

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Αναγνώριση και Κατάταξη Ονομάτων Προσώπων, Οργανισμών και Τοποθεσιών σε Ελληνικά Κείμενα με Χρήση Μηχανών Διανυσμάτων Υποστήριξης»

Ιωάννης Κώνστας

Επιβλέπων: **Ίων Ανδρουτσόπουλος**

ΑΘΗΝΑ, ΑΥΓΟΥΣΤΟΣ 2007

Περίληψη

Στην παρούσα εργασία αναπτύχθηκε ένα σύστημα Αναγνώρισης και Κατάταξης Ονομάτων Οντοτήτων για ελληνικά κείμενα που αποτελεί επέκταση προηγούμενων συστημάτων. Το εν λόγω σύστημα αναγνωρίζει και κατατάσσει ονόματα προσώπων, οργανισμών και τοποθεσιών με τη χρήση τριών ανεξάρτητων Μηχανών Διανυσμάτων Υποστήριξης. Διεξήχθησαν πειράματα τόσο παθητικής όσο και ενεργητικής μάθησης. Στην περίπτωση της ενεργητικής μάθησης, μελετήθηκαν δυο διαφορετικές μέθοδοι επιλογής παραδειγμάτων προς επισημείωση από τον άνθρωπο-εκπαιδευτή. Η εκπαίδευση και ο έλεγχος του συστήματος έγιναν σε μια συλλογή κειμένων του ελληνικού ημερήσιου τύπου. Τα αποτελέσματα έδειξαν ότι εν γένει η ενεργητική μάθηση αποφέρει καλύτερα αποτελέσματα από την παθητική και ότι η κατηγορία των τοποθεσιών χρειάζεται ενδεχομένως ξεχωριστή μελλοντική έρευνα.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών κ. Ίωνα Ανδρουτσόπουλο για την καθοδήγησή του σε όλη τη διάρκεια της παρούσας πτυχιακής εργασίας. Επίσης θα ήθελα να ευχαριστήσω τον υποψήφιο διδάκτορα του ιδίου τμήματος κ. Γεώργιο Λουκαρέλλι για τη συμμετοχή του στην επίβλεψη της εργασίας και ιδιαίτερα για τη βοήθειά του στο ξεκίνημα της εργασίας. Τέλος, θα ήθελα να ευχαριστήσω θερμά τον τελειόφοιτο του Τμήματος Πληροφορικής του Ο.Π.Α. και φίλο κ. Ξενοφώντα Βασιλάκο για την πολύτιμη αρωγή του, τις χρήσιμες συμβουλές του και κυρίως για το χρόνο που διέθεσε σε όλες τις φάσεις της εργασίας αυτής.

Περιεχόμενα

1. Εισαγωγή	
Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων-----	5
Επεκτάσεις προηγούμενων συστημάτων-----	5
Τι θα ακολουθήσει-----	7
2. Περιγραφή Συστήματος	
Μηχανές Διανυσμάτων Υποστήριξης-----	8
Μέτρα Αξιολόγησης-----	8
Αρχιτεκτονική Συστήματος-----	11
Ιδιότητες Διανυσμάτων-----	13
3. Πειράματα	
Συλλογές κειμένων εκπαίδευσης και ελέγχου-----	17
Συντονισμός Παραμέτρων και Επιλογή Ιδιοτήτων-----	18
Παθητική Μάθηση-----	18
Ενεργητική Μάθηση-----	19
Κλασική μέθοδος επιλογής παραδειγμάτων-----	19
Επιγραμμική μέθοδος επιλογής παραδειγμάτων-----	20
Πειραματικά Αποτελέσματα-----	21
Αποτελέσματα Ακρίβειας-----	22
Αποτελέσματα Ανάκλησης-----	24
Αποτελέσματα F-measure-----	26
Μη επικαλυπτόμενες κατηγορίες ονομάτων-----	27
Χρονικά Διαγράμματα Ενεργητικής Μάθησης-----	33
4. Συμπεράσματα και Μελλοντικές Επεκτάσεις-----	36
Βιβλιογραφικές Αναφορές-----	38

1. Εισαγωγή

1.1 Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων

Η Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων (*Named-Entity Recognition and Categorization – NERC*) αποσκοπεί στον εντοπισμό ονομάτων οντοτήτων συγκεκριμένων ειδών σε κείμενα φυσικής γλώσσας. Για παράδειγμα, ένα σύστημα NERC ενδέχεται να έχει εκπαιδευθεί ώστε να εντοπίζει ονόματα προσώπων, οργανισμών, τοποθεσιών, ημερομηνίες κ.ά. σε κείμενα εφημερίδων ή σε επιστημονικά άρθρα. Άλλο σύστημα NERC ενδέχεται να έχει εκπαιδευθεί ώστε να εντοπίζει ονόματα πρωτεϊνών σε ιατρικά κείμενα κ.ο.κ. Η λειτουργία της αναγνώρισης και κατάταξης ονομάτων οντοτήτων αποτελεί απαραίτητο στάδιο προεπεξεργασίας πολλών συστημάτων επεξεργασίας φυσικής γλώσσας, όπως συστήματα ανάκτησης πληροφοριών ή συστήματα ερωταποκρίσεων. Σημαντική έρευνα έχει πραγματοποιηθεί σε αυτό τον τομέα κυρίως σε κείμενα αγγλικής γλώσσας [7].

Στην παρούσα εργασία παρουσιάζουμε ένα ελεύθερα διαθέσιμο σύστημα Αναγνώρισης και Κατάταξης Ονομάτων Οντοτήτων για ελληνικά κείμενα, το οποίο εντοπίζει ονόματα προσώπων, οργανισμών, τοποθεσιών, καθώς επίσης και χρονικές εκφράσεις. Το σύστημα χρησιμοποιεί Μηχανές Διανυσμάτων Υποστήριξης (*MΔΥ, Support Vector Machines, SVMs*) για την αναγνώριση ονομάτων και ημι-αυτόματα παραγόμενα πρότυπα για την αναγνώριση χρονικών εκφράσεων.

1.2 Επεκτάσεις προηγούμενων συστημάτων

Η συγκεκριμένη εργασία αποτελεί επέκταση του ελεύθερα διαθέσιμου συστήματος NERC για ελληνικά κείμενα του Ξενοφώντα Βασιλάκου [3, 4], το οποίο με τη σειρά του αποτέλεσε επέκταση του συστήματος του Γεώργιου Λουκαρέλλι [1, 2].

Το σύστημα του Λουκαρέλλι αναγνώριζε μόνο ονόματα προσώπων και χρονικές εκφράσεις. Οι χρονικές εκφράσεις αναγνωρίζονταν με ημι-αυτόματα παραγόμενα πρότυπα κανονικών εκφράσεων (*regular expression patterns*). Τα ονόματα προσώπων αναγνωρίζονταν με δύο ξεχωριστές σάρωσεις των κειμένων, με τη χρήση δύο ΜΔΥ, μιας για κάθε σάρωση. Κάθε ΜΔΥ εκπαιδευόταν να διαχωρίζει λεκτικές μονάδες που αποτελούν (θετική κατηγορία) μέρη ονομάτων προσώπων από λεκτικές μονάδες που δεν αποτελούν (αρνητική κατηγορία) μέρη ονομάτων προσώπων. Η ΜΔΥ της δεύτερης σάρωσης λάμβανε υπόψη της τα αποτελέσματα της πρώτης, κάτι που μεταξύ άλλων της επέτρεπε να αναγνωρίζει επανεμφάνισης των ονομάτων προσώπων που είχε εντοπίσει η πρώτη σάρωση, αλλά με λιγότερο ενδεικτικά συμφραζόμενα (π.χ. «Κωνσταντίνου» αντί «ο κ. Κωνσταντίνου»).

Το σύστημα του Βασιλάκου μπορούσε επιπλέον να αναγνωρίσει ονόματα οργανισμών. Χρησιμοποιούσε επίσης περισσότερες ιδιότητες (*attributes*) στις ΜΔΥ κατά την αναγνώριση των ονομάτων προσώπων. Για τον εντοπισμό των χρονικών εκφράσεων διατηρήθηκαν τα ημι-αυτόματα παραγόμενα πρότυπα του Λουκαρέλλι, ενώ για τον εντοπισμό και την κατάταξη των ονομάτων οργανισμών προστέθηκαν δύο ακόμη ΜΔΥ, μία για κάθε σάρωση, οδηγώντας σε ένα σύστημα με συνολικά τέσσερις ΜΔΥ. Η εκπαίδευση των δύο ΜΔΥ (ονομάτων προσώπων και οργανισμών)

κάθε σταδίου σάρωσης γινόταν παράλληλα, δηλαδή και οι δύο ΜΔΥ εκπαιδεύονταν στα ίδια παραδείγματα εκπαίδευσης. Επίσης, κατά τη χρήση του συστήματος, αν και οι δύο ΜΔΥ ενός σταδίου σάρωσης θεωρούσαν ότι μια λεκτική μονάδα ανήκει στη θετική τους κατηγορία (όνομα προσώπου και οργανισμού μαζί), η λεκτική μονάδα κατατασσόταν στην κατηγορία ονομάτων της ΜΔΥ με το μεγαλύτερο βαθμό βεβαιότητας (π.χ. μόνο όνομα προσώπου, αν η ΜΔΥ των ονομάτων προσώπων είχε επιστρέψει μεγαλύτερο βαθμό βεβαιότητας). Επομένως, κάθε λεκτική μονάδα κατατασσόταν το πολύ σε μία κατηγορία.

Και τα δύο προηγούμενα συστήματα υποστήριζαν τόσο παθητική μάθηση (PL, *passive learning*) όσο και ενεργητική μάθηση (AL, *active learning*) κατά την εκπαίδευση των ΜΔΥ. Στην παθητική μάθηση, η εκπαίδευση γίνεται σε ένα σύνολο κειμένων στα οποία έχουν επισημειωθεί χειρωνακτικά οι κατηγορίες όλων των λεκτικών μονάδων. Στην ενεργητική μάθηση, αντίθετα, το ίδιο το σύστημα προτείνει στον άνθρωπο-εκπαιδευτή ποιες λεκτικές μονάδες των κειμένων εκπαίδευσης να επισημειώσει, χωρίς να απαιτείται η επισημείωση όλων των λεκτικών μονάδων. Προηγούμενα πειράματα δείχνουν ότι η ενεργητική μάθηση μπορεί να οδηγήσει στις ίδιες ή και καλύτερες επιδόσεις, σε σχέση με την παθητική μάθηση, με λιγότερα παραδείγματα εκπαίδευσης, γεγονός που συνεπάγεται, μεταξύ άλλων, λιγότερη χειρωνακτική δουλειά κατά την εκπαίδευση.

Το σύστημα της παρούσας εργασίας προσθέτει κυρίως τη δυνατότητα αναγνώρισης ονομάτων τοποθεσιών. Λόγω περιορισμένου χρόνου, το σύστημα αναπτύχθηκε ώστε να χρησιμοποιεί μία μόνο σάρωση, αλλά είναι δυνατόν στο μέλλον να επεκταθεί, ώστε να χρησιμοποιεί δύο σαρώσεις όπως τα προηγούμενα συστήματα. Το νέο σύστημα διατηρεί αμετάβλητη τη μέθοδο αναγνώρισης χρονικών εκφράσεων του Λουκαρέλλι, αλλά προσθέτει μία ακόμα ΜΔΥ, για την κατηγορία των τοποθεσιών. Συνολικά πλέον έχουμε τρεις ΜΔΥ, μία για κάθε κατηγορία ονομάτων (πρόσωπα, οργανισμοί, τοποθεσίες) για μία μόνο σάρωση. Κάθε ΜΔΥ εκπαιδεύεται πλέον ξεχωριστά από τις υπόλοιπες (διαφορετικά παραδείγματα εκπαίδευσης αλλά και ξεχωριστές ιδιότητες), κάτι που κάνει ευκολότερη τη μελλοντική υποστήριξη νέων κατηγοριών ονομάτων. Επιπλέον, υποστηρίζεται η κατάταξη ενός ονόματος σε παραπάνω από μία κατηγορίες (π.χ. σε ορισμένες περιπτώσεις «η Βουλή» μπορεί να αποτελεί ταυτόχρονα όνομα οργανισμού και τοποθεσία). Διερευνήθηκε επίσης η χρήση μίας καινούριας επιγραμμικής (online) μεθόδου επιλογής διανυσμάτων προς επισημείωση κατά την ενεργητική μάθηση.

