ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:

Αυτόματη Κατάταξη Ελληνικών Ερωτήσεων σε Κατηγορίες

Ονοματεπώνυμο Φοιτητή: Βρυσαγώτης Χαράλαμπος
Αριθμός Μητρώου:3020010

Επιβλέπων Καθηγητής: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ 2007

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Ο σκοπός αυτής της πτυχιακής εργασίας ήταν η αυτόματη κατάταξη ελληνικών ερωτήσεων σε κατηγορίες, ανάλογα με το αν ζητούν ως απάντηση όνομα προσώπου, τοποθεσίας, οργανισμού, ορισμό ή χρονική έκφραση. Για την κατάταξη των ερωτήσεων χρησιμοποιήθηκαν Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), οι οποίες εκπαιδεύτηκαν και αξιολογήθηκαν σε μία συλλογή 1374 ερωτήσεων που δημιουργήθηκε στη διάρκεια της εργασίας. Δοκιμάστηκαν, επίσης, διαφορετικές μέθοδοι αναπαράστασης των ερωτήσεων, που έχουν χρησιμοποιηθεί σε ανάλογα συστήματα αγγλικών ερωτήσεων, όπως η ανίχνευση μεμονωμένων λέξεων ή συνεχόμενων ακολουθιών λέξεων, η χρήση των θεμάτων των λέξεων (stemming) κλπ. Το ποσοστό ορθότητας (accuracy) του συστήματος είναι κατά μέσο όρο (για όλες τις κατηγορίες ερωτήσεων) περίπου 90%. Υπάρχουν, ωστόσο, αρκετά περιθώρια βελτίωσης των επιδόσεων του συστήματος, αφού δεν κατέστη δυνατή η χρήση ελληνικού ιεραρχικού θησαυρού λέξεων (π.χ. ελληνικό WordNet) ή ελληνικού συντακτικού αναλυτή. Στο μέλλον είναι, επίσης, δυνατόν να προστεθούν νέες κατηγορίες ή υπο-κατηγορίες ερωτήσεων.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της παρούσης εργασίας κ. Ίωνα Ανδρουτσόπουλο για την πολύτιμη καθοδήγησή του. Επίσης οφείλω να ευχαριστήσω τους Πρόδρομο Μαλακασιώτη και Ιωάννη Χρονάκη για την ευγενική παραχώρηση του Συστήματος Αναγνώρισης Μερών του Λόγου που ανέπτυξαν.

1.ΕΙΣΑΓΩΓΗ

1.1 ΑΥΤΟΜΑΤΗ ΚΑΤΑΤΑΞΗ ΕΡΩΤΗΣΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΑ ΕΡΩΤΑΠΟΚΡΙΣΕΩΝ

Οι υπάρχουσες μηχανές αναζήτησης του Παγκόσμιου Ιστού επιτρέπουν στους χρήστες τους να εισάγουν λέξεις-κλειδιά και στη συνέχεια επιστρέφουν μια λίστα από σχετικές ιστοσελίδες ή έγγραφα μέσα στα οποία πρέπει οι ίδιοι οι χρήστες να εντοπίσουν τις πληροφορίες που τους ενδιαφέρουν. Ωστόσο, ο συνεχώς αυξανόμενος όγκος πληροφοριών που συσσωρεύεται στον Παγκόσμιο Ιστό καθιστά όλο και πιο δύσκολο τον εντοπισμό πληροφοριών μέσα στα επιστρεφόμενα έγγραφα από τους ίδιους τους χρήστες. Τα συστήματα ερωταποκρίσεων (question answering systems) επιχειρούν να λύσουν αυτό το πρόβλημα επιτρέποντας στους χρήστες να θέτουν ερωτήσεις φυσικής γλώσσας, αντί για λέξεις-κλειδιά, και προσπαθώντας να επιστρέψουν ακριβείς απαντήσεις (π.χ. το όνομα ενός προσώπου, το οποίο ζητείται από μια ερώτηση), αντί για σχετικά έγγραφα. Τα συστήματα αυτά μπορούν να χρησιμοποιηθούν τόσο στον Παγκόσμιο Ιστό όσο και με άλλες συλλογές εγγράφων (π.χ. αρχεία εφημερίδων).

Η λειτουργία των συστημάτων ερωταποκρίσεων περιλαμβάνει τα στάδια: (α) της επεξεργασίας της ερώτησης που τέθηκε από το χρήστη, (β) της ανάκτησης σχετικών εγγράφων, μέσω μιας υπάρχουσας μηχανής αναζήτησης, (γ) της επεξεργασίας των σχετικών εγγράφων και του εντοπισμού υποψηφίων απαντήσεων και (δ) της επιλογής της τελικής απάντησης. Η παρούσα εργασία επικεντρώνεται στο πρώτο στάδιο (της επεξεργασίας της ερώτησης) και πιο συγκεκριμένα στην αυτόματη κατάταξη της ερώτησης με κριτήριο τον τύπο της απαιτούμενης απάντησης. Δηλαδή η ερώτηση κατατάσσεται σε διαφορετική κατηγορία, ανάλογα με το αν η ζητούμενη απάντηση είναι όνομα προσώπου, τοποθεσίας, κτλ. Η διαδικασία της κατάταξης των ερωτήσεων είναι σημαντική για τα επόμενα στάδια. Αν, για παράδειγμα, η ερώτηση κριθεί ότι ζητά ένα όνομα προσώπου, η αναζήτηση υποψηφίων απαντήσεων του σταδίου (γ) μπορεί να περιορισθεί σε ονόματα προσώπων που έχουν εντοπιστεί μέσα στα σχετικά έγγραφα του σταδίου (β) μέσω ενός συστήματος εντοπισμού ονομάτων οντοτήτων (named-entity recognizer) [3]. Ενώ αν η ερώτηση κριθεί ότι ζητά ορισμό, μπορούν να χρησιμοποιηθούν ειδικές τεχνικές που εντοπίζουν ορισμούς μέσα σε έγγραφα [4,10].

Στη διάρκεια της εργασίας, αναπτύχθηκε ένα σύστημα αυτόματης κατάταξης απλών ελληνικών ερωτήσεων που ζητούν ως απαντήσεις ονόματα προσώπων, τοποθεσιών, οργανισμών, χρονικές εκφράσεις ή ορισμούς. Για την κατάταξη των ερωτήσεων χρησιμοποιήθηκαν Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) [11], οι οποίες εκπαιδεύτηκαν και αξιολογήθηκαν σε μία συλλογή 1374 ελληνικών ερωτήσεων που δημιουργήθηκε στη διάρκεια της εργασίας. Το ποσοστό ορθότητας (accuracy) του συστήματος είναι κατά μέσο όρο (για όλες τις κατηγορίες ερωτήσεων) περίπου 90%. Δοκιμάστηκαν διαφορετικές μέθοδοι αναπαράστασης των ερωτήσεων, που έχουν χρησιμοποιηθεί σε ανάλογα συστήματα αγγλικών ερωτήσεων, όπως η ανίχνευση μεμονωμένων λέξεων ή συνεχόμενων ακολουθιών λέξεων, η

χρήση των θεμάτων των λέξεων (stemming) κλπ.

1.2 ΔΙΑΡΘΡΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Το υπόλοιπο της εργασίας είναι διαρθρωμένο ως εξής:

Στο δεύτερο κεφάλαιο γίνεται μια σύντομη ανασκόπηση των σχετικών αλγορίθμων Μηχανικής Μάθησης που χρησιμοποιήθηκαν, καθώς και μεθόδων κατάταξης ερωτήσεων που έχουν χρησιμοποιηθεί σε αγγλικά συστήματα.

Στο τρίτο κεφάλαιο περιγράφεται η αρχιτεκτονική του συστήματος της εργασίας και ο τρόπος εκπαίδευσης και αξιολόγησής του.

Τέλος, το τέταρτο κεφάλαιο συνοψίζει τα αποτελέσματα και συμπεράσματα της εργασίας και προτείνει κατευθύνσεις για μελλοντικές επεκτάσεις της.

2. ΑΝΑΣΚΟΠΗΣΗ ΣΧΕΤΙΚΩΝ ΕΡΓΑΣΙΩΝ ΚΑΙ ΜΕΘΟΔΩΝ

2.1 TREC

Το Συνέδριο Ανάκτησης Πληροφοριών TREC (Text REtrieval Conference)[1] δημιουργήθηκε το 1992 και βρίσκεται υπό την αιγίδα του National Institute of Standards and Technology και του Υπουργείου Αμύνης των Η.Π.Α. Σκοπός του είναι η ενθάρρυνση και υποβοήθηση της έρευνας προς την κατεύθυνση ολοκληρωμένων συστημάτων ανάκτησης πληροφοριών. Μία από τις δραστηριότητες του TREC είναι ο ετήσιος διαγωνισμός συστημάτων ερωταποκρίσεων (Question Answering Track) [2]. Πολλές από τις μεθόδους που αναφέρουμε στη συνέχεια αναπτύχθηκαν με την ευκαιρία αυτού του διαγωνισμού.

2.2 ΠΡΟΗΓΟΥΜΕΝΕΣ ΜΕΘΟΔΟΙ ΚΑΤΑΤΑΞΗΣ ΕΡΩΤΗΣΕΩΝ

Τα περισσότερα από τα προηγούμενα συστήματα κατάταξης ερωτήσεων υποστηρίζουν ερωτήσεις που είναι γραμμένες στα Αγγλικά. Η παρούσα εργασία είναι μάλλον η πρώτη που ασχολείται με τη δημιουργία ενός αντίστοιχου συστήματος για τα Ελληνικά. Αξίζει, επίσης, να σημειωθεί ότι οι μέχρι τώρα προσεγγίσεις υποστηρίζουν συνήθως και υποκατηγορίες ερωτήσεων, πέρα από τις κύριες κατηγορίες (π.χ. αν ένα τοπωνύμιο είναι πόλη, χώρα, βουνό, κτλ.), που το σύστημα της παρούσης εργασίας δεν αναγνωρίζει.

