις 17, 18 και 19 Μαΐου 1999 στο Παρίσι ή στην ευρύτερη περιοχή του Παρισιού. Την τοπική διοργάνωση έχει αναλάβει η Union Latine. Η πρωινή συνεδρία της 19ης Μαΐου θα είναι αφιερωμένη στην παρουσίαση των συμπερασμάτων του Συνεδρίου (με τη μορφή στρογγυλής τραπέζης), ενώ το απόγευμα θα γίνει η ετήσια Γενική Συνέλευση της EAFT.

Σκοπός του Συνεδρίου δεν είναι η συζήτηση επιστημονικών θεμάτων. Οι στόχοι του είναι πολλαπλοί: μεταξύ άλλων, στο Συνέδριο θα συζητηθούν η συνεργασία μεταξύ των ενώσεων ορολογίας, τα προβλήματα που αντιμετωπίζουν ορολόγοι και ειδικοί στις ορολογικές τους δραστηριότητες, θα προταθούν λύσεις για αυτά τα προβλήματα και θα συζητηθεί η δημιουργία μιας ορολογικής υποδομής στην Ευρώπη.

Το Συνέδριο θα ολοκληρωθεί με τη σύνταξη ενός Σχεδίου Δράσης με βάση τις αποφάσεις που θα έχουν ληφθεί για τις απαιτήσεις των ορολόγων. Το Σχέδιο Δράσης θα προτείνει διάφορες μορφές συνεργασίας, όπως θα έχουν προκύψει από το Συνέδριο.

Στο άμεσο μέλλον, θα δημοσιευθεί Πρόσκληση Υποβολής Ανακοινώσεων. Οι ανακοινώσεις του Συνεδρίου θα περιληφθούν στα Πρακτικά.

Πληροφορίες για το Συνέδριο, μπορείτε να βρείτε και στην εξής σελίδα του διαδικτύου:
http://www.unilat.org/dtil/form/CONFERENCE_Fr.htm

Για περισσότερες πληροφορίες, παρακαλώ επικοινωνήστε με:

Ms. Helmi Sonneveld
President of EAFT
A. van Duinkerkenlaan 39
NL-1187 WD AMSTELVEEN
THE NETHERLANDS
Τηλ.: + 31 20 685 11 94
Τηλεομ.: +31 20 453 75 83
Ηλ. Ταχυδρ.: topterm@euronet.nl
ή:
M. Daniel Prado
Direction Terminologie et Industries de la Langue / Union Latine
131, rue du Bac

75007 Paris / France
Τηλ.: (33 1) 45 49 60 60
Τηλεομ.: (33 1) 45 44 45 97
Ηλ. Ταχυδρ.: dtil@calva.net

ή (για την Ελλάδα):
Κ. Π. Λαμπροπούλου
Ταμίας EAFT
Ινστιτούτο Επεξεργασίας του Λόγου
Επιδαύρου & Αρτέμιδος
151 25 Μαρούσι
Τηλ.: 6800952 - 4, 68 00 959
Τηλεομ.: 6852 620
Ηλ. Ταχυδρ.: penny@ilsp.gr

Συναντήσεις Εργασίας/ Workshops

First International Workshop on
Practical Aspects of Declarative Languages (PADL'99)
Menger Hotel, San Antonio, Texas
Jan. 18-19, 1999
http://www.cs.nmsu.edu/~complog/conferences/padl99

Sponsored by COMPULOG AMERICAS and the Association for Logic Programming

In Cooperation with ACM SIGPLAN

The goal of PADL'99 is to bring together researchers, practitioners and implementors of declarative languages to discuss practical issues and implications of their research results. Thus, papers dealing with practical applications of newly discovered results and techniques in logic, constraint, and functional programming are invited.

Papers dealing with practical applications of theoretical results, new implementation techniques, or innovative applications are particularly welcome. Position papers as well as papers that present work-in-progress are also welcome.

Scope of PADL includes, but is not limited to:
1. Innovative Applications
2. Practical Applications of Theoretical Results
3. Declarative Languages and the Internet
4. Declarative Languages and Software Engineering
5. Declarative Languages and Software Enabled Control
6. Deductive Database Systems
7. Specification and Verification
8. Practical Experiences
9. Innovative Implementation/Compilation Techniques (especially to support applications)

Submission of Papers:
Authors may submit an electronic copy of the full paper, in English, to the e-mail address below (preferred), or submit six copies of the paper to the postal address below.

The paper should reach by Aug 30th. Papers must be no longer than 15 pages, written in 12 point font and with single spacing.

Each copy of the submission must include on an extra sheet:

1. the paper title and the names and affiliations of all authors as they should appear in the advance program, should the paper be accepted;

2. an abstract;

3. three to four keywords

4. contact information: postal address(es),
   telephone number(s),
   fax number(s) (if available),
   e-mail address(es).

Each paper will be reviewed. Authors will be notified of acceptance/rejection by October 10th. Camera ready copies will be due by November 10th. Proceedings will be published as Lecture Notes in Computer Science by Springer Verlag.

