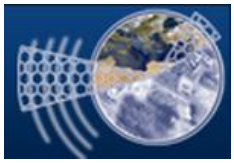
Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης

Σχολή Τεχνολογικών Εφαρμογών
Τμήμα Μηχανικών Πληροφορικής



Πτυχιακή Εργασία

Δημιουργία ενός survey για την Επεξεργασία Φυσικής Γλώσσας με Βάσεις Δεδομένων

Αικατερίνη Αρναουτάκη (ΑΜ: 1884)

Επιβλέπων Καθηγητής: Νικόλαος Παπαδάκης

Ημερομηνία Παρουσίασης: 18 Μαρτίου 2015

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου κ. Νικόλαο Παπαδάκη για την εμπιστοσύνη που μου έδειξε κατά τη διάρκεια υλοποίησης της πτυχιακής μου εργασίας, όπως επίσης και για την πολύτιμη βοήθεια και καθοδήγηση του.

Θα ήθελα να ευχαριστήσω θερμά την κ. Παναγιώτα Σπάλα για την βοήθεια της σχετικά με το αντικείμενο της πτυχιακής μου εργασίας.

Θα ήθελα επίσης να απευθύνω τις ευχαριστίες μου στους γονείς μου και την οικογένειά μου που στήριξαν τις σπουδές μου όλα αυτά τα χρόνια ηθικά και οικονομικά, φροντίζοντας για την καλύτερη μόρφωση μου.

Τέλος θέλω να ευχαριστήσω όλους τους καθηγητές για τις γνώσεις που μου μετέδωσαν, τους φίλους και συναδέλφους μου, όπου με την καθημερινή τους συμπαράσταση συνέβαλαν στην εκπλήρωση του στόχου μου.

Abstract

Natural Language Processing is a branch of Computer Science, within the domains of Artificial Intelligence and Linguistics. This research area deals with the interactions that occur between computer systems and users, who use their natural language as the communication medium. Therefore, Natural Language Processing is closely related with human-computer interaction, language understanding, and the ways of extracting knowledge from any linguistic morphology and composition.

In modern times, novel and innovative expert systems are constantly being developed, to serve Natural Language Processing on computers via databases. This work studies the technologies and innovations that comprise these expert human-computer interaction systems via Natural Language Processing, as well as the expertise and requirements of such interfaces.

Σύνοψη

Η Επεξεργασία του Φυσικού Λόγου είναι ένας κλάδος της επιστήμης της πληροφορικής, που έγκειται στους ερευνητικούς τομείς της τεχνητής νοημοσύνης και της γλωσσολογίας. Ο κλάδος αυτός ασχολείται με τις αλληλεπιδράσεις που πραγματοποιούνται μεταξύ των υπολογιστών και των χρηστών, οι οποίοι επικοινωνούν με τα συστήματα αυτά μέσω της φυσικής τους γλώσσας. Επομένως, η Επεξεργασία του Φυσικού Λόγου συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή, με την κατανόηση της Γλώσσας, και την εξαγωγή γνώσης από την εκάστοτε γλωσσική μορφολογία και σύνθεση.

Στη σύγχρονη εποχή, κατασκευάζονται συνεχώς πρωτότυπα και καινοτόμα έμπειρα συστήματα και εργαλεία, τα οποία εξυπηρετούνται στην επεξεργασία της φυσικής γλώσσας σε υπολογιστές από Βάσεις Δεδομένων. Στην εργασία αυτή ερευνώνται οι τεχνολογίες και καινοτομίες που συνθέτουν ένα τέτοιο έμπειρο σύστημα επικοινωνίας ανθρώπου-υπολογιστή μέσω Φυσικής Γλώσσας, καθώς και η τεχνογνωσία και απαιτήσεις μίας τέτοιας διεπαφής.

Πίνακας Περιεχομένων

Πίνακας Εικόνων

Πίνακας Σχημάτων

ΜΕΡΟΣ Α

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

1 Εισαγωγή

1.1 Περίληψη

Η εργασία αυτή μελετά και αναλύει τα συστήματα και εργαλεία Επεξεργασίας Φυσικού Λόγου που έχουν προταθεί ως διεπαφές εφαρμογών σε συνδυασμό δεδομένων που βρίσκονται καταχωρημένα σε σχεσιακές βάσεις δεδομένων. Αρχικώς, θα παρουσιαστούν οι βασικές έννοιες και απαραίτητοι ορισμοί που αφορούν τον τομέα της Φυσικής Γλώσσας. Αμέσως μετά, αναλύονται τα επιμέρους επίπεδα ανάλυσης της Φυσικής Γλώσσας και τα πεδία στα οποία βρίσκει εφαρμογή. Παράλληλα, αποσαφηνίζονται τα μοντέλα σύνταξης και γραμματικής ανάλυσης της Γλώσσας. Έπειτα, παρουσιάζονται έννοιες και όροι που αφορούν στην μοντελοποίηση δεδομένων κάνοντας χρήση κατάλληλα δομημένων βάσεων δεδομένων, και η χρησιμότητα αυτών σε εφαρμογές Επεξεργασίας Φυσικής Γλώσσας. Αφού ολοκληρωθεί η παρουσίαση των βασικών εννοιών, θα αναλυθεί η ποιοτική και διεξοδική μελέτη και επισκόπηση των πιο σύγχρονων μεθόδων επεξεργασίας του Φυσικού Λόγου και των πεδίων εφαρμογής τους. Τέλος, η εργασία ολοκληρώνεται μέσω της έκθεσης των συμπερασμάτων που προκύπτουν από την έρευνα αυτή, καθώς και τα αποτελέσματα που έχουν αναφερθεί από τους ίδιους τους ερευνητικές των μεθόδων και εργαλείων αυτών.

1.2 Κίνητρο για τη διεξαγωγή της εργασίας

Οι διαδικτυακές διεπαφές υπηρεσιών γίνονται ολοένα και πιο φιλικές προς τους χρήστες. Ταυτόχρονα, η Επεξεργασία της Φυσικής Γλώσσας τείνει να διαδραματίζει ένα σημαντικό ρόλο στη σχεδίαση και την ανάπτυξη επιτυχημένων διαδικτυακών εφαρμογών. Η έρευνα και ανάπτυξη της Επεξεργασίας της Φυσικής Γλώσσας τα τελευταία τουλάχιστον πέντε (5) έτη έχει επικεντρωθεί στα πεδία Κατανόησης του Φυσικού Λόγου, Αυτόματης Παραγωγής Φυσικού Λόγου, Αναγνώρισης Ομιλίας, Μηχανικής Μετάφρασης, και Ορθογραφικής Διόρθωσης και Γραμματικού Ελέγχου της Γλώσσας. Βεβαίως, η Γλώσσα είναι πιο σύνθετη από μία απλή ανταλλαγή πληροφορίας. Η Γλώσσα είναι ένα σύνολο στοιχείων που επιτρέπει στον άνθρωπο την ανταλλαγή σημασιών, αλλά σπάνια θεωρείται ως ένα μέσο κωδικοποίησης σημασιών. Το σύστημα ροής της πληροφορίας κατά την Επεξεργασία του Φυσικού Λόγου παίρνει ως είσοδο δεδομένα Φυσικής Γλώσσας, τα οποία στη συνέχεια επεξεργάζονται, αναλύονται, αξιολογούνται και παράγονται πληροφορίες οι οποίες δίδονται στην έξοδο του συστήματος ως αποτέλεσμα, απάντηση ή δράση.

Τα συστήματα Επεξεργασίας του Φυσικού Λόγου πρέπει να διαθέτουν σημαντική γνώση για την δομή της ίδιας της γλώσσας, τις λέξεις της, του συνδυασμού των λέξεων σε προτάσεις, των εννοιών των λέξεων, το πώς συμμετέχουν οι λέξεις στη σημασία της πρότασης, κοκ. Τα συστήματα αυτά απαιτούν την ύπαρξη μεθόδων κωδικοποίησης και χρήσης της γνώσης που να παράγουν τις κατάλληλες συμπεριφορές. Επίσης, θα πρέπει να λαμβάνεται υπόψη το πλαίσιο συμφραζομένων της Γλώσσας πάνω στο οποίο γίνεται η Επεξεργασία. Ιδανικά, τα εργαλεία Επεξεργασίας Φυσικού Λόγου θα πρέπει να είναι σε θέση να επεξεργάζονται δομημένες και αδόμητες προτάσεις με απλά ερωτήματα, παράγοντας μία ενιαία, συνδυαστική απάντηση, ουδέτερη προς τα δεδομένα. Η ανάγκη δημιουργίας καινοτόμων διαδικτυακών εφαρμογών που συνδυάζουν την Επεξεργασία Φυσικής Γλώσσας με διαδικτυακές υπηρεσίες, προκειμένου να εξυπηρετείται ευκολότερα ο χρήστης, έχει οδηγήσει την ερευνητική κοινότητα στην μελέτη και παραγωγή πρότυπων μεθόδων και εργαλείων στον τομέα Ανάλυσης και Επεξεργασίας του Φυσικού Λόγου. Η μελέτη, ανάλυση, και εξαγωγή συμπερασμάτων από τις σύγχρονες αυτές μεθόδους Επεξεργασίας της Φυσικής Γλώσσας ως διεπαφές επικοινωνίας χρήστη-υπολογιστή είναι το βασικό κίνητρο της συγκεκριμένης εργασίας.

1.3 Σκοπός και στόχοι της εργασίας

Σκοπός της παρούσας εργασίας είναι η συστηματική μελέτη, ανασκόπηση και σύγκριση των νεότερων ερευνητικών μεθόδων που συνθέτουν λύσεις και συστήματα τα οποία έχουν προταθεί για χρήση της επεξεργασίας φυσικής γλώσσας ως διεπαφή σε σχεσιακές βάσεις δεδομένων. Η εργασία στοχεύει στην ποιοτική ανάλυση και μελέτη των πιο σύγχρονων ερευνητικών μεθόδων που

χρησιμοποιούν ή/και εφαρμόζουν την επεξεργασία Φυσικής Γλώσσας προκειμένου να εξυπηρετούνται διεπαφές ανθρώπου-υπολογιστή. Ακόμα, στα πλαίσια της συγκεκριμένης μελέτης εμπίπτει η προσπάθεια εξαγωγής χρήσιμων συμπερασμάτων για την αποτελεσματικότητα των συστημάτων και μεθόδων αυτών, καθώς και για τις μελλοντικές τάσεις του τομέα αυτού.

1.4 Δομή της εργασίας

Η εργασία χωρίζεται σε δύο διακριτά μέρη. Το **Μέρος Α** (Κεφάλαια 1-4) παραθέτει τα απαραίτητα εισαγωγικά στοιχεία και τις βασικές έννοιες Επεξεργασίας Φυσικής Γλώσσας και Βάσεων Δεδομένων, προκειμένου ο αναγνώστης να είναι σε θέση να κατανοήσει το δεύτερο και κυριότερο μέρος. Το πρώτο Κεφάλαιο παρουσιάζει το περιεχόμενο, τα κίνητρα και τους στόχους και σκοπούς διεξαγωγής της εργασίας. Στη συνέχεια, το δεύτερο Κεφάλαιο καταγράφει τις βασικές έννοιες της Φυσικής Γλώσσας και την επεξεργασία αυτής μέσω πληροφοριακών συστημάτων και ερευνητικών πεδίων. Στο τρίτο Κεφάλαιο καταγράφονται τα επίπεδα ανάλυσης της Φυσικής Γλώσσας, καθώς επίσης τα συντακτικά και γραμματικά συστατικά ανάλυσης Γλωσσών. Το τέταρτο Κεφάλαιο κάνει μία επισκόπηση των συστημάτων Βάσεων Δεδομένων και τον τρόπο με τον οποίο τα συστήματα αυτά εξυπηρετούν την ανάλυση και επεξεργασία Φυσικών Γλωσσών.

Το **Μέρος Β** (Κεφάλαια 5-7) περιλαμβάνει την επισκόπηση και σύγκριση των σύγχρονων ερευνητικών μεθόδων και εργαλείων επεξεργασίας της φυσικής γλώσσας που πραγματοποιούνται σε συνδυασμό με σχεσιακές βάσεις δεδομένων. Το πέμπτο Κεφάλαιο παρουσιάζει σύγχρονα ολοκληρωμένα εργαλεία ή συστήματα που χρησιμοποιούν την τεχνολογία της επεξεργασίας Φυσικής Γλώσσας με βάσεις δεδομένων, όπως προκύπτουν από έρευνες, εφαρμοζόμενες στον τομέα στον οποίο προσφέρουν την εκάστοτε λύση. Στο Κεφάλαιο 6, παρουσιάζονται οι νεότερες ερευνητικές μέθοδοι επεξεργασίας της Φυσικής Γλώσσας, που εξάγουν πληροφορία από βάσεις δεδομένων, καθώς και τα πειραματικά τους αποτελέσματα. Το Κεφάλαιο 7 ολοκληρώνει την μελέτη της εργασίας, εκθέτοντας τα συμπεράσματα που προκύπτουν από την συγκεκριμένη μελέτη και επισκόπηση, καθώς και τις μελλοντικές τάσεις στον τομέα αυτό.

2 Φυσική Γλώσσα

2.1 Βασικές έννοιες φυσικής γλώσσας

2.1.1 Γλώσσα και Επικοινωνία

Ο όρος επικοινωνία είναι δύσκολο να οριστεί ακριβώς καθώς απευθύνεται σε πολλά πεδία και περιοχές γνώσης. Γενικότερα, ως επικοινωνία θα μπορούσε να οριστεί η σκόπιμη ανταλλαγή πληροφορίας, μέσω παραγόμενων και κατανοούμενων σημείων (signs) τα οποία επιλέγονται από κάποιο σύστημα συμβατικών σημείων [1]. Ως γνωστό, τα ζώα χρησιμοποιούν ένα περιορισμένο σύνολο τέτοιων σημείων, όπως για παράδειγμα τη κίνηση του σώματος, τις χειρονομίες, κάποια δεδομένη χειραψία κλπ. Το ανθρώπινο είδος, πέραν του περιορισμένου συνόλου αυτού των ζώων, έχει επιπλέον αναπτύξει ένα πολύ πιο περίπλοκο σύνολο σημείων, το οποίο εμπεριέχει συγκεκριμένη δομή και κανόνες. Το ολοκληρωμένο σύστημα αυτό είναι η γλώσσα. Ο άνθρωπος είναι το μόνο είδος που έχει τη δυνατότητα να επικοινωνεί αξιόπιστα χρησιμοποιώντας ένα μεγάλο πλήθος ποιοτικά διαφορετικών μηνυμάτων, για να εκφραστεί και να παραθέτει εμπειρίες, ιδέες, κλπ. Αντιθέτως, τα ζώα διαθέτουν θεωρητικά ένα αντίστοιχα άπειρο πλήθος μηνυμάτων, όχι όμως ποιοτικά διαφορετικών.

2.1.2 Σημασία

Η γλώσσα αποτελείται από θεμελιώδη νοηματικά στοιχεία, τις λέξεις. Οι λέξεις με τη σειρά τους συνδέονται μεταξύ τους παράγοντας τη σημασία (meaning) την οποία μεταφέρουν οι άνθρωποι που επικοινωνούν με το σύνολο των λέξεων αυτών. Η σημασία αυτή προκύπτει δηλωτικά ή μέσω υπονόησης.

Παραδείγματος χάρη, η σημασία διαφέρει μεταξύ των κυρίων ονομάτων ατόμων όπως είναι τα πρόσωπα, οι πόλεις, κλπ., από τα ονόματα που αντιστοιχούν σε διάφορα αντικείμενα. Αντίστοιχα, παράγεται διαφορετική σημασία από τα ρήματα που δείχνουν δράση (π.χ. «τρέχω στο δρόμο»), μία ή περισσότερες καταστάσεις (π.χ. «το παιδί τρώει») ή σε συσχετίσεις των αντικειμένων (π.χ. «το προϊόν κοστίζει τρία ευρώ»). Προφανώς, όπως φαίνεται από την πληθώρα γλωσσών του ανθρώπου, η αντιστοιχία λέξεων και σημασιών δεν είναι ούτε μοναδική ούτε απαραίτητα αναγκαία.

Ο τρόπος με τον οποίο κατανοείται η εκάστοτε γλώσσα είτε σε ανθρώπινη μορφή είτε μέσω υπολογιστικών συστημάτων, στηρίζεται στις παραπάνω συμβάσεις σημασίας των λέξεων. Συνεπώς, τα συστήματα επεξεργασίας φυσικής γλώσσας χρειάζονται την εισαγωγή αυτών των συμβάσεων μεταξύ λέξεων και σημασιών κατά την έναρξή τους, ώστε να επεξεργαστούν την δεδομένη γλώσσα.

Η κατανόηση της γλώσσας (language understanding) μέσω λέξεων προκύπτει αντιστοιχώντας την παραγόμενη σημασία με τις δεδομένες λέξεις της φράσης. Στον χώρο των υπολογιστών, ο τομέας της Τεχνητής Νοημοσύνης (Artificial Intelligence – AI) ασχολείται με την διαδικασία (procedure) ανεύρεσης των δομών αυτών που συντάσσουν τη σημασία της φράσης και της αντιστοίχισής τους με την εκάστοτε σημασία. Κατά τον τρόπο αυτό, προκύπτει η διαδικαστική σημασιολογία (procedural semantics), όπου η διαδικασία που ακολουθείται προκύπτει από τη σημασία της εντολής εισόδου του συστήματος, το οποίο με τη σειρά του εκτελεί μία συγκεκριμένη ενέργεια. Ευρύτερα, δίδεται μία ερώτηση ως είσοδος στο σύστημα, η οποία αντιστοιχεί σε μία απάντηση. Η πληροφορία που προκύπτει κατά την αντιστοίχιση της δεδομένης ερώτησης και απάντησης ονομάζεται δήλωση, εμπλουτίζοντας το σύστημα με νέες σημασίες και έννοιες, με υπολογιστικό τρόπο.

2.1.3 Γνώση

Ο όρος γνώση είναι δύσκολο να οριστεί με συγκεκριμένο τρόπο, για αυτό και στην σύγχρονη βιβλιογραφία δεν υπάρχει ένας ενιαίος και απόλυτος ορισμός. Θα μπορούσε κανείς να ορίσει την γνώση ως την θεωρητική και πρακτική κατανόηση ενός θέματος, δηλαδή την αφομοίωση επεξεργασμένων πληροφοριών, η οποία ενδεχομένως βοηθά στην απόκτηση δεξιοτήτων και ικανοτήτων για συγκεκριμένο σκοπό [2]. Η γνώση αποκτάται τόσο με την εκπαίδευση όσο και με την εμπειρία. Τα συστήματα κατανόησης και επεξεργασίας της γλώσσας θα πρέπει να είναι ικανά να παράγουν ή συνθέτουν γνώση, εντός του γλωσσικού κόσμου που αναφέρονται, αναλύοντας τις δηλωτικές ενέργειες (ερώτηση / απάντηση) με τα μοντέλα που προκύπτουν. Συλλέγοντας τις επεξεργασμένες αυτές δηλώσεις μέσω των μοντέλων, αντλείται η νέα πληροφορία, που συνθέτει τη γνώση του συστήματος.

2.2 Επεξεργασία Φυσικής Γλώσσας

Η Υπολογιστική Γλωσσολογία είναι ο επιστημονικός τομέας της Γλωσσολογίας που ασχολείται με την μοντελοποίηση της φυσικής γλώσσας μέσω υπολογιστικών αλγορίθμων. Ονομάζεται αλλιώς και Επεξεργασία Φυσικής Γλώσσας και διαιρείται σε δύο κλάδους, τον θεωρητικό και πρακτικό. Η θεωρητική προσέγγιση εστιάζει στις θεωρίες αναπαράστασης και γνώσης γλωσσών, όπως επίσης και τους κανόνες που χρησιμοποιούν οι άνθρωποι για τη χρήση της γλώσσας. Στο κλάδο αυτό, η επιστημονική κοινότητα αναπτύσσει σαφώς καθορισμένα μοντέλα, εξομοιώνοντας απόψεις της ανθρώπινης γλωσσολογίας. Τα μοντέλα αυτά υλοποιούνται σε δεύτερη φάση σε υπολογιστικά συστήματα. Η εφαρμοσμένη Υπολογιστική Γλωσσολογία εστιάζει στα πρακτικά αποτελέσματα της μοντελοποίησης που προκύπτει από τις θεωρητικές προσεγγίσεις. Οι μέθοδοι, τεχνικές, τα εργαλεία και οι εφαρμογές του κλάδου αυτού συνθέτουν την γλωσσική μηχανική (language engineering) ή γλωσσική τεχνολογία (language technology). Αν και τα σύγχρονα υπολογιστικά συστήματα αποκλίνουν από την ανθρώπινη δυνατότητα γνώσης και κατανόησης της γλώσσας, η σημερινή έρευνα έχει καταφέρει να αναπτύξει πολλές εφαρμογές Τεχνητής Νοημοσύνης γύρω από την επεξεργασία του Φυσικού Λόγου. Στόχος είναι η ανάπτυξη συστημάτων με κάποια γνώση της ανθρώπινης γλώσσας, εμπλουτίζοντας την επικοινωνία μεταξύ ανθρώπου και μηχανής.

Ο κλάδος της εφαρμοσμένης Υπολογιστικής Γλωσσολογίας καλύπτει:

- Τον **σχεδιασμό και την υλοποίηση υπολογιστικών μοντέλων της φυσικής γλώσσας,** και συγκεκριμένα την αναγνώριση και κατανόηση της φυσικής γλώσσας από το υπολογιστικό σύστημα, καθώς και την παραγωγή φυσικής γλώσσας δια μέσου του υπολογιστή.
- Τις **εφαρμογές που αναπτύσσονται** για την δημιουργία διαλόγων με τον υπολογιστή και τη μηχανική μετάφραση, την εύρεση και φιλτράρισμα κειμένων.

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) αντιστοιχεί στην εφαρμοσμένη Υπολογιστική Γλωσσολογία και ανήκει στον τομέα Τεχνητής Νοημοσύνης. Η Υπολογιστική Γλωσσολογία ως διεπιστημονική γνωστική περιοχή κατατάσσεται μεταξύ Γλωσσολογίας και Πληροφορικής. Ανήκει στις Γνωστικές Επιστήμες (Cognitive Sciences) και επικαλύπτεται με την Τεχνητή Νοημοσύνη (Artificial Intelligence), που ασχολείται με τα υπολογιστικά μοντέλα της ανθρώπινης γνώσης.

2.3 Πεδία έρευνας Επεξεργασίας Φυσικής Γλώσσας

Στην υπο-ενότητα αυτή παρουσιάζονται τα κυριότερα πεδία στα οποία γίνεται εκτεταμένη έρευνα της Επεξεργασίας Φυσικής Γλώσσας. Το κριτήριο διαχωρισμού των πεδίων αυτών είναι το γεγονός ότι για το καθένα από αυτά υπάρχει ένας επίσημα ορισμένος χώρος μελέτης και επίλυσης ζητημάτων, ένα καθιερωμένο μετρικό σύστημα για την αξιολόγηση των ερευνών που προκύπτουν από το πεδίο, κάποια δεδομένα σύνολα κειμένων πάνω στα οποία κάθε πεδίο αξιολογείται και διαγωνισμοί αφιερωμένοι στο κάθε πεδίο [3].

1. **Ανάλυση λόγου**. Αναγνώριση της δομής του λόγου εντός των αναλυόμενων κειμένων, π.χ. την φύση των σχέσεων του λόγου μεταξύ δύο προτάσεων. Επίσης, αναφέρεται στην αναγνώριση και την κατηγοριοποίηση των γλωσσικών πράξεων σε ένα μέρος του κειμένου.
2. **Αυτόματη αναγνώριση ομιλίας.** Η αυτόματη μετατροπή του ανθρώπινου λόγου όπως προφέρεται σε κείμενο από τους τους υπολογιστές.
3. **Αυτόματη ερωταπόκριση**. Η αναζήτηση της σωστής απάντησης σε μία δεδομένη ερώτηση, όπως διαμορφώνεται από την ανθρώπινη γλώσσα.
4. **Αυτόματη μορφολογική τεμαχιοποίηση**. Η κατάτμηση των λέξεων στα μορφήματά τους καθώς και η αναγνώριση και κατηγοριοποίηση αυτών των μορφημάτων. Η δυσκολία του συγκεκριμένου πεδίου μελέτης εξαρτάται σε μεγάλο βαθμό από την περιπλοκότητα της μορφολογίας της εκάστοτε γλώσσας υπό εξέταση.
5. **Αυτόματη περίληψη.** Η παραγωγή μίας αναγνώσιμης (από τον άνθρωπο) περίληψης ενός κειμένου. Συχνά χρησιμοποιείται για να παρέχει περιλήψεις σε κείμενα γνωστής διάταξης, όπως οικονομικά ή πολιτικά ειδησεογραφικά άρθρα.

6. **Εξόρυξη πληροφοριών.** Η ανάκτηση πληροφοριών από μη δομημένα ή ημι-δομημένα δεδομένα (τυπικά κείμενα γραμμένα σε φυσική γλώσσα, ιστοσελίδες κ.α.)
7. **Επίλυση σχέσεων συναναφοράς.** Η αναζήτηση των λέξεων (αναφορές) οι οποίες αναφέρονται στα ίδια υποκείμενα (οντότητες) σε μία δεδομένη πρόταση ή μεγαλύτερο τμήμα κειμένου. Η επίλυση σχέσεων αναφοράς είναι ένα συγκεκριμένο παράδειγμα αυτού του πεδίου και αναφέρεται συγκεκριμένα στην σύνδεση των αντωνυμιών με τα ουσιαστικά ή τα ονόματα στα οποία αναφέρονται.
8. **Επισήμανση των μερών του λόγου.** Ο αυτόματος καθορισμός των μερών του λόγου σε μία δεδομένη πρόταση και η επίλυση της συντακτικής αμφισημίας.
9. **Κατανόηση του φυσικού λόγου.** Η μετατροπή κομματιών κειμένου σε πιο τυπικές αναπαραστάσεις όπως σε δομές λογικής πρώτου βαθμού, οι οποίες μπορούν να μεταχειριστούν ευκολότερα από τους υπολογιστές.
10. **Μηχανική μετάφραση.** Η αυτόματη μετάφραση ενός κειμένου από μία ανθρώπινη γλώσσα σε μία άλλη [4] [5].
11. **Οπτική αναγνώριση χαρακτήρων (Optical Character Recognition - OCR)**. Ο προσδιορισμός του αντίστοιχου κειμένου από μία δεδομένη εικόνα που αναπαριστά κάποιο τυπογραφημένο κείμενο.
12. **Παραγωγή φυσικού λόγου.** Η μετατροπή των πληροφοριών από υπολογιστικές βάσεις δεδομένων σε αναγνώσιμο φυσικό λόγο.
13. **Σύνθεση ομιλίας.** Η αυτόματη, τεχνητή παραγωγή του ανθρώπινου λόγου από υπολογιστές.
14. **Συντακτική ανάλυση.** Ο αυτόματος καθορισμός της σύνταξης μίας δεδομένης πρότασης και η επίλυση των οποιοδήποτε συντακτικών αμφισημιών. Εξαιτίας των πιθανών αμφισημιών που πιθανόν να φέρει μία πρόταση, είναι δυνατόν η εν λόγω πρόταση να αναλυθεί σε παραπάνω από ένα συντακτικά δέντρα.

3 Ανάλυση Φυσικής Γλώσσας

3.1 Επίπεδα Ανάλυσης

Η φυσική γλώσσα υποδιαιρείται σε πέντε (5) διακριτά επίπεδα ανάλυσης, τα οποία είναι:

1. Φωνολογικό
2. Μορφολογικό
3. Συντακτικό
4. Σημασιολογικό
5. Πραγματολογικό

Η σύγχρονη έρευνα της φυσικής γλώσσας εστιάζει σε μεγάλο ποσοστό στις διεπαφές μεταξύ των επιπέδων αυτών. Στο Κεφάλαιο αυτό παρουσιάζονται συνοπτικά τα πέντε επίπεδα αυτά.

3.1.1 Φωνολογικό Επίπεδο

Η γλώσσα εκφράζεται αρχικά μέσω του φυσικού λόγου. Η γραφή είναι προαιρετική για την έκφραση της γλώσσας, είναι λιγότερο εκφραστική, περισσότερο συντηρητική και περιορισμένης χρήσης. Ο λόγος παράγεται σε συνεχή ροή. Σε ένα πρώτο επίπεδο ανάλυσης, ο λόγος αποτελείται από ακολουθίες επαναλαμβανόμενων ήχων, τους φθόγγους. Η έννοια του φθόγγου είναι ανεπαρκής για να καλύψει την κοινή διαίσθηση, δηλαδή ότι κάτι που λέγεται "χι" βρίσκεται στην αρχή των λέξεων "χήρος" και "χώρος" [1]. Η Γλωσσολογία χρησιμοποιεί την έννοια του φωνήματος (phoneme) για την διαίσθηση αυτή. Το φώνημα είναι μία αφηρημένη φωνολογική οντότητα που έχει διαφοροποιητική αξία μέσα στη γλώσσα. Η αξία αυτή διαφαίνεται από την μέθοδο των ελαχίστων ζευγών (minimal pairs). Τα φωνήματα εκφράζονται στον λόγο από έναν ή περισσότερους φθόγγους (αλλόφωνα). Όμως τα φωνήματα και οι φθόγγοι είναι ανεπαρκείς για την έκφραση του λόγου ως φαινόμενο, όπως τις περιπτώσεις ερωτήσεων, θαυμασμού, απλής δήλωσης κλπ. Για κάθε διαφορετική τέτοια εκφορά της ίδιας πρότασης χρησιμοποιείται διαφορετικός επιτονισμός (intonation). Ο επιτονισμός, ο οποίος είναι υπερτμηματικό (supersegmental) φαινόμενο, αποτελεί οργανικό χαρακτηριστικό του λόγου και πολύ ενδιαφέρον πεδίο μελετών και εφαρμογών.