Γενικά, η μέθοδος μηχανικής μάθησης που δείχνει να έχει τα καλύτερα αποτελέσματα στην κατάταξη ερωτήσεων είναι οι Μηχανές Διανυσμάτων Υποστήριξης (Μ.Δ.Υ.), όπως φαίνεται και στην εργασία των Zhang και Lee [1], όπου συγκρίνονται οι Μ.Δ.Υ. με άλλους αλγόριθμους μηχανικής μάθησης στο πρόβλημα της κατάταξης αγγλικών ερωτήσεων. Ωστόσο, έχουν υπάρξει και προσεγγίσεις με διαφορετικούς αλγορίθμους μηχανικής μάθησης. Στην εργασία των Li και Roth [12] παρουσιάζεται ένας ιεραρχικός ταξινομητής που με τη χρήση του αλγορίθμου SNOW (Sparse Network Of Winnows) επιτυγχάνει εντυπωσιακά ποσοστά ορθότητας (92,5% για τις κύριες κατηγορίες, 84,2% για τις υποκατηγορίες).

Οι κυριότερες ιδιότητες που χρησιμοποιούνται για την αναπαράσταση των ερωτήσεων κατά τη χρήση μηχανικής μάθησης σχετίζονται με την ανίχνευση συχνά εμφανιζόμενων λέξεων ή ακολουθιών λέξεων (token n-grams). Επίσης, έχουν χρησιμοποιηθεί ιεραρχικοί θησαυροί λέξεων, όπως το WordNet, αλλά και συντακτικά χαρακτηριστικά [15, 7]. Έχουν, ακόμη, προταθεί ειδικοί πυρήνες για Μ.Δ.Υ., οι οποίοι εκμεταλλεύονται ιεραρχικές επισημειώσεις (annotations) εκφράσεων φυσικής γλώσσας (π.χ. μορφολογικές και συντακτικές πληροφορίες) [6].

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η ιδέα ενός ταξινομητή που δεν εξαρτάται από τη φυσική γλώσσα στην οποία είναι γραμμένες οι ερωτήσεις. Η

1 http://trec.nist.gov

υλοποίησή του μπορεί να επιτευχθεί με εκπαίδευση πάνω σε αποτελέσματα αναζητήσεων στον Παγκόσμιο Ιστό. Στην εργασία των Solorio κ.ά. [8] παρουσιάζεται μια μέθοδος που επιτυγχάνει ικανοποιητικά αποτελέσματα για τρεις γλώσσες (Αγγλικά, Ιταλικά, Ισπανικά).

2.3 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Στην εργασία αυτή χρησιμοποιούμε μεθόδους μηχανικής μάθησης για το πρόβλημα της κατάταξης αντικειμένων (στην περίπτωσή μας, ερωτήσεων) σε κατηγορίες. Χρησιμοποιούμε επιβλεπόμενες (supervised) μεθόδους μάθησης, όπου αρχικά δίνονται στο σύστημα παραδείγματα αντικειμένων προς κατάταξη, το καθένα μαζί με την επιθυμητή απόκριση του συστήματος (κατηγορία). Το σύστημα «μαθαίνει» ένα μοντέλο (π.χ. μια συνάρτηση) κατάταξης, το οποίο μπορεί στη συνέχεια να χρησιμοποιηθεί για να καταταγούν αυτόματα νέα αντικείμενα σε κατηγορίες. Οι περισσότερες μέθοδοι επιβλεπόμενης μάθησης αυτού του είδους απαιτούν τα αντικείμενα (ερωτήσεις) να παριστάνονται ως διανύσματα χαρακτηριστικών (feature vectors), κάτι που προϋποθέτει την επιλογή των πιο χρήσιμων χαρακτηριστικών. Στην παρούσα εργασία χρησιμοποιήθηκαν, όπως προαναφέρθηκε, Μηχανές Διανυσμάτων Υποστήριξης, των οποίων η λειτουργία περιγράφεται παρακάτω.

2.4 ΜΗΧΑΝΕΣ ΔΙΑΝΥΣΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) [11] έχουν χρησιμοποιηθεί με μεγάλη επιτυχία σε πολλά προβλήματα κατάταξης. Στην απλούστερη μορφή τους, κάθε Μ.Δ.Υ. υποστηρίζει μόνο δύο κατηγορίες (θετική και αρνητική). Ένας απλός τρόπος να υποστηριχθούν περισσότερες κατηγορίες (στην περίπτωσή μας, ερωτήσεις προσώπων, οργανισμών, τοποθεσιών κλπ.) είναι να εκπαιδευθεί μία διαφορετική Μ.Δ.Υ. για κάθε κατηγορία (μία Μ.Δ.Υ. για ερωτήσεις προσώπων, μία για ερωτήσεις οργανισμών κλπ.) και η κάθε Μ.Δ.Υ. να αποφαίνεται αν το υπό κατάταξη αντικείμενο (ερώτηση) ανήκει ή όχι σε μία συγκεκριμένη κατηγορία. Με τον τρόπο αυτό είναι δυνατόν να υποστηριχθούν και επικαλυπτόμενες κατηγορίες (π.χ. μία ερώτηση να ανήκει σε περισσότερες από μία κατηγορίες), επιτρέποντας μάλιστα στην κάθε Μ.Δ.Υ. να χρησιμοποιεί διαφορετικά χαρακτηριστικά (features) στα διανύσματα των αντικειμένων.

Η γενική ιδέα στις Μ.Δ.Υ. είναι η προβολή των διανυσμάτων των αντικειμένων σε ένα νέο διανυσματικό χώρο περισσότερων διαστάσεων, με τη χρήση ενός (συνήθως μη γραμμικού) μετασχηματισμού. Με τον τρόπο αυτό, κατηγορίες (θετική και αρνητική) των οποίων τα διανύσματα δεν ήταν γραμμικώς διαχωρίσιμα στον αρχικό χώρο μπορούν να μετατραπούν σε γραμμικώς διαχωρίσιμες, δηλαδή να διαχωριστούν μαθαίνοντας ένα κατάλληλο υπερεπίπεδο στο νέο διανυσματικό χώρο. Η μετάβαση στο νέο διανυσματικό χώρο γίνεται έμμεσα, μέσω μιας συνάρτησης πυρήνα (kernel) που υπολογίζει τα εσωτερικά γινόμενα στο νέο χώρο. Η επιλογή του υπερεπιπέδου διαχωρισμού γίνεται με μεθόδους βελτιστοποίησης, έτσι ώστε να

διαχωρίζονται σωστά τα παραδείγματα εκπαίδευσης αλλά και να μεγιστοποιείται η απόσταση του υπερεπιπέδου από τα κοντινότερα παραδείγματα των δύο κατηγοριών, χωρίς να υπερβαίνουμε ένα μέγιστο επιτρεπόμενο συνολικό σφάλμα. Ο αναγνώστης μπορεί να ανατρέξει για περισσότερες λεπτομέρειες στην εργασία του Γιώργου Λουκαρέλλι [13].

Για τις ανάγκες της εργασίας χρησιμοποιήθηκε η υλοποίηση Μ.Δ.Υ. «LIBSVM»[2] , που διατίθεται ελεύθερα, με συνάρτηση ακτινωτής βάσης (Radial Base Function – RBF) ως πυρήνα.

2http://www.csie.ntu.edu.tw/~cjlin/libsvm

3. ΤΟ ΣΥΣΤΗΜΑ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ

3.1 ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Το σύστημα της εργασίας έχει κατασκευασθεί ούτως ώστε να υποστηρίζει τις εξής κατηγορίες ερωτήσεων φυσικής γλώσσας:

- ερωτήσεις ονομάτων προσώπων (π.χ. «Ποιος είναι ο πρωθυπουργός της Ολλανδίας;»),
- ερωτήσεις ονομάτων τοποθεσιών (π.χ. «Ποια είναι η πρωτεύουσα της Λιβύης;»),
- ερωτήσεις ορισμού (π.χ. «Τι είναι η θαλασσαιμία;»),
- ερωτήσεις χρονικών εκφράσεων (π.χ. «Πότε έγινε η γαλλική επανάσταση;»),
- ερωτήσεις ονομάτων οργανισμών (π.χ. «Ποια εταιρία κατασκευάζει το Χ;»).