Η εκπαίδευση του νέου συστήματος έγινε σε κείμενα του ελληνικού ημερήσιου τύπου με θεματολογία που περιελάμβανε πολιτικές και οικονομικές ειδήσεις, εικαστικά και αθλητικά. Σε γενικές γραμμές, η ενεργητική μάθηση απέφερε υψηλότερες επιδόσεις από την παθητική, γεγονός που επιβεβαιώνει το αντίστοιχο συμπέρασμα των Λουκαρέλλι και Βασιλάκου. Ακόμα, τα αποτελέσματα των πειραμάτων επιβεβαίωσαν την παρατήρηση του Βασιλάκου ότι τα ονόματα των οργανισμών είναι δυσκολότερο να εντοπισθούν από εκείνα των προσώπων. Τα αποτελέσματα δείχνουν, επίσης, ότι η κατηγορία των ονομάτων τοποθεσιών είναι η δυσκολότερη των τριών.

1.3 Τι θα ακολουθήσει

Κεφάλαιο 2: Περιγραφή συστήματος

Σε αυτό το κεφάλαιο θα δοθούν ορισμοί για τις ΜΔΥ και τα χρησιμοποιούμενα μέτρα αξιολόγησης. Θα περιγραφεί, επίσης, συνοπτικά η αρχιτεκτονική του συστήματος. Στη συνέχεια, θα παρουσιαστούν οι δύο βασικές διαφοροποιήσεις από το σύστημα του Βασιλάκου: η απεμπλοκή των κατηγοριών των προσώπων και των οργανισμών και η προσθήκη της κατηγορίας των τοποθεσιών. Θα ακολουθήσει παράθεση και αξιολόγηση των ιδιοτήτων κάθε ΜΔΥ.

Κεφάλαιο 3: Πειράματα

Στο κεφάλαιο αυτό παρέχονται αρχικά πληροφορίες για τις συλλογές κειμένων που χρησιμοποιήθηκαν κατά την εκπαίδευση και την αξιολόγηση του συστήματος. Έπειτα, περιγράφεται η διεξαγωγή των πειραμάτων με τις διάφορες μεθόδους που μελετήθηκαν και παρατίθενται τα διαγράμματα των αποτελεσμάτων μαζί με σχόλια και παρατηρήσεις.

Κεφάλαιο 4: Συμπεράσματα και μελλοντικές επεκτάσεις

Σε αυτό το κεφάλαιο συνοψίζονται τα αποτελέσματα της εργασίας και προτείνονται μελλοντικές βελτιώσεις του συστήματος.

2. Περιγραφή Συστήματος

2.1 Μηχανές Διανυσμάτων Υποστήριξης

Οι Μηχανές Διανυσμάτων Υποστήριξης είναι μία μέθοδος επιβλεπόμενης (supervised) μηχανικής μάθησης. Στην απλούστερη μορφή τους, υποστηρίζουν μόνο δύο κατηγορίες (θετική και αρνητική). Στην περίπτωσή μας, όπου τα αντικείμενα προς κατάταξη είναι λεκτικές μονάδες (tokens) και κάθε ΜΔΥ αντιστοιχεί σε μια κατηγορία ονομάτων (προσώπων, οργανισμών ή τοποθεσιών), η θετική κατηγορία κάθε ΜΔΥ αντιστοιχεί σε λεκτικές μονάδες που ανήκουν στην κατηγορία ονομάτων για την οποία χρησιμοποιείται η ΜΔΥ, ενώ η αρνητική κατηγορία περιλαμβάνει όλες τις υπόλοιπες λεκτικές μονάδες. Η κάθε λεκτική μονάδα αναπαρίσταται εσωτερικά με ένα διάνυσμα ιδιοτήτων που παρέχει πληροφορίες για την ίδια τη λεκτική μονάδα (π.χ. αν αρχίζει με κεφαλαίο γράμμα ή όχι, αν περιλαμβάνεται ή όχι σε λίστα γνωστών κύριων ονομάτων) και τα συμφραζόμενά της (π.χ. γειτονικές λεκτικές μονάδες).

Συνοπτικά, η λειτουργία των ΜΔΥ είναι να προβάλλουν τα διανύσματα ιδιοτήτων των παραδειγμάτων εκπαίδευσης σε ένα χώρο περισσότερων διαστάσεων και στη συνέχεια να βρίσκουν ένα γραμμικό διαχωριστή (εν γένει ένα υπερ-επίπεδο), που να τα διαχωρίζει στις δυο κατηγορίες (θετική και αρνητική). Μετά την εκπαίδευση, η κατάταξη νέων αντικειμένων γίνεται προβάλλοντάς τα στο χώρο περισσότερων διαστάσεων και υπολογίζοντας σε ποια πλευρά του γραμμικού διαχωριστή βρίσκονται. Η προβολή των διανυσμάτων σε χώρο περισσότερων διαστάσεων αυξάνει την πιθανότητα οι δύο κατηγορίες να είναι γραμμικά διαχωρίσιμες. Λεπτομερέστερη εισαγωγή στις ΜΔΥ γίνεται από τον Λουκαρέλλι [1] και από τους Cristianini και Shawe-Taylor [5].

2.2 Μέτρα αξιολόγησης

Η **ακρίβεια** (*Precision*) μίας κατηγορίας ονομάτων είναι ο λόγος των λεκτικών μονάδων που κατετάγησαν ορθά στην κατηγορία (True Positive) προς το σύνολο των λεκτικών μονάδων που κατετάγησαν σε αυτήν είτε ορθά είτε λανθασμένα (True Positive + False Positive).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2.1)$$

Η **ανάκληση** (*Recall*) μίας κατηγορίας ονομάτων είναι ο λόγος των λεκτικών μονάδων που κατετάγησαν ορθά στην κατηγορία (True Positive) προς το σύνολο των λεκτικών μονάδων της κατηγορίας που υπάρχουν στην πραγματικότητα στα κείμενα ελέγχου, δηλαδή προς το σύνολο των λεκτικών μονάδων που κατετάγησαν σωστά στην κατηγορία και των λεκτικών μονάδων που λανθασμένα δεν κατετάγησαν στην κατηγορία (True Positive + False Negative)

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2.2)$$

Το **F-measure** είναι ένας συνδυασμός της ακρίβειας και της ανάκλησης. Ο γενικός τύπος είναι:

$$\text{F-measure} = \frac{(b^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{b^2 \cdot \text{Precision} + \text{Recall}} \quad (2.3)$$

Στην παρούσα εργασία, όπου γίνεται στο εξής αναφορά στο F-measure θα εννοούμε τον τύπο του F_1 , που δίνει ίση βαρύτητα στην ακρίβεια και την ανάκληση:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.4)$$

Για την επιλογή των διανυσμάτων προς επισημείωση στην ενεργητική μάθηση, χρησιμοποιήθηκε το μέτρο του **Πληροφοριακού Οφέλους** (*Informativeness*). Το Πληροφοριακό Όφελος ενός διανύσματος εκπαίδευσης ουσιαστικά είναι μια ένδειξη της βεβαιότητας της ΜΔΥ για την κατηγορία (θετική ή αρνητική) του διανύσματος. Προηγούμενες εργασίες (π.χ. [4, 6]) δείχνουν ότι είναι επωφελές να επιλέγονται παραδείγματα εκπαίδευσης για την κατηγορία των οποίων είναι περισσότερο αβέβαιη η ΜΔΥ.

$$\text{Informativeness}(\vec{u}) = 1 - \text{Dist}(SVM_{hyp, \vec{u}}) \quad (2.5)$$

$$\text{Dist}(SVM_{hyp, \vec{u}}) = |P_{pos}(\vec{u}) - P_{neg}(\vec{u})| \quad (2.6)$$

Στους παραπάνω τύπους, $P_{pos}(\vec{u})$ και $P_{neg}(\vec{u})$ είναι εκτιμήσεις των πιθανοτήτων το διάνυσμα \vec{u} να είναι θετικό ή αρνητικό αντίστοιχα.¹ Εναλλακτικά, θα μπορούσε το πληροφοριακό όφελος να υπολογισθεί κατευθείαν ως η απόσταση του διανύσματος \vec{u} από το υπερ-επίπεδο διαχωρισμού της ΜΔΥ.² Οι παραπάνω τύποι, που χρησιμοποιούν εκτιμήσεις των πιθανοτήτων $P_{pos}(\vec{u})$ και $P_{neg}(\vec{u})$, είχαν χρησιμοποιηθεί τόσο στην εργασία του Λουκαρέλλι [1] όσο και στην εργασία του Βασιλάκου [3].

Για την επιλογή των ιδιοτήτων των διανυσμάτων κάθε μίας ΜΔΥ χρησιμοποιήθηκε το μέτρο του **Πληροφοριακού Κέρδους** (*Information Gain*), όπως ορίζεται, μεταξύ άλλων, στις προηγούμενες εργασίες [1, 3]. Συνοπτικά, το

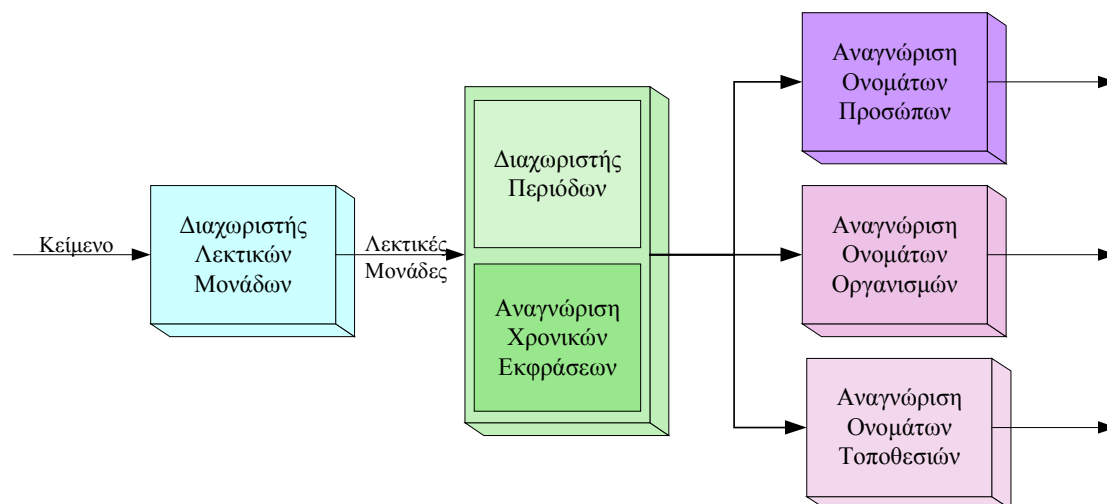
¹ Η υλοποίηση ΜΔΥ που χρησιμοποιήθηκε έχει τη δυνατότητα να επιστρέφει εκτιμήσεις αυτού του είδους, οι οποίες συνήθως παράγονται προσαρμόζοντας μια σιγμοειδή συνάρτηση στις αποστάσεις των διανυσμάτων από το υπερ-επίπεδο της ΜΔΥ. Οι παράμετροι της σιγμοειδούς είναι δυνατόν να επιλεγούν μέσω διασταυρωμένης επικύρωσης (cross-validation) κατά την εκπαίδευση.

