ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα*»

Λουκαρέλλι Γιώργος

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2005

Περίληψη

Η αναγνώριση και κατηγοριοποίηση ονομάτων οντοτήτων είναι μία ιδιαίτερα χρήσιμη υπο-εργασία σε πολλές εφαρμογές επεξεργασίας φυσικής γλώσσας. Σε αυτήν την εργασία παρουσιάζεται μία προσπάθεια αναγνώρισης και κατηγοριοποίησης ονομάτων προσώπων και χρονικών εκφράσεων χρησιμοποιώντας Μηχανές Διανυσμάτων Υποστήριξης και ημι-αυτόματα παραγόμενα πρότυπα αντίστοιχα. Το σύστημα που αναπτύχθηκε ελέγχθηκε σε δύο διαφορετικές συλλογές ελληνικών κειμένων με ικανοποιητικά αποτελέσματα. Επιπλέον, διερευνήθηκαν τα αποτελέσματα της χρήσης ενεργητικής μάθησης και βρέθηκε πως η ενεργητική μάθηση βοηθάει σημαντικά στη μείωση του απαιτούμενου αριθμού των επισημειωμένων κειμένων εκπαίδευσης.

Abstract

Named entity recognition and categorization is a very important subtask in several natural language processing applications. We present an attempt to recognize and categorize person names and temporal expressions by using Support Vector Machines and semi-automatically produced patterns, respectively. The resulting system was tested in two different collections of Greek texts with satisfactory results. Moreover, the effects of active learning were explored, and it was found that active learning helps reduce significantly the required number of tagged training texts.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ίωνα Ανδρουτσόπουλο για τη βοήθεια και την καθοδήγηση που μου πρόσφερε κατά τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Επίσης, ευχαριστώ τον κ. Θεόδωρο Καλαμπούκη που αποδέχτηκε το ρόλο του δεύτερου αξιολογητή και μου διέθεσε μία από τις δύο συλλογές κειμένων που χρησιμοποίησα. Τέλος, ευχαριστώ τον Πρόδρομο (Μάκη) Μαλακασιώτη για τις χρήσιμες και παραγωγικές συζητήσεις μας και την ανταλλαγή ιδεών και απόψεων.

Περιεχόμενα

1 Εισαγωγή

Η εργασία της αναγνώρισης και κατηγοριοποίησης ονομάτων οντοτήτων (named-entity recognition and categorization) αποσκοπεί στον εντοπισμό και την κατάταξη σε κατηγορίες ονομάτων οντοτήτων που εμφανίζονται σε συλλογές κειμένων. Για παράδειγμα, είναι δυνατόν ο στόχος να είναι ο εντοπισμός ονομάτων προσώπων και εταιριών, τοπωνυμίων, ημερομηνιών, αριθμητικών εκφράσεων, ονομάτων πρωτεϊνών σε ιατρικά κείμενα, ονομάτων προϊόντων σε ιστοσελίδες κ.α. Η αναγνώριση και κατηγοριοποίηση ονομάτων οντοτήτων αποτελεί προκαταρκτικό στάδιο σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας, όπως τα συστήματα εξαγωγής πληροφοριών από κείμενα και τα συστήματα ερωταποκρίσεων για συλλογές κειμένων.

Με την αναγνώριση και κατηγοριοποίηση ονομάτων οντοτήτων, κυρίως για αγγλικά κείμενα, έχουν ασχοληθεί αρκετά διεθνή συνέδρια, τα σημαντικότερα των οποίων ήταν τα Message Understanding Conferences (MUC) [27], [28], στα οποία συμμετείχαν αρχικά κυρίως συστήματα που στηρίζονταν σε χειρωνακτικά κατασκευασμένους κανόνες, με τη σταδιακή προσθήκη συστημάτων που χρησιμοποιούσαν μηχανική μάθηση. Το θέμα της αναγνώρισης και κατηγοριοποίησης ονομάτων οντοτήτων έχει απασχολήσει και το συνέδριο Computational Natural Language Learning [1]. Ένα μέτρο αξιολόγησης που χρησιμοποιείται σε αυτή την περιοχή είναι το F-measure, ένας συνδυασμός ανάκλησης και ακρίβειας που προέρχεται από την ανάκτηση πληροφοριών και ορίζεται παρακάτω. Το F-measure των συστημάτων αναγνώρισης και κατηγοριοποίησης ονομάτων οντοτήτων έχει πλέον ξεπεράσει το 93-94% για τα αγγλικά κείμενα. Για τα ελληνικά κείμενα έχουν επίσης γίνει σχετικές προσπάθειες, για παράδειγμα από το Ινστιτούτο Επεξεργασίας του Λόγου και το Ε.Κ.Ε.Φ.Ε. «Δημόκριτος», με μεθόδους παρόμοιες εκείνων που έχουν χρησιμοποιηθεί για αγγλικά κείμενα. Τα περισσότερα από τα ελληνικά συστήματα, όμως, δεν είναι ελεύθερα διαθέσιμα.

Σε αυτήν την εργασία παρουσιάζεται μία προσπάθεια κατασκευής ενός ελεύθερα διαθέσιμου συστήματος αναγνώρισης και κατάταξης ονομάτων οντοτήτων για ελληνικά κείμενα εφημερίδων. Πιο συγκεκριμένα, η προσπάθεια αφορά αναγνώριση ονομάτων προσώπων και χρονικών εκφράσεων. Ελπίζουμε ότι το σύστημα που αναπτύχθηκε θα επεκταθεί στη διάρκεια μελλοντικών εργασιών, ώστε να υποστηρίζει και άλλες κατηγορίες ονομάτων. Το σύστημα χρησιμοποιεί Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines) και ημι-αυτόματα παραγόμενα πρότυπα (patterns). Οι ΜΔΥ χρησιμοποιούνται για την αναγνώριση ονομάτων προσώπων, ενώ τα πρότυπα για τις χρονικές εκφράσεις.

Βασικός στόχος της εργασίας ήταν το σύστημα να είναι δυνατόν να χρησιμοποιηθεί με διαφορετικές συλλογές κειμένων. Για αυτόν το λόγο χρησιμοποιήθηκαν δύο διαφορετικές συλλογές κειμένων για τη διεξαγωγή πειραμάτων, μία ποικίλης θεματολογίας και μία οικονομικών κειμένων. Στην πρώτη συλλογή το F-measure της κατηγορίας των ονομάτων προσώπων πλησιάζει το 87,5%, ενώ της κατηγορίας των χρονικών εκφράσεων ξεπερνάει το 94,5%. Για τη δεύτερη συλλογή τα αντίστοιχα αποτελέσματα είναι 93,34% και 96,46%. Επίσης, στην περίπτωση των ονομάτων προσώπων χρησιμοποιήθηκαν τεχνικές ενεργητικής μάθησης, ώστε να μειωθεί ο αριθμός των κειμένων εκπαίδευσης που πρέπει να επισημειωθούν χειρωνακτικά όταν το σύστημα επανεκπαιδεύεται για νέα συλλογή

κειμένων. Τα πειραματικά αποτελέσματα της εργασίας δείχνουν ότι η χρήση ενεργητικής μάθησης επιτυγχάνει αυτόν τον στόχο.

Στη συνέχεια, στο κεφάλαιο 2, θα γίνει μία σύντομη βιβλιογραφική επισκόπηση της περιοχής της αναγνώρισης και κατάταξης ονομάτων οντοτήτων. Επίσης, θα γίνει μία εισαγωγή στη μηχανική μάθηση και ιδιαίτερα στις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines).

Στο κεφάλαιο 3 θα αναλυθεί η αρχιτεκτονική του συστήματος που αναπτύχθηκε, θα παρουσιαστεί ο τρόπος εφαρμογής της μηχανικής μάθησης, θα παρουσιαστούν οι τεχνικές ενεργητικής μάθησης που χρησιμοποιήθηκαν στην αναγνώριση ονομάτων προσώπων και θα περιγραφεί η διαδικασία κατασκευής προτύπων για την αναγνώριση των χρονικών εκφράσεων.

Στο κεφάλαιο 4 παρουσιάζεται η μορφή των κειμένων που χρησιμοποιήθηκαν, οι κανόνες επισημείωσης των παραδειγμάτων εκπαίδευσης και ελέγχου και τα πειραματικά αποτελέσματα.

Στο κεφάλαιο 5 συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται μελλοντικές επεκτάσεις του συστήματος.

Τέλος, στο παράρτημα δίνονται επιπλέον λεπτομέρειες για την υλοποίηση των Μηχανών Διανυσμάτων Υποστήριξης που χρησιμοποιήθηκε, τις τιμές των παραμέτρων της και τον τρόπο που ενσωματώθηκε στο συνολικό σύστημα.

2 Βιβλιογραφική επισκόπηση

Οι διαφορετικές προσεγγίσεις που ακολουθούνται στη βιβλιογραφία είναι δυνατόν να κατηγοριοποιηθούν ως εξής:

- Συστήματα που χρησιμοποιούν χειρωνακτικά κατασκευασμένους κανόνες (rule-based) και προϋπάρχουσες λίστες ονομάτων (gazetteers).
- Συστήματα που χρησιμοποιούν τεχνικές μηχανικής μάθησης.
- Υβριδικά συστήματα, που συνδυάζουν τις δύο παραπάνω προσεγγίσεις.

Τα βασικά μέτρα επίδοσης των συστημάτων αναγνώρισης και κατηγοριοποίησης ονομάτων οντοτήτων είναι η ανάκληση (recall), η ακρίβεια (precision) και το F-measure, το οποίο αποτελεί ένα συνδυασμό ακρίβειας και ανάκλησης. Τα μέτρα αυτά, όπως χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος που αναπτύξαμε, ορίζονται στην ενότητα 4.2.

2.1 Συστήματα για αγγλικά κείμενα

Αναφέρουμε παρακάτω μερικά από τα πιο γνωστά συστήματα αναγνώρισης και κατάταξης ονομάτων οντοτήτων για αγγλικά κείμενα, που συμμετείχαν στους διαγωνισμούς MUC-6 και MUC-7.

Το σύστημα του Πανεπιστημίου της Νέας Υόρκης για το MUC-6 [20] χρησιμοποιεί χειρωνακτικά κατασκευασμένες κανονικές εκφράσεις (regular expressions), οι οποίες αξιοποιούν πληροφορίες όπως η ύπαρξη κεφαλαίου στην αρχή μιας λέξης ή η ύπαρξη λέξεων κλειδιών στα συμφραζόμενα (όπως “Mr.”, “Co.”, “Inc.”), και πέτυχε F-measure 88% κατά μέσο όρο για όλες τις κατηγορίες ονομάτων οντοτήτων. Για την ανάπτυξή του χρησιμοποιήθηκαν μηχανές πεπερασμένων καταστάσεων, καθώς και λίστες με ονόματα εταιριών, τοποθεσιών και προσώπων.

Το NetOwl [23] της IsoQuest, συμμετείχε στο MUC-7 και έχει γίνει πλέον εμπορικό προϊόν. Χρησιμοποιεί χειρωνακτικά κατασκευασμένα πρότυπα (patterns) και λίστες ονομάτων οντοτήτων και το F-measure του ήταν περίπου 90% στο MUC-7. Επιπλέον το σύστημα παρέχει εξειδικευμένη διαδικασία για κείμενα που είναι γραμμένα πλήρως με κεφαλαία γράμματα.

Το LaSIE του Πανεπιστημίου του Sheffield συμμετείχε στα MUC-6 και MUC-7. Το σύστημα βασίζεται σε 206 χειρωνακτικά κατασκευασμένους κανόνες γραμματικής και η συνολική ανάκληση και ακρίβεια του συστήματος ήταν 92% στο MUC-6. Χρησιμοποιείται επισημειωτής μερών του λόγου (part-of-speech tagger), διαχωριστής περιόδων (sentence splitter), καθώς και λίστες με ονόματα διάφορων κατηγοριών.

Το FACILE [10] του UMIST χρησιμοποιεί, επίσης, προσέγγιση βασισμένη σε κανόνες προτύπων, οι οποίοι κατασκευάστηκαν χειρωνακτικά. Η ακρίβεια του συστήματος ήταν 87% και η ανάκληση 78%. Το κύριο χαρακτηριστικό του συστήματος ήταν ότι τα πρότυπα λαμβάνουν υπόψη τα συμφραζόμενα, ενώ χρησιμοποιούνται βάρη, ούτως ώστε να επιλεχθεί ποιος κανόνας θα εφαρμοστεί.

Το Nymble [9] σε αντίθεση με τα παραπάνω συστήματα χρησιμοποιεί στατιστικά μοντέλα (Κρυφά Μοντέλα Markov – Hidden Markov Models). Χρησιμοποιεί 14 ιδιότητες, όπως αν η λέξη αρχίζει με κεφαλαίο γράμμα, οι οποίες χαρακτηρίζουν κάθε

λέξη. Η επίδοση του συστήματος στο MUC-7 ήταν ιδιαίτερα υψηλή με ανάκληση 89% και ακρίβεια 92%.

Το σύστημα MENE [11] του Πανεπιστημίου της Νέας Υόρκης, που συμμετείχε στο MUC-7, βασίζεται στο μοντέλο της μέγιστης εντροπίας (maximum entropy model). Χρησιμοποιεί συνολικά 29 ετικέτες για τις 8 κατηγορίες (7 κατηγορίες οντοτήτων και η κατηγορία μη-οντότητα) και συγκεκριμένα: αρχή οντότητας, τέλος οντότητας, μέση οντότητας και μονολεκτική οντότητα $(4 \cdot 7 + 1 = 29)$. Χρησιμοποιούνται δυαδικές ιδιότητες παρόμοιες με αυτές του Nymble για να αναπαρασταθούν μορφολογικά χαρακτηριστικά των λέξεων, καθώς και ιδιότητες που προκύπτουν από λίστες ονομάτων προσώπων, οργανισμών και τοποθεσιών. Το F-measure που πέτυχε το σύστημα στην επίσημη αξιολόγηση του MUC-7 είναι 88,8%.

Το σύστημα της ομάδας LTG [24, 25] του Πανεπιστημίου του Εδιμβούργου ακολουθεί την υβριδική μέθοδο. Το σύστημα αυτό χρησιμοποιεί πολλαπλά περάσματα (πέντε), μία ιδέα την οποία δανειστήκαμε και εφαρμόσαμε στην κατηγορία ονομάτων προσώπων (ενότητες 3.4.1 και 3.4.2). Στο πρώτο στάδιο εφαρμόζονται χειρωνακτικά κατασκευασμένοι «σίγουροι» κανόνες, οι οποίοι βασίζονται στην ύπαρξη φράσεων όπως “Mr.”, “Dr.”, “Ltd.”, “Inc.”, και χρησιμοποιούνται λίστες με ονόματα οργανισμών και τοπωνυμίων. Σε αυτό το στάδιο αποφεύγεται να χαρακτηριστούν ως οντότητες ονόματα για τα οποία το σύστημα δεν είναι απολύτως σίγουρο. Για παράδειγμα, η λέξη “Washington” παρόλο που ανήκει στη λίστα με τα τοπωνύμια δεν χαρακτηρίζεται ως τοπωνύμιο σε αυτήν τη φάση, καθώς μπορεί να αποτελεί στη συγκεκριμένη εμφάνισή της επώνυμο ή όνομα οργανισμού. Με βάση τα αποτελέσματα του σταδίου αυτού, στο επόμενο στάδιο συλλέγονται οι εκφράσεις που χαρακτηρίστηκαν ως ονόματα οντοτήτων και γίνεται προσπάθεια να βρεθούν όλες οι διαφορετικές εμφανίσεις τους στο κείμενο. Λέγοντας διαφορετικές εμφανίσεις εννοείται ότι οι εκφράσεις διασπώνται στις διάφορες λέξεις από τις οποίες αποτελούνται, ούτως ώστε να είναι δυνατόν να εντοπιστούν ακόμα και αν δεν εμφανίζονται ολόκληρες σε άλλα σημεία. Για παράδειγμα, αν στο πρώτο στάδιο έχει εντοπιστεί το όνομα του οργανισμού “Adam Kluver Ltd”, τότε εμφανίσεις όπως “Kluver Ltd” ή “Adam Ltd” σημειώνονται στο δεύτερο στάδιο ως πιθανά ονόματα οργανισμών. Στη συνέχεια, εφαρμόζεται ένα πιθανοτικό μοντέλο μέγιστης εντροπίας. Στην περίπτωση όπου το μοντέλο χαρακτηρίσει κάποια φράση, που προέκυψε από το δεύτερο στάδιο, ως οντότητα τότε οριστικοποιείται η κατηγορία της. Στο τρίτο στάδιο εφαρμόζονται και πάλι γραμματικοί κανόνες, με τη διαφορά ότι οι αρχικοί κανόνες χαλαρώνουν, δηλαδή δεν είναι τόσο αυστηροί, και χρησιμοποιούν τα αποτελέσματα των προηγούμενων σταδίων. Σε αυτήν τη φάση χρησιμοποιείται και μία λίστα με ονόματα προσώπων. Η λίστα αυτή δεν χρησιμοποιήθηκε προηγουμένως, καθώς μπορούσε ένα όνομα προσώπου να συμμετέχει σε όνομα οργανισμού. Επίσης, λαμβάνεται απόφαση για περιπτώσεις συνένωσης. Για παράδειγμα, για τη φράση “China International Trust and Investment Corp”, την οποία οι κανόνες της πρώτης φάσης απέφυγαν να σημειώσουν καθώς δεν μπορούσαν να είναι σίγουροι αν πρόκειται για μία ή δύο εταιρίες που συνδέονται με το “and”, αποφασίζεται αν αποτελεί ένα ή δύο οργανισμούς, βάσει άλλων εμφανίσεών της στο ίδιο κείμενο. Στο τέταρτο στάδιο ακολουθείται ακριβώς η ίδια διαδικασία με το δεύτερο, χρησιμοποιώντας τις επιπλέον πληροφορίες του τρίτου σταδίου. Η τελευταία φάση αφορά αποκλειστικά κάποιους τίτλους, οι οποίοι είναι γραμμένοι με κεφαλαία γράμματα. Τελικά, το F-measure του συστήματος είναι περίπου 93%.

Στην πιο πρόσφατη βιβλιογραφία ανήκουν συστήματα που συμμετείχαν στο CoNLL-2003 [1].

Το σύστημα με τη μεγαλύτερη επιτυχία στα δεδομένα του CoNLL-2003 είναι των Florian κ.α. [19] με F-measure κοντά στο 94% στα αγγλικά κείμενα για την κατηγορία των ονομάτων προσώπων και συνολικά περίπου 89% για όλες τις κατηγορίες. Συνδυάζει τέσσερις διαφορετικούς ταξινομητές (κανόνες προτύπων, κρυφά μοντέλα Markov, συμπαγής ταξινομητής ελαχιστοποίησης του ρίσκου – robust risk minimization classifier, μοντέλο μέγιστης εντροπίας), οι οποίοι ψηφίζουν με διαφορετικό βάρος ο καθένας. Χρησιμοποιεί επίσης επισημειωτή μερών του λόγου, καθώς και λίστες με διάφορες κατηγορίες ονομάτων.

Το σύστημα των Chieu και Ng [15] βασίζεται στο μοντέλο της Μέγιστης Εντροπίας. Το ενδιαφέρον στο σύστημα είναι ότι δε λαμβάνει απλώς υπόψη τα συμφραζόμενα της υπό εξέταση λέξης και τα μορφολογικά χαρακτηριστικά της, αλλά χρησιμοποιεί και την πληροφορία προηγούμενων εμφανίσεων της λέξης στο ίδιο κείμενο. Για παράδειγμα, υπάρχει μία λίστα με λέξεις που προηγούνται από ονόματα προσώπων, η οποία κατασκευάζεται δυναμικά για κάθε κείμενο ξεχωριστά. Αν η προηγούμενη λέξη της υπό εξέτασης λέξης ανήκει σε αυτή τη λίστα, τότε ενημερώνεται η τιμή της κατάλληλης ιδιότητας. Στο CoNLL-2003, το F-measure που επιτεύχθηκε για την κατηγορία ονομάτων προσώπων είναι 93,5%, ενώ το συνολικό F-measure για όλες τις κατηγορίες ξεπερνάει το 88%.

2.2 Ελληνικά συστήματα

Τα συστήματα που έχουν αναπτυχθεί για τα ελληνικά κείμενα ακολουθούν παρόμοιες προσεγγίσεις με εκείνες των συστημάτων της προηγούμενης ενότητας. Γενικά, τα περισσότερα ελληνικά συστήματα προϋποθέτουν έναν επισημειωτή μερών του λόγου, γεγονός αναμενόμενο λόγω της περίπλοκης μορφολογίας της ελληνικής γλώσσας.

Στο [22] (Καρκαλέτσης κ.α.) συγκρίνονται δύο διαφορετικές προσεγγίσεις για τις κατηγορίες των ονομάτων προσώπων και των ονομάτων οργανισμών. Η πρώτη βασίζεται σε χειρωνακτικά κατασκευασμένους κανόνες γραμματικής. Τα υποσυστήματα που χρησιμοποιούνται είναι: διαχωριστής λεκτικών μονάδων (tokens), διαχωριστής περιόδων, επισημειωτής μερών του λόγου, αναζήτηση σε λίστες γνωστών ονομάτων και φυσικά επισημειωτής ονομάτων οντοτήτων. Τα κείμενα αποτελούνται από άρθρα γενικών θεμάτων, ενώ η ανάκληση (recall) του συστήματος για την κατηγορία ονομάτων προσώπων είναι 77% και η ακρίβεια (precision) πλησιάζει το 89%. Κατά τη δεύτερη προσέγγιση διερευνάται η χρήση του αλγόριθμου C4.5, η οποία αποφέρει καλύτερα αποτελέσματα, 95% για την ακρίβεια και 80% για την ανάκληση. Ο επισημειωτής μερών του λόγου τροφοδοτεί με ιδιότητες τον C4.5, ενώ χρησιμοποιούνται και οι λίστες ονομάτων.

Το σύστημα των Μπούτση κ.α. [12] χρησιμοποιεί 110 χειρωνακτικά κατασκευασμένους κανόνες για τις εκφράσεις ονομάτων οντοτήτων του MUC-7. Οι υποδιαδικασίες από τις οποίες αποτελείται είναι: διαχωρισμός λεκτικών μονάδων, επισημείωση μερών του λόγου, αποκοπή καταλήξεων, αναζήτηση σε λίστες ονομάτων και εφαρμογή των κανόνων. Ενδιαφέρον παρουσιάζει το γεγονός ότι επιλέχθηκαν κείμενα που περιέχουν μεγάλο αριθμό λέξεων που αρχίζουν με κεφαλαίο γράμμα, καθώς και ότι οι κανόνες λαμβάνουν υπόψη τους αυτό το γεγονός. Τα κείμενα που χρησιμοποιήθηκαν είναι κατά κύριο λόγο οικονομικού περιεχομένου (από οικονομικές εφημερίδες). Το F-measure για την κατηγορία ονομάτων προσώπων είναι 71%, ενώ το συνολικό για όλες τις κατηγορίες 83%.

Το σύστημα των Φαρμακιώτου κ.α. [18] αφορά αποκλειστικά οικονομικά κείμενα, βασίζεται σε χειρωνακτικά κατασκευασμένη γραμματική κανόνων, ενώ αποτελεί μέρος ενός μεγαλύτερου ελληνικού συστήματος εξαγωγής πληροφοριών. Το F-measure για την κατηγορία των ονομάτων προσώπων είναι 81,6%. Και σε αυτήν την περίπτωση χρησιμοποιείται επισημειωτής μερών του λόγου, καθώς και λίστες με ονόματα. Επίσης, αποκόπτεται η κατάληξη των λέξεων, απομακρύνονται οι τόνοι και οι λέξεις μετατρέπονται στις αντίστοιχες με μικρά γράμματα, ούτως ώστε να μειωθεί το μέγεθος των λιστών που χρησιμοποιούνται. Γενικά, το σύστημα χωρίζεται σε δύο στάδια. Στο πρώτο γίνεται προσπάθεια να βρεθούν τα όρια των ονομάτων οντοτήτων χρησιμοποιώντας 3 προκαθορισμένα πρότυπα, ενώ στο δεύτερο (στάδιο κατηγοριοποίησης) εφαρμόζονται οι κανόνες με τη βοήθεια των λιστών ονομάτων οντοτήτων.

Μία ενδιαφέρουσα προσέγγιση είναι το σύστημα των Πετάση κ.α. [29], όπου χρησιμοποιείται μηχανική μάθηση για να ενημερώνονται οι κανόνες της γραμματικής του συστήματος. Ουσιαστικά, πρόκειται για δύο υποσυστήματα, ένα κανόνων γραμματικής (χρησιμοποιείται το σύστημα των Φαρμακιώτου κ.α. που παρουσιάστηκε προηγουμένως) και ένα μηχανικής μάθησης όπου χρησιμοποιείται ο C4.5. Στην αρχή εφαρμόζεται το σύστημα των κανόνων και επισημειώνονται τα κείμενα. Στη συνέχεια, χρησιμοποιώντας τα κείμενα αυτά (δεν χρειάζεται επιπλέον επισημείωση) εκπαιδεύεται ο C4.5. Ακολούθως, σε μία καινούρια συλλογή εφαρμόζονται οι κανόνες και εκτελείται ο C4.5 που εκπαιδεύτηκε προηγουμένως. Τελικά, μελετάται η απόκλιση μεταξύ των δύο συστημάτων και μεταβάλλεται κατάλληλα η γραμματική του πρώτου συστήματος.

2.3 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας από τους παλαιότερους τομείς της τεχνητής νοημοσύνης [26]. Σκοπός της είναι, σε γενικές γραμμές, να κατασκευάσει συστήματα τα οποία να αποκτούν αυτόματα νέες γνώσεις από εμπειρικά δεδομένα του παρελθόντος. Στην εργασία αυτή ασχολούμαστε με μεθόδους επιβλεπόμενης μηχανικής μάθησης για το διαχωρισμό σε κατηγορίες. Αυτού του είδους η μάθηση είναι δυνατόν να χωριστεί σε τρία στάδια. Πρώτον, επισημειώνονται, συνήθως χειρωνακτικά, παραδείγματα εκπαίδευσης με τη ορθή τους κατηγορία. Τα παραδείγματα παριστάνονται στη συνέχεια με τη μορφή διανυσμάτων ιδιοτήτων. Το στάδιο της επισημείωσης των παραδειγμάτων εκπαίδευσης είναι το βασικό μειονέκτημα της επιβλεπόμενης μηχανικής μάθησης στις περισσότερες εφαρμογές επεξεργασίας φυσικής γλώσσας, καθώς σε πολλές εφαρμογές ο όγκος των κειμένων που πρέπει να επισημειωθεί είναι μεγάλος, με αποτέλεσμα η διαδικασία αυτή να είναι κουραστική και ιδιαίτερα χρονοβόρα.