Κατά την εκπαίδευσή του, το σύστημα δέχεται ως είσοδο ερωτήσεις, σε καθεμιά από τις οποίες έχει επισημειωθεί χειρωνακτικά η σωστή κατηγορία. Οι ερωτήσεις είναι αποθηκευμένες σε ένα αρχείο XML του ακόλουθου μορφότυπου:

```
<questions>
    < question question_type=”1” question_id=”105”>
        Ποιος έμεινε στην ιστορία ως "ο πατέρας της νίκης";
    </question>
    ...

    < question question_type=”2” question_id=”511”>
        Ποιος είναι ο ποταμός με το μεγαλύτερο μήκος στον κόσμο;
    </question>
    ...

    < question question_type=”3” question_id=”689”>
        Ποιος είναι ο ορισμός της επιτάχυνσης στη φυσική;
    </question>
    ...

    < question question_type=”4” question_id=”941”>
        Πες την ημερομηνία που υπογράφτηκε η συνθήκη των Σεβρών.
    </question>
    ...

    < question question_type=”5” question_id=”1239”>
        Ποιανής εταιρείας είναι διευθυντής ο Μπιλ Γκέητς;
    </question>
    ...
```

```
</question>
```

Επίσης, είναι πολύτιμη πληροφορία για το σύστημα να γνωρίζει τι μέρος του λόγου είναι η κάθε λέξη της ερώτησης. Για το λόγο αυτό, οι ίδιες ερωτήσεις διαβάζονται και από ένα δεύτερο αρχείο XML, το οποίο είναι η έξοδος από τη χρήση του συστήματος αναγνώρισης μερών του λόγου των Π. Μαλακασιώτη και Γ. Χρονάκη [14, 16] και ακολουθεί τον εξής μορφότυπο:[3]

```
<article>
.....

<sentence>
        <token PoS="pronoun" gender="masc" number="sg"
case="nom">Ποιος</token>
        <token PoS="verb" tense="past" number="sg">έμεινε</token>
         <token PoS="article" function="prep" gender="fem" number="sg"
case="acc">στην  </token>
         <token PoS="noun" gender="fem" number="sg"
case="acc">ιστορία</token>
         <token PoS="preposition">ως</token>
         <token PoS="other" type="symbol">"</token>
         <token PoS="article" function="def" gender="masc" number="sg"
case="nom">ο      </token>
         <token PoS="noun" gender="masc" number="sg"
case="nom">πατέρας</token>
         <token PoS="article" function="def" gender="fem" number="sg"
case="gen">της</token>
         <token PoS="noun" gender="fem" number="sg" case="gen">νίκης</token>
         <token PoS="other" type="symbol">"</token>
         <token PoS="punctuation">.</token>

</sentence>
.....

</article>
```

Με βάση αυτά τα δύο αρχεία, δημιουργούνται αυτόματα οι ανάλογες δομές για την αποθήκευση των πληροφοριών της κάθε ερώτησης. Σε γενικές γραμμές, κάθε ερώτηση διασπάται σε λεκτικές μονάδες, για κάθε μία από τις οποίες αποθηκεύονται πληροφορίες όπως η μορφή της ως συμβολοσειρά, τι μέρος του λόγου είναι και η θέση της μέσα στην ερώτηση.

3 Στο μέλλον, τα δύο αρχεία θα μπορούσαν να ενοποιηθούν.

Ακολουθεί η διαδικασία της εκπαίδευσης, για κάθε μία από τις πέντε Μ.Δ.Υ. (μία για κάθε κατηγορία ερωτήσεων) ξεχωριστά.[4] Κάθε Μ.Δ.Υ. εκπαιδεύεται σε διανυσματικές αναπαραστάσεις των ερωτήσεων εκπαίδευσης οι οποίες περιέχουν τα χαρακτηριστικά που χρησιμοποιεί η συγκεκριμένη Μ.Δ.Υ. Μετά την ολοκλήρωση της εκπαίδευσης, οι Μ.Δ.Υ. μπορούν να χρησιμοποιηθούν για την κατάταξη νέων ερωτήσεων, των οποίων οι κατηγορίες δεν είναι πλέον γνωστές. Σε περίπτωση που θέλουμε να επιτρέψουμε επικαλυπτόμενες κατηγορίες ερωτήσεων (να επιτρέπεται μία ερώτηση να ανήκει σε πολλές κατηγορίες), κάθε ερώτηση κατατάσσεται στις κατηγορίες ερωτήσεων για τις οποίες οι αντίστοιχες Μ.Δ.Υ. θεωρούν ότι η ερώτηση ανήκει στη θετική τους κατηγορία. Αν δεν θέλουμε να επιτρέψουμε επικαλυπτόμενες κατηγορίες ερωτήσεων, εντοπίζουμε τη Μ.Δ.Υ. που κατατάσσει την ερώτηση στη θετική της κατηγορία με τη μεγαλύτερη βεβαιότητα, και κατατάσσουμε την ερώτηση στην κατηγορία ερωτήσεων που αντιστοιχεί σε αυτήν τη Μ.Δ.Υ.[5] Το σύστημα της εργασίας μπορεί να χρησιμοποιηθεί, εν γένει, και με τους δύο τρόπους, αλλά στα πειράματά μας χρησιμοποιήσαμε ερωτήσεις εκπαίδευσης και αξιολόγησης που ανήκαν η κάθε μία ακριβώς σε μία κατηγορία.

Το σύστημα της εργασίας υλοποιήθηκε σε Java και χρησιμοποιεί τη μορφή της LIBSVM που είναι υλοποιημένη στην ίδια γλώσσα.

3.2 Η ΣΥΛΛΟΓΗ ΤΩΝ ΕΡΩΤΗΣΕΩΝ

Το σύνολο εκπαίδευσης αποτελείται από 1374 ερωτήσεις κατανεμημένες στις 5 κατηγορίες όπως φαίνεται παρακάτω:

- Κατηγορία προσώπων: 323 ερωτήσεις
- Κατηγορία τοποθεσιών: 318 ερωτήσεις
- Κατηγορία ορισμών: 244 ερωτήσεις
- Κατηγορία τοποθεσιών: 272 ερωτήσεις
- Κατηγορία οργανισμών: 217 ερωτήσεις

Η παραπάνω συλλογή βασίζεται σε μεγάλο βαθμό στα αντίστοιχα δεδομένα εκπαίδευσης του TREC, δηλαδή οι ερωτήσεις έχουν προκύψει από μετάφρασή τους στα ελληνικά, ενώ παράλληλα υπάρχουν και πολλές ερωτήσεις από διάφορα επιτραπέζια παιχνίδια γνώσεων.

3.3 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ

Στην ενότητα αυτή περιγράφονται τα μέτρα που χρησιμοποιήθηκαν για την πειραματική αξιολόγηση κάθε Μ.Δ.Υ. Τα πρώτα δύο είναι η ακρίβεια (precision) και

4 Χρησιμοποιείται η επιλογή «C-SVC (Support Vector Classification) της LIBSVM.

5 Κατά την κατάταξη ενός αντικειμένου, η υλοποίηση LIBSVM επιστρέφει και το βαθμό βεβαιότητας της απόφασής της, που είναι ένας αριθμός στο διάστημα [-1, +1].

η ανάκληση (recall), που ορίζονται από τις παρακάτω σχέσεις:

$$precision = \frac{\text{ερωτήσεις που κατετάγησαν ορθά στη θετική κατηγορία}}{\text{ερωτήσεις που κατετάγησαν στη θετική κατηγορία}}$$

$$recall = \frac{\text{ερωτήσεις που κατετάγησαν ορθά στη θετική κατηγορία}}{\text{ερωτήσεις της θετικής κατηγορίας}}$$

Τα παραπάνω μεγέθη μπορούν να οριστούν και για την αρνητική κατηγορία της Μ.Δ.Υ. (ερωτήσεις που δεν ανήκουν στην αντίστοιχη κατηγορία ερωτήσεων), ωστόσο τα αποτελέσματα ακρίβειας και ανάκλησης που ακολουθούν είναι μόνο για τη θετική κατηγορία. Επιπλέον, το μέτρο F συμψηφίζει τις δύο παραπάνω ποσότητες, όπως φαίνεται στην παρακάτω σχέση, με την παράμετρο β να καθορίζει το βάρος της ακριβείας σε σχέση με την ανάκληση. Στις επόμενες ενότητες χρησιμοποιούμε β = 1, δίνοντας έτσι το ίδιο βάρος στην ακρίβεια και την ανάκληση.

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall}$$

Επίσης, χρησιμοποιούμε ως μέτρο αξιολόγησης της κάθε Μ.Δ.Υ. και το ποσοστό ορθότητας (accuracy) των ερωτήσεων που κατετάγησαν στη σωστή κατηγορία (είτε αυτή είναι η θετική είτε είναι η αρνητική).

$$accuracy = \frac{\text{ερωτήσεις που κατετάγησαν στη σωστή (αρνητική ή θετική) κατηγορία}}{\text{συνολικό πλήθος ερωτήσεων}}$$

Για περισσότερη αξιοπιστία, η πειραματική αξιολόγηση κάθε Μ.Δ.Υ. έγινε με τη μέθοδο της 20-πλής διασταυρωμένης επικύρωσης (cross-validation). Δηλαδή τα διαθέσιμα δεδομένα εκπαίδευσης (τα διανύσματα των ερωτήσεων) χωρίστηκαν σε 20 ισοπληθείς ομάδες με τυχαίο τρόπο και έγιναν 20 επαναλήψεις. Σε κάθε επανάληψη, τα διανύσματα μιας διαφορετικής ομάδας χρησιμοποιήθηκαν ως δεδομένα αξιολόγησης, ενώ η Μ.Δ.Υ. εκπαιδεύτηκε στα δεδομένα των υπολοίπων 19 ομάδων. Τα αποτελέσματα (ακρίβεια, ανάκληση, ορθότητα) κάθε Μ.Δ.Υ., που παρουσιάζονται στις επόμενες ενότητες, είναι οι μέσοι όροι των αντιστοίχων αποτελεσμάτων των 20 επαναλήψεων.