3.1.2 Μορφολογικό Επίπεδο

Η Μορφολογία ασχολείται με τις λέξεις. Υπάρχουν δύο κύριοι τρόποι προσέγγισης: ο διαδικαστικός (procedural) και ο δηλωτικός (declarative). Ο διαδικαστικός τρόπος προσέγγισης είναι ίσως πιο κοντά στον επιστημονικό τρόπο σκέψης που ασχολείται με υπολογιστές και εφαρμογές. Το ερώτημα που θέτει είναι πώς παράγονται οι λέξεις, μέσω του οποίου ανακαλύπτει τις κατάλληλες διαδικασίες. Ο δηλωτικός τρόπος προσέγγισης απαντά στο ερώτημα ποιες λεκτικές δομές κρίνονται ως σωστές. Δηλαδή, η δηλωτική προσέγγιση διαχωρίζει την περιγραφή του γλωσσικού αντικειμένου από την παραγωγή του ενώ η διαδικαστική τα αναμιγνύει.

Ο διαδικαστικός τρόπος προσέγγισης διακρίνει δύο τρόπους σχηματισμού των λέξεων, την κλίση και την παραγωγή. Τα όρια ανάμεσα σε αυτές τις δύο διαδικασίες είναι δυσδιάκριτα.

3.1.2.1 Κλίση (Inflection)

Η μορφολογική διαδικασία της κλίσης έχει τα εξής χαρακτηριστικά.

- Δεν οδηγεί σε αλλαγή του μέρους του λόγου. Ένα ουσιαστικό που κλίνεται (Ονομαστική, Γενική, Αιτιατική κλπ) παραμένει πάντα ουσιαστικό.
- Εφαρμόζεται σε όλα τα μέλη ενός συνόλου με δεδομένα χαρακτηριστικά. Για παράδειγμα, σχεδόν όλα τα θηλυκά ουσιαστικά που λήγουν σε -η κλίνονται κατά όμοιο τρόπο, έχουν τον ίδιο αριθμό πτώσεων και δύο αριθμούς (ενικό και πληθυντικό).
- Επιφέρει ελάχιστη και συστηματική αλλαγή στη σημασία της λέξης για τη συγκεκριμένη γλώσσα. Για παράδειγμα, κλίνοντας ένα ουσιαστικό, αλλάζει ο αριθμός και η πτώση του αλλά διατηρούνται όλες εκείνες οι ιδιότητες που χαρακτηρίζουν το αντικείμενο αυτό.

3.1.2.2 Παραγωγή (Derivation)

Σε αντιδιαστολή με την κλίση, η παραγωγή έχει τα εξής χαρακτηριστικά.

- Συχνά, αλλά όχι πάντα, επιφέρει αλλαγή του μέρους του λόγου.

- Η εφαρμογή της παρουσιάζει κενά.
- Επιφέρει κατά βάση συστηματική αλλαγή στη σημασία της λέξης αλλά τα προϊόντα της πολύ συχνά λεξικοποιούνται, δηλαδή χρησιμοποιούνται με ίδια ή διαφορετική σημασία της αρχικής.
- Σε αντίθεση με την κλίση που δίνει συνήθως συνεχείς λέξεις, η παραγωγή συχνά δίνει και ασυνεχείς.

3.1.3 Συντακτικό Επίπεδο

Το Συντακτικό ασχολείται με τις φράσεις και τις προτάσεις. Γενικά, όλες οι γραμματικές δίνουν κανόνες φραστικής δομής (phrase structure rules), που δηλώνουν τον τρόπο σύνταξης μίας φράσης. Οι κανόνες αυτοί είναι ανεξάρτητοι των συμφραζόμενων (context free rules). Μέχρι προσφάτως ήταν γενικώς αποδεκτό πως η φυσική γλώσσα περιγράφεται επαρκώς από συμφραστικώς ανεξάρτητους κανόνες. Σήμερα βρίσκονται φαινόμενα της φυσικής γλώσσας που απαιτούν συμφραστικώς εξαρτημένους κανόνες (context sensitive rules) για την παραγωγή ή περιγραφή τους. Οι κανόνες φραστικής δομής χρησιμοποιούνται για να τεχνολογηθούν (parse) ή να συντεθούν (generate) φράσεις και προτάσεις. Διαδικαστικά οι κανόνες φραστικής δομής είναι δυνατόν να ερμηνευτούν ως κανόνες επανεγγραφής (rewrite rules) ή ως περιορισμοί (constraints) στην φραστική δομή.

Η διαδικαστική προσέγγιση δημιουργεί το πρόβλημα της αριστερής αναδρομής, η οποία προκύπτει ερμηνεύοντας διαδικαστικά κανόνες φραστικής δομής, όπου θεωρείται ότι ο κανόνας αυτός παράγει φράσεις. Τέτοιας μορφής κανόνες, όμως, όταν ερμηνεύονται διαδικαστικά, δηλαδή ως κανόνες παραγωγής, οδηγούν σε ατέρμονους βρόχους.

3.1.4 Σημασιολογικό Επίπεδο

Η φυσική γλώσσα έχει τόσο μορφή όσο και σημασία. Η μελέτη της σημασίας της φυσικής γλώσσας παράγει ερωτήματα όπως κατά πόσον η γλώσσα αναφέρεται σε μια αντικειμενική πραγματικότητα, ή σε κάτι που έχει φιλτραριστεί μέσα από τον ανθρώπινο εγκέφαλο, δηλαδή μια υποκειμενική πραγματικότητα. Για την γλωσσολογική μελέτη της γλώσσας, οι επιστήμονες αποδέχονται ότι η γλώσσα έχει αντικειμενική σημασία, η οποία μπορεί να περιγράφει με λογικά συστήματα που στηρίζονται στην αναφορά και στις τιμές αληθείας.

3.1.4.1 Αναφορά και Μοντέλα

Αναφορά είναι η ιδιότητα που έχουν εκφράσεις της φυσικής γλώσσας να εντοπίζουν ένα σύνολο από οντότητες σε έναν συγκεκριμένο κόσμο [1]. Πρωτίστως, ορίζεται ένας συγκεκριμένος κόσμος για την μελέτη σημασιών. Συγκεκριμένα, κατασκευάζεται ένα μαθηματικό αντικείμενο το οποίο ονομάζεται μοντέλο (model) το οποίο απεικονίζει τον κόσμο που βρίσκεται υπό μελέτη. Ένα μοντέλο συνίσταται από το σύνολο των οντοτήτων Ο που περιέχει και από μια συνάρτηση, την ερμηνευτική συνάρτηση (interpretation function) Ε, η οποία απεικονίζει εκφράσεις σε σύνολα.

Οι προτάσεις δεν αναφέρονται σε σύνολα αλλά σε τιμές αληθείας οι οποίες είναι δύο, η Αληθής και η Ψευδής. Η Αληθής τιμή είναι αναφορά των προτάσεων οι οποίες είτε περιγράφουν σχέσεις που υπάρχουν στο μοντέλο ή η τιμή αληθείας τους μπορεί να υπολογισθεί με βάση τους λογικούς συνδέσμους (σύζευξη, άρνηση και τα παράγωγά τους). Αρχικά, βρίσκεται η αναφορά όλων των επιμέρους εκφράσεων που αποτελούν την πρόταση. Στη συνέχεια, οι αναφορές αυτές συνδυάζονται με τον τρόπο που επιβάλει η πρόταση και μελετάται αν η σχέση που προκύπτει δικαιολογείται από το αρχικό μοντέλο. Συνεπώς, η σημασία της πρότασης είναι συνάρτηση των επιμέρους σημασιών. Η ιδιότητα αυτή ονομάζεται συνθετικότητα (compositionality).

Η χρήση μοντέλων για τον υπολογισμό της σημασίας των εκφράσεων της φυσικής γλώσσας προκύπτει από τη σημασιολογία των τυπικών γλωσσών (formal languages) που χρησιμοποιούνται στους υπολογιστές. Το μοντέλο είναι μία δομή δεδομένων που περιέχει οντότητες και συσχετίσεις, όπως και στις Βάσεις Δεδομένων. Η φυσική γλώσσα χρησιμοποιείται για τη διερεύνηση του μοντέλου και την ανάκτηση της κατάλληλης πληροφορίας από αυτό.

3.1.4.2 Κατηγορικός Λογισμός

Η Ερμηνευτική συνάρτηση Ε απεικονίζει εκφράσεις της φυσικής γλώσσας (ονόματα, ρήματα, προτάσεις) σε σύνολα και τιμές αληθείας. Ο συνήθης τρόπος αναπαράστασης της σημασίας της φυσικής γλώσσας πραγματοποιείται με μια ενδιάμεση τεχνητή γλώσσα. Στη συνέχεια, υπολογίζεται η σημασία

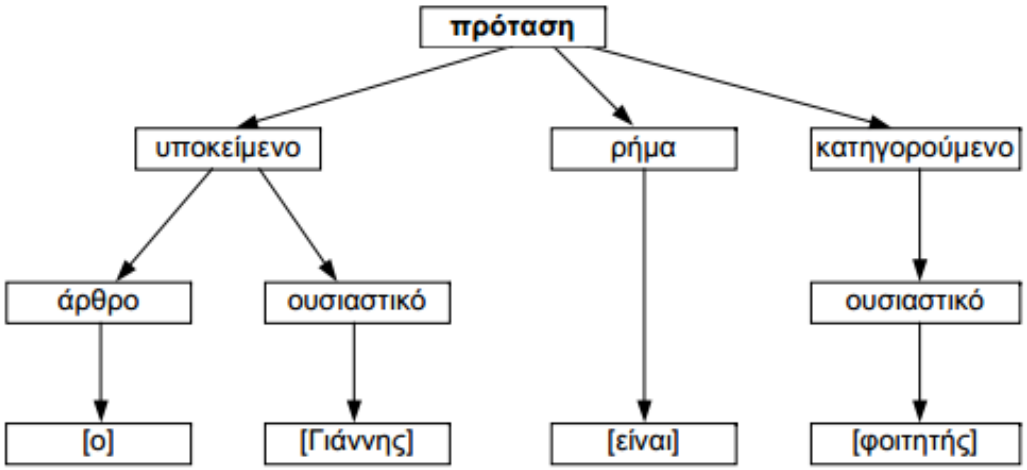
των εκφράσεων της τεχνητής γλώσσας, η οποία επιλέγεται βάσει των κατάλληλων μαθηματικών ιδιοτήτων της. Η ενδιάμεση τεχνητή γλώσσα που χρησιμοποιείται συνήθως είναι ο κατηγορικός λογισμός πρώτης τάξης (first order predicate calculus). Η επιλογή του κατάλληλου μηχανισμού γίνεται με γνώμονα την αξιοπιστία των τελικών συμπερασμάτων, συναρτήσει των δεδομένων που δίδονται ως είσοδος στο μηχανισμό αυτόν.

3.1.5 Πραγματολογικό Επίπεδο

Πέραν του σημασιολογικού επιπέδου, στη φυσική γλώσσα βρίσκονται διάφορες γλωσσικές σημασίες που δεν μπορούν να εκφραστούν μέσω της σημασιολογίας που στηρίζεται στην αναφορά και τις τιμές αληθείας. Τέτοια παραδείγματα είναι οι ερωτήσεις, οι διαταγές, οι παρακλήσεις, και οι ευχές. Οι προτάσεις αυτές δεν έχουν αντιστοιχηθεί σημασιολογικά σε συστήματα που στηρίζονται στην συνθετικότητα και τις τιμές αληθείας. Αυτές οι περιπτώσεις περατώνονται μέσω του Πραγματολογικού Επιπέδου.

3.2 Γραμματικές

Η γραμματική μιας γλώσσας ορίζεται ως το σύνολο των κανόνων που συνθέτουν προτάσεις από μεμονωμένες λέξεις, μαζί με τις όποιες εξαιρέσεις τους. Σε ένα πραγματικό σύστημα υπάρχουν πολλοί περισσότεροι κανόνες, με αποτέλεσμα η αναζήτηση της δομής μίας πρότασης να παρουσιάζει πολύπλοκο πρόβλημα αναζήτησης. Το πρόβλημα αυτό επιλύνεται με τη χρήση του σημασιολογικού και πραγματολογικού επιπέδου ανάλυσης. Κάποια από τα χαρακτηριστικά των γραμματικών βρίσκονται σε αμφισβήτηση αλλά εξακολουθούν να χρησιμοποιούνται καθώς απλοποιούν τον φορμαλισμό χωρίς να μειώνουν την κάλυψη των γλωσσικών δεδομένων. Σε αυτή την ενότητα εξετάζονται τα γραμματικά χαρακτηριστικά της ύπαρξης φραστικών συστατικών, της αναδρομής και της ανεξαρτησίας των συμφραζομένων.



Σχήμα 1 - Συντακτικό Δέντρο ανάλυσης στοιχείων πρότασης [6]

3.2.1 Συστατικά Φράσεων

Κατά την επεξεργασία μίας ακολουθίας συμβόλων ή χαρακτήρων, η μικρότερη δυνατή μονάδα (άτομο) σε αυτή την επεξεργασία είναι το **γράμμα** που αντιστοιχεί σε ένα σύμβολο. Το αμέσως επόμενο επίπεδο ανάλυσης περιλαμβάνει τις **συλλαβές** και τα **μορφήματα** που συγκροτούνται από γράμματα και οι **λέξεις** που αποτελούνται από μορφήματα ή συλλαβές και τελικώς από γράμματα. Η συντακτική δομή μιας **φράσης**, δηλαδή ενός συνόλου από λέξεις, αποτελεί την επόμενη οργανωτική δομή, στη σειρά της ιεραρχίας.

Τα φραστικά συστατικά μίας πρότασης προκύπτουν από την αναγνώριση των συνόλων λέξεων τα οποία αποτελούν λειτουργικά διακριτές ομάδες μέσα στην πρόταση. Για την αναγνώριση των φράσεων χρησιμοποιούνται συγκεκριμένα διαγνωστικά κριτήρια. Στα συστατικά που προκύπτουν από την αναγνώριση αυτή δίνονται διακριτικά ονόματα, όπως Ονοματική Φράση, Ρηματική Φράση κλπ.

3.2.2 Αναδρομή

Στη φυσική γλώσσα παρουσιάζεται το φαινόμενο της αναδρομής, εφόσον θεωρητικά μπορούν να δημιουργηθούν προτάσεις απείρου μήκους. Σε πρακτικό επίπεδο όμως, ο άνθρωπος έχει πεπερασμένες δυνατότητες επεξεργασίας μίας τέτοιας πρότασης. Παρόλα αυτά, τα υπολογιστικά συστήματα είναι ικανά να επεξεργαστούν τέτοιες γραμματικές που περιλαμβάνουν προτάσεις με ακολουθίες συμβόλων απείρου μήκους.

3.2.3 Ανεξαρτησία Συμφραζόμενων

Ως γραμματική ανεξάρτητη συμφραζομένων (context free grammar) είναι μία πλειάδα (tuple) η οποία ορίζεται ως εξής:

$$\mathbf{G = < VT,VN, S, R>}$$

Όπου:
VT : το σύνολο των τερματικών συμβόλων της γλώσσας (πχ ο, η, σκύλος, γάτος, τρώει, κλπ.)
VN : το σύνολο των **μη** τερματικών συμβόλων της γλώσσας
S : το αρχικό σύμβολο της γραμματικής
R : το σύνολο των κανόνων της γλώσσας.

1. Οι κανόνες είναι της μορφής: $A = \psi$ όπου $A \in VN$ και $\psi \in \{VN \cup VT\}^*$
2. Η κενή συμβολοακολουθία e ανήκει στο σύνολο VT ($e \in VT$)
3. Η κενή συμβολοακολουθία χρησιμεύει σε ορισμένες θεωρίες για να αναλύσουν φράσεις όπου υποτίθεται ότι κάποιο συστατικό έχει μεταφερθεί σε άλλο σημείο της φραστικής δομής (α) ή λείπει (β) και έχει αφήσει πίσω του μια κενή υποδοχή.

4 Βάσεις Δεδομένων

Η απαίτηση για αποτελεσματική και αποδοτική διαχείριση της αποθηκευμένης πληροφορίας είναι από τα βασικά χαρακτηριστικά των σύγχρονων εφαρμογών. Στα αρχικά συστήματα, η διαχείριση πληροφορίας πραγματοποιούνταν από τις ίδιες τις εφαρμογές, μέσω των κατάλληλων συστημάτων αρχείων (File System) από το λειτουργικό σύστημα. Ταυτόχρονα, οι μέθοδοι επεξεργασίας και ανάκτησης των δεδομένων αυτών ολοκληρώνονταν από τον εκτελέσιμο κώδικα της εκάστοτε εφαρμογής. Ο παραπάνω τρόπος υλοποίησης ίσως ενδεχομένως να είναι επαρκής σε μικρά συστήματα, ωστόσο παρουσιάζει σοβαρά προβλήματα σε εφαρμογές που φέρουν μεγάλο όγκο πληροφορίας τόσο στην επεξεργασία της πληροφορίας αυτής, όσο και την εύρυθμη διαχείριση και συντήρηση της πληροφορίας αυτής.

Ένα από τα παραδείγματα πολύπλοκων τέτοιων εφαρμογών στον τομέα Επεξεργασίας Φυσικής Γλώσσας, είναι η εφαρμογή οργάνωσης μίας ηλεκτρονικής βιβλιοθήκης ψηφιακών τεκμηρίων. Σε μία τέτοια εφαρμογή απαιτείται η αναζήτηση άρθρων ή βιβλίων βάσει του ονόματος του συγγραφέα, του έτος δημοσίευσης, του τίτλου, με βάση λέξεις-κλειδιά (keywords) που βρίσκονται στο κείμενο και έχουν σημασιολογική αξία για το κείμενο αυτό. Ομοίως, πολλά τέτοια συστήματα παράγουν αυτόματες περιλήψεις των αποθηκευμένων κειμένων, για την εξυπηρέτηση του χρήστη κατά τη διαδικασία εύρεσης των κατάλληλων (ως προς την αναζήτηση του χρήστη) κειμένων. Από την ανάλυση αυτή προκύπτει η πολυπλοκότητα δημιουργίας μίας τέτοιας εφαρμογής, και συγκεκριμένα:

- Το πλήθος των αποθηκευμένων αρχείων είναι μεγάλο και χρειάζονται αποδοτικές και αξιόπιστες μέθοδοι αναζήτηση των αρχείων αυτών.
- Η δομή της πληροφορίας είναι περίπλοκη. Υπάρχουν συγγραφείς με περισσότερα του ενός δημοσιευμένα άρθρα, ενώ παράλληλα ένα άρθρο μπορεί να έχει συνταχθεί από πολλούς συγγραφείς.
- Απαιτείται η δημιουργία ενός αποτελεσματικού και αποδοτικού τρόπου αναζήτησης κειμένων μέσω των λέξεων-κλειδιών. Προφανώς, κάτι τέτοιο δεν μπορεί να πραγματοποιηθεί σειριακά σαρώνοντας κάθε κείμενο που βρίσκεται στο σύστημα, καθότι τα κείμενα είναι μεν πολλά, το δε κάθε κείμενο ενδέχεται να είναι μακροσκελές.
- Δημιουργείται η ανάγκη προσθήκης νέων άρθρων και κειμένων. Το σύστημα θα πρέπει να διασφαλίζει την σωστή καταχώρηση των απαραίτητων στοιχείων που καταχωρούνται, ώστε να εξυπηρετείται και η αναζήτηση του άρθρου αυτού σε επόμενο στάδιο.
- Η εφαρμογή χρησιμοποιείται από μία πληθώρα χρηστών, και πολλές φορές ταυτόχρονα, πράγμα το οποίο πρέπει να προβλέπεται κατά την υλοποίηση ενός τέτοιου συστήματος.

4.1 Συστήματα Βάσεων Δεδομένων

Η προσέγγιση διαχωρισμού των δεδομένων από τις υπόλοιπες μεθόδους επεξεργασίας παρέχει ευελιξία και αποδεσμεύει τους προγραμματιστές των εφαρμογών από την ανάγκη συγχρονισμού των δεδομένων και την υλοποίηση του τρόπου προσπέλασης των δεδομένων αυτών. Η προσέγγιση αυτή στηρίζεται στο λεγόμενο **Σύστημα Βάσης Δεδομένων** το οποίο διαχωρίζει και αναλαμβάνει αποκλειστικά την αποθήκευση, προστασία, ακεραιότητα και επεξεργασία των δεδομένων. Το Σύστημα Βάσεων Δεδομένων παρέχει εξελιγμένους τρόπους πρόσβασης, παροχής δικαιωμάτων και ενημέρωσης των δεδομένων. Τα Συστήματα Βάσεων Δεδομένων αποτελούνται από τα ακόλουθα.

- **Βάση Δεδομένων**

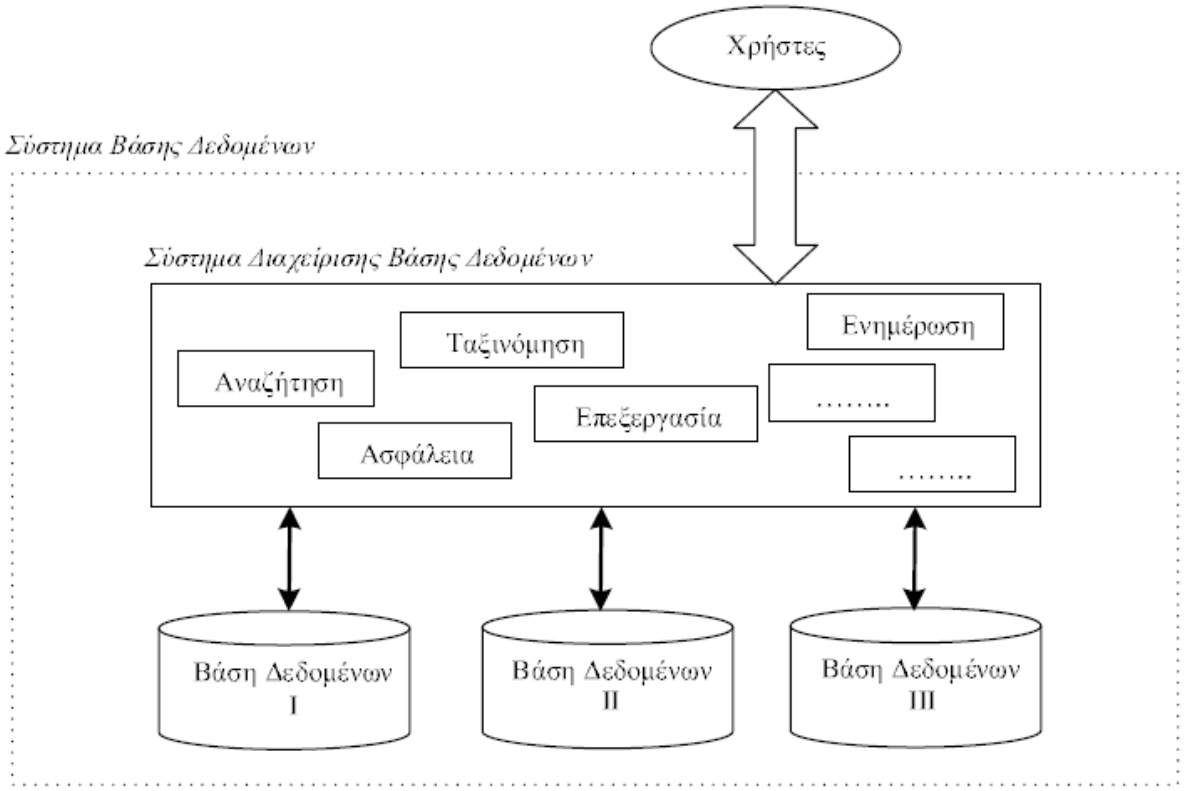
Η **Βάση Δεδομένων (Data Base)** αποτελεί μία συλλογή από στοιχεία τα οποία είναι σχετικά ή όμοια μεταξύ τους, κατάλληλα δομημένα και καταχωρημένα.

- **Σύστημα Διαχείρισης Βάσης Δεδομένων**

Το **Σύστημα Διαχείρισης Βάσης Δεδομένων (Database Management System – DBMS)** είναι το λογισμικό εκείνο σύστημα που υλοποιεί όλες εκείνες τις λειτουργίες που πρέπει να υποστηρίζονται, όπως είναι η αναζήτηση, εισαγωγή, διαγραφή, τροποποίηση, συγχρονισμός, προστασία των δεδομένων κλπ. Ένα Σύστημα Διαχείρισης Βάσης Δεδομένων μπορεί να διαχωρίζεται μία ή περισσότερες Βάσεις Δεδομένων ταυτόχρονα.

Στο επόμενο διάγραμμα παρουσιάζονται οι σχέσεις μεταξύ των επιμέρους Βάσεων Δεδομένων και του καθολικού Συστήματος Διαχείρισης Βάσεων Δεδομένων. Όπως φαίνεται και στην εικόνα, οι

χρήστες δεν έχουν άμεση πρόσβαση στα δεδομένα. Μεταξύ των χρηστών και των Βάσεων Δεδομένων, παρεμβάλλεται το Σύστημα Διαχείρισης της Βάσης Δεδομένων, μέσω του οποίου πραγματοποιείται η προσπέλαση στα δεδομένα.



Σχήμα 2 - Βάση Δεδομένων και Σύστημα Διαχείρισης [7]

4.2 Πλεονεκτήματα χρήσης

Στην ενότητα αυτά αναλύονται τα πλεονεκτήματα χρήσης ενός συστήματος βάσης δεδομένων σε αντίθεση με τις παλαιότερες μεθόδους αρχειοθέτησης (File Systems) [7].

- **Περιγραφή Δεδομένων**

Το Σύστημα Βάσεων Δεδομένων περιέχει τόσο τα ίδια τα δεδομένα όσο και επιπλέον βοηθητικές πληροφορίες για τη περιγραφή των δεδομένων αυτών. Κατά αυτόν τον τρόπο, επιτρέπεται η μεταβολή της δομής και οργάνωσης των δεδομένων σύμφωνα με τις απαιτήσεις των χρηστών. Επομένως, δεν απαιτείται προγραμματιστική επέμβαση στα δεδομένα, καθώς οι μηχανισμοί διαχείρισης δεδομένων υλοποιούνται στο επίπεδο του DBMS.

- **Ανεξαρτησία Δεδομένων και Λειτουργιών**

Τα δεδομένα είναι ανεξάρτητα αποθηκευμένα από τις επιμέρους λειτουργίες που μπορούν να διενεργηθούν σε αυτά. Συνεπώς, η μεταβολή της δομής των δεδομένων δεν απαιτεί μεταβολή στην ίδια την εφαρμογή, χαρίζοντας έτσι ευελιξία στο σύστημα. Ο διαχωρισμός μεταξύ δεδομένων και λειτουργιών διευκολύνει την προσθήκη νέων λειτουργιών, την αποθήκευση των δεδομένων σε διαφορετικά μέσα (συσκευές) και μορφότυπα (format) χωρίς προγραμματιστικές αλλαγές στην εφαρμογή. Τέλος, η εφαρμογή μπορεί να υλοποιηθεί σε οποιαδήποτε γλώσσα προγραμματισμού, αρκεί να επιλεχθεί το κατάλληλο DBMS το οποίο να συνεργάζεται με την επιλεγμένη γλώσσα, χωρίς να απαιτείται οποιαδήποτε τροποποίηση στα δεδομένα της βάσης.

- **Αποδοτική Διαχείριση Δεδομένων**

Η υλοποίηση όλων των λειτουργιών για τα δεδομένα βρίσκεται στο DBMS. Οι αναλυτές και κατασκευαστές του DBMS διαμορφώνουν και υλοποιούν το DBMS ώστε το σύστημα να είναι αποδοτικό και τα ερωτήματα να πραγματοποιούνται με τον καλύτερο δυνατό τρόπο. Ο προγραμματιστής της εκάστοτε εφαρμογής απαλλάσσεται από την ανάλυση, μελέτη και υλοποίηση αποδοτικών αλγόριθμων

και μεθόδων προσπέλασης δεδομένων βάσει της δομής τους, και μπορεί να εστιάσει στην υλοποίηση μόνο των απαραίτητων εκείνων λειτουργιών που αφορούν την συγκεκριμένη εφαρμογή. Η προσπέλαση και παροχή των δεδομένων ανήκει στις αρμοδιότητες του DBMS.

- **Προστασία Δεδομένων και Χρηστών**

Τα DBMS παρέχουν μηχανισμούς προστασίας των δεδομένων με στόχο να αποφεύγονται τυχόν διαγραφές ή τροποποιήσεις των δεδομένων από χρήστες που δεν διαθέτουν τα απαραίτητα δικαιώματα. Το DBMS επιτρέπει τη δυνατότητα αντιστοίχισης διαφορετικών δικαιωμάτων και ρόλων στους χρήστες του, όπως είναι για παράδειγμα το δικαίωμα ανάγνωσης, τροποποίησης, διαγραφής ή εισαγωγής δεδομένων. Ο ρόλος του εκάστοτε χρήστη και τα επιμέρους δικαιώματά του στα δεδομένα και τις λειτουργίες του DBMS μπορούν να μεταβληθούν αναλόγως των αναγκών που παρουσιάζονται.

- **Ταυτόχρονη Προσπέλαση Δεδομένων**

Η δυνατότητα υποστήριξης πολλών χρηστών ταυτόχρονα κρίνεται αναγκαία στις πολύπλοκες εφαρμογές που μελετώνται εδώ. Υπάρχουν περιπτώσεις όπου δύο ή περισσότεροι χρήστες προσπαθούν να προσπελάσουν τα ίδια δεδομένα την ίδια χρονική στιγμή, με σκοπό την τροποποίηση των στοιχείων αυτών. Ομοίως, διακρίνονται περιπτώσεις όπου κάποιος χρήστης επιχειρεί την τροποποίηση του ίδιου στοιχείου που προσπαθεί να ανακτήσει ένας άλλος χρήστης. Προκειμένου να αποφεύγονται τυχόν αστοχίες ή καταστροφή των δεδομένων, το DBMS παρέχει μηχανισμούς που εξασφαλίζουν την ακεραιότητα των δεδομένων που είναι αποθηκευμένα. Συνήθως, εφαρμόζεται κάποια τεχνική κλειδώματος (locking) των εν χρήσει δεδομένων, προκειμένου να μη μπορούν να χρησιμοποιηθούν ή μεταβληθούν από άλλους στην συγκεκριμένη χρονική στιγμή. Οι μηχανισμοί αυτοί προσφέρονται εγγενώς από το DBMS, χωρίς να απαιτείται έλεγχος και συγχρονισμός δεδομένων προγραμματιστικά στην εφαρμογή.