Address for Submission:
Gopal Gupta
Department of Computer Science
Science Hall, Stewart Street,
New Mexico State University
Las Cruces, NM 88003-0001
Telephone: +1 (505) 646 6236

Fax: +1 (505) 646 1002
complog@cs.nmsu.edu

Contact for More Information:
Gopal Gupta
Laboratory for Logic, Databases, and Advanced Programming
Department of Computer Science
Box 30001, Dept. CS
New Mexico State University
Las Cruces, NM 88003-0001
USA

Web: http://www.cs.nmsu.edu/lldap/
e-mail: gupta@nmsu.edu
Telephone: +1 (505) 646 6236
Fax: +1 (505) 646 1002

Θερινά Σχολεία / Summer Schools

Machine Learning and Applications
Advanced Course on Artificial Intelligence 1999 (ACAI-99)
5-16 July 1999, Greece

Preliminary Announcement
(http://www.iit.demokritos.gr/skel/eetn/acai99)

European Coordinating Committee on Artificial Intelligence (ECCAI)
(http://www.eccai.org)

& Hellenic Artificial Intelligence Society (EETN)
(http://www.iit.demokritos.gr/skel/eetn)

Aim of the course:
The ECCAI Advanced Course on Artificial Intelligence for 1999 (ACAI-99) is organised by the Hellenic Artificial Inteligence Society (EETN) on a Greek island.

The goal of ACAI-99 is to present the current state of the art in Machine Learning, as well as to show the potential of Machine Learning in a variety of problems. Towards this goal, the course is structured as a multimodal event, containing plenary sessions from distinguished lecturers, workshops on machine learning applications, student sessions and panel discussions on the success of ML so far and its future directions. ACAI-99 will be of benefit to professionals, who are interested in using machine learning techniques, as well as researchers and postgraduate students, who work or are thinking of working in this exciting field.

List of plenary talks:
T. Mitchell (CMU, USA)
"Machine learning: Setting the scene"
M. van Someren (Amsterdam, Holland)
"Machine learning and knowledge acquisition: from ML technology push to user oriented design"
R. Michalski (George Mason, USA)
"Concept learning"
Y. Kodrattoff (CNRS, France)
"Are ML applications a subfield of KDD?"
R.L. de Mantaras (CSIC, Spain)
"Case-based reasoning"
I. Bratko (Ljublana, Slovenia)
"Noise handling in tree induction" & "Inducing intermediate concepts"
P. Langley (ISLE, USA)
"Machine discovery" & "Formation of probabilistic concept hierarchies"
L. de Raedt (KU Lueven, Belgium)
"Inductive logic programming for data mining and machine learning"
N. Tishby (Hebrew Univ., Israel)
"A unified information theoretic approach to prediction, clustering and learning"
C. Bishop (Microsoft, UK)
"Probabilistic graphical models"
J. Shapiro (Manchester, UK)
"Genetic algorithms in artificial intelligence"
L. Saitta (Torino, Italy)
"Multi-strategy learning"

Important dates:
October 30, 1998
Deadline for Submission of Workshop Proposals
December 1, 1998
Announcement of Selected Workshops
March 1, 1999
Deadline for Grant Applications & Submission of Student Papers
March 15, 1999

Announcement of Grant Offers & Collected Student papers

April 1, 1999
Deadline for Workshop Programmes and Delivery of Workshop Material

April 15, 1999
Deadline of Early Registration

June 15, 1999
Deadline of Late Registration

Committees:

ACAI-99 Chair: C. D. Spyropoulos (costass@iit.demokritos.gr), (NCSR "Demokritos", Greece)

ACAI-99 Co-chair:
N. Fakotakis (fakotaki@wcl.ee.upatras.gr), (Univ. of Patras, Greece)

Organisation Committee:

V. Karkaletsis, (NCSR "Demokritos", Greece)
J. Kontos, (Athens University of Economics & Business, Greece)
I. Malagardi, (ILSP, Greece)
V. Moustakis, (Technical Univ. of Crete, Chania, Greece)
G. Paliouras, (NCSR "Demokritos", Greece)
G. Vouros, (Univ. of Aegean, Greece)

Programme Committee:

Chair: V. Moustakis (moustaki@csi.forth.gr), (Technical Univ. of Crete, Chania, Greece)

Co-chair: G. Paliouras (paliourg@iit.demokritos.gr), (NCSR "Demokritos", Greece)

Members:

C. Bishop, (Microsoft Research Laboratory, UK)
I. Bratko, (Ljubljana University, Slovenia)
M. Giakoumakis, (Athens University of Economics & Business, Greece)
D. Kalles, (Computer Technology Institute, Greece)
V. Karkaletsis, (NCSR "Demokritos", Greece)
Y. Kodratoff, (CNRS, Univ. of Paris-Sud, France)
J. Kontos, (Athens University of Economics & Business, Greece)
P. Langley, (ISLE, USA)
R. Lopez de Mantaras, (CSIC, Spain)
R. Michalski, (George Mason University, USA)
T. Mitchell, (Carnegie Mellon University, USA)
L. de Raedt, (KU Leuven, Belgium)
T. Panayiotopoulos, (University of Piraeus, Greece)
J. Pitas, (Univ. of Thessaloniki, Greece)
L. Saitta, (University of Torino, Italy)
T. Sellis, (Technical Univ. of Athens, Greece)
J. Shapiro, (Manchester University, UK)
M. van Someren, (University of Amsterdam, Holland)
S. Tzafestas, (Technical Univ. of Athens, Greece)
N. Tishby, (The Hebrew University, Israel)
N. Vassilas, (NCSR "Demokritos", Greece)

Workshops:

V. Karkaletsis (vangelis@iit.demokritos.gr), (NCSR "Demokritos", Greece)

Student sessions:

N. Vassilas (nvas@iit.demokritos.gr), (NCSR "Demokritos", Greece)