ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Ανάπτυξη Προσαρμόσιμου Αναγνωριστή Ονομάτων Οντοτήτων*»

Σπυρίδων Αντωνέλλος

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2009

Περίληψη

Στην παρούσα εργασία αναπτύχθηκε ένα σύστημα αναγνώρισης ονομάτων οντοτήτων (named entity recognizer) για κείμενα φυσικής γλώσσας. Το σύστημα είναι δυνατόν να προσαρμοστεί, μέσω επιβλεπόμενης μηχανικής μάθησης, για χρήση με κείμενα διαφορετικών φυσικών γλωσσών και ονόματα οντοτήτων νέων κατηγοριών. Η εκπαίδευση του συστήματος βασίζεται στον αλγόριθμο της Μέγιστης Εντροπίας. Το σύστημα δοκιμάστηκε σε τέσσερις διαφορετικές συλλογές κειμένων, γραμμένων στα ελληνικά, αγγλικά, ισπανικά και ολλανδικά αντίστοιχα. Στα κείμενα αυτά υπάρχουν χειρωνακτικά επισημειωμένες συνολικά εννέα διαφορετικές κατηγορίες ονομάτων οντοτήτων. Οι επιδόσεις του συστήματος ήταν αρκετά ικανοποιητικές και σε αρκετές περιπτώσεις συγκρίσιμες με εκείνες συστημάτων που κατασκευάστηκαν για συγκεκριμένες γλώσσες και κατηγορίες ονομάτων.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας κ. Ίωνα Ανδρουτσόπουλο, επίκουρο καθηγητή στο Τμήμα Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών, για τη συνεχή καθοδήγηση κατά τη διάρκεια αυτής της εργασίας. Επίσης ευχαριστώ τον υποψήφιο διδάκτορα του ιδίου τμήματος κ. Γεώργιο Λουκαρέλλι για τις συμβουλές του στο ξεκίνημα της εργασίας, καθώς επίσης και για τη διάθεση μίας εκ των τεσσάρων συλλογών κειμένων που χρησιμοποιήθηκαν στην εργασία. Τέλος θα ήθελα να ευχαριστήσω όλα τα μέλη της Ομάδας Επεξεργασίας Φυσικής Γλώσσας του ΟΠΑ για τις πολύ χρήσιμες συμβουλές, ιδέες και παρατηρήσεις τους σε όλες τις φάσεις αυτής της εργασίας.

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

Η αναγνώριση ονομάτων οντοτήτων (named entity recognition, NER) περιλαμβάνει τον εντοπισμό λέξεων και κατ' επέκταση φράσεων οι οποίες αποτελούν το όνομα κάποιας οντότητας και τον προσδιορισμό της κατηγορίας της οντότητας αυτής. Ένα τέτοιο σύστημα θα μπορούσε να εκπαιδευτεί, μέσω μηχανικής μάθησης, και να εφαρμοστεί σε οποιαδήποτε γλώσσα και για οποιαδήποτε κατηγορία ονομάτων οντοτήτων. Υπάρχουν συστήματα που έχουν εκπαιδευτεί σε κείμενα διαφόρων γλωσσών και στόχος τους, για παράδειγμα, είναι να αναγνωρίσουν ονόματα προσώπων, τοποθεσιών, οργανισμών, πρωτεϊνών και γενικότερα οτιδήποτε μπορεί να θεωρηθεί ως όνομα οντότητας. Το σύστημα που παρουσιάζεται σε αυτήν την εργασία έχει την ιδιαιτερότητα ότι μπορεί να προσαρμοστεί εύκολα για χρήση με κείμενα διαφορετικών γλωσσών και νέες κατηγορίες ονομάτων οντοτήτων.

Πέρα από την αυτόνομη χρήση της (π.χ. για τη δημιουργία ευρετηρίων ονομάτων), η αναγνώριση ονομάτων οντοτήτων βρίσκει εφαρμογή και ως συστατικό άλλων συστημάτων που σκοπό έχουν την ανάκτηση ή εξαγωγή πληροφοριών (information retrieval and extraction) και γενικότερα την επεξεργασία φυσικής γλώσσας (natural language processing)· χρησιμοποιείται, για παράδειγμα, και σε συστήματα που παράγουν περιλήψεις, συστήματα ερωταποκρίσεων, μηχανικής μετάφρασης κ.ά.

Στο παρελθόν έχουν διοργανωθεί πολλά συνέδρια και διαγωνισμοί αναγνώρισης ονομάτων οντοτήτων και παραλλαγών της. Ιδιαίτερα γνωστοί είναι οι διαγωνισμοί MUC-6[1] και MUC-7[2] της δεκαετίας του '90, στους οποίους χρησιμοποιήθηκαν κείμενα εκπαίδευσης και ελέγχου γραμμένα στην αγγλική γλώσσα· οι κατηγορίες οντοτήτων ήταν ονόματα προσώπων, τοποθεσιών και οργανισμών, καθώς και χρονικές και ποσοτικές εκφράσεις. Τα περισσότερα συστήματα που συμμετείχαν στο διαγωνισμό MUC-7, πέντε από τα οχτώ για την ακρίβεια, βασίζονταν σε χειρωνακτικούς κανόνες και οι επιδόσεις τους ήταν αρκετά υψηλές, με τα περισσότερα να ξεπερνούν το 85% σε F-measure. Την καλύτερη επίδοση είχε το σύστημα των Mikheev κ.ά. [1] με F-measure περίπου 93%. Στους μετέπειτα διαγωνισμούς CoNLL-2002[3] και CoNLL-2003[4] εμφανίστηκαν περισσότερα συστήματα που χρησιμοποιούσαν τεχνικές μηχανικής μάθησης. Σκοπός ήταν τα συστήματα να μπορούν να προσαρμοστούν εύκολα σε πολλές διαφορετικές γλώσσες. Συγκεκριμένα, στο διαγωνισμό του 2002 τα συστήματα δοκιμάστηκαν στα ισπανικά και τα ολλανδικά, ενώ το 2003 στα αγγλικά και τα γερμανικά. Οι κατηγορίες οντοτήτων των οποίων τα ονόματα έπρεπε να αναγνωρίζονται ήταν πρόσωπα, οργανισμοί και τοποθεσίες, καθώς και μία τέταρτη κατηγορία (miscellaneous) για άλλες οντότητες. Το 2002, καλύτερο σύστημα αναδείχθηκε αυτό των Carrera κ.ά. [2], που με τη χρήση κυρίως του αλγορίθμου μάθησης AdaBoost [3] πέτυχε F-measure πάνω από 81% και 77% για τις δύο γλώσσες αντίστοιχα. Η διαφορά της επίδοσης του συστήματος αυτού από τα αντίστοιχα δεύτερα σε επίδοση συστήματα ήταν περίπου 2%. Το 2003 την καλύτερη επίδοση είχε πετύχει και στις δύο γλώσσες το σύστημα των Florian κ.ά. [4], το οποίο με τη χρήση συνδυασμού αλγορίθμων μάθησης,

[1] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
[2] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html~
[3] http://www.cnts.ua.ac.be/conll2002/ner/
[4] http://www.cnts.ua.ac.be/conll2003/ner/

όπως αυτών της Μέγιστης Εντροπίας [5] και των Μοντέλων Markov [6,7], πέτυχε F-measure περίπου 88% και 72% στις αντίστοιχες γλώσσες.

Σκοπός της παρούσας εργασίας ήταν η ανάπτυξη ενός προσαρμόσιμου αναγνωριστή ονομάτων οντοτήτων, ικανού να εκπαιδεύεται σε οποιαδήποτε γλώσσα καθώς και για οποιαδήποτε κατηγορία οντοτήτων. Για παράδειγμα, το σύστημα μπορεί να εκπαιδευτεί να αναγνωρίζει σε ελληνικά ειδησεογραφικά άρθρα ονόματα οργανισμών ή να εκπαιδευτεί να αναγνωρίζει σε αγγλικά κείμενα βιοϊατρικών άρθρων ονόματα πρωτεϊνών κτλ. Όπως σε όλα τα συστήματα αναγνώρισης ονομάτων οντοτήτων, όμως, τα κείμενα (εκπαίδευσης και ελέγχου) είναι καλύτερα να προέρχονται κάθε φορά από μία συγκεκριμένη θεματική περιοχή (π.χ. οικονομία, βιοϊατρική) και να είναι ενός συγκεκριμένου είδους (π.χ. ειδήσεις, άρθρα επιστημονικών περιοδικών).

Η επίδοση που επιτυγχάνει το σύστημά μας είναι όχι μόνο υψηλότερη εκείνης απλοϊκών (baseline) συστημάτων, αλλά σε αρκετές περιπτώσεις πλησιάζει και ξεπερνά την επίδοση συστημάτων που κατασκευάστηκαν για συγκεκριμένη γλώσσα και συγκεκριμένες κατηγορίες ονομάτων οντοτήτων. Το σύστημά μας δοκιμάστηκε σε τέσσερις διαφορετικές γλώσσες, ελληνικά, αγγλικά, ισπανικά και ολλανδικά, με συνολικά εννέα διαφορετικές κατηγορίες ονομάτων οντοτήτων, που συμπεριλαμβάνουν κατηγορίες όπως ονόματα προσώπων και οργανισμών, αλλά και κατηγορίες ονομάτων βιοϊατρικών οντοτήτων όπως πρωτεϊνών.

Στη συνέχεια, στο κεφάλαιο 2, συνοψίζονται τα είδη των συστημάτων αναγνώρισης ονομάτων οντοτήτων και οι μεθόδοι που χρησιμοποιούνται στον τομέα αυτό. Επίσης, παρουσιάζονται συνοπτικά οι συλλογές κειμένων που υπάρχουν διαθέσιμες και χρησιμοποιήθηκαν στα πειράματα της εργασίας. Στο κεφάλαιο 3 περιγράφεται η αρχιτεκτονική του συστήματός μας, παρουσιάζονται αναλυτικά τα σύνολα ιδιοτήτων (attribute sets) που δοκιμάσαμε και οι τεχνικές αξιολόγησης και επιλογής τους. Στο κεφάλαιο 4 περιγράφονται τα πειράματα της εργασίας, συμπεριλαμβανομένων των μέτρων αξιολόγησης, παρατίθενται περισσότερες πληροφορίες για τα κείμενα των πειραμάτων και αναλύονται τα πειραματικά αποτελέσματα. Τέλος, στο κεφάλαιο 5 συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται μελλοντικές βελτιώσεις του συστήματος.

Κεφάλαιο 2: Βιβλιογραφική επισκόπηση

Στη βιβλιογραφία συναντάται πληθώρα διαφορετικών προσεγγίσεων που χρησιμοποιούνται για την αναγνώριση ονομάτων οντοτήτων. Για παράδειγμα υπάρχουν συστήματα που είναι εξολοκλήρου βασισμένα σε χειρωνακτικούς κανόνες και δεν χρησιμοποιούν καμία τεχνική μάθησης. Άλλα ενσωματώνουν τεχνικές μάθησης, αλλά διαφέρουν ως προς τον αλγόριθμο μάθησης που χρησιμοποιούν, το σύνολο των ιδιοτήτων κλπ. Ακολουθεί μια σύντομη περιγραφή των συνηθέστερων προσεγγίσεων που υιοθετούν τα συστήματα αναγνώρισης ονομάτων οντοτήτων, καθώς και μια περιγραφή των ιδιοτήτων που χρησιμοποιούνται στα συστήματα που βασίζονται σε αλγορίθμους μηχανικής μάθησης. Τέλος, παρατίθενται στοιχεία για γνωστές διαθέσιμες συλλογές κειμένων που έχουν χρησιμοποιηθεί σε πειράματα αναγνώρισης ονομάτων οντοτήτων.

2.1 Είδη συστημάτων αναγνώρισης ονομάτων οντοτήτων

Στην περίπτωση των συστημάτων που βασίζονται σε χειρωνακτικούς κανόνες, ο κατασκευαστής του συστήματος μελετά τα κείμενα εκπαίδευσης και δημιουργεί κανόνες οι οποίοι αναγνωρίζουν ονόματα οντοτήτων. Για παράδειγμα, στην περίπτωση των ελληνικών ονομάτων προσώπων, ένας τέτοιος κανόνας θα μπορούσε να είναι ο «εάν η προηγούμενη λεκτική μονάδα της *w* είναι η «κ.», τότε επισημείωσε την *w* ως όνομα προσώπου» ή, στην περίπτωση που αναζητούμε ονόματα οργανισμών, «εάν η επόμενη λεκτική μονάδα της *w* είναι η «Α.Ε.», τότε επισημείωσε την *w* ως όνομα οργανισμού». Επιπροσθέτως συναντάται συχνά η κατασκευή, πριν τη δημιουργία και εφαρμογή των κανόνων, λιστών με λέξεις γνωστών ονομάτων οντοτήτων και η χρήση κανόνων όπως «εάν η *w* ανήκει στη λίστα με τα βαφτιστικά ονόματα, τότε επισημείωσέ την ως όνομα προσώπου». Ένα τέτοιο σύστημα μπορεί μεν να είναι απλό στην κατασκευή και γρήγορο κατά την εφαρμογή του, αλλά υστερεί σε ευελιξία. Οι κανόνες πρέπει να ξαναγραφτούν χειρωνακτικά εξαρχής, αν θέλουμε να αναγνωρίσουμε ονόματα οντοτήτων νέων κατηγοριών ή να μεταβούμε σε κείμενα άλλης γλώσσας. Ακόμα και σε κείμενα της ίδια γλώσσας, είναι αμφίβολη η αποτελεσματικότητα των κανόνων όταν εφαρμοστούν σε κείμενα άλλης θεματικής περιοχής ή είδους. Τέλος, τα συστήματα που βασίζονται σε χειρωνακτικούς κανόνες ενδέχεται να επιτυγχάνουν υψηλή ακρίβεια (precision), δηλαδή να μην επισημειώνουν λανθασμένα ονόματα, αλλά συνήθως υστερούν σε ανάκληση (recall), δηλαδή τους διαφεύγουν πολλά ονόματα.

Μία διαφορετική και πλέον συνηθέστερα χρησιμοποιούμενη προσέγγιση είναι η χρήση μεθόδων μηχανικής μάθησης, κυρίως επιβλεπόμενης. Στην περίπτωση αυτή, το σύστημα εκπαιδεύεται σε κείμενα τα ονόματα των οποίων έχουν επισημειωθεί χειρωνακτικά, ώστε στη συνέχεια να μπορεί να εντοπίζει ονόματα οντοτήτων σε νέα, μη επισημειωμένα κείμενα. Συχνά χρησιμοποιούμενοι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης είναι οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines, SVMs) [8], ο αλγόριθμος της Μέγιστης Εντροπίας (Maximum Entropy, ME) [5], τα Κρυφά Μοντέλα Markov (Hidden Markov Models, HMM) [6,7], ο αλγόριθμος C4.5 [9] και ο αλγόριθμος των *k* κοντινότερων γειτόνων (*k* Nearest Neighbors) [10]. Πολλά συστήματα χρησιμοποιούν συνδυασμούς αλγορίθμων μηχανικής μάθησης ή/και χειρωνακτικούς κανόνες και χαρακτηρίζονται ως υβριδικά. Επίσης, μεταξύ των συστημάτων παρατηρούνται και δομικές διαφορές. Για παράδειγμα, στους διαγωνισμούς CoNLL-

2002 και CoNLL-2003, όπου υπήρχαν τέσσερις κατηγορίες ονομάτων, μερικοί ερευνητές επέλεξαν να χρησιμοποιήσουν ένα μόνο ταξινομητή ο οποίος είχε εκπαιδευτεί να αναγνωρίζει ονόματα και των τεσσάρων κατηγοριών. Σε άλλα συστήματα χρησιμοποιήθηκε σε πρώτο στάδιο ένας ταξινομητής ο οποίος επέλεγε λεκτικές μονάδες που θεωρούσε ότι αποτελούσαν μέρη ονομάτων οντοτήτων· σε δεύτερο στάδιο εφαρμοζόταν ένας άλλος ταξινομητής που είχε εκπαιδευτεί να διαχωρίζει στις τέσσερις κατηγορίες τα ονόματα που εντόπιζε ο πρώτος.

2.2 Είδη ιδιοτήτων

Τα συστήματα αναγνώρισης ονομάτων οντοτήτων που βασίζονται σε αλγορίθμους μηχανικής μάθησης διαφέρουν κυρίως στις ιδιότητες (attributes) που χρησιμοποιούν. Λέγοντας εδώ «ιδιότητα» εννοούμε μία συνάρτηση η οποία εξετάζει την τρέχουσα λέξη *w*, τα συμφραζόμενά της, καταλόγους γνωστών ονομάτων ή/και άλλες διαθέσιμες πληροφορίες και επιστρέφει μία τιμή που σκοπό έχει να βοηθήσει το σύστημα να αποφασίσει αν η *w* αποτελεί όνομα (ή μέρος ονόματος) οντότητας και ποιας κατηγορίας. Η τιμή της ιδιότητας μπορεί να είναι μια λογική τιμή (*true*, *false*), ένας αριθμός (π.χ. ακέραιος, πραγματικός), μία συμβολοσειρά κλπ. Για παράδειγμα, θα μπορούσαμε να είχαμε ορίσει τις παρακάτω ιδιότητες, όπου χάριν συντομίας δείχνουμε ως ορίσματα μόνο την ίδια την *w*:

$$f_1(w) = \begin{cases} 1, \text{ εάν η w αρχίζει με κεφαλαίο γράμμα} \\ 0, \text{ διαφορετικά} \end{cases}$$

$f_2(w) =$ το μέρος του λόγου (π.χ. ρήμα) της προηγούμενης λεκτικής μονάδας της w

Η κάθε λέξη *w* που εξετάζουμε μπορεί να παρασταθεί ως ένα διάνυσμα χαρακτηριστικών (features), όπου κάθε χαρακτηριστικό είναι η τιμή μιας συγκεκριμένης ιδιότητας με όρισμα την *w*.

Οι πιο διαδεδομένες ιδιότητες στην αναγνώριση ονομάτων οντοτήτων είναι οι μορφολογικές. Αυτές εξετάζουν μόνο την ίδια τη λέξη *w* ως συμβολοσειρά. Για παράδειγμα, εξετάζουν εάν η *w* αρχίζει με ή περιέχει κάποιο κεφαλαίο γράμμα, εάν περιέχει αριθμούς, εάν είναι ή περιέχει σημείο στίξης κλπ. Επίσης, συνηθίζεται οι ιδιότητες αυτές να υπολογίζονται και για τις λέξεις που πλαισιώνουν την *w* (π.χ. αν αρχίζει με κεφαλαίο η προηγούμενη λέξη της *w*).

Άλλες συχνά χρησιμοποιούμενες ιδιότητες είναι αυτές που εξετάζουν λίστες γνωστών κυρίων ονομάτων (gazetteers) ή γενικότερα λέξεων που σηματοδοτούν ονόματα οντοτήτων. Για παράδειγμα, χρησιμοποιούνται λίστες με συνηθισμένα βαφτιστικά ονόματα, λίστες που περιέχουν λεκτικές μονάδες (tokens) που συναντώνται πριν από ονόματα προσώπων, όπως το «κ.» στα ελληνικά κείμενα (π.χ. «…ο κ. Κώστας Σημίτης…») κλπ. Επίσης, ενδέχεται να χρησιμοποιούνται λίστες με συχνά προθέματα και καταλήξεις ονομάτων οντοτήτων. Κάθε εγγραφή μιας λίστας δεν είναι υποχρεωτικό να απαρτίζεται από μόνο μία λεκτική μονάδα· συχνά συναντάμε ζευγάρια, τριάδες κλπ. λεκτικών μονάδων, όπως το ζεύγος λέξεων «city of» που συχνά συνοδεύει ονόματα πόλεων (π.χ. «…the city of New York…»).

Η κατασκευή των λιστών γίνεται κυρίως με δύο τρόπους. Ο πρώτος είναι να συμπληρωθούν οι λίστες βάσει των κειμένων εκπαίδευσης. Για παράδειγμα, στην περίπτωση που αναζητούμε ονόματα προσώπων, εξετάζουμε τα κείμενα εκπαίδευσης και αποθηκεύουμε σε μια λίστα λέξεις τις οποίες συναντάμε επισημειωμένες ως ονόματα οντοτήτων. Ο δεύτερος τρόπος είναι να ετοιμαστούν οι λίστες χειρωνακτικά ή εξάγοντας πληροφορίες από προϋπάρχουσες λίστες (π.χ. τηλεφωνικοί κατάλογοι).

Ο απλούστερος τρόπος χρήσης λιστών είναι η προσθήκη ιδιοτήτων στο σύστημα που να εξετάζουν εάν υπάρχει η *w* ή κάποια γειτονική της λεκτική μονάδα (π.χ. η προηγούμενη της *w*) σε κάποια συγκεκριμένη λίστα. Για παράδειγμα:

$$f_3(w) = \begin{cases} 1, \text{ εάν η προηγούμενη λεκτική μονάδα της w ανήκει στη λίστα με τις} \\ \quad \text{λεκτικές μονάδες που εμφανίζονται αμέσως πριν από ονόματα προσώπων} \\ 0, \text{ διαφορετικά} \end{cases}$$

Οι λεκτικές μονάδες μιας τέτοιας λίστας, συνεχίζοντας το παράδειγμα της f_3, δε δείχνουν όλες με την ίδια πιθανότητα ότι μετά από αυτές ακολουθεί όνομα προσώπου. Πολλά συστήματα συσχετίζουν κάθε εγγραφή λίστας με μια εκτίμηση της πιθανότητας η εγγραφή να συνοδεύει όνομα προσώπου συγκεκριμένης κατηγορίας. Για παράδειγμα, θα μπορούσε για κάθε εγγραφή να υπολογίζεται η ακρίβειά της (precision), δηλαδή ο λόγος των εμφανίσεων της εγγραφής σε περιπτώσεις που συνόδευε ονόματα συγκεκριμένης κατηγορίας (π.χ. ονόματα προσώπων) δια του συνολικού αριθμού εμφανίσεων της εγγραφής. Οι εγγραφές της λίστας συχνά ταξινομούνται κατά φθίνουσα σειρά ακρίβειας και κρατιούνται στη λίστα μόνο οι κορυφαίες εγγραφές (π.χ. οι πρώτες 1.000). Τέλος, αντί Boolean ιδιότητας θα μπορούσαμε να χρησιμοποιήσουμε μια ιδιότητα που επιστρέφει την αντίστοιχη τιμή ακρίβειας:

$$f_4(w) = \begin{cases} p, \text{ εάν η προηγούμενη λεκτική μονάδα της w ανήκει (με ακρίβεια p) στη λίστα με τις} \\ \quad \text{λεκτικές μονάδες που εμφανίζονται αμέσως πριν από ονόματα προσώπων} \\ 0, \text{ διαφορετικά} \end{cases}$$

Μια άλλη κατηγορία συχνά χρησιμοποιούμενων ιδιοτήτων είναι οι ιδιότητες που επιστρέφουν τις ετικέτες μερών του λόγου (part-of-speech tags, POS tags) της *w* ή των γειτονικών της λεκτικών μονάδων, ετικέτες που προστίθενται από έναν επισημειωτή μερών του λόγου (POS tagger).