- **Επεκτασιμότητα**

Τα DBMS μπορούν να επεκτείνονται διαρκώς με νέες λειτουργίες και οργανωτικές δομές, επιτρέποντας το σύστημα να προσαρμόζεται συνεχώς με τις όποιες νέες απαιτήσεις μίας εφαρμογής και με τις ανάγκες των χρηστών της. Οι νέες λειτουργίες που κατασκευάζονται αποτελούν μέρος του DBMS, και μπορούν να χρησιμοποιηθούν από όλες τις εφαρμογές που βασίζονται στο DBMS αυτό, χωρίς να απαιτείται περαιτέρω προγραμματισμός στο επίπεδο κάθε εφαρμογής.

4.3 Ανάλυση απαιτήσεων

4.3.1 Το Εννοιολογικό Σχήμα

Το πρώτο βήμα που ακολουθείται κατά τη διαδικασία κατασκευής μίας βάσης δεδομένων αποτελείται από την καταγραφή όλων των λεπτομερειών, των εννοιών και των αντικειμένων που εμφανίζονται, σε συνδυασμό με τις ιδιότητές τους. Ως ιδιότητες ορίζονται οι οντότητες, τα γνωρίσματά τους, ο τρόπος με τον οποίο εμπλέκονται μεταξύ τους, και ο τρόπος λειτουργίας των εφαρμογών στις οποίες εμπίπτει η βάση δεδομένων. Τα παραπάνω στοιχεία πρέπει να αναλυθούν και να καταγραφούν όσο το δυνατό λεπτομερέστερα στο στάδιο αυτό, για να αποφευχθούν τυχόν λάθη και αποκλίσεις στα επόμενα βήματα.

Η παραπάνω διαδικασία, με την οποία καταγράφονται μία προς μία, όλες τις έννοιες ή τα αντικείμενα που θα αποθηκεύονται στη βάση και είναι απαραίτητα στην όποια εφαρμογή, καθώς και τον τρόπο λειτουργίας της, αποκαλείται **εννοιολογική ανάλυση της βάσης δεδομένων.** Η εννοιολογική ανάλυση της βάσης δεδομένων είναι η συστηματική και λεπτομερής μελέτη του πραγματικού κόσμου, όπου πραγματικός κόσμος είναι το πλαίσιο γύρω από το οποίο γίνεται η μελέτη και για το οποίο κατασκευάζεται η βάση δεδομένων. Παραδείγματα τέτοιων κόσμων για την πτυχιακή αυτή θα μπορούσαν να είναι η βιβλιοθήκη που μελετάται προς ψηφιοποίηση, η γλώσσα κλπ.

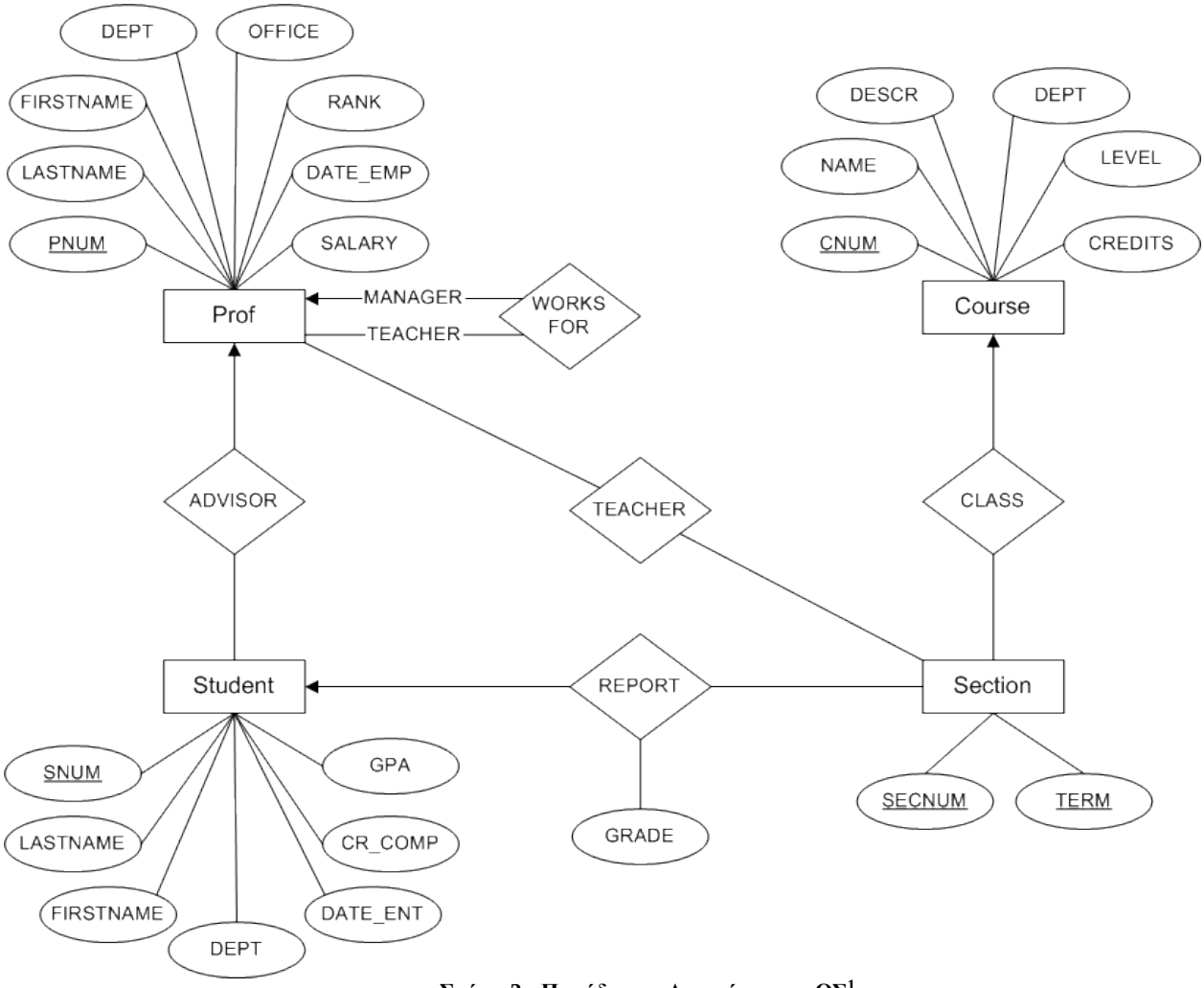
Μετά την ολοκλήρωση του βήματος εννοιολογικής ανάλυσης, στο επόμενο βήμα παραλείπονται οι λεπτομέρειες και εκφράζονται οι οντότητες, οι ιδιότητές τους και οι περιορισμοί ακεραιότητας που επιβάλλονται πάνω στις οντότητες, με απλό και εύχρηστο τρόπο, κατανοητό από τους χρήστες της βάσης. Η καταγραφή των συνόλων οντοτήτων με τις ιδιότητες τους και τους περιορισμούς ακεραιότητας, σε μορφή κατανοητή από τους χρήστες της βάσης, η οποία προκύπτει από την εννοιολογική ανάλυση, ονομάζεται **Εννοιολογική Βάση Δεδομένων.** Η Εννοιολογική Βάση Δεδομένων περιλαμβάνει την αφαίρεση (από πλευράς των χρηστών) του πραγματικού κόσμου, που περιγράφεται στην εννοιολογική ανάλυση. Οι πληροφορίες για τις οντότητες, τις ιδιότητές τους και τον τρόπο σύνδεσής τους που έχουν

καταγραφεί στην Εννοιολογική βάση δεδομένων, κωδικοποιούνται με κατάλληλα σύμβολα και εκφράζονται όλες μαζί μέσω ενός σχεδίου, το οποίο ονομάζεται **Εννοιολογικό Σχήμα (Conceptual Design)** της βάσης.

Οι συμβολισμοί που χρησιμοποιούνται για το σχέδιο της εννοιολογικής βάσης παρέχονται από ένα μοντέλο δεδομένων υψηλού επιπέδου, πολύ κοντά στη φυσική γλώσσα του χρήστη, που είναι γνωστό ως Μοντέλο Οντοτήτων-Συσχετίσεων, και περιγράφεται στην επόμενη υπο-ενότητα. Οι περιορισμοί ακεραιότητας που επιβάλλονται στα δεδομένα πρέπει να εμφανίζονται στην Εννοιολογική Βάση Δεδομένων και όχι στο Εννοιολογικό Σχήμα.

4.3.2 Μοντέλο Οντοτήτων – Συσχετίσεων

Το **Μοντέλο Οντοτήτων–Συσχετίσεων (ΟΣ*)*** βασίζεται στη θεωρία ότι ο πραγματικός κόσμος αποτελείται από διαφορετικές **οντότητες**, οι οποίες διαθέτουν συγκεκριμένα **χαρακτηριστικά ή ιδιότητες**, ενώ συνδέονται μεταξύ τους με **συσχετίσεις**. Το μοντέλο ΟΣ προτάθηκε και αναπτύχθηκε από τον Chen το 1976 [8] και είναι το επικρατέστερο μοντέλο ανάλυσης και σχεδίασης Βάσεων Δεδομένων. Το μοντέλο αυτό γεφυρώνει το χάσμα μεταξύ των σχεδιαστών της βάσης και των χρηστών. Ο συνήθης τρόπος σχεδίασης του μοντέλου ΟΣ είναι με τη χρήση του λεγόμενου **διαγράμματος Οντοτήτων-Συσχετίσεων (Entity-Relationship Diagram - ER)**. Στο επόμενο σχήμα δίνεται ένα τέτοιο παράδειγμα διαγράμματος ER.



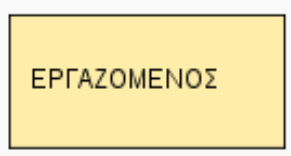
Σχήμα 3 - Παράδειγμα Διαγράμματος ΟΣ[1]

Αναλυτικά, τα επιμέρους στοιχεία που συνθέτουν το διάγραμμα ER είναι:

[1] Πηγή: http://163.238.35.144/~chi/MySQL/Univ12/University%20ER%20Diagram.gif

4.3.2.1 Οντότητα

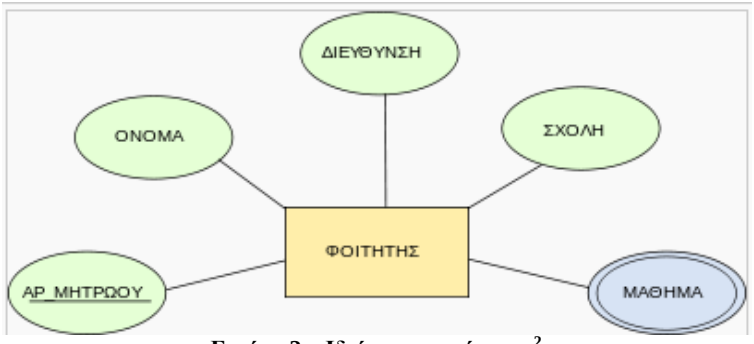
Ως **οντότητα (entity)** ορίζεται κάθε αντικείμενο του πραγματικού κόσμου που μπορεί να διαχωριστεί από άλλα αντικείμενα. Πρόκειται για οποιοδήποτε αντικείμενο με ανεξάρτητη υπόσταση από το οποίο μπορεί να εξαχθεί πληροφορία [9]. Ένα **σύνολο οντοτήτων (entity set)** είναι μία συλλογή από ίδιου τύπου οντότητες με ίδιες ιδιότητες ή χαρακτηριστικά. Παραδείγματα οντοτήτων στον «πραγματικό κόσμο» μίας εταιρίας θα μπορούσαν να είναι: *Εργαζόμενος, Εργοδότης, Μισθός, Τμήμα* κλπ. Μια βάση δεδομένων συνήθως περιέχει πολλές διαφορετικές οντότητες, οι οποίες συσχετίζονται μεταξύ τους, ενώ σχεδιάζονται με ορθογώνια παραλληλόγραμμα, όπως στο επόμενο παράδειγμα.



Εικόνα 1 – Οντότητα[2]

4.3.2.2 Ιδιότητα ή Χαρακτηριστικό

Οι οντότητες μπορούν να έχουν μία ή περισσότερες **ιδιότητες ή χαρακτηριστικά (attributes).** Οι τιμές των ιδιοτήτων προσδιορίζουν με μοναδικό τρόπο μία οντότητα του συνόλου, και βοηθούν στον διαχωρισμό της μίας οντότητας από την άλλη. Κάθε οντότητα περιγράφεται από το όνομά της, τον τύπο της, και από τις ιδιότητες που περιέχει. Οι ιδιότητες σχεδιάζονται με τη μορφή της έλλειψης, η οποία συνδέεται με το ορθογώνιο παραλληλόγραμμο της οντότητας. Η επόμενη εικόνα παραθέτει ένα τέτοιο παράδειγμα, όπου η οντότητα *Φοιτητής* χαρακτηρίζεται από τις ιδιότητες *Διεύθυνση, Σχολή, Όνομα, ΑΜ*, και το *σύνολο των μαθημάτων* που παρακολουθεί.



Εικόνα 2 – Ιδιότητες οντότητας[2]

4.3.2.3 Συσχέτιση

Συσχέτιση (relationship) είναι η σύνδεση δύο ή περισσοτέρων τύπων οντοτήτων, που παρουσιάζουν κάποιο σχεδιαστικό και λογικό νόημα. Ταυτόχρονα, συσχετίσεις μπορούν να συνδέουν και τις ιδιότητες των οντοτήτων. Σχεδιαστικά, μία συσχέτιση παριστάνεται με έναν ρόμβο που συνδέεται μεταξύ των στοιχείων που αναφέρεται. Οι συσχετίσεις είναι απαραίτητες κατά το σχεδιασμό για την άντληση της σχετικής πληροφορίας που αφορά τις οντότητες ή τις ιδιότητες που σχετίζονται. Η λειτουργία που επιτελεί μία οντότητα σε μία συσχέτιση ονομάζεται **ρόλος (role).** Ο δε **βαθμός (degree)** μιας συσχέτισης ισούται με το πλήθος των οντοτήτων που συμμετέχουν σ' αυτήν. Για την σωστή σχεδίαση μιας βάσης δεδομένων, θα πρέπει να αντιστοιχούνται ιδιότητες μόνο σε οντότητες και στη συνέχεια να παράγονται συσχετίσεις μεταξύ των οντοτήτων αυτών, όχι μέσω των ιδιοτήτων τους.

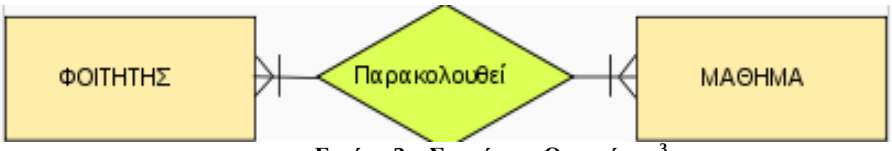
4.3.2.4 Πληθικότητες συσχετίσεων

Τα τρία βασικά ήδη διμελών συσχετίσεων, βάσει του **λόγου πληθικότητας (cardinality ratio)** τους, μεταξύ οντοτήτων είναι:

- **Ένα-προς-ένα (1:1).** Σε αυτή τη περίπτωση μια οντότητα συνδέεται ακριβώς μόνο με μία άλλη οντότητα.

[2] Πηγή: http://el.wikipedia.org/wiki/Μοντέλο_Οντοτήτων-Συσχετίσεων

- **Ένα-προς-πολλά (1:M)**. Στη περίπτωση αυτή τα στοιχεία της μίας οντότητα συνδέονται με πολλά στοιχεία της άλλης οντότητας, αλλά κάθε στοιχείο της δεύτερης οντότητας αντιστοιχεί σε ένα και μόνο ένα στοιχείο της πρώτης. Η περίπτωση αυτή συναντάται πιο συχνά από τις υπόλοιπες στις βάσεις δεδομένων.
- **Πολλά-προς-πολλά (M:N).** Σε αυτή τη περίπτωση οι εγγραφές της μίας οντότητας αντιστοιχούν σε πολλές εγγραφές της άλλης οντότητες, και το ανάποδο (οι εγγραφές της δεύτερης οντότητας αντιστοιχούν σε πολλές εγγραφές της πρώτης). Σε αυτές τις περιπτώσεις, δημιουργούνται ενδιάμεσοι πίνακες-δείκτες για να αντιστοιχούν τις εγγραφές μεταξύ των εγγραφών αυτών.



Εικόνα 3 – Συσχέτιση Οντοτήτων[3]

4.3.2.5 Κλειδιά

Η ιδιότητα που προσδιορίζει μοναδικά μία οντότητα ονομάζεται **κλειδί (key)**, και έχει ένα πεδίο ορισμού (domain). Μπορεί περισσότερα από ένα κλειδιά να έχουν την ιδιότητα του κλειδιού. Όλα αυτά τα πιθανά κλειδιά ονομάζονται **υποψήφια κλειδιά (candidate keys)**. Κατά τη διάρκεια σχεδιασμού και ανάλυσης της βάσης δεδομένων, επιλέγεται ένα από αυτά τα κλειδιά ως **πρωτεύον κλειδί (primary key)** και τα υπόλοιπα υποψήφια κλειδιά χαρακτηρίζονται ως **εναλλακτικά κλειδιά (alternative keys)**. Το επιλεγμένο πρωτεύον κλειδί σχεδιάζεται στο διάγραμμα ΟΣ ως υπογραμμισμένο (στην έλλειψη της ιδιότητας). Στις περιπτώσεις όπου δεν μπορεί να επιλεχθεί ένα μοναδικό κλειδί ως πρωτεύον, καθώς ενδεχομένως να μην μπορεί να προσδιορίζει μοναδικά μία οντότητα, τότε ο σχεδιαστής επιλέγει δύο ή περισσότερες ιδιότητες ως κλειδιά ταυτόχρονα, οι οποίες συνδυαστικά προσδιορίζουν την οντότητα. Το συνδυαστικό αυτό κλειδί ονομάζεται **σύνθετο κλειδί (composite key)**.

Το Μοντέλο ΟΣ περιγράφει με λογικό τρόπο τα στοιχεία που θα συμμετέχουν σε μια βάση δεδομένων, με αφαιρετικό τρόπο, χωρίς να καθορίζει μέσω της σχεδίασής του τις λεπτομέρειες που χρειάζονται για την υλοποίηση της ίδιας της βάσης. Συνοπτικά, το διάγραμμα ΟΣ κατασκευάζεται με τα εξής βήματα :

1) Ορίζονται οι οντότητες που είναι μοναδικές και προκύπτουν από τον «πραγματικό κόσμο» που μελετάται.
2) Καταγράφονται τα σύνολα των ιδιοτήτων ή χαρακτηριστικών για κάθε δεδομένη οντότητα.
3) Βρίσκονται τα υποψήφια κλειδιά κάθε οντότητας, και επιλέγεται το πρωτεύον κλειδί.
4) Σχεδιάζονται οι συσχετίσεις που προκύπτουν μεταξύ των οντοτήτων, και ο λόγος πληθικότητάς τους.

[3] Πηγή: http://el.wikipedia.org/wiki/Μοντέλο_Οντοτήτων-Συσχετίσεων

ΜΕΡΟΣ Β

ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΕΡΓΑΛΕΙΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

5 Συστήματα Επεξεργασίας Φυσικής Γλώσσας

Τα συστήματα Επεξεργασίας Φυσικής Γλώσσας έχουν περιορισμένο υπόβαθρο φυσικής γλώσσας, με αποτέλεσμα να απαιτείται ορισμένη εκπαίδευση για την χρήση τους. Υπό περιπτώσεις, ο χρήστης δυσκολεύεται περισσότερο να κατανοήσει τους περιορισμούς που προκύπτουν στην Φυσική Γλώσσα. Αντίθετα, ο ίδιος χρήστης βρίσκει πιο εύκολη την εκμάθηση και χρήση ενός ολοκληρωμένου συστήματος που διαθέτει γραφικό περιβάλλον, για να εξάγει πληροφορία από δεδομένα, μέσω μηχανισμών Επεξεργασίας Φυσικής Γλώσσας.

Στο σημείο αυτό, αξίζουν να αναφερθούν δύο (2) από τα πιο αξιοσημείωτα πρότυπα συστήματα Επεξεργασίας Φυσικής Γλώσσας. Το πρώτο ονομάζεται ELIZA [10], το οποίο είναι ένα εξαιρετικά απλό υπολογιστικό πρόγραμμα που εστιάζει στην επικοινωνία μέσω φυσικής γλώσσας. Αναλαμβάνει το ρόλο ενός Ρογεριανού ψυχαναλυτή [11], όπου βασικά επαναλαμβάνει οτιδήποτε πει ο χρήστης σε μορφή ερώτησης. Η ψυχιατρική κοινότητα δέχτηκε με ενθουσιασμό τη συγκεκριμένη εφαρμογή και πρότεινε να χρησιμοποιηθεί ως ένα από τα εργαλεία κατά τη θεραπεία των ασθενών. Το δεύτερο αξιοσημείωτο έργο κατασκευάστηκε από την NASA, ονόματι LUNAR [12]. Το πρότυπο LUNAR επέτρεπε σε γεωλόγους να κάνουν ερωτήσεις για την χημική σύσταση σεληνιακών πετρωμάτων και δειγμάτων του εδάφους. Για παράδειγμα, ένας γεωλόγος μπορούσε να υποβάλει στο σύστημα ερωτήσεις, όπως: «Ποια είναι η μέση συγκέντρωση πλαγιοκλασικής σύστασης σε σεληνιακά δείγματα που περιέχουν ρουβίδιο;». Αν και το σύστημα δεν χρησιμοποιήθηκε σε πραγματικές συνθήκες, παρά μόνο στο πιλοτικό του στάδιο, τα αποτελέσματα της μελέτης έδειξαν ποσοστά επιτυχίας της τάξης του 78%.

Στην ενότητα αυτή παρουσιάζονται σύγχρονες έρευνες, βάσει των βιβλιογραφικών τους αποτελεσμάτων, που υλοποιούν συστήματα ή εργαλεία Επεξεργασίας Φυσικής Γλώσσας, τα οποία βασίζονται σε κάποιο Σύστημα Βάσεων Δεδομένων.

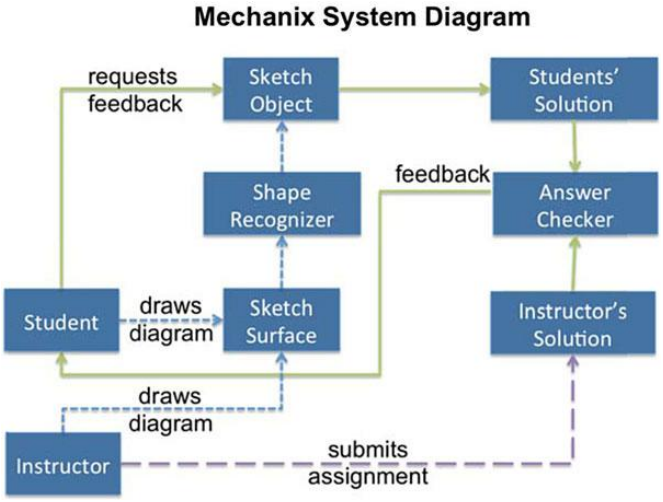
5.1 Mechanix

Στο [13], παρουσιάζεται ένα καινοτόμο βοηθητικό εργαλείο εκμάθησης στατικής μηχανικής, ονόματι Mechanix. Όπως αναφέρουν οι συγγραφείς, το Mechanix αξιοποιεί νέες τεχνολογίες τεχνητής νοημοσύνης για την εκπαίδευση των χρηστών στον κλάδο της στατικής μηχανικής, παρέχοντας άμεση ανάδραση προς το χρήστη και αυτοματοποιημένη διόρθωση αποτελεσμάτων. Πιο συγκεκριμένα, το Mechanix είναι ένα εργαλείο σκιαγράφησης στατικών στηριγμάτων τα οποία συνθέτουν μοντέλα στατικής μηχανικής (π.χ. οδικές γέφυρες), εκπαιδεύοντας τους μαθητές μηχανικής στις έννοιες των δοκών, στηριγμάτων και διαγράμματα ελευθέρων σωμάτων (FBD).

Συγκεκριμένα, η εφαρμογή του Mechanix, λειτουργεί ως εξής:

Οι μαθητές μπορούν να σκιαγραφήσουν ένα στήριγμα FBD απευθείας είτε μέσω συσκευής tablet και βοηθητικής έξυπνης πένας (smart pen), είτε μέσω του ποντικιού και της οθόνης του υπολογιστή. Το σύστημα αναγνωρίζει το σωστά σκιαγραφημένο σχέδιο FBD, τοποθετεί τις κατάλληλες ετικέτες στους κόμβους μεταξύ των δοκών, και παρέχει την κατάλληλη πληροφορία – ανάδραση στο μαθητή, εφόσον τη ζητήσει. Παράλληλα, το εργαλείο βαθμολογεί το πρόβλημα που έχει τεθεί. Αντίστοιχα, οι εκπαιδευτές μπορούν να δημιουργούν προβλήματα προς διερεύνηση από τους μαθητές και να εισάγουν τις κατάλληλες απαντήσεις στα ερωτήματα αυτά.

Στην παρακάτω εικόνα, απεικονίζεται ένα απλουστευμένο διάγραμμα των διεργασιών που χρησιμοποιούνται από το σύστημα για την αναγνώριση των σχημάτων, την παροχή αλληλεπίδρασης και διάδρασης με το μαθητή, και την εισαγωγή των προβλημάτων και λύσεων από τους εκπαιδευτές. Όταν ο μαθητής ή εκπαιδευτής σχεδιάσει ένα διάγραμμα στην επιφάνεια σχεδιασμού, κάθε σχεδιασμένη ακμή αποστέλλεται απευθείας στους ανιχνευτές γεωμετρικών σχημάτων. Το Mechanix προσθέτει τα αντικείμενα σχημάτων που αναγνώρισε στο αντικείμενο περιεχομένου σκίτσων (Sketch container). Όταν ο μαθητής ζητάει τη γνώμη του συστήματος για το διάγραμμα που σχεδίασε, το Mechanix μετασχηματίζει το σκίτσο σε ένα αντικείμενο λύσης (Solution object) που αποτελεί το FBD. Στη συνέχεια, το Mechanix αποστέλλει το FBD του μαθητή στο τμήμα ελέγχου λύσεων, το οποίο με τη σειρά του συγκρίνει τη λύση που δόθηκε από τον μαθητή σε συγκρίσει με την λύση που έχει καταχωρηθεί από τον διδάσκοντα. Η τελική σύγκριση αποστέλλεται ως απάντηση από το τμήμα ελέγχου λύσεων στο μαθητή, ώστε να ελέγξει τα αποτελέσματα του.



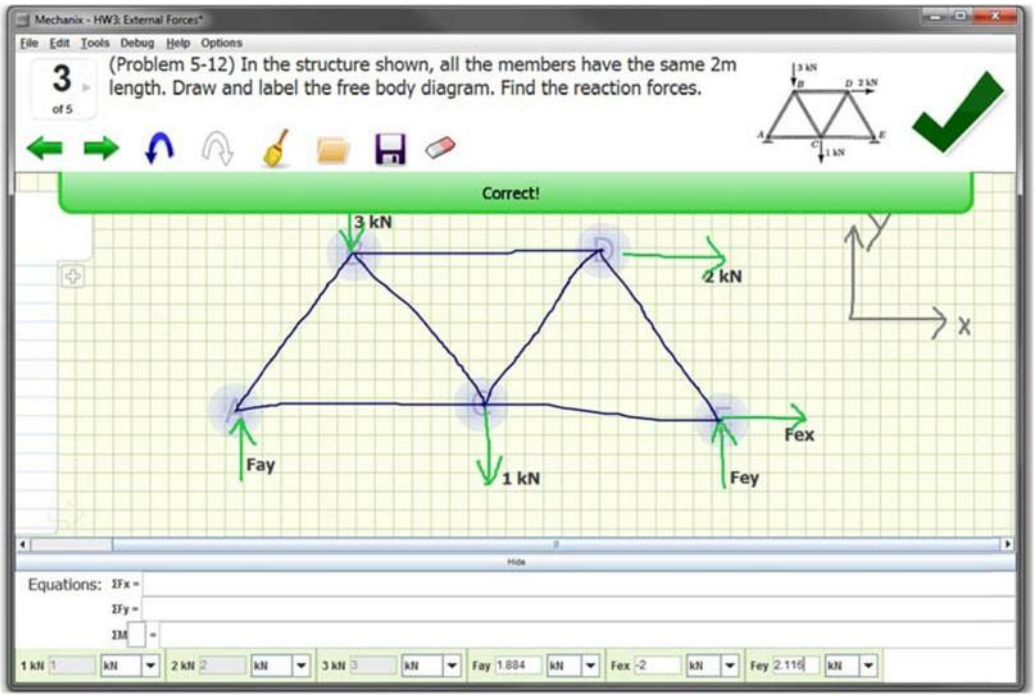
Εικόνα 4 – Διάγραμμα Συστήματος Mechanix [13]

Αυτός ο νέος τρόπος εκμάθησης έχει ως αποτέλεσμα την άμεση εκπαίδευση των μαθητών, ενώ παράλληλα οι μαθητές δείχνουν περισσότερο ενδιαφέρον και ενθουσιασμό για το αντικείμενο της στατικής.