Σημειωτέον ότι σε κάθε επανάληψη της διασταυρωμένης επικύρωσης, χρησιμοποιήσαμε κατά την εκπαίδευση της Μ.Δ.Υ. και το μηχανισμό ρύθμισης παραμέτρων (parameter tuning) που παρέχει η LIBSVM. Ο μηχανισμός αυτός επιχειρεί να εντοπίσει τον καλύτερο συνδυασμό τιμών των παραμέτρων C και γ της Μ.Δ.Υ. (η πρώτη παράμετρος ρυθμίζει τη βαρύτητα που δίδεται στα σφάλματα της Μ.Δ.Υ., ενώ η δεύτερη είναι παράμετρος του πυρήνα RBF) χρησιμοποιώντας μια αναζήτηση πλέγματος (grid search) και τη δική του διασταυρωμένη επικύρωση. Σε κάθε επανάληψη της διασταυρωμένης επικύρωσης των πειραμάτων μας, μόνο τα

δεδομένα εκπαίδευσης εκείνης της επανάληψης ήταν διαθέσιμα στο μηχανισμό ρύθμισης παραμέτρων (και άρα η δική του εσωτερική διασταυρωμένη επικύρωση εκτελείτο μόνο σε εκείνα τα δεδομένα).

3.4 ΔΙΑΝΥΣΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΕΡΩΤΗΣΕΩΝ

Στο υπόλοιπο αυτού του κεφαλαίου περιγράφουμε τις διαφορετικές μορφές διανυσματικής αναπαράστασης των ερωτήσεων που δοκιμάσαμε, μαζί με τα αντίστοιχα πειραματικά αποτελέσματα. Συνοπτικά, δοκιμάσαμε να περιλάβουμε τα εξής χαρακτηριστικά στα διανύσματα των ερωτήσεων:

- δυαδικά χαρακτηριστικά που δείχνουν το καθένα αν μία διαφορετική λέξη (ή η ρίζα της) που είναι συχνή σε μια κατηγορία ερωτήσεων (π.χ. αντωνυμίες όπως «ποιος», «πότε») εμφανίζεται ή όχι στην ερώτηση,
- αριθμητικά χαρακτηριστικά που δείχνουν το καθένα την απόσταση από την αρχική ερωτηματική αντωνυμία της ερώτησης (π.χ. «ποιος») μιας λέξης που εμφανίζεται στην ερώτηση και είναι συχνή σε μια κατηγορία ερωτήσεων (π.χ. «πρόεδρος» στην ερώτηση «Ποιος είναι ο πρόεδρος της Γαλλίας;»),
- δυαδικά χαρακτηριστικά που δείχνουν το καθένα αν μια διαφορετική ακολουθία συνεχόμενων λέξεων (ή των ριζών τους) που είναι συχνή σε μια κατηγορία ερωτήσεων εμφανίζεται ή όχι στην ερώτηση,
- δυαδικά χαρακτηριστικά που δείχνουν το καθένα αν μια διαφορετική ακολουθία όχι απαραιτήτως συνεχόμενων λέξεων (ή των ριζών τους) που είναι συχνή σε μια κατηγορία ερωτήσεων εμφανίζεται ή όχι στην ερώτηση,
- χαρακτηριστικά που παρέχουν επιπλέον πληροφορίες για το «κεντρικό» ουσιαστικό της ερώτησης (π.χ. «Ποια είναι η πρωτεύουσα της Λιβύης;»).

Αρκετά από τα παραπάνω χαρακτηριστικά είχαν χρησιμοποιηθεί και στην πτυχιακή εργασία του Δημήτρη Μαυροειδή [15], η οποία όμως είχε εστιαστεί σε αγγλικές ερωτήσεις.

Τα δυαδικά χαρακτηριστικά παριστάνονται με τις τιμές -1 (δεν εμφανίζεται) και +1 (εμφανίζεται), ενώ τα χαρακτηριστικά που παριστάνουν αποστάσεις κανονικοποιούνται στο διάστημα [-1,1], ούτως ώστε να ακολουθούνται οι σχετικές συμβουλές των κατασκευαστών του LIBSVM.

3.5 ΧΡΗΣΗ ΣΥΧΝΑ ΕΜΦΑΝΙΖΟΜΕΝΩΝ ΛΕΞΕΩΝ

Ένα από τα πιο σημαντικά χαρακτηριστικά μιας ερώτησης, που θα μπορούσε να αποκαλύψει την κατηγορία της, είναι η παρουσία συγκεκριμένων λέξεων μέσα στην ερώτηση. Για παράδειγμα, ερωτήσεις που ξεκινούν με «Πού» ή «Πότε» ανήκουν σίγουρα στις κατηγορίες των τοπικών και χρονικών ερωτήσεων αντίστοιχα. Ερωτήσεις, όμως, που ξεκινούν, για παράδειγμα, με «Ποιος» ή «Τι» μπορούν να

ανήκουν σε πολλές κατηγορίες (π.χ. «Ποιος κατασκευάζει το προϊόν Χ;»), οπότε δεν αρκεί να εξετάζει κανείς την αρχική λέξη της ερώτησης. Υπάρχουν, πάντως, αρκετές λέξεις (κυρίως ερωτηματικές αντωνυμίες και ουσιαστικά) που είναι στενά συνδεδεμένες με κάποια συγκεκριμένη κατηγορία ερωτήσεων. Για παράδειγμα, λέξεις όπως «χώρα», «πόλη», «βουνό», κτλ. είναι πολύ πιο πιθανό να βρίσκονται μέσα σε ερωτήσεις τοποθεσιών παρά σε ερωτήσεις οποιασδήποτε άλλης κατηγορίας. Επομένως, είναι εύλογο να αντιστοιχήσει κανείς κάθε μία από αυτές τις λέξεις σε μία διαφορετική δυαδική ιδιότητα που θα δείχνει αν η αντίστοιχη λέξη εμφανίζεται ή όχι στην ερώτηση.

Οι λέξεις για τις οποίες θα υπάρχουν δυαδικές ιδιότητες μπορούν να είναι διαφορετικές σε κάθε μία από τις πέντε Μ.Δ.Υ. (κατηγορία ονομάτων) και μπορούν να επιλεγούν με κριτήρια όπως το κέρδος πληροφορίας (information gain), το μέτρο χ^2 κλπ. Το απλούστερο κριτήριο επιλογής, που χρησιμοποιήσαμε στην εργασία, είναι να επιλέγονται σε κάθε Μ.Δ.Υ. λέξεις που εμφανίζονται πάνω από ν φορές στις ερωτήσεις εκπαίδευσης της αντίστοιχης κατηγορίας. Στα πειράματα αυτής της εργασίας θέσαμε ν = 4. Για την ακρίβεια, σε κάθε Μ.Δ.Υ. και σε κάθε επανάληψη διασταυρωμένης επικύρωσης, επιλέγονται οι λέξεις που εμφανίζονται πάνω από ν = 4 φορές στα δεδομένα εκπαίδευσης εκείνης της επανάληψης.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	78,63	84,82	23,47	36,77
40	128	81,17	92,46	32,02	47,57
60	193	83,89	90,94	43,93	59,24
80	258	85,44	92,05	50,36	65,11
100	323	86,28	89,3	55,71	68,62

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	87,23	92,59	54,4	68,61
40	126	90,95	95,19	67,62	79,07
60	190	93,96	94,58	80,85	87,18
80	254	94,99	95,43	84,54	89,65
100	318	95,39	95,66	85,89	90,51

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	92,58	89,57	71,18	79,32
40	97	94,17	91,68	78,71	84,71
60	146	94,74	91,91	81,12	86,18
80	195	94,91	92,3	81,98	86,84
100	244	94,99	93,16	82,14	87,3

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	96,09	97,98	82,77	89,73
40	109	97,36	98,85	88,18	93,21
60	163	97,74	99,66	89,31	94,2
80	218	98,64	100	93,37	96,57
100	272	99,09	99,64	95,93	97,75

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	85,29	67,58	20,81	31,82
40	87	89,4	87,23	48,31	62,19
60	130	90,79	75,98	55,27	63,99
80	173	92,43	91,02	65,5	76,18
100	217	93,1	90,57	68,72	78,15

Αποτελέσματα ερωτήσεων οργανισμών

Στους παραπάνω πίνακες φαίνονται τα αποτελέσματα των αντιστοίχων πειραμάτων, όταν χρησιμοποιείται σε κάθε επανάληψη διασταυρωμένης επικύρωσης το x% των δεδομένων εκπαίδευσης. Οι αντίστοιχες καμπύλες μάθησης (βλ. το CD που συνοδεύει το παρόν κείμενο) είναι γενικά αύξουσες συναρτήσεις του πλήθους των ερωτήσεων εκπαίδευσης, κάτι που δείχνει ότι τα αποτελέσματα του συστήματος θα μπορούσαν να βελτιωθούν περαιτέρω χρησιμοποιώντας περισσότερες ερωτήσεις εκπαίδευσης.

Τα υψηλότερα ποσοστά επιτυχίας εμφανίζονται στις χρονικές ερωτήσεις, με τις ερωτήσεις τοποθεσιών και ορισμών να ακολουθούν. Οι ερωτήσεις προσώπων και οργανισμών, ωστόσο, παρουσιάζουν χαμηλότερα ποσοστά επιτυχίας, ενδεχομένως λόγω της μεγαλύτερης ποικιλίας λέξεων που μπορούν να χρησιμοποιηθούν για τη διατύπωσή τους, κάτι που έχει ως αποτέλεσμα οι περισσότερες από τις λέξεις τους να μην υπερβαίνουν το κατώφλι εμφανίσεων ν και να μην επιλέγονται ως ιδιότητες. Τα χαμηλότερα αποτελέσματα αυτών των δύο κατηγοριών ενδέχεται να οφείλονται και στο ότι οι ερωτήσεις τους είναι δυσκολότερο να διαχωριστούν.