2.3 Συλλογές επισημειωμένων κειμένων

Για να είναι δυνατή η σύγκριση των επιδόσεων διαφορετικών συστημάτων αναγνώρισης ονομάτων οντοτήτων, πρέπει να γίνουν πειράματα με τα συστήματα στις ίδιες συλλογές κειμένων. Σε κάθε κείμενο των συλλογών πρέπει να έχουν επισημειωθεί χειρωνακτικά τα ονόματα οντοτήτων και οι κατηγορίες τους, ώστε να είναι δυνατή η εκπαίδευση και αξιολόγηση των συστημάτων. Τα συστήματα εκπαιδεύονται και αξιολογούνται σε διαφορετικά κείμενα των συλλογών. Ακριβέστερα, κάθε συλλογή κειμένων χωρίζεται συνήθως σε τρία μη επικαλυπτόμενα τμήματα, που χρησιμοποιούνται αντίστοιχα για

εκπαίδευση (training data), ρύθμιση παραμέτρων και προκαταρκτικές δοκιμές (development data, δεδομένα ανάπτυξης) και τελική αξιολόγηση (test data).

Μεταξύ των πιο συχνά χρησιμοποιούμενων συλλογών κειμένων είναι αυτές των συνεδρίων/διαγωνισμών MUC-7, CoNLL-2002, CoNLL-2003, καθώς και η συλλογή κειμένων GENIA[5]. Τα κείμενα του MUC-7 είναι στην αγγλική γλώσσα και είναι επισημειωμένα με ονόματα προσώπων, τοποθεσιών και οργανισμών. Τα κείμενα του CoNLL-2002 είναι στα ισπανικά και ολλανδικά, ενώ του CoNLL-2003 στα αγγλικά και γερμανικά· στα κείμενα των δύο αυτών συλλογών έχουν επισημειωθεί τα ονόματα προσώπων, τοποθεσιών και οργανισμών, όπως στο MUC-7, καθώς και μια τέταρτη κατηγορία ονομάτων (miscellaneous) που περιλαμβάνει ονόματα οντοτήτων άλλων ειδών. Η συλλογή GENIA περιλαμβάνει 2.000 περιλήψεις βιοϊατρικών δημοσιεύσεων. Είναι στην αγγλική γλώσσα και έχουν επισημειωθεί σε αυτή τα ονόματα τριάντα έξι κατηγοριών βιοϊατρικών οντοτήτων.

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

Κεφάλαιο 3: Το σύστημα της εργασίας

Κατά την εκπαίδευση του συστήματος, δεδομένης της ζητούμενης κατηγορίας ονομάτων και κειμένων της επιθυμητής γλώσσας, δημιουργούνται αρχικά τα διανύσματα εκπαίδευσης για τις λεκτικές μονάδες των δεδομένων εκπαίδευσης (training data)· τα διανύσματα περιλαμβάνουν χαρακτηριστικά για όλες τις ιδιότητες που υποστηρίζει το σύστημα. Έπειτα, χρησιμοποιώντας τα δεδομένα ανάπτυξης (development data), γίνεται η επιλογή των πιο χρήσιμων ιδιοτήτων. Κατόπιν το σύστημα εκπαιδεύεται πάνω στα διανύσματα των δεδομένων εκπαίδευσης, λαμβάνοντας υπόψη μόνο τις επιλεγμένες ιδιότητες που προέκυψαν από το προηγούμενο βήμα, χρησιμοποιώντας τον αλγόριθμο της Μέγιστης Εντροπίας, ώστε να προκύψει ο «κύριος ταξινομητής» του συστήματος.[6] Ο ταξινομητής αυτός είναι εκπαιδευμένος να εντοπίζει λεκτικές μονάδες που αποτελούν ονόματα (ή μέρη ονομάτων) της ζητούμενης κατηγορίας, κατατάσσοντας κάθε μία στις δύο πιθανές κατηγορίες: «ανήκει σε όνομα οντότητας» ή «δεν ανήκει σε όνομα οντότητας». Στη συνέχεια, ο κύριος ταξινομητής εφαρμόζεται στα δεδομένα ανάπτυξης. Ένα ζεύγος ταξινομητών δευτέρου επιπέδου εκπαιδεύεται πάνω στα λάθη που κάνει ο κύριος ταξινομητής στα δεδομένα ανάπτυξης· σκοπός του ζεύγους των ταξινομητών δευτέρου επιπέδου είναι να μάθει να διορθώνει τα λάθη του κύριου ταξινομητή του συστήματος. Τέλος, το συνολικό σύστημα αξιολογείται στα δεδομένα αξιολόγησης (test data) εφαρμόζοντας πρώτα τον κύριο ταξινομητή και κατόπιν τον ταξινομητή δευτέρου επιπέδου.

3.1 Ιδιότητες του συστήματος

Παρακάτω περιγράφονται οι ιδιότητες που υποστηρίζονται από το σύστημα.

3.1.1 Μορφολογικές ιδιότητες

Αρχικά θα περιγράψουμε τις μορφολογικές ιδιότητες που χρησιμοποιήσαμε. Οι ιδιότητες αυτές υπολογίζονται κάθε φορά για μία συγκεκριμένη λεκτική μονάδα, που μπορεί να είναι είτε η τρέχουσα λεκτική μονάδα *w*, για την οποία προσπαθούμε να αποφασίσουμε αν ανήκει σε κάποια κατηγορία ονομάτων (και ποια) ή όχι, είτε οποιαδήποτε γειτονική λεκτική μονάδα της *w*, μέχρι και τρεις λεκτικές μονάδες πριν και τρεις μετά την *w*. Ως λεκτική μονάδα θεωρούμε εν γένει κάθε ομάδα χαρακτήρων που χωρίζεται από τις υπόλοιπες με κενό. Μια λεκτική μονάδα μπορεί να αποτελείται από γράμματα, αριθμητικά ψηφία, σημεία στίξης και άλλους χαρακτήρες. Συναντάμε συχνά, όμως, λεκτικές μονάδες που θα θέλαμε να διαιρεθούν σε μικρότερες. Στο παράδειγμα «… ο κ. Κώστας Σημίτης, που έφτασε …» ο χαρακτήρας «,» δεν θα θέλαμε να θεωρείται μέρος της ίδιας λεκτικής μονάδας με το όνομα «Σημίτης» και γι' αυτό θα πρέπει να τις χωρίσουμε σε δύο ξεχωριστές λεκτικές μονάδες («Σημίτης» και «,»), σε αντίθεση με το «κ.» που αντιμετωπίζεται ως μία λεκτική μονάδα. Ο τρόπος διαίρεσης λεκτικών μονάδων αυτού του είδους σε μικρότερες περιγράφεται στην ενότητα 4.1.

[6] Χρησιμοποιούμε την υλοποίηση του αλγορίθμου Μέγιστης Εντροπίας του Πανεπιστημίου Stanford (βλ. http://nlp.stanford.edu/software/classifier.shtml).

Οι μορφολογικές ιδιότητες που χρησιμοποιήσαμε είναι οι εξής έντεκα και οι τιμές τους είναι *αληθές* (1) ή *ψευδές* (0).

1. Αρχίζει η λεκτική μονάδα με κεφαλαίο γράμμα;
   Πολλές φορές τα ονόματα οντοτήτων αρχίζουν με κεφαλαίο γράμμα, όπως αυτά των ονομάτων προσώπων, οργανισμών και τοποθεσιών.
2. Η λεκτική μονάδα αποτελείται μόνο από κεφαλαία γράμματα, εξαιρουμένων σημείων στίξης και αριθμητικών ψηφίων;
   Μία λεκτική μονάδα της οποίας όλα τα γράμματα είναι κεφαλαία μάλλον αξίζει ιδιαίτερης προσοχής. Εάν περιέχεται οποιουδήποτε είδους σημείο στίξης στη λεκτική μονάδα, όπως μια παύλα ή μια τελεία, ή κάποιο αριθμητικό ψηφίο, αυτά αγνοούνται κατά τον υπολογισμό της τιμής της ιδιότητας, καθώς εξετάζονται μόνο οι χαρακτήρες που είναι γράμματα.
3. Η λεκτική μονάδα αποτελείται μόνο από μικρά γράμματα, εξαιρουμένων σημείων στίξης και αριθμητικών ψηφίων;
   Λεκτικές μονάδες των οποίων όλα τα γράμματα είναι μικρά συχνά δεν αποτελούν ονόματα κάποιων κατηγοριών (π.χ. ονόματα προσώπων). Κατά τον υπολογισμό της ιδιότητας αυτής εξετάζονται μόνο οι χαρακτήρες που αποτελούν γράμματα, ενώ οποιοιδήποτε άλλοι αγνοούνται.
4. Η λεκτική μονάδα περιέχει κάποιο μη αρχικό κεφαλαίο γράμμα και δεν αποτελείται εξολοκλήρου από κεφαλαία γράμματα;
   Η ιδιότητα αυτή εντοπίζει λεκτικές μονάδες που περιέχουν κεφαλαία γράμματα χωρίς να εμπίπτουν στις περιπτώσεις των ιδιοτήτων 1 και 2.
5. Η λεκτική μονάδα αποτελείται μόνο από αριθμητικά ψηφία;
   Λεκτικές μονάδες αυτού του είδους ενδέχεται να είναι π.χ. χρηματικά ποσά, χρονολογίες κλπ.
6. Η λεκτική μονάδα περιέχει κάποιο αριθμητικό ψηφίο μαζί με γράμματα ή/και σημεία στίξης;
   Η ιδιότητα αυτή ξεχωρίζει τις λεκτικές μονάδες που δεν περιέχουν μόνο αριθμητικά ψηφία, όπως στην περίπτωση της προηγούμενης ιδιότητας, αλλά επίσης περιέχουν γράμματα ή/και κάποια σημεία στίξης.
7. Η λεκτική μονάδα είναι σημείο στίξης;
   Συγκεκριμένα η ιδιότητα αυτή ελέγχει εάν η λεκτική μονάδα είναι ενός μόνο χαρακτήρα και εάν είναι κάποιος από τους χαρακτήρες «.», «,», «-» «;», «:», «?» ή «!».
8. Η λεκτική μονάδα περιέχει κάποιον από τους χαρακτήρες «;», «:», «?» ή «!» και δεν αποτελείται μόνο από έναν χαρακτήρα;
   Στην περίπτωση που η εξεταζόμενη λεκτική μονάδα αποτελείται από περισσότερους του ενός χαρακτήρες, η ιδιότητα αυτή εξετάζει εάν κάποιος από αυτούς είναι ο χαρακτήρας «;», «:», «?» ή «!».
9. Η λεκτική μονάδα περιέχει τον χαρακτήρα «.» και δεν αποτελείται μόνο από έναν χαρακτήρα;
   Ομοίως με την ιδιότητα 8, αλλά εξετάζεται μόνο η ύπαρξη του χαρακτήρα «.».
10. Η λεκτική μονάδα περιέχει τον χαρακτήρα «,» και δεν αποτελείται μόνο από έναν χαρακτήρα;

Ομοίως με την ιδιότητα 8, αλλά εξετάζεται μόνο η ύπαρξη του χαρακτήρα «,».

11. Η λεκτική μονάδα περιέχει τον χαρακτήρα «-» και δεν αποτελείται μόνο από έναν χαρακτήρα;
    Ομοίως με την ιδιότητα 8, αλλά εξετάζεται μόνο η ύπαρξη του χαρακτήρα «-».

3.1.2 Ιδιότητες λιστών

Το επόμενο είδος ιδιοτήτων που υποστηρίζεται περιλαμβάνει ιδιότητες που χρησιμοποιούν λίστες. Το σύστημα κατασκευάζει έντεκα διαφορετικές λίστες, αποκλειστικά από τα δεδομένα εκπαίδευσης. Υπενθυμίζεται ότι το σύστημα εκπαιδεύεται και χρησιμοποιείται ξεχωριστά για κάθε κατηγορία ονομάτων· τα περιεχόμενα των λιστών είναι εν γένει διαφορετικά για κάθε κατηγορία ονομάτων. Οι λίστες αυτές είναι οι ακόλουθες:

1. Λίστα ονομάτων οντοτήτων.
   Σε αυτή τη λίστα αποθηκεύονται λεκτικές μονάδες που εμφανίστηκαν τουλάχιστον μία φορά στα δεδομένα εκπαίδευσης ως ονόματα οντοτήτων (ή μέρη ονομάτων οντοτήτων) της ζητούμενης κατηγορίας.
2. Λίστα προηγούμενων λεκτικών μονάδων.
   Στη λίστα αυτή αποθηκεύονται λεκτικές μονάδες που εμφανίστηκαν αμέσως πριν από λεκτικές μονάδες της ζητούμενης κατηγορίας ονομάτων.
3. Λίστα επόμενων λεκτικών μονάδων.
   Η λίστα αυτή περιέχει λεκτικές μονάδες που εμφανίστηκαν αμέσως μετά από λεκτικές μονάδες της ζητούμενης κατηγορίας ονομάτων.
4. Λίστα προπροηγούμενων λεκτικών μονάδων
   Όπως η λίστα 2, αλλά για τις προπροηγούμενες λεκτικές μονάδες.
5. Λίστα μεθεπόμενων λεκτικών μονάδων.
   Όπως η λίστα 3, αλλά για τις μεθεπόμενες λεκτικές μονάδες.
6. Λίστες καταλήξεων μήκους δύο, τριών και τεσσάρων χαρακτήρων.
   Οι λίστες αυτές περιέχουν όλες τις καταλήξεις μήκους 2, 3 ή 4 χαρακτήρων των ονομάτων της ζητούμενης κατηγορίας που υπάρχουν στα δεδομένα εκπαίδευσης. Υπάρχουν τρεις ξεχωριστές λίστες, μία για κάθε μήκος καταλήξεων.
7. Λίστες προθεμάτων μήκους δύο, τριών και τεσσάρων χαρακτήρων.
   Όπως οι λίστες 6, αλλά για τους αρχικούς 2, 3 ή 4 χαρακτήρες.

Κατά την κατασκευή των παραπάνω λιστών, μετατρέπουμε τις λεκτικές μονάδες ώστε να περιέχουν μόνο μικρά γράμματα.

Κάθε εγγραφή της λίστας αξιολογείται βάσει κάποιου από τα τρία παρακάτω μέτρα. Η τιμή που προκύπτει από την αξιολόγηση της εγγραφής χρησιμοποιείται ως η τιμή της αντίστοιχης ιδιότητας, όταν το όρισμα της ιδιότητας είναι η εγγραφή. Τα μέτρα που υποστηρίζει το σύστημα είναι η ακρίβεια (precision), ανάκληση (recall) και το F-measure. Ο τρόπος υπολογισμού τους φαίνεται στους παρακάτω τύπους:

$$\text{precision} = \frac{\text{αριθμός εμφανίσεων της εγγραφής που ικανοποιούν τα κριτήρια της λίστας}}{\text{αριθμός εμφανίσεων της εγγραφής}}$$

$$\text{recall} = \frac{\text{αριθμός εμφανίσεων της εγγραφής που ικανοποιούν τα κριτήρια της λίστας}}{\text{αριθμός εμφανίσεων λεκτικών μονάδων της συγκεκριμένης κατηγορίας ονομάτων}}$$

$$\text{F - measure} = \frac{\left(\beta^2 + 1\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

Για παράδειγμα, στην περίπτωση της κατηγορίας των ονομάτων προσώπων, αν η λίστα 2 (των λεκτικών μονάδων που εμφανίζονται αμέσως πριν από ονόματα προσώπων) περιέχει ως εγγραφή τη λεκτική μονάδα «κύριος», η λεκτική αυτή μονάδα εμφανίζεται 10 φορές αμέσως πριν από ονόματα προσώπων στα δεδομένα εκπαίδευσης (ικανοποιώντας τα κριτήρια της λίστας), εμφανίζεται συνολικά 50 φορές στα δεδομένα εκπαίδευσης και τα δεδομένα εκπαίδευσης περιλαμβάνουν 100 λεκτικές μονάδες που είναι ονόματα προσώπων (ή μέρη ονομάτων προσώπων), τότε η ακρίβεια (precision) της εγγραφής «κύριος» της λίστας αυτής είναι 10/50 και η ανάκλησή της (recall) είναι 10/100.

Κατά τον υπολογισμό του F-measure θέτουμε β=1, τιμή που δίνει ίσο βάρος στην ακρίβεια και την ανάκληση. Προκύπτει έτσι το μέτρο που είναι γνωστό και ως F1:

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Όταν το σύστημα καλείται να δημιουργήσει το διάνυσμα χαρακτηριστικών της λεκτικής μονάδας *w*, περιλαμβάνονται στο διάνυσμα οι τιμές των ακόλουθων ιδιοτήτων που χρησιμοποιούν τις λίστες:

1. Η ακρίβεια (precision) της *w*, όπως προκύπτει από τη λίστα των ονομάτων οντοτήτων.
2. Το F-measure της προηγούμενης λεκτικής μονάδας της *w*, όπως προκύπτει από τη λίστα των προηγούμενων λεκτικών μονάδων.
3. Το F-measure της επόμενης λεκτικής μονάδας της *w*, όπως προκύπτει από τη λίστα των επόμενων λεκτικών μονάδων.
4. Το F-measure της λεκτικής μονάδας που βρίσκεται δύο θέσεις αριστερά της *w*, όπως προκύπτει από τη λίστα των προπροηγούμενων λεκτικών μονάδων.
5. Το F-measure της λεκτικής μονάδας που βρίσκεται δύο θέσεις δεξιά της *w*, όπως προκύπτει από τη λίστα των μεθεπόμενων λεκτικών μονάδων.
6. Το F-measure της κατάληξης δύο χαρακτήρων της *w*, όπως προκύπτει από τη λίστα των καταλήξεων μήκους δύο χαρακτήρων.
7. Το F-measure της κατάληξης τριών χαρακτήρων της *w* , όπως προκύπτει από τη λίστα των καταλήξεων μήκους τριών χαρακτήρων.
8. Το F-measure της κατάληξης τεσσάρων χαρακτήρων της *w*, όπως προκύπτει από τη λίστα των καταλήξεων μήκους τεσσάρων χαρακτήρων.

9. Το F-measure του προθέματος δύο χαρακτήρων της *w*, όπως προκύπτει από τη λίστα των προθεμάτων μήκους δύο χαρακτήρων.
10. Το F-measure του προθέματος τριών χαρακτήρων της *w*, όπως προκύπτει από τη λίστα των προθεμάτων μήκους τριών χαρακτήρων.
11. Το F-measure του προθέματος τεσσάρων χαρακτήρων της *w*, όπως προκύπτει από τη λίστα των προθεμάτων μήκους τεσσάρων χαρακτήρων.

Οι τιμές των παραπάνω ιδιοτήτων είναι πραγματικοί αριθμοί στο διάστημα [0,1). Αν κατά τον υπολογισμό κάποιας από τις τιμές των ιδιοτήτων δεν βρεθεί η αναζητούμενη εγγραφή στην αντίστοιχη λίστα, τότε η ιδιότητα παίρνει την τιμή 0.

Όπως παρατηρούμε παραπάνω, χρησιμοποιήσαμε το F-measure σε όλες τις ιδιότητες λιστών, εκτός από την ιδιότητα που χρησιμοποιεί τη λίστα με τα ονόματα οντοτήτων. Η ακρίβεια μιας λεκτικής μονάδας που περιέχεται στη λίστα αυτή μας δείχνει πόσο συχνά συναντάμε τη λεκτική μονάδα ως όνομα οντότητας (ή μέρος ονόματος οντότητας) επί του συνόλου των εμφανίσεών της. Η ανάκληση της λεκτικής μονάδας θα μας έδειχνε το ποσοστό που αυτή καταλαμβάνει επί του συνόλου των λεκτικών μονάδων που εμφανίστηκαν ως ονόματα οντοτήτων στα δεδομένα εκπαίδευσης, πληροφορία όχι και τόσο χρήσιμη.

3.1.3 Ιδιότητες που χρησιμοποιούν ετικέτες μερών του λόγου

Το επόμενο είδος ιδιοτήτων χρησιμοποιεί τις ετικέτες μερών του λόγου (part-of-speech tags) της τρέχουσας λέξης *w* και των τριών προηγούμενων και επόμενων λεκτικών μονάδων. Υπάρχουν συνολικά επτά ιδιότητες αυτού του είδους, κάθε μία από τις οποίες παίρνει ως τιμή την ετικέτα της αντίστοιχης λεκτικής μονάδας· για την ακρίβεια, οι τιμές είναι ακέραιοι αριθμοί που παριστάνουν τις ετικέτες. Οι ετικέτες που υποστηρίζονται από το σύστημα φαίνονται στον παρακάτω πίνακα.