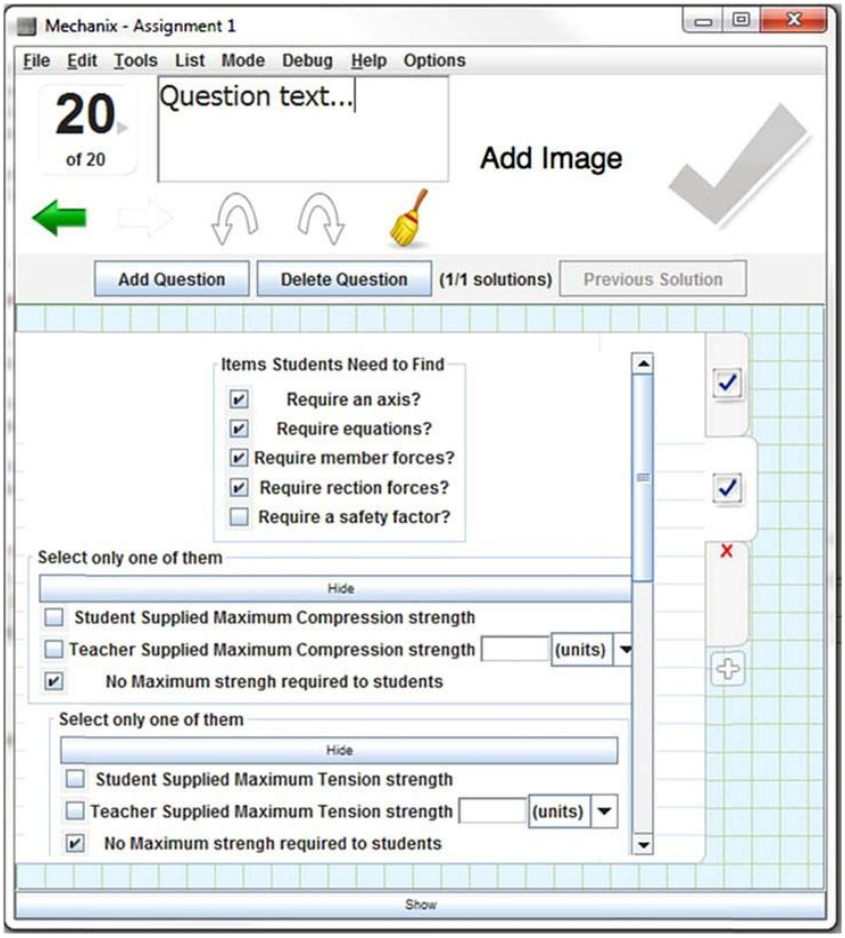
Το κύριο πλεονέκτημα του Mechanix, έναντι των προηγουμένων διαθέσιμων προγραμμάτων, είναι η άμεση και περιγραφική διάδραση που παρέχει στους μαθητές. Επιπλέον, το πρόγραμμα επιτρέπει στους μαθητές να σκιαγραφούν τα διαγράμματα ελευθέρων σωμάτων (FBD) με φυσικό τρόπο, όπως στο χαρτί. Παράλληλα, το Mechanix βαθμολογεί αυτόματα τις απαντήσεις των μαθητών στα προβλήματα που τίθενται, μειώνοντας έτσι τον απαιτούμενο χρόνο ενασχόλησης των καθηγητών και των εκπαιδευτικών βοηθών. Αυτή η λειτουργία του συστήματος είναι ιδιαίτερα βοηθητική σε περιπτώσεις όπου το πλήθος των μαθητών είναι μεγάλο και στις περιπτώσεις των ολοένα και δημοφιλέστερων online διαδραστικών μαθημάτων (ηλεκτρονική εκπαίδευση – e-learning).

Όπως αναφέρουν οι συγγραφείς, υπάρχει μία πληθώρα υπάρχοντων εφαρμογών για την εκπαίδευση μαθητών σε θέματα ανάλυσης δοκών / στηριγμάτων, FBDs, καθώς και άλλα προβλήματα στατικής μηχανικής. Σε αυτά συμπεριλαμβάνονται τα WinTruss [14], το Open Learning Initiative από το πανεπιστήμιο Carnegie Mellon [15], το σύστημα εκμάθησης φυσικής Andes [16], το VaNTH ERC Free-Body Diagram Assistant [17], InTEL [18], Newton's Pen [19], Interactive Physics [20], και το M-MODEL8 [21]. Τα συστήματα αυτά βοηθούν τους μαθητές στην επίλυση προβλημάτων στατικής μηχανικής και καθοδηγούν τους μαθητές στα βήματα επίλυσης τους. Όμως, κανένα από αυτά δεν δίνει τη δυνατότητα στους μαθητές να επιλύσουν το ζητούμενο πρόβλημα μόνοι τους. Αντιθέτως, δίνουν τη μερική λύση εξ αρχής στους μαθητές και ζητούν από αυτούς να διευκρινίσουν τις υπόλοιπες τιμές, την κατεύθυνση των δυνάμεων, και το κρίσιμο σημείο. Παράλληλα, παρέχουν πληροφορίες για την ορθή ή μη απάντηση του μαθητή στα σημεία που ζητούνται. Αντίθετα, κανένα από τα παραπάνω δεν αξιολογεί το σκίτσο FBD που σχεδίασε ο μαθητής.

Τα πειραματικά αποτελέσματα από τη σύγκριση που διενέργησαν οι συγγραφείς μεταξύ του Mechanix και του WinTruss σε ευρεία γκάμα μαθητών, έδειξαν ότι οι μαθητές πέτυχαν μεγαλύτερες βαθμολογίες έχοντας εκπαιδευτεί με το Mechanix σε θέματα στατικής. Οι μαθητές έδειξαν αρκετά ικανοποιημένοι από την άμεση ανταπόκριση του συστήματος κατά την επίλυση του εκάστοτε προβλήματος που επιχείρησαν, όπως επίσης από την ευκολία χρήσης του γραφικού περιβάλλοντος του Mechanix.



Εικόνα 5 – Γραφικό περιβάλλον επίλυσης προβλήματος στο Mechanix (μαθητής) [13]



Εικόνα 6 – Γραφικό περιβάλλον εισαγωγής νέου προβλήματος στο Mechanix (εκπαιδευτής) [13]

5.2 SIMBA

Η έρευνα στο [22] παρουσιάζει το SIMBA, ένα σύστημα εξαγωγής περιλήψεων πολλαπλών εγγράφων για την Πορτογαλική γλώσσα. Το σύστημα αυτό λαμβάνει ένα σύνολο Πορτογαλικών κειμένων από οποιοδήποτε τομέα, και παράγει ενημερωτικές περιλήψεις για ένα γενικό ακροατήριο. Το μήκος των περιλήψεων καθορίζεται από ένα ρυθμό συμπίεσης που ορίζεται από το χρήστη. Η περίληψη πραγματοποιείται κάνοντας χρήση μίας ρηχής πλην όμως αποτελεσματικής μεθόδου που βασίζεται στον υπολογισμό στατιστικών χαρακτηριστικών από τα στοιχεία του κειμένου. Η μέθοδος που εφαρμόζεται για την εξαγωγή των περιλήψεων περιλαμβάνει μία προσέγγιση ομαδοποίησης δύο-φάσεων. Κατά την πρώτη φάση ομαδοποιούνται οι προτάσεις βάσει της ομοιότητάς τους και επιλέγεται μία από αυτές τις προτάσεις κάθε ομάδας, για την μείωση του πλεονασμού της ίδιας πληροφορίας. Στη συνέχεια, οι προτάσεις αυτές ομαδοποιούνται βάσει μίας συλλογής από λέξεις-κλειδιά, ώστε να κατηγοριοποιηθούν σε θεματικές ενότητες. Η διαδικασία περίληψης περιλαμβάνει ένα ακόμα στάδιο απλοποίησης της πρότασης, κατά το οποίο δημιουργούνται πιο απλές και ακριβείς προτάσεις, ενώ ταυτόχρονα επιτρέπει στην εισαγωγή όσο το δυνατόν περισσότερο σχετικού περιεχομένου.

Η σύνοψη πραγματοποιείται με την διαδοχική εκτέλεση πέντε διακριτών σταδίων: το στάδιο αναγνώρισης, το στάδιο ταυτοποίησης, το στάδιο φιλτραρίσματος, το στάδιο περιορισμού αποτελεσμάτων και το τελικό στάδιο παρουσίασης. Τα πέντε αυτά στάδια κατατάσσονται σε τρεις γενικές φάσεις της διαδικασίας περίληψης. Η φάση ανάλυσης περιλαμβάνει το στάδιο αναγνώρισης, η φάση μετασχηματισμού περιλαμβάνει τα στάδια ταυτοποίησης και φιλτραρίσματος, και τέλος η φάση σύνθεσης περιλαμβάνει τα στάδια περιορισμού αποτελεσμάτων και παρουσίασης.

5.2.1 Στάδιο Αναγνώρισης

Το στάδιο αναγνώρισης εκτελείται σε δύο μέρη. Το πρώτο μέρος χειρίζεται τα έγγραφα που υποβάλλονται στο σύστημα από το χρήστη, τα μετατρέπει στο ίδιο μορφότυπο (format) και αφαιρεί τον υπάρχον θόρυβο. Κατά το δεύτερο μέρος, τα κείμενα αυτά σχολιάζονται σημασιολογικά, αναγνωρίζονται οι επιμέρους προτάσεις και παράγραφοι που συνθέτουν το κείμενο και τοποθετούνται σημασιολογικές ετικέτες σε συγκεκριμένες λέξεις. Επίσης, δημιουργείται ένας γράφος ανάλυσης του κειμένου που προσδιορίζει τη συντακτική δομή των προτάσεων. Στο τέλος του σταδίου αυτού εξάγονται πλέον τα κατάλληλα σύνολα προτάσεων που προκύπτουν από το συνολικό κείμενο.

5.2.2 Στάδιο Ταυτοποίησης

Το στάδιο αυτό στοχεύει στον προσδιορισμό της σχετικής πληροφορίας από το σύνολο των κειμένων. Αρχικά, βαθμονομούνται οι προτάσεις που έχει εξάγει το στάδιο Αναγνώρισης. Εν συνεχεία, οι προτάσεις ομαδοποιούνται βάσει ομοιότητας ώστε να αφαιρεθεί ο πλεονασμός από το σύνολο των στοιχείων. Τέλος, οι προτάσεις αυτές ομαδοποιούνται εκ νέου βάσει των λέξεων-κλειδιών ώστε να προσδιοριστούν εκείνες με την πιο σχετική πληροφορία. Για την βαθμολόγηση της ομοιότητας, λαμβάνονται υπόψη τρεις επιμέρους βαθμολογίες: η κύρια βαθμολογία, η επιπλέον βαθμολογία και η συνολική βαθμολογία. Η κύρια βαθμολογία αντιστοιχεί στη σχετικότητα της εκάστοτε πρότασης ως προς το γενικότερο σύνολο των προτάσεων. Η επιπλέον βαθμολογία χρησιμοποιείται κατά τη διαδικασία εξαγωγής περίληψης για την επιβράβευση ή μη της πρότασης βάσει της σχετικότητάς της, προσθαφαιρώντας προκαθορισμένες τιμές βαθμολόγησης. Η συνολική βαθμολογία προκύπτει από το άθροισμα της κύριας και της επιπλέον βαθμολογίας.

5.2.3 Στάδιο Φιλτραρίσματος

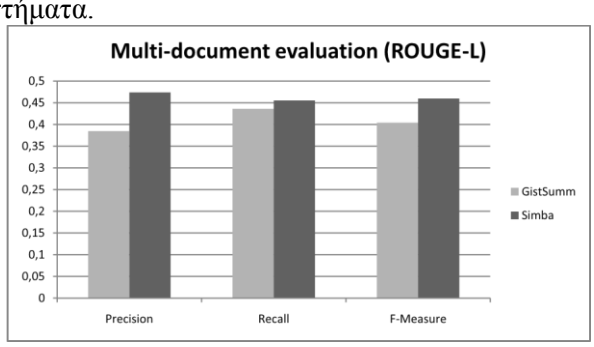
Στο στάδιο αυτό λαμβάνονται υπόψη οι προτάσεις εκείνες που έχουν ομαδοποιηθεί βάσει των λέξεων-κλειδιών. Οι προτάσεις που συντάσσονται με λιγότερες από δέκα (10) λέξεις επιδέχονται επιπλέον κύρωσης, αφαιρώντας μία προκαθορισμένη τιμή από την επιπλέον βαθμολόγησή τους. Αντίστοιχα, προτάσεις που συντάσσονται με πάνω από δέκα λέξεις, βραβεύονται αυξάνοντας την επιπλέον βαθμολόγησή τους. Η τελική βαθμολογία χρησιμοποιείται για την ταξινόμηση των προτάσεων, ορίζοντας έτσι τη σειρά με την οποία θα επιλεχθούν οι προτάσεις της τελικής σύνοψης.

5.2.4 Στάδιο Περιορισμού

Στο τελικό στάδιο περιορισμού αποτελεσμάτων στοχεύει στη μείωση του αρχικού περιεχομένου, ώστε να παραχθεί η τελική περίληψη του κειμένου με πιο απλές και ενημερωτικές προτάσεις. Το στάδιο

αυτό αποτελείται από δύο επιμέρους μέρη, εκείνο της απλοποίησης και εκείνο της συμπίεσης. Η απλοποίηση προτάσεων πραγματοποιείται αφαιρώντας εκφράσεις ή φράσεις των οποίων η αφαίρεση δεν επηρεάζει το νόημα του κειμένου. Στη συνέχεια εφαρμόζεται στο σύνολο των προτάσεων ο ρυθμός συμπίεσης, που ορίστηκε είτε από τον χρήστη κατά την είσοδο των κειμένων είτε αυτόματα από το σύστημα στο 70%. Οι προτάσεις αυτές προστίθενται στην τελική περίληψη βάσει του συνολικού αριθμού λέξεων. Αν ο συνολικός αριθμός λέξεων των προτάσεων που έχουν ήδη προστεθεί στην τελική περίληψη ξεπερνούν το συνολικό αριθμό λέξεων που ορίζει ο ρυθμός συμπίεσης, τότε δεν προστίθενται άλλες προτάσεις και παράγεται έτσι η τελική περίληψη. Τέλος, η σύνοψη που προκύπτει διανέμεται στο χρήστη μέσω αρχείου κειμένου.

Από την αξιολόγηση του συστήματος προκύπτουν ενθαρρυντικά αποτελέσματα, καθώς ο συνδυασμός ομαδοποίησης κατά σχετικότητα και κατά λέξεις-κλειδιά φαίνεται να παράγει πολύ κατανοητές περιλήψεις. Οι παραγόμενες συνόψεις καταφέρνουν να διατηρήσουν το αρχικό νόημα των κειμένων, ενώ ταυτόχρονα καλύπτουν τα πιο σημαντικά θέματα που αναφέρονται στη συλλογή των κειμένων. Οι περιλήψεις συμπιεστεί κατά 85%, δηλαδή περιλαμβάνουν το 85% των λέξεων που αναφέρονται στο σύνολο των κειμένων, ρυθμός ο οποίος θεωρείται ως ο μέσος ρυθμός συμπίεσης σε ιδανικά κατασκευασμένες περιλήψεις. Ο ρυθμός ακρίβειας (precision) είναι αρκετά υψηλός, το οποίο σημαίνει ότι εξάγεται η πιο σχετική πληροφορία από το σύνολο των κειμένων. Αντίθετα, ο ρυθμός ανάκλησης (recall) δεν είναι τόσο υψηλός, καθώς πραγματοποιείται η διαδικασία απλοποίησης των προτάσεων από το συγκεκριμένο σύστημα, επομένως είναι δύσκολο να επιτευχθούν τα ίδια αποτελέσματα με εκείνα μίας ιδανικής περίληψης. Όμως, λαμβάνοντας υπόψη τις τιμές αξιολόγησης αυτές σε συγκρίσει με την αξιολόγηση άλλων παρόμοιων συστημάτων, συμπεραίνεται ότι οι περιλήψεις που εξάγονται από το σύστημα SIMBA καλύπτουν τα σημαντικά θέματα των συνόλων των κειμένων που συνοψίζουν σε σχέση με άλλα συστήματα.



Εικόνα 7 – Μέτρηση ROUGE-L για τις περιλήψεις gistsumm και simba [22]

5.3 ATLAS

Το ερευνητικό πρόγραμμα ATLAS [23], το οποίο ξεκίνησε τον Μάρτιο του 2010, αποσκοπεί στην δημιουργία ενός πλαισίου επεξεργασίας πολυγλωσσικότητας. Το συγκεκριμένο πρόγραμμα προσφέρει τη δυνατότητα να δοκιμαστεί η διαλειτουργικότητα διάφορων εργαλείων NLP για την Πολωνική γλώσσα σε συνδυασμό με άλλες γλώσσες σε πολύγλωσσα συστήματα πληροφοριών, όπως την Βουλγάρικη, Αγγλική, Γερμανική, Ελληνική και Ρουμάνικη γλώσσα. Το ATLAS εστιάζει σε Πολωνικά κείμενα και στην πρακτική εφαρμογή διάφορων προηγμένων γλωσσικών ζητημάτων, όπως την ανάλυση των σχετικών αναφορών, την περίληψη, την αυτοματοποιημένη μετάφραση ή κατηγοριοποίηση κειμένων, δημιουργώντας συχνά συνέργειες με άλλες τρέχουσες καινοτομίες. Έχοντας επιλέξει ένα κοινό πλαίσιο ενοποίησης και σχολιασμού, στη συνέχεια οι συγγραφείς του [23] αξιολογούν και προσαρμόζουν τα κατάλληλα γλωσσικά εργαλεία παραγωγής της απαραίτητης γλωσσικής πληροφορίας σε έτοιμες προς χρήση αλυσίδες επεξεργασίας.

Η αρχιτεκτονική που επιλέγουν εδώ οι συγγραφείς είναι η Αρχιτεκτονική Διαχείρισης Αδόμητης Πληροφορίας (Unstructured Information Management Architecture - UIMA), έχοντας καταλήξει σε αυτήν κατόπιν έρευνας σε μία πληθώρα διαθέσιμων και αξιόπιστων αρχιτεκτονικών, όπως για παράδειγμα την Γενική Αρχιτεκτονική για Μηχανική Κειμένου (General Architecture for Text Engineering). Η επιλογή της συγκεκριμένης αρχιτεκτονικής βασίζεται στην δυνητική επεκτασιμότητά της

μέσω της αποσύνθεσης των εφαρμογών επεξεργασίας γλώσσας σε επιμέρους συστατικά τα οποία μπορούν να αναπαραχθούν μέσω δικτυακών κόμβων. Τα εργαλεία NLP ενοποιούνται στο πλαίσιο της αρχιτεκτονικής UIMA συνθέτοντάς τα σε πρωτόγονες μηχανές συμβατές με την UIMA, διασυνδεδεμένες μεταξύ τους. Σε ορισμένες περιπτώσεις απαιτείται η τεχνική προσαρμογή για τη συνεχή χρήση τους (χωρίς να γίνεται συχνή φόρτωση των μοντέλων, των κανόνων επεξεργασίας, κ.λπ.) και την επιβολή πολύ-νηματικών κανόνων για την απρόσκοπτη λειτουργίας τους. Πέραν τούτου, το σύστημα ATLAS αναπτύχθηκε μέσω των κατάλληλων υποδομών για να διατηρείται ένα ομοιόμορφο σύστημα, χρησιμοποιώντας ιδιότητες τόσο σε επίπεδο εγγράφου όσο και επίπεδο κειμένου. Το επίπεδο κειμένου περιλαμβάνει τα σχόλια των επιμέρους παραγράφων, σημασιολογικές ενδείξεις (ετικέτες POS και μορφοσυντακτικές κατηγορίες), φράσεις ουσιαστικών (με σημασιολογικές κεφαλές) και ονομασμένες οντότητες. Ταυτόχρονα, στο ATLAS ενσωματώνονται εργαλεία κατηγοριοποίησης ανεξαρτήτου γλώσσας για ετερογενείς τομείς.

Η γλωσσική μηχανή χρησιμοποιεί διάφορους αλγόριθμους κατηγοριοποίησης, όπως τον Naive Bayesian, την σχετική εντροπία, το γνώρισμα Class-Feature Centroid, τις μηχανές υποστήριξης γράφων (Support Vector Machines – SVM), και την λανθάνουσα κατανομή Dirichlet, τα αποτελέσματα των οποίων ενοποιούνται μέσω ενός συστήματος ψηφοφορίας. Η μηχανή αυτόματης μετάφρασης συνδυάζει ένα συστατικό βασισμένο σε παραδείγματα μαζί με μία στατιστική προσέγγιση τροφοδοτούμενη από το σύστημα Moses [24]. Αφού επιλεγεί το καταλληλότερο μοντέλο μετάφρασης και το υποσύστημα παραδείγματος κειμένου, βάσει των αποτελεσμάτων από την μηχανή κατηγοριοποίησης, χρησιμοποιείται η βάση δεδομένων μετάφεσης και ο μηχανισμός στατιστικής ανάλυσης για να συνθέσουν το μεταφρασμένο κείμενο.

Αυτή τη στιγμή, υπάρχουν τρεις (3) online υπηρεσίες που κάνουν χρήση της λειτουργίας αυτής: το i-Publisher, το οποίο είναι ένα σύστημα διαχείρισης online περιεχομένου, το i-Librarian που συνθέτει μία ψηφιακή βάση επιστημονικών έργων, καθώς επίσης και οι ιστότοποι EUDocLib/PLDocLib, οι οποίοι χρησιμεύουν στην αναζήτηση εγγράφων EUR-LEX και ενέργειες του Πολωνικού κοινοβουλίου. Η ανάπτυξη των NLP εργαλείων μέσω του ATLAS αποδεικνύει έμπρακτα ότι η διαλειτουργικότητα των εργαλείων αυτών μπορεί να επιτευχθεί και είναι μάλιστα και σχετικά εύκολη στην υλοποίηση.

5.4 CoreNLP

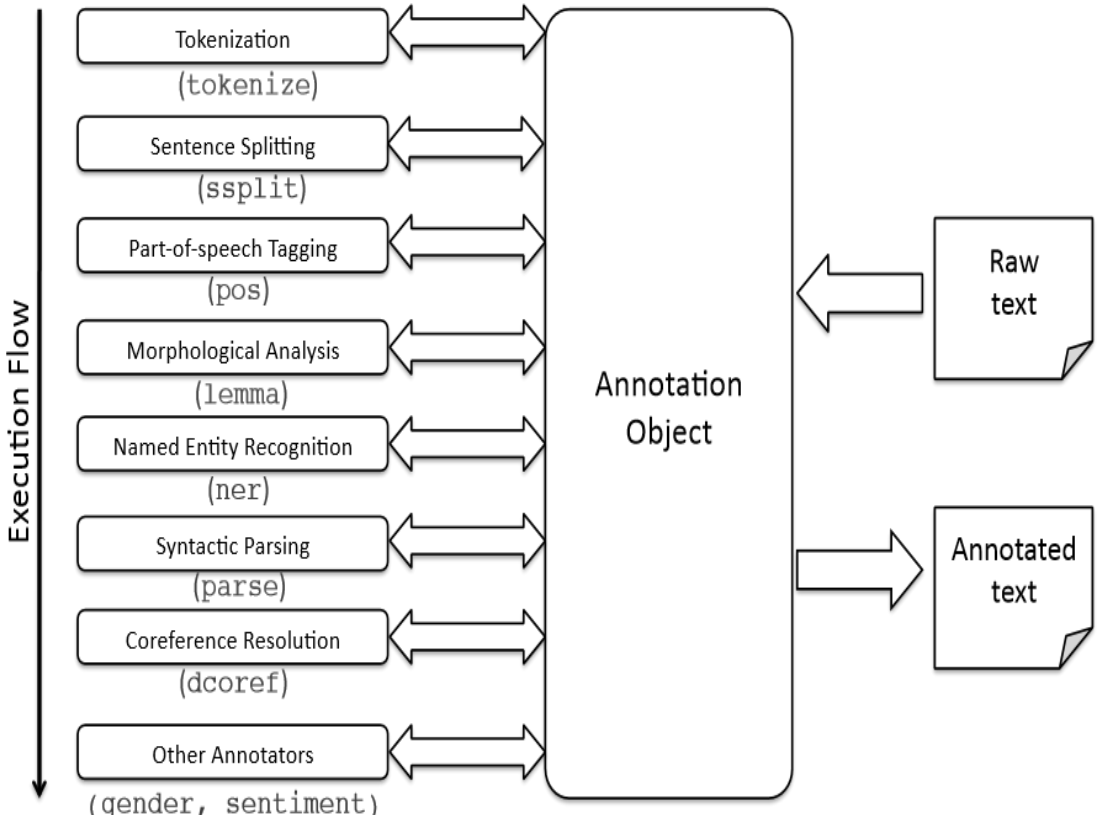
Στο [25], περιγράφεται ο σχεδιασμός και η ανάπτυξη του Stanford CoreNLP, ένα πλαίσιο ροής σχολιασμού βασισμένο σε κώδικα Java, το οποίο παρέχει τα περισσότερα από τα πιο κοινά βήματα του πυρήνα Επεξεργασίας Φυσικής Γλώσσας, από την διάσπαση έως την συσχέτιση αναφορών. Οι συγγραφείς περιγράφουν τον αρχικό σχεδιασμό του συστήματος και τις δυνατότητές του, απλά υποδείγματα χρήσης, το σύνολο των διαθέσιμων σχολιαστών και τον τρόπο διαχείρισής τους μέσω κατάλληλων ιδιοτήτων, τους τρόπους επέκτασης του συστήματος με επιπλέον σχολιαστές, και τέλος, παραθέτουν σχόλια υψηλού επιπέδου και παραρτήματα σχολιασμών. Αν και υπάρχουν αρκετά καλά εργαλεία ανάλυσης της Γλώσσας, το Stanford CoreNLP παραμένει ένα από τα πιο ευρέως χρησιμοποιούμενα.

Το σύστημα ροής του CoreNLP σχεδιάστηκε αρχικώς για εσωτερική χρήση. Το 2006 έγινε η πρώτη προσπάθεια ανάπτυξης μίας ροής σχολιασμού η οποία να μπορεί να χρησιμοποιηθεί ανεξαρτήτως εφαρμογής. Στο πλαίσιο αυτό, κατασκευάστηκε μία ενιαία διεπαφή για έναν Σχολιαστή (Annotator), ο οποίος προσθέτει κάποιου είδους πληροφορία ανάλυσης στο κείμενο. Αυτό επιτυγχάνεται δίνοντας ως είσοδο στον Σχολιαστή ένα αντικείμενο Σχολιασμού, το οποίο προσδίδει επιπλέον πληροφορίες. Ο σχολιασμός αποθηκεύεται σε ένα ετερογενή χάρτη, ακολουθώντας τις ιδέες για την επιλογή των τύπων δεδομένων όπως παρουσιάζονται στο [26]. Η βασική αρχιτεκτονική έχει αποδειχθεί αρκετά επιτυχής, και εξακολουθεί να είναι η βάση του συστήματος που περιγράφεται εδώ. Στην επομένη εικόνα παρουσιάζεται η αρχιτεκτονική του συστήματος, όπου το ακατέργαστο κείμενο δίδεται σε ένα αντικείμενο Σχολιασμού και στη συνέχεια μία ακολουθία Σχολιαστών προσθέτει πληροφορία στην ροή ανάλυσης. Ο παραγόμενος Σχολιασμός εξάγεται είτε σε XML είτε σε μορφότυπα απλού κειμένου, περιλαμβάνοντας όλες τις πληροφορίες ανάλυσης που προστέθηκαν από τους Σχολιαστές.

Τα κίνητρα κατασκευής του CoreNLP περιλαμβάνουν:

- Η δυνατότητα εξαγωγής γλωσσολογικών σχολίων από κείμενα, με γρήγορο και εύκολο τρόπο.
- Η απόκρυψη παρεκκλίσεων μεταξύ διαφορετικών συστατικών μέσω ενός κοινού API.

- Η ύπαρξη ενός ελάχιστου εννοιολογικού αποτυπώματος, ώστε να γίνεται ευκολότερη η εκμάθηση του συστήματος.
- Η παροχή ενός ελαφρούς πλαισίου, χρησιμοποιώντας απλά αντικείμενα Java, αντί για αντικείμενα όπως XML ή UIMA.



Εικόνα 8 – Αρχιτεκτονική Συστήματος CoreNLP

Το 2009, αρχικώς στο πλαίσιο ενός ερευνητικού προγράμματος, το υπάρχον σύστημα επεκτάθηκε ώστε να μπορεί να χρησιμοποιηθεί από ευρύτερη γκάμα χρηστών. Κατασκευάστηκε μία διεπαφή χρήσης του συστήματος μέσω γραμμής εντολών (Command-Line), με τη δυνατότητα εξαγωγής ενός Σχολίου σε διάφορα μορφότυπα, συμπεριλαμβανομένου και του XML. Επιπλέον εργασίες ανάπτυξης οδήγησαν στην έκδοση του συστήματος ως λογισμικό Ανοιχτού Πηγαίου κώδικα (Open Source) τελικώς το 2010.