3.6 ΑΠΟΚΟΠΗ ΚΑΤΑΛΗΞΕΩΝ ΣΥΧΝΩΝ ΛΕΞΕΩΝ

Στην περίπτωση της κατάταξης ερωτήσεων με την ανίχνευση συχνά εμφανιζόμενων λέξεων της προηγούμενης ενότητας, οι λέξεις «ηθοποιού» και «ηθοποιός» στις δύο παρακάτω ερωτήσεις θεωρούνται διαφορετικές λέξεις.

Ποιου ηθοποιού είναι σύζυγος η Τζένιφερ Άννιστον;
Ποιος ηθοποιός πρωταγωνίστησε στην ταινία Καζαμπλάνκα;

Αν ωστόσο στις λέξεις αυτές είχε εφαρμοστεί προηγουμένως αποκοπή καταλήξεων (stemming), θα θεωρούνταν η ίδια λέξη (ακριβέστερα, θέμα ή ρίζα) και θα καταχωρούνταν με μεγαλύτερο αριθμό εμφανίσεων. Ωστόσο, η αποκοπή καταλήξεων δεν έδειξε να βελτιώνει τα αποτελέσματα όλων των κατηγοριών, όπως φαίνεται και στους παρακάτω πίνακες αποτελεσμάτων.[6] Για την ακρίβεια, υπήρχε μια μικρή άνοδος στις κατηγορίες ερωτήσεων προσώπων και οργανισμών, ενώ στις υπόλοιπες κατηγορίες έδειξε να προσθέτει περισσότερο θόρυβο παρά βελτίωση.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	97,38	85,74	24,24	37,79
40	128	82,42	86,97	40,27	55,05
60	193	83,89	89,57	44,87	59,79
80	258	85,46	89,4	52,09	65,83
100	323	86,6	90,68	55,97	69,22

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	85	88,97	47,2	61,68
40	126	88,01	91,85	57,83	70,97
60	190	90,48	92,89	67,37	78,1
80	254	90,87	93,2	69,2	79,43
100	318	91,04	93,49	70,2	80,19

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	92,77	90,07	72,17	80,14
40	97	94,49	91,39	80,67	85,7
60	146	94,74	91,06	81,89	86,23
80	195	94,9	91,37	82,66	86,79
100	244	94,9	91,56	82,75	86,94

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	95,42	96,93	80,52	87,97
40	109	96,62	96,61	86,81	91,45
60	163	97,07	97,62	87,96	92,54
80	218	97,52	96,44	91,7	94,01
100	272	97,44	97,24	90,43	93,72

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

6 Χρησιμοποιήσαμε μια μονάδα αποκοπής ελληνικών καταλήξεων που είχε υλοποιηθεί από τη Σταματίνα Χαραμή.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	86,76	75,9	30,72	43,74
40	87	91,29	88,06	59,13	70,75
60	130	93,1	91,79	68,45	78,42
80	173	93,91	90,9	74,63	81,97
100	217	94,74	90,93	79,36	84,754

Αποτελέσματα ερωτήσεων οργανισμών

3.7 ΧΡΗΣΗ ΑΠΟΣΤΑΣΕΩΝ ΑΠΟ ΤΗΝ ΑΡΧΙΚΗ ΛΕΞΗ

Παρ' όλο που η χρήση των συχνών λέξεων κάθε κατηγορίας φαίνεται αρκετά αποδοτική, είναι πολλές φορές δυνατό να προκληθεί θόρυβος από ερωτήσεις που τυχαίνει να περιέχουν μια συχνά εμφανιζόμενη λέξη μιας άλλης διαφορετικής κατηγορίας. Για παράδειγμα, η παρακάτω χρονική ερώτηση:

Ποιο έτος ιδρύθηκε η πόλη της Κολωνίας από τους Ρωμαίους;

περιέχει τη λέξη «πόλη» που εμφανίζεται συχνότερα στις ερωτήσεις τοποθεσιών. Για το λόγο αυτό, στα πειράματα αυτής της ενότητας, εκτός από τις δυαδικές ιδιότητες της ενότητας 3.5, χρησιμοποιήθηκαν και αριθμητικές ιδιότητες , των οποίων οι τιμές ήταν οι αποστάσεις (μετρούμενες σε λέξεις) των αντιστοίχων (συχνών) λέξεων από την αρχή της ερώτησης. Αν μια συχνή λέξη δεν εμφανίζεται στην ερώτηση, η τιμή της αντίστοιχης αριθμητικής ιδιότητας είναι -1. Οι αποστάσεις κανονικοποιούνται στο διάστημα [–1, +1]. Η κανονικοποίηση γίνεται διαιρώντας την κάθε απόσταση με ένα σταθερό αριθμό, που είναι το μήκος (σε λέξεις) της μεγαλύτερης ερώτησης των δεδομένων εκπαίδευσης. Με τον τρόπο αυτό, στο παραπάνω παράδειγμα λαμβάνουμε υπόψη ότι η λέξη «έτος» βρίσκεται πιο κοντά στην αρχή της ερώτησης από τη λέξη «πόλη».

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	79,04	84,7	24,76	38,32
40	128	83,32	93,8	40,25	56,33
60	193	84,3	88,84	46,74	61,25
80	258	84,96	88,65	50,55	64,38
100	323	89,06	93,25	64,04	75,93

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	88,01	94,48	55,99	70,32
40	126	90,4	92,22	67,7	78,08
60	190	92,22	92,17	75,81	83,19
80	254	93,88	94,68	80,52	87,02
100	318	95,31	94,4	87,43	90,78

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	92,27	89,34	70,89	79,06
40	97	94,31	90,03	81,44	85,52
60	146	94,65	89,82	83,17	86,37
80	195	94,73	90,14	83,17	86,51
100	244	95,15	90,05	85,99	87,97

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	95,55	97,27	80,74	88,24
40	109	96,99	98,9	86,31	92,18
60	163	97,97	100	90,08	94,78
80	218	98,27	98,51	93,04	95,7
100	272	99,32	100	96,73	98,33

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	85,05	70,69	22,22	33,81
40	87	89,97	90,08	50,54	64,75
60	130	90,87	86,82	58,59	69,96
80	173	91,69	87,22	63,4	73,43
100	217	93,5	93,65	68,09	78,85

Αποτελέσματα ερωτήσεων οργανισμών

Στους παραπάνω πίνακες φαίνεται πως η χρήση των αποστάσεων βελτιώνει εν γένει τα αποτελέσματα, με μόνη εξαίρεση τις ερωτήσεις οργανισμών.

3.8 ΧΡΗΣΗ ΑΚΟΛΟΥΘΙΩΝ ΣΥΝΕΧΟΜΕΝΩΝ ΛΕΞΕΩΝ

Στην περίπτωση αυτή, εκτός από τις δυαδικές ιδιότητες των συχνών λέξεων και τις ιδιότητες που δείχνουν τις αποστάσεις των συχνών λέξεων (αν εμφανίζονται) από την αρχή της ερώτησης, χρησιμοποιήθηκαν και πρόσθετες δυαδικές ιδιότητες που

δείχνουν αν υπάρχουν ή όχι συγκεκριμένες ακολουθίες συνεχόμενων λέξεων (token n-grams). Ακολουθώντας το ίδιο σκεπτικό που χρησιμοποιήθηκε για την εύρεση των συχνών λέξεων των κατηγοριών, εντοπίσαμε για κάθε κατηγορία ερωτήσεων τις ακολουθίες συνεχόμενων λέξεων μήκους από 2 ως 4 λέξεις που εμφανίζονταν περισσότερες από ν = 4 φορές στις ερωτήσεις εκπαίδευσης της κατηγορίας αυτής. (Ο υπολογισμός γίνεται εκ νέου σε κάθε επανάληψη διασταυρωμένης επικύρωσης.) Το σκεπτικό είναι πως κάποιες συχνές ακολουθίες λέξεων, όπως «Τι είναι το», «Πώς λέγεται», «Πώς ορίζεται» κ.ά. συχνά παραπέμπουν σε συγκεκριμένες κατηγορίες ερωτήσεων ή βοηθούν στο να αποκλείσουμε κάποιες άλλες.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	81,43	78,91	42,61	55,43
40	128	84,47	82,48	53,51	64,91
60	193	85,78	88,11	54,17	67,09
80	258	87,5	92,55	57,84	71,19
100	323	88,33	93,1	60,79	73,55

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	86,42	85,09	57,43	68,58
40	126	90,07	93,43	65,68	77,14
60	190	92,85	97,64	73,58	83,92
80	254	94,68	96,95	81,75	88,7
100	318	95,07	96,68	83,66	89,7

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	93,58	90,89	76,15	82,87
40	97	94,41	92,04	79,55	85,34
60	146	95,06	93,35	82,33	87,5
80	195	95,47	94,38	83,2	88,44
100	244	95,23	92,56	83,17	87,61

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	96,01	99,64	80,79	89,23
40	109	96,99	98,06	87,17	92,29
60	163	97,74	98,92	90,02	94,26
80	218	98,94	99,64	95,16	97,35
100	272	98,72	100	93,73	96,76

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	86,03	80,53	25,9	39,19
40	87	88	90,14	38,36	53,82
60	130	91,12	92,24	55,77	69,51
80	173	92,35	91,55	64,13	75,43
100	217	93,17	95,7	65,4	77,7

Αποτελέσματα ερωτήσεων οργανισμών

Αν και τα αποτελέσματα των πειραμάτων ήταν εν γένει καλύτερα από εκείνα των αρχικών πειραμάτων, όπου είχαν χρησιμοποιηθεί μόνο ιδιότητες που αντιστοιχούσαν στις συχνές λέξεις, δεν ήταν καλύτερα από τα αποτελέσματα της προηγούμενης ενότητας, όπου είχαν προστεθεί και οι αποστάσεις από την αρχή της ερώτησης.