POS tag	Σημασία	Περιγραφή
N	Noun	Ουσιαστικό
V	Verb	Ρήμα
Adj	Adjective	Επίθετο
Adv	Adverb	Επίρρημα
Art	Article	Άρθρο
Prep	Preposition	Πρόθεση
Pron	Pronoun	Αντωνυμία
Conj	Conjunction	Σύνδεσμος
Punc	Punctuation	Σημείο στίξης
Other	Other	Οτιδήποτε δεν ανήκει στα παραπάνω

Είδη μερών του λόγου που υποστηρίζει το σύστημα

Η επισημείωση των λεκτικών μονάδων με ετικέτες μερών του λόγου είναι δουλειά ενός εργαλείου επισημείωσης μερών του λόγου (POS tagger). Στο σύστημά μας δεν έχει

ενσωματωθεί κάποιο τέτοιο εργαλείο, αλλά το σύστημα μπορεί να αξιοποιήσει τις παραπάνω ετικέτες, αν έχουν προστεθεί στα κείμενα από ένα εξωτερικό εργαλείο επισημείωσης μερών του λόγου. Αν δεν έχουν προστεθεί τέτοιες ετικέτες στα κείμενα, οι ιδιότητες αυτής της ενότητας δεν χρησιμοποιούνται.

3.1.4 Ιδιότητα συντελεστή συσχέτισης

Τέλος, στο σύστημα προσθέσαμε μία ιδιότητα η οποία βασίζεται στο Συντελεστή Συσχέτισης (Correlation Coefficient, *CC*) [11]. Ο συντελεστής αυτός υπολογίζεται για την τρέχουσα λεκτική μονάδα *w*, τις τρεις προηγούμενες λεκτικές μονάδες της *w* και τις τρεις επόμενες. Ο τρόπος υπολογισμού του *CC* για μια λεκτική μονάδα *t* φαίνεται παρακάτω:

$$CC = \frac{\left(N_{r+}N_{n-} - N_{r-}N_{n+}\right)\sqrt{N}}{\sqrt{\left(N_{r+} + N_{r-}\right)\left(N_{n+} + N_{n-}\right)\left(N_{r+} + N_{n+}\right)\left(N_{r-} + N_{n-}\right)}}$$

N_{r+} είναι ο αριθμός των προτάσεων εκπαίδευσης οι οποίες περιέχουν όνομα οντότητας της ζητούμενης κατηγορίας ονομάτων και στις οποίες εμφανίστηκε η *t*. N_{r-} είναι ο αριθμός των προτάσεων οι οποίες περιέχουν όνομα οντότητας και στις οποίες η *t* δεν εμφανίστηκε. Ομοίως τα N_{n+} και N_{n-} είναι ο αριθμός των προτάσεων που δεν περιέχουν όνομα οντότητας και στις οποίες η λεκτική μονάδα *t* εμφανίστηκε ή δεν εμφανίστηκε αντίστοιχα. *N* είναι ο συνολικός αριθμός των προτάσεων εκπαίδευσης.

Οι τιμές του συντελεστή είναι πραγματικοί αριθμοί, θετικοί ή αρνητικοί. Πιο θετικές τιμές αντιστοιχούν σε μεγαλύτερη βεβαιότητα ότι η *t* συνοδεύει όνομα οντότητας, ενώ πιο αρνητικές τιμές αντιστοιχούν σε απουσία βεβαιότητας. Ο συντελεστής στην πραγματικότητα προ-υπολογίζεται, κατά την εκπαίδευση, για όλες τις λεκτικές μονάδες που εμφανίζονται έστω και μία φορά στα δεδομένα εκπαίδευσης. Λεκτικές μονάδες που δεν έχουν εμφανιστεί στα δεδομένα εκπαίδευσης θεωρούμε ότι έχουν $CC = 0$.

Τελικά στο σύστημα χρησιμοποιείται η ιδιότητα που έχει ως τιμή το μέγιστο *CC* των λεκτικών μονάδων του παραθύρου της τρέχουσας λέξης *w*. Ως «παράθυρο» εννοούμε την ίδια την *w*, τις προηγούμενες τρεις λεκτικές μονάδες της *w* και τις επόμενες τρεις. Κατά τον υπολογισμό της τιμής της ιδιότητας, εξαιρούνται από το παράθυρο λεκτικές μονάδες που έχουν εμφανιστεί στα δεδομένα εκπαίδευσης ως ονόματα οντοτήτων της ζητούμενης κατηγορίας τουλάχιστον μία φορά.

3.2 Αυτόματη επιλογή ιδιοτήτων

Οι υποστηριζόμενες ιδιότητες που περιγράφηκαν στις προηγούμενες ενότητες είναι 96. Από αυτές, το σύστημα επιλέγει τις πιο χρήσιμες χρησιμοποιώντας μια παραλλαγή της αναζήτησης Beam Search.

Πιο συγκεκριμένα, η αναζήτηση γίνεται σε ένα χώρο καταστάσεων στον οποίο κάθε κατάσταση είναι ένα υποσύνολο των υποστηριζόμενων ιδιοτήτων. Η αναζήτηση προχωρά είτε αφαιρώντας είτε προσθέτοντας ιδιότητες. Στην πρώτη περίπτωση, η αρχική κατάσταση περιέχει όλες τις υποστηριζόμενες ιδιότητες (96). Οι καταστάσεις-παιδιά της

αρχικής κατάστασης αντιστοιχούν σε όλα τα υποσύνολα 95 υποστηριζόμενων ιδιοτήτων, δηλαδή είναι όλες οι καταστάσεις που προκύπτουν αφαιρώντας μία ιδιότητα από την αρχική κατάσταση. Κάθε κατάσταση-παιδί αξιολογείται υπολογίζοντας το F-measure που επιτυγχάνει ο ταξινομητής Μέγιστης Εντροπίας στα δεδομένα ανάπτυξης, όταν εκπαιδεύεται στα δεδομένα εκπαίδευσης και χρησιμοποιεί τις ιδιότητες της κατάστασης-παιδιού. Μετά την αξιολόγηση των καταστάσεων-παιδιών, επιλέγονται τα *k* καλύτερα (με τα υψηλότερα F-measure) από αυτά, παράγονται και αξιολογούνται τα δικά τους (μόνο) παιδιά, επιλέγονται τα *k* καλύτερα κ.ο.κ. Στη δεύτερη περίπτωση, όπου η αναζήτηση προχωρά προσθέτοντας ιδιότητες, η αρχική κατάσταση δεν περιέχει καμία ιδιότητα. Οι καταστάσεις-παιδιά της περιέχουν η κάθε μία μόνο μία (διαφορετική) ιδιότητα. Οι καταστάσεις-παιδιά αξιολογούνται υπολογίζοντας πάλι τα αντίστοιχα F-measure στα δεδομένα ανάπτυξης. Κατόπιν επιλέγονται τα *k* καλύτερα παιδιά, παράγονται και αξιολογούνται τα δικά τους (μόνο) παιδιά κ.ο.κ.

Η διαφορά με τον κλασικό αλγόριθμο Beam Search έγκειται στο ότι στη δική μας περίπτωση η αναζήτηση συνεχίζεται πάντα μέχρι να μην υπάρχουν επόμενες καταστάσεις-παιδιά (π.χ. μέχρι να έχουν προστεθεί όλες οι υποστηριζόμενες ιδιότητες, στην περίπτωση όπου η αναζήτηση προχωρά προσθέτοντας ιδιότητες). Κατά τη διάρκεια της αναζήτησης αποθηκεύουμε την κατάσταση που έχει μέχρι στιγμής την καλύτερη επίδοση F-measure στα δεδομένα ανάπτυξης. Οι ιδιότητες αυτής της καλύτερης κατάστασης επιστρέφονται τελικά ως το καλύτερο υποσύνολο ιδιοτήτων.

3.3 Ταξινομητές δεύτερου επιπέδου

Στη βιβλιογραφία πολλές φορές συναντούμε συστήματα που χρησιμοποιούν χειρωνακτικούς κανόνες ως ένα τελευταίο στάδιο, προκειμένου να διορθώνουν λάθη ταξινομητών που εκπαιδεύονται μέσω μηχανικής μάθησης. Για παράδειγμα, στο σύστημα των Michailidis κ.ά. [12] χρησιμοποιούνται κανόνες όπως αυτοί του παρακάτω πίνακα. Στο συγκεκριμένο σύστημα (και πολλά άλλα) οι λεκτικές μονάδες κατατάσσονται ως B-C όταν είναι οι πρώτες λεκτικές μονάδες ενός ονόματος κατηγορίας C, ως I-C όταν είναι άλλες λεκτικές μονάδες ονόματος κατηγορίας C και ως O διαφορετικά.

Ακολουθία επισημείωσης ταξινομητή	Διώρθωση σε
B-MISC O I-MISC	B-MISC I-MISC I-MISC
B-PER O I-PER	B-PER I-PER I-PER
B-ORG I-ORG O O I-ORG	B-ORG I-ORG I-ORG I-ORG I-ORG
…	…

Παραδείγματα διορθωτικών κανόνων

Το μειονέκτημα των χειρωνακτικών κανόνων είναι πως είναι χρονοβόρα η ενημέρωσή τους όταν μεταφέρουμε το σύστημα σε άλλες γλώσσες, κείμενα άλλων θεματικών περιοχών με διαφορετικές κατηγορίες ονομάτων κλπ. Στο δικό μας σύστημα, αντί χειρωνακτικών κανόνων διόρθωσης, χρησιμοποιείται ένα ζεύγος ταξινομητών (πάλι Μέγιστης Εντροπίας) δευτέρου επιπέδου, που επιχειρεί να μάθει να διορθώνει τα λάθη του κύριου ταξινομητή του συστήματος. Ο πρώτος ταξινομητής του ζεύγους εκπαιδεύεται στις λεκτικές μονάδες των δεδομένων ανάπτυξης που ο κύριος ταξινομητής κατέταξε λανθασμένα ως ονόματα οντοτήτων (false positives). Καλείται δε να εκφέρει γνώμη κατά την αξιολόγηση (ή όταν το σύστημα προσπαθεί να εντοπίσει ονόματα σε μη επισημειωμένα κείμενα) μόνο για λεκτικές μονάδες που ο κύριος ταξινομητής κατέταξε ως ονόματα· αν αποκριθεί ότι μια λεκτική μονάδα κατετάγη λανθασμένα ως όνομα από τον κύριο ταξινομητή, η λεκτική μονάδα επισημειώνεται ως μη όνομα. Αντίστοιχα, ο δεύτερος ταξινομητής του ζεύγους εκπαιδεύεται στις λεκτικές μονάδες των δεδομένων ανάπτυξης που ο κύριος ταξινομητής θα έπρεπε να είχε κατατάξει ως ονόματα οντοτήτων, αλλά δεν το έκανε (false negatives). Καλείται δε να εκφέρει γνώμη κατά την αξιολόγηση (ή όταν το σύστημα προσπαθεί να εντοπίσει ονόματα σε μη επισημειωμένα κείμενα) μόνο για λεκτικές μονάδες που ο κύριος ταξινομητής *δεν* κατέταξε ως ονόματα· αν αποκριθεί ότι μια λεκτική μονάδα θα έπρεπε να είχε καταταγεί ως όνομα, η λεκτική μονάδα επισημειώνεται ως όνομα. Παρόμοιοι ταξινομητές είχαν χρησιμοποιηθεί και σε προηγούμενες εργασίες [13,14].

Οι ταξινομητές δευτέρου επιπέδου κάθε κατηγορίας ονομάτων χρησιμοποιούν τις παρακάτω Boolean ιδιότητες, που είναι κοινές και για τους δύο ταξινομητές του ζεύγους:

1. Η τρέχουσα λεκτική μονάδα *w* αρχίζει με κεφαλαίο γράμμα;
2. Η *w* αρχίζει με μικρό γράμμα;
   Η ιδιότητα αυτή δεν είναι απλά το λογικό συμπλήρωμα της προηγούμενης, αφού μπορεί μια λεκτική μονάδα να ξεκινά π.χ. με αριθμητικό ψηφίο, οπότε και οι δύο ιδιότητες θα έχουν τιμή *false* (0).
3. Η *w* υπάρχει στην λίστα με τα ονόματα οντοτήτων της ζητούμενης κατηγορίας ονομάτων;
4. Η προηγούμενη λεκτική μονάδα της *w* έχει καταταγεί ως όνομα οντότητας;
5. Η προπροηγούμενη λεκτική μονάδα της *w* έχει καταταγεί ως όνομα οντότητας;
6. Η επόμενη λεκτική μονάδα της *w* έχει καταταγεί ως όνομα οντότητας;
7. Η μεθεπόμενη λεκτική μονάδα της *w* έχει καταταγεί ως όνομα;
8. Η *w* είναι σημείο στίξης;
   Συγκεκριμένα ελέγχουμε εάν η *w* αποτελείται από ένα μόνο χαρακτήρα και εάν αυτός είναι ένας από τους ‘.’, ‘,’, ‘(’ ή ‘)’.
9. Η προηγούμενη λεκτική μονάδα της *w* είναι ο χαρακτήρας ‘.’;
10. Η προηγούμενη λεκτική μονάδα της *w* είναι ο χαρακτήρας ‘-’;

Κεφάλαιο 4: Πειράματα και αποτελέσματα

Στο κεφάλαιο αυτό περιγράφονται τα πειράματα που έγιναν κατά τη διάρκεια αυτής της εργασίας, καθώς και τα αποτελέσματά τους.

4.1 Συλλογές κειμένων

Στα πειράματα της εργασίας χρησιμοποιήθηκαν τέσσερις συλλογές κειμένων. Το σύστημα της εργασίας δοκιμάστηκε συνολικά σε τέσσερις διαφορετικές γλώσσες και σε εννέα διαφορετικά είδη ονομάτων οντοτήτων.

Η πρώτη συλλογή κειμένων είναι στην ισπανική γλώσσα και δημιουργήθηκε για τους σκοπούς του διαγωνισμού CoNLL-2002 (βλ. και ενότητα 2.3). Περιέχει κείμενα ειδήσεων στα οποία έχουν επισημειωθεί τα ονόματα προσώπων, οργανισμών, τοποθεσιών, καθώς και ονόματα οντοτήτων που δεν μπορούν να καταταγούν σε κάποια από τις τρεις προηγούμενες κατηγορίες· τέτοια ονόματα είναι διευθύνσεις ιστοσελίδεων, όροι όπως «Internet», «Champions League» κ.ά. Τα ίδια ακριβώς ισχύουν και για τη δεύτερη συλλογή, που δημιουργήθηκε για τον ίδιο διαγωνισμό, με τη διαφορά ότι τα κείμενα της δεύτερης συλλογής είναι στα ολλανδικά. Μία πρόσθετη διαφορά είναι ότι στη δεύτερη συλλογή οι λεκτικές μονάδες των κειμένων είναι επισημειωμένες με ετικέτες μερών του λόγου, κάτι που δεν ισχύει για την πρώτη συλλογή. Τα αρχεία που περιέχουν τα κείμενα των δύο συλλογών έχουν σε κάθε γραμμή μόνο μία λεκτική μονάδα και πληροφορίες για αυτήν (π.χ. ετικέτα μέρους του λόγου, την κατηγορία ονομάτων στην οποία ανήκει η λεκτική μονάδα κλπ.). Οι παρεμβάσεις που έγιναν από πλευράς μας στις παραπάνω συλλογές ήταν η αφαίρεση των κενών γραμμών που υπήρχαν, η αφαίρεση γραμμών που περιείχαν διάφορα άλλα στοιχεία, όπως ενδείξεις για την εναλλαγή των ειδησεογραφικών άρθρων, και γενικότερα η αφαίρεση γραμμών που δεν αφορούσαν λεκτικές μονάδες.

Η τρίτη συλλογή κειμένων είναι στα ελληνικά και είχε χρησιμοποιηθεί στη διάρκεια προηγούμενων εργασιών [13,15,16]. Αποτελείται από τετρακόσια ειδησεογραφικά άρθρα των εφημερίδων «ΤΑ ΝΕΑ» και «ΤΟ ΒΗΜΑ», στα οποία έχουν επισημειωθεί τα ονόματα προσώπων, οργανισμών και τοποθεσιών, καθώς και οι χρονικές εκφράσεις (π.χ. ημερομηνίες)· οι χρονικές εκφράσεις δεν χρησιμοποιήθηκαν στα πειράματα αυτής της εργασίας. Τα κείμενα της συλλογής αυτής είναι σε μορφή XML. Για τους σκοπούς της εργασίας, μετατράπηκαν στη μορφή των αρχείων του διαγωνισμού CoNLL-2002, που περιγράφηκε παραπάνω. Επίσης, έγινε χειρωνακτική επισημείωση των χαρακτήρων «,» και «.». Επισημειώθηκε, δηλαδή, στα κείμενα η διαφορετική χρήση αυτών των σημείων στίξης ως μέρη λεκτικών μονάδων (π.χ. η τελεία της λεκτικής μονάδας «Κ.» στην περίπτωση του «Κ. Σημίτης», καθώς και το κόμμα στο ποσοστό «3,25%») ή ως αυτούσιες λεκτικές μονάδες, δηλαδή ως χαρακτήρων σημείων στίξης. Χωρίσαμε τυχαία τη συλογή αυτή σε εφτά μέρη ίσου μεγέθους, εκ των οποίων πέντε χρησιμοποιήθηκαν ως δεδομένα εκπαίδευσης, ένα ως δεδομένα ανάπτυξης και ένα ως δεδομένα ελέγχου.

Η τέταρτη συλλογή είναι στην αγγλική γλώσσα και χρησιμοποιήθηκε στο διαγωνισμό BioNLP/NLPBA 2004[7]. Αποτελείται από δύο χιλιάδες περιλήψεις βιοϊατρικών δημοσιεύσεων. Στα κείμενα αυτά είναι επισημειωμένες πέντε κατηγορίες

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

βιοϊατρικών ονομάτων οντοτήτων και συγκεκριμένα οι κατηγορίες «πρωτεΐνη», «DNA», «RNA», «cell type» και «cell line». Η δομή των κειμένων ήταν ίδια με αυτή των δύο πρώτων συλλογών και έγιναν οι αντίστοιχες αλλαγές. Στο διαγωνισμό αυτό δε διατέθηκαν δεδομένα ανάπτυξης. Ως εκ τούτου, αποκόψαμε περίπου το τελευταίο 20% των κειμένων εκπαίδευσης για να δημιουργήσουμε δεδομένα ανάπτυξης.

Στους παρακάτω πίνακες δίνονται περισσότερες πληροφορίες για τις τέσσερις συλλογές κειμένων. Οι γραμμές αντιστοιχούν στις κατηγορίες ονομάτων οντοτήτων που έχουν επισημειωθεί σε κάθε συλλογή. Οι στήλες αντιστοιχούν στα τμήματα των συλλογών. Οι πίνακες δείχνουν πόσα ονόματα οντοτήτων κάθε είδους έχουν επισημειωθεί σε κάθε τμήμα της συλλογής (αριστερά της καθέτου) και από πόσες συνολικά λεκτικές μονάδες αποτελούνται αυτά (δεξιά της καθέτου).