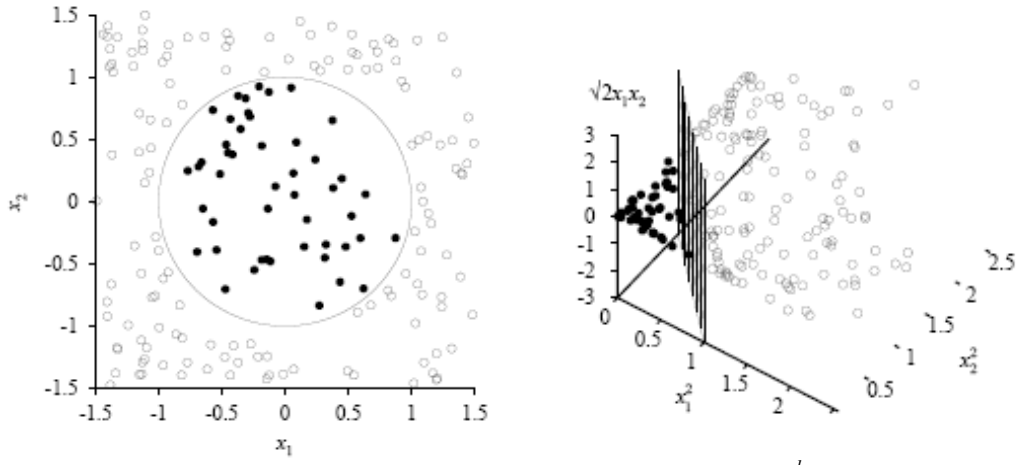
Στο δεύτερο στάδιο χρησιμοποιείται ένας αλγόριθμος μάθησης, ο οποίος επεξεργάζεται τα παραδείγματα εκπαίδευσης προκειμένου να κατασκευάσει έναν ταξινομητή που θα είναι σε θέση να διαχωρίζει παραδείγματα διαφορετικών κατηγοριών με βάση τις τιμές των ιδιοτήτων τους. Έχουν προταθεί πολλοί αλγόριθμοι, όπως ο C4.5 (δημιουργία δέντρων απόφασης), τα νευρωνικά δίκτυα, το μοντέλο μέγιστης εντροπίας, οι Μηχανές Διανυσμάτων Υποστήριξης κλπ.

Το τελευταίο στάδιο αφορά την κατηγοριοποίηση νέων περιπτώσεων, για τις οποίες δεν είναι γνωστή η ορθή κατηγορία. Οι νέες περιπτώσεις αναπαρίστανται επίσης με τη μορφή διανυσμάτων ιδιοτήτων και κατατάσσονται στις κατηγορίες χρησιμοποιώντας τον ταξινομητή του προηγούμενου σταδίου.

2.3.1 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines, SVMs [17, 16, 33]) είναι μία σχετικά καινούρια μέθοδος επιβλεπόμενης μηχανικής μάθησης, η οποία μπορεί να εφαρμοστεί και σε προβλήματα κατηγοριοποίησης και η οποία έχει επιτύχει εξαιρετικά αποτελέσματα σε πολλές εφαρμογές. Στην απλούστερή τους μορφή, που χρησιμοποιούμε εδώ, οι ΜΔΥ μαθαίνουν να διαχωρίζουν περιπτώσεις δύο κατηγοριών. Ουσιαστικά προβάλλουν, με τη χρήση μίας συνάρτησης μετασχηματισμού, τα διανύσματα ιδιοτήτων σε ένα χώρο περισσοτέρων διαστάσεων και στη συνέχεια προσπαθούν να βρουν ένα γραμμικό διαχωριστή, δηλαδή ένα υπερεπίπεδο, που να διαχωρίζει τις δύο κατηγορίες με μέγιστο περιθώριο (margin) στο νέο διανυσματικό χώρο.

Η μετάβαση στο νέο χώρο περισσοτέρων διαστάσεων διευκολύνει την εύρεση γραμμικού διαχωριστή. Για παράδειγμα, στο παρακάτω σχήμα αριστερά φαίνεται μία περίπτωση, όπου δεν υπάρχει γραμμικός διαχωριστής (ευθεία) στο επίπεδο (εδώ υπάρχουν δύο μόνο ιδιότητες). Χρησιμοποιώντας όμως τη συνάρτηση μετασχηματισμού $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2} \cdot x_1 \cdot x_2 \right\rangle$ και μεταβαίνοντας στις τρεις διαστάσεις παρατηρούμε ότι υπάρχει ένα επίπεδο που διαχωρίζει τα διανύσματα (σχήμα στα δεξιά). Στην περίπτωση περισσοτέρων ιδιοτήτων, ο διαχωριστής θα είναι ένα υπερεπίπεδο.



Μετασχηματισμός από τις δύο διαστάσεις στις τρεις [1]

Γενικά, η εξίσωση του υπερεπιπέδου διαχωρισμού θα είναι της ακόλουθης μορφής, όπου F η συνάρτηση μετασχηματισμού:

$$\vec{w} \cdot \vec{F}(\vec{x}) + b = 0$$

Το υπερεπίπεδο διαχωρισμού τοποθετείται στο μέσον της απόστασης δύο παράλληλων υπερεπιπέδων, τα οποία διαχωρίζουν πλήρως τα παραδείγματα εκπαίδευσης και εφάπτονται με τουλάχιστον ένα παράδειγμα εκπαίδευσης,

[1] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, 2002.

διαφορετικής κατηγορίας για το κάθε ένα από τα δύο υπερεπίπεδα. Τα $\vec{w}$ ($\vec{w} \in R^l$, όπου l ο αριθμός των ιδιοτήτων στο νέο χώρο) και b μπορούν να επιλεγούν (με scaling) ώστε τα δύο παράλληλα εφαπτόμενα υπερεπίπεδα να έχουν εξισώσεις:

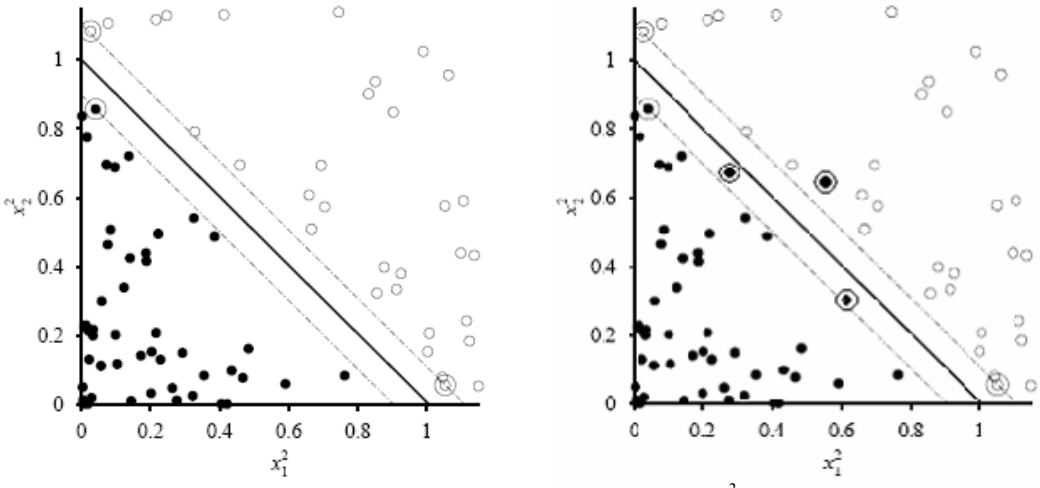
$$\vec{w} \cdot \vec{F}(\vec{x}) + b = \pm 1$$

οπότε η απόσταση μεταξύ των δύο εφαπτόμενων υπερεπιπέδων είναι $2/\left\|\vec{w}\right\|$. Η απόσταση αυτή είναι το «περιθώριο» του υπερεπιπέδου διαχωρισμού, που τοποθετείται στο μέσον της απόστασης των δύο εφαπτόμενων υπερεπιπέδων. Όπως αναφέρθηκε ήδη, ο στόχος των ΜΔΥ είναι να βρουν το υπερεπίπεδο διαχωρισμού με το μέγιστο περιθώριο. Οπότε, προκύπτει τα παρακάτω πρόβλημα βελτιστοποίησης:

$$\min \left\|\vec{w}\right\|^2 / 2$$
$$(\vec{w} \cdot \vec{F}(\vec{x_j}) + b) \cdot y_j \geq 1$$

όπου x_j με $1 \leq j \leq n$ είναι το διάνυσμα του j -οστού παραδείγματος εκπαίδευσης και $y_j \in \{1, -1\}$ είναι η κατηγορία του j -οστού διανύσματος εκπαίδευσης.

Οι περιορισμοί του παραπάνω προβλήματος βελτιστοποίησης επιβάλλουν όλα τα διανύσματα εκπαίδευσης να βρίσκονται έξω ή το πολύ στα όρια του περιθωρίου και από τη σωστή πλευρά του υπερεπιπέδου, ανάλογα με την κατηγορία τους, όπως φαίνεται στο παρακάτω σχήμα αριστερά. Οι περιορισμοί αυτοί, όμως, είναι πολύ αυστηροί. Για παράδειγμα, ενδέχεται να μην είναι δυνατή η εύρεση γραμμικού διαχωριστή που να διαχωρίζει πλήρως τα παραδείγματα εκπαίδευσης, παρά τη μετάβαση στο νέο χώρο διαστάσεων. Ή ενδέχεται να προτιμούμε ένα υπερεπίπεδο διαχωρισμού που έχει μεγαλύτερο περιθώριο αλλά κατατάσσει λανθασμένα ή εντός του περιθωρίου κάποια παραδείγματα εκπαίδευσης (όπως στο παρακάτω σχήμα στα δεξιά) από κάποιο άλλο που ικανοποιεί όλους τους περιορισμούς αλλά έχει μικρότερο περιθώριο.



Υπερεπίπεδο με μέγιστο περιθώριο[2]

[2] Τα σχήματα είναι από το βιβλίο των Stuart Russell και Peter Norvig *"Artificial Intelligence: A Modern Approach (Second Edition)"*, Prentice Hall, 2002.

Για τους λόγους αυτούς, είναι δυνατόν οι περιορισμοί να χαλαρώσουν, με αποτέλεσμα να κατασκευαστεί ένα ανεκτικότερο πρόβλημα βελτιστοποίησης, το οποίο ορίζεται ως εξής:

$$\min \left\| \vec{w} \right\|^2 / 2 + C \cdot \sum_j \xi_j$$
$$(\vec{w} \cdot \vec{F}(\vec{x_j}) + b) \cdot y_j \geq 1 - \xi_j$$
$$\xi_j \geq 0$$

όπου το ξ_j είναι το σφάλμα για κάθε διάνυσμα εκπαίδευσης (το πόσο απέχουμε από το να ικανοποιείται ο αντίστοιχος περιορισμός) και C το κόστος (ανοχή) που δίνεται στο συνολικό σφάλμα. Στην περίπτωση αυτή, όπως φαίνεται στο παραπάνω σχήμα δεξιά, υπάρχει καλύτερη δυνατότητα γενίκευσης και ο διαχωριστής είναι πιο ανεκτικός σε λάθη επισημείωσης των δεδομένων εκπαίδευσης.

Τελικά, επιλύνοντας το παραπάνω πρόβλημα ελαχιστοποίησης, προκύπτει $\vec{w}$ της μορφής:

$$\vec{w} = \sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j})$$

Οπότε η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j}) \right) \cdot \vec{F}(\vec{x}) + b = 0$$

ή

$$\left(\sum_j a_j \cdot y_j \cdot \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x}) \right) + b = 0$$

όπου οι τιμές των a_j είναι διάφορες του μηδενός μόνο για τα «διανύσματα υποστήριξης», δηλαδή τα διανύσματα εκπαίδευσης που βρίσκονται πάνω στα δύο εφαπτόμενα υπερεπίπεδα και (στην περίπτωση που ανεχόμαστε σφάλματα) τα διανύσματα που κατατάσσονται λανθασμένα ή εντός του περιθωρίου. Τα παραδείγματα που δεν είναι διανύσματα υποστήριξης ουσιαστικά αγνοούνται.

Η συνάρτηση μετασχηματισμού συμμετέχει μόνο σε εσωτερικά γινόμενα $\vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$. Ορίζουμε, λοιπόν, ως πυρήνα της ΜΔΥ τη συνάρτηση:

$$K(\vec{x_j}, \vec{x_i}) = \vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$$

Σύμφωνα με το θεώρημα του Mercer, κάθε συνάρτηση $K(\vec{x_i}, \vec{x_j})$ για την οποία ο πίνακας $K_{ij} = K(\vec{x_i}, \vec{x_j})$ είναι θετικά ορισμένος[3] υπολογίζει το εσωτερικό γινόμενο των $\vec{x_i}, \vec{x_j}$ σε κάποιο νέο διανυσματικό χώρο, δηλαδή μπορεί να χρησιμοποιηθεί ως πυρήνας μιας ΜΔΥ. Το ενδιαφέρον είναι ότι σε πολλές περιπτώσεις είναι δυνατόν να υπολογιστούν οι τιμές του πυρήνα χωρίς να υπολογιστεί πρώτα η τιμή των $\vec{F}(\vec{x_j})$ και $\vec{F}(\vec{x_i})$, δηλαδή χωρίς να υπολογίσουμε τις (συνήθως πολύ περισσότερες) ιδιότητες των διανυσμάτων στο νέο χώρο, κάτι που επιτρέπει τη χρήση πυρήνων που υπολογίζουν εσωτερικά γινόμενα σε νέους χώρους πολύ μεγάλου αριθμού διαστάσεων. Για παράδειγμα, στην περίπτωση του αρχικού παραδείγματος μετασχηματισμού με $\vec{F}(\vec{x}) = \left\langle x_1^2, x_2^2, \sqrt{2} \cdot x_1 \cdot x_2 \right\rangle$ ο πυρήνας έχει τη μορφή $K(\vec{x_i}, \vec{x_j}) = F(\vec{x_i}) \cdot F(\vec{x_j}) = (\vec{x_i} \cdot \vec{x_j})^2$, δηλαδή οι τιμές του μπορούν να υπολογισθούν με βάση μόνο τις τιμές των ιδιοτήτων του αρχικού χώρου. Τελικά, αντικαθιστώντας το εσωτερικό γινόμενο $\vec{F}(\vec{x_j}) \cdot \vec{F}(\vec{x_i})$ με τη συνάρτηση πυρήνα $K(\vec{x_j}, \vec{x_i})$ η εξίσωση του υπερεπιπέδου γίνεται:

$$\left(\sum_j a_j \cdot y_j \cdot K(\vec{x_j}, \vec{x}) \right) + b = 0$$

Παραδείγματα πυρήνων που χρησιμοποιούνται είναι τα εξής:

- γραμμικός: $K(\vec{x_i}, \vec{x_j}) = \vec{x_i} \cdot \vec{x_j}$
- πολυωνυμικός: $K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} \cdot \vec{x_j} + 1)^d$
- ακτινωτής βάσης (radial base function – RBF):
  $K(\vec{x_i}, \vec{x_j}) = \exp\left(-\gamma \cdot \left\| \vec{x_i} - \vec{x_j} \right\|^2 \right), \gamma > 0$
- σιγμοειδής: $K(\vec{x_i}, \vec{x_j}) = \tanh(\vec{x_i} \cdot \vec{x_j} + r)$

όπου τα γ, r και d είναι παράμετροι κάθε πυρήνα. Ο γραμμικός πυρήνας (που δεν προκαλεί μετάβαση σε νέο διανυσματικό χώρο) είναι ειδική περίπτωση του πυρήνα ακτινωτής βάσης (Keerthi και Lin 2003). Επίσης, ο σιγμοειδής πυρήνας συμπεριφέρεται όπως ο πυρήνας ακτινωτής βάσης για συγκεκριμένες παραμέτρους (Lin και Lin 2003).

Τέλος, η απόφαση για την κατηγορία ενός καινούριου διανύσματος, δεδομένου ενός εκπαιδευμένου ταξινομητή, λαμβάνεται με βάση το πρόσημο της ακόλουθης παράστασης:

$$sign\left(\sum_j a_j \cdot y_j \cdot K(\vec{x}_j, \vec{x}) + b \right)$$

---

[3] Ο πίνακας $A \in R^{n \times n}$ είναι ένας θετικά ορισμένος πίνακας αν για όλα τα μη μηδενικά διανύσματα $x \in R^n$ ισχύει $x^T \cdot A \cdot x > 0$, όπου x^T είναι το ανάστροφο διάνυσμα.

3 Αρχιτεκτονική του συστήματος

Η επιλογή των κατηγοριών των ονομάτων οντοτήτων έγινε με βάση τις οδηγίες του διαγωνισμού MUC-7 [28]. Πιο συγκεκριμένα, το σύστημα σχεδιάσθηκε ώστε να υποστηρίζει μελλοντικά τις εξής κατηγορίες:

- Κύρια ονόματα (ENAMEX)
  - Ονόματα προσώπων (PERSON)
  - Ονόματα οργανισμών (ORGANIZATION)
  - Τοπωνύμια (LOCATION)
- Χρονικές εκφράσεις (TIMEX)
  - Ημερομηνίες (DATE)
  - Εκφράσεις ώρας (TIME)
- Αριθμητικές εκφράσεις (NUMEX)
  - Νομισματικές εκφράσεις (MONEY)
  - Ποσοστά (PERCENT)

Στην τρέχουσα μορφή του το σύστημα υποστηρίζει μόνο τρεις κατηγορίες οντοτήτων: ονόματα προσώπων (enamex person), ημερομηνίες (timex date) και εκφράσεις ώρας (timex time). Σημειώνεται ότι το σύστημα εντοπίζει όλες τις χρονικές εκφράσεις, αλλά τις μαρκάρει ως date, χωρίς διάκριση μεταξύ date και time. Όπως αναφέρθηκε και προηγουμένως, ελπίζουμε ότι το σύστημα θα επεκταθεί σε μελλοντικές εργασίες, ώστε να υποστηρίζει και τις υπόλοιπες παραπάνω κατηγορίες.

Τα βασικά στάδια επεξεργασίας του συστήματος, όπως φαίνονται και στο παρακάτω σχήμα, είναι τα εξής:

- Διαχωρισμός σε λεκτικές μονάδες (tokens)
- Διαχωρισμός σε περιόδους
- Αναγνώριση και επισημείωση χρονικών εκφράσεων
- Αναγνώριση και επισημείωση ονομάτων προσώπων

Τα στάδια του διαχωρισμού σε περιόδους και της αναγνώρισης των χρονικών εκφράσεων είναι ανεξάρτητα μεταξύ τους. Η σειρά, δηλαδή, με την οποία θα εκτελεστούν μπορεί να είναι οποιαδήποτε (στην παρούσα μορφή της υλοποίησης, ο διαχωριστής περιόδων προηγείται της αναγνώρισης χρονικών εκφράσεων).

Το στάδιο της αναγνώρισης ονομάτων προσώπων εξαρτάται από τα δύο προηγούμενα, από τα οποία αντλεί ιδιότητες. Για αυτόν το λόγο, είναι απαραίτητο να εκτελεστεί τελευταίο.

Στάδια επεξεργασίας του συστήματος

Στη συνέχεια θα παρουσιαστούν αναλυτικά τα παραπάνω στάδια επεξεργασίας.

3.1 Διαχωρισμός σε λεκτικές μονάδες

Ο διαχωριστής λεκτικών μονάδων (tokenizer) δέχεται ως είσοδο ένα κείμενο. Το χωρίζει σε λεκτικές μονάδες (tokens) και δημιουργεί μία δομή που τις περιέχει. Η δομή αυτή χρησιμοποιείται και από τα υπόλοιπα στάδια επεξεργασίας, όπως θα εξηγηθεί παρακάτω. Οι λεκτικές μονάδες ορίζονται σε αυτή την εργασία ως εξής:

- Κάθε ακολουθία ελληνικών ή λατινικών χαρακτήρων θεωρείται λεκτική μονάδα. Μία ακολουθία που περιέχει μαζί ελληνικούς και λατινικούς χαρακτήρες χωρίζεται σε λεκτικές μονάδες ανάλογες με τον αριθμό των εναλλαγών μεταξύ ελληνικών και λατινικών χαρακτήρων. Για παράδειγμα η ακολουθία χαρακτήρων «Eurogνώση» αποτελείται από δύο λεκτικές μονάδες, τις «Euro» και «γνώση». Γενικά, η περίπτωση αυτή είναι ιδιαίτερα σπάνια.
- Κάθε ακολουθία αριθμητικών χαρακτήρων θεωρείται λεκτική μονάδα.
- Κάθε άλλος μη κενός (non-whitespace) χαρακτήρας αποτελεί λεκτική μονάδα από μόνος του, ακόμα και αν δύο ή περισσότεροι συνεχόμενοι μπορούν να θεωρηθούν γραμματικά ή συντακτικά ως μία οντότητα. Για παράδειγμα, οι τρεις τελείες (…) θεωρούνται ως τρεις ξεχωριστές λεκτικές μονάδες. Επίσης, το κόμμα και η τελεία που μπορούν να περιέχονται σε έναν αριθμό θεωρούνται ξεχωριστές λεκτικές μονάδες. Παραδείγματος χάριν, ο αριθμός «3,14» αποτελείται από τρεις λεκτικές μονάδες, το 3, το «,» και το 14.

Η δομή που τελικά δημιουργείται είναι μία λίστα που περιέχει τις λεκτικές μονάδες του κειμένου εισόδου, με τη σειρά που συναντούνται στο κείμενο. Κάθε κόμβος της λίστας αντιστοιχεί σε μία λεκτική μονάδα και περιέχει τις παρακάτω πληροφορίες για αυτήν:

- λεκτική μονάδα
- αν ανήκει ή όχι σε κάποια κατηγορία ονομάτων οντοτήτων, και αν ναι σε ποια
- αν αποτελεί τέλος περιόδου
- αν περιέχεται σε τίτλο ή υπότιτλο.

Δεν υπάρχουν πληροφορίες στη λίστα που να δείχνουν αν μια λεκτική μονάδα αποτελεί την αρχή ή το τέλος ενός ονόματος οντότητας. Θεωρούμε ότι συνεχόμενες λεκτικές μονάδες που ανήκουν στην ίδια κατηγορία ονομάτων οντοτήτων (π.χ. *«Γιώργος Παπαδημητρίου»*, όπου έχουμε δύο συνεχόμενες λεκτικές μονάδες κατηγορίας person) αποτελούν μέρος του ιδίου ονόματος. Γενικά, η παραδοχή αυτή ισχύει, εκτός από ελάχιστες εξαιρέσεις, όπως για παράδειγμα η φράση *«ο πατέρας του Δημήτρη Γιάννης»*, όπου τα *«Δημήτρης»* και *«Γιάννης»* θα θεωρηθούν λανθασμένα ότι αποτελούν ένα όνομα.

Ο διαχωριστής λεκτικών μονάδων έχει, επίσης, τη δυνατότητα να δέχεται ως είσοδο επισημειωμένα κείμενα, όπως το παρακάτω. Στην περίπτωση αυτή, οι ετικέτες των επισημειώσεων μετατρέπονται σε αντίστοιχες πληροφορίες κατηγορίας, τέλους περιόδου ή τίτλου, που εισάγονται στους αντίστοιχους κόμβους της λίστας.

```
<SUBTITLE>
«Θα παραιτηθώ αν αποδειχθεί ότι ζημιώθηκε έστω και μία δραχμή το
<ENAMEX TYPE="ORGANIZATION">Δημόσιο</ENAMEX>»
</SUBTITLE>
Στα άκρα είναι αποφασισμένος να φθάσει ο υπουργός ΠΕΧΩΔΕ κ. <ENAMEX
TYPE="PERSON">Κ. Λαλιώτης</ENAMEX> τον... αεροπορικό πόλεμο που του
έχει κηρύξει η <ENAMEX TYPE="ORGANIZATION">Νέα Δημοκρατία</ENAMEX> με
αφορμή την ολοκλήρωση και επικείμενη λειτουργία του νέου διεθνούς
αεροδρομίου της <ENAMEX TYPE="LOCATION">Αθήνας</ENAMEX> <ENAMEX
TYPE="LOCATION">«Ελευθέριος Βενιζέλος»</ENAMEX> στα <ENAMEX
TYPE="LOCATION">Σπάτα</ENAMEX>. Την ίδια στιγμή έκθεση της <ENAMEX
TYPE="ORGANIZATION">Salomon Smith</ENAMEX> δικαιώνει τον υπουργό για
τους χειρισμούς του στο θέμα του αεροδρομίου.
<PARAGRAPH>
Στον πίνακα φαίνονται οι διαφορές της σύμβασης που ετοίμασε το <TIMEX
TYPE="DATE">1993</TIMEX> η <ENAMEX TYPE="ORGANIZATION">Ν.Δ.</ENAMEX>
με τη σύμβαση που υπέγραψε το <TIMEX TYPE="DATE">1995</TIMEX> η
κυβέρνηση του <ENAMEX TYPE="ORGANIZATION">ΠΑΣΟΚ</ENAMEX> για το νέο
αεροδρόμιο στα <ENAMEX TYPE="LOCATION">Σπάτα</ENAMEX>
<PARAGRAPH>
```

3.2 Διαχωριστής Περιόδων

Ο διαχωριστής περιόδων (sentence splitter) δέχεται ως είσοδο τη δομή που προέκυψε από το προηγούμενο βήμα, αποφασίζει αν κάποια λεκτική μονάδα (συγκεκριμένα αν κάποια τελεία) αποτελεί τέλος περιόδου και συμπληρώνει το κατάλληλο πεδίο της δομής.

Ο σκοπός της ανάπτυξης του διαχωριστή περιόδων είναι διπλός. Πρώτον, είναι ιδιαίτερα χρήσιμος ως στάδιο επεξεργασίας σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας. Δεύτερον, τροφοδοτεί τη βασική διαδικασία του συστήματος, τη διαδικασία της αναγνώρισης ονομάτων προσώπων, με ιδιότητες, όπως αν η τρέχουσα λεκτική μονάδα (ή κάποια άλλη γύρω από αυτήν σε ένα παράθυρο 5 λεκτικών μονάδων) αποτελεί τέλος περιόδου (παράγραφος 3.4.2). Επίσης, η τιμή πολλών ιδιοτήτων της ΜΔΥ για την αναγνώριση ονομάτων προσώπων επηρεάζεται έμμεσα από το διαχωριστή περιόδων: οι ιδιότητες που αφορούν λεκτικές μονάδες οι οποίες δεν βρίσκονται στην ίδια περίοδο με την υπό εξέταση λεκτική μονάδα παίρνουν τιμή -1 (ψευδές). Για παράδειγμα, η ιδιότητα «ο τύπος της λεκτικής μονάδας με απόσταση -2 από την τρέχουσα είναι *ακολουθία ελληνικών χαρακτήρων*;» παίρνει τιμή -1 αν η λεκτική μονάδα με απόσταση -1 από την τρέχουσα αποτελεί τέλος περιόδου, ακόμα και αν ο τύπος της λεκτικής μονάδας με απόσταση -2 από την τρέχουσα είναι όντως *«ακολουθία ελληνικών χαρακτήρων»*.