² Χρησιμοποιώντας απλώς την απόσταση από το υπερ-επίπεδο διαχωρισμού, θα αποφεύγαμε τα σφάλματα και την επιπλέον καθυστέρηση κατά την εκπαίδευση (λόγω της απαιτούμενης διασταυρωμένης επικύρωσης) που εισάγει η εκτίμηση των πιθανοτήτων. Θα ήταν σκόπιμο να δοκιμασθεί αυτή η μέθοδος υπολογισμού του πληροφοριακού οφέλους σε μελλοντική εργασία.

πληροφοριακό κέρδος μιας ιδιότητας είναι η αναμενόμενη μείωση της εντροπίας της τυχαίας μεταβλητής της κατηγορίας (στην περίπτωσή μας, θετική ή αρνητική), που προκαλείται όταν γίνει γνωστή η τιμή της ιδιότητας. Ιδιότητες με μεγάλο πληροφοριακό κέρδος είναι γενικά πιο χρήσιμες.

2.3 Αρχιτεκτονική Συστήματος

Το σύστημα της παρούσας εργασίας αποτελείται από 6 διαφορετικές μονάδες, όπως φαίνεται και στο σχήμα 2.1.



Σχήμα 2.1

Κάθε κείμενο (εκπαίδευσης ή αξιολόγησης) περνάει πρώτα από το **Διαχωριστή Λεκτικών Μονάδων**, ο οποίος το διασπά σε λεκτικές μονάδες (tokens), αποθηκεύοντας αυτές σε μία λίστα. Ως λεκτική μονάδα θεωρούμε κάθε ακολουθία ελληνικών, λατινικών ή αριθμητικών χαρακτήρων. Οποιοσδήποτε άλλος μη κενός χαρακτήρας αποτελεί λεκτική μονάδα από μόνος του. Παραδείγματος χάριν τα αποσιωπητικά («...») θεωρούνται τρεις λεκτικές μονάδες.

Στη συνέχεια το κείμενο περνάει από το **Διαχωριστή Περιόδων** και τη μονάδα **Αναγνώρισης Χρονικών Εκφράσεων**. Επειδή και οι δύο μονάδες επιδρούν μόνο στο αποτέλεσμα του Διαχωριστή Λεκτικών Μονάδων, δεν έχει σημασία η σειρά εκτέλεσής τους. Ο Διαχωριστής Περιόδων, που χρησιμοποιεί μια δική του ΜΔΥ, εντοπίζει τα τέλη των περιόδων διαχωρίζοντας τις τελείες που σηματοδοτούν τέλη περιόδων από τελείες που σηματοδοτούν συντομογραφίες (π.χ. «κ. Σημίτης») ή που έχουν άλλες χρήσεις (π.χ. διαχωριστικά ημερομηνιών κλπ.). Φυσικά είναι δυνατόν κάποιες περιόδους να μην τελειώνουν με τελείες (π.χ. να τελειώνουν με ερωτηματικά), αλλά αυτές οι περιπτώσεις είναι πολύ σπάνιες στα κείμενα ειδήσεων με τα οποία κυρίως ασχολούμαστε και δεν επηρεάζουν σημαντικά τις επιδόσεις του συνολικού συστήματος, όπως εξηγείται στην εργασία του Λουκαρέλλι [1]. Η μονάδα Αναγνώρισης Χρονικών Εκφράσεων εντοπίζει χρονικές εκφράσεις χρησιμοποιώντας ημι-αυτόματα παραγόμενα πρότυπα κανονικών εκφράσεων. Ο Διαχωριστής Λεκτικών Μονάδων, ο Διαχωριστής Περιόδων και η μονάδα Αναγνώρισης Χρονικών Εκφράσεων λειτουργούν ακριβώς όπως στο σύστημα του Λουκαρέλλι [1] και δεν περιγράφονται περαιτέρω στην παρούσα εργασία.

Τέλος, η λίστα των λεκτικών μονάδων, στην οποία έχουν προστεθεί πληροφορίες για τα τέλη των περιόδων και τις χρονικές εκφράσεις, περνάει από τις μονάδες **Αναγνώρισης Ονομάτων Προσώπων, Οργανισμών και Τοποθεσιών**. Και πάλι δεν είναι απαραίτητο η εκτέλεση των τριών αυτών μονάδων να γίνει με συγκεκριμένη σειρά, καθώς οι τρεις μονάδες είναι εντελώς ανεξάρτητες μεταξύ τους. Είναι, επίσης, δυνατόν να παραληφθεί η εκτέλεση κάποιων από τις τρεις αυτές μονάδες, αν δεν μας ενδιαφέρει η αναγνώριση των ονομάτων των αντιστοίχων κατηγοριών. Η λειτουργία των μονάδων Αναγνώρισης Ονομάτων Προσώπων και Οργανισμών είναι πολύ παρόμοια με τη λειτουργία των αντιστοίχων μονάδων των εργασιών των Λουκαρέλλι και Βασιλάκου [1, 3]. Ως εκ τούτου, οι δύο αυτές μονάδες περιγράφονται πολύ περιληπτικά στην παρούσα εργασία. Η αναγνώριση ονομάτων τοποθεσιών γίνεται με μία ξεχωριστή ΜΔΥ, όπως ακριβώς στις άλλες δύο κατηγορίες ονομάτων, αλλά χρησιμοποιώντας διαφορετικές ιδιότητες.³ Οι ιδιότητες και των τριών ΜΔΥ παρατίθενται στην επόμενη ενότητα.

Δύο σημαντικές διαφορές από τα συστήματα των προηγούμενων εργασιών είναι πως χρησιμοποιείται μόνο μία σάρωση του κειμένου και οι ΜΔΥ των μονάδων Αναγνώρισης Ονομάτων Προσώπων, Οργανισμών και Τοποθεσιών εκπαιδεύονται πλέον εντελώς ανεξάρτητα, με διαφορετικά εν γένει παραδείγματα εκπαίδευσης. Επίσης, κάθε μία ΜΔΥ χρησιμοποιεί μόνο ιδιότητες που σχετίζονται άμεσα με την αναγνώριση ονομάτων της κατηγορίας της. Αντίθετα, στις προηγούμενες εργασίες η ΜΔΥ των ονομάτων οργανισμών χρησιμοποιούσε, για παράδειγμα, και ιδιότητες που έλεγχαν αν η προς κατάταξη λεκτική μονάδα έχει στα αριστερά της εκφράσεις όπως «κ.» (κύριος), οι οποίες συνοδεύουν συνήθως ονόματα προσώπων. Αντίστοιχα, η ΜΔΥ των ονομάτων προσώπων χρησιμοποιούσε και ιδιότητες που έλεγχαν αν η προς κατάταξη λεκτική μονάδα ακολουθείται από εκφράσεις όπως «Α.Ε.», οι οποίες σηματοδοτούν συνήθως ονόματα οργανισμών. Η απεμπλοκή των ΜΔΥ (ξεχωριστή εκπαίδευση και ξεχωριστές ιδιότητες) έγινε για να είναι ευκολότερη η προσθήκη ΜΔΥ αναγνώρισης νέων κατηγοριών ονομάτων στο μέλλον, χωρίς να απαιτείται η επανεκπαίδευση των ΜΔΥ των υπαρχόντων κατηγοριών ονομάτων. Η χρήση μιας μόνο σάρωσης έγινε λόγω έλλειψης χρόνου κατά την ανάπτυξη του νέου συστήματος. Στο μέλλον θα ήταν σκόπιμο να προστεθεί και δεύτερη σάρωση, όπως στις προηγούμενες εργασίες.

Μια ακόμη διαφορά από τις προηγούμενες εργασίες είναι πως πλέον είναι δυνατόν μια λεκτική μονάδα να καταταγεί σε περισσότερες από μία κατηγορίες ονομάτων. Για την ακρίβεια, κάθε λεκτική μονάδα κατατάσσεται στις κατηγορίες ονομάτων των οποίων οι ΜΔΥ κατέταξαν τη λεκτική μονάδα στη θετική τους κατηγορία. Εναλλακτικά, ο χρήστης μπορεί να επιλέξει κάθε λεκτική μονάδα να κατατάσσεται σε μία μόνο κατηγορία ονομάτων, όπως στις προηγούμενες εργασίες. Στην περίπτωση αυτή, αν περισσότερες από μία ΜΔΥ κατατάξουν μια λεκτική μονάδα στη θετική τους κατηγορία, η λεκτική μονάδα κατατάσσεται στην κατηγορία ονομάτων που αντιστοιχεί στην ΜΔΥ που επέστρεψε το μεγαλύτερο βαθμό βεβαιότητας για την απόφασή της.

³ Κατά την αναγνώριση ονομάτων τοποθεσιών χρησιμοποιούνται, επίσης, «ασφαλείς κανόνες» και λίστες «τετριμμένων λέξεων», αντίστοιχες εκείνων που χρησιμοποιούνταν κατά την αναγνώριση ονομάτων προσώπων και οργανισμών στις προηγούμενες εργασίες [1, 3].

2.4 Ιδιότητες Διανυσμάτων

Για τις κατηγορίες των προσώπων και των οργανισμών έχουμε διατηρήσει τις περισσότερες ιδιότητες που είχε χρησιμοποιήσει ο Βασιλάκος στο σύστημά του. Ωστόσο, από τις ιδιότητες που χρησιμοποιούσε η ΜΔΥ των προσώπων έχουμε αφαιρέσει 15 ιδιότητες που αναφέρονταν στους οργανισμούς και αντίστοιχα έχουμε αφαιρέσει 28 ιδιότητες που χρησιμοποιούσε η ΜΔΥ των οργανισμών. Πιο αναλυτικά, από τις 165 ιδιότητες πλέον έχουμε 150 για τα πρόσωπα και για τους οργανισμούς από 170 έχουμε συνολικά 142. Για την κατηγορία των τοποθεσιών χρησιμοποιούμε 133 ιδιότητες για κάθε διάνυσμα.

Ακολουθούν πίνακες με τις ιδιότητες που χρησιμοποιεί η κάθε μία από τις τρεις ΜΔΥ (ονομάτων προσώπων, οργανισμών, τοποθεσιών). Συμβολίζουμε με t_0 τη λεκτική μονάδα που έχουμε να κατατάξουμε, με t_1 την επόμενη της λεκτική μονάδα, με t_{-1} την προηγούμενη κ.ο.κ. Για παράδειγμα, η τρίτη ιδιότητα του πρώτου πίνακα εξετάζει αν η προηγούμενη λεκτική μονάδα είναι κόμμα. Κάθε κελί περιέχει μια κουκίδα, το μέγεθος της οποίας αναπαριστά το Πληροφοριακό Κέρδος της αντίστοιχης ιδιότητας. Όσο μεγαλύτερη είναι η κουκίδα, τόσο μεγαλύτερο είναι και το πληροφοριακό κέρδος της ιδιότητας. Αν στη θέση της κουκίδας υπάρχει το «0», τότε η ιδιότητα αυτή έχει πολύ μικρό πληροφοριακό κέρδος και δε χρησιμοποιείται στο σύστημα καθόλου (βλ. ενότητα 3.2). Οι ιδιότητες που χρησιμοποιούνταν στις ΜΔΥ των ονομάτων προσώπων και οργανισμών εξηγούνται αναλυτικά στις προηγούμενες εργασίες [1, 3].