Βέβαια, σε επίπεδο αρχιτεκτονικής, το Stanford CoreNLP δεν επιχειρεί να κάνει τα πάντα. Δεν είναι τίποτα παραπάνω από μία απλή αρχιτεκτονικής δομημένης ροής, παρέχοντας ένα συμπαγές Java API. Δεν επιχειρεί να καλύψει όλες τις διαφορετικές εφαρμογές, ούτε λαμβάνει υπόψη την αρχιτεκτονική των μηχανημάτων υποστήριξής τους (π.χ. κατανομή πόρων συστήματος). Όμως, οι απαιτήσεις αυτές ικανοποιούν μία ευρεία γκάμα πιθανών χρηστών, και η απλότητα του συστήματος βοηθά τους χρήστες να ξεκινήσουν να χρησιμοποιούν το πλαίσιο αυτό. Αυτό είναι και το κυριότερο πλεονέκτημα του CoreNLP σε αντιδιαστολή με μεγαλύτερα πλαίσια όπως το UIMA [27] ή το GATE [28]. Άλλα συστήματα επιχειρούν να παρέχουν περισσότερα, όπως το UIUC Curator [29], το οποίο περιλαμβάνει τη δυνατότητα επικοινωνίας χρήστη-εξυπηρετητή σε διαφορετικές μηχανές για την επεξεργασία και την χρήση της κρυφής μνήμης για τις αναλύσεις της Φυσικής Γλώσσας. Βέβαια, η λειτουργία κοστίζει. Το σύστημα είναι δύσκολο να εγκατασταθεί και να κατανοηθεί. Επιπλέον, σε πρακτικό επίπεδο, ένας οργανισμός ενδέχεται να έχει επιλέξει μία κλιμακούμενη (scale-out) λύση, η οποία είναι διαφορετική από εκείνη που παρέχεται από το εργαλείο ανάλυσης της Φυσικής Γλώσσας. Για παράδειγμα, ένας οργανισμός μπορεί να χρησιμοποιεί το Kryo ή το Google protobuf αντί για το Apache Thrift που περιλαμβάνεται στο UIUC Curator. Σε αυτές τις περιπτώσεις, είναι προτιμότερο για τον χρήστη να επιλέξει ένα πιο μικρό και αυτάρκες σύστημα ανάλυσης Φυσικής Γλώσσας, από ένα μεγαλύτερο που εμπεριέχει επιπλέον δυνατότητες και στοιχεία που να μην είναι απαραίτητα στη χρήση.

Οι σχολιαστές που παρέχονται με το Stanford CoreNLP μπορούν να λειτουργήσουν με οποιαδήποτε κωδικοποίηση χαρακτήρων, είτε μέσω της Java που παρέχει ένα πολύ καλό σύστημα κωδικοποίησης, είτε μέσω της εγγενής UTF-8 κωδικοποίησης. Οι σχολιαστές επίσης υποστηρίζουν την επεξεργασία διαφόρων ανθρωπίνων γλωσσών, εφόσον υπάρχουν τα κατάλληλα μοντέλα και οι πόροι για τις συγκεκριμένες γλώσσες. Το σύστημα υποστηρίζει εγγενώς την Αγγλική γλώσσα. Επιπλέον πακέτα μοντέλων του CoreNLP παρέχουν υποστήριξη για την Κινέζικη γλώσσα, όπως επίσης για την επεξεργασία Αγγλικών χωρίς τη διάκριση πεζών-κεφαλαίων. Η υποστήριξη άλλων γλωσσών είναι σε περιορισμένο επίπεδο, αλλά πολύ Σχολιαστές υποστηρίζουν μοντέλα της Γαλλικής, Γερμανικής και Αραβικής Γλώσσας. Επιπλέον υποστήριξη άλλων γλωσσών μπορεί να κατασκευαστεί μέσω των εσωτερικών εργαλείων του πλαισίου. Αξίζει να σημειωθεί ότι κάποιοι από τους Σχολιαστές των επιμέρους μοντέλων εκπαιδεύονται από σχολιασμένα σώματα (corpora) με τη χρήση μηχανών εκμάθησης υπό επίβλεψη (Supervised Machine Learning), ενώ κάποια άλλα αποτελούν στοιχεία βασισμένα σε κανόνες, τα οποία ενδέχεται να χρειαστούν ορισμένες φορές επιπλέον στοιχεία της γλώσσας.

5.5 *λόγος*: Ένα Σύστημα Μετάφρασης Ερωτημάτων σε Αφηγήσεις

Οι συγγραφείς του [30] παρουσιάζουν το σύστημα «λόγος», το οποίο παράγει μεταφράσεις SQL ερωτημάτων σε φυσική γλώσσα. Η επεξήγηση ερωτημάτων σε κείμενο είναι χρήσιμη σε πληθώρα σεναρίων. Για παράδειγμα, πολλές εφαρμογές (π.χ. ιστότοποι μουσείων, ηλεκτρονικές βιβλιοθήκες, εμπορικοί ιστότοποι, κλπ.) προσφέρουν ένα περιβάλλον σε μορφή φόρμας για την παραγωγή ερωτημάτων σε διαδικτυακές βάσεις αναζήτησης. Επίσης, τα ολοένα και περισσότερο αυξανόμενα συστήματα παραγωγής ιστοχώρων με βάσεις δεδομένων, για την κατασκευή των οποίων δεν απαιτείται ιδιαίτερη προγραμματιστική γνώση, επιτρέπουν στους κατασκευαστές τη δημιουργία τέτοιων εφαρμογών, βάσει των απαιτήσεων του πελάτη, μέσω της επεξεργασίας οπτικών συστατικών (π.χ. drag and drop). Σε όλα αυτά τα σενάρια, που αφορούν την αναζήτηση και τον προγραμματισμό σε βάσεις δεδομένων, οι διαδράσεις του χρήστη με την εφαρμογή μετασχηματίζονται σε δομημένα ερωτήματα. Η επεξήγηση των εγγενών ερωτημάτων χωρίς να παραθέτονται οι λεπτομέρειες της εκάστοτε γλώσσας ερωτημάτων κρίνεται ιδιαίτερα σημαντική, ειδικά όταν η εκτέλεση ενός ερωτήματος μπορεί να παράγει διαφορετικό αποτέλεσμα από αυτό που περίμενε ο χρήστης. Η μετάφραση των επιλογών του χρήστη από μία φόρμα σε αφηγηματικές λύσεις μπορεί να βοηθήσει στην παραγωγή των σωστών ερωτημάτων, χωρίς να απαιτείται ιδιαίτερη τεχνογνωσία από το χρήστη σε ειδικευμένες γλώσσες ή διεπαφές ερωτημάτων.

Η μετάφραση ερωτημάτων μπορεί να είναι βοηθητική όταν οι χρήστες χρησιμοποιούν δομημένα ερωτήματα μέσω γλώσσας ερωτημάτων. Προτού ο προγραμματιστής εκτελέσει την εντολή, μπορεί να δει την αφηγηματική της περιγραφή, η οποία του είναι πιο οικεία, ώστε να ελέγξει την εγκυρότητα της σημασίας του ερωτήματος που ετοιμάζεται να εκτελέσει. Ένας χρήστης που προσπαθεί να καταλάβει ένα μήνυμα σφάλματος σχετικά με ένα λάθος ερώτημα μπορεί να προτιμήσει να έχει μια εξήγηση αυτού του ερωτήματος σε μια οικεία γλώσσα, αντί να πάρει πίσω έναν κωδικό σφάλματος και μια γενική περιγραφή του σφάλματος. Ως άλλο παράδειγμα, όταν ένα ερώτημα επιστρέφει μια κενή απάντηση, μια εξήγηση του ερωτήματος μπορεί να βοηθήσει στον εντοπισμό των τμημάτων της επερώτησης που είναι υπεύθυνα για την αποτυχία. Ομοίως, όταν το ερώτημα επιστρέφει ένα πολύ μεγάλο αριθμό των απαντήσεων, μια εξήγηση του ερωτήματος μπορεί να βοηθήσει το χρήστη να κατανοήσει τους λόγους και, ενδεχομένως, να ξαναγράψει το ερώτημα σε ένα που επιστρέφει λιγότερα αποτελέσματα. Ως ακόμη ένα άλλο παράδειγμα, η μετάφραση ερωτημάτων σε περιγραφές της φυσικής γλώσσας μπορεί να είναι βολική σε αυτο-οργανωμένες ασκήσεις ως μέρος μιας τάξης μάθησης βάσεων δεδομένων για τους μαθητές που μπορεί να μην είναι εξοικειωμένοι με τις γλώσσες επερωτήσεων.

Παραδοσιακά, η εφαρμογή των τεχνικών φυσικής γλώσσας για το front-end του συστήματος πληροφόρησης επικεντρώθηκε κυρίως στην αντίθετη κατεύθυνση από εκείνες που μελετώνται εδώ: από αιτήματα σε φυσική γλώσσα προς ερωτήματα της βάσης (π.χ., [31] [32]). Άλλες ερευνητικές προσπάθειες έχουν προσπαθήσει να παρέχουν οπτικές εξηγήσεις στα ερωτήματα (π.χ., [33]). Βρίσκοντας μια σωστή και ουσιαστική μετάφραση για ένα ερώτημα δεν είναι ασήμαντη εργασία. Μερικά ερωτήματα δεν έχουν προφανή σημασιολογία. Άλλα ερωτήματα περιέχουν διαφορετικό είδος δυσκολίας. Ένα ερώτημα με πολλές συνενώσεις μπορεί να είναι απλό να μεταφραστεί απλά ακολουθώντας συσχετίσεις πρωτεύοντος κλειδιού (PK) - ξένου κλειδιού (FK), αλλά μια ουσιαστική μετάφραση δεν μπορεί να είναι κάτι

τετριμμένο. Το σύστημα λόγος, ασχολείται με τέτοιες δυσκολίες με δύο βασικούς μηχανισμούς: (α) ένα μηχανισμό προτυποποίησης, που επιτρέπει τη μετάφραση των ιδιόμορφων συντακτικά προτύπων και την παραγωγή τους σε φυσικό κείμενο, και (β) ένα σύνολο στρατηγικών διάσχισης γράφου ερωτημάτων που παράγουν κείμενο από ένα ερώτημα αποφεύγοντας την επανάληψη ορισμένων φράσεων ή ουσιαστικών και μακροσκελείς αφηγήσεις.

5.5.1 Σχήμα και Ερωτήματα της βάσης ως γράφοι

Το σύστημα αυτό λαμβάνει μια γραφο-θεωρητική προσέγγιση για την αναπαράσταση ενός σχήματος βάσης δεδομένων και τις διάφορες μορφές των δομημένων ερωτημάτων ως κατευθυνόμενους γράφους. Το γράφημα του σχήματος της βάσης δεδομένων είναι ένα κατευθυνόμενο γράφημα που αποτυπώνει τους βασικούς ρόλους των σχέσεων και αποδίδει ερωτήματα πάνω από τη βάση δεδομένων. Οι κόμβοι του γράφου είναι οι σχέσεις και τα χαρακτηριστικά, και τα άκρα του αποτελούν είτε ένταξη (σύνδεση χαρακτηριστικών στις σχέσεις), επιλογή (από τις σχέσεις με χαρακτηριστικά) ή άκρα κατηγορήματος (να αποδίδουν ενώσεις από το χαρακτηριστικό). Ένα απλό ερώτημα γράφου αποτυπώνει τις πιθανές σημασιολογίες μιας επερώτησης SPJ και είναι μια επέκταση του γραφήματος του σχήματος της βάσης δεδομένων. Οι κόμβοι του είναι σχέσεις (μία για κάθε μεταβλητή πλειάδα στο ερώτημα), χαρακτηριστικά (μία για κάθε χαρακτηριστικό του περιστατικού), και κόμβοι αξίας (ένα για κάθε αξία ή ένα σύνολο αξιών που προσδιορίζονται στην ερώτηση). Οι ακμές του είναι τα μέλη (σύλληψη προβολών), το κατηγόρημα (σύλληψη της ένωσης), και άκρα επιλογής (σύλληψη συνθηκών επιλογής). Για πιο σύνθετους τύπους ερωτημάτων, ο γράφος του ερωτήματος μπορεί να περιέχει άλλο άκρο και ο κόμβος των τύπων που λειτουργούν για τη σύλληψη καθώς και ρήτρες ταξινόμησης (order-by), ομαδοποίησης (group-by) και περιεκτικότητας (having), όπως αναφέρονται στο [34].

5.5.2 Γράφοι Σχολιασμών και Πρότυπα

Το νόημα δίνεται στα διάφορα τμήματα ενός ερωτήματος σχολιάζοντας τα άκρα του γράφου του ερωτήματος με ετικέτες του προτύπου, χρησιμοποιώντας ένα επεκτάσιμο μηχανισμό προτυποποίησης. Κάθε κόμβος που μπορεί να είναι μέρος ενός γραφήματος μπορεί να συμπληρώνεται από μια ετικέτα η οποία σηματοδοτεί την έννοια του κόμβου σε φυσική γλώσσα. Ομοίως, κάθε ακμή (ή διαδρομή) σύνδεσης δύο κόμβων μπορεί να συμπληρώνεται από μια ετικέτα η οποία σηματοδοτεί την έννοια, σε φυσική γλώσσα, από τη σχέση μεταξύ του κόμβου της πηγής και του κόμβου προορισμού. Οι ετικέτες αποθηκεύονται στο γράφημα του σχήματος της βάσης δεδομένων για τους δύο κόμβους και τις ακμές. Ένα γράφημα κληρονομεί αυτά τα άκρα από τη γραφική παράσταση της βάσης δεδομένων.

Οι μέθοδοι μετάφρασης διασχίζουν το γράφημα του ερωτήματος και δημιουργούν φράσεις συνθέτοντας τις ετικέτες που βρίσκονται στη διαδρομή. Για την παραγωγή πιο φυσικών αποτελεσμάτων, χρησιμοποιείται το πρότυπο ετικετών σε διαφορετικά επίπεδα λεπτομέρειας και ένας επεκτάσιμος μηχανισμός προτύπου για την συγχώνευση αυτών των ετικετών. Μια ετικέτα πρότυπο, l ((v, u)), έχει ανατεθεί σε μια ακμή (v, u), ή σε μια διαδρομή που συνδέει v στο u. Αυτό το πρότυπο χρησιμοποιείται για την ερμηνεία της σχέσης μεταξύ V και U σε μια αφήγηση. Μια ετικέτα πρότυπο μπορεί να έχει τη μορφή:

$$\mathrm{l}\left((\mathrm{v},\mathrm{u})\right) \;=\; \mathrm{expr1} \;+\; \mathrm{L}\left(\mathrm{v}\right) \;+\; \mathrm{expr2} \;+\; \mathrm{l}\left(\mathrm{u}\right) \;+\; \mathrm{expr3}$$

Τα πρότυπα ετικέτας δημιουργούνται από τον άνθρωπο, και μπορούν να παράγουν συνοπτικό κείμενο υψηλής ποιότητας.

5.5.3 Μετάφραση Ερωτημάτων ως Διάσχιση Γράφου

Χρησιμοποιούμε τρία τομέα ανεξάρτητες στρατηγικές διάσχισης για την αποτελεσματική διερεύνηση γραφήματα ερώτημα και συνθέτοντας περιγραφές ερώτημα ως αφηγήσεις.

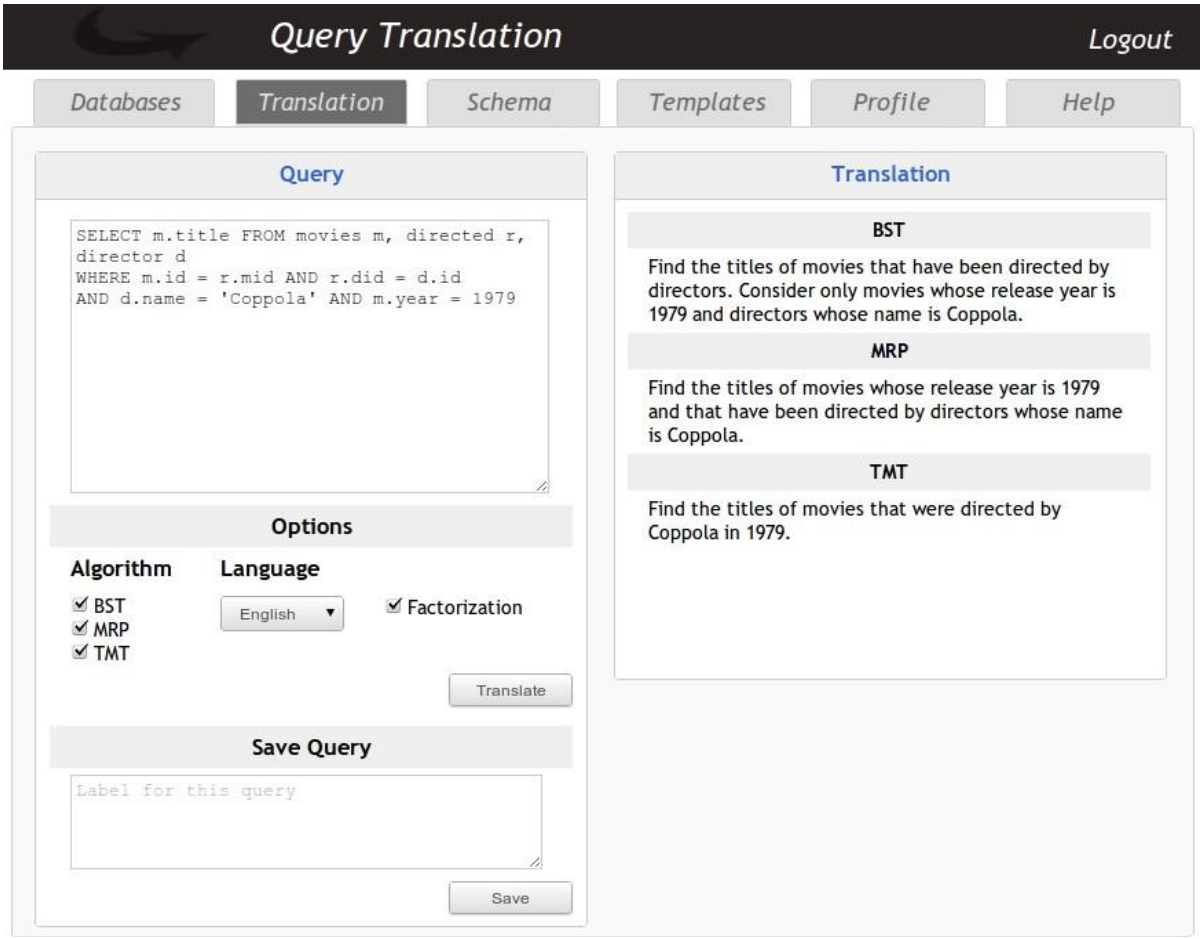
Η πρώτη στρατηγική (BST αλγόριθμο) συνθέτει ξεχωριστές ρήτρες για κάθε τμήμα του ερωτήματος. Πρώτον, μεταφράζει τα άκρα των μελών, στη συνέχεια συνδέει όλες τις σχέσεις του ερωτήματος για το θέμα του ερωτήματος μέσω των ενώσεων του ερωτήματος, και τέλος, διαβάζει τις διαδρομές που συνδέουν τις σχέσεις με τις τιμές των κόμβων που έχουν καθοριστεί για τα χαρακτηριστικά αυτών των σχέσεων. Η μετάφραση πραγματοποιείται σε πρώτο επίπεδο από την γραφική παράσταση του ερωτήματος ξεκινώντας από το ερώτημα που υπόκεινται σε όλες τις σχέσεις μέσω των ενώσεών τους στο γράφημα. Το θέμα του ερωτήματος αντιπροσωπεύει σε τι αναφέρεται το ερώτημα. Ο

προσδιορισμός του αντικειμένου του ερωτήματος είναι σημαντικό γιατί καθορίζει το πώς θα διασχίσει το γράφημα του ερωτήματος, δηλαδή, τη κατεύθυνση της μετάφρασης του ερωτήματος, και τι είδους ρήτρες παράγονται. Φυσικά, είναι μια σχέση με τα χαρακτηριστικά που προβλέπονται στον όρο select, που κατέχουν τη κεντρική θέση στο γράφημα του ερωτήματος [34].

Στη δεύτερη στρατηγική (αλγόριθμος MRP), η μετάφραση πραγματοποιείται κατά ολιστικό τρόπο, όπου οι πληροφορίες από όλα τα μέρη του γραφήματος του ερωτήματος είναι αναμεμειγμένες στη μετάφραση καθώς διασχίζεται ο γράφος. Η βασική ιδέα εδώ είναι ότι, ενώ στο BST όλα τα εξαρτήματα του ερωτήματος αναφέρονται πάντοτε στο θέμα του ερωτήματος, στο MRP χρησιμοποιούνται επιπλέον σημεία αναφοράς για να αποφευχθούν οι μεγάλες, πιθανώς αφύσικες, ποινές. Με αυτόν τον τρόπο, θα χωριστεί σημασιολογικά η μετάφραση σε πολλαπλά σημεία.

Τέλος, η τρίτη στρατηγική (αλγόριθμος TMT) επιτρέπει τη χρήση προκαθορισμένων, πιο πλούσιων, πρότυπα για τα μέρη του ερωτήματος, σε μια προσπάθεια να παράγονται πιο συνοπτικές μεταφράσεις.

Παραδείγματα τέτοιων μεταφράσεων φαίνονται στην επόμενη εικόνα.



Εικόνα 9 – λόγος: Φόρμα μετάφρασης [30]

5.5.4 Διάδραση με το σύστημα *λόγος*

5.5.4.1 Οπτική Χρήστη.

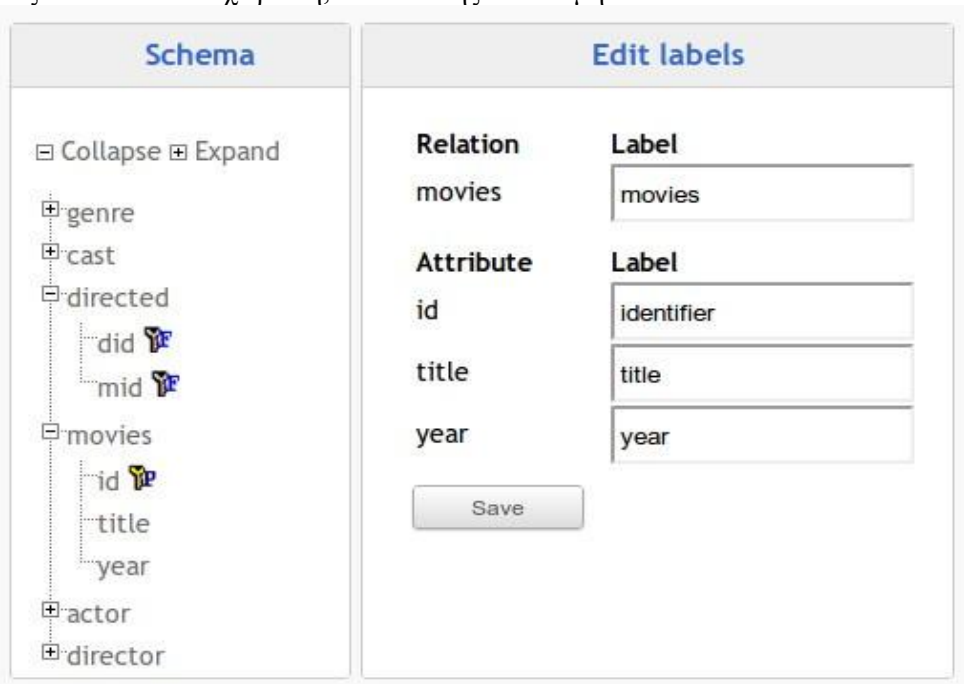
Τυπικά, ένας τακτικός χρήστης ενδιαφέρεται για τις μεταφράσεις ερωτημάτων και το προφίλ των φορμών του.

- **Μεταφράσεις ερωτημάτων.**

Η μετάφραση ενός ερωτήματος επιτυγχάνεται μέσω μιας φόρμας, όπου ο χρήστης παρέχει ένα ερώτημα και βλέπει τη μετάφρασή της σε φυσική γλώσσα. Η μετάφραση μπορεί να προσαρμοστεί με τη ρύθμιση κατάλληλων παραμέτρων, όπως την προτιμώμενη γλώσσα, τον αλγόριθμο της μετάφρασης, και

την παραγοντοποίηση. Η παραγοντοποίηση ελέγχει αν οι ειδικές λέξεις, όπως άρθρα ή αναφορικές αντωνυμίες επαναλαμβάνονται ή όχι, ώστε να ρυθμίζεται η μετάφραση. Είναι δυνατόν να παραχθούν περισσότερες από μία μεταφράσεις με την ίδια μορφή με την επιλογή πολλαπλών αλγορίθμων μετάφρασης. Στη συνέχεια, ο χρήστης μπορεί να εκδώσει ένα νέο ερώτημα ή να βελτιώσει την υπάρχουσα μετάφραση τροποποιώντας τις παραμέτρους. Είναι επίσης δυνατή η επισήμανση και αποθήκευση ενός ερωτήματος για μεταγενέστερη χρήση.

Για την έκδοση ενός ερωτήματος, πρέπει κανείς να γνωρίζει το σχήμα της βάσης δεδομένων. Το «λόγος» διαθέτει μία δεντρική δομή του σχήματος της βάσης, η οποία επιτρέπει μια γρήγορη επισκόπηση του σχήματος της βάσης δεδομένων. Πρώτον, ο χρήστης παρουσιάζεται με τους κόμβους του ανώτατου επιπέδου του δέντρου δομής, δηλαδή, τις σχέσεις της βάσης δεδομένων. Επεκτείνοντας τους κόμβους σχέσεων, εμφανίζονται τα παιδιά των κόμβων, ενώ επιλέγοντας ένα συγκεκριμένο κόμβο - είτε σχέση είτε γνώρισμα – εμφανίζεται η ετικέτα της. Στην τελευταία περίπτωση, η επεξεργασία της ετικέτας, που γίνεται συνήθως από κάποιο διαχειριστή, είναι επίσης διαθέσιμη.



Εικόνα 10 – λόγος: Εύρεση σχήματος και ετικετοποίηση [30]

- **Προφίλ Χρήστη.**

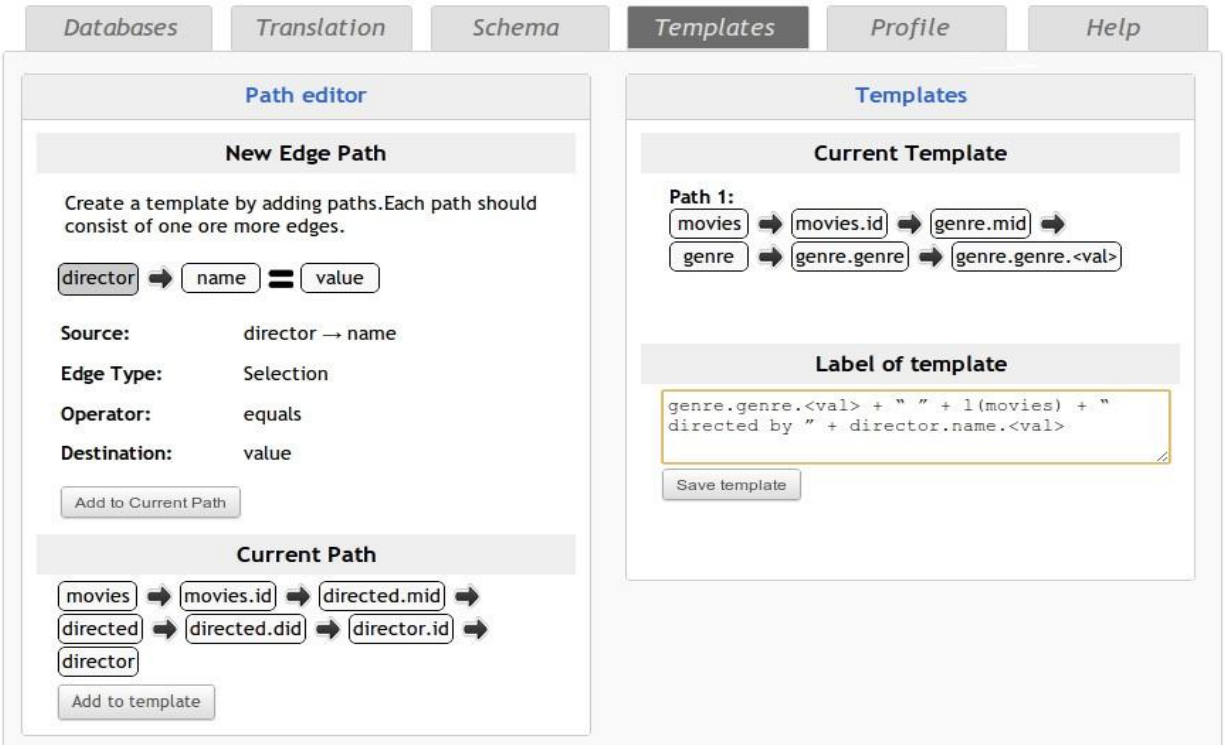
Το σύστημα «λόγος» διαθέτει μια διαδραστική διεπαφή ιστού που επιτρέπει την προσαρμοσμένη μετάφραση των ερωτημάτων. Προκειμένου να παρέχει μια απρόσκοπτη εμπειρία χρήστη, αποθηκεύονται τα δεδομένα του προφίλ. Τα δεδομένα αυτά περιλαμβάνουν τις προτιμήσεις μετάφρασης, όπως τη γλώσσα, τον αλγόριθμος της μετάφρασης, την παραγοντοποίηση, καθώς και μια λίστα αποθηκευμένων ερωτημάτων. Μέσω της σελίδας του προφίλ χρήστη, οι προτιμήσεις μετάφρασης μπορούν να τροποποιηθούν και, στη συνέχεια, μπορούν να επιλέγονται ερωτήματα από το αρχείο ιστορικού για μετάφραση. Οι χρήστες μπορούν επίσης να εκχωρούν ετικέτες σε ερωτήματα του από το ιστορικό.

Εικόνα 11 – λόγος: Ρυθμίσεις χρήστη [30]

5.5.4.2 Οπτική Διαχειριστή.

Το «λόγος» μπορεί να χρησιμοποιηθεί για την βελτίωση ενός σχήματος βάσης δεδομένων με τρόπο που να επιτρέπει ουσιαστικές και χρήσιμες μεταφράσεις. Το σύστημα συνδέεται με ένα DBMS και εξάγει τα μεταδεδομένα του σχήματος μιας βάσης δεδομένων. Αφού συνδεθεί με μια ειδική βάση δεδομένων, ο ορισμός των προτύπων πραγματοποιείται με τη χρήση ολοκληρωμένης κονσόλας διαχείρισης του συστήματος. Η εφαρμογή υποστηρίζει προεπιλεγμένες ετικέτες, αλλά εφόσον ο σχεδιαστής παρέχει το σύστημα με πιο κατάλληλες ετικέτες, τα αποτελέσματα της μετάφρασης γίνονται ακόμα πιο περιγραφικά.