Εκτός από τις τελείες υπάρχουν και άλλα σύμβολα που σηματοδοτούν τέλος περιόδου, όπως τα θαυμαστικά ή τα ερωτηματικά. Τα σύμβολα αυτά δεν εμφανίζονται συχνά στα κείμενα εφημερίδων που χρησιμοποιήθηκαν κατά τη διεξαγωγή των πειραμάτων και για αυτόν το λόγο τα αγνοήσαμε (η αναλογία των εμφανίσεων των τελείων προς το άθροισμα των εμφανίσεων των υπόλοιπων συμβόλων που είναι δυνατόν να σηματοδοτούν τέλος περιόδου είναι μεγαλύτερη από 100 προς 1). Η επίπτωση αυτού του γεγονότος είναι ιδιαίτερα μικρή, καθώς, εκτός από το ότι τα άλλα σύμβολα είναι ιδιαίτερα σπάνια, η πληροφορία που προσφέρει ο διαχωριστής περιόδων αποτελεί μόνο μία από τις πολλές ιδιότητες που έχει στη διάθεσή της η ΜΔΥ της αναγνώρισης ονομάτων προσώπων.

Ο διαχωριστής περιόδων χρησιμοποιεί και αυτός μια ΜΔΥ. Κάθε τελεία παριστάνεται με τη μορφή ενός διανύσματος ιδιοτήτων (συνολικά 17 ιδιότητες). Χρησιμοποιήθηκαν οι εξής ιδιότητες:

- Ο τύπος της προηγούμενης / επόμενης λεκτικής μονάδας και συγκεκριμένα: λέξη, σύμβολο, αριθμός (6 δυαδικές ιδιότητες). Συνήθως, τέλος περιόδου σηματοδοτούν οι τελείες, οι οποίες έπονται και ακολουθούνται από λέξεις. Για παράδειγμα, αν μία τελεία βρίσκεται ανάμεσα σε δύο αριθμούς τότε είναι σχεδόν σίγουρο ότι αποτελεί μέρος μίας αριθμητικής έκφρασης και όχι τέλος περιόδου.
- Αν η προηγούμενη / επόμενη λεκτική μονάδα αρχίζει με κεφαλαίο γράμμα (2 δυαδικές ιδιότητες). Η πρώτη λέξη κάθε πρότασης αρχίζει με κεφαλαίο γράμμα. Στην περίπτωση, όμως, που η υπό εξέταση τελεία έπεται και ακολουθείται από λέξεις που αρχίζουν με κεφαλαίο γράμμα, το πιο πιθανό είναι να μη σηματοδοτεί τέλος περιόδου, όπως για παράδειγμα στη φράση «Κων. Σημίτης».
- Αν η προηγούμενη / επόμενη λεκτική μονάδα είναι γραμμένη με κεφαλαίους χαρακτήρες (2 δυαδικές ιδιότητες). Οι ιδιότητες αυτές έχουν ως στόχο τους τίτλους, όπου όλες οι λέξεις είναι γραμμένες με κεφαλαίους χαρακτήρες. Επίσης, αποσκοπούν σε συντομογραφίες, όπως για παράδειγμα η φράση «ΕΛ.ΤΑ.», όπου οι τελείες δε σηματοδοτούν τέλος περιόδου.
- Απόσταση από την προηγούμενη / επόμενη τελεία (2 ιδιότητες με κανονικοποιημένες τιμές στο διάστημα [-1, 1]). Αναμένουμε ότι όσο πιο κοντά στην υπό εξέταση τελεία βρίσκεται κάποια άλλη τελεία τόσο πιο απίθανο είναι να σηματοδοτεί τέλος περιόδου. Για παράδειγμα, οι τρεις τελείες (…) δεν είναι δυνατόν να σηματοδοτούν τέλος περιόδου, με εξαίρεση την τελευταία.
- Το μήκος (σε χαρακτήρες) της προηγούμενης / επόμενης λεκτικής μονάδας (2 ιδιότητες με κανονικοποιημένες τιμές στο διάστημα [-1, 1]). Συνήθως οι προτάσεις αρχίζουν με άρθρο ή πρόθεση μήκους 2-3 χαρακτήρες. Επίσης, όταν χρησιμοποιούνται αρχικά ονομάτων το μήκος των λεκτικών μονάδων είναι μικρό, όπως για παράδειγμα η φράση «Κ. Σημίτης».
- Το γράμμα με το οποίο τελειώνει η προηγούμενη λεκτική μονάδα. Οι ελληνικές λέξεις στην πλειοψηφία τους τελειώνουν σε φωνήεν, «ν» και «ς» (3 δυαδικές ιδιότητες). Επομένως, αν πριν από την υπό εξέταση τελεία υπάρχει ελληνική λέξη που τελειώνει σε κάποιον άλλο χαρακτήρα, τότε η τελεία πιθανότατα δε σηματοδοτεί τέλος περιόδου, όπως για παράδειγμα στη φράση «κ. Σημίτης».

Για την ανάπτυξη του διαχωριστή περιόδων χρησιμοποιήσαμε τη βιβλιοθήκη για ΜΔΥ libSVM [3, 14], επιλέγοντας ως συνάρτηση πυρήνα τον πυρήνα ακτινωτής βάσης. Σημειώνεται ότι ακολουθήθηκαν οι οδηγίες των δημιουργών της συγκεκριμένης υλοποίησης των ΜΔΥ (βλ. παράρτημα). Πιο συγκεκριμένα, οι ιδιότητες παίρνουν τιμές στο διάστημα [-1, 1]. Επίσης, με βάση τα δεδομένα εκπαίδευσης έγινε ρύθμιση των παραμέτρων της ΜΔΥ. Οι τιμές των παραμέτρων που προέκυψαν είναι $C = 16$ και $\gamma = 0{,}125$.

Το ποσοστό επιτυχίας του διαχωριστή περιόδων είναι ιδιαίτερα ικανοποιητικό. Συγκεκριμένα επιτυγχάνει ορθότητα[4] (accuracy) 98,89%. Το αποτέλεσμα αυτό έχει προκύψει με τη διαδικασία της 10-πλής διασταυρωμένης επικύρωσης (10-fold cross-validation). Κατά την *n*-πλή διασταυρωμένη επικύρωση, τα δεδομένα χωρίζονται σε *n* ίσα μέρη και εκτελούνται *n* επαναλήψεις, όπου σε κάθε επανάληψη *n*-1 μέρη χρησιμοποιούνται ως δεδομένα εκπαίδευσης και ένα μέρος (διαφορετικό κάθε φορά) ως δεδομένα ελέγχου. Το τελικό αποτέλεσμα είναι ο μέσος όρος των *n* επιμέρους αποτελεσμάτων.

Τα κείμενα που χρησιμοποιήθηκαν κατά τη διαδικασία της επικύρωσης είναι 340 κείμενα καθημερινών και οικονομικών εφημερίδων, επισημειωμένα κατάλληλα, συνολικού μεγέθους 1,38M, με συνολικά 10210 τελείες.

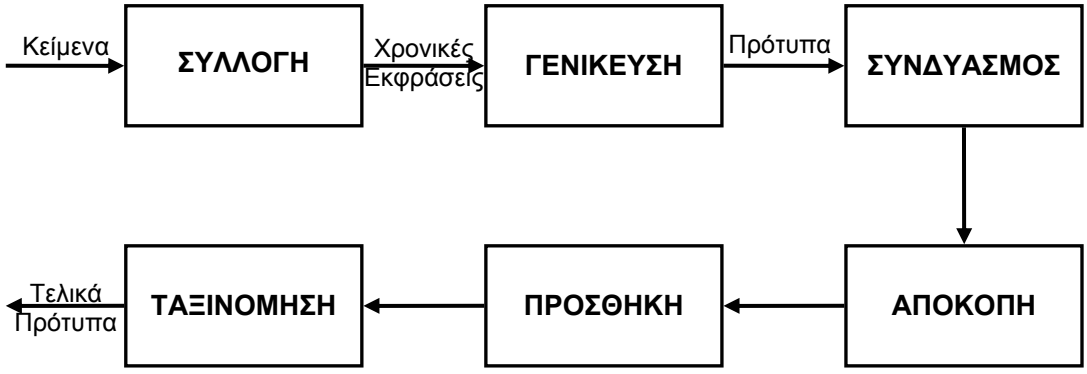
3.3 Αναγνώριση χρονικών εκφράσεων

Μία σημαντική διαφορά τόσο με το προηγούμενο στάδιο επεξεργασίας (διαχωρισμός περιόδων) όσο και με το επόμενο (αναγνώριση ονομάτων προσώπων) είναι ότι στο στάδιο της αναγνώρισης χρονικών εκφράσεων δε χρησιμοποιείται κάποιος καθιερωμένος αλγόριθμος μηχανικής μάθησης αλλά πρότυπα (patterns) χρονικών εκφράσεων, τα οποία όμως προκύπτουν σε μεγάλο βαθμό αυτόματα.

Το στάδιο αναγνώρισης χρονικών εκφράσεων δέχεται ως είσοδο την κοινή δομή που έχει προκύψει από το πρώτο στάδιο επεξεργασίας (λίστα λεκτικών μονάδων). Τα πρότυπα των χρονικών εκφράσεων δεν εφαρμόζονται στο κείμενο εισόδου αλλά στην κοινή αυτή δομή.

3.3.1 Ημι-αυτόματη διαδικασία δημιουργίας προτύπων για χρονικές εκφράσεις

Η αναγνώριση των χρονικών εκφράσεων βασίζεται σε πρότυπα (patterns), τα οποία παράγονται σε μεγάλο βαθμό αυτόματα, με τη διαδικασία του παρακάτω σχήματος. Η ημι-αυτόματη διαδικασία δημιουργίας προτύπων για χρονικές εκφράσεις επαναλαμβάνεται ξεχωριστά για κάθε μία συλλογή κειμένων που χρησιμοποιήθηκε (ενότητα 4.1). Στη συνέχεια, αναλύονται οι επιμέρους φάσεις αυτής της διαδικασίας.



[4] Η ορθότητα ορίζεται ως το ποσοστό των ορθά ταξινομημένων περιπτώσεων (εδώ τελειών) προς τις συνολικές περιπτώσεις.

3.3.1.1 Συλλογή χρονικών εκφράσεων

Η φάση αυτή είναι καθαρά προγραμματιστική και δε χρειάζεται ιδιαίτερη επεξήγηση. Οι χρονικές εκφράσεις συλλέγονται από τα κείμενα εκπαίδευσης.

3.3.1.2 Γενίκευση

Κατά τη φάση αυτή γενικεύονται οι χρονικές εκφράσεις του προηγούμενου βήματος και δημιουργούνται πρότυπα (patterns). Η γενίκευση γίνεται αντικαθιστώντας λεκτικές μονάδες των παρακάτω τύπων με τους τύπους τους (π.χ. το «Δευτέρας» γίνεται day). Οι τύποι και οι λεκτικές μονάδες που ανήκουν σε αυτούς ενδέχεται να πρέπει να προσαρμοσθούν χειρωνακτικά όταν η διαδικασία δημιουργίας προτύπων χρονικών εκφράσεων εφαρμόζεται σε νέες συλλογές κειμένων και για αυτό χαρακτηρίζουμε τη διαδικασία ημι-αυτόματη.

- **separator**: «/» και «.» Αυτός ο τύπος περιέχει διαχωριστές τμημάτων ημερομηνιών. Σε αυτόν τον τύπο είναι δυνατόν να ενταχθεί και το σύμβολο «-». Το γεγονός αυτό, όμως, απορρίφθηκε καθώς δημιουργούνται πολλά λάθη σε εκφράσεις διαστημάτων, όπως για παράδειγμα «3-5%». Αντίθετα, η τελεία στην ελληνική αριθμητική σημειογραφία χρησιμοποιείται ως διαχωριστής χιλιάδων (π.χ. 3.000), με αποτέλεσμα το ένα τουλάχιστον μέρος (αριστερά ή δεξιά της τελείας) να είναι ένας τριψήφιος αριθμός. Σχεδόν καμία χρονική έκφραση δεν περιέχει τριψήφιους αριθμούς και σε συνδυασμό με τον τρόπο αναπαράστασης των αριθμητικών εκφράσεων που χρησιμοποιούμε (περιγράφεται παρακάτω) δεν δημιουργούνται λάθη με την ένταξη του συμβόλου της τελείας σε αυτόν το τύπο.
- **article**: τον, του, το, της, την, τη, τα, ΤΟΝ, ΤΟΥ… Ασχολούμαστε μόνο με αυτά τα άρθρα, γιατί κυρίως αυτά χρησιμοποιούνται σε χρονικές εκφράσεις.
- **day**: Δευτέρα, Δευτέρας, ΔΕΥΤΕΡΑ, ΔΕΥΤΕΡΑΣ, Τρίτη… στην ονομαστική, γενική και αιτιατική του ενικού.
- **month**: Ιανουάριος, Ιανουαρίου, Ιανουάριο, ΙΑΝΟΥΑΡΙΟΣ, ΙΑΝΟΥΑΡΙΟΥ, ΙΑΝΟΥΑΡΙΟ, Γενάρης, Γενάρη, ΓΕΝΑΡΗΣ, ΓΕΝΑΡΗ, Φεβρουάριος…
- **part of the day**: πρωί, απόγευμα, βράδυ, μεσάνυχτα, μεσημέρι, ξημερώματα, νύχτα, πρωινό, πρωινού,…, ΠΡΩΙ,…
- **season**: άνοιξη, καλοκαίρι, φθινόπωρο, χειμώνας, άνοιξης,…, ΑΝΟΙΞΗ,…
- **special day**: Πρωτοχρονιά, Χριστούγεννα, Πάσχα, Πρωτομαγιά, Πρωταπριλιά, Σαββατοκύριακο, Δεκαπενταύγουστος, Πρωτοχρονιάς,…, ΠΡΩΤΟΧΡΟΝΙΑ,…
- **spell**: αιώνας, χιλιετία, μήνας, χρόνος, δεκαετία, ημέρα, δίμηνο, τρίμηνο, τετράμηνο, εξάμηνο, οκτάμηνο, δεκαήμερο, δεκαπενθήμερο, διήμερο, τριήμερο, πενθήμερο, εβδομάδα, αιώνα,…, ΑΙΩΝΑΣ,…
- **ordinal number**: πρώτος, δεύτερος, τρίτος, τέταρτος, πέμπτος, έκτος, έβδομος, όγδοος, ένατος, δέκατος…
- **ending**: -ού, -ου, -ός, -ος, -ό, -ο, -ής, -ης, -ή, -η, -ΟΥ, -ΟΣ… Ο στόχος είναι εκφράσεις όπως *«20ος αιώνας»*.

- **abbreviation**: πμ, μμ, πΧ, μΧ, π, μ, Χ. Ο διαχωριστής λεκτικών μονάδων θεωρεί τη συντομογραφία «π.Χ.» (προ Χριστού) ως τέσσερις ξεχωριστές λεκτικές μονάδες, όπως αναφέρθηκε στην ενότητα 3.1.
- **attributive adjective**: καθαρή, καθαρής, μεγάλη, μεγάλης, ΚΑΘΑΡΗ, ΚΑΘΑΡΗΣ… Ο στόχος είναι εκφράσεις όπως *«Καθαρή Δευτέρα»*, *«Μεγάλη Τρίτη»*.
- **relative**: αρχή, τέλος, νωρίς, μέσα, περασμένος, προηγούμενος, επόμενος, αρχής,…, ΑΡΧΗ…

Επίσης, οι αριθμοί αντικαθίσταται από μία έγκυρη κανονική έκφραση της Java, στην οποία καθορίζεται ο αριθμός των ψηφίων. Για παράδειγμα, το «*2000»* αντικαθίσταται από το *[0-9]{4}*.

Στην περίπτωση που μια λεκτική μονάδα χρονικής έκφρασης δεν ανήκει σε καμία από τις παραπάνω περιπτώσεις, διατηρείται ως έχει.

3.3.1.3 Συνδυασμός αριθμητικών εκφράσεων

Κατά τη φάση αυτή επεξεργαζόμαστε μόνο πρότυπα που περιέχουν αριθμούς. Πρότυπα που διαφέρουν μόνο σε αριθμούς συνδυάζονται, χρησιμοποιώντας διαζεύξεις. Έστω, για παράδειγμα, ότι έχουμε αρχικά τις παρακάτω δύο ημερομηνίες και το αντίστοιχο πρότυπο, το οποίο προκύπτει μετά τη γενίκευση:

- `1/12/2005` ➔ `[0-9]{1} SEPARATOR [0-9]{2} SEPARATOR [0-9]{4}`
- `10.1.2005` ➔ `[0-9]{2} SEPARATOR [0-9]{1} SEPARATOR [0-9]{4}`

Τα δύο πρότυπα συνδυάζονται ως εξής:

```
([0-9]{1}|[0-9]{2}) SEPARATOR ([0-9]{1}|[0-9]{2}) SEPARATOR [0-9]{4}
```

3.3.1.4 Αποκοπή σπάνια εμφανιζόμενων προτύπων

Η φάση αυτή είναι προαιρετική και δε χρησιμοποιείται στα πειράματα που θα παρουσιαστούν στο επόμενο κεφάλαιο (ενότητα 4.2). Ο σκοπός της είναι η απομάκρυνση των προτύπων που δεν έχουν μεγάλο αριθμό εμφανίσεων στα κείμενα, ούτως ώστε να μειωθεί ο χρόνος αναζήτησης χρονικών εκφράσεων σε νέα κείμενα, που απαιτεί ελέγχους ταιριάσματος με όλα τα πρότυπα που έχουν παραχθεί κατά την εκπαίδευση. Το πρόβλημα, βέβαια, που δημιουργείται με τη μείωση του αριθμού των προτύπων είναι η μείωση της ανάκλησης (ενότητα 4.2) του συστήματος, καθώς και το ότι δεν είναι προφανές ποιο πρέπει να είναι το κατώφλι του αριθμού εμφανίσεων για την αποκοπή.

3.3.1.5 Προσθήκη επιπλέον προτύπων

Στη διάρκεια αυτής της φάσης, που είναι προαιρετική, ο χρήστης μπορεί να εισαγάγει χειρωνακτικά επιπλέον πρότυπα, πιθανώς παραλλαγές των προτύπων που έχουν παραχθεί αυτόματα, ώστε να καλυφθούν, για παράδειγμα, περιπτώσεις που δεν εμφανίζονται στα κείμενα εκπαίδευσης αλλά θεωρεί ότι ενδέχεται να εμφανιστούν σε

νέα κείμενα. Σε όλα τα πειράματα που διεξήχθησαν και θα παρουσιαστούν στο επόμενο κεφάλαιο δεν προστέθηκε κανένα χειρωνακτικά κατασκευασμένο πρότυπο.

3.3.1.6 Ταξινόμηση

Η ταξινόμηση γίνεται ως προς το μήκος των προτύπων κατά φθίνουσα σειρά. Έχει ως σκοπό να εξετάζονται πρώτα τα μεγαλύτερα σε μήκος πρότυπα. Έστω, για παράδειγμα, τα δύο παρακάτω πρότυπα:

```
(a) ([0-9]{2}|[0-9]{1}) MONTH
(b) ([0-9]{2}|[0-9]{1}) MONTH [0-9]{4}
```

Αν εξεταστεί πρώτα το (a), δε θα εντοπιστεί ολόκληρη η έκφραση *«31 Ιανουαρίου 2001»*.

Επιπλέον, ως δευτερεύον κλειδί για την ταξινόμηση χρησιμοποιείται η συχνότητα του προτύπου στα κείμενα εκπαίδευσης.

3.4 Αναγνώριση ονομάτων προσώπων

Η αναγνώριση των ονομάτων προσώπων, όπως έχει ήδη αναφερθεί, προϋποθέτει και τα τρία πρώτα στάδια της διαδικασίας που παρουσιάζεται στην αρχή του κεφαλαίου. Ειδικότερα, το πρώτο στάδιο δημιουργεί τη δομή πάνω στην οποία εκτελούνται οι διάφορες αναζητήσεις και πράξεις, ενώ το δεύτερο και το τρίτο στάδιο τροφοδοτούν με ιδιότητες τη αναγνώριση ονομάτων προσώπων.

3.4.1 Αρχική προσέγγιση

Αρχικά, είχε ακολουθηθεί κοινή προσέγγιση για τα ονόματα οντοτήτων τύπου enamex. Υπήρχε, δηλαδή, ένας ενιαίος ταξινομητής που κατέτασσε κάθε λεκτική μονάδα σε ακριβώς μία από τις κατηγορίες: όχι όνομα οντότητας, όνομα προσώπου, όνομα οργανισμού, όνομα τοποθεσίας. Το θετικό με αυτήν την προσέγγιση είναι ότι κάθε λεκτική μονάδα στο τέλος της ταξινόμησης ανήκει σε μόνο μία κατηγορία, ενώ στην περίπτωση που υπάρχει ένας διαφορετικός ταξινομητής για κάθε κατηγορία (π.χ. ένας ταξινομητής για την κατηγορία των ονομάτων προσώπων, που αποφαίνεται για κάθε λεκτική μονάδα αν αποτελεί μέρος ονόματος προσώπου ή όχι, ένας ταξινομητής για την κατηγορία των ονομάτων οργανισμών κ.ο.κ.) είναι δυνατόν μία λεκτική μονάδα να καταλήξει να καταταγεί σε πολλές κατηγορίες ταυτόχρονα. Από την άλλη πλευρά, όμως, χρησιμοποιώντας έναν ξεχωριστό ταξινομητή για κάθε κατηγορία είναι δυνατόν να χρησιμοποιηθεί σε κάθε ταξινομητή διαφορετικό σύνολο ιδιοτήτων, που να εξειδικεύεται στην παροχή πληροφοριών που είναι χρήσιμες για τον εντοπισμό ονομάτων της συγκεκριμένης κατηγορίας. Επίσης, η ανάπτυξη των ταξινομητών μπορεί να γίνει ανεξάρτητα.

Τελικά, αποφασίσαμε να ακολουθήσουμε την προσέγγιση του ενός ταξινομητή ανά κατηγορία και εστιαστήκαμε στην ανάπτυξη του ταξινομητή που εντοπίζει ονόματα προσώπων, αφήνοντας για επόμενες εργασίες την ανάπτυξη των ταξινομητών για τις άλλες υποκατηγορίες της enamex. Στο κεφάλαιο 5 παρατίθενται

σκέψεις για το πώς θα μπορούσαν να συνδεθούν σε ένα ενιαίο σύστημα οι ταξινομητές όλων των υποκατηγοριών της enamex.

3.4.2 Εντοπισμός ονομάτων προσώπων – 1^{ο} πέρασμα

Για τον εντοπισμό των ονομάτων προσώπων χρησιμοποιήθηκε μία ΜΔΥ. Είναι απαραίτητο, επομένως, και σε αυτήν την περίπτωση οι λεκτικές μονάδες να παρασταθούν ως διανύσματα ιδιοτήτων.

Ένα σημαντικό πρόβλημα είναι ότι δε διαθέταμε επισημειωτή μερών του λόγου (part-of-speech tagger), ο οποίος είναι βασικό κομμάτι των περισσοτέρων ελληνικών συστημάτων της βιβλιογραφίας. Για αυτόν το λόγο χρησιμοποιήσαμε πολλές ιδιότητες που παρέχουν πληροφορίες για τις καταλήξεις των λέξεων. Με αυτές τις ιδιότητες γίνεται προσπάθεια κυρίως να καθοριστεί ο αριθμός (ενικός / πληθυντικός) των λέξεων και η πτώση όσο είναι δυνατόν.

Γενικά, οι ιδιότητες που χρησιμοποιήθηκαν είναι δυνατόν να χωριστούν σε δύο είδη. Πρώτον, σε ιδιότητες που δεν εξαρτώνται από τη φύση της συλλογής των κειμένων, όπως:

1. Τύπος της λεκτικής μονάδας, όπως λέξη με ελληνικούς ή λατινικούς χαρακτήρες (δεν υπάρχει περίπτωση στην ίδια λεκτική μονάδα να υπάρχουν και ελληνικοί και λατινικοί χαρακτήρες, λόγω του τρόπου λειτουργίας του διαχωριστή λεκτικών μονάδων, ενότητα 3.1), σύμβολο (τελεία, κόμμα), ακολουθία αριθμών. Συνήθως τα ονόματα προσώπων αποτελούνται από ακολουθίες ελληνικών ή λατινικών χαρακτήρων, ενώ πολύ συχνά σε κείμενα εφημερίδων (σε τέτοια κείμενα διεξήχθησαν τα πειράματα) τα ονόματα προσώπων περιέχουν τελεία «.», για παράδειγμα «Κ. Σημίτης».
2. Μήκος της λεκτικής μονάδας (σε χαρακτήρες). Η ιδιότητα αυτή βοηθάει στον εντοπισμό αρχικών ονομάτων προσώπων, για παράδειγμα «Κ.Χ. Μύρης».
3. Απόσταση από την αρχή ονόματος προσώπου. Για παράδειγμα, στη φράση *«..Ο κ. <ENAMEX TYPE="PERSON">Κ. Σημίτης</ENAMEX>...»* η λεκτική μονάδα «Σημίτης» έχει απόσταση δύο από την αρχή του ονόματος προσώπου. Κατά μέσο όρο τα ονόματα προσώπων αποτελούνται από τρεις λεκτικές μονάδες και συνήθως από λιγότερες από πέντε λεκτικές μονάδες. Στη φάση της εκπαίδευσης, η τιμή της ιδιότητας αυτής υπολογίζεται εύκολα από τις επισημειώσεις των κειμένων. Στη φάση χρήσης[5] του συστήματος, η τιμή της ιδιότητας καθορίζεται από τις προηγούμενες αποφάσεις του ταξινομητή. Στην περίπτωση όπου δεν υπάρχει «ανοιγμένο» όνομα προσώπου η τιμή της ιδιότητας είναι -1.
4. Υπάρχει το «κ.» πριν από την υπό εξέταση λεκτική μονάδα; Συνήθως τα κείμενα των εφημερίδων χρησιμοποιούν έναν τυποποιημένο τρόπο γραφής, με αποτέλεσμα πολλά ονόματα προσώπων να έπονται του «κ.».
5. Υπάρχει το «κ.κ.» ή το «κκ.» ή το «κκ» πριν από την υπό εξέταση λεκτική μονάδα (μία δυαδική ιδιότητα); Ομοίως με το 4.
6. Η λεκτική μονάδα σηματοδοτεί τέλος περιόδου; Η πληροφορία αυτή έχει ήδη προστεθεί στη δομή από τον διαχωριστή περιόδων (ενότητα 3.1) ή έχει

[5] Ως φάση χρήσης του συστήματος ορίζεται η περίοδος λειτουργίας του σε αντιδιαστολή με τη φάση εκπαίδευσης. Η φάση χρήσης δε σχετίζεται με τη φάση αξιολόγησης.