Ισπανικά	Training data	Development data	Test data
Person	4321/8224	1222/2081	735/1369
Location	4913/6804	984/1321	1084/1409
Organization	7390/12382	1700/3066	1400/2504
Miscellaneous	2173/5385	445/1099	339/896

Ολλανδικά	Training data	Development data	Test data
Person	4716/7601	703/1127	1098/1905
Location	3208/3676	479/543	774/823
Organization	2082/3282	686/1084	882/1433
Miscellaneous	3338/4743	748/963	1187/1597

Ελληνικά	Training data	Development data	Test data
Person	3019/4510	805/1256	991/1503
Location	2217/2684	765/970	600/695
Organization	2653/4249	828/1275	783/1202

Βιοϊατρικά	Training data	Development data	Test data
Protein	25806/46119	4463/9065	5067/9849
DNA	7446/19788	2088/5581	1056/2852
RNA	713/1796	238/685	118/305
Cell type	5708/13257	1010/2209	1921/4912
Cell line	3115/9181	715/2036	500/1489

4.2 Μέτρα αξιολόγησης

Για την αξιολόγηση των επιδόσεων του συστήματος εφαρμόσαμε τρία μέτρα που χρησιμοποιούνται πολύ συχνά στην αξιολόγηση συστημάτων αναγνώρισης ονομάτων οντοτήτων και γενικότερα στην επεξεργασία φυσικής γλώσσας, αλλά και την ανάκτηση πληροφοριών. Τα μέτρα αυτά είναι η ακρίβεια (precision), η ανάκληση (recall) και το F-measure. Ορίζονται για κάθε κατηγορία ονομάτων ως εξής:

$$\text{precision} = \frac{\text{πλήθος λεκτικών μονάδων που κατετάγησαν σωστά ως ονόματα της κατηγορίας}}{\text{πλήθος λεκτικών μονάδων που κατετάγησαν ως ονόματα της κατηγορίας}}$$

$$\text{recall} = \frac{\text{πλήθος λεκτικών μονάδων που κατετάγησαν σωστά ως ονόματα της κατηγορίας}}{\text{πλήθος λεκτικών μονάδων ονομάτων της κατηγορίας}}$$

$$\text{F - measure} = \frac{\left(\beta^2 + 1\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

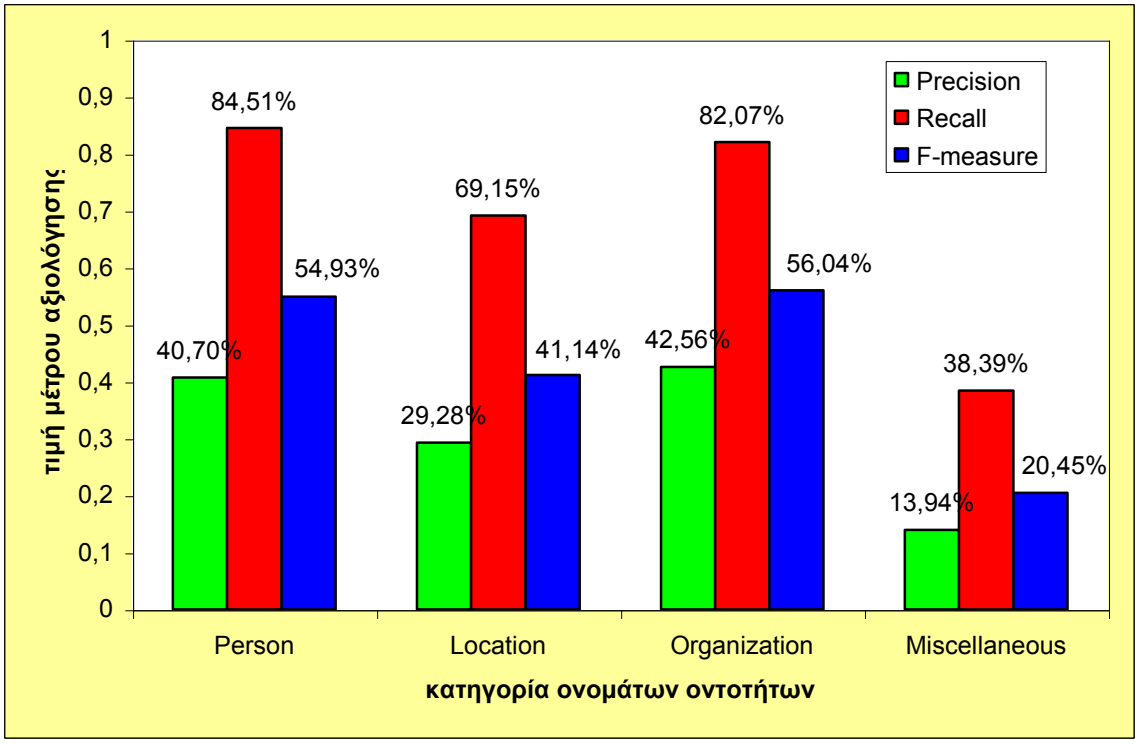
Στα πειράματα που ακολουθούν έχουμε θέσει $\beta = 1$, κάτι που δίνει ίσο βάρος στην ακρίβεια και την ανάκληση. Προκύπτει έτσι το μέτρο που είναι γνωστό και ως F1:

$$\text{F1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

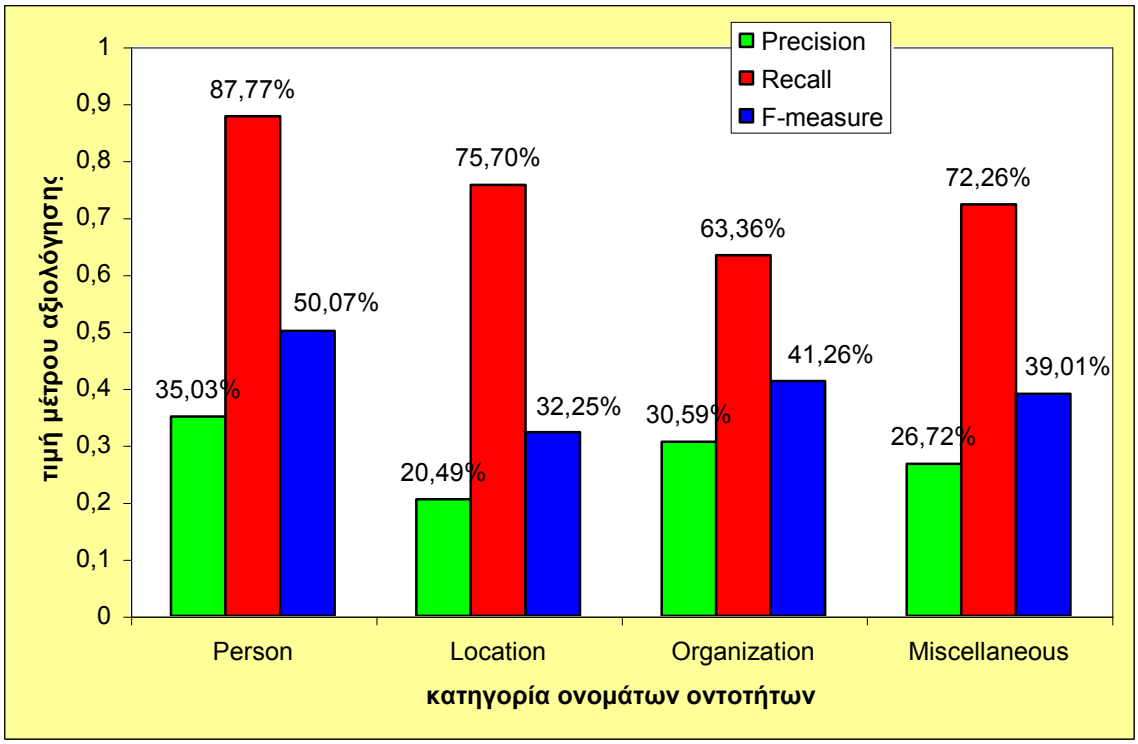
Παρακάτω, όποτε αναφέρεται το F-measure εννοείται το F1.

4.3 Βασική μορφή του συστήματος

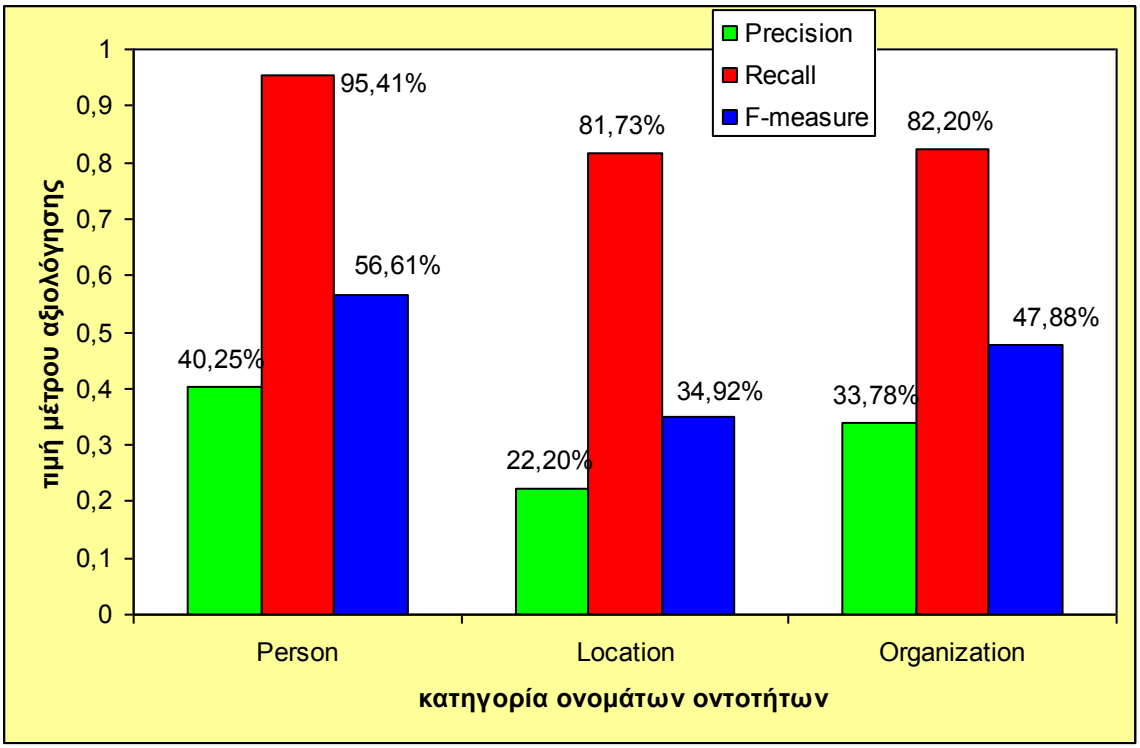
Αρχικά παραθέτουμε τα πειραματικά αποτελέσματα μιας βασικής μορφής του συστήματος της εργασίας, η οποία χρησιμοποιεί μόνο μορφολογικές ιδιότητες, χωρίς περαιτέρω επιλογή ιδιοτήτων, και μόνο τον κύριο ταξινομητή. Τα αποτελέσματα φαίνονται στα παρακάτω διαγράμματα. Σε όλες τις περιπτώσεις το σύστημα πέτυχε υψηλότερη ανάκληση από ό,τι ακρίβεια. Αυτό οφείλεται στο ότι με τις μορφολογικές πληροφορίες που έχει (μόνο) στη διάθεσή του το σύστημα, δεν μπορεί να ξεχωρίσει αρκετά καλά τις διαφορετικές κατηγορίες ονομάτων, με αποτέλεσμα να κατατάσσει π.χ. ως ονόματα προσώπων και πολλά ονόματα τοποθεσιών και αντίστροφα.



Διάγραμμα 1. Επιδόσεις της **βασικής μορφής** του συστήματος της εργασίας (μορφολογικές ιδιότητες, ένα επίπεδο ταξινομητών, χωρίς επιλογή ιδιοτήτων) στην **ισπανική συλλογή κειμένων**.



Διάγραμμα 2. Επιδόσεις της **βασικής μορφής** του συστήματος της εργασίας (μορφολογικές ιδιότητες, ένα επίπεδο ταξινομητών, χωρίς επιλογή ιδιοτήτων) στην **ολλανδική συλλογή κειμένων**.

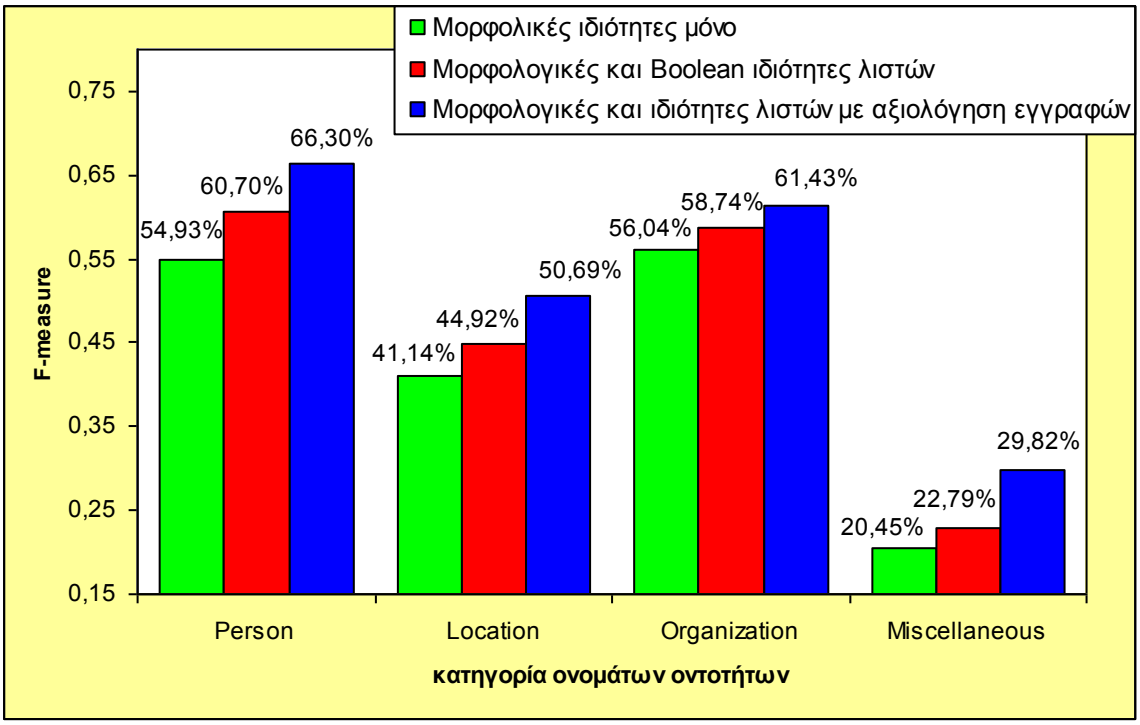


Διάγραμμα 3. Επιδόσεις της **βασικής μορφής** του συστήματος της εργασίας (μορφολογικές ιδιότητες, ένα επίπεδο ταξινομητών, χωρίς επιλογή ιδιοτήτων) στην **ελληνική συλλογή κειμένων.**

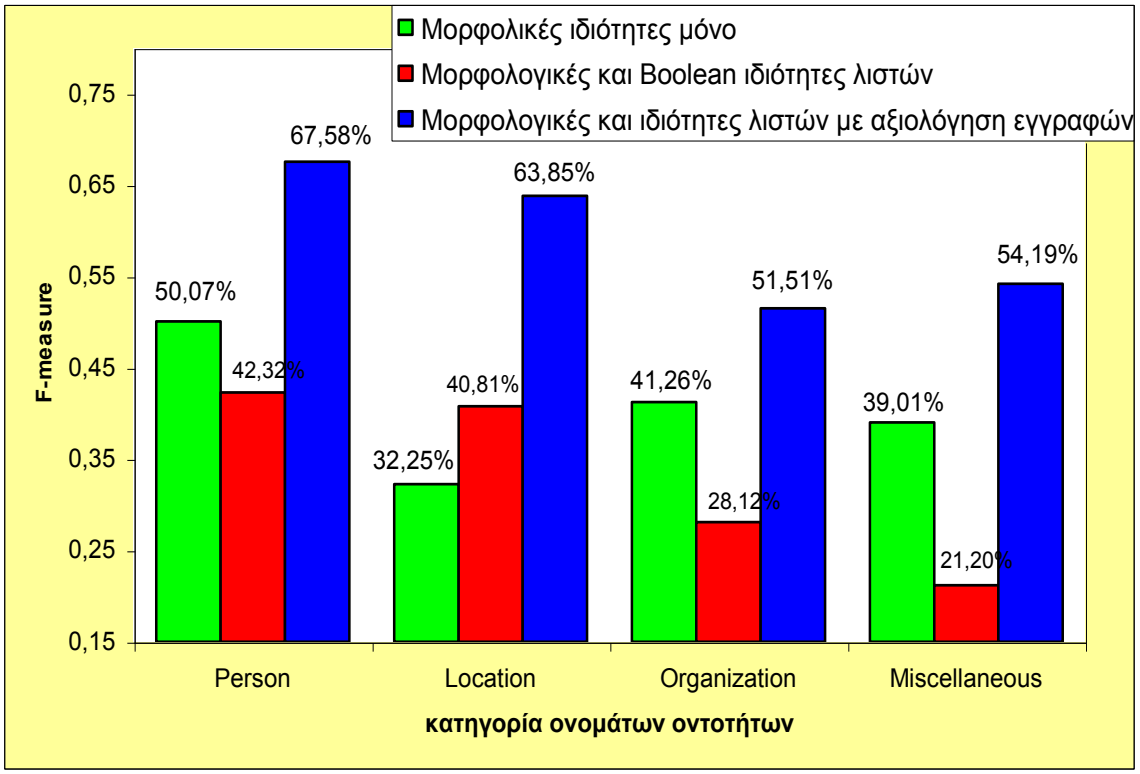
4.4 Προσθήκη ιδιοτήτων λιστών

Στα παρακάτω διαγράμματα φαίνονται οι επιδόσεις του συστήματος όταν προστεθούν οι ιδιότητες που χρησιμοποιούν λίστες, όπως αυτές περιγράφηκαν στην παράγραφο 3.1.2. Δοκιμάσαμε τις ιδιότητες αυτές με δύο διαφορετικούς τρόπους. Στην πρώτη περίπτωση, οι ιδιότητες ήταν Boolean και έδειχναν αν υπήρχε ή όχι η αντίστοιχη εγγραφή στην αντίστοιχη λίστα. Στη δεύτερη περίπτωση, οι τιμές των ιδιοτήτων ήταν τα αποτελέσματα της αξιολόγησης των αντίστοιχων εγγραφών (τιμές ακρίβειας ή F-measure) των λιστών, όπως εξηγήθηκε στην ενότητα 3.1.2.

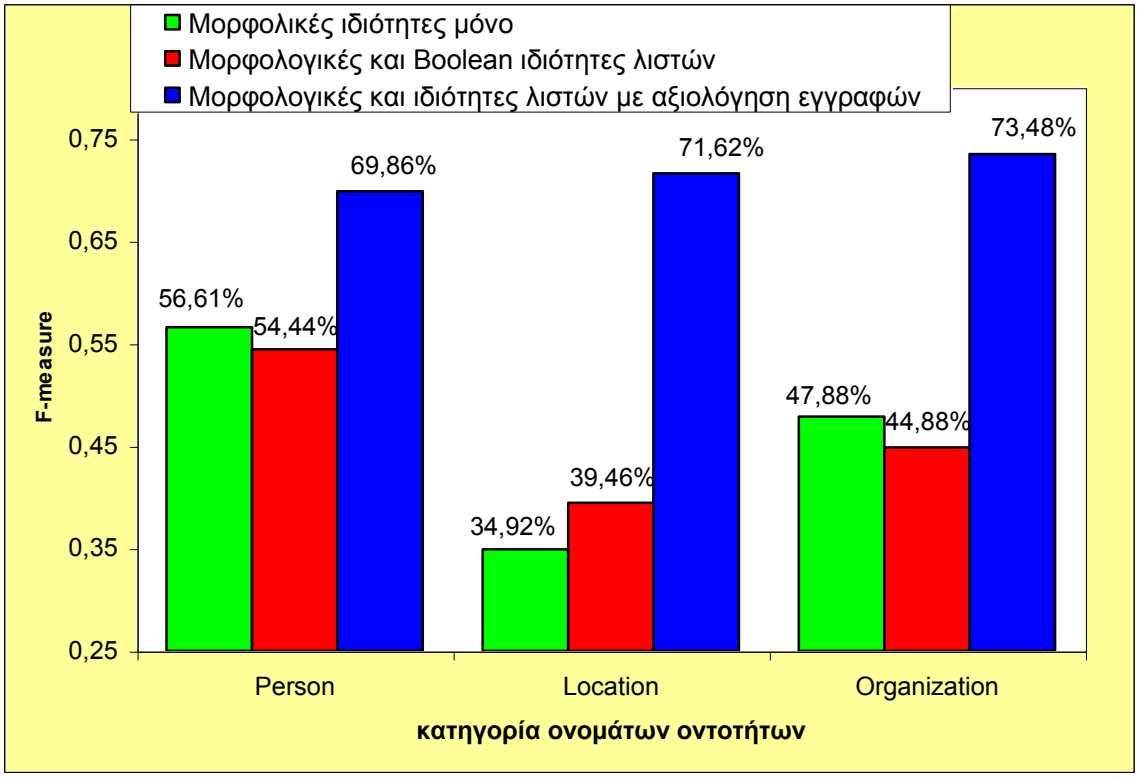
Όπως προκύπτει από τα διαγράμματα, σε όλες τις περιπτώσεις τα καλύτερα αποτελέσματα (και με σημαντική διαφορά) επιτυγχάνονται με την προσθήκη ιδιοτήτων λιστών και αξιολόγηση των εγγραφών των λιστών. Χωρίς την αξιολόγηση των εγγραφών, η προσθήκη ιδιοτήτων λιστών δεν βελτιώνει πάντα τα αποτελέσματα. Αυτό οφείλεται στο ότι, για παράδειγμα, στην περίπτωση της λίστας των προηγούμενων λεκτικών μονάδων, προστίθενται στη λίστα όλες οι λεκτικές μονάδες που βρέθηκαν στα δεδομένα εκπαίδευσης έστω και μόνο μία φορά πριν από όνομα οντότητας και τους δίνεται το ίδιο βάρος (Boolean τιμή *true*) με άλλες λεκτικές μονάδες που «προβλέπουν» πολύ πιο αξιόπιστα την εμφάνιση ονομάτων οντοτήτων.