3.9 ΧΡΗΣΗ ΑΚΟΛΟΥΘΙΩΝ ΜΗ ΣΥΝΕΧΟΜΕΝΩΝ ΛΕΞΕΩΝ

Ως επέκταση της προηγούμενης μεθόδου, χρησιμοποιήθηκαν σε αυτή την περίπτωση, αντί των πρόσθετων ιδιοτήτων της προηγούμενης ενότητας, δυαδικές ιδιότητες που αντιστοιχούσαν σε συχνές ακολουθίες λέξεων που δεν ήταν απαραίτητα συνεχόμενες μέσα στην ερώτηση.
Για παράδειγμα, μία τέτοια πιθανή ακολουθία της ερώτησης:

Ποιος είναι ο μεγαλύτερος κόλπος στον κόσμου;

είναι η: «*Ποιος μεγαλύτερος κόλπος κόσμου*». Ωστόσο, επειδή δεν κατέστη δυνατό να εξεταστούν όλοι οι δυνατοί συνδυασμοί τέτοιων ακολουθιών για κάθε ερώτηση, επιλέξαμε μόνο τις ακολουθίες που προέκυψαν παραλείποντας λέξεις όπως προθέσεις, συνδέσμους, άρθρα, κτλ., που είναι γνωστές ως stop-words.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	82,33	81,05	44,59	57,53
40	128	84,22	81,9	52,27	63,82
60	193	84,8	85,38	53,23	65,58
80	258	85,61	91,24	51,37	65,73
100	323	86,35	91,69	53,49	67,56

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	86,82	90,61	53,83	67,54
40	126	90,78	93,02	69,14	79,32
60	190	91,89	95,12	72	81,96
80	254	94,35	96,27	80,77	87,84
100	318	95,15	97	83,66	89,84

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	93,42	91,6	75	82,47
40	97	94,42	91,71	80,09	85,51
60	146	94,26	91,27	79,61	85,04
80	195	95,14	93,11	82,75	87,63
100	244	95,06	92,55	82,24	87,09

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	95,72	99,09	79,86	88,44
40	109	97,14	99,14	86,75	92,53
60	163	97,37	99,25	87,91	93,24
80	218	98,57	99,64	93,43	94,63
100	272	99,02	99,28	95,98	97,6

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	85,62	89,79	21,22	34,33
40	87	88,91	87,31	42,59	57,25
60	130	91,04	92,19	56,18	69,81
80	173	92,44	89,86	65,81	75,98
100	217	93,25	94,35	66,63	78,11

Αποτελέσματα ερωτήσεων οργανισμών

Σε γενικές γραμμές, τα αποτελέσματα αυτής της μεθόδου δε διέφεραν ιδιαίτερα από τα αντίστοιχα με συνεχόμενες ακολουθίες λέξεων, όπως είχε επισημανθεί και στην εργασία του Δημήτρη Μαυροειδή [15] . Πιο συγκεκριμένα, υπήρχε μια αισθητή μείωση στην κατηγορία των προσώπων, ενώ για τις υπόλοιπες κατηγορίες σημειώθηκε ανεπαίσθητη βελτίωση. Τα αποτελέσματα εικονίζονται και στους παραπάνω πίνακες.

3.10 ΑΠΟΚΟΠΗ ΚΑΤΑΛΗΞΕΩΝ ΣΤΙΣ ΑΚΟΛΟΥΘΙΕΣ ΛΕΞΕΩΝ

Σε μια προσπάθεια να επεκταθεί η χρήση των ακολουθιών λέξεων, δοκιμάσαμε να αποκόψουμε τις καταλήξεις των λέξεων των συχνών συνεχόμενων ακολουθιών. Ωστόσο, τα αποτελέσματα δεν ήταν ιδιαιτέρως καλύτερα από εκείνα που είχαμε πάρει χρησιμοποιώντας ακολουθίες λέξεων χωρίς αποκοπή καταλήξεων, πιθανότατα για τους ίδιους λόγους που δεν απέδωσε και η χρήση της αποκοπής καταλήξεων στις μεμονωμένες λέξεις. Στον παρακάτω πίνακα βλέπουμε ότι με εξαίρεση κάποιες μικρές βελτιώσεις στις ερωτήσεις ορισμών και τοποθεσιών, δεν υπήρξε ουσιαστική

άνοδος της συνολικής επίδοσης.

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	81,58	83,66	40,07	54,18
40	128	84,14	81,37	55,18	65,76
60	193	85,29	86,52	54,46	66,85
80	258	86,69	91,92	55,4	69,13
100	323	87,83	94,22	57,75	71,61

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	88,18	91,39	59,72	72,24
40	126	90,71	96,56	66,06	78,45
60	190	93,41	95,64	77,7	85,74
80	254	94,6	96,54	81,79	88,55
100	318	95,55	98,49	83,95	90,64

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	93,85	91,49	77,08	83,67
40	97	94,49	90,94	80,73	85,53
60	146	95,15	93,43	81,63	87,13
80	195	95,98	95,7	84,03	89,49
100	244	96,38	97,71	83,97	90,32

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	96,01	99,58	80,85	89,24
40	109	97,06	99,58	85,96	92,27
60	163	97,52	99,64	88,24	93,59
80	218	98,8	99,64	94,61	97,06
100	272	98,41	100	92,22	95,95

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	85,45	72,25	22,27	34,05
40	87	89,81	92,11	47,59	62,75
60	130	92,19	92,17	61,86	74,03
80	173	93,26	93,64	67,59	78,51
100	217	93,26	92,01	68,68	78,65

Αποτελέσματα ερωτήσεων οργανισμών

3.11 ΧΡΗΣΗ ΤΟΥ ΚΕΝΤΡΙΚΟΥ ΟΥΣΙΑΣΤΙΚΟΥ

Στην ενότητα 3.5 είχε αναφερθεί το πρόβλημα (που εντοπίζεται κυρίως στις ερωτήσεις προσώπων) της αδυναμίας κάποιων σημαντικών λέξεων (δηλαδή λέξεων που σηματοδοτούν την κατηγορία στην οποία ανήκει η ερώτηση) να ξεπεράσουν το κατώφλι αριθμού εμφανίσεων που απαιτείται για να αντιπροσωπευτούν από ιδιότητες. Η αντιμετώπιση αυτού του προβλήματος γίνεται μέσω του εντοπισμού των λέξεων αυτών, που είναι τις περισσότερες φορές τα «κεντρικά ουσιαστικά» (π.χ. «Ποια είναι η πρωτεύουσα της Λιβύης;») των ερωτήσεων, με το σύστημα αναγνώρισης μερών του λόγου (POS Tagger) των Μαλακασιώτη και Χρονάκη [14, 16]. Το κεντρικό ουσιαστικό κάθε ερώτησης θεωρήσαμε ότι είναι το πρώτο ουσιαστικό μετά την ερωτηματική αντωνυμία (αν αυτή υπάρχει, διαφορετικά το πρώτο ουσιαστικό της ερώτησης) και αυτή η προσέγγιση αποδεικνύεται ικανοποιητική στην πλειοψηφία των ερωτήσεων.

Η εύρεση του κεντρικού ουσιαστικού της ερώτησης έχει αποδειχθεί ότι βοηθάει στην αύξηση της επίδοσης ενός συστήματος κατάταξης ερωτήσεων [7], κυρίως με την ταυτόχρονη χρήση ιεραρχικού θησαυρού (π.χ. Wordnet). Ωστόσο, επειδή δεν διαθέταμε ελληνικό ιεραρχικό θησαυρό, απλά προσθέσαμε τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (κάθε επανάληψης διασταυρωμένης επικύρωσης) στις συχνά εμφανιζόμενες λέξεις (για τις οποίες έχουμε ιδιότητες), ανεξάρτητα του αριθμού εμφανίσεων των κεντρικών ουσιαστικών στις ερωτήσεις εκπαίδευσης. Στον παρακάτω πίνακα παρατίθενται τα αντίστοιχα αποτελέσματα, από τα οποία γίνεται εμφανές ότι υπάρχει αισθητή βελτίωση σε όλες τις κατηγορίες και σε όλα τα μεγέθη. Στην περίπτωση αυτή χρησιμοποιούμε τις ιδιότητες των ενοτήτων 3.5 (συχνές λέξεις), 3.7 (αποστάσεις από την αρχή της ερώτησης) και 3.8 (ακολουθίες συχνών λέξεων), αλλά οι συχνές λέξεις περιλαμβάνουν και όλα τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (κάθε επανάληψης διασταυρωμένης επικύρωσης).