αντληθεί στη φάση του διαχωρισμού σε λεκτικές μονάδες από το κείμενο με τη βοήθεια της ετικέτας <PARAGRAPH> (ενότητα 4.1.1).

7. Η λεκτική μονάδα βρίσκεται μέσα σε τίτλο ή υπότιτλο; Η πληροφορία αυτή αντλείται κατά το στάδιο του διαχωρισμού σε λεκτικές μονάδες με τη βοήθεια της ετικέτας <SUBTITLE> που πιθανότατα υπάρχει στο κείμενο (ενότητα 4.1.1).
8. Η λεκτική μονάδα αρχίζει με κεφαλαίο γράμμα ή όλη η λέξη είναι γραμμένη με κεφαλαία γράμματα (δύο διαφορετικές δυαδικές ιδιότητες);
9. Η κατάληξή της είναι χαρακτηριστική ελληνικών επωνύμων (-άκης, -ίδης, …);
10. Η αρχή της είναι χαρακτηριστική ελληνικών επωνύμων (παπά-, χατζή-, …);
11. Η κατάληξή της είναι (-ς, -ου, -οι, -ων); Συνήθως τα ονόματα προσώπων είναι στον ενικό αριθμό, ενώ συχνότερα εμφανίζονται στην ονομαστική πτώση (κατάληξη -ς).

Οι παραπάνω ιδιότητες, εκτός από τις 3, 4 και 5, δημιουργούνται για κάθε λεκτική μονάδα σε ένα παράθυρο 5 λεκτικών μονάδων γύρω από την υπό κατάταξη λεκτική μονάδα, δηλαδή λαμβάνεται υπόψη η τρέχουσα λεκτική μονάδα ±2. Για παράδειγμα, στην περίπτωση της ιδιότητας 8, χρησιμοποιούνται στην πραγματικότητα πέντε ιδιότητες: «αρχίζει η υπό κατάταξη / η προηγούμενή της / η προπροηγούμενή της / η επόμενή της / η μεθεπόμενή της λεκτική μονάδα με κεφαλαίο γράμμα;». Η ιδιότητα 4 απαιτεί την ύπαρξη του «κ.» ακριβώς πριν από την υπό κατάταξη λεκτική μονάδα. Η ιδιότητα 5 παίρνει αληθή τιμή (+1) αν υπάρχουν οι φράσεις «κ.κ.», «κκ.» και «κκ» σε ένα παράθυρο μέχρι 7 λεκτικές μονάδες πριν από την υπό κατάταξη λεκτική μονάδα. Επίσης, οι ιδιότητες είναι δυαδικές (αληθές / ψευδές), εκτός από αυτές που αφορούν μήκη ή αποστάσεις, που παίρνουν τιμές στο διάστημα [-1, 1].

Επιπλέον, υπάρχει μία ταξινομημένη λίστα με 350 ελληνικά βαφτιστικά ονόματα προσώπων και υποκοριστικά τους στην ονομαστική πτώση και μια επιπλέον ιδιότητα που δείχνει αν η υπό κατάταξη λεκτική μονάδα υπάρχει ή όχι μέσα σε αυτή τη λίστα. Πιο συγκεκριμένα, η ιδιότητα αυτή παίρνει τιμές στο [-1, 1], αναλόγως με το κατά πόσο η υπό κατάταξη λεκτική μονάδα ταιριάζει με την πιο κοντινή της από τις λέξεις στη λίστα. Για παράδειγμα, έστω ότι η υπό κατάταξη λεκτική μονάδα είναι η λέξη «Μιλτιάδη». Οι πιο κοντινές της λέξεις στην ταξινομημένη λίστα είναι οι «Μηνάς» και «Μιλτιάδης», αν δηλαδή ανήκε στη λίστα θα τοποθετούνταν ανάμεσά τους. Στη συνέχεια, υπολογίζουμε την ομοιότητα της υπό κατάταξη λέξης με κάθε μία από τις δύο κοντινές της και επιλέγουμε τη μεγαλύτερη τιμή. Η ομοιότητα δύο λέξεων παίρνει ως τιμή τη θέση του τελευταίου κοινού τους χαρακτήρα, ξεκινώντας από τα αριστερά προς τα δεξιά. Στο παράδειγμα η ομοιότητα είναι $\max(1,8) = 8$. Η τιμή αυτή κανονικοποιείται στο διάστημα [-1, 1]. Με αυτόν τον τρόπο γίνεται προσπάθεια να εντοπιστούν γνωστά βαφτιστικά ονόματα προσώπων, ακόμα και αν δεν είναι γραμμένα στην ονομαστική πτώση.

Το δεύτερο είδος ιδιοτήτων περιλαμβάνει ιδιότητες που εξαρτώνται από τη φύση των κειμένων. Πιο συγκεκριμένα, δημιουργούνται κάποιες λίστες από τα κείμενα εκπαίδευσης. Οι λίστες αυτές περιέχουν συχνές λέξεις που εμφανίζονται πριν από τα ονόματα προσώπων, σε ένα παράθυρο 1 ή 7 λεκτικών μονάδων. Οι λίστες αποθηκεύουν λέξεις αναλόγως του μήκους των λέξεων σε χαρακτήρες και του μεγέθους του παραθύρου. Συγκεκριμένα, υπάρχουν έξι λίστες. Η πρώτη περιέχει μικρού μήκους λέξεις (1-2 χαρακτήρες) οι οποίες βρίσκονται αμέσως πριν από ονόματα προσώπων στα κείμενα εκπαίδευσης (δηλαδή το μέγεθος του παραθύρου είναι 1). Η δεύτερη λίστα περιέχει λέξεις μικρού μήκους που εμφανίζονται πριν από

τα ονόματα προσώπων σε ένα παράθυρο 7 λεκτικών μονάδων στα κείμενα εκπαίδευσης. Η τρίτη λίστα περιέχει λέξεις μεσαίου μήκους (3-4 χαρακτήρες) οι οποίες βρίσκονται αμέσως πριν από ονόματα προσώπων στα κείμενα εκπαίδευσης κ.ο.κ. Ο στόχος όσον αφορά το παράθυρο μεγέθους 1 είναι να δοθεί έμφαση σε άρθρα καθώς και λέξεις όπως «δρ», «σερ», «στρατηγός», «αρχιεπίσκοπος» κ.α. Το μεγαλύτερο παράθυρο αποσκοπεί στη συλλογή λέξεων όπως «πρωθυπουργός», «υπουργός», «πρόεδρος», «διευθυντής», οι οποίες συνήθως δεν εμφανίζονται ακριβώς πριν από ονόματα προσώπων. Η απόφαση για το διαχωρισμό των λιστών με βάση το μέγεθος του παραθύρου, όπως επίσης και τα μεγέθη των παραθύρων που επιλέχθηκαν, αποτελούν απόρροια πειραματικών δοκιμών. Επιπλέον, για κάθε λέξη που περιλαμβάνεται στις λίστες υπολογίζεται και ο αριθμός εμφανίσεών της στο παράθυρο 1 ή 7 λεκτικών μονάδων πριν από τα ονόματα προσώπων στα κείμενα εκπαίδευσης. Ο διαχωρισμός με βάση το μήκος των λέξεων είναι απαραίτητος γιατί οι μικρού μήκους λέξεις (κυρίως άρθρα) εμφανίζονται συχνότερα με αποτέλεσμα να επικαλύπτουν τις μεγαλύτερες σε μήκος λέξεις (ο λόγος διαχωρισμού φαίνεται καθαρότερα στη συνέχεια).

Στη φάση χρήσης του συστήματος, αν η υπό κατάταξη λεκτική μονάδα έχει συμφραζόμενα που ανήκουν στη λίστα, τότε η ιδιότητα αυτής της λίστας (υπάρχει μία ιδιότητα ανά λίστα) παίρνει ως τιμή το άθροισμα των αριθμών εμφανίσεων των συμφραζομένων στη λίστα, κανονικοποιημένο στο διάστημα [-1, 1] (ο αριθμός εμφανίσεων, όπως έχει ήδη αναφερθεί, υπολογίζεται από τα κείμενα εκπαίδευσης). Ως συμφραζόμενα εννοούνται οι λέξεις που υπάρχουν σε ένα παράθυρο 1 ή 7 λεκτικών μονάδων πριν από την υπό κατάταξη λεκτική μονάδα. Σε αυτό το σημείο φαίνεται καθαρότερα ο λόγος του διαχωρισμού των λέξεων με βάση το μήκος τους. Για παράδειγμα, ο αριθμός εμφανίσεων του άρθρου «ο» είναι της τάξης των εκατοντάδων, ενώ της λέξης «πρωθυπουργός» της τάξης των δεκάδων. Οπότε, στην περίπτωση όπου δεν υπήρχε διαχωρισμός των λιστών με βάση το μήκος των λέξεων που περιέχουν, η τιμή των ιδιοτήτων θα εξαρτιόταν κυρίως από την ύπαρξη ή όχι στα συμφραζόμενα της υπό κατάταξης λεκτικής μονάδας του άρθρου «ο», ενώ η ύπαρξη της λέξης «πρωθυπουργός» θα είχε σχεδόν μηδενική επίδραση.

Στην πράξη οι λίστες έχουν αποδειχθεί ιδιαίτερα χρήσιμες. Είναι χαρακτηριστικό πως περιέχουν με μεγάλη συχνότητα λέξεις όπως «πρόεδρος», «πρωθυπουργός», «υπουργός», «διευθυντής», «καθηγητής», καθώς και άρθρα και προθέσεις που εμφανίζονται συχνά στα συμφραζόμενα ονομάτων προσώπων.

Κατά την επιλογή των ιδιοτήτων χρησιμοποιήθηκε κώδικας του συστήματος WEKA [5], ο οποίος αξιολογεί κάθε ιδιότητα μετρώντας το πληροφοριακό κέρδος (information gain) που παρέχει, με βάση τα δεδομένα εκπαίδευσης. Πριν ορίσουμε το πληροφοριακό κέρδος είναι απαραίτητο να οριστεί η έννοια της εντροπίας. Η εντροπία H(C) της τυχαίας μεταβλητής C (στην περίπτωση μας η C εκφράζει την απόκριση του ταξινομητή) δείχνει πόσο αβέβαιοι είμαστε για την τιμή της C. Η αβεβαιότητα μπορεί να οριστεί ως ο αναμενόμενος αριθμός δυφίων που πρέπει να μεταδοθούν ούτως ώστε να καθοριστεί η τιμή της C. Η εντροπία υπολογίζεται με τον τύπο:

$$H(C) = -\sum_{c} P(C = c) \cdot \log_2 P(C = c)$$

όπου c οι διάφορες τιμές που μπορεί να πάρει η C. Η εκτίμηση των πιθανοτήτων γίνεται από τα παραδείγματα εκπαίδευσης. Στην περίπτωση μας, όπου έχουμε μόνο

δύο κατηγορίες, όνομα προσώπου ή μη όνομα προσώπου, ο τύπος της εντροπίας γίνεται:

$$H(C) = -P(C=-1)\cdot \log_2 P(C=-1) - P(C=1)\cdot \log_2 P(C=1)$$

Αν γνωρίζουμε ότι η τιμή της ιδιότητας X είναι x, τότε η εντροπία υπολογίζεται από τον τύπο:

$$H(C \mid X=x) = -\sum_c P(C=c \mid X=x)\cdot \log_2 P(C=c \mid X=x)$$

όπου $P(C=c \mid X=x)$ είναι η δεσμευμένη πιθανότητα, εκφράζει δηλαδή την πιθανότητα η τιμή της C να είναι c αν γνωρίζουμε ότι η τιμή της X είναι x. Τελικά, το πληροφοριακό κέρδος IG της ιδιότητας X ορίζεται ως εξής:

$$IG(C,X) = H(C) - \sum_x P(X=x)\cdot H(C \mid X=x)$$

Με γνώμονα αυτήν την αξιολόγηση απομακρύνθηκαν κάποιες ιδιότητες. Για παράδειγμα, οι ιδιότητες «είναι ο τύπος της προπροηγούμενης / μεθεπόμενης λέξης ακολουθία αριθμών / κόμμα;» απομακρύνθηκαν (δηλαδή μειώθηκε το παράθυρο από 5 σε 3 για τη συγκεκριμένη ομάδα ιδιοτήτων), λόγω του ιδιαίτερα χαμηλού πληροφοριακού κέρδους. Η αξιολόγηση των ιδιοτήτων εκτελέστηκε μία φορά πριν αρχίσουν τα πειράματα που θα παρουσιαστούν στην ενότητα 4.2, με τη χρήση της πρώτης από τις δύο συλλογές κειμένων που θα παρουσιαστεί στην ενότητα 4.1.

Τελικά, χρησιμοποιούνται συνολικά 65 ιδιότητες, εκ των οποίων οι 59 είναι ανεξάρτητες της συλλογής κειμένων και οι υπόλοιπες έξι αφορούν τις λίστες που κατασκευάζονται από τα δεδομένα εκπαίδευσης, όπως περιγράφηκε παραπάνω.

Ένα σημαντικό πρόβλημα, το οποίο αντιμετωπίστηκε, είναι ο μεγάλος αριθμός λεκτικών μονάδων που ανήκουν στην κατηγορία των μη ονομάτων προσώπων σε σχέση με τον αριθμό των λεκτικών μονάδων της κατηγορίας των ονομάτων προσώπων, σε αναλογία 42:1. Το γεγονός αυτό ωθεί τη ΜΔΥ στο να μάθει να κατατάσσει όλα τα διανύσματα ως μη ονόματα προσώπων, καθώς κατά την εκπαίδευση «προτιμάει» την ύπαρξη των λίγων ξ_j των λανθασμένων διανυσμάτων της κατηγορίας των ονομάτων προσώπων από την ύπαρξη των πολλών ξ_j των λανθασμένων διανυσμάτων της κατηγορίας μη ονομάτων προσώπων (υπενθυμίζεται ότι ξ_j είναι το σφάλμα για κάθε διάνυσμα εκπαίδευσης x_j, δηλαδή το πόσο απέχουμε από το να ικανοποιείται ο αντίστοιχος περιορισμός, ενότητα 2.3.1). Το γεγονός αυτό έχει ως αποτέλεσμα τη μεταφορά του υπερεπιπέδου διαχωρισμού προς όφελος της κατηγορίας των μη ονομάτων προσώπων. Επίσης, ο μεγάλος αριθμός διανυσμάτων, πολλά από τα οποία πιθανότατα δε βοηθούν στον εντοπισμό ονομάτων προσώπων, επιβαρύνει ιδιαίτερα τον χρόνο εκπαίδευσης. Για αυτούς τους λόγους, έγινε προσπάθεια λεκτικές μονάδες που σίγουρα δεν ανήκουν στην κατηγορία των ονομάτων προσώπων να εντοπίζονται σε ένα αρχικό στάδιο, χρησιμοποιώντας κάποιους «ασφαλείς κανόνες», και να μην ερωτάται για αυτές η ΜΔΥ. Ο σκοπός είναι να μειωθούν τα διανύσματα της συχνής κατηγορίας (μη ονόματα προσώπων) που συναντά ο αλγόριθμος μάθησης κατά την εκπαίδευση και την επεξεργασία νέων

κειμένων, αποκρύπτοντάς του περιπτώσεις λεκτικών μονάδων για τις οποίες μπορούμε να είμαστε σίγουροι ότι δεν αποτελούν ονόματα προσώπων.

Αρχικά, λοιπόν, αποκλείονται από τον αλγόριθμο μηχανικής μάθησης τα σύμβολα, οι αριθμοί και οι λεκτικές μονάδες που συμμετέχουν σε χρονικές εκφράσεις. Επίσης, ο αλγόριθμος μάθησης ερωτάται μόνο για λέξεις που αρχίζουν με κεφαλαίο γράμμα ή είναι γραμμένες πλήρως με κεφαλαίους χαρακτήρες (του ζητείται να μάθει να κατατάσσει μόνο λέξεις αυτού του είδους). Επιπλέον, δημιουργήθηκε από τα κείμενα εκπαίδευσης μία λίστα με συχνά εμφανιζόμενες λέξεις (stop words). Η λίστα αυτή περιέχει λέξεις οι οποίες εμφανίζονται στα κείμενα εκπαίδευσης τουλάχιστον τόσες φορές όσα είναι τα κείμενα. Η ΜΔΥ δεν ερωτάται για λέξεις που ανήκουν σε αυτήν την λίστα, ούτε και εκπαιδεύεται στην κατάταξή τους. Επίσης, έχουν επιλεχθεί κάποιες καταλήξεις ρημάτων, όπως -ωνω, -μαι, -σαι, -ται, και οι λέξεις με αυτές τις καταλήξεις θεωρούνται ρήματα, άρα όχι ονόματα προσώπων, και δεν ερωτάται για αυτές η ΜΔΥ. Όλα τα παραπάνω ανατρέπονται στην περίπτωση που η προηγούμενη λεκτική μονάδα της υπό κατάταξης λεκτικής μονάδας έχει αποφασιστεί ότι είναι όνομα προσώπου. Στην περίπτωση, δηλαδή, που γνωρίζουμε ότι η προηγούμενη λεκτική μονάδα έχει καταταγεί ως όνομα προσώπου, το διάνυσμα της υπό κατάταξης λεκτικής μονάδας, είναι ορατό από τη ΜΔΥ, ακόμα και αν ισχύει κάποιος από τους παραπάνω κανόνες. Με αυτές τις παραδοχές η αναλογία μειώνεται εντυπωσιακά σε 3,5:1.

Κατά την αξιολόγηση του συστήματος, οι λεκτικές μονάδες που δεν είναι ορατές από τη ΜΔΥ μετά την εφαρμογή των «ασφαλών» κανόνων (οι οποίες κατατάσσονται ως μη ονόματα προσώπων) συμπεριλαμβάνονται στον υπολογισμό των μέτρων επίδοσης.

Στην πράξη οι ασφαλείς κανόνες λειτουργούν ιδιαίτερα θετικά. Συγκεκριμένα, αποκλείουν από τη ΜΔΥ μόλις το 0,2% των λεκτικών μονάδων που αποτελούν όνομα προσώπου. Τα λάθη αυτά παρατηρούνται κυρίως σε ονόματα προσώπων που περιέχουν σύμβολα ή αριθμούς, για παράδειγμα «Γιάννα Αγγελοπούλου – Δασκαλάκη», «Λουδοβίκος ο 16ος». Το πρόβλημα αυτό δημιουργείται μόνο στην περίπτωση όπου δεν έχουν εντοπιστεί οι προηγούμενες λεκτικές μονάδες (πριν από το σύμβολο ή τον αριθμό) του ονόματος προσώπου, αν, δηλαδή, δεν έχουν εντοπιστεί τα «Γιάννα Αγγελοπούλου» και «Λουδοβίκος ο» στα παραδείγματα αντίστοιχα. Τέλος, η μείωση των διανυσμάτων εκπαίδευσης, ελαττώνει σημαντικά τον χρόνο εκπαίδευσης.

3.4.3 Εντοπισμός ονομάτων προσώπων – 2^{ο} πέρασμα

Η διαδικασία που περιγράψαμε στην προηγούμενη ενότητα (3.4.2) αφορά την εκπαίδευση μίας ΜΔΥ για την αναγνώριση ονομάτων προσώπων. Σε αυτήν την ενότητα θα περιγραφεί η διαδικασία κατασκευής μίας δεύτερης ΜΔΥ, η οποία θα συμπληρώνει την πρώτη ΜΔΥ, θα εκτελείται μετά από αυτήν και θα εξαρτάται από αυτήν. Στη συνέχεια, ως πρώτο πέρασμα εννοούμε τη ΜΔΥ της προηγούμενης ενότητας, ενώ ως δεύτερο πέρασμα τη ΜΔΥ που θα περιγράψουμε σε αυτήν την ενότητα.

Η τεχνική των πολλαπλών περασμάτων (στην περίπτωσή μας έχουμε δύο περάσματα) έχει εφαρμοστεί σε συστήματα της βιβλιογραφίας, με χαρακτηριστικότερο παράδειγμα το σύστημα του Εδιμβούργου [24], όπου υπάρχουν πέντε περάσματα (ενότητα 2.1). Η λογική των πολλαπλών περασμάτων αποβλέπει στην χρησιμοποίηση της «γνώσης» που αποκτάται σε προηγούμενα περάσματα. Στην περίπτωσή μας υπάρχουν δύο ταξινομητές, ένας για το πρώτο και ένας για το δεύτερο

πέρασμα, όπου ο πρώτος τροφοδοτεί με ιδιότητες τον δεύτερο. Πιο συγκεκριμένα, η δεύτερη ΜΔΥ λαμβάνει υπόψη της το αν η πρώτη κατέταξε την υπό εξέταση λεκτική μονάδα αλλού στο ίδιο κείμενο ως όνομα προσώπου με υψηλή βεβαιότητα. Ο βασικός στόχος του δεύτερου περάσματος είναι η αύξηση του ποσοστού της ανάκλησης (ορίζεται στην ενότητα 4.2) λεκτικών μονάδων που συμμετέχουν σε ονόματα προσώπων.

Η υλοποίηση των ΜΔΥ που χρησιμοποιήσαμε (libSVM [3]) έχει τη δυνατότητα να επιστρέφει κατά την κατάταξη όχι μόνο την κατηγορία στην οποία ανήκει η υπό κατάταξη λεκτική μονάδα (βασιζόμενη στον τελευταίο τύπο της ενότητας 2.3.1), αλλά και το βαθμό βεβαιότητας της κατάταξης της λεκτικής μονάδας σε αυτήν την κατηγορία. Για τον υπολογισμό του βαθμού βεβαιότητας χρησιμοποιείται η απόσταση του διανύσματος της λεκτικής μονάδας από το υπερεπίπεδο διαχωρισμού.

Στο δεύτερο πέρασμα χρησιμοποιούνται όλες οι ιδιότητες του πρώτου, απλά προστίθενται και μερικές ακόμα (προστίθενται 6, συνολικά 71 ιδιότητες). Οι καινούριες ιδιότητες προκύπτουν από ένα παράθυρο 3 λεκτικών μονάδων (η τρέχουσα ±1) και είναι δυνατόν να χωριστούν σε δύο είδη. Πρώτον, σε αυτές που αφορούν την απόσταση από το υπερεπίπεδο που υπολογίστηκε από το πρώτο πέρασμα. Οι ιδιότητες αυτές αφορούν την ετυμηγορία του ταξινομητή του πρώτου σταδίου για την υπό εξέταση λεκτική μονάδα, καθώς και για την προηγούμενή της και την επόμενή της (τρεις ξεχωριστές ιδιότητες στο διάστημα [-1, 1]).

Μία λεκτική μονάδα είναι δυνατόν να εμφανίζεται πολλές φορές σε ένα κείμενο και σε μερικές θέσεις η ΜΔΥ του πρώτου περάσματος μπορεί να είχε για παράδειγμα μεγάλη βεβαιότητα ότι πρόκειται για όνομα προσώπου ενώ σε άλλες όχι. Κατά τη διάρκεια του πρώτου περάσματος, όσες λεκτικές μονάδες κατηγοριοποιηθούν ως ονόματα προσώπων με αρκετά μεγάλη σιγουριά προστίθενται σε μια λίστα. Στο δεύτερο πέρασμα, ελέγχεται αν η υπό εξέταση λεκτική μονάδα ανήκει στη λίστα και αυτό βοηθά να εντοπιστούν και εμφανίσεις λεκτικών μονάδων για τις οποίες η ΜΔΥ του πρώτου περάσματος δεν ήταν βέβαιη αν αποτελούν ονόματα προσώπων.

Η λίστα αυτή δημιουργείται για κάθε κείμενο ξεχωριστά. Δεν υπάρχει, δηλαδή, μια κοινή λίστα για όλα τα κείμενα. Ελπίζουμε ότι οι εμφανίσεις ενός ονόματος οντότητας στο ίδιο κείμενο θα ανήκουν όλες στην ίδια κατηγορία. Για παράδειγμα, ένα κείμενο που αναφέρεται στον Ελευθέριο Βενιζέλο, ως όνομα προσώπου, είναι δύσκολο να αναφέρεται και στο αεροδρόμιο «Ελευθέριος Βενιζέλος» (τοπωνύμιο).

Επίσης, πρέπει να διευκρινίσουμε τι εννοούμε λέγοντας ότι στη λίστα προστίθενται οι λεκτικές μονάδες που τις κατηγοριοποίησε «με μεγάλη σιγουριά» το πρώτο πέρασμα ως ονόματα προσώπων. Χρησιμοποιείται και σε αυτήν την περίπτωση ο βαθμός βεβαιότητας της ΜΔΥ του πρώτου περάσματος. Οι λεκτικές μονάδες των οποίων κάποια εμφάνιση μέσα στο κείμενο κατετάγη ως όνομα προσώπου με βαθμό βεβαιότητας που υπερβαίνει κάποιο κατώφλι εισάγονται στη λίστα. Η επιλογή του κατωφλίου αφήνεται στο χρήστη (παίρνει τιμές στο [0, 1]). Η προκαθορισμένη τιμή που χρησιμοποιείται είναι 0,8 και για την κατηγορία ονομάτων προσώπων λειτουργεί ικανοποιητικά. Όσο μικρότερη είναι η τιμή του κατωφλίου, τόσο βελτιώνεται η ανάκληση (αφού κατατάσσονται περισσότερες λεκτικές μονάδες ως ονόματα προσώπων) και μειώνεται η ακρίβεια (αυξάνεται η πιθανότητα να κατατάξουμε λανθασμένα μια λεκτική μονάδα ως όνομα προσώπου).

Τελικά, το δεύτερο είδος ιδιοτήτων περιλαμβάνει 3 δυαδικές ιδιότητες, των οποίων οι τιμές εξαρτώνται από το αν η τρέχουσα / προηγούμενή της / επόμενή της λεκτική μονάδα έχει εμφανιστεί στο κείμενο και σε άλλη θέση στην οποία το πρώτο πέρασμα την κατέταξε ως όνομα προσώπου με μεγάλη βεβαιότητα.

Θα μπορούσε να ισχυριστεί κάποιος ότι το δεύτερο πέρασμα δεν είναι τίποτα άλλο από ένα είδος πρωτόγονου Boosting [30]. Αυτό, όμως, δεν ισχύει, καθώς οι δύο ταξινομητές έχουν διαφορετικό σύνολο ιδιοτήτων και ο δεύτερος ταξινομητής χρησιμοποιεί την απόφαση του πρώτου.