SVM for Person Names									
no.	Feature Descriptions	t_{-3}	t_{-2}	t_{-1}	t_0	t_{+1}	t_{+2}	t_{+3}	
1-7	comma?	●	•	●	●	•	•	•	B
8-14	full stop?	●	•	●	●	●	•	•	B
15-21	dash?	•	•	•	•	•	•	•	B
22-28	slash?	•	0	•	0	•	•	0	B
29-35	number?	•	•	•	0	●	•	•	B
36-42	Greek characters?	●	●	•	●	●	●	•	B
43-49	Latin characters?	•	•	•	●	•	•	•	B
50-56	first character capital?	●	●	•	●	●	●	•	B
57-63	all characters capital?	•	●	●	●	•	•	•	B
64-70	length in characters?	●	●	●	●	●	•	•	B
71-77	common surname prefix?	•	•	•	●	•	•	•	B
78-84	common surname suffix?	●	●	•	●	●	•	•	B
85-91	common person first name?	●	●	●	●	•	0	•	B
92-98	common last character?	●	•	•	•	•	•	•	B
99-105	common sing. adj. ending?	•	•	•	●	•	•	•	B
106-112	plural noun/adj. ending?	•	●	•	•	•	•	•	B
113-119	common sing. gen. ending?	•	•	•	•	•	•	•	B
120-126	ends in final sigma?	•	•	•	•	•	•	•	B
127-133	is last token of sentence?	•	•	●	•	•	•	•	B
134-140	part of article's title?	•	•	•	•	•	•	•	B
141-147	distance from start of name?				●				B
148	directly preceded by Mr(s)?				●				B
149	preceded by plural Mr(s)?				●				B
150	in P_{1-2}^{t-1} list?			●					n
151	in P_{3-4}^{t-1} list?			•					n
152	in $P_{>4}^{t-1}$ list?			●					n
153	prev. tokens in $P_{1-2}^{t-7, \dots, t-1}$?				●				n
154	prev. tokens in $P_{3-4}^{t-7, \dots, t-1}$?				•				n
155	prev. tokens in $P_{>4}^{t-7, \dots, t-1}$?				•				n

B: Boolean, n: numeric, 0: $IG=0$, •: $0 < IG \leq 0.01$, ●: $0.01 < IG \leq 0.1$, ●: $IG > 0.1$

Πίνακας 2.5.1 – Ιδιότητες της ΜΔΥ Ονομάτων Προσώπων

Οι ιδιότητες 150–152 αντιστοιχούν σε λίστες λεκτικών μονάδων που συναντώνται συχνά αμέσως πριν από λεκτικές μονάδες ονομάτων προσώπων στα δεδομένα εκπαίδευσης. Η λίστα P_{1-2}^{t-1} περιέχει λεκτικές μονάδες μήκους ενός ή δύο χαρακτήρων, η λίστα P_{3-4}^{t-1} περιέχει λεκτικές μονάδες μήκους τριών ή τεσσάρων χαρακτήρων και η λίστα $P_{>4}^{t-1}$ λεκτικές μονάδες με μήκος μεγαλύτερο των τεσσάρων χαρακτήρων. Οι ιδιότητες 153–155 αντιστοιχούν σε παρόμοιες με τις προηγούμενες τρεις λίστες, αλλά στην περίπτωση αυτή οι λίστες περιλαμβάνουν τις συχνές λεκτικές μονάδες που βρίσκονται σε ένα παράθυρο 7 λεκτικών μονάδων πριν από λεκτικές μονάδες ονομάτων προσώπων στα δεδομένα εκπαίδευσης. Περισσότερες

πληροφορίες για αυτές τις ιδιότητες παρέχονται στις εργασίες των Λουκαρέλλι και Βασιλάκου [1, 3].

SVM for Organization Names									
no.	Feature Descriptions	t_{-3}	t_{-2}	t_{-1}	t_0	t_{+1}	t_{+2}	t_{+3}	
1-7	comma?	•	•	●	●	•	•	•	B
8-14	full stop?	●	•	•	•	•	●	●	B
15-21	dash?	•	•	•	•	•	•	•	B
22-28	slash?	•	•	•	•	•	•	0	B
29-35	number?	•	•	•	•	•	•	•	B
36-42	Greek characters?	●	●	●	●	•	●	•	B
43-49	Latin characters?	•	●	●	●	●	●	●	B
50-56	first character capital?	●	●	•	●	●	●	●	B
57-63	all characters capital?	•	•	•	●	•	●	●	B
64-70	length in characters?	●	●	●	●	0	0	0	n
71-77	common last character?	●	●	●	●	•	●	●	B
78-84	common sing. adj. ending?	●	●	•	●	•	•	•	B
85-91	plural noun/adj. ending?	•	●	●	•	•	•	•	B
92-98	common sing. gen. ending?	•	•	●	●	●	•	•	B
99-105	ends in final sigma?	●	●	●	•	•	•	•	B
106-112	distance from "A.E." etc.?	●	●	•	●	●	●	●	n
113-119	starts with "ministry" etc.?	•	•	•	●	●	•	•	B
120-126	is last token of sentence?	●	●	●	●	•	•	•	B
127-133	part of article's title?	•	•	•	●	●	●	●	B
134-140	distance from start of name?				●				B
141	in R_{1-2}^{t-1} list?			●					n
142	in R_{3-4}^{t-1} list?			●					n
143	in $R_{>4}^{t-1}$ list?			•					n
144	prev. tokens in $R_{1-2}^{t-7, \dots, t-1}$?				●				n
145	prev. tokens in $R_{3-4}^{t-7, \dots, t-1}$?				●				n
146	prev. tokens in $R_{>4}^{t-7, \dots, t-1}$?				●				n

B: Boolean, n: numeric, 0: $IG=0$, •: $0 < IG \leq 0.01$, ●: $0.01 < IG \leq 0.1$, ●: $IG > 0.1$

Πίνακας 2.5.2 – Ιδιότητες της ΜΔΥ Ονομάτων Οργανισμών

Οι ιδιότητες 141–146 είναι αντίστοιχες των ιδιοτήτων 150–155 του προηγούμενου πίνακα, αλλά σε αυτή την περίπτωση οι λίστες περιέχουν λεκτικές μονάδες που εμφανίζονται συχνά πριν από ονόματα οργανισμών.

SVM for Location Names									
no.	Feature Descriptions	t_{-3}	t_{-2}	t_{-1}	t_0	t_{+1}	t_{+2}	t_{+3}	
1-7	comma?	•	•	•	•	•	•	•	B
8-14	full stop?	•	•	•	•	•	•	•	B
15-21	dash?	•	•	•	•	•	•	•	B
22-28	slash?	•	0	•	0	•	•	•	B
29-35	number?	•	•	•	•	•	•	•	B
36-42	Greek characters?	●	●	●	•	•	•	•	B
43-49	Latin characters?	•	•	•	•	•	•	•	B
50-56	first character capital?	•	•	•	•	•	•	•	B
57-63	all characters capital?	•	•	•	●	•	•	•	B
64-70	length in characters?	●	●	●	●	●	•	•	n
71-77	common location first name?	0	•	•	●	0	0	0	n
78-84	common last character?	●	●	●	•	•	•	•	B
85-91	common sing. adj. ending?	•	•	•	•	•	•	•	B
92-98	plural noun/adj. ending?	•	•	•	•	•	•	•	B
99-105	common sing. gen. ending?	•	•	•	•	•	•	•	B
106-112	ends in final sigma?	•	•	•	•	•	•	•	B
113-119	is last token of sentence?	•	•	●	•	•	•	•	B
120-126	part of article's title?	•	•	•	•	•	•	•	B
127-133	distance from start of name?				•				B
134	preceded by "avenue" etc.?				0				B
135	in L_{1-2}^{t-1} list?			•					n
136	in L_{3-4}^{t-1} list?			●					n
137	in $L_{>4}^{t-1}$ list?			•					n
138	prev. tokens in $L_{1-2}^{t-7, \dots, t-1}$?				•				n
139	prev. tokens in $L_{3-4}^{t-7, \dots, t-1}$?				●				n
140	prev. tokens in $L_{>4}^{t-7, \dots, t-1}$?				•				n

B: Boolean, n: numeric, 0: $IG=0$, •: $0 < IG \leq 0.01$, ●: $0.01 < IG \leq 0.1$, ●: $IG > 0.1$

Πίνακας 2.5.3 – Ιδιότητες της ΜΔΥ Ονομάτων Τοποθεσιών

Οι ιδιότητες 135–140 είναι αντίστοιχες των ιδιοτήτων 141–146 του προηγούμενου πίνακα. Παρατηρούμε στον πίνακα 2.5.3 ότι οι ιδιότητες για την κατηγορία των τοποθεσιών έχουν πολύ χαμηλότερο Πληροφοριακό Κέρδος από ό,τι οι αντίστοιχες των άλλων κατηγοριών.

3. Πειράματα

3.1 Συλλογές κειμένων εκπαίδευσης και ελέγχου

Η διεξαγωγή των πειραμάτων έγινε σε κείμενα από τον ημερήσιο τύπο με άρθρα από τις εφημερίδες «Το Βήμα» και «Τα Νέα». Όπως προαναφέρθηκε, η θεματολογία ποικίλλει από πολιτικές και οικονομικές ειδήσεις μέχρι εικαστικά, αθλητικά κ.ά. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι δύο δεξαμενές κειμένων των 400 και 5000 κειμένων αντίστοιχα από την εργασία του Λουκαρέλλι [1]. Η πρώτη δεξαμενή αποτελείται από 200 κείμενα από την εφημερίδα «Το Βήμα» και 200 κείμενα από την εφημερίδα «Τα Νέα», με μέσο μέγεθος άρθρου 6,3 KB και είναι πλήρως επισημειωμένη ως προς τα ονόματα προσώπων, οργανισμών και τοποθεσιών. Ωστόσο, ενώ η επισημείωση του Λουκαρέλλι επέτρεπε σε κάθε λεκτική μονάδα να ανήκει σε μία το πολύ κατηγορία ονομάτων, πλέον υπάρχει η δυνατότητα μια λεκτική μονάδα να ανήκει σε περισσότερες από μία κατηγορίες ονομάτων. Ως εκ τούτου, χρειάστηκε να επανεξεταστούν τα κείμενα αυτής της δεξαμενής και να τροποποιηθούν οι επισημειώσεις τους.

Η γενική πολιτική που ακολουθήσαμε ήταν να κατατάσσουμε κάθε λεκτική μονάδα σε μία μόνο κατηγορία, ανάλογα με τα συμφραζόμενα. Για παράδειγμα, στην πρόταση «χθες συνεδρίασε η Βουλή για την ψήφιση του νέου νόμου για το ασφαλιστικό», η λεκτική μονάδα «Βουλή» επισημειώθηκε μόνο ως όνομα οργανισμού, ενώ στην πρόταση «διαδήλωση συνταξιούχων πραγματοποιήθηκε σήμερα το πρωί μπροστά από τη Βουλή» επισημειώθηκε μόνο ως όνομα τοποθεσίας. Επίσης, στην πρόταση «η μετοχή της Παπαδόπουλος Α.Ε. σημείωσε σήμερα άνοδο 5,6%», η λέξη «Παπαδόπουλος» επισημειώθηκε ως οργανισμός, παρόλο που αποτελεί σύνηθες όνομα προσώπου. Σε περιπτώσεις, όμως, που θεωρούσαμε ότι κάποιες λεκτικές μονάδες μπορούσαν να ανήκουν σε περισσότερες από μία κατηγορίες ονομάτων, τις κατατάσσαμε σε όλες αυτές τις κατηγορίες. Για παράδειγμα στην πρόταση «ακυρώθηκαν 13 πτήσεις σήμερα στο αεροδρόμιο Ελ. Βενιζέλος, προκαλώντας αναστάτωση στους επιβάτες», το «Ελ. Βενιζέλος» επισημειώθηκε τόσο ως όνομα τοποθεσίας όσο και ως όνομα οργανισμού.

Για τους σκοπούς των πειραμάτων η πρώτη δεξαμενή χωρίστηκε σε δύο τμήματα με 198 και 202 άρθρα αντίστοιχα. Το πρώτο τμήμα χρησιμοποιήθηκε για την εκπαίδευση του συστήματος κατά την παθητική μάθηση, για το συντονισμό των παραμέτρων (parameter tuning) του συστήματος, καθώς και για την επιλογή των ιδιοτήτων σύμφωνα με το πληροφοριακό τους κέρδος. Το δεύτερο τμήμα χρησιμοποιήθηκε για τον έλεγχο του συστήματος.