Διάγραμμα 4. F-measure του συστήματος της εργασίας με και χωρίς **προσθήκη λιστών** στην **ισπανική συλλογή κειμένων.**

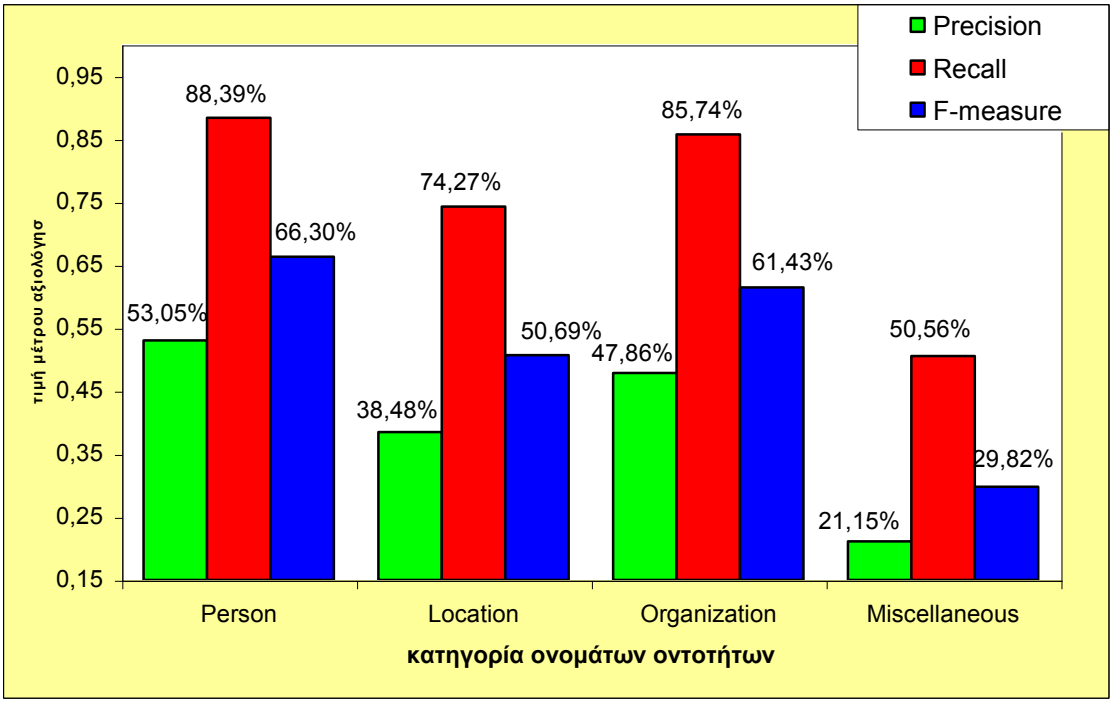


Διάγραμμα 5. F-measure του συστήματος της εργασίας με και χωρίς **προσθήκη λιστών** στην **ολλανδική συλλογή κειμένων.**

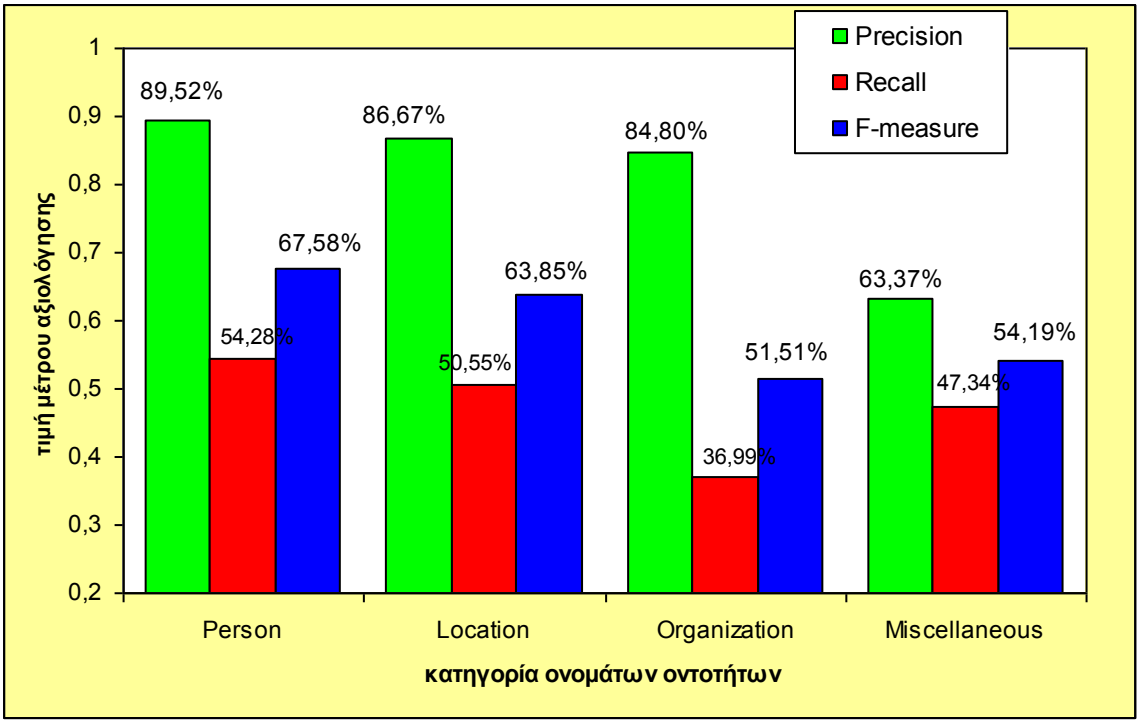


Διάγραμμα 6. F-measure του συστήματος της εργασίας με και χωρίς **προσθήκη λιστών** στην **ελληνική συλλογή κειμένων.**

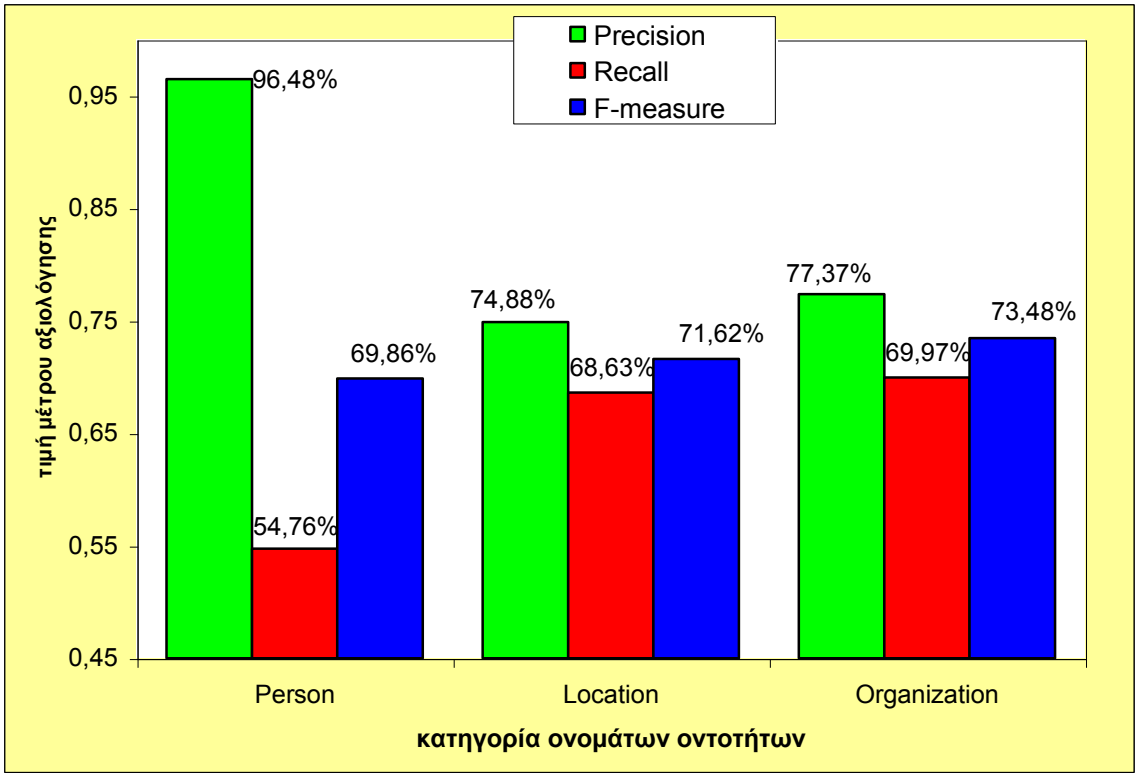
Για λόγους πληρότητας, παρατίθενται και διαγράμματα με αποτελέσματα ακρίβειας, ανάκλησης και F-measure, για την περίπτωση όπου χρησιμοποιούνται μαζί οι μορφολογικές ιδιότητες και οι λίστες ιδιοτήτων με αξιολόγηση εγγραφών.



Διάγραμμα 7. Επιδόσεις του συστήματος της εργασίας με χρήση **μορφολογικών ιδιοτήτων και λιστών** με αξιολόγηση εγγραφών στην **ισπανική συλλογή κειμένων**.



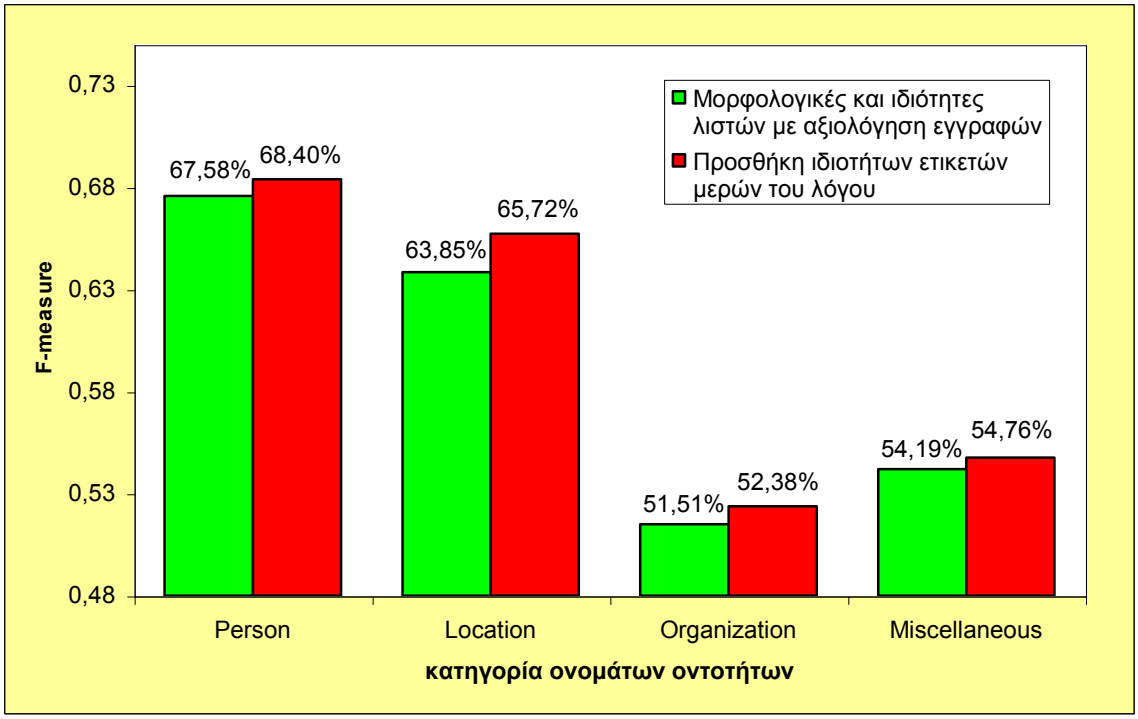
Διάγραμμα 8. Επιδόσεις του συστήματος της εργασίας με χρήση **μορφολογικών ιδιοτήτων και λιστών** με αξιολόγηση εγγραφών στην **ολλανδική συλλογή κειμένων**.



Διάγραμμα 9. Επιδόσεις του συστήματος της εργασίας με χρήση **μορφολογικών ιδιοτήτων και λιστών** με αξιολόγηση εγγραφών στην **ελληνική συλλογή κειμένων.**

4.5 Προσθήκη ιδιοτήτων ετικετών μερών του λόγου

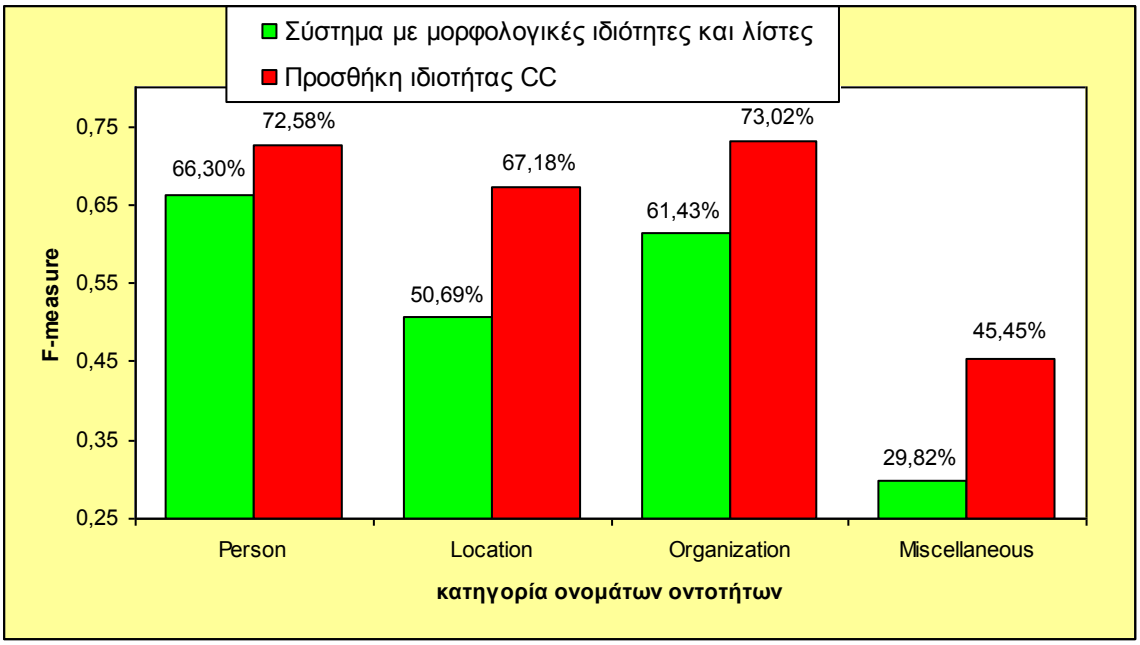
Η μόνη συλλογή κειμένων στην οποία έγιναν πειράματα με ιδιότητες ετικετών μερών του λόγου είναι η συλλογή ολλανδικών κειμένων. Στο διάγραμμα που ακολουθεί φαίνεται η σύγκριση των αποτελεσμάτων του καλύτερου συστήματος της προηγούμενης ενότητας (μορφολογικές ιδιότητες και ιδιότητες λιστών με αξιολόγηση εγγραφών) με τα αποτελέσματα που προκύπτουν όταν προστεθούν στο ίδιο σύστημα οι ιδιότητες ετικετών μερών του λόγου. Παρατηρούμε μικρή βελτίωση των αποτελεσμάτων σε όλες τις περιπτώσεις.



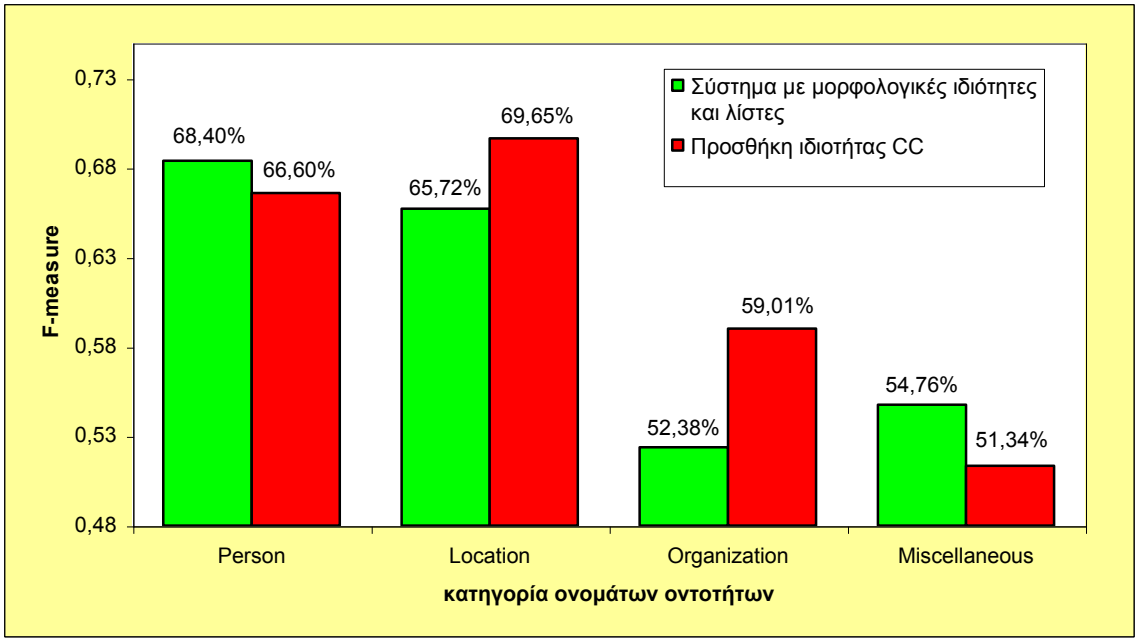
Διάγραμμα 10. F-measure του συστήματος της εργασίας στην **ολλανδική συλλογή** με και χωρίς **ιδιότητες ετικετών μερών του λόγου**.

4.6 Προσθήκη ιδιότητας συντελεστή συσχέτισης

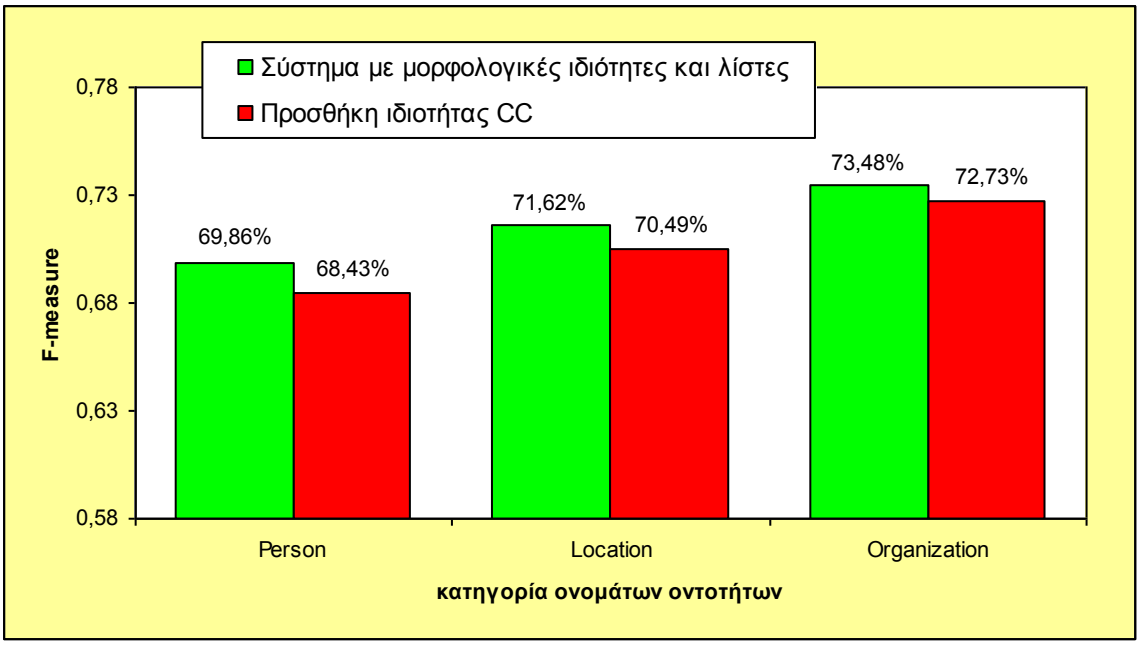
Στα παρακάτω διαγράμματα φαίνονται οι διαφορές στις επιδόσεις του συστήματος που παρατηρήθηκαν με την προσθήκη της ιδιότητας του συντελεστή συσχέτισης στο καλύτερο σύστημα της ενότητας 4.4 (μορφολογικές ιδιότητες και ιδιότητες λιστών με αξιολόγηση εγγραφών). Όπως παρατηρούμε, η προσθήκη των ιδιοτήτων του συντελεστή συσχέτισης δεν βελτιώνει πάντα τις επιδόσεις: στα ισπανικά κείμενα, τα αποτελέσματα βελτιώθηκαν σε όλες τις κατηγορίες ονομάτων, στα ελληνικά χειροτέρευσαν σε όλες τις κατηγορίες, ενώ στα ολλανδικά σε άλλες κατηγορίες βελτιώθηκαν και σε άλλες χειροτέρευσαν.



Διάγραμμα 11. F-measure του συστήματος της εργασίας με και χωρίς την **ιδιότητα του συντελεστή συσχέτισης** στην **ισπανική συλλογή κειμένων**.

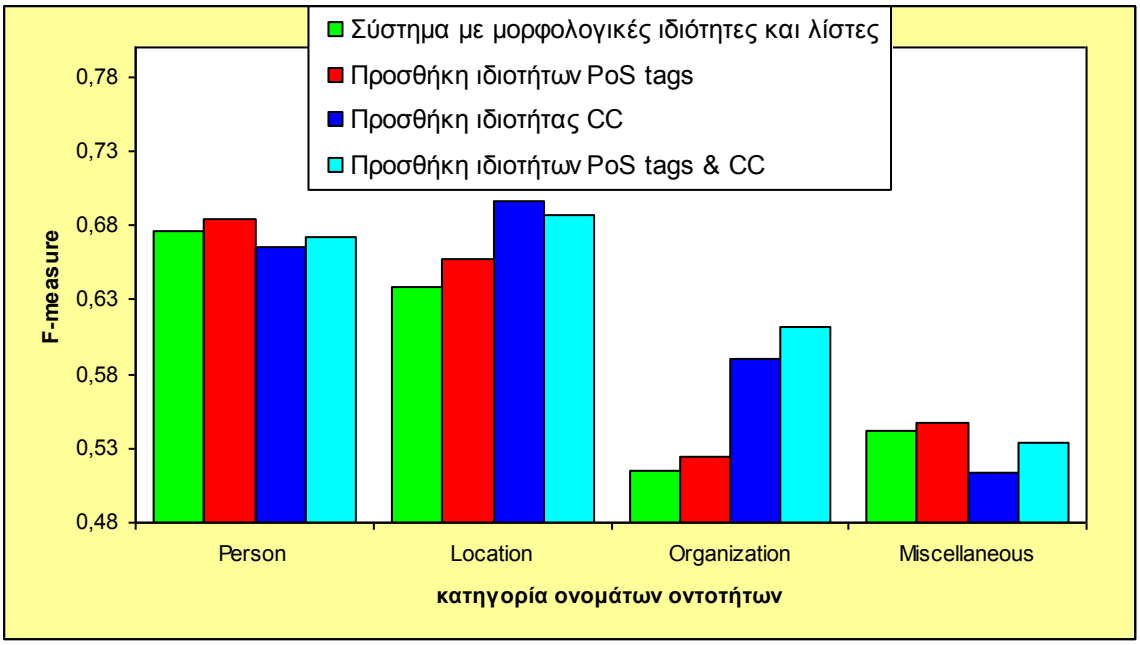


Διάγραμμα 12. F-measure του συστήματος της εργασίας με και χωρίς την **ιδιότητα του συντελεστή συσχέτισης** στην **ολλανδική συλλογή κειμένων**.



Διάγραμμα 13. F-measure του συστήματος της εργασίας με και χωρίς την **ιδιότητα του συντελεστή συσχέτισης** στην **ελληνική συλλογή κειμένων**.