Τελικά, όπως θα φανεί και από τα αποτελέσματα (ενότητα 4.2.1), η κύρια επίδραση του δεύτερου περάσματος είναι η αύξηση του ποσοστού της ανάκλησης. Για να επιτευχθεί αυτό, βέβαια, είναι απαραίτητο να έχουμε μία ΜΔΥ στο πρώτο πέρασμα την οποία να μπορούμε να εμπιστευτούμε για τις λεκτικές μονάδες που κατατάσσει με υψηλή βεβαιότητα ως ονόματα προσώπων. Επιθυμούμε, δηλαδή, υψηλή ακρίβεια από το πρώτο πέρασμα.

3.4.4 Ενεργητική μάθηση

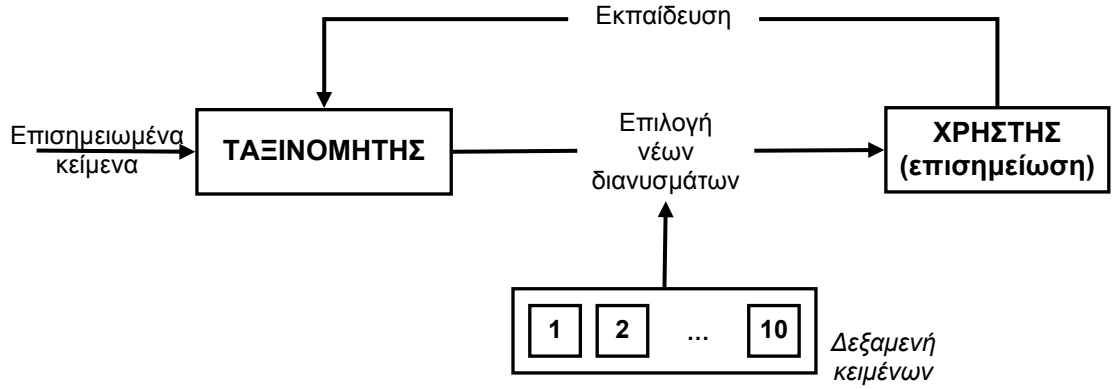
Ως ενεργητική μάθηση (active learning) θεωρούνται διάφορες μέθοδοι που επιτρέπουν στο ίδιο το σύστημα να προτείνει παραδείγματα (στην περίπτωσή μας λεκτικές μονάδες) που πρέπει να επισημειωθούν ώστε να προστεθούν στα δεδομένα εκπαίδευσης. Από την άλλη πλευρά, στην παθητική μάθηση (passive learning), η οποία χρησιμοποιήθηκε στις δύο προηγούμενες παραγράφους, τα προς επισημείωση παραδείγματα εκπαίδευσης επιλέγονται τυχαία. Τόσο η ενεργητική όσο και η παθητική μάθηση αποτελούν περιπτώσεις επιβλεπόμενης μάθησης. Η χρησιμότητα της ενεργητικής μάθησης έγκειται στο γεγονός ότι μειώνεται ο αριθμός των παραδειγμάτων που πρέπει να επισημειωθούν, αφού δεν επισημειώνονται περιπτώσεις που δεν ωφελούν ιδιαίτερα τη μάθηση. Υπάρχουν πολλές εργασίες στις οποίες περιγράφονται σχετικές μέθοδοι ενεργητικής μάθησης [32, 31, 13]. Μία ιδέα ιδιαίτερα διαδεδομένη στην περίπτωση των ΜΔΥ, η οποία περιγράφεται εκτενέστερα στη συνέχεια, είναι να επιλέγονται περιπτώσεις που βρίσκονται πολύ κοντά στο τρέχον υπερεπίπεδο διαχωρισμού.

Στην περίπτωσή μας, αρχικά η ΜΔΥ εκπαιδεύεται σε λίγα πλήρως επισημειωμένα κείμενα, δηλαδή κείμενα στα οποία έχουν επισημειωθεί όλα τα ονόματα προσώπων. Στη συνέχεια, επιλέγονται από το σύστημα οι πιο «αμφιλεγόμενες» Χ λεκτικές μονάδες από άλλα μη επισημειωμένα κείμενα και προτείνεται στο χρήστη να τις επισημειώσει με τις σωστές τους κατηγορίες. Οι επισημειωμένες αυτές λεκτικές μονάδες προστίθενται στα δεδομένα εκπαίδευσης και το σύστημα επανεκπαιδεύεται. Ακολούθως, επιλέγονται Χ νέες μη επισημειωμένες λεκτικές μονάδες, επισημειώνονται και το σύστημα επανεκπαιδεύεται. Η διαδικασία αυτή επαναλαμβάνεται έως όταν ικανοποιηθεί κάποιο κριτήριο τερματισμού, όπως για παράδειγμα η αύξηση των μέτρων επίδοσης κατά συγκεκριμένες ποσοστιαίες μονάδες ή η σταθεροποίηση των μέτρων επίδοσης παρατηρώντας την καμπύλη μάθησης ή η προσθήκη συγκεκριμένου αριθμού νέων διανυσμάτων.

Για να λειτουργήσει καλά η μέθοδος αυτή χρειάζεται μία αρκετά μεγάλη δεξαμενή μη επισημειωμένων κειμένων, ούτως ώστε το σύστημα να έχει τη δυνατότητα να επιλέξει περιπτώσεις που πρέπει να επισημειωθούν από ένα μεγάλο εύρος περιπτώσεων. Αν, όμως, η δεξαμενή είναι πολύ μεγάλη, αυξάνεται πολύ ο χρόνος που απαιτείται για την αξιολόγηση των υποψηφίων παραδειγμάτων εκπαίδευσης. Για αυτόν το λόγο, στα πειράματα που θα παρουσιαστούν στη συνέχεια, η δεξαμενή κειμένων έχει χωριστεί σε 10 ίσα μέρη. Σε κάθε επανάληψη επιλέγεται κυκλικά το επόμενο μέρος. Με αυτόν τον τρόπο μειώνεται ο χρόνος που απαιτείται για την αξιολόγηση των υποψηφίων παραδειγμάτων εκπαίδευσης, ενώ

εξακολουθούμε να αντλούμε παραδείγματα εκπαίδευσης από ολόκληρη τη δεξαμενή κειμένων.

Σε αυτό το σημείο είναι απαραίτητο να καθορίσουμε τι εννοούμε με τη φράση «πιο αμφιλεγόμενες λεκτικές μονάδες». Υπάρχουν διάφορα μέτρα [32] για να αποφασιστεί αν μία λεκτική μονάδα (γενικότερα ένα υποψήφιο παράδειγμα) πρέπει να δοθεί στο χρήστη για επισημείωση. Στα πειράματα αυτής της εργασίας, εκμεταλλευόμενοι τη φύση των ΜΔΥ, χρησιμοποιούμε ως μέτρο επιλογής την απόσταση του υποψηφίου παραδείγματος από το υπερεπίπεδο διαχωρισμού που έχει προκύψει από την τελευταία εκπαίδευσης της ΜΔΥ. Όσο πιο κοντά στο υπερεπίπεδο βρίσκεται ένα υποψήφιο παράδειγμα (το διάνυσμα ιδιοτήτων του), τόσο πιο αμφιλεγόμενο θεωρείται.



Ενεργητική Μάθηση

Η μοναδική παράμετρος της μεθόδου ενεργητικής μάθησης που παρουσιάστηκε παραπάνω είναι ο αριθμός Χ των διανυσμάτων που προστίθενται σε κάθε επανάληψη (batch size). Ο αριθμός αυτός παίζει σημαντικό ρόλο, λόγω της απλότητας του κριτηρίου επιλογής υποψηφίων παραδειγμάτων. Παρατηρήσαμε στα πειράματα της εργασίας (ενότητα 4.2.1) ότι αν το Χ λάβει σχετικά μικρή τιμή, επιτυγχάνεται ο στόχος μας, δηλαδή η αύξηση των μέτρων επίδοσης του συστήματος, χρησιμοποιώντας λιγότερα παραδείγματα εκπαίδευσης, καθώς περιορίζεται η επιλογή ακριβώς ίδιων διανυσμάτων κατά τη διάρκεια της ίδιας επανάληψης. Για παράδειγμα, παρατηρήσαμε ότι σε κάποια επανάληψη πολλά από τα παραδείγματα που προτάθηκαν από το σύστημα προς επισημείωση έμοιαζαν μεταξύ τους, καθώς αφορούσαν σχεδόν όλα την πρώτη λεκτική μονάδα του κειμένου (οι λίστες δεν παίζουν κανένα ρόλο αφού δεν υπάρχουν λεκτικές μονάδες πριν από αυτήν), η οποία βρισκόταν μέσα σε τίτλο και γενικά προέκυψαν διανύσματα με ακριβώς ίδιες τιμές ιδιοτήτων (ίδια διανύσματα έχουν ίδια απόσταση από το υπερεπίπεδο διαχωρισμού). Οπότε, σε αυτήν την επανάληψη προστέθηκαν (σχεδόν) Χ ακριβώς ίδια διανύσματα (χειρότερα και από παθητική μάθηση). Η συχνότητα εμφάνισης αυτών των περιπτώσεων περιορίζεται για σχετικά μικρό Χ. Από την άλλη πλευρά, αν ο αριθμός Χ είναι σχετικά μεγάλος θα χρειαστούν λιγότερες επανεκπαιδεύσεις της ΜΔΥ για να προστεθεί τελικά ο ίδιος αριθμός διανυσμάτων, γεγονός που σημαίνει ότι απαιτείται συνολικά λιγότερος χρόνος για τις επανεκπαιδεύσεις της ΜΔΥ. Στα πειράματα που θα παρουσιαστούν η τιμή της παραμέτρου είναι Χ = 100. Η επιλογή της έγινε με βάση τον μέσο αριθμό διανυσμάτων που υπάρχουν στα κείμενα που έχουν επισημειωθεί. Ένα κείμενο της πρώτης από τις δύο συλλογές (ενότητα 4) που χρησιμοποιήθηκαν (τα πειράματα για την επίδραση της ενεργητικής μάθησης διεξήχθησαν με την πρώτη συλλογή, ενότητα 4.2.1) περιέχει κατά μέσο όρο περίπου

100 διανύσματα, τα οποία είναι «ορατά» από τον αλγόριθμο μάθησης μετά την εφαρμογή των «ασφαλών» κανόνων. Επομένως, θέτοντας X = 100 είναι σαν να προσθέτουμε ένα κείμενο στα δεδομένα εκπαίδευσης (οι καμπύλες της παθητικής μάθησης που θα παρουσιαστούν στην ενότητα 4.2.1 προκύπτουν προσθέτοντας σταδιακά ένα επιπλέον κείμενο στα δεδομένα εκπαίδευσης).

4 Δεδομένα – Αποτελέσματα

Στο κεφάλαιο αυτό θα παρουσιαστούν τα αποτελέσματα των πειραμάτων που εκτελέστηκαν. Πρώτα, όμως, είναι απαραίτητο να γίνει μία αναφορά στα κείμενα των πειραμάτων, τη μορφή τους, την προέλευση και το είδος τους, καθώς και τον τρόπο που επισημειώθηκαν χειρωνακτικά με τις ορθές αποκρίσεις.

4.1 Κείμενα των πειραμάτων

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήθηκαν δύο διαφορετικές συλλογές κειμένων.

Τα κείμενα της πρώτης συλλογής προέρχονται από τις εφημερίδες «ΤΑ ΝΕΑ» [6] και «ΤΟ ΒΗΜΑ» [7]. Πιο συγκεκριμένα, από την δικτυακή σελίδα *http://ta-nea.dolnet.gr* ανακτήθηκαν από το αρχείο της εφημερίδας τα κείμενα της περιόδου Μάρτιος 2001 – Ιούλιος 2002. Επίσης, από την ιστοσελίδα *http://tovima.dolnet.gr* ανακτήθηκαν τα κείμενα της περιόδου Ιούλιος 2000 – Οκτώβριος 2001.

Οι δύο εφημερίδες, από τις οποίες προέρχονται τα άρθρα είναι ποικίλης ύλης. Επίσης, πολλά κείμενα προέρχονται από τα ένθετα περιοδικά των εφημερίδων. Γίνεται εύκολα αντιληπτό, λοιπόν, ότι η μορφή των κειμένων διαφέρει αρκετά. Επιπλέον, τα θέματά τους ποικίλλουν. Πιο συγκεκριμένα, υπάρχουν πολιτικά, οικονομικά, αθλητικά, πολιτιστικά κείμενα, αναλύσεις βιβλίων και θεατρικών έργων, συνεντεύξεις, αφηγήσεις ιστορικών γεγονότων κ.α. Ενδεικτικά αναφέρεται ότι υπάρχουν ακόμα και ενδεκάδες ομάδων και εκλογικά αποτελέσματα.

Τα κείμενα της δεύτερης συλλογής αποτελούνται από σύντομες ειδήσεις οικονομικού περιεχομένου από μία μόνο πηγή, που συγκεντρώθηκαν στη διάρκεια του ερευνητικού έργου «ΜΙΤΟΣ». Χαρακτηριστικά παραδείγματα αυτής της συλλογής αφορούν τιμές μετοχών, άρθρα για συγχωνεύσεις – εξαγορές εταιριών ή για τον κρατικό προϋπολογισμό. Η συλλογή περιέχει 715 επισημειωμένα κείμενα (βάσει των κατηγοριών του MUC-7) με μέσο μέγεθος 2,1K.

4.1.1 Προεπεξεργασία

Το στάδιο της προεπεξεργασίας αφορά μόνο την πρώτη συλλογή κειμένων, η οποία ανακτήθηκε από τις ιστοσελίδες των δύο εφημερίδων σε μορφή HTML. Οι περισσότερες ετικέτες της HTML δεν είναι χρήσιμες για τους σκοπούς της αναγνώρισης και κατάταξης ονομάτων οντοτήτων. Για αυτόν το λόγο, ήταν απαραίτητο να απομακρυνθούν. Οι μόνες ετικέτες που διατηρήθηκαν είναι αυτές που αφορούν επικεφαλίδες (<H>) και αυτές που αφορούν αλλαγές παραγράφων (<P>), οι οποίες μετατράπηκαν σε ετικέτες που υποστηρίζει το σύστημα (<SUBTITLE> και <PARAGRAPH> αντίστοιχα). Η πληροφορία που μας δίνεται από τις ετικέτες που διατηρούνται αφορά τη σηματοδότηση τέλους περιόδων (ετικέτα <PARAGRAPH>) και την ένταξη λεκτικών μονάδων σε τίτλους (ετικέτα <SUBTITLE>). Χρησιμοποιείται με τη μορφή ιδιοτήτων (ενότητα 3.4.2) των ΜΔΥ κατά το στάδιο της αναγνώρισης ονομάτων προσώπων, ενώ συμπληρώνει και το διαχωριστή περιόδων.

Επίσης, εντοπίστηκαν στα κείμενα πληροφορίες όπως ο συγγραφέας, η ημερομηνία, ο αριθμός φύλλου, οι οποίες εκφράζονται με σχετικά τυποποιημένους τρόπους σε κάθε εφημερίδα και μπορούν να εντοπιστούν εύκολα. Οι πληροφορίες αυτές σημειώθηκαν με ειδικές ετικέτες (<AUTHOR>, <DATE>, <ID>, <URL>), εντάχθηκαν σε κάθε κείμενο ως επικεφαλίδα (<HEAD>) και αγνοούνται στα επόμενα στάδια επεξεργασίας.

Η μορφή των κειμένων κάθε εφημερίδας είναι σχετικά τυποποιημένη, οπότε η προεπεξεργασία των κειμένων είναι αρκετά εύκολη. Παρόλα αυτά, παρατηρούνται κάποια λάθη, κυρίως όσον αφορά τον εντοπισμό των συγγραφέων. Επίσης, σε περίπτωση χρήσης του συστήματος με κείμενα από άλλες πηγές, πρέπει να προσαρμοστεί ο κώδικας του προεπεξεργαστή ανάλογα.

4.1.2 Επισημείωση κειμένων εκπαίδευσης

Από τα κείμενα της πρώτης συλλογής επιλέχθηκαν τυχαία και επισημειώθηκαν 400 (200 από κάθε εφημερίδα). Το μέσο μέγεθος κάθε κειμένου είναι περίπου 6,3K, από τα οποία περίπου τα 0,3K αφορούν την επικεφαλίδα που αναφέρθηκε στην προηγούμενη παράγραφο (καθαρά 6K). Ένα μέρος των επισημειωμένων κειμένων χρησιμοποιήθηκε για την εκπαίδευση του συστήματος και ένα μέρος για τον έλεγχο της απόδοσής του, όπως θα εξηγηθεί σε επόμενες ενότητες. Όπως αναφέρθηκε παραπάνω τα κείμενα της δεύτερης συλλογής είναι ήδη επισημειωμένα. Στον παρακάτω πίνακα φαίνεται ο αριθμός των ονομάτων οντοτήτων ανά κατηγορία στα επισημειωμένα κείμενα και στις δύο συλλογές κειμένων που χρησιμοποιήθηκαν.

		1η Συλλογή	2η Συλλογή
ENAMEX	*PERSON*	4797	1046
	ORGANIZATION	4265	4067
	LOCATION	3583	1107
TIMEX	*DATE*	1454	1244
	TIME	109	0
NUMEX	*MONEY*	517	0
	PERCENT	642	0

Αριθμός ονομάτων οντοτήτων ανά κατηγορία στα αρχικά επισημειωμένα κείμενα εκπαίδευσης

Κατά την επισημείωση των κειμένων της πρώτης συλλογής λάβαμε υπόψη τους κανόνες που προτείνονται από το MUC-7 [4]. Οι κανόνες αυτοί προσαρμόστηκαν κατάλληλα για τα ελληνικά και θα παρουσιαστούν συνοπτικά στη συνέχεια ανά κατηγορία ονομάτων οντοτήτων. Η επισημείωση των ονομάτων οργανισμών, των τοπωνυμίων και των αριθμητικών εκφράσεων έγινε με στόχο την διευκόλυνση ενδεχόμενης μελλοντικής επέκτασης του συστήματος, αφού στην τρέχουσα μορφή του το σύστημα δεν εντοπίζει ονόματα οντοτήτων αυτών των κατηγοριών. Οι κανόνες επισημείωσης που έχουν ακολουθηθεί για τη δεύτερη συλλογή είναι παρόμοιοι με αυτούς που θα περιγραφούν στη συνέχεια.

4.1.2.1 Ονόματα προσώπων

Στην κατηγορία των ονομάτων προσώπων (person) ανήκουν ονόματα φυσικών προσώπων ή οικογενειών, ψευδώνυμα, υποκοριστικά, καθώς και συντομογραφίες

των ονομάτων. Οι τίτλοι, όπως «κ.», «δρ.», «πρόεδρος», κλπ. θεωρείται ότι δεν αποτελούν ονόματα προσώπων ούτε μέρη τους. Τα ονόματα προσώπων όταν χρησιμοποιούνται ως επωνυμίες εταιριών ανήκουν στην κατηγορία organization. Ακολουθούν παραδείγματα:

κ. <ENAMEX TYPE="PERSON">Πέτρος Παπαδόπουλος</ENAMEX>
<ENAMEX TYPE=" PERSON">Κ. Χ. Μύρης</ENAMEX>
κυβέρνηση <ENAMEX TYPE=" PERSON">Σημίτη</ENAMEX>
οικογένεια <ENAMEX TYPE=" PERSON">Πετροβασίλη</ENAMEX>
ο <ENAMEX TYPE=" PERSON">Μ.Π.</ENAMEX> (Μάριος Πλωρίτης)
η <ENAMEX TYPE="ORG">Νικ. Ι. Θεοχαράκης</ENAMEX>

4.1.2.2 Ονόματα οργανισμών

Στην κατηγορία των ονομάτων οργανισμών (organization) ανήκουν ονόματα εταιριών, ομίλων, υπουργείων, χρηματιστηρίων, πανεπιστημίων, αθλητικών ομάδων, νοσοκομείων, πολιτικών κομμάτων, συλλόγων, τηλεοπτικών ή ραδιοφωνικών σταθμών κλπ., όπως επίσης και συντομογραφίες και παραφράσεις αυτών.

Στην περίπτωση κτητικής ή ιεραρχικής δομής δύο οργανισμών (π.χ. όμιλος και εταιρεία του ομίλου, εταιρεία και τμήμα της), τα ονόματα των δύο οργανισμών επισημειώνονται ξεχωριστά. Επιπλέον, ονόματα τοποθεσιών σε μορφή εμπρόθετων προσδιορισμών δεν θεωρείται ότι αποτελούν μέρη ονομάτων οργανισμών.

Λέξεις όπως «εταιρεία», «οργανισμός» δεν θεωρείται ότι αποτελούν ονόματα οργανισμών ούτε μέρη τους, εκτός αν αρχίζουν με κεφαλαίο γράμμα. Από αυτόν τον κανόνα εξαιρούνται οι λέξεις «υπουργείο» και «χρηματιστήριο», εφόσον προηγούνται του ονόματος του υπουργείου ή του χρηματιστηρίου αντίστοιχα.

Τέλος, συντομογραφίες του τύπου «Αφοι», «Α.Ε.», κλπ. θεωρείται ότι αποτελούν μέρη των ονομάτων των οργανισμών. Ακολουθούν παραδείγματα:

<ENAMEX TYPE="ORG">ΔΕΗ Α.Ε.</ENAMEX>
<ENAMEX TYPE="ORG">ΧΑΑ</ENAMEX>
<ENAMEX TYPE="ORG">Χρηματιστήριο Αξιών Αθηνών</ENAMEX>
της <ENAMEX TYPE="ORG">Σοφοκλέους</ENAMEX>
στη <ENAMEX TYPE="LOC">Σοφοκλέους</ENAMEX>
<ENAMEX TYPE="ORG">υπουργείο Εξωτερικών</ENAMEX>
<ENAMEX TYPE="ORG">Εθνικό Μουσείο της Κίνας</ENAMEX>
<ENAMEX TYPE="ORG">Εθνικό Μουσείο</ENAMEX> στην <ENAMEX TYPE="LOC">Κίνα</ENAMEX>
<ENAMEX TYPE="ORG">Τμήμα Πληροφορικής</ENAMEX> του <ENAMEX TYPE="ORG">ΟΠΑ</ENAMEX>

4.1.2.3 Τοπωνύμια

Στην κατηγορία των τοπωνυμίων (location) ανήκουν ονόματα χωρών, πόλεων, ηπείρων, ωκεανών, περιοχών, βουνών, ποταμών, λιμνών, δρόμων, αεροδρομίων, γηπέδων, κτηρίων, μνημείων, κ.α. Τα ονόματα κρατών και πόλεων, ακόμα και όταν χρησιμοποιούνται ως οργανισμοί, ανήκουν πάντα στην κατηγορία των τοπωνυμιών. Ακολουθούν παραδείγματα:

<ENAMEX TYPE="LOC">Ε.Ο. Αθηνών – Πατρών</ENAMEX>
οδός <ENAMEX TYPE="LOC">Κυδαθηναίων 10</ENAMEX>
γήπεδο <ENAMEX TYPE="LOC">«Νίκος Γκούμας»</ENAMEX>
<ENAMEX TYPE="LOC">Λάρισα</ENAMEX> – <ENAMEX TYPE="LOC">Βόλος</ENAMEX> 0 – 0

4.1.2.4 Ημερομηνίες

Στην κατηγορία των ημερομηνιών (date) ανήκουν εκφράσεις που υποδηλώνουν μέρες, μήνες, έτη, εποχές, περιόδους, γιορτές κ.λ.π. Οι εκφράσεις αυτές ενδέχεται να περιέχουν και λέξεις όπως «επόμενος», «προηγούμενος», «αρχές», «τέλος» κ.λ.π. (για τη δεύτερη, ήδη επισημειωμένη, συλλογή κειμένων αυτός ο κανόνας δεν ισχύει) Ειδικά εκφράσεις όπως «χθες», «αύριο», «σήμερα» κ.λ.π. δεν θεωρούνται ημερομηνίες. Ακολουθούν παραδείγματα:

<TIMEX TYPE="DATE">Τρίτη, 10 Φεβρουαρίου 2004</TIMEX>
<TIMEX TYPE="DATE">10/2/2004</TIMEX>
<TIMEX TYPE="DATE">Πάσχα του 2000</TIMEX>
<TIMEX TYPE="DATE">3ος αιώνας π.Χ.</TIMEX>
<TIMEX TYPE="DATE">τέλος Αυγούστου</TIMEX>
<TIMEX TYPE="DATE">επόμενο χειμώνα</TIMEX>
περίοδος <TIMEX TYPE="DATE">2000 - 2001 </TIMEX>

4.1.2.5 Εκφράσεις ώρας

Στην κατηγορία των εκφράσεων ώρας (time) ανήκουν χρονικές εκφράσεις που δηλώνουν ώρα. Εξαιρούνται σχετικές εκφράσεις, όπως «αύριο το πρωί», οι οποίες δεν περιέχουν ούτε συγκεκριμένη ώρα ούτε κάποια συγκεκριμένη ημέρα. Υπενθυμίζεται ότι οι εκφράσεις ώρας υποστηρίζονται από το σύστημα και επισημειώνονται από το χρήστη ως τύπου time, αλλά κατά τη φάση χρήσης χαρακτηρίζονται, λόγω της κοινής προσέγγισης της κατηγορίας timex, από το σύστημα ως date. Ακολουθούν παραδείγματα:

<TIMEX TYPE="TIME">3:30</TIMEX>
<TIMEX TYPE="TIME">3.30 μ.μ.</TIMEX>
<TIMEX TYPE="TIME">12 το μεσημέρι</TIMEX>
<TIMEX TYPE="TIME">πρωί της Δευτέρας</TIMEX>

4.1.2.6 Χρηματικές εκφράσεις

Ως χρηματικές εκφράσεις (money) θεωρούνται εκείνες που δηλώνουν χρηματικά ποσά και απαραιτήτως περιέχουν χρηματικές μονάδες. Ακολουθούν παραδείγματα:

<NUMEX TYPE="MONEY">2000 €</NUMEX>
<NUMEX TYPE="MONEY">1.200 δολάρια Η.Π.Α.</NUMEX>

4.1.2.7 Ποσοστά

Στην κατηγορία των ποσοστών (percent) ανήκουν εκφράσεις που εκφράζουν ποσοστά και απαραιτήτως περιέχουν το σύμβολο «%». Ακολουθούν παραδείγματα:

<NUMEX TYPE="PERCENT">3%</NUMEX>
<NUMEX TYPE="PERCENT">3.5 %</NUMEX>
<NUMEX TYPE="PERCENT">3 – 5 %</NUMEX>

4.1.2.8 Μη ονόματα οντοτήτων

Οι λεκτικές μονάδες που δεν ανήκουν στις παραπάνω κατηγορίες δεν θεωρούνται ονόματα οντοτήτων ούτε μέρη τους. Επομένως, σε αυτήν την κατηγορία τοποθετούνται και ονόματα προϊόντων, βιβλίων, βραβείων, αγώνων, ασθενειών κλπ. Επιπλέον, δεν επισημειώνονται φωλιασμένες (nested) εκφράσεις. Για παράδειγμα, η επισημείωση της φράσης «το <TIMEX TYPE="TIME">πρωί της <TIMEX TYPE="DATE">Δευτέρας</TIMEX></TIMEX>» δεν υποστηρίζεται από το σύστημα. Τέλος, δεν θεωρούνται ονόματα οντοτήτων (ούτε μέρη τους) τίτλοι προσώπων, όπως «Πρωθυπουργός», «Δρ», «Πατριάρχης», «Πάπας», «Σερ», ακόμα και αν αρχίζουν με κεφαλαίο γράμμα.