Η δεύτερη δεξαμενή, των 5000 κειμένων, αποτελείται από μη επισημειωμένα άρθρα. Από αυτήν επιλέγονταν τα προς επισημείωση παραδείγματα εκπαίδευσης της ενεργητικής μάθησης. Παρακάτω θα δοθούν περισσότερες λεπτομέρειες σχετικά με τις διάφορες μεθόδους επιλογής παραδειγμάτων ενεργητικής μάθησης που μελετήθηκαν σε αυτή την εργασία.

3.2 Συντονισμός Παραμέτρων και Επιλογή Ιδιοτήτων

Χρησιμοποιήσαμε ΜΔΥ με πυρήνα Radial Basis Function, με αποτέλεσμα να υπάρχουν δύο παράμετροι (C και γ) των οποίων έπρεπε να επιλέξουμε τις τιμές σε κάθε ΜΔΥ. Η επιλογή αυτών των τιμών (*parameter tuning*, συντονισμός παραμέτρων) έγινε όπως στις εργασίες των Λουκαρέλλι και Βασιλάκου [1, 3], χρησιμοποιώντας τη μέθοδο αναζήτησης πλέγματος (*grid search*) και πενταπλής διασταυρωμένης επικύρωσης (*5-fold cross-validation*) που παρέχεται από την υλοποίηση LibSVM [8] που χρησιμοποιήσαμε. Κατά το συντονισμό παραμέτρων χρησιμοποιήθηκαν 14 κείμενα από το τμήμα εκπαίδευσης της πρώτης δεξαμενής, από όπου αντλήθηκαν 1115 διανύσματα για την κατηγορία των προσώπων, 1033 για την κατηγορία των οργανισμών και 969 για την κατηγορία των τοποθεσιών. Τα αποτελέσματα του συντονισμού για κάθε ΜΔΥ ήταν τα ακόλουθα:

Κατηγορία	Παράμετρος C	Παράμετρος γ
Πρόσωπα	4.0	0.018581361171917516
Οργανισμοί	1.6817928305074292	0.022097086912079608
Τοποθεσίες	2.378414230005442	0.03125

Πίνακας 3.2.1 – Παράμετροι Συντονισμού

Οι ιδιότητες των ΜΔΥ προσώπων και οργανισμών διατηρήθηκαν ίδιες όπως στο προηγούμενο σύστημα του Βασιλάκου, αλλά αφαιρέθηκαν ιδιότητες που δεν ήταν άμεσα σχετικές με τον εντοπισμό ονομάτων των αντιστοίχων κατηγοριών (βλ. ενότητα 2.3) Για την επιλογή των ιδιοτήτων της ΜΔΥ των τοποθεσιών υιοθετήθηκε και πάλι η μέθοδος του προηγούμενου συστήματος. Εν συντομία, οι ιδιότητες ταξινομούνται κατά φθίνουσα σειρά βάσει του πληροφοριακού τους κέρδους και απορρίπτονται οι k χειρότερες. Επιλέγεται η τιμή του k που μεγιστοποιεί το F-measure της ΜΔΥ σε ένα τμήμα του συνόλου εκπαίδευσης, το οποίο χρησιμοποιείται ως σύνολο επικύρωσης (*validation set*) και εξαιρείται από την εκπαίδευση της ΜΔΥ. Για περισσότερες πληροφορίες, συμβουλευτείτε την εργασία του Βασιλάκου [3].

3.3 Παθητική Μάθηση

Τα πειράματα παθητικής μάθησης (*passive learning*, PL) πραγματοποιήθηκαν εξολοκλήρου στα κείμενα της πρώτης δεξαμενής. Έγιναν τρεις ξεχωριστές σειρές πειραμάτων, μία για κάθε κατηγορία ονομάτων. Σε κάθε σειρά πειραμάτων, επιλεγόταν κάθε φορά τυχαία ένα κείμενο (πλήρως επισημειωμένο) από το τμήμα εκπαίδευσης της δεξαμενής και εξάγονταν από αυτό τα διανύσματα των λεκτικών μονάδων που περνούσαν τους ασφαλείς κανόνες και δεν περιλαμβάνονταν στις λίστες τετριμμένων λέξεων. Τα διανύσματα αυτά προσθέτονταν στα δεδομένα εκπαίδευσης της ΜΔΥ. Η ΜΔΥ επανεκπαιδεύονταν στα δεδομένα εκπαίδευσης και καλείτο να ταξινομήσει τις λεκτικές μονάδες του τμήματος ελέγχου της πρώτης δεξαμενής. Σε αυτό το σημείο υπολογίζονταν η Ακρίβεια, η Ανάκληση και του F-measure της ΜΔΥ στο τμήμα ελέγχου και η σειρά πειραμάτων προχωρούσε με την επιλογή ενός νέου κειμένου από το τμήμα εκπαίδευσης, που συνεισέφερε πρόσθετα δεδομένα εκπαίδευσης. Με τον τρόπο αυτό έγινε δυνατή η κατασκευή καμπυλών μάθησης, που δείχνουν πώς μεταβάλλονται οι επιδόσεις της κάθε ΜΔΥ όσο συσσωρεύονται περισσότερα δεδομένα εκπαίδευσης.

3.4 Ενεργητική Μάθηση

Για τη διεξαγωγή των πειραμάτων ενεργητικής μάθησης (*active learning*, AL) χρησιμοποιήθηκαν και οι δύο δεξαμενές κειμένων. Η επιλογή των διανυσμάτων προς επισημείωση γινόταν από τα 5000 κείμενα της δεύτερης δεξαμενής, ενώ ο έλεγχος του συστήματος γινόταν στα 202 κείμενα του τμήματος ελέγχου της πρώτης δεξαμενής, όπως ακριβώς και στην παθητική μάθηση.

Πιο αναλυτικά, μελετήθηκαν δύο διαφορετικές μέθοδοι⁴ επιλογής διανυσμάτων προς επισημείωση, οι οποίες περιγράφονται στις επόμενες ενότητες. Για κάθε μία από τις δύο μεθόδους επιλογής, πραγματοποιήθηκαν τρεις σειρές πειραμάτων, αντίστοιχες εκείνων της παθητικής μάθησης, δηλαδή με διαδοχικά περισσότερα παραδείγματα εκπαίδευσης, μία σειρά πειραμάτων για κάθε κατηγορία ονομάτων.

Προτού ξεκινήσει οποιαδήποτε σειρά πειραμάτων ενεργητικής μάθησης, κάθε ΜΔΥ εκπαιδεύεται με παθητική μάθηση στα 14 κείμενα που χρησιμοποιήθηκαν για το συντονισμό τους (βλ. ενότητα 3.2). Στη συνέχεια, σε κάθε βήμα ενεργητικής μάθησης μιας σειράς πειραμάτων, αναζητείται μια δέσμη (batch) των «καλύτερων» παραδειγμάτων προς επισημείωση από τη δεύτερη δεξαμενή, χρησιμοποιώντας το μέτρο του πληροφοριακού οφέλους. Το μέγεθος της δέσμης ποικίλλει, για λόγους που θα δούμε στις επόμενες ενότητες. Ακολούθως, η δέσμη παραδειγμάτων δίνεται προς επισημείωση στον άνθρωπο-εκπαιδευτή και τα παραγόμενα διανύσματα εκπαίδευσης προστίθενται στα υπάρχοντα δεδομένα εκπαίδευσης⁵. Η ΜΔΥ επανεκπαιδεύεται, αξιολογείται στα δεδομένα εκπαίδευσης και η σειρά πειραμάτων προχωρά με την επιλογή μιας νέας δέσμης παραδειγμάτων προς επισημείωση.

3.4.1 Κλασική μέθοδος επιλογής παραδειγμάτων

Η πρώτη μέθοδος επιλογής παραδειγμάτων προς επισημείωση, την οποία ονομάζουμε χάριν συντομίας «κλασική», είναι ουσιαστικά η ίδια που είχε χρησιμοποιηθεί από τους Λουκαρέλλι και Βασιλάκο [1, 3].

Συγκεκριμένα, η δεύτερη δεξαμενή των 5000 κειμένων χωρίζεται σε 5 διαμερίσεις των 1000 κειμένων η κάθε μία. Σε κάθε βήμα μιας σειράς πειραμάτων, όποτε χρειάζεται να επιλεγεί μια νέα δέσμη παραδειγμάτων προς επισημείωση, το σύστημα εξετάζει τα κείμενα μιας διαφορετικής (κυκλικά επιλεγόμενης) διαμέρισης, υπολογίζει το πληροφοριακό όφελος κάθε υποψηφίου διανύσματος εκπαίδευσης της διαμέρισης, ταξινομεί τα διανύσματα κατά φθίνουσα σειρά πληροφοριακού οφέλους και επιλέγει ως δέσμη τα διανύσματα με το υψηλότερο πληροφοριακό όφελος. Ως μέγεθος δέσμης σε αυτή τη μέθοδο επιλογής επιλέχθηκαν τα 400 διανύσματα,

⁴ Σημειώνεται ότι για τα ονόματα οργανισμών, πραγματοποιήθηκε μία επιπλέον σειρά πειραμάτων ενεργητικής μάθησης, όπου η επιλογή των διανυσμάτων γινόταν με μια τρίτη μέθοδο, η οποία περιγράφεται στην εργασία [9]. Ωστόσο, τα αποτελέσματα αυτής της σειράς πειραμάτων δεν ήταν ενθαρρυντικά, όπως είχε παρατηρηθεί και στην εργασία του Βασιλάκου, γεγονός που απέτρεψε την περαιτέρω χρήση της τρίτης μεθόδου επιλογής.

⁵ Στο σημείο αυτό επανυπολογίζονται οι λίστες των συχνών λέξεων, βάσει και των νέων παραδειγμάτων εκπαίδευσης. Επανυπολογίζονται, επίσης, όλες οι ιδιότητες των διανυσμάτων εκπαίδευσης (παλαιών και νέων) που εξαρτώνται από τις λίστες συχνών λέξεων. Αντιθέτως, στα προηγούμενα δύο συστήματα [1, 3] οι λίστες συχνών λέξεων υπολογίζονταν μόνο βάσει των 14 κειμένων παθητικής μάθησης.

μέγεθος στο οποίο είχε καταλήξει ο Βασιλάκος [3], ο οποίος είχε πειραματιστεί με διαφορετικά μεγέθη δέσμης. Σημειώνεται ότι όσο μικρότερο είναι το μέγεθος της δέσμης τόσο περισσότερα είναι τα βήματα μιας σειράς πειραμάτων και άρα τόσο περισσότερες είναι οι επανεκπαιδεύσεις της ΜΔΥ που απαιτούνται μέχρι να φτάσουμε σε έναν επιθυμητό συνολικό αριθμό επισημειωμένων παραδειγμάτων εκπαίδευσης, με αποτέλεσμα να απαιτείται περισσότερος χρόνος για την εκτέλεση των πειραμάτων και γενικότερα την εκπαίδευση του συστήματος. Από την άλλη πλευρά, όσο μεγαλύτερο είναι το μέγεθος της δέσμης, τόσο περισσότερο κινδυνεύουμε η δέσμη να περιέχει πολύ παρόμοια, και άρα λιγότερο χρήσιμα, διανύσματα εκπαίδευσης ή γενικότερα διανύσματα που είναι λιγότερο χρήσιμα δεδομένων των υπολοίπων διανυσμάτων της δέσμης.