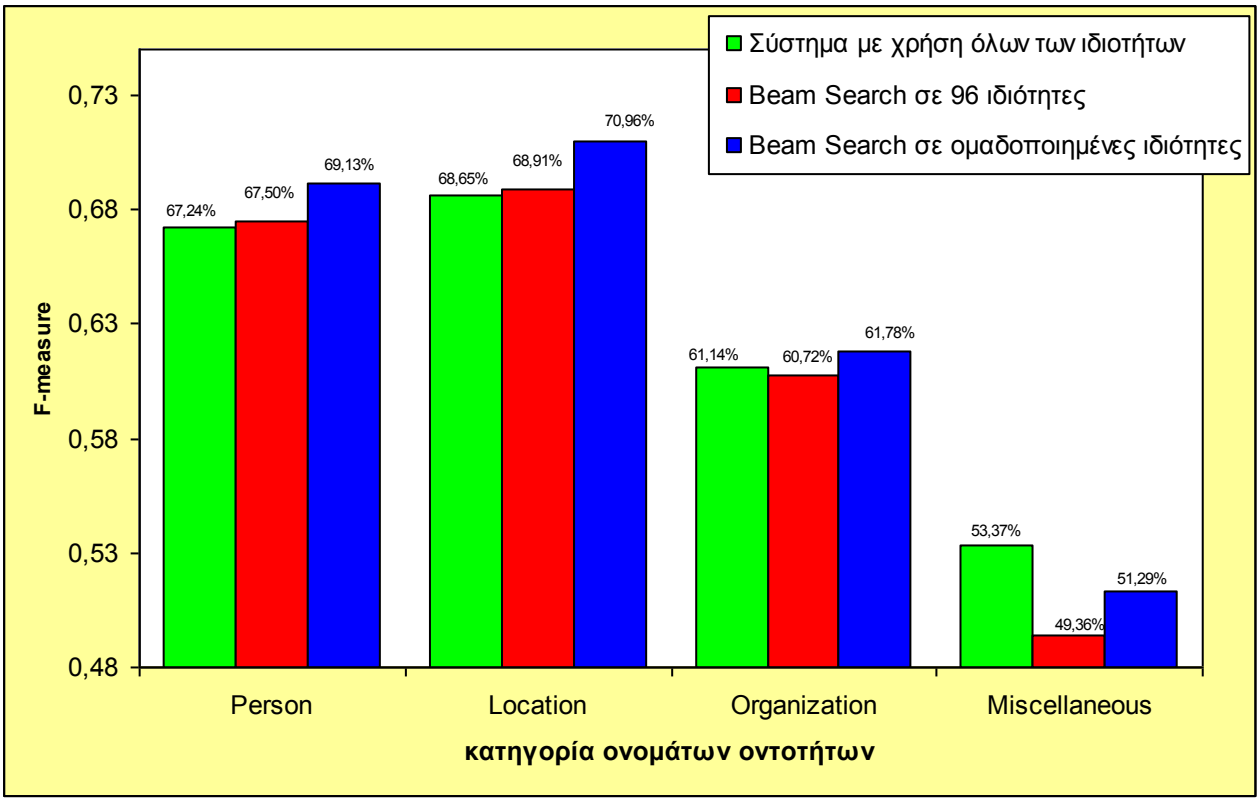
Το επόμενο διάγραμμα δείχνει τις επιδόσεις του συστήματος στην ολλανδική συλλογή, όταν προστίθενται οι ιδιότητες του συντελεστή συσχέτισης με ή χωρίς την ταυτόχρονη προσθήκη των ιδιοτήτων των ετικετών μερών του λόγου. Παρατηρούμε ότι όταν το σύστημα αναζητά ονόματα οργανισμών, είναι καλύτερο να χρησιμοποιηθούν όλες οι διαθέσιμες ιδιότητες. Αντίθετα, στην περίπτωση των ονομάτων προσώπων επιτυγχάνονται καλύτερα αποτελέσματα χωρίς την ιδιότητα του συντελεστή συσχέτισης, ενώ στην περίπτωση των ονομάτων τοποθεσιών είναι καλύτερο να αφαιρεθούν οι ιδιότητες των ετικετών μερών του λόγου. Αυτό κάνει φανερή την ανάγκη της επιλογής, για κάθε κατηγορία ονομάτων (και κάθε γλώσσα), ενός υποσυνόλου των διαθέσιμων ιδιοτήτων.



Διάγραμμα 14. Σύγκριση των επιδόσεων των συστημάτων εκπαιδευμένα στην **ολλανδική συλλογή κειμένων** όταν προστίθενται σε αυτά οι ιδιότητες των POS tags και του CC

4.7 Αυτόματη επιλογή ιδιοτήτων

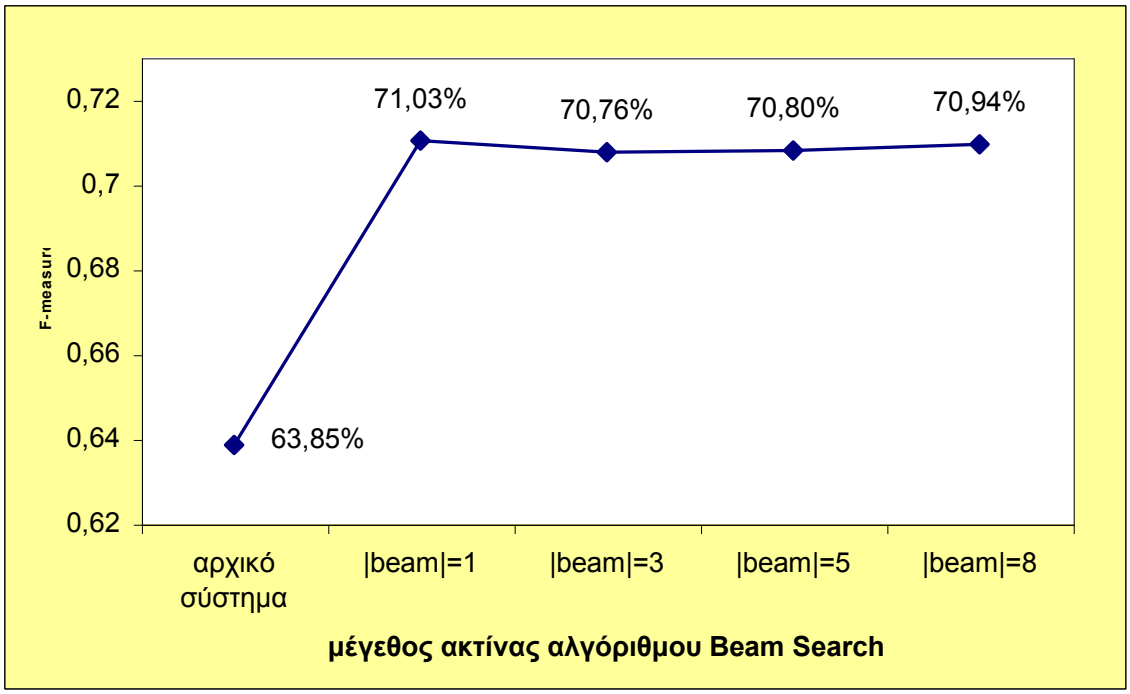
Τα παρακάτω πειράματα αφορούν την εφαρμογή της παραλλαγής του αλγόριθμου Beam Search που περιγράψαμε στην παράγραφο 3.2. Στα πειράματα αυτά ξεκινάμε από μια αρχική κατάσταση που περιλαμβάνει όλες τις υποστηριζόμενες ιδιότητες και αφαιρούμε σταδιακά ιδιότητες. Δοκιμάσαμε δύο τρόπους αφαίρεσης ιδιοτήτων. Στον πρώτο τρόπο η κάθε ιδιότητα είναι δυνατόν να αφαιρεθεί μεμονωμένα, ενώ στο δεύτερο, οι ιδιότητες που εφαρμόζονται σε παράθυρο, δηλαδή οι μορφολογικές και αυτές των ετικετών μερών του λόγου, μπορούν να ομαδοποιηθούν ώστε ο αλγόριθμος να αποφασίζει ταυτόχρονα για την αφαίρεση και των εφτά αντίστοιχων ιδιοτήτων που εφαρμόστηκαν στο παράθυρο. Στη δεύτερη περίπτωση, εάν ο αλγόριθμος αποφασίσει, για παράδειγμα, να αφαιρέσει τις ιδιότητες των ετικετών μερών του λόγου, τότε θα αφαιρεθούν ταυτόχρονα και οι 7 ιδιότητες αυτού του είδους. Δημιουργείται έτσι ένας πολύ μικρότερος χώρος αναζήτησης, σε σχέση με αυτόν του πρώτου τρόπου, με αποτέλεσμα η επιλογή υποσυνόλου ιδιοτήτων να είναι ταχύτερη.



Διάγραμμα 15. Σύγκριση των **διαφορετικών τρόπων επιλογής ιδιοτήτων με Beam Search** στην **ολλανδική συλλογή κειμένων**

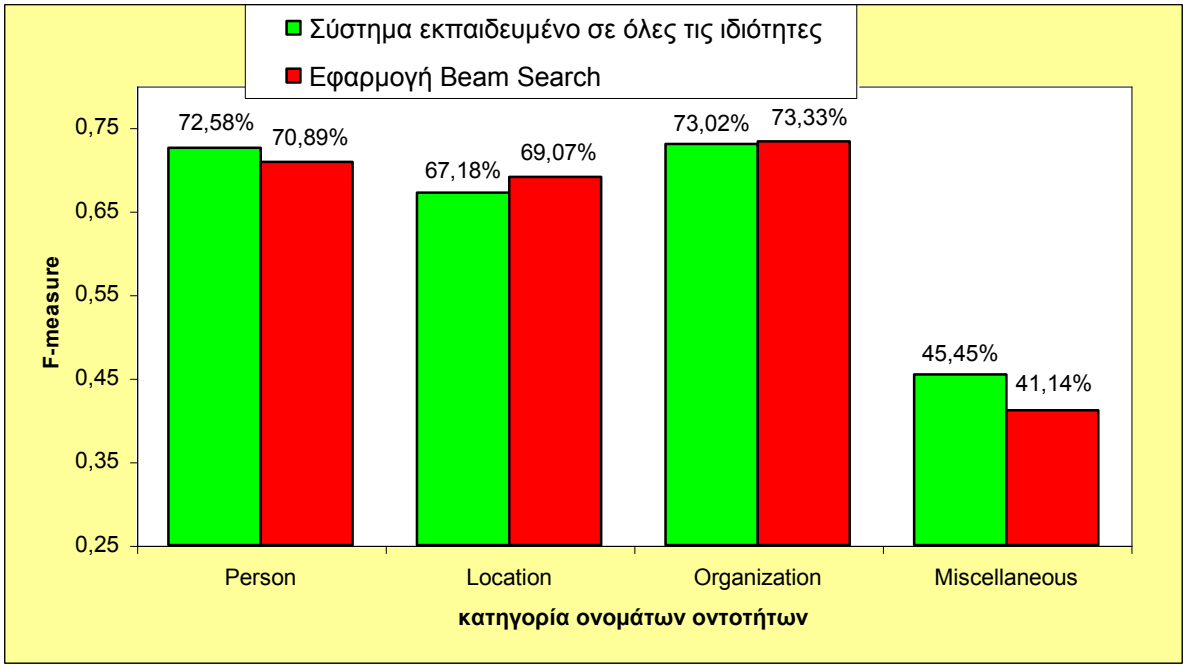
Όπως παρατηρούμε από το παραπάνω διάγραμμα, όπου χρησιμοποιήθηκε η ολλανδική συλλογή κειμένων, με το δεύτερο τρόπο χρήσης του Beam Search, δεν έχουμε μόνο ταχύτερη επιλογή ιδιοτήτων, αλλά επιτυγχάνουμε και καλύτερα αποτελέσματα. Μόνη εξαίρεση ήταν η κατηγορία «Miscellaneous», στην οποία μετρήσαμε χαμηλότερα αποτελέσματα και με τις δύο προσεγγίσεις του Beam Search.

Στο παρακάτω διάγραμμα φαίνονται τα αποτελέσματα για τα ονόματα τοποθεσιών, πάντα στην ολλανδική συλλογή κειμένων, όταν μεταβάλλουμε το μέγεθος του μετώπου αναζήτησης (παράμετρος *k* της ενότητας 3.2). Σημειώνουμε πως στην περίπτωση αυτή χρησιμοποιούμε το σύστημα της παραγράφου 4.4, δεν έχουμε προσθέσει δηλαδή ακόμα τις ιδιότητες των ετικετών μερών του λόγου, ούτε του συντελεστή συσχέτισης.

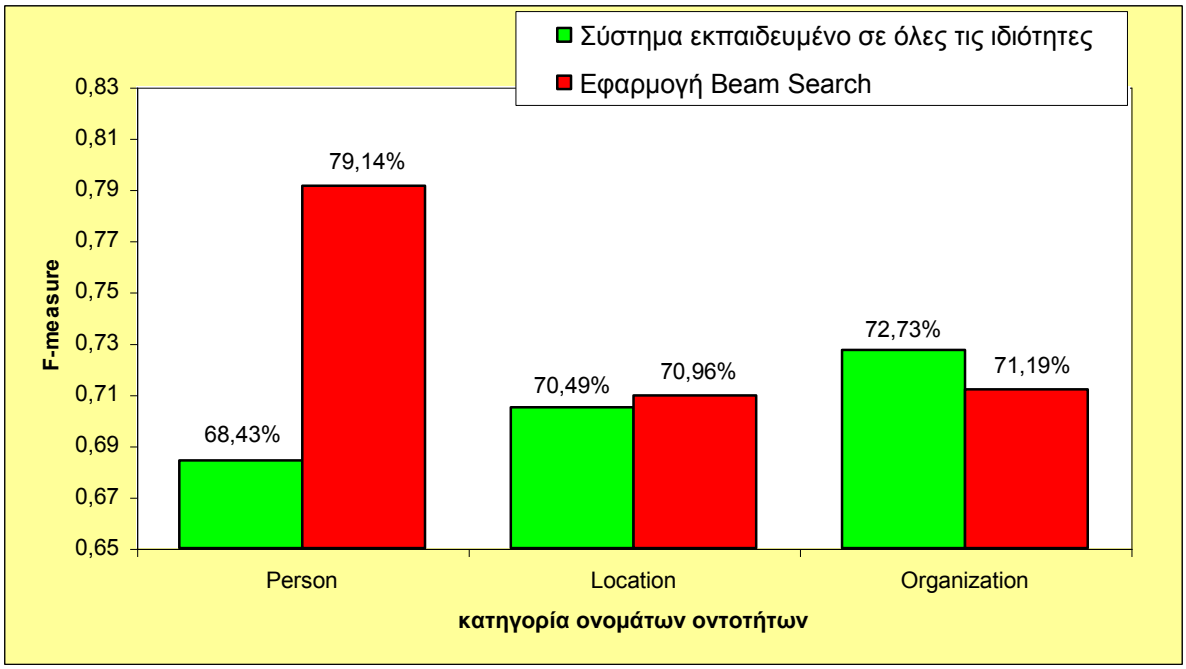


Διάγραμμα 16. Η επίδραση του **μεγέθους του μετώπου αναζήτησης** του **Beam Search** στα αποτελέσματα της αναγνώρισης **ονομάτων τοποθεσιών** στην **ολλανδική** συλλογή κειμένων.

Όπως παρατηρούμε, το μέγεθος του μετώπου δεν επηρεάζει σημαντικά το τελικό αποτέλεσμα. Αντίστοιχα αποτελέσματα προέκυψαν και με πειράματα άλλων κατηγοριών ονομάτων και κειμένων άλλων γλωσσών. Έτσι στα παρακάτω πειράματα με κείμενα άλλων γλωσσών χρησιμοποιούμε $k = 1$, που είναι η απλούστερη επιλογή. Στα πειράματα αυτά χρησιμοποιήθηκε το σύστημα της παραγράφου 4.6, ώστε η αναζήτηση Beam Search να έχει στη διάθεσή της όλες τις υποστηριζόμενες ιδιότητες και να επιλέξει μόνη της τις πιο χρήσιμες.



Διάγραμμα 17. Σύγκριση αποτελεσμάτων πριν και μετά την εφαρμογή του αλγόριθμου **Beam Search** στην **ισπανική συλλογή κειμένων** με το σύστημα της ενότητας 4.6

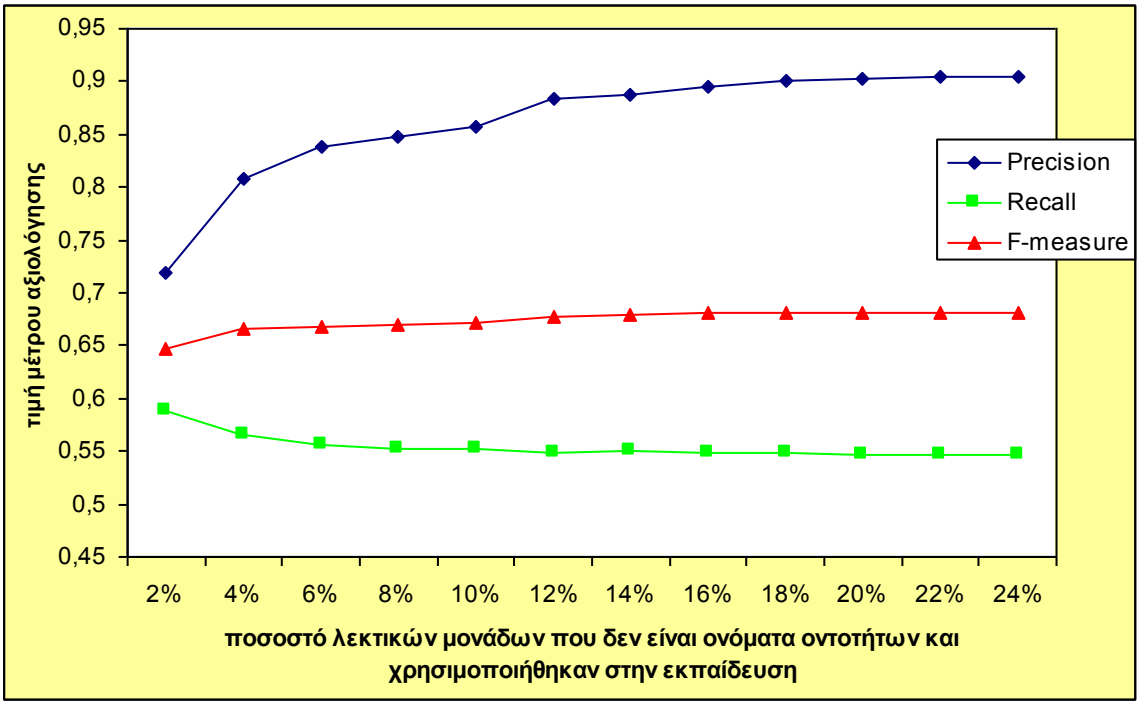


Διάγραμμα 18. Σύγκριση αποτελεσμάτων πριν και μετά την εφαρμογή του αλγόριθμου **Beam Search** στην **ελληνική συλλογή κειμένων** με το σύστημα της ενότητας 4.6

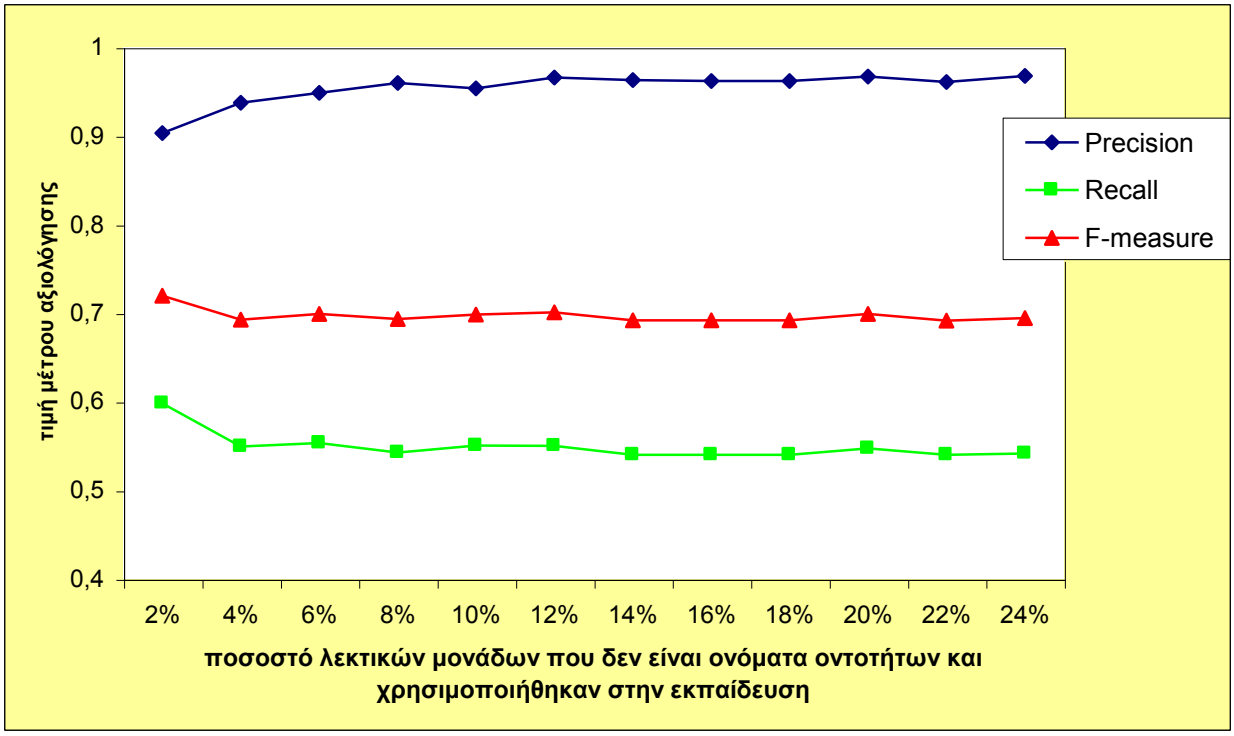
Όπως παρατηρούμε στα παραπάνω διαγράμματα των τριών συλλογών κειμένων, η εφαρμογή του Beam Search για την επιλογή υποσυνόλου ιδιοτήτων δεν βελτιώνει πάντα τα αποτελέσματα, αν και τα βελτιώνει στην πλειοψηφία των περιπτώσεων.