Για την εξομάλυνση των βραχυχρόνιων διακυμάνσεων της ρευστότητας, το πρότυπο σύνθεσης του γράφου γίνεται σε διαδοχικά βήματα με τη βοήθεια του συστήματος «λόγος». Το πρότυπο του γραφήματος που χρησιμοποιείται, κάνει δύο διαδρομές. Η πρώτη ξεκινά από τον κόμβο σχέσης, πηγαίνει σε άλλους κόμβους, χρησιμοποιώντας τους κόμβους των χαρακτηριστικών PK-FK, στη συνέχεια κινείται προς την ιδιότητα ονόματος του κόμβου σχέσης και, τέλος, στην τιμή του κόμβου. Η δεύτερη διαδρομή ξεκινάει από τον κόμβο σχέσης, παίρνει την τιμή του κόμβου χρησιμοποιώντας τα χαρακτηριστικά PK-FK του κόμβου, αμέσως μετά μεταβαίνει στην ιδιότητα του κόμβου σχέσης και, τέλος, την τιμή του κόμβου.



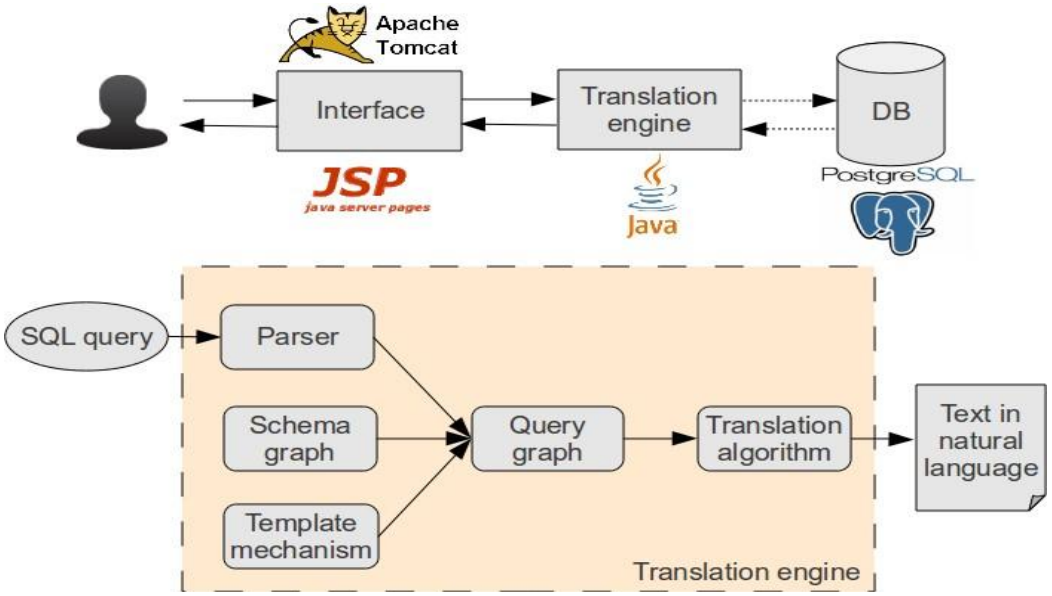
Εικόνα 12 – λόγος: Κατασκευή Template [30]

5.5.5 Πολυγλωσσικές μεταφράσεις ερωτημάτων.

Ο «λόγος» προσφέρει μεταφράσεις και σε άλλες γλώσσες εκτός της αγγλικής. Σε πολλές απλές ρυθμίσεις τοπικότητας, η υποστήριξη μιας γλώσσα συνίσταται μόνο στην αντιστοίχιση των αλφαριθμητικών. Η παραγωγή φυσικής γλώσσας, ωστόσο, απαιτεί την πλήρη εκφραστικότητα και την ευελιξία. Δεδομένου ότι κάθε γλώσσα έχει τις δικές γραμματικές και συντακτικές ιδιομορφίες, η απλή εκχώρηση αλφαριθμητικού δεν αρκεί σε αυτή την περίπτωση. Το σύστημα αυτό αντιμετωπίζει αυτή την πρόκληση εξετάζοντας την επεκτασιμότητα της γλώσσας, σε ένα πιο αφηρημένο επίπεδο, έτσι ώστε να εσωκλείει τις γραμματικές και συντακτικές πτυχές της. Στο API του, κάθε γλώσσα καθορίζεται μέσα από μια προγραμματική ενότητα, η οποία καθορίζει τις συγκεκριμένες ιδιότητες και τη συμπεριφορά του συστήματος σε ότι αφορά την μετάφραση. Επιπλέον, η μονάδα κληρονομικότητας επιτρέπει την ευελιξία μέσω της επαναχρησιμοποίησης υπάρχοντος κώδικα. Με αυτή την προσέγγιση, μια νέα γλώσσα μπορεί να υποστηριχθεί με τη δημιουργία μιας νέας μονάδας και την ενσωμάτωσή της στο σύστημα. Αυτό φαίνεται και σε πρακτικό επίπεδο καθώς, εκτός από τα αγγλικά, το σύστημα «λόγος» μπορεί επίσης να δημιουργήσει μεταφράσεις στα ισπανικά και ελληνικά.

5.5.6 Αρχιτεκτονική συστήματος

Το σύστημα «λόγος» απαρτίζεται από δύο ενότητες, την *Μηχανή Μετάφρασης* και το *Γραφικό Περιβάλλον (Graphical User Interface – GUI)*. Η Μηχανή Μετάφρασης περιλαμβάνει πέντε διακριτές ενότητες: τους γράφους Σχήματος και Ερωτημάτων, τον Parser, το Μηχανισμό Προτύπων, και την ενότητα του αλγορίθμου μετάφρασης. Το σύστημα έχει κατασκευαστεί σε Java, χρησιμοποιώντας τη βιβλιοθήκη Apache collections. Ως είσοδο δέχεται ένα σχήμα βάσης, το SQL ερώτημα, και τις προτιμήσεις μετάφρασης, και παράγει ως έξοδο το ερώτημα σε φυσική γλώσσα. Η βάση δεδομένων που χρησιμοποιείται είναι η PostgreSQL.



Εικόνα 13 – Αρχιτεκτονική του συστήματος λόγος [30]

- Οι *γράφοι Σχήματος και Ερωτημάτων* αντιστοιχούν στα σχήματα της βάσης και τα SQL ερωτήματά της. Οι κόμβοι και ακμές του γράφου ζωγραφίζονται με ετικέτες που αντιστοιχούν στο πραγματικό τους νόημα στη φυσική γλώσσα.
- Ο *Parser* είναι μία ενότητα που διενεργεί λεξική και συντακτική ανάλυση στο ερώτημα και παράγει τμήματα του γράφου του ερωτήματος, χρησιμοποιώντας πληροφορίες από τον γράφο του Σχήματος και το μηχανισμό προτύπων.
- Ο *μηχανισμός Προτύπου* είναι μία δυναμική λειτουργία που ορίζει τις ετικέτες των προτύπων και βοηθά στην παραγωγή υψηλής ποιότητας, συνοπτικού κειμένου.
- Ο *αλγόριθμος Μετάφρασης* module εμφωλεύει τις διάφορες στρατηγικές διάσχισης του γράφου, μοντελοποιημένος σε ξεχωριστές κλάσεις. Η προσέγγιση αυτή προσφέρει την επεκτασιμότητα του συστήματος, καθώς επιτρέπει την εισαγωγή νέων αλγορίθμων με απλό τρόπο.

5.6 Ανάπτυξη Ευφυούς Εικονικού Περιβάλλοντος με NLP

Ο συνδυασμός ευφυών τεχνικών και εργαλείων, μαζί με αποτελεσματικά μέσα για την γραφική αναπαράσταση και την αλληλεπίδραση των διαφόρων ειδών τους, έχει οδηγήσει σε μια ένα νέο τομέα μελέτης που ονομάζεται «Ευφυές Εικονικό Περιβάλλον» (Intelligent Virtual Environment - IVE). Το περιβάλλον αυτό προσφέρει μια πολύ πιο σύνθετη προσέγγιση στην αλληλεπίδραση του χρήστη [35]. Σε αυτήν την προτεινόμενη έρευνα αναπτύσσεται μια σουίτα λογισμικού που ενσωματώνει Ευφυή Εικονικά Περιβάλλοντα (IVE) που παρέχουν Διεπαφή Φυσικής Γλώσσας για την άμεση επεξεργασία της εικονικής πραγματικότητας (VR).

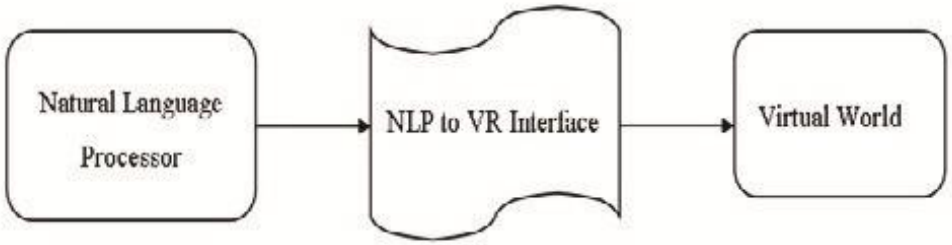
Η σουίτα IVE της προτεινόμενης μελέτης [36] είναι ένα σύστημα οδηγούμενο από κείμενα φυσικής γλώσσας για την οπτική προσομοίωση του μοντέλου λειτουργίας σε τρισδιάστατο (3D) εικονικό χώρο [37]. Αυτή η σουίτα είναι σε θέση να λαμβάνει ως είσοδο κείμενα στην Αγγλική γλώσσα. Το τμήμα NLP της σουίτας εξάγει εντολές από το δεδομένο κείμενο. Ο parser το κειμένου αναλύει το κείμενο που δίδεται και δημιουργεί μία βάση κανόνων. Η βάση κανόνων διατηρεί μία ακολουθία ενεργειών που πρέπει να εκτελεστούν. Η διεπαφή κατευθύνει όλες αυτές τις λεπτομέρειες απεικόνισης στον renderer. Ο renderer εμφανίζει το αντικείμενο με επιπλέον χαρακτηριστικά, βάσει δεδομένων που καθορίζονται από το χρήστη.

Ο σκοπός της παρούσας έρευνας είναι να προωθήσει μια ικανότητα στους χρήστες να επικοινωνούν σε φυσική γλώσσα με συστήματα VR. Οι στόχοι της προτεινόμενης έρευνας είναι οι εξής:

- Να διευρυνθούν τα οφέλη της NLP και του VR μέσω της κατασκευασμένης σουίτας IVE.

- Η κύρια προσπάθεια της προτεινόμενης έρευνας είναι να αναπτύξει μια διεπαφή NLP για τα συστήματα εικονικής πραγματικότητας. Με τη χρήση της NLP, οι υπολογιστές κατανοούν τις φυσικές γλώσσες δίνοντας στους χρήστες πιο βολικό τρόπο για να επικοινωνούν.
- Ο στόχος της έρευνας αυτής είναι να αναλύσει το κείμενο που εισάγεται από το χρήστη και να αποκτήσει την επιθυμητή απεικόνιση με την ενεργοποίηση του εικονικού ρεαλισμού.
- Έλεγχος της σκοπιμότητας της ενσωμάτωσης NLP και VR που οδηγεί σε μια σειρά από τεχνολογικές εξελίξεις. Η συντόμευση του χρονοδιαγράμματος για την κατασκευή του εικονικού κόσμου με τη χρήση του NLP περιβάλλοντος.

Η Εικονική Πραγματικότητα είναι μια τεχνολογία που επιτρέπει την αλληλεπίδραση με ένα τρισδιάστατο περιβάλλον, προσομοιωμένο στον υπολογιστή, σαν να ήταν πραγματικό, είτε το περιβάλλον αυτό είναι προσομοίωση του πραγματικού ή του φανταστικού κόσμου [38]. Αυτό το εικονικό περιβάλλον μπορεί να παρέχει μία διαδραστική εμπειρία και να το χειριστεί ο χρήστης σε πραγματικό χρόνο. Οι δυνατότητες που παρέχονται από τις συσκευές VR έχουν αποδειχθεί καλά υποσχόμενες ως οικονομικά εργαλεία εκπαίδευσης και εκμάθησης [39]. Στην απλούστερη μορφή του, το VR είναι η παρουσίαση και η αλληλεπίδραση με ένα συνθετικό, ηλεκτρονικό 3D κόσμο μέσω υπολογιστή, τόσο ρεαλιστικό ώστε ο χρήστης να αισθάνεται σαν να αντιμετωπίζει το πραγματικό αντικείμενο. Αυτή η διεπαφή επιτρέπει στο χρήστη να αλληλεπιδρά με το σύστημα VR με ένα φυσικό και έξυπνο τρόπο. Η ικανότητα διεπαφών φυσικής γλώσσας για τα συστήματα VR φέρει νέα προοπτική στον τομέα του IVE. Το εννοιολογικό μοντέλο της προτεινόμενης έρευνας φαίνεται στην επόμενη εικόνα.



Εικόνα 14 – Εννοιολογικό μοντέλο του συστήματος Intelligent Virtual Environment (IVE) [36]

Η σουίτα IVE αποτελείται κυρίως από τρία διακριτά τμήματα, και συγκεκριμένα:

- **Επεξεργαστής Φυσικής Γλώσσας**. Αυτή η φάση αφιερώνεται για την επεξεργασία των εισροών που δίδονται σε φυσική γλώσσα. Τα δεδομένα είναι αποθηκευμένα σε μια δομή που προσεγγίζει τη φυσική γλώσσα, προσπαθώντας να προσομοιωθεί ο τρόπος αποθήκευσης των σχετικών πληροφοριών στον ανθρώπινο εγκέφαλο.
- **Διεπαφή NLP σε VR**. Το τμήμα φροντίζει για την πραγματική διαδικασία διαβίβασης της «Γνώσης» που ανακτάται από το εισαγόμενο σενάριο, για τη δημιουργία της τελικής «Εικονικής Διάδρασης».
- **Εικονικός Κόσμος**. Σε έναν υπολογιστή, η εικονική πραγματικότητα βιώνεται κυρίως μέσα από την όραση. Η απλούστερη μορφή του 3D είναι η στερεοσκοπική απεικόνιση, στην οποία μία 3D εικόνα μπορεί να εμφανιστεί διαδραστικά στο χρήστη. Ο εικονικός κόσμος είναι το τελικό στάδιο της IVE, κατά το οποίο οι οθόνες των υπολογιστών μοντελοποιούν στην οθόνη με τις σχετικές ιδιότητες. Θα παρέχεται ένα πλήρως λειτουργικό σύστημα 3D, επιτρέποντας στο χρήστη να μπει στον κόσμο της εικονικής πραγματικότητας. Θα έχει άμεση διασύνδεση με VR μέσω στεροσκοπικών γυαλιών.

5.7 Κατηγοριοποίηση Συνόλων Εικόνας σε Ταξινόμηση μέσω μεταδεδομένων και της μηχανής Wikipedia

Οι συγγραφείς του [40] παρουσιάζουν μία μέθοδο ιεραρχικής ταξινόμησης εικόνων, επικεντρώνοντας την έρευνά τους σε μία συγκεκριμένη περίπτωση κατηγοριοποίησης εικόνας, ονόματι ταξινόμηση γκαλερί εικόνων. Μία γκαλερί εικόνων αποτελεί μέρος μίας ιστοσελίδας που εμφανίζει μία συλλογή εικόνων, η οποία έχει συνήθως ένα συγκεκριμένο θεματικό πεδίο όπως για παράδειγμα κάποιο άτομο, σύνολο ατόμων, μία εκδήλωση ή ένα τόπο. Οι συγγραφείς εστιάζουν κυρίως στις συλλογές αυτές καθώς είναι κατάλληλες για την απεικόνιση online ειδησεογραφικών στοιχείων. Με την κατηγοριοποίηση εικόνων, οι συγγραφείς ευελπιστούν στην βελτίωση της ακρίβειας της μεθόδου η οποία βρίσκει τις κατάλληλες εικονογραφήσεις για ειδησεογραφικά άρθρα. Χρησιμοποιούν μόνο μεταδεδομένα κειμένου, όπως τον τίτλο και τις λεζάντες της εκάστοτε εικόνας για την κατηγοριοποίηση. Οι γκαλερί εικόνων κατηγοριοποιούνται σε ένα ιεραρχικά δομημένο σύνολο από προκαθορισμένες κατηγορίες που αποτελούν την ταξινόμηση των συγγραφέων. Οι κατηγορίες έχουν επιλεχθεί βάσει την σχετικότητάς τους για την ταξινόμηση συλλογών εικόνων, καθώς επίσης βάσει της εφαρμογής Kalooga [41], η οποία προτείνει αυτόματα τέτοιου είδους συλλογές για ειδησεογραφικά άρθρα. Κάθε κατηγορία στην ταξινόμηση συνδέεται με μία κατηγορία από τον γράφο κατηγοριών της Wikipedia [42]. Η χαρτογράφηση αυτή χρησιμοποιείται στη διαδικασία κατηγοριοποίησης οντοτήτων. Η παραδοσιακά στατιστική προσέγγιση για την ταξινόμηση εικόνας απαιτεί ένα τρέχον ανανεωμένο και σωστά προσημασμένο σύστημα εκπαίδευσης (training-set), το οποίο όμως είναι δύσκολο να παραχθεί και να διατηρηθεί. Εδώ, οι συγγραφείς αποφεύγουν τη χρήση τέτοιου συστήματος εκπαίδευσης, υιοθετώντας μία τεχνική κατηγοριοποίησης βάσει οντολογίας, όπως περιγράφεται στο [43]. Η οντολογία που χρησιμοποιείται έχει εξαχθεί από το σύστημα κατηγοριοποίησης της Wikipedia. Οι οντότητες που βρίσκονται από τα μεταδεδομένα της εικόνας συνδέονται με σελίδες της Wikipedia, και οι πιο σημαντικές κατηγορίες της Wikipedia από τις οντότητες αυτές επιλέγονται τελικά ως κατηγορίες για τη συλλογή της εικόνας.

Η μέθοδος ταξινόμησης της συγκεκριμένης έρευνας αναλύεται σε πέντε (5) διακριτά στάδια. Συνοπτικά, τα στάδια αυτά είναι:

Εξαγωγή κειμένου της γκαλερί. Στο στάδιο αυτό ενοποιούνται σε κείμενο μεταδεδομένα όπως το URL της συλλογής, ο τίτλος, οι λεζάντες και οι λέξεις-κλειδιά της σελίδας HTML. Ο τίτλος και το URL περιλαμβάνονται δύο (2) φορές για να αποκτήσουν μεγαλύτερο βάρος κατά την ταξινόμηση. Επίσης, αφαιρούνται λέξεις όπως «εικόνα» και «gallery», που χρησιμοποιούνται ως stopwords στο σύστημα.

Εξαγωγή οντοτήτων. Στο στάδιο αυτό χρησιμοποιείται ένας εσωτερικός «wikifier» ο οποίος αναγνωρίζει τις σχετικές οντότητες από το κείμενο του παραπάνω σταδίου και τις συνδέει με τις αντίστοιχες Wikipedia σελίδες. Επίσης, κατασκευάζεται ένας γράφος σχετικότητας, στον οποίο οι κόμβοι είναι οι οντότητες αυτές και οι ακμές απεικονίζουν τη σημασιολογική συσχέτιση μεταξύ των συνδεόμενων οντοτήτων. Για τον υπολογισμό της σχετικότητας χρησιμοποιείται το μέτρο Wikipedia link-based measure (WLM) [12].

Βαθμολόγηση οντοτήτων βάσει σπουδαιότητας. Εδώ το σύστημα χρησιμοποιεί μία παραλλαγή της μεθόδου στάθμισης μέσου όρου Averaged PageRank (APW) [44], η οποία λαμβάνει υπόψη την κεντρικότητα της οντότητας στον σημασιολογικό γράφο σε συναρτήσει με την σχετική συχνότητα εμφάνισης της οντότητας. Αφού βαθμολογηθούν οι οντότητες κατά αυτό τον τρόπο, εφαρμόζεται ένα φίλτρο αποκοπής θορύβου, ώστε να αγνοούνται μη-σχετικές οντότητες.

Εύρεση και βαθμολόγηση Οντοτήτων Κατηγοριών. Στο στάδιο αυτό, το σύστημα προσπαθεί να βρει κατηγορίες ταξινόμησης για κάθε οντότητα από το σχετικό της άρθρο στην Wikipedia, το οποίο είναι συνδεδεμένο με τις κατηγορίες του από τον γράφο κατηγοριών της Wikipedia. Το σύστημα προσπαθεί να βρει ευρύτερες κατηγορίες ταξινόμησης αναζητώντας τους προγόνους των κατηγοριών του Wikipedia γράφου. Για την αποφυγή υπερ-γενίκευσης ή κατάταξης σε άσχετες κατηγορίες, η αναζήτηση αυτή πραγματοποιείται στο μέγιστο πέντε (5) φορές. Κάθε σχετική κατηγορία που αντιστοιχεί σε μία κατηγορία ταξινόμησης λαμβάνει τη σχετική βαθμονόμηση.

Επιλογή Κατηγορίας Γκαλερί. Πρόκειται για το τελευταίο στάδιο του συστήματος. Για κάθε κατηγορία, υπολογίζεται η τελική βαθμολογία της συλλογής (gallery category score – GCS), λαμβάνοντας υπόψη το πλήθος των οντοτήτων που μετέχουν σε μία εν δυνάμει κατηγορία, την σημασία

που έχουν οι οντότητες αυτές στη ταξινόμηση και την σχετικότητα κάθε κατηγορίας προς την οντότητα αυτή. Μετά τον υπολογισμό την τελικής βαθμολόγησης, εφαρμόζεται μία ζώνη αποκοπής για να διατηρηθούν μόνο οι πιο σχετικές κατηγορίες.

Το παραγόμενο σύστημα αποδίδει καλά πάνω από μία τυχαία βάση αναφοράς, και επιτυγχάνει μία βαθμολογία (ονόματι F-score) της τάξης του 0.59 στις 9 κορυφαίες κατηγορίες του συστήματος και 0.40 όταν χρησιμοποιούνται και οι 57 συνολικά κατηγορίες του συστήματος.

5.8 Διαδραστικό σύστημα ερωτο-απαντήσεων μέσω αντιστοίχισης προτύπων και SQL για την NLP

Η αλληλεπίδραση μεταξύ των υπολογιστών και των ανθρώπων είναι πάντα ενδιαφέρουσα και προκλητική. Διαφορετικοί μηχανισμοί για τους υπολογιστές προκειμένου να απαντούν σε ερωτήσεις έχουν γίνει πολύ ενδιαφέροντες με την αυξημένη χρήση των ηλεκτρονικών υπολογιστών. Τα μηχανήματα χρησιμοποιούν αυτόν τον μηχανισμό για να απαντήσουν στις ερωτήσεις που τίθενται σε φυσική γλώσσα ή τη γλώσσα SMS. Οι χρήστες θέτουν το ερώτημα ή στέλνουν το ερώτημα με το δικό τους τρόπο και με τη χρήση φυσικής γλώσσας επεξεργάζονται οι απαντήσεις του υπολογιστή με το καλύτερα συμφωνημένο αποτέλεσμα. Η κύρια πρόκληση σε αυτόν τον μηχανισμό είναι να λάβει ο χρήστης τη σωστή απάντηση για την ερώτηση. Με άλλα λόγια, οι ερωτήσεις θα πρέπει να γίνονται κατανοητές με το σωστό τρόπο από τους υπολογιστές για να δίνεται η σωστή έξοδος. Ως εκ τούτου, η επεξεργασία της φυσικής γλώσσας θα πρέπει να είναι αποτελεσματική και αποδοτική.

Το σύστημα ερωτο-απαντήσεων (QA) μπορεί να ικανοποιήσει καλύτερα τις ανάγκες των χρηστών, καθώς θα παρέχει ένα ακριβή, γρήγορο, βολικό και αποτελεσματικό τρόπο για να δώσει απαντήσεις στα ερωτήματα των χρηστών. Η προσέγγιση που χρησιμοποιείται στην έρευνα του [45] είναι ότι οι ερωτήσεις που τίθενται από τους χρήστες θα έχουν ήδη αποθηκευτεί σε μια βάση δεδομένων. Ο χρήστης δημιουργεί την ερώτηση με τον τρόπο του, είτε σε SMS ή σε φυσική γλώσσα. Αυτό το ερώτημα θα πρέπει να ταιριάζει με το ήδη αποθηκευμένο ερώτημα χρησιμοποιώντας ένα πρότυπο αλγόριθμο που ταιριάζει το κατάλληλο ερώτημα με την καλύτερη απάντηση, η οποία ανακτάται και αποστέλλεται στον χρήστη. Η αντιστοίχιση ερωτήσεων με ερωτήσεις που έχουν ήδη αποθηκευτεί στη βάση δεδομένων θα πρέπει να εκτελούνται από τον αλγόριθμο ταιριάσματος προτύπων.

Υπάρχουν τρεις γενικές μέθοδοι με τις οποίες η απάντηση μπορεί να δημιουργηθεί χρησιμοποιώντας αποθηκευμένες ερωτήσεις και απαντήσεις [46], οι οποίες είναι:

1. Προσέγγιση τεχνητής νοημοσύνης
2. Στατιστικές τεχνικές
3. Η αντιστοίχιση ενός template.

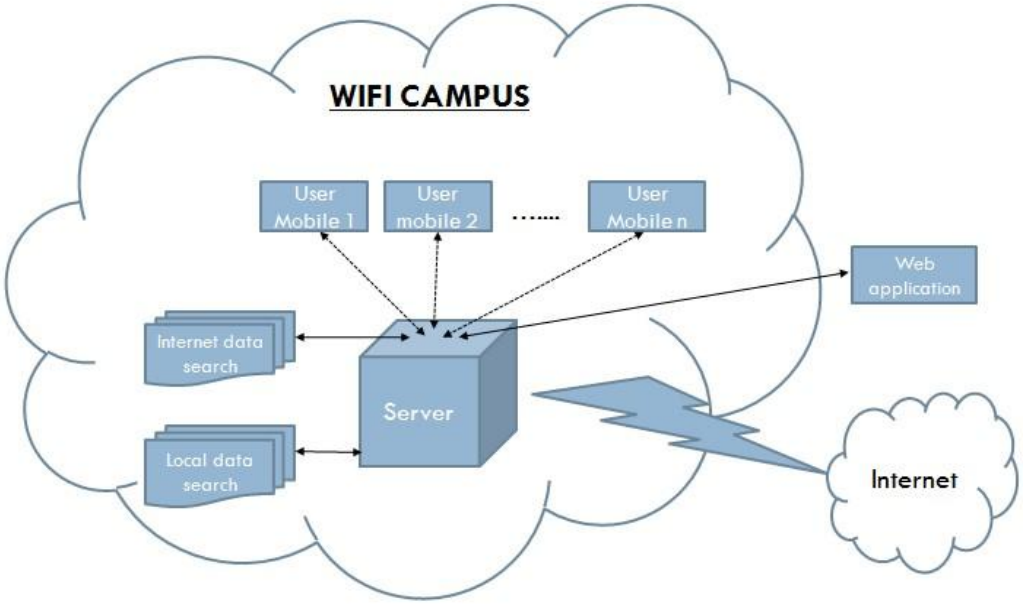
Ένα κλειστό σύστημα του τομέα των ερωτο-απαντήσεων σε συνδυασμό με το ταίριασμα των προτύπων σε συνδυασμό με την χαρτογράφηση SQL έχει προσαρμοστεί στους χρήστες των υπηρεσιών στο σύστημα της παρούσας μελέτης. Η χαρτογράφηση SQL χρησιμοποιείται για να χαρτογραφηθούν τα ερωτήματα μετά την αντιστοίχιση προτύπων, για να ληφθεί το καλύτερο αποτέλεσμα από τη βάση δεδομένων που βρίσκεται σε τοπικό διακομιστή. Οι χρήστες μπορούν να έχουν πρόσβαση στο σύστημα αυτό μέσω διαδικτύου από Android κινητά είτε μέσω ασύρματου δικτύου Wi-Fi. Ο χρήστης μπορεί επίσης να στέλνει και να ελέγχει τα email του και να ενημερώνεται με τα τελευταία νέα από τα RSS feeds. Έτσι, το διαδραστικό σύστημα ερώτησης-απάντησης αφορά την προσπέλαση των τοπικών βάσεων δεδομένων και του διαδικτύου, χωρίς την απαραίτητη σύνδεση στο διαδίκτυο στις συσκευές τους. Αυτή η αρχιτεκτονική που βασίζεται στο Wi-Fi είναι βασικά σχεδιασμένη για τους χώρους της συγκεκριμένης πανεπιστημιούπολης. Το διαδίκτυο παρέχεται στον server και οι χρήστες είναι σε θέση να έχουν πρόσβαση μέσω WIFI. Εδώ, ο κύριος στόχος αυτού του έργου είναι να παρέχει πληροφορίες για τους χρήστες σύμφωνα με τις απαιτήσεις τους.

Το κινητό τηλέφωνο είναι η πιο ευρέως χρησιμοποιούμενη κινητή συσκευή. Κάθε χρήστης κινητού τηλεφώνου μπορεί εύκολα να επικοινωνεί μέσω μηνυμάτων κειμένου SMS (υπηρεσία σύντομων μηνυμάτων). Το σύστημά περιλαμβάνει επίσης τα κινητά τηλέφωνα μέσω των οποίων ο χρήστης θα κάνει ερωτήματα και οι απαντήσεις των ερωτημάτων ανακτώνται και εμφανίζονται πίσω στο χρήστη μέσω του ίδιου κινητού τηλεφώνου. Τα ερωτήματα μπορούν να γίνουν είτε σε SMS ή φυσική γλώσσα.