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	64	85,36	79,12	62,51	69,84
40	128	88	85,47	66,91	75,06
60	193	88,98	85,45	71,83	78,05
80	258	89,56	84,86	74,54	79,36
100	323	90,63	88,6	75,22	81,36

Αποτελέσματα ερωτήσεων προσώπων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	63	91,11	93,99	70,06	80,28
40	126	93,01	93,77	78,02	85,17
60	190	94,28	93,41	83,56	88,21
80	254	95,47	94,33	87,68	90,88
100	318	96,03	95,08	89,37	92,13

Αποτελέσματα ερωτήσεων τοποθεσιών

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	48	94,23	91,06	79,83	85,08
40	97	94,9	91,11	83,2	86,98
60	146	95,15	91,45	84,45	87,81
80	195	95,89	94,77	84,77	89,49
100	244	96,05	94,66	85,54	89,87

Αποτελέσματα ερωτήσεων ορισμού

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	54	97,22	98,8	87,55	92,83
40	109	98,35	98,66	93,46	95,99
60	163	99,09	99,66	95,96	97,77
80	218	99,09	99,66	95,93	97,76
100	272	99,62	100	98,15	99,07

Αποτελέσματα ερωτήσεων χρονικών εκφράσεων

% ερωτήσεων εκπαίδευσης	#ερωτήσεων εκπαίδευσης	Accuracy	Precision	Recall	f-measure
20	43	88,2	84,9	42,49	56,64
40	87	91,86	87,67	63,68	73,77
60	130	93,58	87,28	75,9	81,2
80	173	94,49	90,34	77,49	83,43
100	217	95,23	91,42	81,04	85,92

Αποτελέσματα ερωτήσεων οργανισμών

Ωστόσο, συνεχίζουν να υπάρχουν ερωτήσεις που είναι δύσκολο να ταξινομηθούν σωστά, όπως οι παρακάτω,

Τι ανακάλυψε ο Χριστόφορος Κολόμβος;
Τι ανακάλυψε η Μαρία Κιουρί;

οι οποίες έχουν σχεδόν την ίδια διατύπωση και πιθανόν να αντιστοιχούσαν στο ίδιο διάνυσμα ιδιοτήτων. Ωστόσο, οι κατηγορίες τους είναι εντελώς διαφορετικές και για το διαχωρισμό τους απαιτούνται εγκυκλοπαιδικές γνώσεις που δεν διαθέτει το σύστημα της εργασίας.

3.12 ΤΕΛΙΚΑ ΠΕΙΡΑΜΑΤΑ

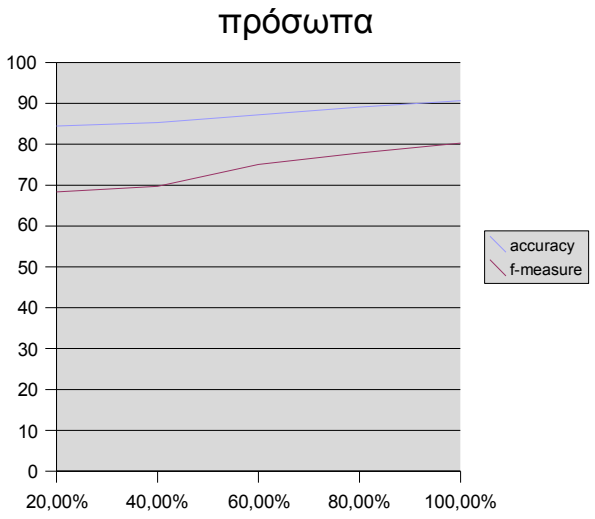
Στα τελικά πειράματα αυτής της ενότητας διαλέξαμε για κάθε μία από τις πέντε Μ.Δ.Υ. το συνδυασμό ιδιοτήτων που την είχε οδηγήσει στα καλύτερα αποτελέσματα:

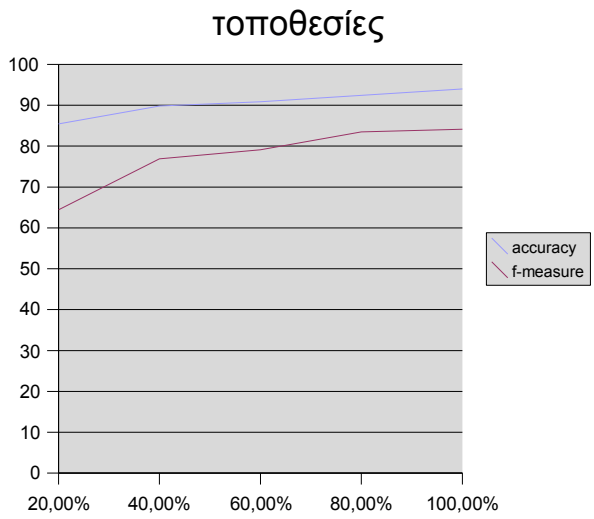
- Στη Μ.Δ.Υ. των ερωτήσεων προσώπων, χρησιμοποιήσαμε τις ιδιότητες των συχνά εμφανιζόμενων λέξεων (ενότητα 3.5), τις ιδιότητες των συχνά εμφανιζόμενων λέξεων χωρίς καταλήξεις (ενότητα 3.6), τις ιδιότητες των αποστάσεων των συχνά εμφανιζόμενων λέξεων (με καταλήξεις) από την αρχή της ερώτησης (ενότητα 3.7) και τις ιδιότητες των συνεχόμενων ακολουθιών λέξεων (ενότητα 3.8), σε κάθε περίπτωση συμπεριλαμβάνοντας στις συχνές λέξεις τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (ενότητα 3.11).
- Στη Μ.Δ.Υ. των ερωτήσεων τοποθεσιών, χρησιμοποιήσαμε τις ιδιότητες των συχνά εμφανιζόμενων λέξεων (ενότητα 3.5), τις ιδιότητες των αποστάσεων των συχνά εμφανιζόμενων λέξεων (με καταλήξεις) από την αρχή της ερώτησης (ενότητα 3.7) και τις ιδιότητες των συνεχόμενων ακολουθιών λέξεων (ενότητα 3.8), σε κάθε περίπτωση συμπεριλαμβάνοντας στις συχνές λέξεις τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (ενότητα 3.11).
- Στη Μ.Δ.Υ. των ερωτήσεων ορισμού, χρησιμοποιήσαμε τις ιδιότητες των συχνά εμφανιζόμενων λέξεων (ενότητα 3.5), τις ιδιότητες των αποστάσεων των συχνά εμφανιζόμενων λέξεων (με καταλήξεις) από την αρχή της ερώτησης (ενότητα 3.7) και τις ιδιότητες των συνεχόμενων ακολουθιών λέξεων (ενότητα 3.8), σε κάθε περίπτωση συμπεριλαμβάνοντας στις συχνές λέξεις τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (ενότητα 3.11).
- Στη Μ.Δ.Υ. των ερωτήσεων χρονικών εκφράσεων, χρησιμοποιήσαμε τις ιδιότητες των συχνά εμφανιζόμενων λέξεων (ενότητα 3.5), τις ιδιότητες των αποστάσεων των συχνά εμφανιζόμενων λέξεων (με καταλήξεις) από την αρχή της ερώτησης (ενότητα 3.7) και τις ιδιότητες των συνεχόμενων ακολουθιών λέξεων (ενότητα 3.8), σε κάθε περίπτωση συμπεριλαμβάνοντας στις συχνές λέξεις τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (ενότητα 3.11).
- Στη Μ.Δ.Υ. των ερωτήσεων οργανισμών, χρησιμοποιήσαμε τις ιδιότητες των συχνά εμφανιζόμενων λέξεων (ενότητα 3.5), τις ιδιότητες των συχνά εμφανιζόμενων λέξεων χωρίς καταλήξεις (ενότητα 3.6), τις ιδιότητες των αποστάσεων των συχνά εμφανιζόμενων λέξεων (με καταλήξεις) από την αρχή της ερώτησης (ενότητα 3.7) και τις ιδιότητες των συνεχόμενων ακολουθιών λέξεων (ενότητα 3.8), σε κάθε περίπτωση συμπεριλαμβάνοντας στις συχνές λέξεις τα κεντρικά ουσιαστικά των ερωτήσεων εκπαίδευσης (ενότητα 3.11).

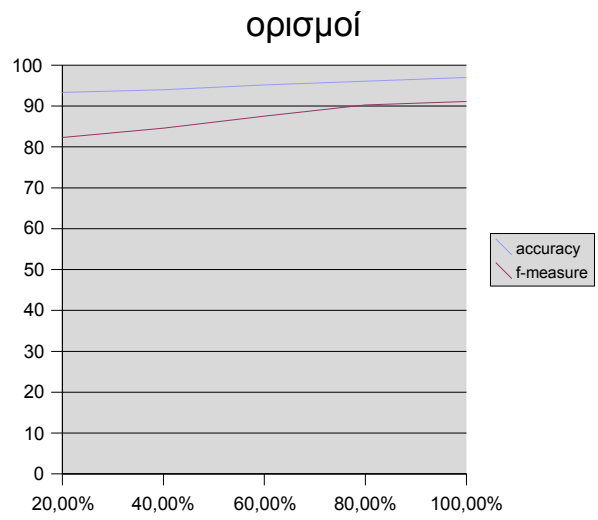
Επιπλέον θεωρήσαμε ότι μία ερώτηση δεν είναι δυνατόν να ανήκει σε περισσότερες από μία κατηγορίες. Όποτε έπρεπε να καταταγεί μια ερώτηση, τη δίνουμε (τα αντίστοιχα διανύσματα) στις πέντε Μ.Δ.Υ. και κατατάσσουμε την ερώτηση στην κατηγορία ερωτήσεων που αντιστοιχούσε στη Μ.Δ.Υ. που επέστρεψε το μεγαλύτερο βαθμό βεβαιότητας πως η ερώτηση ανήκει στη θετική της κατηγορία. Στον παρακάτω πίνακα φαίνονται αναλυτικά τα αποτελέσματα για κάθε μία κατηγορία ερωτήσεων, συνοδευόμενα από τις αντίστοιχες καμπύλες μάθησης.