4.8 Πειράματα με μεταβαλλόμενο αριθμό αρνητικών παραδειγμάτων εκπαίδευσης

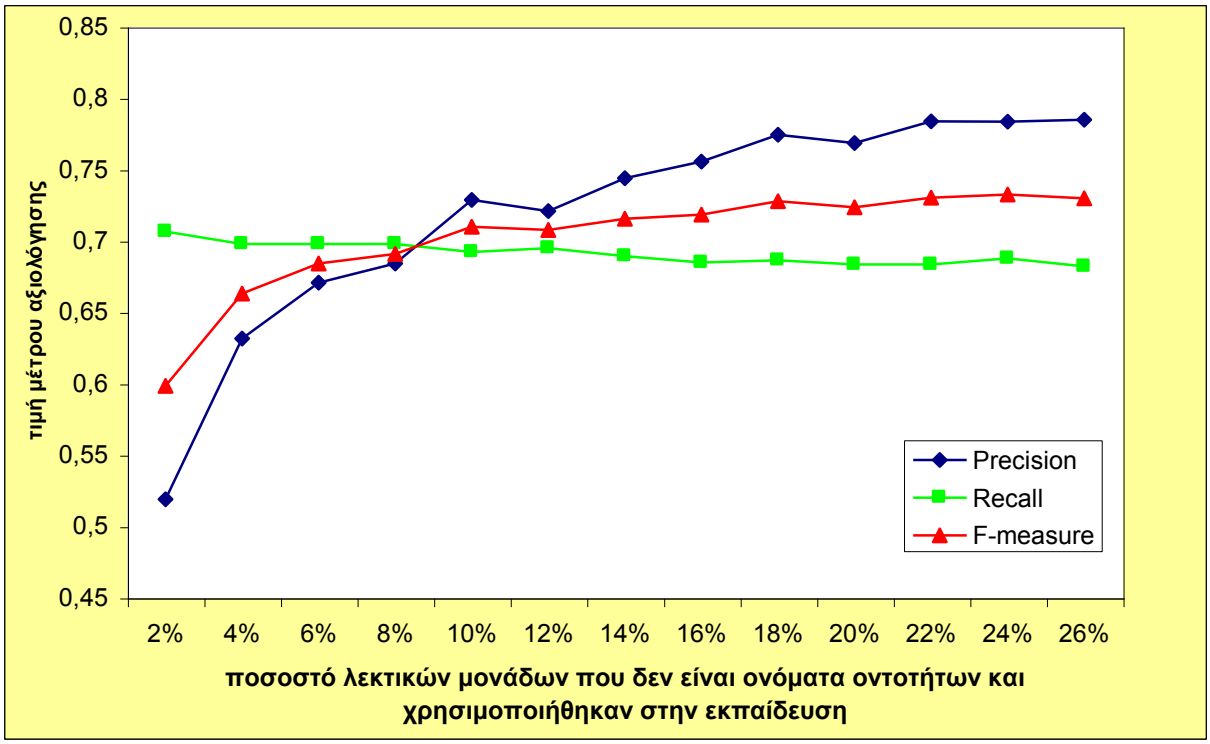
Οι συλλογές κειμένων των πειραμάτων αποτελούνται από χιλιάδες λεκτικές μονάδες και μόνο ένα μικρό ποσοστό αυτών αποτελούν ονόματα οντοτήτων. Στα πειράματα αυτής της ενότητας χρησιμοποιήσαμε ως παραδείγματα εκπαίδευσης όλες τις λεκτικές μονάδες ονομάτων οντοτήτων (θετική κατηγορία) των δεδομένων εκπαίδευσης και μόνο ένα μεταβαλλόμενο ποσοστό των λεκτικών μονάδων που δεν ανήκουν σε ονόματα οντοτήτων (αρνητική κατηγορία). Το να μη χρησιμοποιούνται όλες οι λεκτικές μονάδες της αρνητικής κατηγορίας βοηθά στην εξοικονόμηση μνήμης και χρόνου κατά την εκπαίδευση του ταξινομητή και ενδεχομένως κατά τη χρήση του αργότερα. Διερευνούμε, επίσης, το κατά πόσον η αναλογία θετικών και αρνητικών παραδειγμάτων επηρεάζει τις επιδόσεις του συστήματος.



Διάγραμμα 19. Πειράματα με **μεταβαλλόμενο αριθμό αρνητικών παραδειγμάτων** εκπαίδευσης στην περίπτωση **ονομάτων προσώπων** της **ολλανδικής συλλογής κειμένων**.



Διάγραμμα 20. Πειράματα με **μεταβαλλόμενο αριθμό αρνητικών παραδειγμάτων** εκπαίδευσης στην περίπτωση **ονομάτων προσώπων** της **ελληνικής συλλογής κειμένων**.

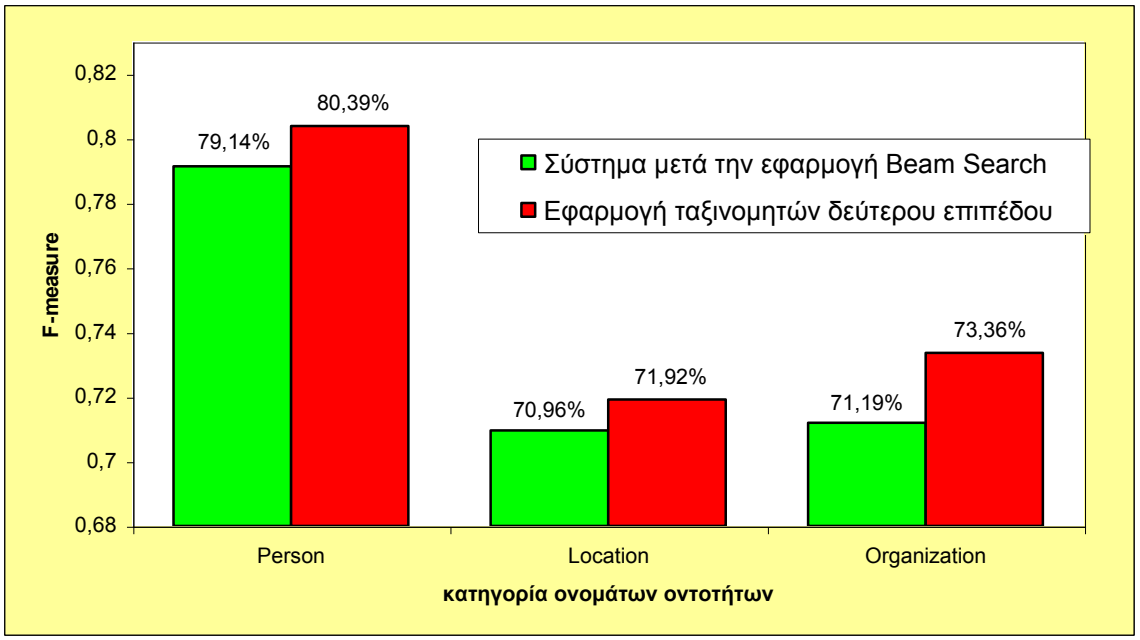


Διάγραμμα 21. Πειράματα με **μεταβαλλόμενο αριθμό αρνητικών παραδειγμάτων** εκπαίδευσης στην περίπτωση **ονομάτων τοποθεσιών** της **ελληνικής συλλογής κειμένων**.

Όπως προκύπτει από τα παραπάνω διαγράμματα, η επίδοση F-measure του συστήματος είναι σχεδόν ανεξάρτητη του ποσοστού των αρνητικών παραδειγμάτων εκπαίδευσης. Στις περισσότερες περιπτώσεις, όμως, όσο μεγαλώνει το ποσοστό των αρνητικών παραδειγμάτων, αυξάνεται η ακρίβεια και μειώνεται η ανάκληση. Για την εκτέλεση όλων των υπολοίπων πειραμάτων αυτού του κεφαλαίου χρησιμοποιήσαμε το 15% των αρνητικών παραδειγμάτων κατά την εκπαίδευση.

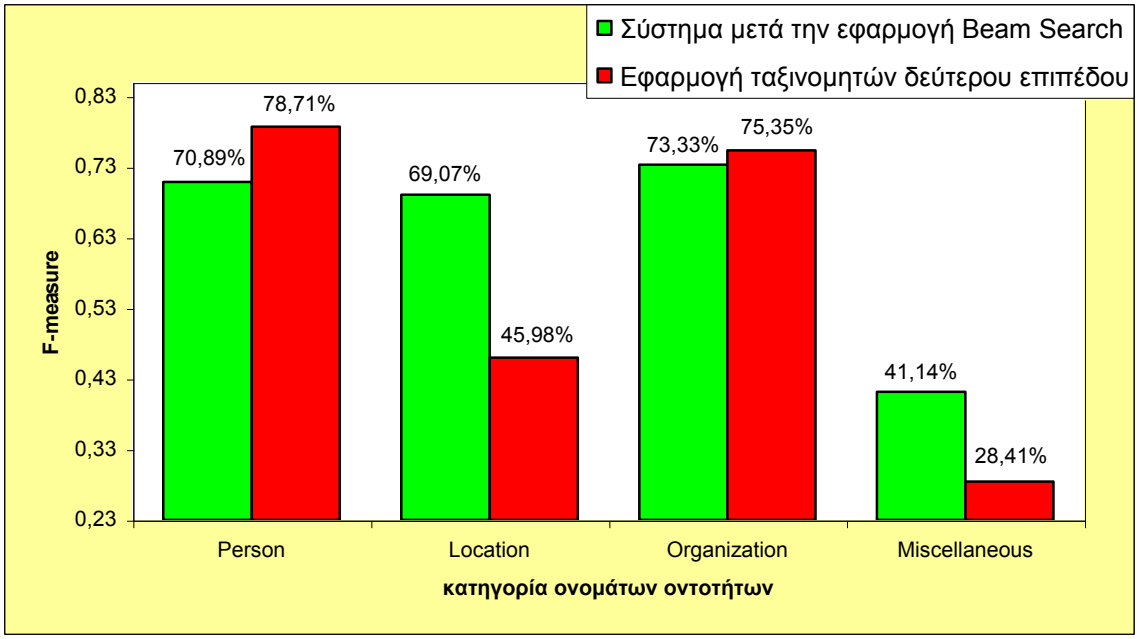
4.9 Ταξινομητές δευτέρου επιπέδου

Σε αυτή την ενότητα φαίνονται τα αποτελέσματα που λαμβάνουμε, εάν εφαρμόσουμε τους ταξινομητές δευτέρου επιπέδου στα συστήματα της προηγούμενης ενότητας, εφόσον δηλαδή έχουμε εφαρμόσει τον αλγόριθμο Beam Search.

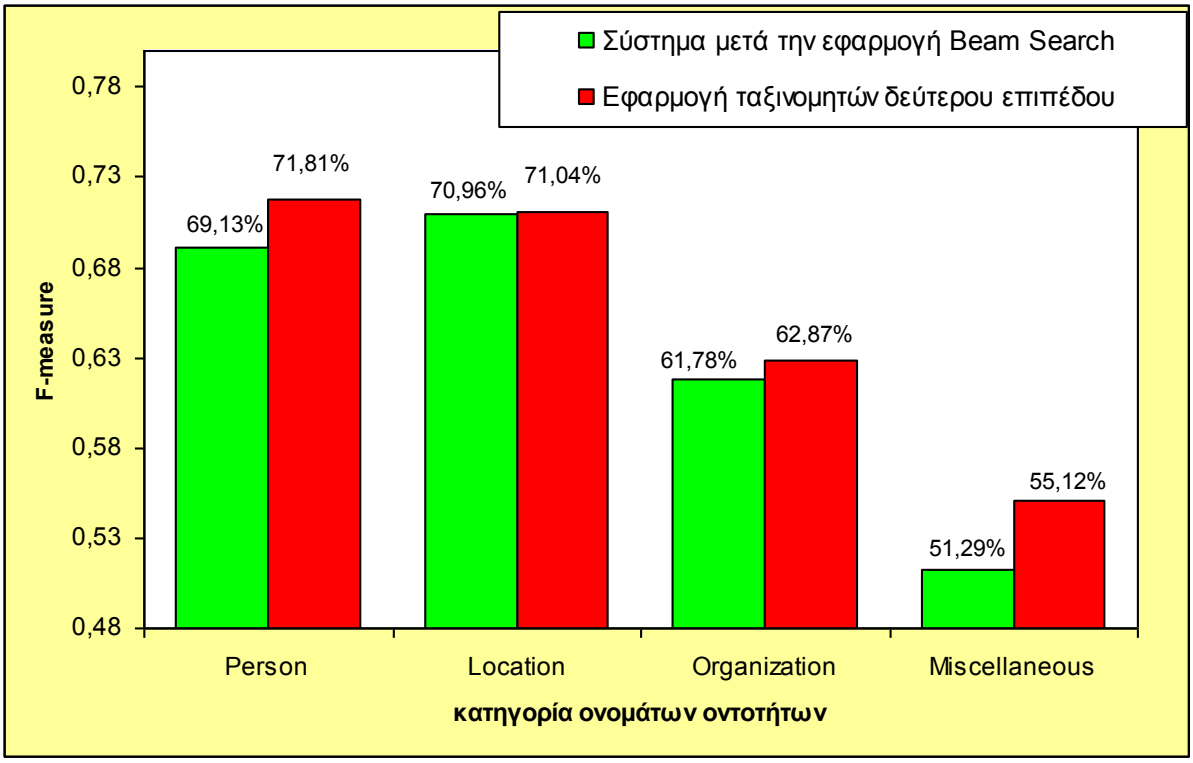


Διάγραμμα 22. Σύγκριση συστήματος πριν και μετά την εφαρμογή των **ταξινομητών δευτέρου επιπέδου** στην **ελληνική συλλογή κειμένων**

Παρατηρούμε πως στις περισσότερες περιπτώσεις οι ταξινομητές δευτέρου επιπέδου αυξάνουν αρκετά τις επιδόσεις των συστημάτων. Υπάρχουν βέβαια και περιπτώσεις, όπως αυτές των κατηγοριών «location» και «miscellaneous» της ισπανικής συλλογής, στις οποίες παρατηρούμε αρκετά μεγάλη μείωση της επίδοσης. Αυτό το πρόβλημα θα μπορούσε να αντιμετωπιστεί δοκιμάζοντας πρώτα τους ταξινομητές δευτέρου επιπέδου σε κάποια δοκιμαστικά κείμενα, όπως αυτά των δεδομένων ανάπτυξης, ώστε να συμπεράνουμε το εάν θα τους χρησιμοποιήσουμε ή όχι στο τελικό σύστημα.



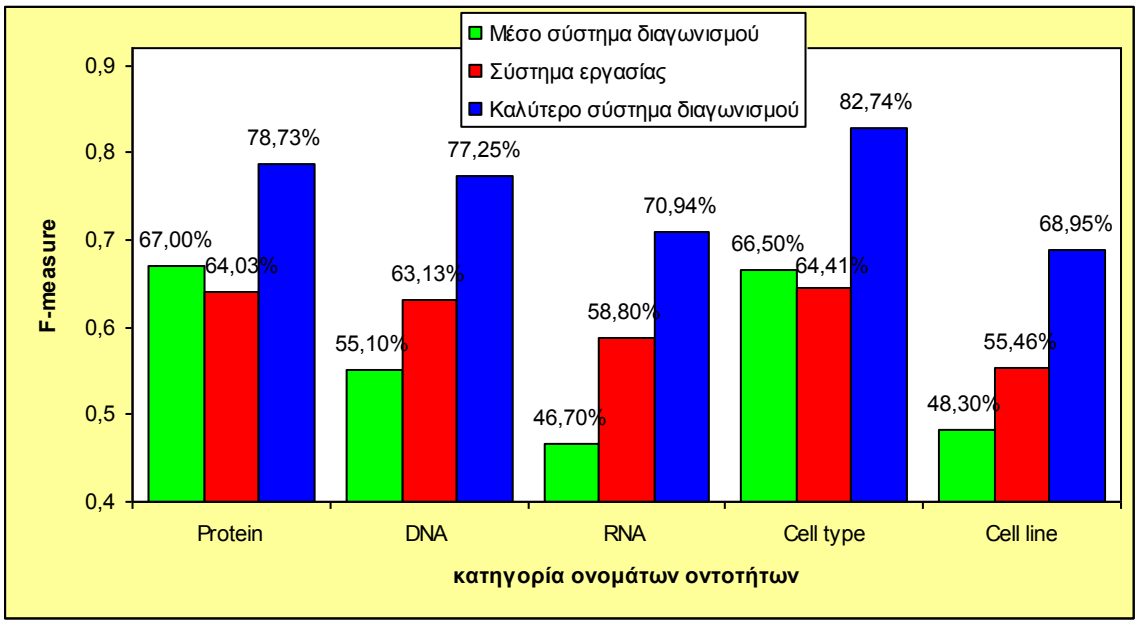
Διάγραμμα 23. Σύγκριση συστήματος πριν και μετά την εφαρμογή των ταξινομητών δευτέρου επιπέδου στην **ισπανική συλλογή κειμένων**



Διάγραμμα 24. Σύγκριση συστήματος πριν και μετά την εφαρμογή των ταξινομητών δευτέρου επιπέδου στην **ολλανδική συλλογή κειμένων**

4.10 Βιοϊατρικά ονόματα οντοτήτων

Στην προσπάθειά μας να δοκιμάσουμε τις επιδόσεις του συστήματος σε μία επιπλέον γλώσσα, τα αγγλικά, και σε διαφορετικά είδη οντοτήτων από αυτά που αναφέρονται στις προηγούμενες παραγράφους, χρησιμοποιήσαμε την τέταρτη συλλογή κειμένων που περιγράψαμε στην παράγραφο 4.1. Στα παρακάτω διαγράμματα έχουν χρησιμοποιηθεί όλα τα στάδια που αναφέρθηκαν στις προηγούμενες παραγράφους, μέχρι και τους ταξινομητές δευτέρου επιπέδου. Επίσης στο παρακάτω διάγραμμα εμφανίζονται οι επιδόσεις ενός μέσου [17] και του καλύτερου [18] συστήματος που έλαβαν μέρος στο διαγωνισμό από τον οποίο προέρχεται η συγκεκριμένη συλλογή κειμένων. Σημειώνεται ότι τα συστήματα του διαγωνισμού καλούνταν να κατατάξουν κάθε λεκτική μονάδα σε μία από τρεις διαφορετικές κατηγορίες (αρχική λεκτική μονάδα ονόματος, μη αρχική λεκτική μονάδα ονόματος, λεκτική μονάδα που δεν αποτελεί μέρος ονόματος), ενώ το σύστημα της παρούσας εργασίας κατατάσσει τις λεκτικές μονάδες μόνο σε δύο κατηγορίες (λεκτική μονάδα που αποτελεί ή όχι μέρος ονόματος. Έτσι καταλαβαίνουμε πως τα αποτελέσματα του διαγωνισμού δεν είναι άμεσα συγκρίσιμα με αυτά του συστήματός μας. Επίσης, όπως περιγράψαμε στην ενότητα 4.1, δε χρησιμοποιήσαμε όλα τα δεδομένα εκπαίδευσης, γιατί αποκόψαμε μέρος αυτών (περίπου 20%) για τη δημιουργία των δεδομένων ανάπτυξης, οπότε το σύστημά μας έχει εκπαιδευτεί σε λιγότερα δεδομένα από ό,τι τα συστήματα του διαγωνισμού.

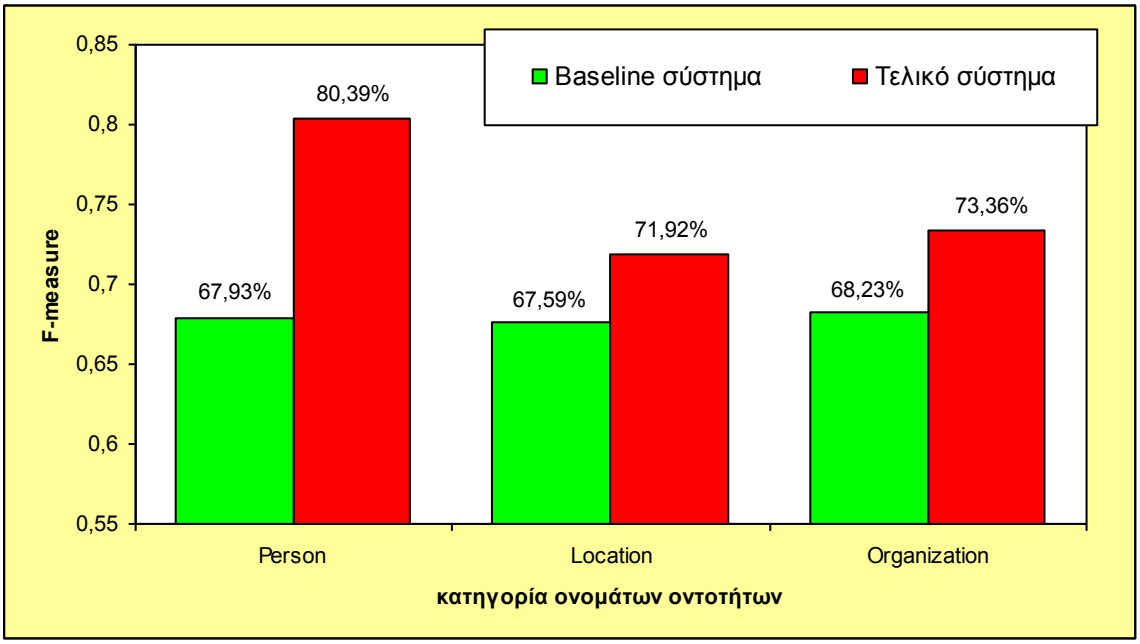


Διάγραμμα 25. Επιδόσεις του συστήματος στην **αγγλική συλλογή βιοϊατρικών κειμένων** και επιδόσεις άλλων συστημάτων του διαγωνισμού.