4.2 Αποτελέσματα πειραμάτων

Ως μέτρα αξιολόγησης της διαδικασίας της αναγνώρισης ονομάτων προσώπων και χρονικών εκφράσεων χρησιμοποιήθηκαν τα εξής:

$$\text{ακρίβεια (precision)} = \frac{\text{λεκτικές μονάδες που κατετάγησαν ορθά στην κατηγορία}}{\text{λεκτικές μονάδες που κατετάγησαν στην κατηγορία}}$$

$$\text{ανάκληση (recall)} = \frac{\text{λεκτικές μονάδες που κατετάγησαν ορθά στην κατηγορία}}{\text{σύνολο λεκτικών μονάδων της κατηγορίας}}$$

$$\text{F-measure} = \frac{(\beta^2 + 1) \cdot \text{ακρίβεια} \cdot \text{ανάκληση}}{\beta^2 \cdot \text{ακρίβεια} + \text{ανάκληση}}$$

Από τους παραπάνω ορισμούς φαίνεται ότι η ακρίβεια και η ανάκληση υπολογίζονται ξεχωριστά για κάθε κατηγορία. Στο F-measure, η παράμετρος β χρησιμοποιείται για να δώσει μεγαλύτερο βάρος στην ακρίβεια σε σχέση με την ανάκληση ή αντίστροφα. Στα αποτελέσματα που θα ακολουθήσουν χρησιμοποιήσαμε $\beta = 1$, ούτως ώστε να δώσουμε ίσο βάρος στην ακρίβεια και την ανάκληση.

Τα διαγράμματα που θα παρουσιαστούν περιλαμβάνουν και τα διαστήματα εμπιστοσύνης των μετρήσεων, με επίπεδο εμπιστοσύνης 0.99. Τα διαστήματα εμπιστοσύνης, λόγω του μεγάλου πλήθους των περιπτώσεων ελέγχου (του δείγματος), υπολογίζονται με βάση τον τύπο:

$$p \pm z_{a/2}\sqrt{\frac{p \cdot (1-p)}{n}}$$

όπου n το πλήθος των περιπτώσεων ελέγχου[6], a το περιθώριο (πιθανότητα) σφάλματος (χρησιμοποιούμε $a = 0{,}01$), $z_{a/2}$ η τιμή της τυποποιημένης κανονικής μεταβλητής Z με περιθώριο σφάλματος a[7] (προκύπτει από τους πίνακες της κανονικής κατανομής, χρησιμοποιούμε $z_{0{,}005} = 2{,}576$) και p το ποσοστό της ακρίβειας, της ανάκλησης ή του F-measure.

Κατά τον υπολογισμό των αποτελεσμάτων (ακρίβεια, ανάκληση, F-measure) της κατηγορίας των ονομάτων προσώπων, θεωρούμε ότι οι «ασφαλείς» κανόνες (συμπεριλαμβανομένης και της αναγνώρισης χρονικών εκφράσεων) δεν κάνουν ποτέ λάθος, δηλαδή όποτε κατατάσσουν μια λεκτική μονάδα ως όνομα προσώπου, αυτή είναι πράγματι όνομα προσώπου. Αυτή η παραδοχή δεν απέχει πολύ από την πραγματικότητα, αφού οι «ασφαλείς» κανόνες χαρακτηρίζουν λανθασμένα ως μη ονόματα προσώπων μόνο περίπου το 0,2% των λεκτικών μονάδων που αποτελούν ονόματα προσώπων.

Ως βάση σύγκρισης (baseline) για τα αποτελέσματα στις κατηγορίες των ονομάτων προσώπων και των χρονικών εκφράσεων θεωρούνται τα παρακάτω ποσοστά.

	Ακρίβεια	**Ανάκληση**	**F-measure**
PERSON	91,14%	26,65%	41,25%
TIMEX	1,20%	0,61%	0,81%

Baseline για τις κατηγορίες ονομάτων προσώπων και χρονικών εκφράσεων

Για την κατηγορία των ονομάτων προσώπων, τα παραπάνω ποσοστά έχουν προκύψει χρησιμοποιώντας τους εξής δύο κανόνες, οι οποίοι εμφανίζονται αρκετά συχνά στη βιβλιογραφία:

1. κ. [(X(X|x)* |.)*]
2. [name_in_list (X(X|x)*)*]

όπου X και x οποιοσδήποτε (ελληνικός ή λατινικός) κεφαλαίος ή μικρός χαρακτήρας αντίστοιχα, name_in_list μία λεκτική μονάδα που περιέχεται στη λίστα με 350 ελληνικά βαφτιστικά ονόματα προσώπων που διαθέτουμε (ενότητα 3.4.2), το | σημαίνει ή, το * σημαίνει ότι η προηγούμενη έκφραση επαναλαμβάνεται μηδέν ή περισσότερες φορές και οι παρενθέσεις [] σηματοδοτούν τα όρια του ονόματος προσώπου. Ο πρώτος κανόνας εντοπίζει εκφράσεις όπως *«κ. [Κώστας Σημίτης]»*, *«κ. [Κ. Σημίτης]»*, *«κ. [Ιωάννης Π. Παπαγεωργίου]»* κ.α. Ο δεύτερος κανόνας εντοπίζει εκφράσεις όπως *«[Κώστας Σημίτης]»*, *«[Άννα Ψαρούδα Μπενάκη]»* κ.α., όπου υπογραμμισμένες είναι οι λεκτικές μονάδες που περιέχονται στη λίστα με τα ονόματα.

Για την κατηγορία των χρονικών εκφράσεων, τα ποσοστά έχουν προκύψει κατατάσσοντας με τυχαίο τρόπο κάθε λεκτική μονάδα ως χρονική έκφραση ή όχι, με πιθανότητες 1% και 99% αντίστοιχα (η αναλογία των λεκτικών μονάδων που αποτελούν χρονικές εκφράσεις σε σχέση με αυτές που δεν αποτελούν χρονικές εκφράσεις στα κείμενα εκπαίδευσης είναι 1:99).

[6] Κατά τον υπολογισμό του n δεν συμπεριλαμβάνονται οι λεκτικές μονάδες που κατετάγησαν ως μη ονόματα προσώπων από τους ασφαλείς κανόνες.

[7] Το σφάλμα a μοιράζεται στα δύο άκρα της κανονικής κατανομής, δηλαδή $a/2$ στο αριστερό άκρο και $a/2$ στο δεξί άκρο. Ισχύει ότι $z_{1-a/2} = -z_{a/2}$.

4.2.1 Πείραμα 1^{ο}: Εκπαίδευση και έλεγχος στην 1^{η} συλλογή

Όσον αφορά την κατηγορία των χρονικών εκφράσεων, ακολουθούν αποτελέσματα που έχουν προκύψει με 10-πλή διασταυρωμένη επικύρωση στα 400 επισημειωμένα κείμενα της πρώτης συλλογής. Κατά την αυτόματη δημιουργία των προτύπων δεν εφαρμόστηκαν τα στάδια της αποκοπής των προτύπων με μικρό αριθμό εμφάνισης και της προσθήκης επιπλέον προτύπων. Ακολουθεί ο πίνακας με τα αποτελέσματα, ο οποίος περιλαμβάνει και τις μέγιστες και ελάχιστες τιμές που σημειώθηκαν στις 10 επαναλήψεις της διασταυρωμένης επικύρωσης.

		min	max
Ακρίβεια	96,62%	91,64%	99,06%
Ανάκληση	92,95%	78,42%	98,43%
F-measure	94,75%	85,54%	98,43%

Αποτελέσματα κατηγορίας timex στα 400 κείμενα με 10-πλή διασταυρωμένη επικύρωση

Μια διερεύνηση των εκφράσεων που το σύστημα χαρακτήρισε λανθασμένα ως χρονικές (false positives) έδειξε ότι πολλά λάθη (περίπου 50%) αφορούν αριθμητικά δεδομένα (βλ. παρακάτω πίνακα). Πιο συγκεκριμένα, το 14% των λαθών αφορά 4-ψήφιους αριθμούς, οι οποίοι κατετάγησαν ως χρονολογίες. Επίσης, το 35% αφορά αριθμητικές εκφράσεις, όπως για παράδειγμα «3.14». Υπήρξαν ακόμα λεκτικές μονάδες ονομάτων οργανισμών («17 Νοέμβρη», «Οργανωτική Επιτροπή Αθήνα 2004») ή τοποθεσιών («Αγία Παρασκευή») που κατετάγησαν λανθασμένα ως χρονικές εκφράσεις. Στα 400 κείμενα που χρησιμοποιήθηκαν δεν υπήρξαν περιπτώσεις λεκτικών μονάδων ονομάτων προσώπων που να κατετάγησαν ως χρονικές εκφράσεις. Παρ' όλα αυτά είναι φανερό ότι μπορεί να υπάρξουν και τέτοια λάθη. Ένα παράδειγμα είναι τα κύρια ονόματα Παρασκευή, Κυριακή ή Ιούλιος, Αύγουστος, τα οποία αποτελούν και ημέρες της εβδομάδας ή μήνες.

Κατηγορία	4-ψήφιοι	οργανισμοί	τοποθεσίες	αριθμοί	λοιπά
Ποσοστό	14%	27%	6%	35%	8%

Ανάλυση εκφράσεων που χαρακτηρίστηκαν λανθασμένα ως χρονικές

Από την άλλη πλευρά, από τη διερεύνηση των χρονικών εκφράσεων που το σύστημα δεν κατάφερε να εντοπίσει (false negatives) προκύπτει ότι η μη επιτυχημένη ταξινόμησή τους οφείλεται στην έλλειψη σπάνιων και συνήθως μεγάλου μήκους (σε λεκτικές μονάδες) προτύπων, όπως για παράδειγμα τα πρότυπα που αντιστοιχούν στις εκφράσεις «19ος – 20ος αιώνας» (εντοπίζεται μόνο η φράση «20ος αιώνας»), «9 έως τις 2 το μεσημέρι» (εντοπίζεται μόνο η φράση «2 το μεσημέρι»).

Όσον αφορά την κατηγορία ονομάτων προσώπων, σε αυτό το πείραμα χρησιμοποιήθηκαν τόσο ως δεδομένα εκπαίδευσης όσο και ως δεδομένα ελέγχου τα κείμενα από τις εφημερίδες «ΤΑ ΝΕΑ» και «ΤΟ ΒΗΜΑ». Συγκεκριμένα, επιλέχθηκαν τυχαία τα μισά (200) από τα αρχικά επισημειωμένα κείμενα (ενότητα 4.1.2) ως δεδομένα ελέγχου. Τα υπόλοιπα κείμενα χρησιμοποιήθηκαν για την εκπαίδευση του συστήματος, όπως θα εξηγηθεί αναλυτικότερα παρακάτω. Συγκεκριμένα, ο αριθμός των διανυσμάτων στα κείμενα εκπαίδευσης που είναι ορατά στη ΜΔΥ (δεν έχουν χαρακτηριστεί από τους «ασφαλείς» κανόνες ως μη ονόματα προσώπων) είναι περίπου 13.000.

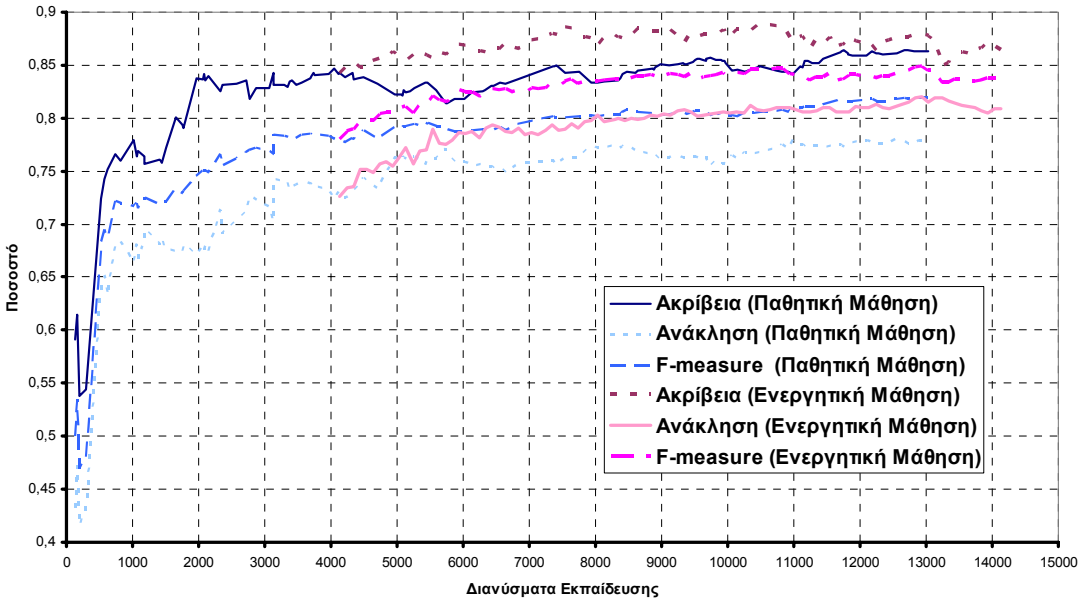
Υπενθυμίζεται ότι οι «ασφαλείς» κανόνες, που χρησιμοποιούνται κατά την αναγνώριση ονομάτων προσώπων, περιλαμβάνουν και τα πρότυπα που εντοπίζουν χρονικές εκφράσεις. Για την αξιολόγηση του συστήματος στην κατηγορία των ονομάτων προσώπων, εφαρμόστηκε αρχικά στα κείμενα εκπαίδευσης η διαδικασία της αυτόματης εξαγωγής προτύπων για χρονικές εκφράσεις, χωρίς να εφαρμοστούν τα στάδια της αποκοπής και της προσθήκης προτύπων. Προέκυψαν 99 πρότυπα. Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα που προκύπτουν με την εφαρμογή των προτύπων αυτών στα δεδομένα ελέγχου.

	Ακρίβεια	**Ανάκληση**	**F-measure**
TIMEX	98,03%	93,93%	95,93%

Αποτελέσματα χρονικών εκφράσεων στο 1ο πείραμα με 200 κείμενα εκπαίδευσης και 200 κείμενα ελέγχου

Στην γραφική παράσταση που ακολουθεί παρουσιάζονται τα αποτελέσματα για την κατηγορία των ονομάτων προσώπων. Συγκεκριμένα, φαίνονται αρχικά οι καμπύλες μάθησης για τα τρία μέτρα αξιολόγησης (ακρίβεια, ανάκληση και F-measure), όταν χρησιμοποιείται μόνο το πρώτο πέρασμα με παθητική μάθηση. Ο οριζόντιος άξονας αντιστοιχεί στον αριθμό των παραδειγμάτων εκπαίδευσης (δε συνυπολογίζονται στον οριζόντιο άξονα τα παραδείγματα εκπαίδευσης που έχουν αποκλειστεί από τη ΜΔΥ με τη χρήση των «ασφαλών» κανόνων). Στην παθητική μάθηση, σε κάθε σημείο του οριζόντιου άξονα (π.χ. 7000) χρησιμοποιούνται μόνο τα αντίστοιχα πρώτα παραδείγματα των επισημειωμένων κειμένων εκπαίδευσης (π.χ. τα διανύσματα των πρώτων 7000 λεκτικών μονάδων), τα οποία δεν έχουν καταταγεί ως μη ονόματα προσώπων από τους «ασφαλείς» κανόνες και την αναγνώριση χρονικών εκφράσεων. Πρόκειται για μια προσομοίωση της διαδικασίας της παθητικής μάθησης. Θεωρούμε δηλαδή ότι η επισημείωση των κειμένων εκπαίδευσης θα γινόταν εξαντλητικά, ξεκινώντας από το πρώτο κείμενο εκπαίδευσης, και ότι θα είχε σταματήσει μετά την επισημείωση του αντίστοιχου αριθμού των παραδειγμάτων εκπαίδευσης.

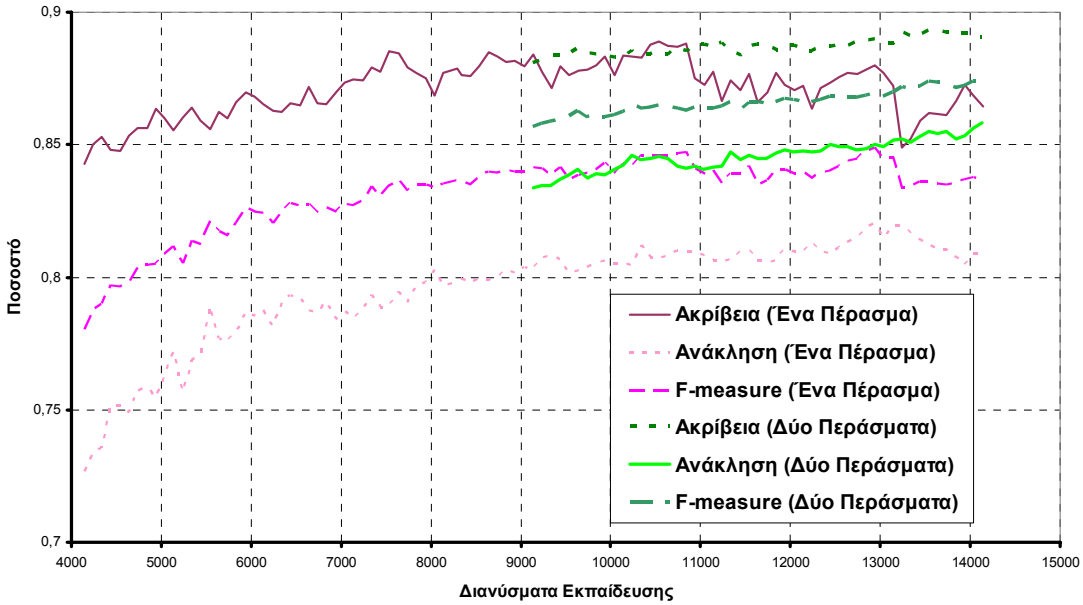
Μετά τα πρώτα 4000 περίπου παραδείγματα εκπαίδευσης, δοκιμάσαμε παράλληλα τη χρήση της ενεργητικής μάθησης (καμπύλες με τη παχιά γραμμή). Στην περίπτωση της ενεργητικής μάθησης, τα νέα διανύσματα εκπαίδευσης δεν επιλέγονται με τη σειρά από τα ήδη επισημειωμένα κείμενα εκπαίδευσης αλλά τα επιλέγει το ίδιο το σύστημα από όλα τα διαθέσιμα κείμενα των εφημερίδων (που αποτελούν τη «δεξαμενή κειμένων» της ενότητας 3.4.4). Παρατηρούμε ότι υπάρχει μία πιο έντονη άνοδος του F-measure στην καμπύλη της ενεργητικής μάθησης σε σχέση με την παθητική μάθηση. Η διαφορά τους φθάνει περίπου στο 4,5%, κάτι που δείχνει ότι με την ενεργητική μάθηση μπορούμε να επιτύχουμε καλύτερα αποτελέσματα από ό,τι με την παθητική όταν χρησιμοποιείται ο ίδιος αριθμός παραδειγμάτων εκπαίδευσης. Αντίστοιχη διαφορά παρατηρείται στην περίπτωση της ανάκλησης και της ακρίβειας, αν και η διαφορά στην περίπτωση της ακρίβειας μειώνεται για μεγάλο αριθμό παραδειγμάτων εκπαίδευσης.



Παθητική και ενεργητική μάθηση για τον εντοπισμό ονομάτων προσώπων στο 1ο πείραμα

Στο 23% των δεδομένων εκπαίδευσης (3100 διανύσματα), πριν αρχίσει η ενεργητική μάθηση, έγινε ρύθμιση των παραμέτρων της ΜΔΥ (βλ. παράρτημα). Οι τιμές των παραμέτρων που προέκυψαν ($C = 2^{1,25}$ και $\gamma = 2^{-4,25}$) χρησιμοποιήθηκαν και στα δύο είδη μάθησης από αυτό το σημείο και μετά.

Το επόμενο διάγραμμα δείχνει τη βελτίωση που προκύπτει χρησιμοποιώντας δύο περάσματα (ενότητα 3.4.3) αντί για μόνο ένα. Με τον όρο «Ένα Πέρασμα» στο διάγραμμα εννοείται ότι χρησιμοποιήθηκε μόνο ο ταξινομητής του πρώτου περάσματος (μωβ καμπύλες, συμπίπτουν με τις μωβ καμπύλες του παραπάνω διαγράμματος). Με τον όρο «Δύο Περάσματα» στο διάγραμμα εννοείται ότι το σύστημα χρησιμοποιεί και τους δύο ταξινομητές που περιγράφηκαν στις ενότητες 3.4.2 και 3.4.3. Εδώ και στις δύο περιπτώσεις έχει χρησιμοποιηθεί ενεργητική μάθηση. Η βελτίωση φτάνει για το F-measure μέχρι και 4,5%, ενώ παρατηρούμε ότι για την ακρίβεια αρχικά δεν υπάρχει μεγάλη διαφορά. Όπως φαίνεται από τα αποτελέσματα, η μεγαλύτερη επίδραση του δεύτερου περάσματος είναι στην ανάκληση, όπου η βελτίωση ξεπερνάει το 5%. Είναι, επίσης, αξιοσημείωτο ότι οι επιδόσεις του συστήματος με δύο περάσματα παρουσιάζουν σταθερή βελτίωση όσο αυξάνεται ο αριθμός των παραδειγμάτων εκπαίδευσης, ενώ στην περίπτωση του ενός περάσματος οι επιδόσεις δείχνουν τάσης αστάθειας ή χειροτέρευσης μετά τα περίπου 11,000 διανύσματα εκπαίδευσης.



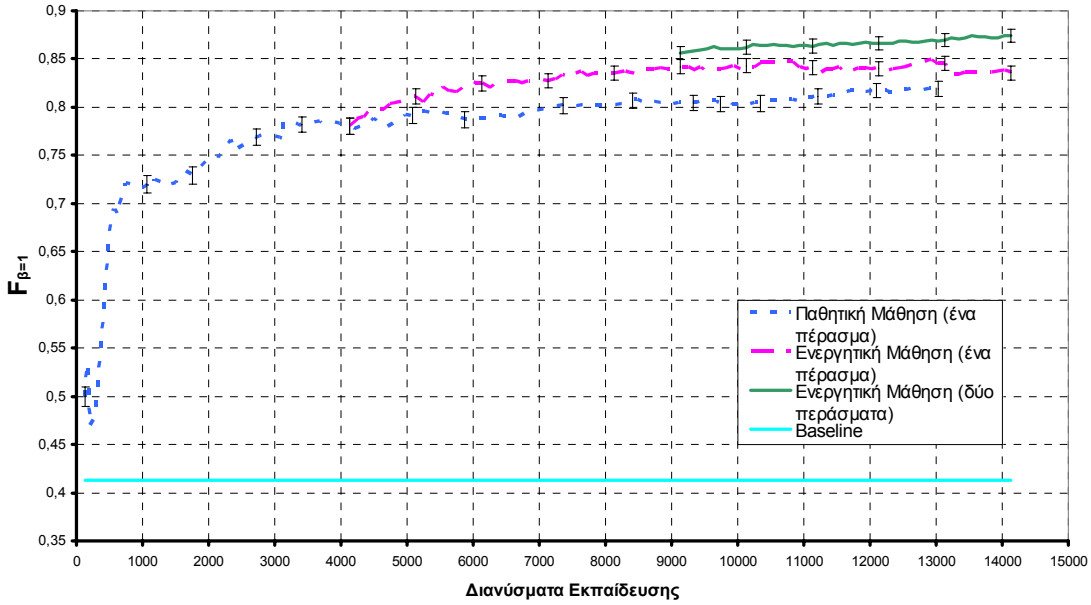
Η επίδραση του 2ου περάσματος στο 1ο πείραμα, κατά την αναγνώριση ονομάτων προσώπων, όταν χρησιμοποιείται ενεργητική μάθηση

Κατά την έναρξη της χρήσης του δεύτερου περάσματος (αριστερότερα σημεία των πράσινων καμπυλών, 9100 διανύσματα εκπαίδευσης) έγιναν τα ακόλουθα. Τα 9100 διανύσματα που χρησιμοποιήθηκαν σε αυτό το σημείο για την εκπαίδευση της ΜΔΥ του πρώτου περάσματος εμπλουτίσθηκαν με τις επιπλέον ιδιότητες του δεύτερου περάσματος (ενότητα 3.4.3). (Κατά τη διάρκεια των πειραμάτων με το πρώτο πέρασμα ως τα 9100 διανύσματα, αποθηκεύονταν πληροφορίες για κάθε διάνυσμα που έδειχναν τη θέση της αντίστοιχης λεκτικής μονάδας μέσα στα κείμενα εκπαίδευσης, οπότε ο εμπλουτισμός των διανυσμάτων ήταν εφικτός.) Από τα 9100 εμπλουτισμένα διανύσματα επιλέχθηκαν τυχαία 3100 διανύσματα, με τη χρήση των οποίων έγινε η ρύθμιση των παραμέτρων της ΜΔΥ του δεύτερου περάσματος (βλ. παράρτημα· η ρύθμιση των παραμέτρων της ΜΔΥ του πρώτου περάσματος είχε γίνει επίσης χρησιμοποιώντας 3100 διανύσματα, όπως αναφέρθηκε παραπάνω). Οι τιμές των παραμέτρων που προέκυψαν είναι $C = 2^{3,25}$ και $\gamma = 2^{-5,5}$. Τέλος, η ΜΔΥ του δεύτερου περάσματος εκπαιδεύτηκε με τα 9100 εμπλουτισμένα διανύσματα και τις τιμές των παραμέτρων που προέκυψαν.