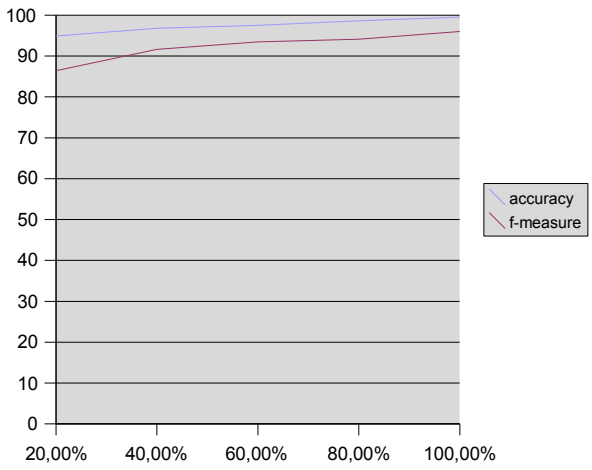
	πρόσωπα		τοποθεσίες		ορισμοί		χρονικ. εκφράσ.		οργανισμοί	
%ερωτήσεων εκπαίδευσης	accuracy	f-measure	accuracy	f-measure	accuracy	f-measure	accuracy	f-measure	accuracy	f-measure
20%	84,47	68,36	85,47	64,43	93,35	82,31	94,89	86,39	84,38	28,52
40%	85,29	69,74	89,84	76,86	94	84,61	96,77	91,62	87,01	46,81
60%	87,17	75,08	90,87	79,09	95,15	87,54	97,51	93,49	88,98	60,16
80%	89,05	77,85	92,45	83,47	96,05	90,23	98,61	94,14	90,14	71,28
100%	90,67	80,25	94,02	84,14	97,01	91,13	99,51	96,03	92,11	83,65

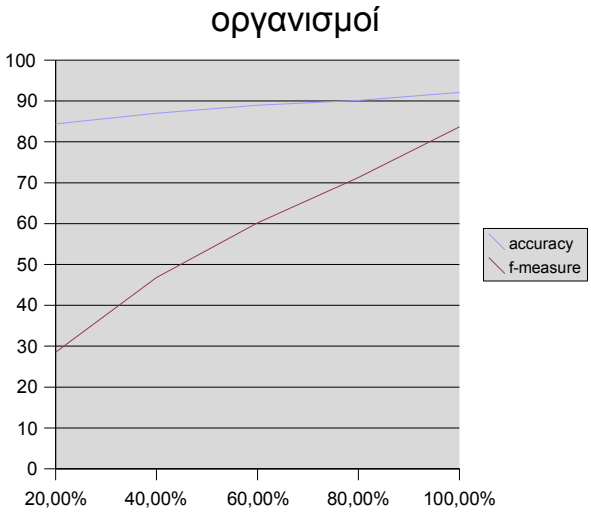
Αποτελέσματα Τελικών Πειραμάτων





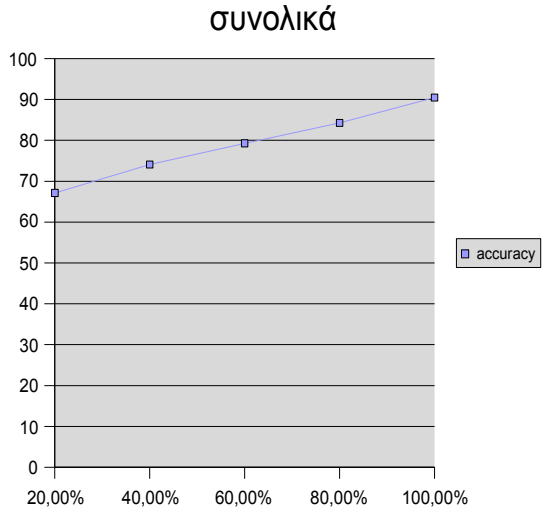






Επιπλέον, παρατίθενται στη συνέχεια τα ποσοστά των ερωτήσεων που κατετάγησαν σωστά από το συνολικό σύστημα, το οποίο κατατάσσει κάθε ερώτηση σε μία μόνο κατηγορία. Η καμπύλη μάθησης του συνολικού συστήματος είναι σαφώς αύξουσα συνάρτηση του πλήθους των ερωτήσεων εκπαίδευσης. Τα αποτελέσματα δείχνουν πως οι επιδόσεις του συστήματος μπορούν να βελτιωθούν περαιτέρω χρησιμοποιώντας περισσότερες ερωτήσεις εκπαίδευσης.

#ερωτ. Εκπ.	Accuracy
20,00%	67,1
40,00%	74,09
60,00%	79,25
80,00%	84,27
100,00%	90,45



4.ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΒΕΛΤΙΩΣΕΙΣ

Τα αποτελέσματα της εργασίας ήταν αρκετά ενθαρρυντικά και επιβεβαίωσαν σε μεγάλο βαθμό τα ποσοστά επιτυχίας που είχαν ανακοινώσει οι Li και Roth [12] για τις κύριες (coarse) κατηγορίες αγγλικών ερωτήσεων. Το τελικό σύστημα της παρούσας εργασίας κατατάσσει σωστά πάνω από το 90% των ελληνικών ερωτήσεων.

Το σύστημα έχει ακόμα πολλά περιθώρια βελτίωσης, τόσο μέσω της αύξησης του πλήθους των ερωτήσεων εκπαίδευσης, όσο και με τη χρήση πιο εξειδικευμένων γλωσσικών εργαλείων και πόρων. Για παράδειγμα, η χρήση ενός ιεραρχικού θησαυρού της ελληνικής γλώσσας ήταν αδύνατη στην παρούσα εργασία, αλλά ενδέχεται να προσφέρει πολλά στο μέλλον, όπως έχουν δείξει αντίστοιχες εργασίες για αγγλικές ερωτήσεις [6]. Το ίδιο ισχύει και για τους ελληνικούς συντακτικούς αναλυτές. Επίσης, η χρήση του διαδικτύου ή οποιασδήποτε μεγάλης συλλογής εγγράφων μπορεί να φανεί χρήσιμη, όπως έχουν δείξει ανάλογες εργασίες [8, 5], αφού με αυτόν τον τρόπο το σύστημα έχει πρόσβαση σε ένα ευρύ πεδίο γνώσεων και πληροφοριών.

ΑΝΑΦΟΡΕΣ

[1] D. Zhang και W.S. Lee, "Question Classification using Support Vector Machines". Πρακτικά του *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, σελ. 26–32, Τορόντο, Καναδάς, 2003.

[2] E.M. Voorhees και H.T. Dang, "The TREC Question-Answering Track". *Natural Language Engineering*, 7(4):361–378, 2001.

[3] G. Lucarelli, X. Vasilakos και I. Androutsopoulos, "Named Entity Recognition in Greek Texts with an Ensemble of SVMs and Active Learning". *International Journal on Artificial Intelligence Tools*, υπό δημοσίευση.

[4] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos και P. Stamatopoulos, "Stacking Classifiers for Anti-Spam Filtering of E-Mail". Πρακτικά του *6th Conference on Empirical Methods in Natural Language Processing* (EMNLP 2001), Carnegie Mellon University, Pittsburgh, PA, ΗΠΑ, σελ. 44–50, 2001.

[5] H. T. Ng, J. L. P. Kwan, Y. Xin, "Question Answering Using a Large Text Database: A Machine Learning Approach". Πρακτικά του *6th Conference on Empirical Methods in Natural Language Processing* (EMNLP 2001), Carnegie Mellon University, Pittsburgh, PA, ΗΠΑ, σελ. 67-73, 2001

[6] J. Suzuki, H. Taira, Y. Sasaki και E. Maeda, "Question Classification using HDAG Kernel". Πρακτικά του *ACL Workshop on Multilingual Summarization and Question Answering*, Sapporo, Ιαπωνία, σελ. 61–68, 2003.

[7] S. M. Harabagiu, S. J. Maiorano και M. A. Pasca, "Open-Domain Textual Question Answering Techniques". *Natural Language Engineering*, 9(3):231–267, 2003.

[8] T. Solorio, M. Perez-Coutino, M. Montes-y-Gomez, L. Vilasenor-Pineda και A. Lopez-Lopez, "A Language Independent Method for Question Classification". Πρακτικά του *20th International Conference on Computational Linguistics,* (COLING-04), Γενεύη, Ελβετία, σελ. 1374–1380, 2004.

[9] V. Krishnan, S. Das, και S. Chakrabarti, "Enhanced Answer Type Inference from Questions using Sequential Models". Πρακτικά του *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP 2005), Βανκούβερ, Καναδάς, σελ. 315–322, 2005.

[10] S. Miliaraki και I. Androutsopoulos, “Learning to Identify Single-Snippet Answers to Definition Questions”. Πρακτικά του *20th International Conference on Computational Linguistics* (COLING 2004), Γενεύη, Ελβετία, σελ. 1360–1366, 2004.

[11] V.P. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[12] X. Li και D. Roth, “Learning Question Classifiers: the Role of Semantic Information”. *Natural Language Engineering*, 12(3):229–249, 2006.

[13] Γ. Λουκαρέλλι, *Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων σε Ελληνικά Κείμενα*. Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[14] Π. Μαλακασιώτης, *Αναγνώριση Μερών του Λόγου σε Ελληνικά Κείμενα με Τεχνικές Ενεργητικής Μάθησης*. Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[15] Δ. Μαυροειδής, *Αυτόματη Κατάταξη Ερωτήσεων Φυσικής Γλώσσας σε Κατηγορίες*. Πτυχιακή Εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[16] Ι. Χρονάκης, *Επεκτάσεις και Περαιτέρω Αξιολόγηση Συστήματος Αναγνώρισης Μερών του Λόγου για Ελληνικά Κείμενα.* Πτυχιακή Εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.