Το σύστημά αυτό είναι βασικά ένα διαδραστικό σύστημα ερωταποκρίσεων στο οποίο ο χρήστης μπορεί να θέσει ένα ερώτημα στον server στην τοπική βάση δεδομένων ή από οποιοδήποτε είδος

πληροφοριών που μπορούν να ληφθούν από την μηχανή αναζήτησης της Google στο σύστημα. Το σύστημα αποτελείται από δύο κύρια μέρη στα οποία ένα μέρος αντιπροσωπεύει την τοπική βάση δεδομένων και το άλλο την αναζήτηση στο διαδίκτυο.

Τα σημερινά συστήματα έχουν πρόσβαση στην τοπική βάση δεδομένων, χρησιμοποιώντας την επεξεργασία φυσικής γλώσσας, αλλά απαιτούν ένα συγκεκριμένο πρότυπο. Δεν υπάρχει σύστημα που επιτρέπει επίσης πρόσβαση στο διαδίκτυο μέσω του ίδιου συστήματος. Αντιστρόφως, το σύστημα αποτελείται από δύο χαρακτηριστικά, αλλά και τα επιπλέον χαρακτηριστικά, όπως RSS feeds που θα ενημερώνουν το χρήστη χρόνο με το χρόνο.



Εικόνα 15 – Αρχιτεκτονική Συστήματος [45]

Υπάρχουν συστήματα που χρησιμοποιούν την επεξεργασία φυσικής γλώσσας για να απαντήσει στις ερωτήσεις, αλλά δεν είναι τόσο ακριβή στην ανάκτηση της κατάλληλης απάντησης και επίσης απαιτούν συγκεκριμένα προκαθορισμένα ερωτήματα με συγκεκριμένη μορφή. Για τους χρήστες των συστημάτων αυτών πρέπει δίνουν τις ερωτήσεις μόνο με τη συγκεκριμένη μορφή, αλλά, το σύστημά που παρουσιάζεται εδώ έχει πολλές εξελίξεις σε αυτά τα συστήματα. Οι χρήστες μπορούν να εισάγουν τα ερωτήματα με τον τρόπο που επιθυμούν, και στη χειρότερη περίπτωση, λαμβάνοντας υπόψη αν το ερώτημα δεν ταιριάζει με οποιοδήποτε προκαθορισμένο ερώτημα, θα προταθεί στο χρήστη το καλύτερο μήνυμα απάντησης που ταιριάζει με το ερώτημά του. Επίσης, το σύστημα αυτό παρέχει την προσπέλαση των δεδομένων μέσω του διαδικτύου, χωρίς να απαιτείται η σύνδεση στο κινητό τους τηλέφωνο, την αποστολή και λήψη μηνυμάτων, RSS feeds που κρατούν τους χρήστες ενήμερους με τον πραγματικό κόσμο, συμπεριλαμβανομένης βέβαια και της προσπέλασης της βάσης δεδομένων από τον τοπικό διακομιστή.

6 Ερευνητικές μέθοδοι με επεξεργασία Φυσικής Γλώσσας

Στην Ενότητα αυτή παρουσιάζονται οι βιβλιογραφικές μελέτες των ερευνών που έχουν πραγματοποιηθεί πρόσφατα στον τομέα Επεξεργασίας Φυσικής Γλώσσας για την επίλυση διαφορετικών προβλημάτων, στα οποία εστιάζει η σύγχρονη ερευνητική κοινότητα.

6.1 Σύγκριση διαφορετικών μεθόδων για την εξαγωγή γνώμης από Ειδησεογραφικά Άρθρα

Οι συγγραφείς του [47] παρουσιάζουν κάποια ιδιαίτερα χαρακτηριστικά της εξόρυξης απόψεων (opinion mining) σε ειδησεογραφικά άρθρα, σε αντιδιαστολή με την ανάλυση συναισθήματος (sentiment analysis) σε κριτικές προϊόντων. Εισάγουν νέες μεθόδους για τον προσδιορισμό της πολικότητας του συναισθήματος (polarity) σε δηλώσεις που περιέχονται σε ειδησεογραφικά άρθρα. Ως πολικότητα ορίζεται η θετική ή αρνητική άποψη των χρηστών σε κάποια δήλωση ή κριτική τους σε online περιεχόμενο. Οι συγγραφείς επικεντρώνουν την έρευνά τους στην εξόρυξη κριτικών από ειδησεογραφικά άρθρα καθώς, όπως αναφέρουν, συστήματα όπως online καταστήματα, μηχανές κρατήσεων, ιστοσελίδες εύρεσης ταινιών κ.ά., έχουν ήδη αυτοματοποιημένους τρόπους εξόρυξης των κριτικών των χρηστών. Ένα τέτοιο σύστημα, το οποίο παρουσιάζει μία Ανάλυση Απόκρισης Μέσων (Media Response Analysis - MRA) [48], μπορεί να παρέχει ένα ξεχωριστό επαγγελματικό τομέα για τις υπηρεσίες ανάλυσης και έχει ιδιαίτερο ενδιαφέρον για διάφορες επιχειρήσεις και οργανισμούς.

Σήμερα, καταβάλλεται ένα τεράστιο ποσό ανθρώπινης προσπάθειας για την εκτέλεση ενός MRA στις εταιρίες παρακολούθησης μέσων. Αφού συλλεχθούν από τους ανιχνευτές Ιστού (Web Crawlers) τα ειδησεογραφικά κείμενα που περιέχουν ένα συγκεκριμένο αλφαριθμητικό αναζήτησης, οι αναλυτές μέσων διαβάζουν τα κείμενα αυτά και επισημαίνουν τις σχετικές δηλώσεις των καταναλωτών, καταχωρώντας και την πολικότητα της γνώμης. Η περίπτωση αυτή αναφέρεται ως τονικότητα (tonality). Σε αντιδιαστολή με τις κριτικές χρηστών ή ορισμένες συμβολές στα μέσα κοινωνικής δικτύωσης, οι ειδήσεις δεν είναι τόσο υποκειμενικές όσο οι κριτικές αυτές [49]. Επίσης, μπορεί να υπάρχουν διαφορετικές πολικότητες απόψεων σε διάφορα μέρη του ίδιου άρθρου. Επομένως, οι συγγραφείς εστιάζουν στην ταξινόμηση των δηλώσεων αυτών ως αρνητικές ή θετικές.

Αρχικώς, επικυρώνονται οι κατηγορίες των λέξεων που κρίνονται ως πιο σημαντικές, χρησιμοποιώντας διαφορετικές μεθόδους μηχανικής μάθησης (machine learning). Προστίθενται βάρη σημαντικότητας στις τέσσερις κατηγορίες που θεωρούνται κρίσιμες για την τονικότητα [50] [51]. Οι κατηγορίες αυτές είναι τα ρήματα, τα ουσιαστικά, τα επίθετα και τα επιρρήματα. Έχει βρεθεί ότι οι πιο σημαντικές λέξεις στις ειδήσεις δεν ανήκουν στην κατηγορία των επιθέτων, το οποίο είναι η συνήθης υπόθεση που γίνεται στις κριτικές χρηστών. Αντίθετα, στα ειδησεογραφικά άρθρα το πιο σημαντικό ρόλο έχουν τα ουσιαστικά και ρήματα που χρησιμοποιούνται.

Συγκεκριμένα, το παραγόμενο σύστημα υπολογίζει αρχικά ένα βαθμό τονικότητας (Tonality Score – TS), για κάθε λέξη που εμπεριέχεται στις σημαντικές κατηγορίες (επίθετα, ρήματα, ουσιαστικά, επιρρήματα). Επομένως, εξάγονται τέσσερις βαθμοί TS για κάθε δήλωση. Κάθε βαθμός υπολογίζεται ως ο μέσος όρος TS όλων των λέξεων κάθε κατηγορίας. Για τον υπολογισμό του βαθμού TS χρησιμοποιούνται ήδη υπάρχοντες μέθοδοι όπως το chi-square [52], η μέθοδος PMI [52] [53], ή ακόμα και το Association Rule Mining [54]. Επιπροσθέτως, οι συγγραφείς προτείνουν από την έρευνά τους δύο ακόμα μεθόδους, εκείνη της Εντροπίας (Entropy) και της Απόκτησης Γνώσης (Information Gain). Η απόκτηση γνώσης προκύπτει κατόπιν του υπολογισμού της εντροπίας, η οποία ορίζεται ως θετική εφόσον η πιθανότητα εμφάνισης μία λέξης σε θετικές δηλώσεις είναι μεγαλύτερη από τον βαθμό TS, ή αρνητική αν η πιθανότητα εμφάνισης είναι μικρότερη από τον όρο TS.

Οι δύο αυτές μέθοδοι προτείνονται ως συμπληρωματικοί των ήδη υπαρχόντων ώστε να επιτευχθεί σωστότερη ανάλυση στο συγκεκριμένο τομέα των ειδησεογραφικών κειμένων. Η αξιολόγηση του συστήματος αυτού στο οποίο συμπεριλαμβάνονται οι μέθοδοι αυτές, πραγματοποιήθηκε σε ένα σύνολο πραγματικών δεδομένων ανάλυσης αντιδράσεων σε Γερμανικά μέσα (MRA), από την οποία προκύπτει ότι οι μέθοδοι αυτές έχουν καλύτερες επιδόσεις από τις υπάρχουσες προσεγγίσεις και τρόπους εξόρυξης. Τα αποτελέσματα της έρευνας έδειξαν ότι η συγκεκριμένη μέθοδος εντροπίας ξεπερνά τόσο τις άλλες μεθόδους που βασίζονται σε λέξεις, όσο και τις πολύπλοκες προσεγγίσεις που βασίζονται σε συνδυασμούς λέξεων.

6.2 Παραγωγή ερωτημάτων SQL χρησιμοποιώντας συντακτικές εξαρτήσεις και μεταδεδομένα NLP

Οι διεπαφές διασύνδεσης βάσεων δεδομένων με τη φυσική γλώσσα (Natural Language Interface to DataBases – NLIDB) προτείνουν ένα μεγάλο εύρος χειρωνακτικής εργασίας για προδιαγραφές

γραμματικής και σημασιολογίας δεδομένων. Αντιθέτως, το έργο της απάντησης σε ερωτήσεις, της μετάφρασης της φυσικής γλώσσας (Natural Language – NL) σε μορφή αναγνώσιμη από υπολογιστές με αυτοματοποιημένο τρόπο είναι αρκετά περίπλοκο, καθώς δεν δύναται να καθοριστούν απόλυτα όλοι οι απαραίτητοι κανόνες.

Οι συγγραφείς του [55] προτείνουν μία λύση στο συγκεκριμένο πρόβλημα, εξάγοντας τα κατάλληλα SQL ερωτήματα των οποίων η δομή και τα συστατικά ταιριάζουν με NL έννοιες (που εκφράζεται ως λέξεις) και εξαρτήσεις της γραμματικής. Η ιδέα τους στηρίζεται στο πως και που μπορεί να βρεθεί η αντιστοίχιση της πληροφορίας, όπως για παράδειγμα με την αξιοποίηση υπάρχουσας γνώσης των βάσεων, γνωστά ως μεταδεδομένα. Η αντιστοίχιση που προκύπτει μεταξύ των μεταδεδομένων και των λέξεων επιτρέπει τη κατασκευή ενός συνόλου από στοιχεία ερωτημάτων (ρήτρες). Αυτά τα στοιχεία συνδυάζονται μεταξύ τους μέσω ενός έξυπνου αλγόριθμου, λαμβάνοντας υπόψη και την αρχική δομή της φυσικής γλώσσας. Η έρευνα αυτή εστιάζει στη μεταγλώττιση ερωτημάτων Φυσικής Γλώσσας σε SQL ερωτήματα αξιοποιώντας την MySQL τόσο για την κατασκευή υποθέσεων όσο και για την επαλήθευση του ζητούμενου κατά την απάντηση στα ερωτήματα που τίθενται. Χρησιμοποιούνται γλωσσικές εξαρτήσεις και μεταδεδομένα για την κατασκευή ενός συνόλου από πιθανά ερωτήματα SELECT και WHERE. Επίσης, δημιουργήθηκε ένας αλγόριθμος αντιστοίχισης εξαρτήσεων μεταξύ στοιχείων της φυσικής γλώσσας και της δομής της SQL, ο οποίος επιτρέπει την κατασκευή μίας σειράς από πιθανά ερωτήματα τα οποία απαντούν στην δεδομένη ερώτηση. Για την απεικόνιση των λεξικών σχέσεων χρησιμοποιήθηκαν συσχετισμοί εξαρτήσεων κειμένου, και συγκεκριμένα η απεικόνιση Stanford Dependencies Collapsed (SDC) [56]. Το πρώτο βήμα πριν να παραχθούν τα πιθανά ερωτήματα μίας ερώτησης q είναι η κατασκευή των στοιχείων S, F, και W που αντιστοιχούν στις δομές SELECT, FROM και WHERE, ξεκινώντας από ένα κατάλογο εξαρτήσεων SDCq. Ο κατάλογος αυτός θα πρέπει να είναι (α) προ-επεξεργασμένος χρησιμοποιώντας τεχνικές κλαδέματος, ανακοπής και προσθήκης συνώνυμων λέξεων, (β) αναλυμένος ώστε να έχουν δημιουργηθεί τα σύνολα στελεχών που χρησιμοποιούνται για τα S, W και (γ) τροποποιημένος / εκκαθαρισμένος ώστε να διατηρείται η εξάρτηση που θα χρησιμοποιηθεί σε ενδεχόμενα αναδρομικά βήματα, ώστε να παραχθούν ένθετα ερωτήματα. Το επόμενο στάδιο εκμεταλλεύεται τα μεταδεδομένα για την παραγωγή στοιχείων FROM τα οποία εμπλουτίζονται με σημασιολογικές ενώσεις. Στο τελευταίο στάδιο συνδυάζονται όλα τα στοιχεία (SELECT, FROM, WHERE) για να παραχθούν όλα τα πιθανά SQL ερωτήματα που μπορούν να δώσουν απάντηση στο ερώτημα.

Ο συγκεκριμένος αλγόριθμος μπορεί να χρησιμοποιηθεί αναδρομικά για την επίλυση σύνθετων ερωτημάτων που απαιτούν εμφωλευμένες δομές SELECT. Επιπροσθέτως, προτείνει ένα σύστημα αξιολόγησης των απαντήσεων χρησιμοποιώντας βάρη, ώστε να βαθμονομήσει τα αποτελέσματα βάσει της πιθανότητας να είναι ορθά. Το βάρος αυτό προκύπτει με απλό τρόπο, μετρώντας από πόσα στελέχη προέκυψε η ρήτρα (ερώτημα). Στις περιπτώσεις όπου γίνεται συνδυασμός των στοιχείων, το συνολικό βάρος υπολογίζεται από το άθροισμα των επιμέρους στοιχείων που απαρτίζουν το τελικό ερώτημα και στο τελικό στάδιο κατατάσσεται με σειρά πιθανότητας τα αποτελέσματα, από το πιο πιθανό στο πιο απίθανο. Σωστή απάντηση θεωρείται εκείνη με το υψηλότερο βάρος.

Παρόμοια αποτελέσματα έχουν βρεθεί χρησιμοποιώντας διάφορες τεχνικές, όπως: την προσέγγιση βάσει της σημασίας της γραμματικής των λέξεων, η οποία αποδίδεται από κάποιον εμπειρογνώμονα [57], τον εμπλουτισμό της πληροφορίας που βρίσκεται στα παραγόμενα ζεύγη ερωτημάτων [58], καθώς και εφαρμόζοντας κάποιους κατά περίπτωση κανόνες μέσω ειδικού σημασιολογικού μεταγλωττιστή [59] [60]. Σε συγκρίσει με τις τεχνικές αυτές, οι συγγραφείς αναφέρουν ότι το σύστημά τους δεν απαιτεί καμίας μορφής παρέμβαση, καθώς τα μεταδεδομένα που βρίσκονται στη βάση τους εμπεριέχουν ήδη όλη την απαιτούμενη πληροφορία. Το αποτέλεσμα αυτό είναι ενθαρρυντικό, δεδομένου ότι συγκρίνεται ευνοϊκά με την τρέχουσα τεχνογνωσία. Συγκεκριμένα, με το σύστημα Precise [58] εμφανίζει 100% ακρίβεια (Precision) και 77% ακρίβεια (Recall), ενώ με το σύστημα Krisp [59] παρουσιάζει 94% ακρίβεια (Precision) και 78% ανάκληση (Recall). Το σύστημα αυτό προσφέρει απαντήσεις στο 88% των σεναρίων που αποτυπώνονται, με ακρίβεια 81%. Λαμβάνοντας υπόψη ότι η σωστή απάντηση βρίσκεται μέσα στις πρώτες τρεις, τότε η ακρίβεια αυξάνεται στο 95%, και λαμβάνει 99% στα πρώτα δέκα (10) αποτελέσματα που έχουν ταξινομηθεί. Η ακρίβεια του συστήματος μπορεί να αυξηθεί χρησιμοποιώντας ένα σύστημα αυτόματης εκπαίδευσης και ταξινόμησης [61], το οποίο μπορεί να μαθαίνει τις σωστές απαντήσεις και να τις επιστρέφει απευθείας στα πρώτα αποτελέσματα.

6.3 Αυτόματη ανάκτηση πόρων μέσω NLP αιτημάτων του χρήστη

Στη βιβλιογραφία υπάρχουν προσεγγίσεις οι οποίες εξετάζουν την ανάκτηση πόρων που απευθύνεται στους τελικούς χρήστες [62] [63]. Σκοπός των προσεγγίσεων αυτών είναι η αντικατάσταση της επίσημης έκφρασης με απλά αιτήματα κατασκευασμένα με τη φυσική γλώσσα. Τα αιτήματα αυτά διαμορφώνονται μέσω διαφορετικών διαδικασιών γλωσσικής ανάλυσης, όπως την κατάτμηση κειμένου, την αφαίρεση μη σχετικών λέξεων, της αποκοπής και γραμματικής διόρθωσης [62], της ανάκτησης πληροφορίας [63], και την αντιστοίχιση της δομής του αιτήματος με προκαθορισμένα μοντέλα μέσω λέξεων-κλειδιά. Οι μέθοδοι αυτές επιχειρούν να υπολογίσουν την ομοιότητα μεταξύ των όρων που λαμβάνονται από το αίτημα του χρήστη και των διαθέσιμων πόρων του συστήματος. Οι διαφορετικές αυτές προσεγγίσεις παρουσιάζουν μειονεκτήματα καθώς μερικές από αυτές χρησιμοποιούν κοινές οντολογίες που μπορεί να μην αντιστοιχούν στο λεξιλόγιο που χρησιμοποιεί ο τελικός χρήστης. Επίσης, κάποιες από αυτές τις προσεγγίσεις περιορίζουν την χρήση της φυσικής γλώσσας στα απλά αιτήματα ενώ δεν επιτρέπουν την εξαγωγή μίας ροής υπηρεσιών από την αίτηση. Αντίθετα, η προσέγγιση που παρουσιάζεται στο [64] επιτρέπει την παραγωγή ευέλικτων αιτήσεων στα Αγγλικά και την κατασκευή μίας γενικής ροής ελέγχου από λέξεις-κλειδιά από το αίτημα που δημιουργεί ο χρήστης. Παράλληλα, η έρευνα που πραγματοποιείται στο [64] προτείνει την χρήση κοινότυπων ταξινομημένων εννοιών για την αύξηση της δυναμικότητας στην προσαρμογή των αλλαγών που πραγματοποιούνται στο λεξιλόγιο του χρήστη. Για την προσέγγιση αυτή υλοποιήθηκαν διάφορες τεχνικές NLP οι οποίες επιτρέπουν την προηγμένη ανάλυση των αιτημάτων που γίνονται σε φυσική γλώσσα [65]. Οι συγγραφείς προτείνουν μία νέα προσέγγιση που υποστηρίζει την αυτοματοποιημένη ανάκτηση πόρων σε συγκλίνοντα περιβάλλοντα [65], στοχεύοντας στην υποστήριξη των τελικών χρηστών να προσδιορίζουν τα αιτήματά τους σε φυσική γλώσσα, χωρίς να απαιτείται ειδικευμένη τεχνική ή προγραμματιστική γνώση. Στην εργασία τους αυτή, ορίζουν τρεις βασικές οντότητες: τον χρήστη (user – U), τους πόρους (resources – R), και τις περιγραφές αυτών των πόρων και των λειτουργικών παραμέτρων τους, τις ετικέτες (tags – T). Αξίζει να σημειωθεί ότι το συγκεκριμένο σύστημα εκμεταλλεύεται το αίτημα του χρήστη ώστε να αντιστοιχίσει τους πόρους με τις κατάλληλες ετικέτες, καθώς δεν είναι κάτι που γίνεται συνήθως από τους χρήστες. Επομένως, οι χρήστες παραθέτουν ετικέτες στους πόρους, δημιουργώντας μία τριπλή συσχέτιση μεταξύ του χρήστη, του πόρου και της ετικέτας.

Η τριπλή αυτή συσχέτιση καθορίζει την σημασιολογική περιγραφή στην πρόταση που παραθέτει η έρευνα του [64], η οποία μπορεί να οριστεί μέσω ενός συνόλου σχολιασμών. Συγκεκριμένα, εντοπίζονται τρεις διαφορετικές ταξινομήσεις λέξεων από το αίτημα του χρήστη: (α) οι λέξεις Ελέγχου, οι οποίες χρησιμεύουν στον καθορισμό της τελικής ροής των επιθυμητών υπηρεσιών, (β) οι Λειτουργικές λέξεις, που ταξινομούνται σε κατηγορίες Συμπεριφοράς και Εισόδου / Εξόδου, και αντιπροσωπεύουν τις λειτουργικές παραμέτρους που χρησιμοποιούνται για τη διαδικασία ανάκτησης, και (γ) τις Μη-Λειτουργικές λέξεις, που αντιπροσωπεύουν μη λειτουργικές ιδιότητες των πόρων. Το αίτημα του χρήστη εκφρασμένο σε φυσική γλώσσα παρέχεται ως είσοδος στο σύστημα, ενώ ο σκοπός της μονάδας είναι η αναγνώριση και επιλογή των κατάλληλων όρων που παράγονται από το μετά-μοντέλο του αιτήματος του χρήστη. Οι όροι αυτοί κατατάσσονται σε διάφορες κατηγορίες και χρησιμοποιούνται για την ανάκτηση των πόρων. Οι πόροι αναπαριστούν ένα υποσύνολο υφιστάμενων πραγματικών υπηρεσιών (ιστού και τηλεπικοινωνιών), σύμφωνα με το μοντέλο Πόρων.

Βάσει των αποτελεσμάτων των δοκιμών που πραγματοποιήθηκαν από την παρούσα έρευνα, εκτιμάται ότι η προσέγγιση αυτή αποδίδει καλά σε γενικές γραμμές, λαμβάνοντας υπόψη ξεχωριστά την πολυπλοκότητα κάθε επιμέρους μονάδας. Ωστόσο, ορισμένοι κανόνες και φίλτρα που συστάθηκαν για την αναγνώριση συγκεκριμένων χαρακτηριστικών και κατηγοριών των αιτημάτων των χρηστών μπορούν να ενισχυθούν, λαμβάνοντας υπόψη περισσότερες περιπτώσεις ώστε να αυξηθεί το επίπεδο ακρίβειας των αποτελεσμάτων.

6.4 Χρήση της NLP για την βελτίωση της κατηγοριοποίησης Εγγράφων με Συσχετιζόμενα Δίκτυα

Η ταξινόμηση μεγάλων συνόλων εγγράφων σε μία ιεραρχική δομή κατηγοριών, όπως για παράδειγμα τα άρθρα της Wikipedia, επιτρέπει στους χρήστες να βρίσκουν ευκολότερα τις κατάλληλες πληροφορίες. Ωστόσο, η αλλαγή της φύσης των εν λόγω δεδομένων προσφέρει ιδιαίτερες προκλήσεις

όσον αφορά τη δημιουργία και τη διατήρηση αυτής της ιεραρχίας. Η διατήρηση αυτής της ιεραρχίας μπορεί να επιτευχθεί με τη χρήση ενός συνδεδεμένου δικτύου (associative network). Ως συνδεδεμένο δίκτυο ορίζεται ένα μοντέλο διασυνδεδεμένης γλώσσας, το οποίο δύναται να κατηγοριοποιεί αυτόματα βιβλιοθήκες. Τα συνδεδεμένα δίκτυα μοντελοποιούνται ως γράφοι, όπου κάθε κόμβος αντιπροσωπεύει μία παρατήρηση, όπως την εμφάνιση μίας συγκεκριμένης έννοιας, ενώ κάθε ακμή αντιστοιχεί στην σύνδεση μεταξύ τέτοιων παρατηρήσεων. Οι ακμές με τα μεγαλύτερα βάρη εκπροσωπούν παρατηρήσεις που συνδέονται πιο στενά μεταξύ τους. Χρησιμοποιώντας την ίδια βάση διασύνδεσης [66] όπως στα νευρωνικά δίκτυα, κάθε κόμβος μπορεί να ενεργοποιηθεί, εξαπλώνοντας το σήμα εισόδου στους συνδεδεμένους κόμβους, ανάλογα με το βάρος της ακμής τους. Επομένως, οι ακμές με μεγαλύτερο βάρος εξαπλώνουν το σήμα πιο έντονα από εκείνες με μικρότερο βάρος.

Αντίστοιχα, η μέθοδος συχνότητας όρου προς την αντίστροφη συχνότητα εγγράφου TD-IDF (Term Frequency – Inverted Document Frequency) [67] καθορίζει τη σημαντικότητα μίας λέξης σε ένα έγγραφο ή σε ένα σύνολο εγγράφων. Οι συγγραφείς του [68], προσπάθησαν να αποδείξουν ότι τα συνδεδεμένα δίκτυα αποδίδουν καλύτερα από ότι η βάση TF-IDF στην κατηγοριοποίηση εγγράφων, ενώ παράλληλα τα αποτελέσματα κατηγοριοποίησης μπορούν να βελτιωθούν περαιτέρω κάνοντας χρήση τεχνικών NLP για την αποσαφήνιση και την ανίχνευση σχετικών στοιχείων του κειμένου. Στην έρευνα αυτή μελετήθηκε η χρήση τεχνικών επεξεργασίας φυσικής γλώσσας σε συνδυασμό με Συνδεδεμένα Δίκτυα για την κατηγοριοποίηση κειμένων και τη σύγκριση των αποτελεσμάτων αυτών με μία βάση TF-IDF. Στο σύστημα που παρουσιάζεται στο [68], η εκπαίδευση πραγματοποιείται με τη δημιουργία μίας βιβλιοθήκης εκμάθησης που αποτελείται από τριάντα (30) χειρωνακτικά επιλεγμένα άρθρα της Wikipedia, κάθε ένα από τα οποία συνδέεται στενά με ακριβώς ένα άλλο από τα άρθρα αυτά και δεν σχετίζεται καθόλου με τα υπολειπόμενα είκοσι οκτώ (28). Κατά αυτόν τον τρόπο, δημιουργούνται δεκαπέντε (15) ζεύγη συσχετισμένων άρθρων. Τα άρθρα επιλέχθηκαν από έγγραφα που ανήκουν στην ίδια κατηγορία με το test set, αλλά από διαφορετικές υποκατηγορίες. Τα συνδεδεμένα δίκτυα δοκιμάστηκαν χωρίς τη χρήση NLP, με ένα σύστημα προσθήκης γλωσσικής ετικέτας και με την πλήρη ανάλυση της γλώσσας, τα αποτελέσματα των οποίων συγκρίθηκαν μεταξύ τους αλλά και με τις συχνότητες TF-IDF. Για τις περιπτώσεις δοκιμών με NLP χρησιμοποιήθηκε ο Stanford Natural Language Parser [69] και ο Stanford Part-Of-Speech Tagger [70] για την ενίσχυση της ενεργοποίησης από λέξεις-κλειδιά σε κάθε πρόταση. Με απλά λόγια, αυτές οι τεχνικές παρουσιάζουν τους κόμβους του δικτύου που ενεργοποιούνται και το ποσοστό ενεργοποίησής τους. Σε όλες τις περιπτώσεις χρησιμοποιήθηκε το Princeton WordNet [71] [72] για τη διασύνδεση των λέξεων και των εννοιών τους. Στην συγκεκριμένη έρευνα ένα έγγραφο ενεργοποιεί κάθε έννοια λέξης που βρίσκεται στο έγγραφο. Έννοιες που έχουν μεγαλύτερη συσχέτιση με το έγγραφο ενεργοποιούνται περισσότερο από άλλες, εξαπλώνοντας έτσι περισσότερο και ισχυρότερα στον παραγόμενο γράφο. Η ποιότητα του μοντέλου ενεργοποίησης, δηλαδή οι κόμβοι του δικτύου που ενεργοποιούνται βάσει μίας δεδομένης εισόδου, είναι σημαντική για την παραγωγή βέλτιστων αποτελεσμάτων.

Η επεξεργασία φυσικής γλώσσας μπορεί να προ-φιλτράρει την πληροφορία πριν εισαχθεί στο διασυνδεδεμένο δίκτυο, φιλτράροντας τις ανεπιθύμητες παρατηρήσεις και επιλέγοντας εξ αρχής τα σχετικά δεδομένα βάσει της δομής της πρότασης, βελτιώνοντας έτσι τα αποτελέσματα. Λόγω των συσχετισμών που παράγονται από το δίκτυο, οι έννοιες που είναι σχετικότερες με το κείμενο ταξινομούνται πρώτες, ακόμα και αν δεν αναφέρονται μέσα στο ίδιο το κείμενο ή αναφέρονται μόνο μερικές φορές. Σύμφωνα με την έρευνα αυτή, τα συνδεδεμένα δίκτυα παράγουν καλύτερο συνειρμό του κειμένου από ότι η τεχνική TF-IDF. Αντίθετα, η χρήση NLP σε συνδυασμό με την τεχνική TF-IDF παρήγαγε χειρότερα αποτελέσματα, καθώς οι σχετικές λέξεις του κειμένου ενεργοποίησαν πολύ μικρότερο ποσοστό κόμβων, έχοντας ως συνέπεια την αντιστοίχιση με λιγότερα άρθρα, αν και συσχετιζόμενα. Αντίθετα, τα συνδεδεμένα δίκτυα επωφελούνται σαφώς από την χρήση NLP στην αντιστοίχιση λέξεων και κόμβων του γράφου, κερδίζοντας 6% ακρίβεια κατά τη χρήση του Stanford Part-Of-Speech Tagger και 3% χρησιμοποιώντας τον Stanford Natural Language Parser.