Στη συνέχεια, άρχισε η διαδικασία της ενεργητικής μάθησης για τον ταξινομητή του δεύτερου περάσματος. Αρχικά, οι δύο ταξινομητές (πρώτου και δεύτερου περάσματος) ήταν εκπαιδευμένοι σε 9100 διανύσματα. Με τη βοήθειά τους επιλέχθηκαν από τη δεξαμενή κειμένων τα 100 πιο κοντινά διανύσματα στο υπερεπίπεδο της δεύτερης ΜΔΥ. Τα 100 διανύσματα επισημειώθηκαν και επανεκπαιδεύτηκε μόνο η δεύτερη ΜΔΥ με 9200 διανύσματα κ.ο.κ. Η ΜΔΥ του πρώτου περάσματος, η οποία εκπαιδεύτηκε σε 9100 διανύσματα, διατηρήθηκε απαράλλακτη (χωρίς επανεκπαίδευση) σε όλη τη διάρκεια του πειράματος για το δεύτερο πέρασμα. Η επιλογή αυτή έγινε επειδή αν κατά τη διάρκεια του πειράματος για το δεύτερο πέρασμα η ΜΔΥ του πρώτου περάσματος μεταβαλλόταν, τότε σε κάθε 100άδα νέων παραδειγμάτων εκπαίδευσης οι τιμές των 6 επιπλέον ιδιοτήτων του δεύτερου περάσματος (ενότητα 3.4.3) θα προέκυπταν από διαφορετική ΜΔΥ πρώτου περάσματος. Για παράδειγμα, έστω ότι επανεκπαιδευόταν η ΜΔΥ του πρώτου περάσματος με 9200 διανύσματα και στη συνέχεια επανεκπαιδευόταν η ΜΔΥ του δεύτερου περάσματος με 9200 διανύσματα. Στα 9100 πρώτα διανύσματα που

χρησιμοποιήθηκαν για την εκπαίδευση της δεύτερης ΜΔΥ, οι ιδιότητες του δεύτερου περάσματος (που αφορούν αποστάσεις από το υπερεπίπεδο της ΜΔΥ του πρώτου περάσματος) παίρνουν τιμές από μία ΜΔΥ πρώτου περάσματος που έχει εκπαιδευτεί σε 9100 διανύσματα, ενώ στα υπόλοιπα 100 διανύσματα οι ιδιότητες αυτές παίρνουν τιμές από μία ΜΔΥ πρώτου περάσματος που έχει εκπαιδευτεί σε 9200 διανύσματα (δηλαδή παίρνουν τιμές σε σχέση με διαφορετικό υπερεπίπεδο).

Μία άλλη πιθανή λύση αυτού του προβλήματος, εκτός από το να διατηρείται η ΜΔΥ του πρώτου περάσματος σταθερή, είναι να επαναταξινομούνται από τη ΜΔΥ του πρώτου περάσματος όλα τα υπάρχοντα παραδείγματα εκπαίδευσης μετά από κάθε επανεκπαίδευσή της, ώστε να ενημερώνονται οι τιμές των ιδιοτήτων που αφορούν απόσταση από το υπερεπίπεδο διαχωρισμού της πρώτης ΜΔΥ. Αυτό, όμως, είναι χρονοβόρο και ανατρέπει τη λογική του χωρισμού της δεξαμενής κειμένων σε 10 μέρη, καθώς για να ενημερωθούν οι τιμές παλαιών διανυσμάτων εκπαίδευσης πρέπει να επανα-επεξεργαστούμε και κείμενα από άλλα μέρη της δεξαμενής (από τα οποία προέρχονταν τα παλαιά διανύσματα εκπαίδευσης), πέρα από το τρέχον μέρος της δεξαμενής.



Παθητική μάθηση, ενεργητική μάθηση και ενεργητική μάθηση με δύο περάσματα κατά την αναγνώριση ονομάτων προσώπων στο 1^{ο} πείραμα

Στο προηγούμενο διάγραμμα φαίνονται τα συγκριτικά αποτελέσματα του F-measure για τις τρεις περιπτώσεις (παθητική μάθηση με ένα πέρασμα, ενεργητική μάθηση με ένα πέρασμα και ενεργητική μάθηση με δύο περάσματα), καθώς και τα διαστήματα εμπιστοσύνης για κάθε περίπτωση. Ως πλήθος περιπτώσεων ελέγχου κατά τον υπολογισμό των διαστημάτων εμπιστοσύνης χρησιμοποιήθηκαν μόνο τα διανύσματα που είναι ορατά στις ΜΔΥ, δηλαδή εξαιρούνται διανύσματα που αντιστοιχούν σε λεκτικές μονάδες τις οποίες οι «ασφαλείς» κανόνες κατέταξαν ως μη ονόματα προσώπων. Το πλήθος των διανυσμάτων αυτών είναι περίπου 16.000.

Στηριζόμενοι στα διαστήματα εμπιστοσύνης και παρατηρώντας ότι σε καμία περίπτωση δεν υπάρχει επικάλυψη, συμπεραίνουμε ότι τόσο η ενεργητική μάθηση όσο και το δεύτερο πέρασμα έχουν θετικά αποτελέσματα στην επίδοση του συστήματος. Τα τελικά αποτελέσματα για την κατηγορία ονομάτων προσώπων και στις τρεις περιπτώσεις φαίνονται καθαρότερα στον επόμενο πίνακα. Όσον αφορά τις

δύο πρώτες περιπτώσεις, παθητική μάθηση με ένα πέρασμα και ενεργητική μάθηση με ένα πέρασμα, τα αποτελέσματα που παρατίθενται προκύπτουν από ΜΔΥ που έχουν εκπαιδευτεί με 13.000 και 14.000 διανύσματα αντίστοιχα. Στην περίπτωση της ενεργητικής μάθησης με δύο περάσματα, η ΜΔΥ του πρώτου περάσματος έχει εκπαιδευτεί με 9.000 διανύσματα, ενώ η ΜΔΥ του δεύτερου περάσματος με 14.000 διανύσματα.

	Ακρίβεια (%)	**Ανάκληση (%)**	**F-measure (%)**
Παθητική Μάθηση	86,29	77,91	81,89
Ενεργητική Μάθηση	86,43	80,85	83,56
Δύο Περάσματα	89,06	85,83	87,42

Αποτελέσματα για την κατηγορία των ονομάτων προσώπων στο 1ο πείραμα

Ο επόμενος πίνακας δείχνει πόσες από τις λεκτικές μονάδες που κατετάγησαν λανθασμένα ως ονόματα προσώπων (false positives) ανήκαν στις άλλες κατηγορίες, στην περίπτωση όπου χρησιμοποιούνται δύο περάσματα (9.000 και 14.000 παραδείγματα εκπαίδευσης για το πρώτο και δεύτερο πέρασμα αντίστοιχα).

Κατηγορία Λαθών	
Σύμβολα	5,5% (23)
Λέξεις με κεφαλαία	5,5% (22)
Τίτλοι προσώπων - Εθνικότητες	8% (33)
Αγγλικές λέξεις	4% (15)
Οργανισμοί	32% (129)
Τοπωνύμια	31% (124)
Λοιπά	14% (55)

Ανάλυση λεκτικών μονάδων που χαρακτηρίσθηκαν λανθασμένα ως ονόματα προσώπων στο 1ο πείραμα, όταν χρησιμοποιούνται δύο περάσματα

Όπως φαίνεται από τον πίνακα δημιουργείται σοβαρό πρόβλημα ανάμιξης μεταξύ των τριών κατηγοριών (ονόματα προσώπων, ονόματα οργανισμών, τοπωνύμια) enamex. Κατετάγησαν λανθασμένα ως ονόματα προσώπων περίπου 250 λεκτικές μονάδες που ανήκαν στην πραγματικότητα στις κατηγορίες των ονομάτων οργανισμών και τοπωνυμίων (αποτελούν το 63% των λεκτικών μονάδων που κατετάγησαν λανθασμένα ως ονόματα προσώπων).

Μία άλλη ενδιαφέρουσα κατηγορία λαθών είναι η «Τίτλοι προσώπων - Εθνικότητες». Ως τίτλοι θεωρούνται λέξεις όπως «Πατριάρχης», «Πρωθυπουργός», «Δρ», ενώ ως εθνικότητες λέξεις όπως «Ιταλός», «Παλαιστίνιος», κ.λ.π. Οι λέξεις αυτές χρησιμοποιούνται συχνά ως αναφορές σε ονόματα προσώπων, με αποτέλεσμα ο ταξινομητής να θεωρεί λανθασμένα ότι αποτελούν όντως ονόματα προσώπων. Ειδικότερα όσον αφορά τους τίτλους προσώπων, κάποιες προσεγγίσεις επισημείωσης στη βιβλιογραφία θεωρούν ότι συμμετέχουν στα ονόματα προσώπων, ενώ κάποια συστήματα δημιουργούν ειδική κατηγορία για τους τίτλους προσώπων.

Το 5,5% των λαθών αφορά λέξεις που δεν είναι ονόματα προσώπων παρόλο που αρχίζουν με κεφαλαίο γράμμα. Οι λέξεις αυτές μπορεί να αποτελούν αρχή προτάσεων. Επίσης, στον πίνακα εμφανίζεται ένα μικρό ποσοστό (2%) που δεν αντιστοιχεί σε πραγματικά λάθη του συστήματος, αλλά σε περιπτώσεις όπου οι χειρωνακτικά επισημειωμένες κατηγορίες των κειμένων ελέγχου ήταν λανθασμένες.

Ο επόμενος πίνακας δείχνει τι είδους λεκτικές μονάδες που αποτελούν ονόματα προσώπων το σύστημα δεν κατάφερε να εντοπίσει (false negatives).

Κατηγορία Λαθών	
Σύμβολα	5% (27)
Ονόματα σε αθλητικά κείμενα	34% (184)
Αγγλικές λέξεις	15% (81)
Λοιπά ονόματα	46% (248)

Ανάλυση λεκτικών μονάδων ονομάτων προσώπων που το σύστημα δεν κατάφερε να εντοπίσει στο 1ο πείραμα, όταν χρησιμοποιούνται δύο περάσματα

Το ένα τρίτο των ονομάτων προσώπων που δεν εντοπίστηκαν εμφανίζονται σε αθλητικά κείμενα και ιδιαίτερα σε παραγράφους αθλητικών κειμένων που αναφέρουν ενδεκάδες ομάδων. Σε αυτές τις περιπτώσεις, όπου απαριθμούνται ονόματα προσώπων (ποδοσφαιριστών) χωρισμένα με κόμματα, οι ιδιότητες που παίρνουν τιμές χρησιμοποιώντας τα συμφραζόμενα της υπό κατάταξης λεκτικής μονάδας (περιγράφονται αναλυτικά στην ενότητα 3.4.2) δεν μπορούν να βοηθήσουν σχεδόν καθόλου τον σύστημα. Πολλές ιδιότητες, δηλαδή, δε λειτουργούν καλά λόγω έλλειψης κατάλληλων συμφραζομένων.

Επίσης, το 15% περίπου των λαθών αφορά αγγλικές λέξεις. Αυτό οφείλεται στον προσανατολισμό των ιδιοτήτων, οι οποίες αφορούν ελληνικά κείμενα· αντίστοιχο πρόβλημα υπήρχε και στον προηγούμενο πίνακα. Τέλος, το μεγαλύτερο ποσοστό (46%) των περιπτώσεων ονομάτων προσώπων που δεν κατάφερε να εντοπίσει το σύστημα αφορά ονόματα διαφόρων ειδών, τα οποία δεν είναι δυνατόν να ομαδοποιηθούν.

4.2.2 Πείραμα 2ο: Εκπαίδευση και έλεγχος σε διαφορετικές συλλογές

Στο δεύτερο πείραμα χρησιμοποιήθηκε το τελικό σύστημα με τα δύο περάσματα (9.000 και 14.000 διανύσματα εκπαίδευσης για τις ΜΔΥ του πρώτου και δεύτερου περάσματος αντίστοιχα) που προέκυψε από την εκπαίδευση στην πρώτη συλλογή κειμένων και αξιολογήθηκε το κατά πόσον μπορεί να εντοπίσει σωστά χρονικές εκφράσεις και ονόματα προσώπων στη δεύτερη συλλογή κειμένων. Χρησιμοποιήθηκαν, δηλαδή, δύο διαφορετικές συλλογές κειμένων για εκπαίδευση και έλεγχο. Τα αποτελέσματα που προκύπτουν φαίνονται στον παρακάτω πίνακα.

		Επίδοση	Διάστημα Εμπιστοσύνης
Πρώτο Πέρασμα	*Ακρίβεια*	74,95%	±0,83%
	Ανάκληση	88,70%	±0,61%
	F-measure	81,25%	±0,75%
Δεύτερο Πέρασμα	*Ακρίβεια*	77,95%	±0,8%
	Ανάκληση	87,21%	±0,64%
	F-measure	82,29%	±0,73%

2ο Πείραμα: Αποτελέσματα αναγνώρισης ονομάτων προσώπων

	Ακρίβεια	Ανάκληση	F-measure
TIMEX	84,21%	89,03%	86,55%

2ο Πείραμα: Αποτελέσματα αναγνώρισης χρονικών εκφράσεων

Παρατηρούμε σημαντική πτώση της απόδοσης του συστήματος, σε σχέση με τα αποτελέσματα της προηγούμενης ενότητας, της τάξης του 5% για την κατηγορία

ονομάτων προσώπων και 10% για τις χρονικές εκφράσεις. Η πτώση αυτή οφείλεται στις διαφορές που υπάρχουν μεταξύ των κειμένων των δύο συλλογών. Για παράδειγμα, στα κείμενα της δεύτερης συλλογής υπάρχουν ονόματα προσώπων γραμμένα στη μορφή «Κων. Σημίτης», ενώ στα κείμενα της πρώτης συλλογής τα ονόματα αυτά γράφονται μόνο ως «Κ. Σημίτης». Επίσης, η συλλογή που χρησιμοποιήθηκε για την εκπαίδευση περιέχει κείμενα ποικίλης ύλης, σε αντίθεση με τη δεύτερη συλλογή που περιέχει μόνο κείμενα οικονομικής φύσης. Αυτό έχει ως αποτέλεσμα πολλές ιδιότητες που ήταν χρήσιμες σε μη οικονομικά κείμενα στην πρώτη συλλογή να είναι άχρηστες στη δεύτερη συλλογή. Επίσης, η κατανομή των λεκτικών μονάδων ανά κατηγορία είναι πολύ διαφορετική μεταξύ των δύο συλλογών.

4.2.3 Πείραμα 3ο: Εκπαίδευση και έλεγχος στη 2η συλλογή

Σε αυτό το πείραμα τόσο η εκπαίδευση όσο και ο έλεγχος έγιναν στα κείμενα της δεύτερης συλλογής. Σημειώνεται ότι τα κείμενα αυτά δεν έχουν βοηθητικές ετικέτες για αλλαγή παραγράφου και για τίτλους, με αποτέλεσμα κάποιες ιδιότητες να μη «λειτουργούν» σωστά.

Τα πειραματικά αποτελέσματα που ακολουθούν για τις κατηγορίες ονομάτων προσώπων και χρονικών εκφράσεων έχουν προκύψει με τη διαδικασία του 10-πλής διασταυρωμένης επικύρωσης επί του συνόλου των κειμένων της δεύτερης συλλογής.

		Επίδοση	Διάστημα Εμπιστοσύνης
Πρώτο Πέρασμα	*Ακρίβεια*	94,96%	±1,28%
	Ανάκληση	88,95%	±1,83%
	F-measure	91,86%	±1,6%
Δεύτερο Πέρασμα	*Ακρίβεια*	95,76%	±1,18%
	Ανάκληση	91,05%	±1,67%
	F-measure	93,34%	±1,46%

3ο Πείραμα: Ονόματα Προσώπων

	Ακρίβεια	Ανάκληση	F-measure
TIMEX	97,59%	95,35%	96,46%

3ο Πείραμα: Χρονικές Εκφράσεις

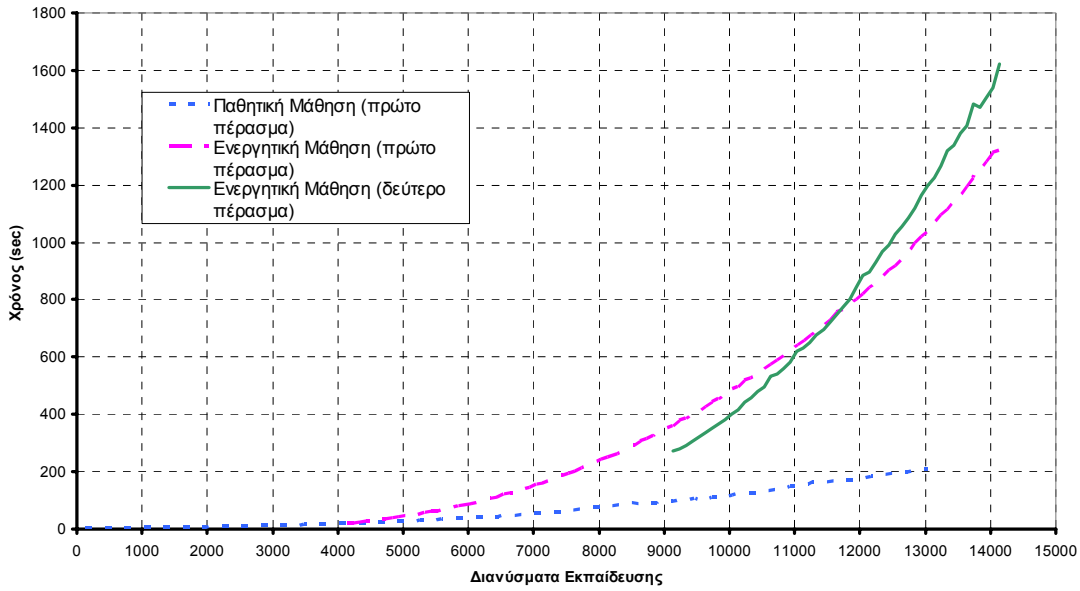
Όπως παρατηρούμε, τα αποτελέσματα είναι ιδιαίτερα βελτιωμένα σε σχέση με το προηγούμενο πείραμα, γεγονός αναμενόμενο, αλλά παρουσιάζουν βελτίωση, κυρίως στα ονόματα προσώπων, και σε σχέση με το πρώτο πείραμα. Το τελευταίο το αποδίδουμε στην περιορισμένη θεματολογία των κειμένων της δεύτερης συλλογής και στον πιο τυποποιημένο τρόπο γραφής τους.

4.2.4 Ταχύτητα του συστήματος

Στη διάρκεια του πρώτου πειράματος μετρήθηκε, επίσης, η ταχύτητα του συστήματος κατά την εκπαίδευση και τον εντοπισμό χρονικών εκφράσεων και ονομάτων οντοτήτων σε νέα κείμενα. Όλα τα πειράματα διεξήχθησαν στον ίδιο υπολογιστή (2,66 GHz, 512M RAM) και όσο το δυνατόν υπό παρόμοιες συνθήκες.

Σύμφωνα με το [8], η πολυπλοκότητα των ΜΔΥ με γραμμικό πυρήνα κατά την εκπαίδευση είναι $O(m \cdot N^2)$, όπου m ο αριθμός των ιδιοτήτων και N ο αριθμός των

διανυσμάτων εκπαίδευσης. Στα συγκεκριμένα πειράματα χρησιμοποιήθηκε πυρήνας ακτινωτής βάσης. Παρόλα αυτά, από το επόμενο διάγραμμα φαίνεται να επιβεβαιώνεται η παραπάνω πολυπλοκότητα. Οι χρόνοι που παρουσιάζονται στο παρακάτω διάγραμμα αφορούν αποκλειστικά την εκπαίδευση των ΜΔΥ. Η πράσινη καμπύλη αντιστοιχεί στο χρόνο εκπαίδευσης της ΜΔΥ του 2ου περάσματος (δεν περιλαμβάνει το χρόνο εκπαίδευσης της ΜΔΥ του 1ου περάσματος).

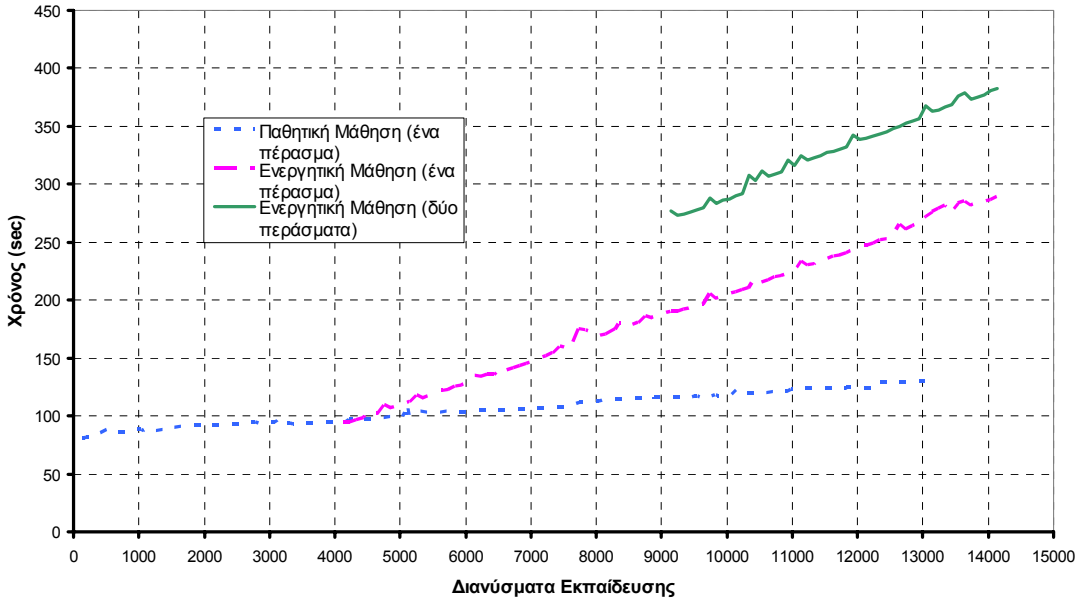


Χρόνος εκπαίδευσης των ΜΔΥ αναγνώρισης ονομάτων προσώπων στο 1ο πείραμα

Όπως φαίνεται από το σχήμα υπάρχει πολύ μεγάλη διαφορά μεταξύ του χρόνου εκπαίδευσης της παθητικής σε σχέση με την ενεργητική μάθηση. Ενδεχομένως, η διαφορά αυτή οφείλεται στο γεγονός ότι στην ενεργητική μάθηση τα περισσότερα διανύσματα είναι κοντά στο υπερεπίπεδο, είναι, δηλαδή, διανύσματα υποστήριξης και άρα δεν μπορούν να αγνοηθούν.

Το περίεργο στο παραπάνω σχήμα είναι ότι ο ταξινομητής του δεύτερου περάσματος, που διαθέτει περισσότερες ιδιότητες, φαίνεται αρχικά να εκπαιδεύεται πιο γρήγορα σε σχέση με τον ταξινομητή του πρώτου περάσματος. Ενδέχεται αυτό να οφείλεται στο γεγονός ότι τα αρχικά δεδομένα εκπαίδευσης του δεύτερου περάσματος έχουν επιλεχθεί με ενεργητική μάθηση βάσει των αποστάσεών τους από το υπερεπίπεδο διαχωρισμού της ΜΔΥ του 1ου περάσματος. Κάποια από τα αρχικά παραδείγματα εκπαίδευσης, δηλαδή, ενδέχεται να ήταν διανύσματα υποστήριξης για τη ΜΔΥ του πρώτου περάσματος, αλλά όχι για τη ΜΔΥ του δεύτερου.

Το παρακάτω διάγραμμα δείχνει το χρόνο που απαιτείται για την επεξεργασία της συλλογής ελέγχου, η οποία περιλαμβάνει 200 κείμενα συνολικού μεγέθους 1,22Μ (συνολικά 167.167 διανύσματα). Οι χρόνοι του σχήματος περιλαμβάνουν διάφορες διαδικασίες, συγκεκριμένα διαχωρισμό σε λεκτικές μονάδες, διαχωρισμό περιόδων, αναγνώριση χρονικών εκφράσεων και αναγνώριση ονομάτων προσώπων με τις ΜΔΥ. Όλοι οι χρόνοι, όμως, είναι δυνατόν να θεωρηθούν σταθεροί ως προς τον αριθμό των διανυσμάτων εκπαίδευσης, εκτός από το χρόνο των ΜΔΥ που χρησιμοποιούνται για την αναγνώριση ονομάτων προσώπων. Επομένως, οι διαφορές στους χρόνους οφείλονται μόνο σε αυτές τις ΜΔΥ.



Χρόνος ελέγχου των ΜΔΥ αναγνώρισης ονομάτων προσώπων στο 1ο πείραμα

Σύμφωνα με το [8], ο χρόνος ελέγχου για τις ΜΔΥ με γραμμικό πυρήνα είναι $O(m \cdot N)$, όπου m ο αριθμός των ιδιοτήτων και N ο αριθμός των διανυσμάτων εκπαίδευσης. Από τη στιγμή που ο αριθμός των ιδιοτήτων, δηλαδή, είναι σταθερός, ο χρόνος ελέγχου είναι γραμμικός σε σχέση με τα διανύσματα εκπαίδευσης, κάτι που φαίνεται να επιβεβαιώνεται από τις καμπύλες του παραπάνω διαγράμματος, παρ' όλο που χρησιμοποιήσαμε πυρήνα ακτινωτής βάσης.

Στην περίπτωση της ενεργητικής μάθησης η καμπύλη του χρόνου έχει μεγαλύτερη κλίση σε σχέση με την παθητική μάθηση. Αυτό οφείλεται στο γεγονός ότι στην ενεργητική μάθηση τα περισσότερα διανύσματα είναι διανύσματα υποστήριξης. Όπως αναφέρθηκε στην ενότητα 2.3.1, τα διανύσματα εκπαίδευσης που δεν είναι διανύσματα υποστήριξης αγνοούνται κατά τον υπολογισμό της απόκρισης της ΜΔΥ. Επομένως, στην παθητική μάθηση, όπου υπάρχουν λιγότερα διανύσματα υποστήριξης σε σχέση με την ενεργητική μάθηση, ο χρόνος ελέγχου είναι μικρότερος.

Στο παραπάνω διάγραμμα, αντίθετα από το προηγούμενο, η πράσινη καμπύλη περιλαμβάνει το χρόνο κατάταξης και των δύο ΜΔΥ. Η αύξηση του χρόνου ελέγχου στην περίπτωση του δεύτερου περάσματος οφείλεται στο ότι στην περίπτωση αυτή γίνεται κατάταξη από δύο ΜΔΥ.

5 Επίλογος

Παρουσιάσαμε ένα σύστημα αναγνώρισης και κατηγοριοποίησης χρονικών εκφράσεων και ονομάτων προσώπων για ελληνικά κείμενα. Το σύστημα χρησιμοποιεί ημι-αυτόματα παραγόμενα πρότυπα για την αναγνώριση των χρονικών εκφράσεων και Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) για την αναγνώριση των ονομάτων προσώπων. Στην περίπτωση της αναγνώρισης ονομάτων προσώπων, τα αποτελέσματα F-measure που επιτεύχθηκαν ήταν 87,5% και 93,34% για δύο διαθέσιμες συλλογές κειμένων. Στην αναγνώριση χρονικών εκφράσεων, τα αποτελέσματα ήταν 94,5% και 96,46% αντίστοιχα. Τα αποτελέσματα ήταν υψηλότερα στη δεύτερη συλλογή (οικονομικές ειδήσεις), λόγω της μικρότερης θεματικής ποικιλίας των κειμένων της και του πιο τυποποιημένου τρόπου γραφής τους. Σε ένα επιπλέον πείραμα, το σύστημα εκπαιδεύθηκε στην πρώτη συλλογή και αξιολογήθηκε στη δεύτερη. Το F-measure ήταν χαμηλότερο (82,29% για τα ονόματα προσώπων και 86,55% για τις χρονικές εκφράσεις), λόγω των διαφορών μεταξύ των κειμένων των δύο συλλογών.