Από πλευράς διεπαφής χρήστη, σε κάθε βήμα ο άνθρωπος-εκπαιδευτής βλέπει μία-μία τις λεκτικές μονάδες που περιέχονται στη δέσμη, μαζί με μια γειτονιά 16 λεκτικών μονάδων πριν και μετά τη λεκτική μονάδα που αποτελεί το παράδειγμα εκπαίδευσης, και απαντά «Ναι» ή «Όχι», ανάλογα με το αν η λεκτική μονάδα ανήκει ή όχι στην κατηγορία ονομάτων της ΜΔΥ.

3.4.2 Επιγραμμική μέθοδος επιλογής παραδειγμάτων

Η δεύτερη μέθοδος επιλογής παραδειγμάτων, την οποία καλούμε «επιγραμμική» (online) [10], συμπληρώνει σε κάθε βήμα τη δέσμη των παραδειγμάτων προς επισημείωση επιλέγοντας τυχαία ένα ή περισσότερα κείμενα από τη δεύτερη δεξαμενή (των μη επισημειωμένων κειμένων) και κρατώντας από κάθε κείμενο τα καλύτερα διανύσματα, βάσει του πληροφοριακού τους οφέλους.

Πιο αναλυτικά, στην αρχή των πειραμάτων ο άνθρωπος-εκπαιδευτής ορίζει τον επιθυμητό τελικό συνολικό αριθμό διανυσμάτων εκπαίδευσης (*διανύσματα-στόχος*) και το σύστημα υπολογίζει το συνολικό αριθμό των διαθέσιμων υποψηφίων διανυσμάτων εκπαίδευσης της δεύτερης δεξαμενής (*διανύσματα στη δεξαμενή*). Κάθε φορά που χρειάζεται να συμπληρωθεί μια δέσμη, το σύστημα επιλέγει τυχαία ένα διαφορετικό κείμενο της δεύτερης δεξαμενής και από αυτό επιλέγει τα k καλύτερα διανύσματα, όπου το k ορίζεται στη σχέση (3.1). Ανάλογα, δηλαδή, με τον αριθμό των διανυσμάτων του κειμένου (*διανύσματα στο κείμενο*) επιλέγονται περισσότερα ή λιγότερα διανύσματα από αυτό ως παραδείγματα εκπαίδευσης. Αν η δέσμη δεν έχει συμπληρωθεί, επιλέγεται τυχαία και νέο κείμενο της δεύτερης δεξαμενής, από αυτό επιλέγονται πάλι k διανύσματα (το k επανυπολογίζεται) κ.ο.κ., μέχρι να συμπληρωθεί η δέσμη.

$$k = \left\lceil \frac{\text{διανύσματα-στόχος} \cdot \text{διανύσματα στο κείμενο}}{\text{διανύσματα στη δεξαμενή}} \right\rceil \quad (3.1)$$

Το κυριότερο ίσως πλεονέκτημα της επιγραμμικής μεθόδου είναι ότι επιτυγχάνει δραματική μείωση του χρόνου επιλογής των παραδειγμάτων εκπαίδευσης, αφού σε κάθε βήμα ο υπολογισμός του πληροφοριακού οφέλους γίνεται μόνο για τα διανύσματα των (λίγων) τυχαία επιλεγμένων κειμένων από τα οποία εξάγονται τα παραδείγματα της δέσμης. Αντιθέτως, στην κλασική μέθοδο, ο επανυπολογισμός γίνεται για όλα τα διανύσματα της διαμέρισης (1/5 της δεύτερης δεξαμενής). Ένα άλλο πλεονέκτημα της επιγραμμικής μεθόδου είναι ότι αποφεύγεται ο κίνδυνος η δέσμη να αποτελείται από πολύ παρόμοια διανύσματα, τα οποία προέρχονται από πολλαπλές εμφανίσεις των ιδίων λεκτικών μονάδων με τα ίδια συμφραζόμενα στη

δεύτερη (μεγάλη) δεξαμενή κειμένων. Κι αυτό γιατί στην περίπτωση της επιγραμμικής μεθόδου, τα διανύσματα επιλέγονται από λίγα, τυχαία επιλεγμένα κείμενα της δεύτερης δεξαμενής, όχι μια ολόκληρη διαμέριση (1/5) της δεύτερης δεξαμενής, και η πιθανότητα να εμφανιστούν (και να επιλεγούν) πολλές φορές οι ίδιες λεκτικές μονάδες με τα ίδια συμφραζόμενα στα λίγα αυτά κείμενα είναι πολύ μικρότερη από ό,τι στην περίπτωση μιας ολόκληρης διαμέρισης της δεύτερης δεξαμενής.

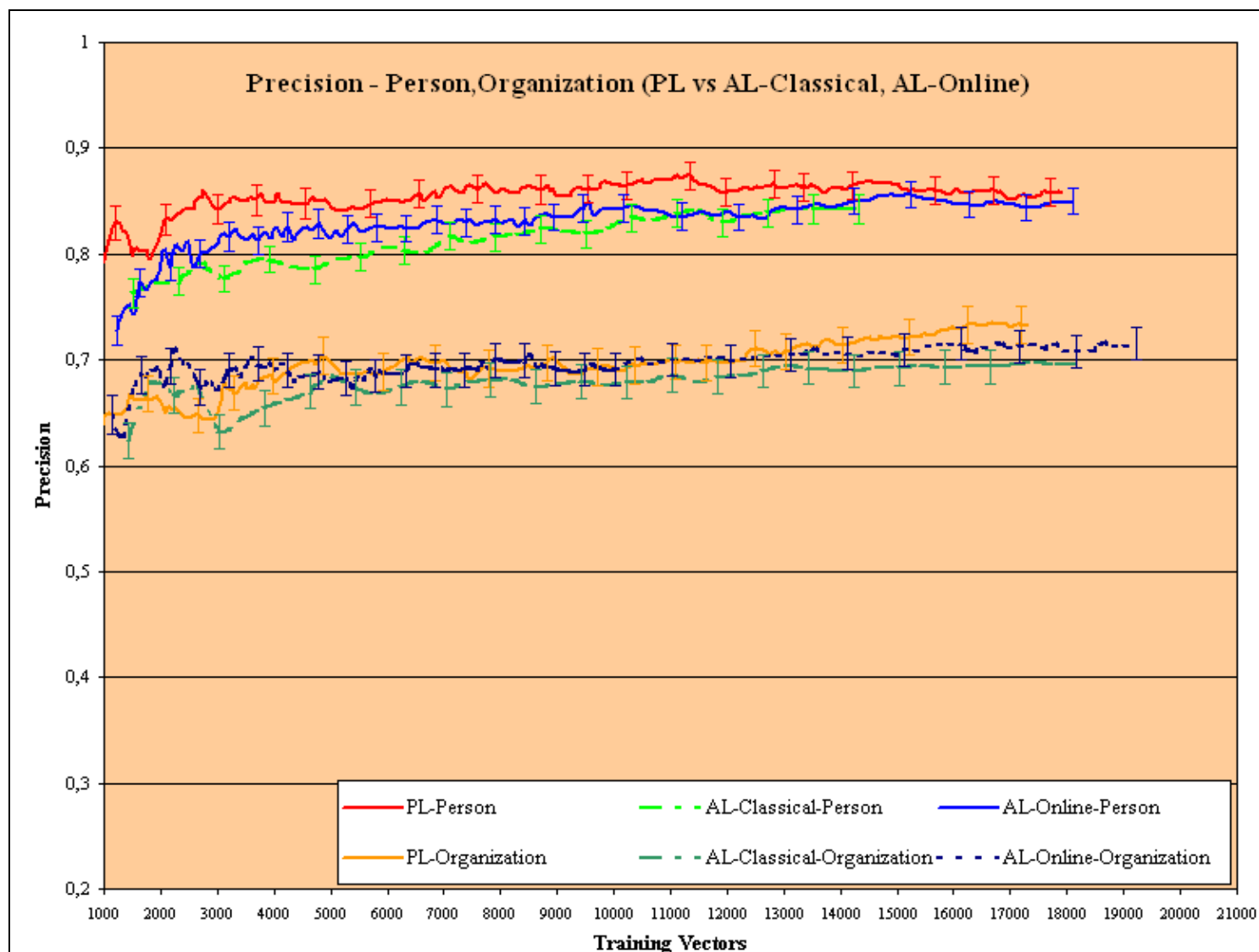
Η σημαντική μείωση του χρόνου επιλογής παραδειγμάτων κατά την κατασκευή κάθε δέσμης επιτρέπει τη χρήση μικρότερου μεγέθους δέσμης (συνολικά περισσότερες δέσμες μέχρι να φτάσουμε στον επιθυμητό συνολικό αριθμό παραδειγμάτων εκπαίδευσης). Στα πρώτα περίπου 10000 διανύσματα εκπαίδευσης κάθε σειράς πειραμάτων, η δέσμη είχε μέγεθος 100 διανυσμάτων, ενώ στα επόμενα είχε μέγεθος 200 διανυσμάτων.

Στη διεπαφή χρήστη, ο άνθρωπος-εκπαιδευτής βλέπει πλέον στην οθόνη ολόκληρα τα κείμενα από τα οποία επιλέγονται τα παραδείγματα της δέσμης. Σε κάθε κείμενο, το σύστημα σημειώνει τις k λεκτικές μονάδες που πρέπει να επισημειωθούν.

3.5 Πειραματικά αποτελέσματα

Στις επόμενες υποενότητες παρουσιάζονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν στην παρούσα εργασία, με τη μορφή διαγραμμάτων ακρίβειας, ανάκλησης και f -measure. Στα διαγράμματα της ακρίβειας και της ανάκλησης υπάρχουν ράβδοι λάθους (*error bars*) που απεικονίζουν τα διαστήματα εμπιστοσύνης 95% (βλ. παράρτημα εργασίας Βασιλάκου [3]).