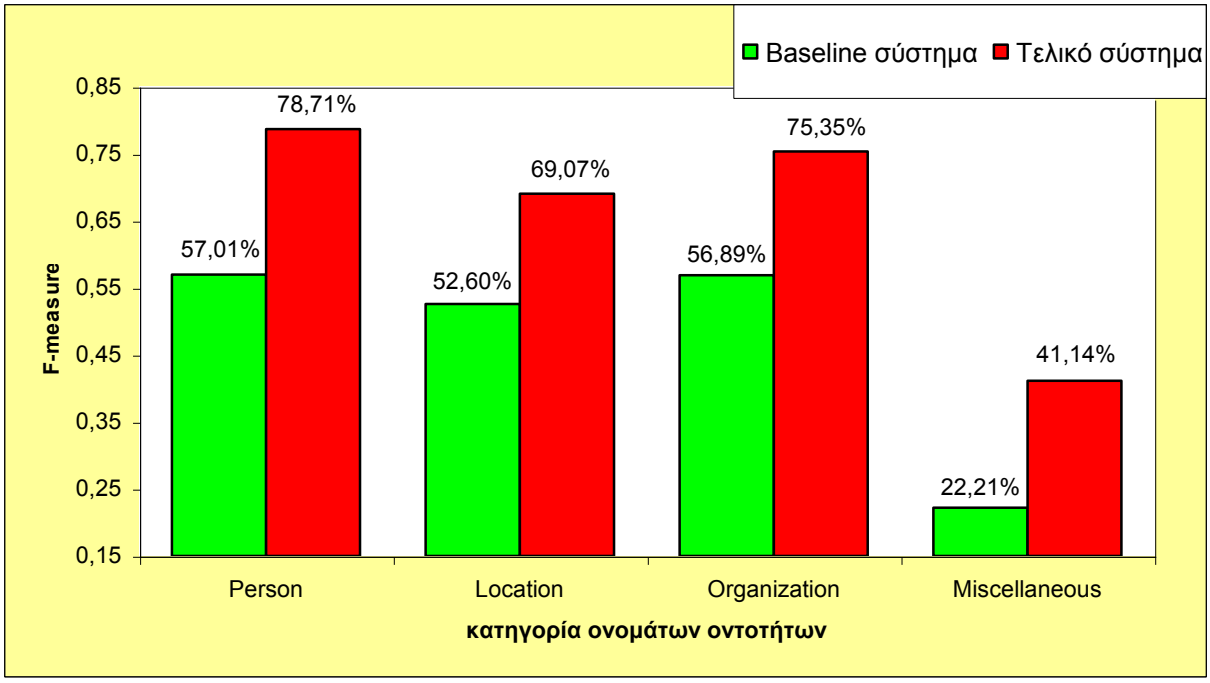
4.11 Σύγκριση συστήματος με άλλα συστήματα

Στην ενότητα αυτή συγκρίνουμε τις επιδόσεις του δικού μας συστήματος με εκείνες απλοϊκών (baseline) συστημάτων και συστημάτων άλλων.

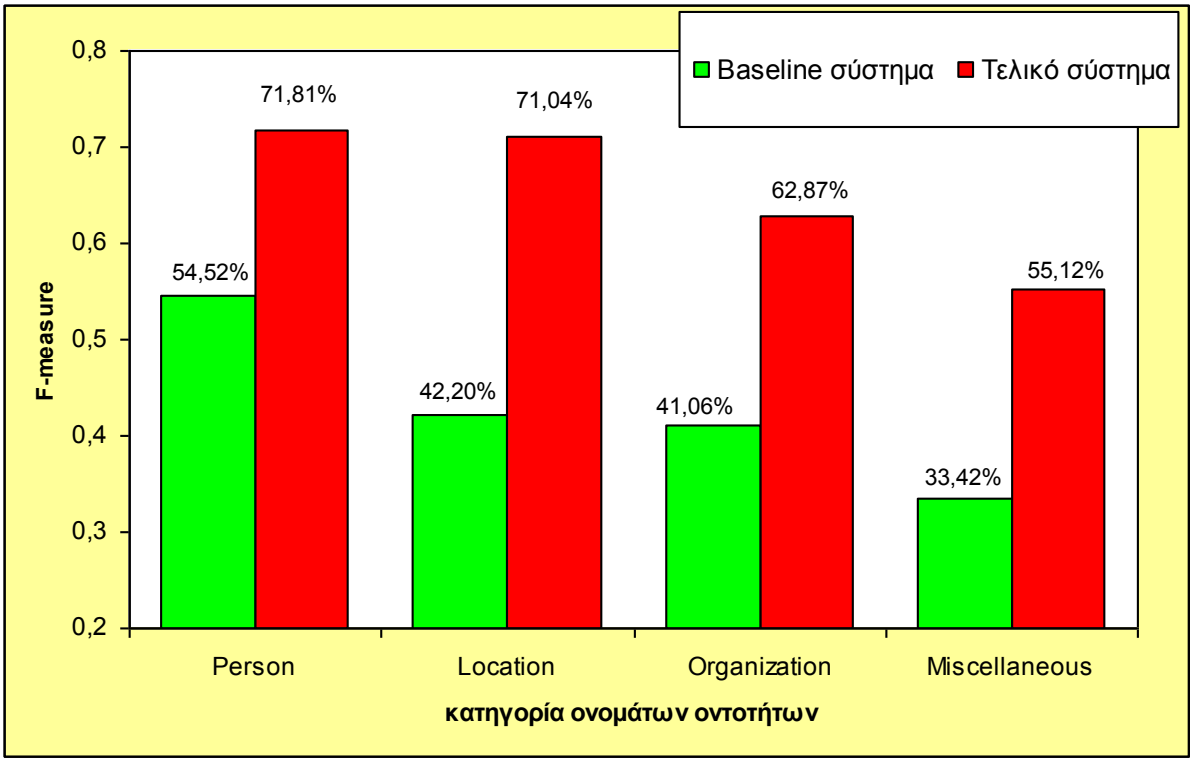
Τα baseline συστήματα, των οποίων οι επιδόσεις φαίνονται παρακάτω, έχουν προκύψει ως εξής. Αρχικά κατασκευάζεται μία λίστα με όλες τις λεκτικές μονάδες των δεδομένων εκπαίδευσης που ανήκουν στην επιθυμητή κατηγορία ονομάτων οντοτήτων. Από αυτή τη λίστα αφαιρούμε τις λεκτικές μονάδες που αποτελούνται μόνο από μικρά γράμματα. Για να κατατάξει το σύστημα μια νέα λεκτική μονάδα ως όνομα οντότητας, θα πρέπει αυτή να υπάρχει μέσα στη λίστα, διαφορετικά την κατατάσσει ως μη όνομα οντότητας. Στα συστήματά μας των παρακάτω διαγραμμάτων έχουμε εφαρμόσει τους ταξινομητές δευτέρου επιπέδου μόνο στις περιπτώσεις (κατηγορίες ονομάτων και γλώσσες) όπου οδηγούσαν σε βελτίωση των αποτελεσμάτων (*τελικά συστήματα*).



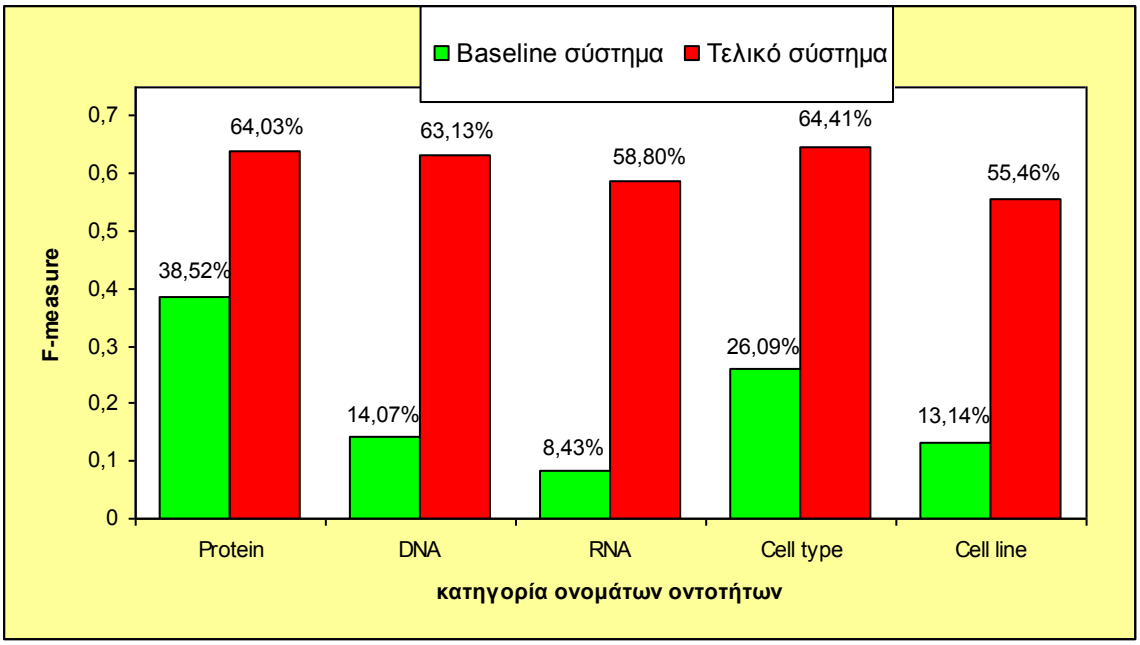
Διάγραμμα 26. Σύγκριση του **τελικού συστήματος** της εργασίας με τα αντίστοιχα **baseline** συστήματα στην **ελληνική συλλογή κειμένων**.



Διάγραμμα 27. Σύγκριση του **τελικού συστήματος** της εργασίας με τα αντίστοιχα **baseline** συστήματα στην **ισπανική συλλογή κειμένων**.



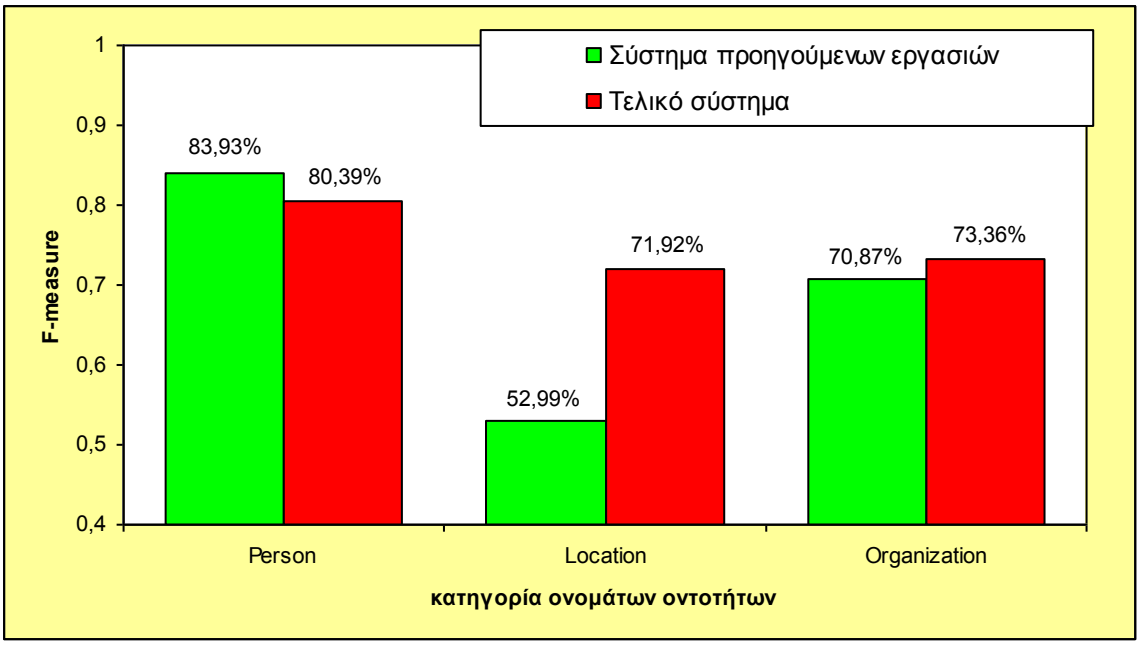
Διάγραμμα 28. Σύγκριση του **τελικού συστήματος** της εργασίας με τα αντίστοιχα **baseline** συστήματα στην **ολλανδική συλλογή κειμένων**.



Διάγραμμα 29. Σύγκριση του **τελικού συστήματος** της εργασίας με τα αντίστοιχα **baseline** συστήματα στην **αγγλική συλλογή κειμένων.**

Παρατηρούμε ότι οι επιδόσεις του συστήματός μας ξεπερνούν κατά πολύ αυτές των baseline συστημάτων. Επίσης, από τις επιδόσεις των baseline συστημάτων μπορούμε να βγάλουμε ένα πρώτο συμπέρασμα για τη δυσκολία μιας κατηγορίας ονομάτων οντοτήτων ανά συλλογή κειμένων. Τα baseline συστήματα της ελληνικής συλλογής κειμένων έχουν αρκετά καλές επιδόσεις, παρ' όλη την απλοϊκότητα της κατασκευής τους. Από την άλλη, οι κατηγορίες της βιοϊατρικής συλλογής κειμένων είναι πιο δύσκολες. Αυτό είναι αρκετά ενθαρρυντικό, μιας και βλέπουμε πως το σύστημά μας έχει προσαρμοστεί αρκετά καλά σε αυτά τα δύσκολα είδη οντοτήτων.

Στο παρακάτω διάγραμμα φαίνεται η σύγκριση του συστήματός μας με το ήδη υπάρχον σύστημα που κατασκευάστηκε αρχικά από τον Γ. Λουκαρέλλι [13] και στη συνέχεια επεκτάθηκε και βελτιώθηκε από τους Ξ. Βασιλάκο [15] και Ι. Κώνστα [16]. Λόγω του ότι χρησιμοποιήθηκαν οι ίδιες συλλογές κειμένων για την εκπαίδευση και αξιολόγηση και των δύο συστημάτων μπορούμε να τα συγκρίνουμε άμεσα. Παρατηρούμε πως το σύστημά μας έχει επιτύχει παρόμοιες επιδόσεις στις κατηγορίες των ονομάτων προσώπων και οργανισμών και μία αρκετά καλύτερη επίδοση σε αυτή των ονομάτων τοποθεσιών.



Διάγραμμα 30. Σύγκριση των επιδόσεων του **τελικού συστήματος** αυτής της εργασίας με το προϋπάρχον σύστημα στην **ελληνική συλλογή κειμένων.**

Κεφάλαιο 5: Συμπεράσματα και μελλοντικές κατευθύνσεις

Στην εργασία αυτή αναπτύχθηκε ένα σύστημα αναγνώρισης ονομάτων οντοτήτων, το οποίο μπορεί να προσαρμοστεί, μέσω μηχανικής μάθησης, για χρήση με κείμενα διαφορετικών φυσικών γλωσσών και ονόματα οντοτήτων νέων κατηγοριών. Το σύστημα χρησιμοποιεί δύο επίπεδα ταξινομητών, με τους ταξινομητές του δευτέρου επιπέδου να διορθώνουν λάθη των ταξινομητών του πρώτου. Όλοι οι ταξινομητές χρησιμοποιούν τον αλόριθμο μάθησης της Μέγιστης Εντροπίας και ιδιότητες που βασίζονται στη μορφολογία των λέξεων, λίστες που κατασκευάζονται αυτόματα κατά την εκπαίδευση, ετικέτες μερών του λόγου και ένα συντελεστή συσχέτισης. Το σύστημα περιλαμβάνει μηχανισμούς επιλογής υποσυνόλων ιδιοτήτων, διαφορετικών ανά κατηγορία ονομάτων, οι οποίοι βασίζονται σε αναζήτηση Beam Search. Το σύστημα δοκιμάστηκε σε τέσσερις διαφορετικές συλλογές κειμένων, γραμμένων στα ελληνικά, αγγλικά, ισπανικά και ολλανδικά αντίστοιχα. Στα κείμενα αυτά υπάρχουν χειρωνακτικά επισημειωμένες συνολικά εννέα διαφορετικές κατηγορίες ονομάτων οντοτήτων. Οι επιδόσεις του συστήματος ήταν αρκετά ικανοποιητικές και σε αρκετές περιπτώσεις συγκρίσιμες με εκείνες συστημάτων που κατασκευάστηκαν για συγκεκριμένες γλώσσες και κατηγορίες ονομάτων.

Παρακάτω παραθέτουμε μερικές ιδέες για μελλοντικές επεκτάσεις του συστήματος.

Αρχικά θα μπορούσαμε να δοκιμάσουμε την εκπαίδευση του συστήματος σε παραπάνω από τις δύο κατηγορίες που δηλώνουν απλώς εάν μια λεκτική μονάδα είναι ή δεν είναι όνομα οντότητας. Υπάρχουν πολλές παραλλαγές αυτής της προσέγγισης και πολλοί κατασκευαστές χρησιμοποιούν τρεις κατηγορίες, που δηλώνουν εάν μια λεκτική μονάδα είναι αρχή ονόματος οντότητας, εάν είναι μέρος ονόματος οντότητας αλλά όχι η πρώτη του λεκτική μονάδα και εάν δεν είναι όνομα οντότητας (ούτε μέρος του). Άλλες προσεγγίσεις προτείνουν τη χρήση πέντε κατηγοριών, προσθέτοντας μια κατηγορία που δηλώνει εάν μια λεκτική μονάδα βρίσκεται στο τέλος ενός ονόματος οντότητας και μια ξεχωριστή κατηγορία για τα ονόματα οντοτήτων που αποτελούνται από μόνο μία λεκτική μονάδα.

Πολύ χρήσιμος θα ήταν ο εμπλουτισμός των ταξινομητών δευτέρου επιπέδου με περισσότερες ιδιότητες. Για παράδειγμα, χρησιμοποιώντας το συντελεστή συσχέτισης θα μπορούσαμε να ορίσουμε μια νέα ιδιότητα που να εξετάζει αν η κατηγορία στην οποία κατετάγη μια λεκτική μονάδα από το πρώτο επίπεδο συμφωνεί με την κατηγορία που υποδηλώνουν τα συμφραζόμενά της.

Επίσης, ίσως θα ήταν δυνατό να εμπλουτιστεί ο χώρος καταστάσεων που εξερευνά η επιλογή ιδιοτήτων με Beam Search, ώστε να εξετάζεται όχι μόνο το εάν θα πρέπει να προστεθεί (ή αφαιρεθεί) κάποια ιδιότητα στο σύνολο ιδιοτήτων, αλλά στην περίπτωση των ιδιοτήτων λιστών να εξετάζεται και ποιο μέτρο (π.χ. ακρίβεια, ανάκληση, F-measure) είναι προτιμότερο να χρησιμοποιηθεί για την αξιολόγηση των εγγραφών της λίστας.

Τέλος θα ήταν ενδιαφέρον να προστεθεί υποστήριξη και για άλλους αλγορίθμους μάθησης, όπως τα Hidden Markov Models ή τα Conditional Random Fields [19] και να γίνουν πειράματα και με αυτούς.

6 Αναφορές

[1] A. Mikheev, C. Grover and M. Moens, *Description of the LTG System used for MUC-7*, Proceedings of Seventh Message Understanding Conference, Faifax, Virginia USA, 1997.

[2] X. Carreras, L. Marques and L. Padro, *Named entity extraction using adaboost*, Proceedings of CoNLL-2002, pages 167-170, Taipei, Taiwan, 2002.

[3] Y. Freund and R.E. Schapire, *Experiments with a new boosting algorithm*, Machine Learning: Proceedings of the Thirteenth International Conference, pages 148–156, Bari, Italy 1996.

[4] R. Florian, A. Ittycheriah, H. Jing and T. Zhang, *Named Entity Recognition through Classifier Combination*, Proceedings of CoNLL-2003, pages 168-171, Edmonton, Canada, 2003.

[5] *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, R.D. Rosenkrantz ed., D. Reidel Publishing Co., Dordrecht-Holland, 1983.

[6] S.J. DeRose, *Grammatical category disambiguation by statistical optimization*, Computational Linguistics, pages 14:31-39, 1988.

[7] Eugene Charniak, *Statistical Language Learning*, MIT Press, Cambridge, MA, 1993.

[8] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[9] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.

[10] B.V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, Los Alamitos, California: IEEE Computer Society Press, 1991.

[11] H.T. Ng, W.B. Goh and K.L. Low, *Feature selection, perceptron learning, and a usability case study for text categorization*, In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 67-73, Philadelphia, Pennsylvania, USA, 1997.

[12] I. Michailidis, K. Diamantaras, S. Vasileiadis and Y. Frere, *Greek named entity recognition using Support Vector Machines, Maximum Entropy and Onetime*, Proceedings of the 5th International Conference on Language Resources and Evaluation, pages 45–72, Genoa, Italy, 2006.

[13] G. Lucarelli, *Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα*, Διπλωματική εργασία Προγράμματος Μεταπτυχιακών Σπουδών στην Επιστήμη των Υπολογιστών, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.

[14] G. Lucarelli, X. Vasilakos and I. Androutsopoulos, *Named Entity Recognition in Greek Texts with an Ensemble of SVMs and Active Learning*, International Journal on Artificial Intelligence Tools, pages 16(6):1015-1045, 2007.

[15] Ξ. Βασιλάκος, *Αναγνώριση και κατάταξη ονομάτων οντοτήτων σε ελληνικά κείμενα με Μηχανές Διανυσμάτων Υποστήριξης*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.

[16] Ι. Κώνστας, *Αναγνώριση και κατάταξη ονομάτων προσώπων, οργανισμών και τοποθεσιών σε ελληνικά κείμενα με χρήση Μηχανών Διανυσμάτων Υποστήριξης*, πτυχιακή εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2007.

[17] M. Rossler, *Adapting an NER-System for German to the Biomedical Domai*,In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), Geneva, Switzerland, 2004.

[18] GD. Zhou and J. Su, *Exploring Deep Knowledge Resources in Biomedical Name Recognition*, In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), Geneva, Switzerland, 2004.

[19] J. Lafferty, A. McCallum and F. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, In Proceedings of the 18th International Conference on Machine Learning, pages 282–289, Morgan Kaufmann, San Francisco, CA, USA, 2001.