6.5 Coh-Metrix

Το COH-Metrix είναι ένα αυτοματοποιημένο εργαλείο που παρέχει γλωσσικούς δείκτες για το κείμενο και τον λόγο [73]. Το COH-Metrix αναπτύχθηκε για να ικανοποιήσει τρεις πρακτικές ανάγκες. Πρώτον, όταν το ερευνητικό πρόγραμμα του COH-Metrix ξεκίνησε το 2002, δεν υπήρχαν άμεσα διαθέσιμα εργαλεία που να παρέχουν μια σειρά από δείκτες σε λέξεις ή κείμενα. Για παράδειγμα, εάν

ένας ερευνητής χρειάζεται τις τιμές συχνότητας μίας λέξης ως προς τις λέξεις ή φράσεις σε ένα κείμενο, μπορεί να έβρισκε ένα διαθέσιμο εργαλείο (αν και δύσκολο να βρεθεί). Αλλά θα έπρεπε να χρησιμοποιηθεί ένα άλλο εργαλείο για την μέτρηση της συγκεκριμενοποίησης της λέξης, της οικειότητας, της εικονοποίησης, της συντακτικής πολυπλοκότητας, και ούτω καθεξής. Με άλλα λόγια, δεν υπήρχαν γλωσσικά εργαλεία ικανά να παρέχουν ένα ευρύ φάσμα μέτρων για τη γλώσσα και την ομιλία. Δεύτερον, τα παραδοσιακά μέτρα της δυσκολίας του κειμένου, που αναφέρονται ως αναγνωσιμότητα, ήταν παρωχημένα λόγω της ωρίμανσης των γνώσεών μας σχετικά με το κείμενο και τον λόγο [74] [75] [76]. Υπήρξε μια αυξανόμενη αναγνώριση ενός αριθμού παραγόντων που συμβάλλουν στη δυσκολία του κειμένου που δεν αναγνωρίζονται από τα παραδοσιακά μέτρα της ανάγνωσης του κειμένου. Τρίτον, δεν υπήρχαν αυτοματοποιημένα μέτρα για τη συνοχή του κειμένου. Αν και η αναγνώριση της σημασίας της συνοχής άκμασε στη δεκαετία του '80 και του '90 [77] [78] [79] [80] [81], δεν υπήρχαν αντικειμενικά, εφαρμοζόμενα μέτρα για τη συνοχή του κειμένου. Έτσι, με τον πρωταρχικό στόχο της παροχής πιο κατατοπιστικών μέτρων πολυπλοκότητας του κειμένου, λαμβάνοντας ιδιαίτερα υπόψη τη συνοχή του κειμένου, οι συγγραφείς ξεκίνησαν το 2002 να αναπτύξουν το COH-Metrix (αρχικά χρηματοδοτήθηκε από ένα Ινστιτούτο Επιστημών της Εκπαίδευσης).

Τα διάφορα μέτρα της αναγνωσιμότητας συσχετίζονται σε μεγάλο βαθμό μεταξύ τους διότι βασίζονται στις ίδιες δομές: τη δυσκολία των μεμονωμένων λέξεων και την πολυπλοκότητα των επιμέρους προτάσεων στο κείμενο. Ωστόσο, ο τρόπος με τον οποίο αυτές οι δομές θα συγκεκριμενοποιηθούν και οι στατιστικές υποθέσεις διαφέρουν κάπως μεταξύ των μετρήσεων αναγνωσιμότητας. Τα μέτρα αναγνωσιμότητας που βασίζονται στα χαρακτηριστικά των προτάσεων και λέξεων (δηλαδή, μετρήσει συνήθως μήκους) ισχύουν ως δείκτες της δυσκολίας του κειμένου. Όταν οι λέξεις περιέχουν περισσότερα γράμματα ή συλλαβές, τείνουν να χρησιμοποιούνται λιγότερο συχνά σε μια γλώσσα. Οι αναγνώστες πρέπει να έχουν μεγαλύτερη έκθεση στη γλώσσα και το κείμενο, προκειμένου να αντιμετωπίσουν λιγότερο συχνές λέξεις και να γνωρίζουν τη σημασία τους. Σαφώς, μια προϋπόθεση για την κατανόηση είναι να γνωρίζουν το νόημα των λέξεων σε ένα κείμενο. Αντίστοιχα, στο βαθμό που η πρόταση περιέχει περισσότερες λέξεις, υπάρχει μεγαλύτερη πιθανότητα η φράση να είναι πιο περίπλοκη συντακτικά. Οι αναγνώστες που έχουν μικρότερη έκθεση στη γλώσσα και το κείμενο (δηλαδή, οι νεότεροι και λιγότερο έμπειροι αναγνώστες) αντιμετωπίζουν μεγαλύτερες προκλήσεις στη κατανόηση συντακτικά σύνθετων προτάσεων.

Οι συνεκτικές νύξεις βοηθούν τον αναγνώστη να κατανοήσει τις συνδέσεις μεταξύ των προτάσεων και παραγράφων. Αυτό, με τη σειρά του, διευκολύνει την κατανόηση των λέξεων και προτάσεων, και ενισχύει την πλήρη κατανόηση του κειμένου από τον αναγνώστη. Πολλές και ποικίλες μελέτες έχουν δείξει ότι οι συνεκτικές νύξεις στο κείμενο διευκολύνουν την κατανόηση της ανάγνωσης και βοηθούν τους αναγνώστες στην κατασκευή πιο συνεκτικών νοητικών αναπαραστάσεων του περιεχομένου του κειμένου [82] [83] [84]. Για το λόγο αυτό, η κύρια βάση δεικτών που προβλέπονται από COH-Metrix αξιολογεί τη συνοχή του κειμένου.

Το COH-Metrix παρέχει δείκτες για μια σειρά από μέτρα που ενδεχομένως σχετίζονται με την κατανόηση του αναγνώστη. Αυτά τα μέτρα περιλαμβάνουν την αιτιότητα, όπως βαθμολογίες επίπτωσης για αιτιακά ρήματα, εκ προθέσεως ρήματα, και τα αιτιώδη σώματα (π.χ. σύνδεσμοι), και οι αναλογίες των αιτιωδών σωμάτων προς τα αιτιακά ρήματα. Υπάρχουν μέτρα επικάλυψης των ρημάτων. Οι δείκτες αυτοί είναι ενδεικτικοί του βαθμού στον οποίο τα ρήματα (τα οποία έχουν εμφανείς συνδέσεις με δράσεις, γεγονότα και καταστάσεις) επαναλαμβάνονται κατά μήκος του κειμένου. Υπάρχουν μέτρα χρονικής συνοχής και λογικής συνοχής. Αυτά και άλλα μέτρα αντανακλούν στοιχεία του κειμένου που είναι πιθανό να υποστηρίξουν την νόηση του αναγνώστη σε ένα συνεκτικό μοντέλο.

Έτσι, το COH-Metrix, σε αντίθεση με τα παραδοσιακά μέτρα ανάγνωσης του κειμένου, έχει τη δυνατότητα να προσφέρει μια πιο ολοκληρωμένη εικόνα των πιθανών προκλήσεων που ενδέχεται να αντιμετωπίσει ο αναγνώστης, καθώς και τα πιθανά ικριώματα που μπορεί να προσφέρονται από το κείμενο. Το COH-Metrix υποκινείται από τις θεωρίες του λόγου και την κατανόηση κειμένου. Τέτοιες θεωρίες περιγράφουν την κατανόηση σε πολλαπλά επίπεδα, από τα ρηχά, που βασίζονται στη κατανόηση του κειμένου, έως βαθύτερα επίπεδα κατανόησης που ενσωματώνουν πολλαπλές ιδέες στο κείμενο και φέρουν πληροφορίες που αναπτύσσουν τις ιδέες του κειμένου χρησιμοποιώντας την γνώση του κόσμου και του συγκεκριμένου τομέα [85]. Το COH-Metrix αξιολογεί τις προκλήσεις που ενδέχεται να προκύψουν στα επίπεδα λέξεων και φράσεων. Επιπλέον, είναι σε θέση να εκτιμήσει βαθύτερα επίπεδα της γλώσσας όσον αφορά τη συνοχή και το μοντέλο κατάστασης. Με αυτόν τον τρόπο, τείνει να παρέχει

τη δυνατότητα να εκτιμήσει πόσο καλά ο αναγνώστης θα κατανοήσει ένα κείμενο σε βαθύτερα επίπεδα της γνωστικής λειτουργίας.

Ο στόχος των συγγραφέων ήταν να δώσουν έμφαση πέρα από τα παραδοσιακά μέτρα της αναγνωσιμότητας που εστιάζουν στα επιφανειακά χαρακτηριστικά των κειμένων, τα οποία με τη σειρά τους έχουν την τάση να επηρεάζουν κυρίως την επιφανειακή κατανόηση. Πράγματι, η επικύρωση των παραδοσιακών αλγορίθμων αναγνωσιμότητας χρησιμοποιώντας τα χαρακτηριστικά λέξεων και φράσεων (π.χ. αριθμός των γραμμάτων ή λέξεων) έχει γίνει σχεδόν αποκλειστικά με τη χρήση εκτιμήσεων που βασίζονται κυρίως στην επιφανειακή κατανόηση (π.χ. δοκιμασίες cloze). Το COH-Metrix μπορεί να χρησιμοποιηθεί για να κατανοηθούν καλύτερα οι διαφορές μεταξύ των κειμένων και να διερευνηθεί ο βαθμός στον οποίο τα χαρακτηριστικά της γλώσσας και του λόγου μπορούν επιτυχώς να διακρίνουν τύπους κειμένου. Το COH-Metrix μπορεί επίσης να χρησιμοποιηθεί για την ανάπτυξη και τη βελτίωση των προσεγγίσεων της επεξεργασίας της φυσικής γλώσσας.

7 Συμπεράσματα

Αν και η έρευνα και ανάπτυξη στον τομέα της Επεξεργασίας του Φυσικού Λόγου συνεχίζεται για πάνω από εξήντα (60) χρόνια, τα παραγόμενα συστήματα εξακολουθούν να έχουν πολύπλοκες σχεδιαστικές απαιτήσεις. Στη σημερινή εποχή επικρατεί μία πληθώρα από μοντέλα και αλγόριθμους. Προφανώς, τα συστήματα NLP δεν μπορούν να τελειοποιηθούν λόγω της πολυπλοκότητας της ανθρώπινης γλώσσας, επομένως είναι δύσκολο να καταγραφεί πλήρως η γνώση της γλωσσολογίας ώστε να παραχθεί 100% ακρίβεια επεξεργασίας. Για παράδειγμα, αν και εκατοντάδες εταιρίες έχουν προβεί στην αντικατάσταση των τηλεφωνικών κέντρων τους με αυτοματοποιημένες υπηρεσίες ομιλίας μέσω λογισμικού, ευαίσθητες τηλεφωνικές υπηρεσίες όπως είναι η το Κέντρο Άμεσης Δράσης (100) εξακολουθούν να προσφέρονται μέσω κατάλληλα εκπαιδευμένου προσωπικού, ακόμα και σε παγκόσμιο επίπεδο. Επιπροσθέτως, δεν διαφαίνεται να αντικατασταθούν κάποια στιγμή σύντομα με αυτοματοποιημένες λύσεις, λόγω του ευαίσθητου χαρακτήρα τους, καθώς και της εξειδικευμένης γνώσης που απαιτείται στο χειρισμό των επειγόντων περιστατικών.

Συστήματα αναγνώρισης και επεξεργασίας ομιλίας, χρησιμοποιώντας την ομιλία αυτή ως ερωτήματα, χρησιμοποιούνται ήδη και βρίσκονται σε αρκετά εμπορική άνθιση. Ενδεικτικά, αναφέρονται οι εφαρμογές Google Now [86], Google Voice Search [87], Apple Siri [88], και Ask Ziggy [89] ως οι πιο ευρέως διαδεδομένες εμπορικές εφαρμογές μέχρι στιγμής, οι οποίες χρησιμοποιούνται από πλήθος κόσμου σε καθημερινή βάση, κυρίως μέσω κινητών συσκευών. Όμως, τα τρέχοντα συστήματα αναγνώρισης και επεξεργασίας ομιλίας εξακολουθούν να απαιτούν επιπλέον προσαρμογές. Κάποια από αυτά δεν μπορούν να κατανοήσουν δύσκολες ή βαριές προφορές, εγγενή προβλήματα λόγου και ομιλίας (π.χ. τραύλισμα), ή χαμηλές φωνές. Παράλληλα, τα περισσότερα από τα συστήματα αυτά είναι ιδιόκτητα και υπόκεινται σε πνευματικά δικαιώματα και περιορισμούς χρήσης από τρίτους. Συνεπώς, είναι δύσκολο να δοθεί η άδεια λειτουργίας τους σε άλλες εφαρμογές εκτός της ιδιοκτήτριας εταιρίας ή οργανισμού, αυξάνοντας το κόστος υλοποίησης ολοκληρωμένων εφαρμογών που εμπεριέχουν αυτές τις λειτουργίες.

Το μέλλον της χρήσης εργαλείων NLP έγκειται στο κατά πόσο η κοινότητα της Πληροφορικής έχει σκοπό την ανάπτυξη τέτοιων προγραμμάτων χρησιμοποιώντας τεχνικές Ανοιχτού Πηγαίου Κώδικα (Open Source), ώστε να διευρυνθεί και η συμμετοχή περισσοτέρων εταιριών και οργανισμών στον τομέα αυτό. Παρόλα αυτά, οι πιθανότητες ύπαρξης διεπαφών που επιτρέπουν την αλληλεπίδραση ανθρώπου – υπολογιστή με απόλυτα φυσικό λόγο φαίνονται μικρές, παρά τις εξελίξεις που έχουν γίνει στο χώρο. Αυτό συμβαίνει καθώς είναι λίαν δύσκολο να εξαχθεί τέτοια γνώση από ένα σύστημα, όπως εξάγεται από τον ανθρώπινο εγκέφαλο, σε συνδυασμό με άλλες παραμέτρους όπως τη διαφοροποίηση που υπάρχει στο περιβάλλον και τη κουλτούρα ενός λαού. Μακροπρόθεσμα, η συνεργασία μεταξύ του συστήματος και του φυσικού χρήστη θα βοηθήσει εκατέρωθεν στην εκμάθηση τρόπων φυσικής επικοινωνίας και παραγωγής γνώσης.

Βιβλιογραφία

[1] Σ. Μαρκαντωνάτου και Γ. Μαϊστρος, «Επεξεργασία Φυσικής Γλώσσας - Εισαγωγή,» [Ηλεκτρονικό]. Available: http://courses.dbnet.ntua.gr/fsr/2524/lecture1.pdf. [Πρόσβαση 15 01 2015].

[2] Wikipedia, «Γνώση,» Wikipedia, [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/%CE%93%CE%BD%CF%8E%CF%83%CE%B7. [Πρόσβαση 28 02 2015].

[3] Wikipedia, «Υπολογιστική Γλωσσολογία,» Wikipedia, [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/%CE%A5%CF%80%CE%BF%CE%BB%CE%BF%CE%B3%CE%B9%CF%83%CF%84%CE%B9%CE%BA%CE%AE_%CE%B3%CE%BB%CF%89%CF%83%CF%83%CE%BF%CE%BB%CE%BF%CE%B3%CE%AF%CE%B1. [Πρόσβαση 01 03 2015].

[4] Google, «Google Translate,» Google, [Ηλεκτρονικό]. Available: https://translate.google.com/. [Πρόσβαση 01 03 2015].

[5] World Lingo, «World Lingo,» World Lingo, [Ηλεκτρονικό]. Available: http://www.worldlingo.com/. [Πρόσβαση 01 03 2015].

[6] Ι. Βλαχαβάς, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου, «Τεχνητή Νοημοσύνη - Β΄Έκδοση, Κεφάλαιο 30,» [Ηλεκτρονικό]. Available: http://aibook.csd.auth.gr/include/slides/Chap30.pdf. [Πρόσβαση 11 02 2015].

[7] Ι. Μανωλόπουλος και Α. Ν. Παπαδόπουλος, Συστήματα Βάσεων Δεδομένων, Θεωρία και Πρακτική Εφαρμογή, Αθήνα.

[8] S. Chen και P. Pin, «The entity-relationship model—toward a unified view of data,» ACM Transactions on Database Systems (TODS) - Special issue: papers from the international conference on very large data bases, τόμ. 1, αρ. 1, pp. 9-36, 1976.

[9] Κ. Π. Ν. ΦΛΩΡΙΝΑΣ, «Η Θεωρία των Βάσεων Δεδομένων,» [Ηλεκτρονικό]. Available: http://dide.flo.sch.gr/Plinet/Tutorials/Tutorials-DataBasesTheory.html.

[10] J. Weizenbaum, «ELIZA—A Computer Program For the Study of Natural Language Communication Between Man And Machine,» Communications of the ACM, pp. 36-45, 1966.

[11] Wikipedia, «Rogerian Argument,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Rogerian_argument. [Πρόσβαση 11 03 2015].

[12] W. A. Woods, R. M. Kaplan και B. Nash-Webber, «The Lunar Sciences Natural Language Information System: Final Report,» Bolt Beranek and Newman, 1972.

[13] S. Valentine, F. Vides, G. Lucchese, D. Turner, H.-h. Kim, W. Li, J. Linsey και T. Hammond, «Mechanix: A Sketch-Based Tutoring System for Statics Courses.,» IAAI, 2012.

[14] M. G. Sutton και I.-C. Jong, «A truss analyzer for enriching the learning experience of students.,» ASEE Annual Conference, 2000.

[15] C. Mellon, «A Game Changer: The Open Learning Initiative,» presidential perspectives No. 6, 2012.

[16] K. Vanlehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, ... και M. Wintersgill, «The Andes physics tutoring system: Lessons learned,» International Journal of Artificial Intelligence in Education, τόμ. 15, αρ. 3, pp. 147-204, 2005.

[17] R. Roselli, L. Howard και S. Brophy, «A Computer-Based Free Body Diagram Assistant,» Computer Applications in Engineering Education, pp. 281-290, 2006.

[18] S. Rosser, «InTEL: Interactive toolkit for engineering learning,» Georgia Tech, [Ηλεκτρονικό]. Available: http://intel.gatech.edu/toolkit/. [Πρόσβαση 8 03 2015].

[19] W. Lee, R. de Silva, E. Peterson, R. Calfee και T. Stahovich, «Newton’s pen: a pen-based tutoring system for statics,» Computers & Graphics, pp. 511-524, 2008.

[20] DST, «Interactive Physics,» Design Simulation Technologies, [Ηλεκτρονικό]. Available: http://www.design-simulation.com/ip/. [Πρόσβαση 8 03 2015].

[21] E. Anderson, «M-MODEL8,» 2011. [Ηλεκτρονικό]. Available: http://aln.coe.ttu.edu/anderson/premier/Default.html. [Πρόσβαση 8 03 2015].

[22] S. B. Silveira και A. Branco, «Extracting multi-document summaries with a double clustering approach.,» Natural Language Processing and Information Systems, pp. 70-81, 2012.

[23] M. Ogrodniczuk και A. Przepiórkowski, «Polish language processing chains for multilingual information systems.,» Natural Language Processing and Information Systems, pp. 152-157, 2012.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin και E. Herbst, «Moses: open source toolkit for statistical machine translation,» Proceeedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177-180, 2007.

[25] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard και D. McClosky, «The Stanford CoreNLP natural language processing toolkit,» σε 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014.

[26] J. Bloch, Effective Java, Upper Sadle River, NJ: Addison Wesley, 2008.

[27] D. Ferrucci και A. Lally, «UIMA: an architectural approach to unstructured information processing in the corporate research environment,» Natural Language Engineering, pp. 327-348, 2004.

[28] H. Cunningham, D. Maynard, K. Bontcheva και V. Tablan, «GATE: an architecture for development of robust HLT applications,» ACL, 2002.

[29] J. Clarke, V. Srikumar, M. Sammons και D. Roth, «An NLP Curator,» LREC, 2012.

[30] A. Kokkalis, P. Vagenas, A. Zervakis, A. Simitsis, G. Koutrika και Y. Ioannidis, «λόγος: A System for Translating Queries into Narratives».

[31] M. Minock, «A phrasal approach to natural language interfaces over databases,» NLDB, 2005.

[32] V. C. Storey, R. C. Goldstein και H. Ullrich, «Naive semantics to support automated database design,» IEEE TKDE, 2002.

[33] J. Danaparamita και W. Gatterbauer, «QueryViz: Helping Users Understand SQL Queries and their Patterns,» EDBT, 2011.

[34] G. Koutrika, A. Simitsis και Y. Ioannidis, «Explaining structured queries in natural language,» ICDE, 2010.

[35] C. I. Guinn και R. J. Montoya, «Natural language processing in Virtual Reality training environments,» σε The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), 1997.

[36] R. S. Kamath, «Development of Intelligent Virtual Environment by Natural Language Processing,» Special issue of International Journal of Latest Trends in Engineering and Technology, 2013.

[37] J. Veras, S. Labidi, N. Costa και T. Pinheiro, «Development of an intelligent virtual environment applied to mastology for diagnosis and training,» σε IEEE International Conference on Computer Medical Applications (ICCMA), 2013.

[38] R. Kamath και R. Kamat, «Development of Virtual Reality Interface for Architectural Visualization,» International Journal of Computer Science and Knowledge Engineering, pp. 27-31, 2012.

[39] D. Wormell, E. Foxlin και P. Katzman, «Improved 3D Interactive Devices for Passive and Active Stereo Virtual Environments,» σε 13th. Eurographics Workshop on Virtual Environments, 2007.

[40] G. Kramer, G. Bouma, D. Hendriksen και M. Homminga, «Classifying image galleries into a taxonomy using metadata and wikipedia,» Natural Language Processing and Information Systems, pp. 191-196, 2012.

[41] Kalooga, «Kalooga,» [Ηλεκτρονικό]. Available: http://www.kalooga.com/. [Πρόσβαση 10 02 2015].

[42] Wikipedia, «Wikipedia Category,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Help:Category. [Πρόσβαση 10 02 2015].

[43] M. Janik και K. Kochut, «Training-less ontology-based text categorization,» ECIR Workshop on

Exploiting Semantic Annotations in Information Retrieval, pp. 3-17, 2008.

[44] G. Tsatsaronis, I. Varlamis και K. Nørv°ag, «Semanticrank: ranking keywords and sentences using semantic graphs,» Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1074-1082, 2010.

[45] D. Tyagi, T. Joshi, D. Ghule και A. Joshi, «An Interactive Answering System using Template Matching and SQL Mapping for Natural Language Processing,» International Journal, 2014.

[46] A. Andrenucci και E. Sneiders, «Automated Question Answering: Review of the Main Approaches,» σε International Conference on Information Technology and Applications, 2005.

[47] T. Scholz, S. Conrad και I. Wolters, «Comparing different methods for opinion mining in newspaper articles,» Natural Language Processing and Information Systems, pp. 259-264, 2012.

[48] T. Watson και P. Noble, «Evaluating public relations: a best practice guide to public relations planning, research & evaluation,» PR in practice series, 2007.

[49] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen και J. Belyaeva, «Sentiment analysis in the news,» Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010.

[50] D. Bollegala, D. Weir και J. Carroll, «Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification,» Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies, τόμ. 1, pp. 132-141, 2011.

[51] R. Remus, U. Quasthoff και G. Heyer, «SentiWS – a publicly available german-language resource for sentiment analysis,» Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010.

[52] N. Kaji και M. Kitsuregawa, «Building lexicon for sentiment analysis from massive collection of html documents,» Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

[53] K. Church και P. Hanks, «Word association norms, mutual information, and lexicography,» Proceedings of the 27th Annual Meeting on ACL, pp. 76-83, 1989.

[54] A. Harb, M. Planti´e, G. Dray, M. Roche, F. Trousset και P. Poncelet, «Web opinion mining: how to extract opinions from blogs?,» Proc. of the 5th Intl. Conf. on Soft Computing as Transdisciplinary Science and Technology, pp. 211-217, 2008.

[55] A. Giordani και A. Moschitti, «Generating SQL queries using natural language syntactic dependencies and metadata,» Natural Language Processing and Information Systems, pp. 164-170, 2012.

[56] B. Marie-Catherine de Marneffe και C. Manning, «Generating typed dependency parses from phrase structure parses,» Proceedings LREC, 2006.

[57] A. Giordani και A. Moschitti, «Corpora for automatically learning to map natural language questions into sql queries,» Proceedings of LREC 2010. European Language Resources Association (ELRA), 2010.

[58] A. Popescu, O. Etzioni και H. Kautz, «Towards a theory of natural language interfaces to databases,» Proceedings of the 2003 International Conference on Intelligent User Interfaces. Association for Computational Linguistics, 2003.

[59] R. Kate και R. Mooney, «Using string-kernels for learning semantic parsers,» Proceedings of the 21st ICCL and 44th Annual Meeting of the ACL, 2006.

[60] S. Ruwanpura, «Sq-hal: Natural language to sql translator,» [Ηλεκτρονικό]. Available: http://www.csse.monash.edu.au/hons/projects/2000/Supun.Ruwanpura/. [Πρόσβαση 10 03 2015].

[61] A. Giordani και A. Moschitti, «Syntactic Structural Kernels for Natural Language Interfaces to Databases,» ECML PKDD, pp. 391-406, 2009.

[62] F.-C. Pop, M. Cremene, M.-F. Vaida και M. Riveill, «On-demand service composition based on natural language requests,» WONS 2009, pp. 41-44, 2009.

[63] S. Kirati, «A demonstration on service compositions based on natural language request and user contexts,» Master's thesis, Norwegian University of Science and Technology (NTNU), Department

of Telematics, 2008.

[64] E. C. Pedraza, J. A. Zúñiga, L. J. Suarez-Meza και J. C. Corrales, «User-Driven automatic resource retrieval based on natural language request,» Natural Language Processing and Information Systems, pp. 203-209, 2012.

[65] E. Pedraza, J. Zuniga, L. Suarez Meza και J. Corrales, «Automatic service retrieval in converged environments based on natural language request,» SERVICE COMPUTATION 2011, pp. 52-56, 2011.

[66] W. Bechtel, «Connectionism and the philosophy of mind: an overview,» The Southern Journal of Philosophy, pp. 17-41, 1988.

[67] J. Ramos, «Using TF-IDF to Determine Word Relevance in Document Queries,» Proceedings of the First InstructionalConference onMachine Learning, iCML, 2003.

[68] N. Bloom, «Using natural language processing to improve document categorization with associative networks,» Natural Language Processing and Information Systems, pp. 177-182, 2012.

[69] D. Klein και C. Manning, «Fast Exact Inference with a Factored Model for Natural Language Parsing,» Adv. in Neural Information Processing Systems, pp. 3-10, 2003.

[70] R. Schank και R. Abelson, «Scripts, Plans, Goals and Understanding,» Erlbaum, Hillsdale, 1977.

[71] C. Fellbaum, «WordNet: An Electronic Lexical Database,» MIT Press, Cambridge, 1998.

[72] G. Miller, «WordNet: A Lexical Database for English,» Communications of the ACM, pp. 39-41, 1995.

[73] A. Graesser, D. McNamara, M. Louwerse και Z. Cai, «Coh-Metrix: Analysis of text on cohesion and language,» Behavior Research Methods, Instruments, & Computers, pp. 193-202, 2004.

[74] H. Clark, Using language, Cambridge: Cambridge University Press., 1996.

[75] A. C. Graesser, M. A. Gernsbacher και S. Goldman, Handbook of discourse processes., Erlbaum, 2003.

[76] W. Kintsch, Comprehension: A paradigm for cognition, Cambridge, MA: Cambridge University Press, 1998.

[77] M. Gernsbacher, Language comprehension as structure building, Hillsdale, NJ: Erlbaum, 1990.

[78] S. Goldman, A. C. Graesser και P. van den Broek, Narrative comprehension, causality, and coherence, Mahwah, NJ: Erlbaum, 1999.

[79] M. Louwerse, «An analytic and cognitive parameterization of coherence relations,» Cognitive Linguistics, pp. 291-315, 2001.

[80] D. McNamara και W. Kintsch, «Learning from text: Effects of prior knowledge and text coherence,» Discourse Processes, pp. 247-287, 1996.

[81] T. J. M. Sanders και L. G. M. Noordman, «The role of coherence relations and their linguistic markers in text processing,» Discourse Processes, pp. 37-60, 2000.

[82] B. K. Britton και S. Gulgoz, «Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures,» Journal of Educational Psychology, pp. 329-345, 1991.

[83] D. McNamara, «Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge,» Canadian Journal of Experimental Psychology, pp. 51-62, 2001.

[84] R. Zwaan και G. Radvansky, «Situation models in language comprehension and memory,» Psychological Bulletin, pp. 162-185, 1998.

[85] J. Weston, S. Crossley και D. S. McNamara, «Computationally assessing human judgments of freewriting quality,» Applied natural language processing: Identification, investigation, and resolution.

[86] Google Now, «Google Now,» Google, [Ηλεκτρονικό]. Available: http://www.google.com/landing/now/. [Πρόσβαση 12 03 2015].

[87] Google InsideSearch, «Google Inside Search,» Google, [Ηλεκτρονικό]. Available: http://www.google.com/insidesearch/features/voicesearch/. [Πρόσβαση 12 03 2015].

[88] Apple Siri, «Apple Siri,» Apple, [Ηλεκτρονικό]. Available: http://www.apple.com/ios/siri/. [Πρόσβαση 12 03 2015].

[89] Ask-Ziggy, «Ask Ziggy,» Ask Ziggy Inc., [Ηλεκτρονικό]. Available: http://www.ask-ziggy.com/. [Πρόσβαση 12 03 2015].