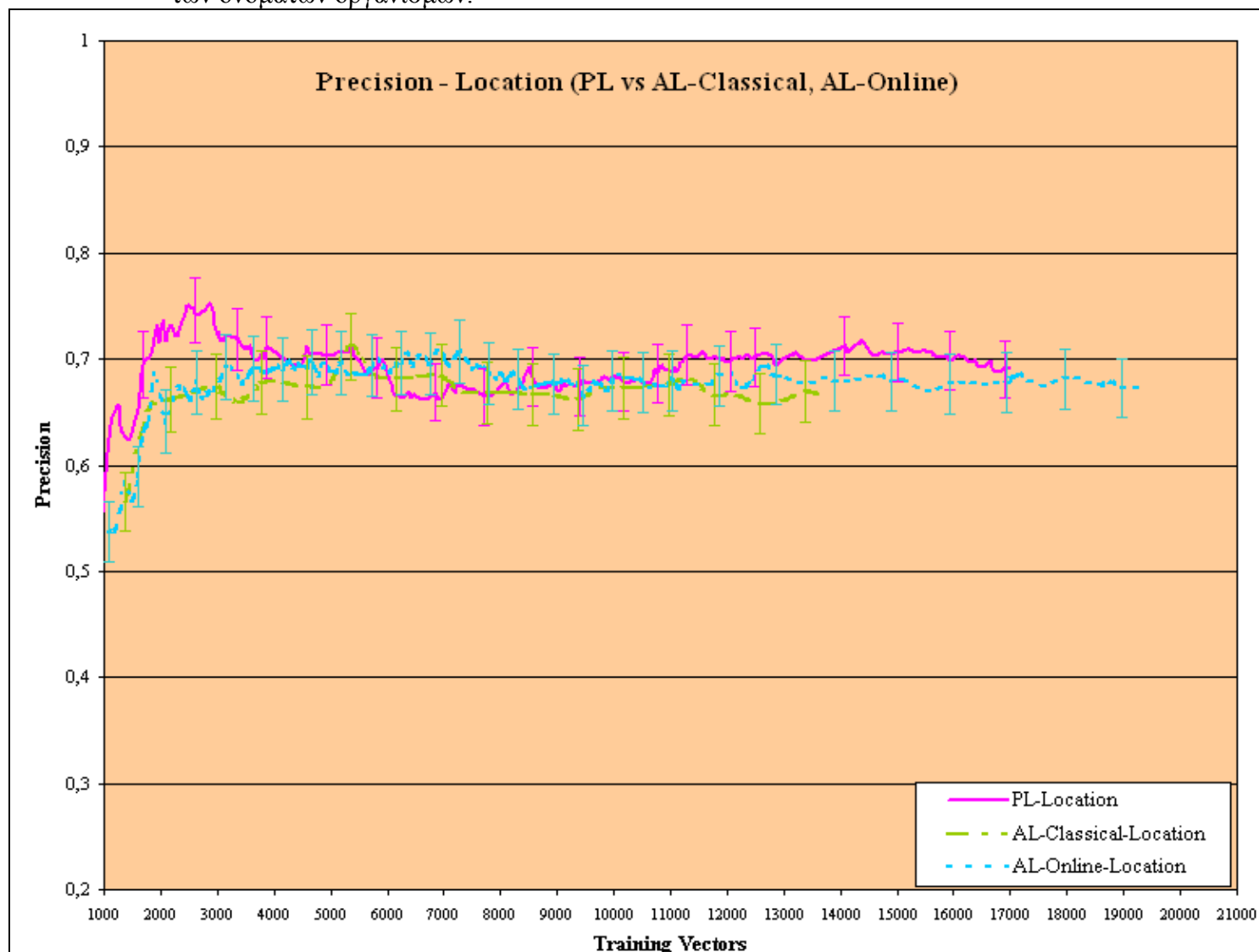
3.5.1 Αποτελέσματα Ακρίβειας



3.5.1 – Διάγραμμα ακρίβειας για τις κατηγορίες προσώπων και οργανισμών

Το διάγραμμα 3.5.1 παρουσιάζει τα αποτελέσματα ακρίβειας (precision) των πειραμάτων παθητικής και ενεργητικής μάθησης για τις κατηγορίες των ονομάτων προσώπων και οργανισμών. Εξετάζοντας τα διαστήματα εμπιστοσύνης, παρατηρούμε ότι στην κατηγορία των οργανισμών υπάρχει μεγάλη επικάλυψη μεταξύ παθητικής και ενεργητικής μάθησης, ενώ αντίθετα η επικάλυψη είναι πολύ μικρότερη στα ονόματα προσώπων. Και στις δύο κατηγορίες, ιδιαίτερα στην κατηγορία των προσώπων, η παθητική μάθηση φαίνεται περιέργως να οδηγεί σε καλύτερα αποτελέσματα ακρίβειας, κάτι που έρχεται σε αντίθεση με τα αποτελέσματα των προηγούμενων εργασιών [1, 3], αν και τα αποτελέσματα αυτής της εργασίας δεν είναι άμεσα συγκρίσιμα με των προηγούμενων, επειδή οι ΜΔΥ εκπαιδεύονται πλέον ανεξάρτητα. Ως προς τις δύο μεθόδους ενεργητικής μάθησης, σε γενικές γραμμές η καμπύλη της κλασικής μεθόδου είναι πιο κάτω από την αντίστοιχη της επιγραμμικής,

αν και οι διαφορές δεν είναι πάντα στατιστικά σημαντικές, ιδιαίτερα στην περίπτωση των ονομάτων οργανισμών.

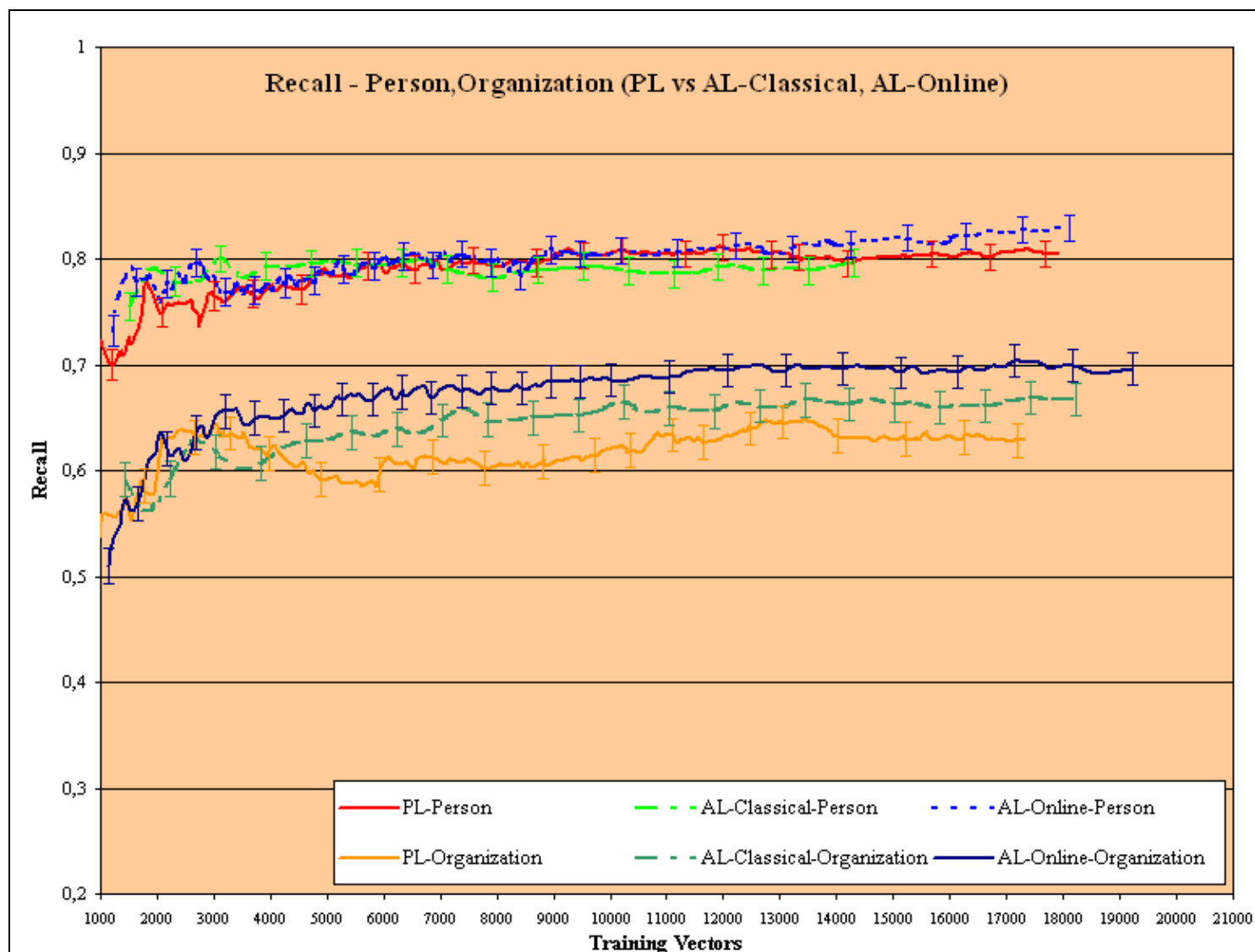


3.5.2 – Διάγραμμα ακρίβειας για την κατηγορία των τοποθεσιών

Το διάγραμμα 3.5.2 παρουσιάζει τα αποτελέσματα ακρίβειας για την κατηγορία των τοποθεσιών. Η επικάλυψη μεταξύ όλων των μεθόδων είναι πολύ μεγαλύτερη σε αυτή την κατηγορία ονομάτων, σε σχέση με τις προηγούμενες δύο. Επίσης οι ράβδοι λάθους έχουν μεγαλύτερο εύρος, γεγονός που οφείλεται στο μικρότερο αριθμό διαθέσιμων παραδειγμάτων ελέγχου (2094 διανύσματα ελέγχου, έναντι 3846 για τα πρόσωπα και 3560 για τους οργανισμούς).

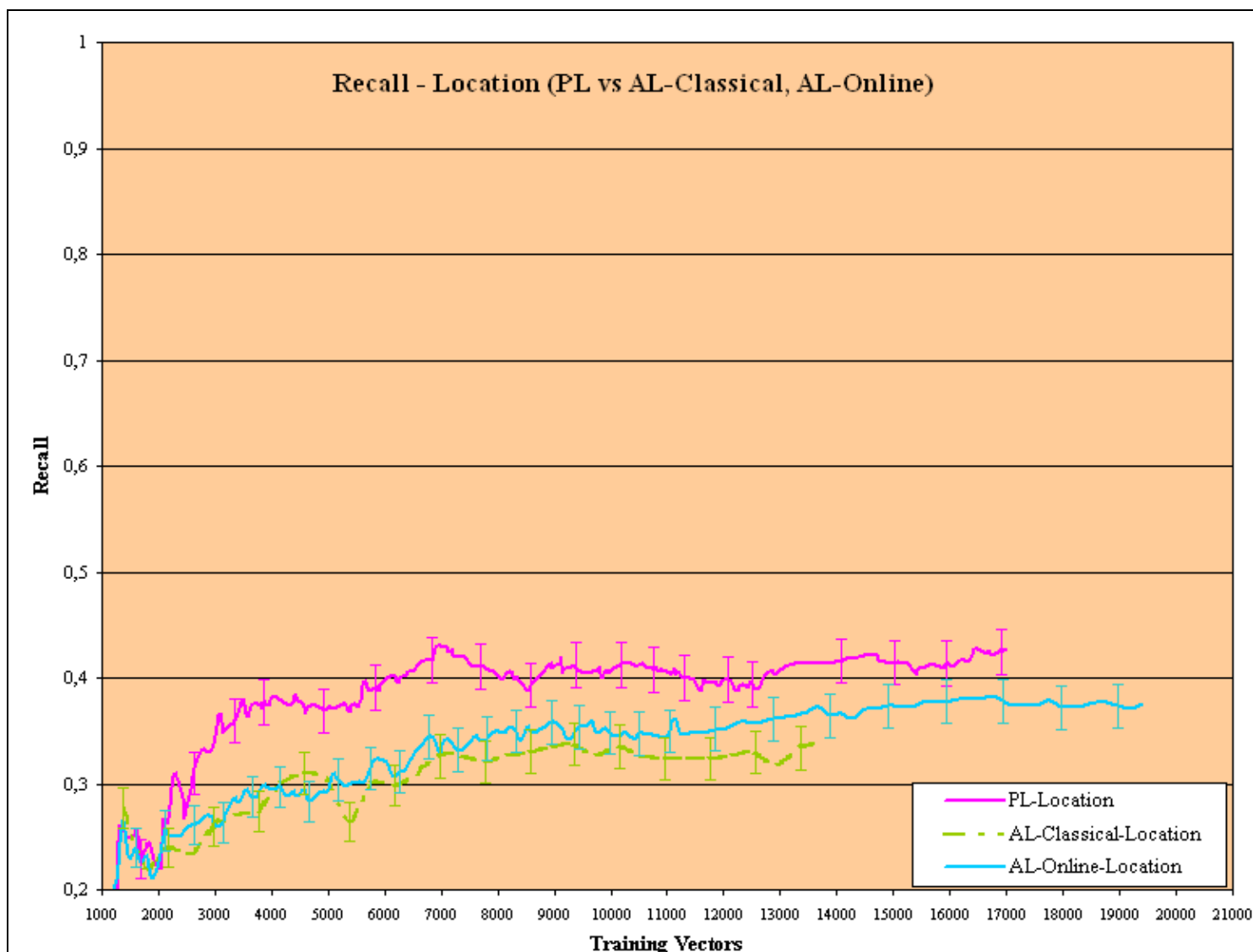
Η γενική εικόνα που προκύπτει ως προς την ακρίβεια είναι ότι η υπολογιστικά φθηνότερη επιγραμμική μέθοδος ενεργητικής μάθησης είναι εξίσου καλή ή και καλύτερη της κλασικής. Καμία από τις δύο μεθόδους ενεργητικής μάθησης, όμως, δεν φαίνεται να οδηγεί σε αποτελέσματα καλύτερα της παθητικής μάθησης. Στην περίπτωση των ονομάτων προσώπων, μάλιστα, η παθητική μάθηση φαίνεται να οδηγεί σε σημαντικά καλύτερα αποτελέσματα ακρίβειας.