Επίσης, χρησιμοποιήθηκαν στην αναγνώριση ονομάτων προσώπων τεχνικές ενεργητικής μάθησης, όπου το ίδιο το σύστημα προτείνει τα παραδείγματα εκπαίδευσης. Ως μέτρο αξιολόγησης των υποψηφίων παραδειγμάτων εκπαίδευσης χρησιμοποιήθηκε η απόσταση από το υπερεπίπεδο της ΜΔΥ. Τα πειραματικά αποτελέσματα δείχνουν ότι η ενεργητική μάθηση οδηγεί σε καλύτερα αποτελέσματα από ό,τι η παθητική όταν χρησιμοποιείται ο ίδιος αριθμός παραδειγμάτων εκπαίδευσης.

Κατά την αναγνώριση ονομάτων προσώπων δοκιμάστηκαν δύο ΜΔΥ στη σειρά, όπου η δεύτερη λαμβάνει υπόψη της το αν η πρώτη κατέταξε την υπό εξέταση λεκτική μονάδα αλλού στο ίδιο κείμενο ως όνομα προσώπου με υψηλή βεβαιότητα. Τα πειραματικά αποτελέσματα δείχνουν ότι η χρήση των δύο ΜΔΥ οδηγεί σε βελτίωση του F-measure μέχρι και 5% σε σχέση με τη χρήση μόνο μίας ΜΔΥ.

5.1 Μελλοντικές επεκτάσεις

Η προφανέστερη επέκταση που θα μπορούσε να γίνει στο σύστημα είναι η τροποποίησή του, ώστε να έχει τη δυνατότητα να υποστηρίζει όλες τις κατηγορίες ονομάτων οντοτήτων του MUC-7 (βλ. ενότητα 3). Ουσιαστικά, αρκεί να εκτελεστούν αρκετά πειράματα και να βρεθούν κατάλληλες ιδιότητες για τις κατηγορίες ονομάτων οργανισμών και τοπωνυμιών, καθώς κατά τα άλλα μπορεί να χρησιμοποιηθεί ο υπάρχων κώδικας.

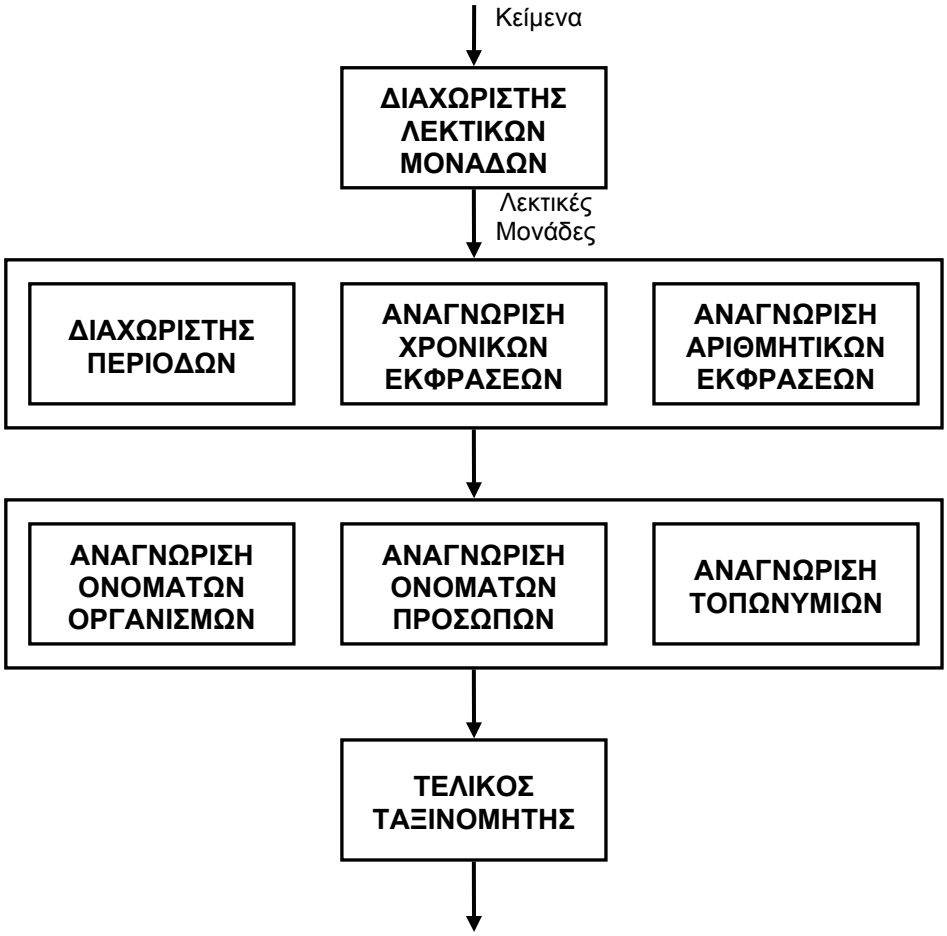
Για την κατηγορία των ονομάτων οργανισμών, από κάποια προκαταρκτικά πειράματα που έγιναν στη διάρκεια της εργασίας προέκυψε ότι είναι απαραίτητο να μεγαλώσει το παράθυρο από το οποίο αντλούνται οι ιδιότητες από 5 λεκτικές μονάδες σε 7. Η κατηγορία αυτή φαίνεται ότι είναι η δυσκολότερη και για αυτόν το λόγο απαιτείται να αυξηθεί ο αριθμός των ιδιοτήτων.

Η κατηγορία των τοπωνυμίων φαίνεται ότι χρειάζεται μία λίστα με τοπωνύμια. Υπάρχει ήδη στο σύστημα μία λίστα τοπωνυμίων, η οποία περιέχει ονόματα κρατών, πρωτευουσών και ηπείρων και δεν χρησιμοποιείται προς το παρόν. Είναι απαραίτητο να επεκταθεί, κυρίως με ελληνικά τοπωνύμια, όπως ονόματα νομών πόλεων, ποταμών, λιμνών, καθώς τα κείμενα είναι στα ελληνικά και αναμένεται να περιέχουν

κυρίως ελληνικές περιοχές. Επίσης, καλό θα ήταν να συμπεριληφθούν περισσότερες μεγάλες ξένες πόλεις και άλλα γνωστά παγκόσμια τοπωνύμια.

Κατά τη χειρωνακτική επισημείωση των κειμένων εκπαίδευσης παρατηρήθηκαν δυσκολίες στην κατάταξη (επισημείωση) κάποιων ονομάτων οντοτήτων, κυρίως των κατηγοριών των ονομάτων οργανισμών και τοπωνυμίων. Πολλές φορές δεν είναι σαφές σε ποια από τις δύο κατηγορίες πρέπει να ενταχθεί το όνομα. Για παράδειγμα, ο όρος «Βουλή» όταν χρησιμοποιείται στη φράση «ο πρόεδρος της Βουλής» θεωρείται πιο πιθανόν να ανήκει στην κατηγορία ονομάτων οργανισμών, ενώ όταν αναφέρεται «έξω από τη Βουλή» χρησιμοποιείται ως τοπωνύμιο. Ένα άλλο χαρακτηριστικό παράδειγμα είναι η φράση «Τρίκαλα – Ιωνικός 0 - 0». Όπου η μία ομάδα (Τρίκαλα) σύμφωνα με τους κανόνες επισημείωσης θεωρείται τοπωνύμιο, ενώ η άλλη (Ιωνικός) όνομα οργανισμού. Η αξιολόγηση του συστήματος θα έπρεπε να λαμβάνει υπόψη της ότι και οι άνθρωποι αντιμετωπίζουν προβλήματα κατά την επισημείωση τέτοιων περιπτώσεων. Μία πιθανή λύση είναι να επιτρέπεται στους ανθρώπους επισημειωτές να επισημειώσουν κάποιες λεκτικές μονάδες ως «οργανισμούς ή τοπωνύμια» και η απόκριση του συστήματος να θεωρείται σωστή αν αυτές οι λεκτικές μονάδες καταταγούν είτε ως οργανισμοί είτε ως τοπωνύμια.

Φυσικά, είναι απαραίτητο να υποστηριχθούν και οι ποσοτικές εκφράσεις (χρηματικές εκφράσεις και ποσοστά). Ενδέχεται να είναι δυνατόν εκφράσεις αυτών των κατηγοριών να μπορούν να εντοπιστούν με ημι-αυτόματα παραγόμενα πρότυπα, όπως οι χρονικές εκφράσεις.



Τελική προτεινόμενη αρχιτεκτονική

Στην περίπτωση που χρησιμοποιηθούν διαφορετικοί ταξινομητές για κάθε κατηγορία, δημιουργείται όπως έχει ήδη αναφερθεί, το πρόβλημα ότι μία λεκτική μονάδα μπορεί να καταταγεί ταυτόχρονα σε πολλές κατηγορίες. Μία πιθανή λύση είναι να υπάρξει ένας τελικός ταξινομητής, ο οποίος θα λαμβάνει υπόψη του τις αποφάσεις των ταξινομητών των διαφόρων κατηγοριών και θα αποφαίνεται για την κατηγορία που είναι περισσότερο πιθανή. Η τελική αρχιτεκτονική που προτείνεται φαίνεται στο παραπάνω σχήμα.

Μια άλλη πιθανή επέκταση είναι η βελτίωση του κριτηρίου που χρησιμοποιείται στην ενεργητική μάθηση για την αξιολόγηση των υποψηφίων παραδειγμάτων εκπαίδευσης. Ενδιαφέρουσες προτάσεις, οι οποίες, όμως, έχουν περιθώρια βελτίωσης και προσαρμογής είναι αυτές της εργασίας [32].

Ενδιαφέρον θα ήταν, επίσης, να προστεθεί στο σύστημα ένας επισημειωτής μερών του λόγου, ούτως ώστε να αντληθούν ιδιότητες από τα αποτελέσματά του. Θα μπορούσαν, για παράδειγμα, τα ρήματα να θεωρούνται πάντα μη ονόματα.

Τέλος, η μετατροπή του κώδικα του συστήματος από JAVA σε C++ θα αύξανε την ταχύτητά του. Η υλοποίηση των ΜΔΥ που χρησιμοποιήθηκε (βλ. παράρτημα) είναι διαθέσιμη και σε C++.

Παράρτημα

Στο παράρτημα αυτό θα παρουσιαστούν επιπλέον στοιχεία για την υλοποίηση των ΜΔΥ που χρησιμοποιήθηκε, δηλαδή τη βιβλιοθήκη libSVM [3, 14, 21].

Χρησιμοποιήθηκε ο πυρήνας ακτινωτής βάσης (RBF), του οποίου η συνάρτηση είναι:

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = \exp\left(-\gamma \cdot \left\|\overrightarrow{x_i} - \overrightarrow{x_j}\right\|^2\right), \gamma > 0$$

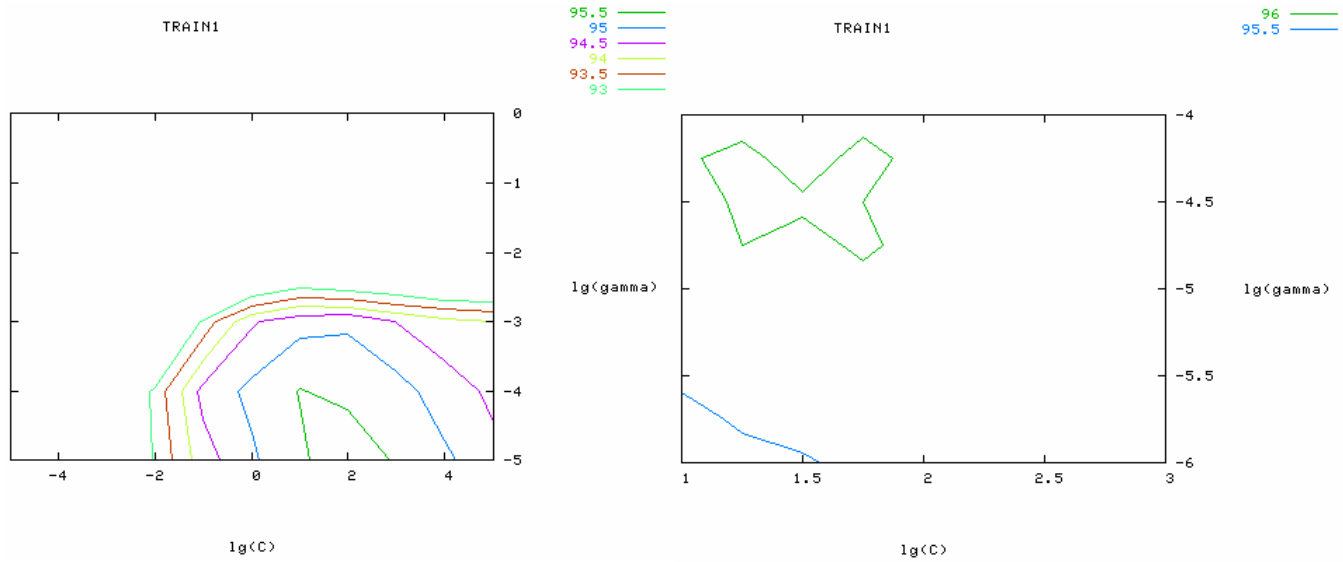
Όπως παρατηρούμε, ο πυρήνας αυτός έχει μία παράμετρο, το γ, την τιμή της οποίας επιλέγει ο χρήστης. Επίσης, υπάρχει η παράμετρος κόστους C (ενότητα 2.3.1), η οποία δείχνει την ανοχή επί του συνολικού σφάλματος. Οι δημιουργοί της libSVM προσφέρουν ένα εργαλείο, σε Python, το οποίο επιλέγει τις τιμές αυτών των παραμέτρων που είναι οι καταλληλότερες για ένα σύνολο δεδομένων εκπαίδευσης. Το εργαλείο εκτελεί διασταυρωμένη επικύρωση (cross-validation, ενότητα 3.2) στα δεδομένα εκπαίδευσης με διάφορες τιμές των παραμέτρων, και επιλέγει το συνδυασμό τιμών που οδηγεί στο μεγαλύτερο ποσοστό ορθότητας (accuracy). Πιο συγκεκριμένα, οι δημιουργοί της libSVM προτείνουν να δοκιμαστούν για κάθε παράμετρο τιμές που είναι δυνάμεις του 2. Για παράδειγμα, να ελεγχθούν οι τιμές $2^{-5}, 2^{-4}, 2^{-3}, ..., 2^{14}, 2^{15}$. Στη συνέχεια, αφού τελειώσει η αναζήτηση και δοθούν αρχικές τιμές στις δύο παραμέτρους, προτείνουν να εκτελεστεί μία πιο εξειδικευμένη, όσον αφορά το χώρο τιμών, αναζήτηση στη συγκεκριμένη γειτονιά που προέκυψε από το πρώτο βήμα. Για παράδειγμα, αν η τιμή που προέκυψε για το C είναι 2^2 να δοκιμαστούν για αυτή την παράμετρο οι τιμές $2^1, 2^{1,25}, 2^{1,5}, ..., 2^{2,75}, 2^3$. Αντίστοιχα, ενεργούμε και για την ρύθμιση του γ.

Η εντολή που πρέπει να εκτελεστεί για τη ρύθμιση των παραμέτρων (προϋποθέτει την εγκατάσταση της Python και του εργαλείου gnuplot) είναι η εξής:

```
python grid.py -log2c -5,5,1 -log2g -4,0,1 -svmtrain [path]\svmtrain
               -gnuplot [path]\pgnuplot -v 10 TRAIN
```

Η εντολή αυτή αφορά την πρώτη (γενικότερη) αναζήτηση. Πιο συγκεκριμένα, στο C (παράμετρος -log2c) δίνονται οι τιμές από 2^{-5} μέχρι 2^5 με βήμα 2^1 ενώ στο γ (παράμετρος -log2g) οι τιμές από 2^{-4} μέχρι 2^0 με βήμα 2^1. Σε όλες τις περιπτώσεις εκπαίδευσης μίας ΜΔΥ (ενότητες 3.2 και 4.2) διεξήχθη ρύθμιση παραμέτρων (γενική και εξειδικευμένη αναζήτηση) χρησιμοποιώντας τα δεδομένα εκπαίδευσης.

Στα σχήματα που ακολουθούν φαίνονται τα αποτελέσματα της αναζήτησης βέλτιστου συνδυασμού παραμέτρων που διεξήχθη κατά τη διάρκεια του πρώτου πειράματος (ενότητα 4.2) και αφορά την περίπτωση της παθητικής μάθησης του πρώτου περάσματος. Στο πρώτο σχήμα το C παίρνει τιμές από το σύνολο $\{2^{-5}, 2^{-4}, ..., 2^5\}$, ενώ το γ από το σύνολο $\{2^{-5}, 2^{-4}, ..., 2^0\}$. Στο δεύτερο σχήμα φαίνεται η πιο εξειδικευμένη αναζήτηση, γύρω από τις τιμές 2^2 και 2^{-5} που προέκυψαν από την πρώτη αναζήτηση, για τα C και γ αντίστοιχα.



Ρύθμιση παραμέτρων της υλοποίησης ΜΔΥ

Επιπλέον, οι κατασκευαστές του libSVM προτείνουν πριν από τη ρύθμιση των παραμέτρων οι τιμές των ιδιοτήτων να κανονικοποιηθούν στο διάστημα [-1, 1] ή [0, 1]. Αυτό είναι χρήσιμο για να μη δίνεται μεγαλύτερη βαρύτητα σε κάποιες ιδιότητες σε σχέση με κάποιες άλλες. Οι τιμές των ιδιοτήτων σε όλες τις περιπτώσεις που χρησιμοποιήθηκαν οι ΜΔΥ της βιβλιοθήκης libSVM (διαχωριστής περιόδων στην ενότητα 3.2 και αναγνώριση ονομάτων προσώπων στις ενότητες 3.4.2 και 3.4.3) κανονικοποιήθηκαν στο διάστημα [-1, 1].

Τέλος, πρέπει να αναφερθεί ότι η βιβλιοθήκη libSVM είναι διαθέσιμη τόσο σε Java όσο και σε C++. Το υπόλοιπο σύστημα της εργασίας αναπτύχθηκε σε Java. Παρ' όλα αυτά, η εκπαίδευση της ΜΔΥ με την υλοποίηση της libSVM που παρέχεται σε C++ είναι πολύ ταχύτερη. Για αυτόν το λόγο αποφασίστηκε κατά την εκπαίδευση να χρησιμοποιείται η υλοποίηση της libSVM που παρέχεται σε C++, καλώντας μία συνάρτηση C++ μέσα από τη Java, με τη χρήση του JNI [2] (Java Native Interface). Τα βήματα για να επιτευχθεί αυτό είναι τα εξής:

- Στον κώδικα της Java μπαίνει η δήλωση της συνάρτησης που θα γραφεί σε C++, χρησιμοποιώντας τη δεσμευμένη λέξη native. Π. χ.

```
public native void train(list of parameters);
```

- Εκτελείται η εντολή:

```
javah -jni trainer.java
```

  η οποία δημιουργεί το αρχείο κεφαλίδας trainer.h για την αντίστοιχη συνάρτηση σε C++.
- Στη συνέχεια, υλοποιείται η συνάρτηση train στη C++ και μεταγλωττίζεται το αρχείο trainer.cpp που την περιέχει. Αν χρησιμοποιείται ο μεταγλωττιστής GCC σε Windows μέσω της συλλογής Cygwin, αυτό γίνεται με την ακόλουθη εντολή:

```
g++ -O3 -Wl,--kill-at -mno-cygwin -I [path]\include
      -I [path]\include\win32 –shared
      -o trainer.dll trainer.cpp
```

- Τέλος, στον κώδικα της Java πρέπει να μπουν οι εντολές:

```
static {
        System.loadLibrary("trainer");
}
```

Η επικοινωνία μεταξύ της Java και της C++ γίνεται μέσω αρχείων. Πιο συγκεκριμένα, μία από τις παραμέτρους της native συνάρτησης είναι το όνομα του αρχείου που περιέχει τα διανύσματα εκπαίδευσης (πέρασμα από Java σε C++), το οποίο συμπληρώνεται από τη Java (ο διαχωριστής λεκτικών μονάδων και γενικότερα όλο το σύστημα είναι σε Java) και διαβάζεται από τη C++, ούτως ώστε να γίνει η εκπαίδευση. Μετά την εκπαίδευση, που εκτελείται με τον κώδικα που είναι γραμμένος σε C++, αποθηκεύεται σε ένα αρχείο ο ταξινομητής που προέκυψε, με τη μορφή διανυσμάτων υποστήριξης. Το αρχείο αυτό δημιουργείται από τη C++ και διαβάζεται από τη Java, ούτως ώστε να φορτωθεί ο ταξινομητής κατά τη φάση ελέγχου και τη φάση χρήσης.

Όσον αφορά τη διαδικασία ελέγχου, επιλέχθηκε να γίνεται με την υλοποίηση της libSVM που παρέχεται σε Java. Δεν είναι εύκολο σε αυτή την περίπτωση να χρησιμοποιηθεί η υλοποίηση σε C++, γιατί δημιουργείται πρόβλημα με την τιμή της ιδιότητας «απόσταση από την αρχή του ονόματος προσώπου» (ενότητα 3.4.2). Όπως έχει αναφερθεί, η τιμή αυτής της ιδιότητας στη φάση εκπαίδευσης ανακτάται από τα επισημειωμένα κείμενα, οπότε είναι δυνατή η δημιουργία ενός αρχείου με τα διανύσματα εκπαίδευσης πριν αρχίσει η διαδικασία της εκπαίδευσης. Αντίθετα, στη φάση ελέγχου η τιμή της ιδιότητας εξαρτάται από την απόφαση του ταξινομητή για τις προηγούμενες λεκτικές μονάδες. Δεν είναι, λοιπόν, δυνατόν να δημιουργηθεί ένα αρχείο με τα διανύσματα ελέγχου, όπως γίνεται στην εκπαίδευση. Επιπλέον, είναι ιδιαίτερα χρονοβόρο να μεταφέρεται ο έλεγχος (παράμετροι κ.λ.π.) για κάθε διάνυσμα ξεχωριστά από τη Java στη C++ και το αντίστροφο.

Βιβλιογραφία

[1] *CoNLL-2003*, *http://www.cnts.ua.ac.be/conll2003/*.
[2] *JNI - Java Native Interface*, http://java.sun.com/j2se/1.5.0/docs/guide/jni/.
[3] *LibSVM*, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[4] *MUC-7 Named Entity Task Definition*, *http://www.itl.nist.gov/894.02/related_projects/muc/proceedings/ne_task.html*.
[5] *WEKA*, http://www.cs.waikato.ac.nz/~ml/weka/index.html.
[6] *"TA NEA"*, http://ta-nea.dolnet.gr.
[7] *"TO BHMA"*, http://tovima.dolnet.gr.
[8] I. Androutsopoulos, G. Paliouras and E. Michelakis, *Learning to Filter Unsolicited Commercial E-Mail*.
[9] D. Bikel, S. Miller, R. Schwartz and R. Weischedel, *Nymble: A High-Performance Learning Name-finder*, Conference on Applied Natural Language Processing (1997).
[10] W. Black, F. Rinaldi and D. Mowatt, *FACILE: Description of the NE System used for MUC-7*, Proceedings of Seventh Message Understanding Conference (1998).
[11] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, *NYU: Description of the MENE Named Entity System as Used in MUC-7*, Proceedings of Seventh Message Understanding Conference (1997).
[12] S. Boutsis, I. Demiros, V. Giouli, M. Liakata, H. Papageorgiou and S. Piperidis, *A System Recognition of Named Entities in Greek*, Lecture Notes in Computer Science (2000).
[13] K. Brinker, *Incorporating Diversity in Active Learning with Support Vector Machines*, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003) (2003).
[14] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*.
[15] H. L. Chieu, Hwee Tou Ng, *Named Entity Recognition with a Maximum Entropy Approach*, Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003) (2003).
[16] C. Cortes and V. P. Vapnik, *Support-vector networks*, Machine Learning, 20(3) (1995), pp. 273-297.
[17] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.
[18] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. Spyropoulos and P. Stamatopoulos, *Rule-based Named Entity Recognition for Greek Financial Texts*, Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMPLEX 2000) (2000), pp. 75-78.
[19] R. Florian, A. Ittycheriah, H. Jing and T. Zhang, *Named Entity Recognition through Classifier Combination*, Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003) (2003).
[20] R. Grishman, *The NYU System for MUC-6 or Where's the Syntax*, Proceedings of Sixth Message Understanding Conference (1995).
[21] C.-W. Hsu, C.-C. Chang and C.-J. Lin, *A Practical Guide to Support Vector Classification*.

[22] V. Karkaletsis, G. Paliouras, G. Petasis, N. Manousopoulou and C. D. Spyropoulos, *Named-Entity Recognition from Greek and English Texts*, Journal of Intelligent and Robotic Systems (1999), pp. 123-135.
[23] G. R. Krupka and K. Hausman, *IsoQuest, Inc: Description of the NetOwl Extractor System as Used for MUC-7*, Proceedings of Seventh Message Understanding Conference (1998).
[24] A. Mikheev, C. Grover and M. Moens, *Description of the LTG System used for MUC-7*, Proceedings of Seventh Message Understanding Conference (1997).
[25] A. Mikheev, M. Moens, C. Grover, *Named Entity Recognition without Gazetteers*, Proceedings of EACL '99 (1999).
[26] T. M. Mitchell, *Machine Learning*, McGraw-Hill International Editions, 1997.
[27] MUC-6, *Proceedings of the Sixth Message Understanding Conference*, 1995.
[28] MUC-7, *Proceedings of the Seventh Message Understanding Conference*, 1998.
[29] G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis and C. Spyropoulos, *Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems*, Meeting of the Association for Computational Linguistics (ACL) (2002).
[30] R. Schapire and Y. Singer, *BoosTexter: A Boosting-based System for Text Categorization*, *Machine Learning*, 2000, pp. 135-168.
[31] G. Schohn and D. Cohn, *Less is more: Active learning with support vector machines*, Seventeenth International Conference on Machine Learning (2000).
[32] D. Shen, J. Zhang, J. Su, G. Zhou and C.-L. Tan, *Multi-Criteria-based Active Learning for Named Entity Recognition*, Association for Computational Linguistics (ACL) (2004).
[33] V. P. Vapnik, *Statistical Learning Theory*, (1998).