3.5.2 Αποτελέσματα Ανάκλησης



3.5.3 – Διάγραμμα ανάκλησης για τις κατηγορίες προσώπων και οργανισμών

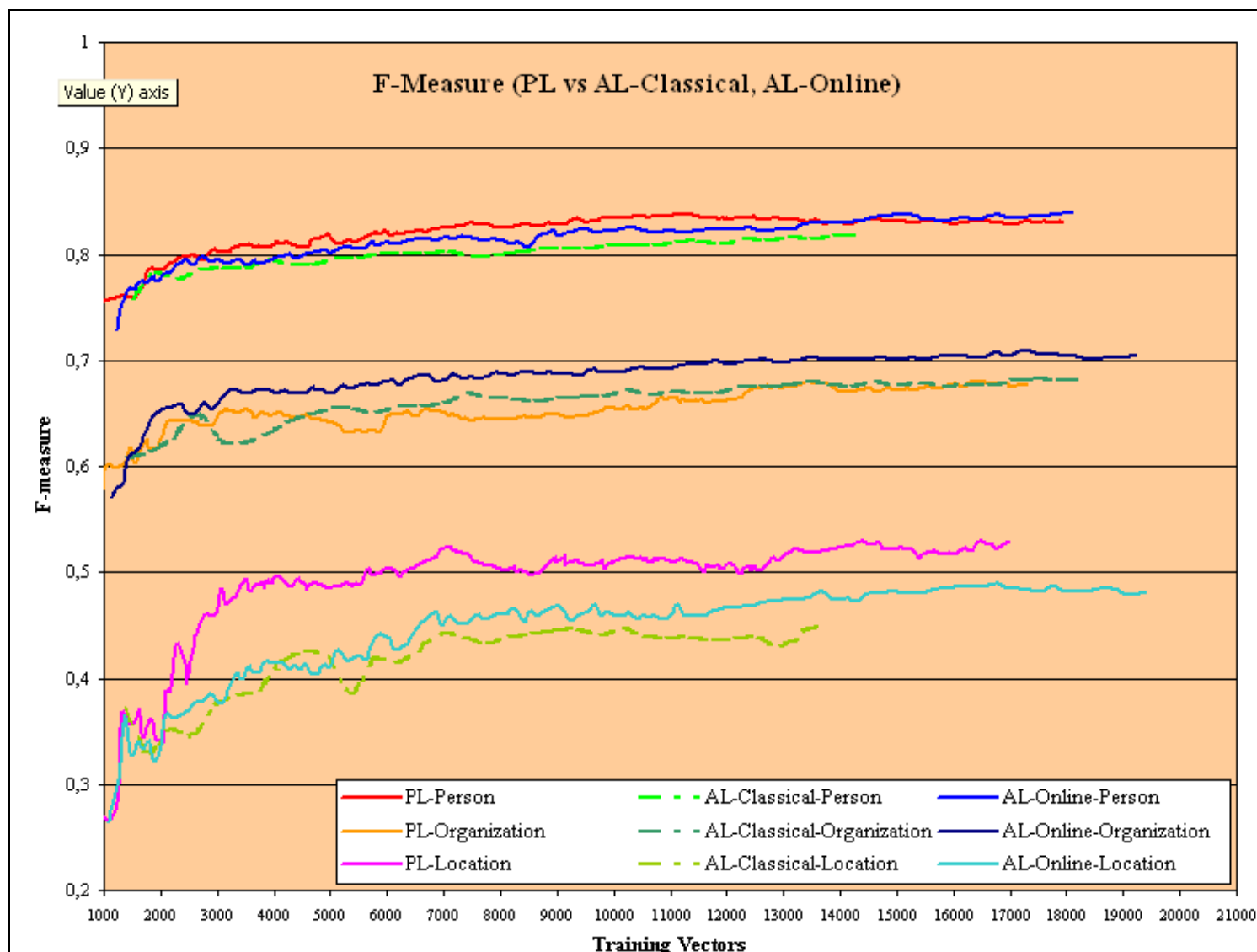
Το διάγραμμα 3.5.3 παρουσιάζει τα αποτελέσματα ανάκλησης (recall) για τα ονόματα προσώπων και οργανισμών. Και στις δύο κατηγορίες, τα καλύτερα αποτελέσματα επιτυγχάνονται με την επιγραμμική ενεργητική μάθηση, η οποία φαίνεται να υπερτερεί της κλασικής, αν και οι διαφορές είναι πολύ μικρές στην περίπτωση των ονομάτων προσώπων. Επιπροσθέτως, στην περίπτωση των ονομάτων οργανισμών οι δύο μέθοδοι ενεργητικής μάθησης οδηγούν σε σαφώς καλύτερα αποτελέσματα, σε σχέση με την παθητική μάθηση, κάτι που ενδεχομένως να αποτελεί ένδειξη πως η ενεργητική μάθηση βοηθά το σύστημα να μάθει καλύτερα τη μεγάλη ποικιλία μορφών που παρουσιάζουν τα ονόματα οργανισμών. Δεν παρουσιάζεται, όμως, το ίδιο φαινόμενο στην κατηγορία των ονομάτων τοποθεσιών, τα αποτελέσματα ανάκλησης της οποίας φαίνονται στο διάγραμμα 3.5.4. Η επιγραμμική μέθοδος ενεργητικής μάθησης φαίνεται και πάλι να υπερτερεί της κλασικής, αλλά τα καλύτερα αποτελέσματα επιτυγχάνονται με την παθητική μάθηση.



3.5.4 – Διάγραμμα ανάκλησης για την κατηγορία των τοποθεσιών

Η γενική εικόνα που προκύπτει ως προς την ανάκληση είναι ότι η υπολογιστικά φθηνότερη επιγραμμική μέθοδος ενεργητικής μάθησης είναι εξίσου καλή ή και καλύτερη της κλασικής, όπως και στην περίπτωση της ακρίβειας. Οι δύο μέθοδοι ενεργητικής μάθησης φαίνεται, επίσης, να οδηγούν σε υψηλότερη ανάκληση από ό,τι η παθητική μάθηση στα ονόματα οργανισμών, αλλά οι διαφορές είναι πολύ μικρές στα ονόματα προσώπων, ενώ στα ονόματα τοποθεσιών η παθητική μάθηση υπερτερεί.

3.5.3 Αποτελέσματα F-measure



3.5.5 – Διάγραμμα F-measure και για τις τρεις κατηγορίες ονομάτων

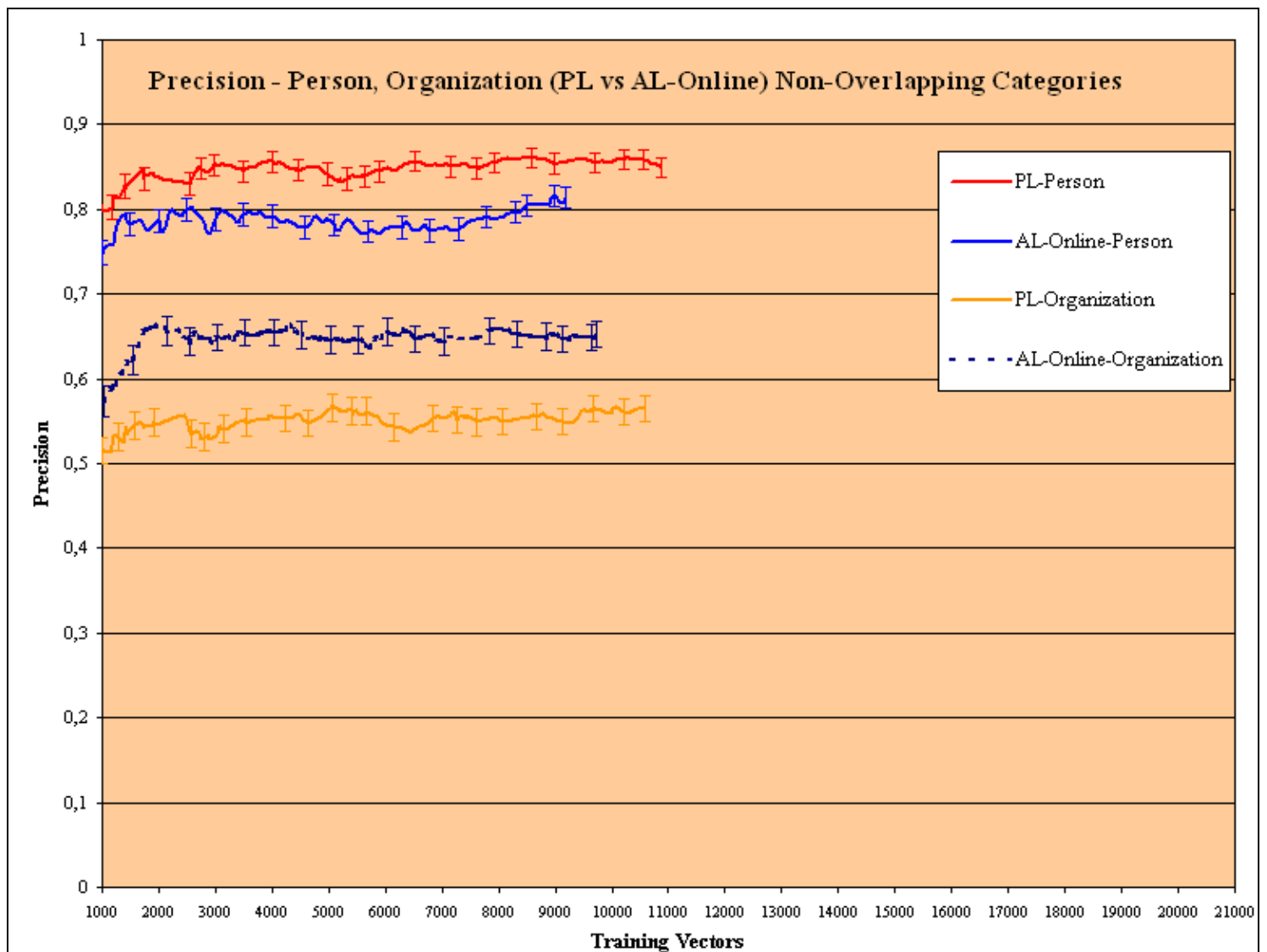
Το διάγραμμα 3.5.5 παρουσιάζει τα αποτελέσματα F-measure των τριών κατηγοριών ονομάτων. Βλέπουμε και πάλι ότι η επιγραμμική μέθοδος ενεργητικής μάθησης υπερτερεί της κλασικής. Η επιγραμμική μέθοδος επιτυγχάνει, επίσης, τα καλύτερα τελικά αποτελέσματα μεταξύ των τριών μεθόδων στα ονόματα προσώπων και οργανισμών, αν και η διαφορά στην περίπτωση των ονομάτων προσώπων είναι πάρα πολύ μικρή. Αντιθέτως, στα ονόματα τοποθεσιών υπερτερεί η παθητική μάθηση.

Σημειώνεται ότι στην περίπτωση της κλασικής μεθόδου ενεργητικής μάθησης, στην κατηγορία των προσώπων και των τοποθεσιών εκτελέστηκαν λιγότερα πειράματα από ό,τι στις άλλες δύο μεθόδους, καθότι παρατηρήθηκε μεγάλη διαφορά από τα αποτελέσματα της παθητικής μάθησης και η επιλογή διανυσμάτων προς επισημείωση ήταν χρονικά αρκετά δαπανηρή.

3.5.4 Μη επικαλυπτόμενες κατηγορίες ονομάτων

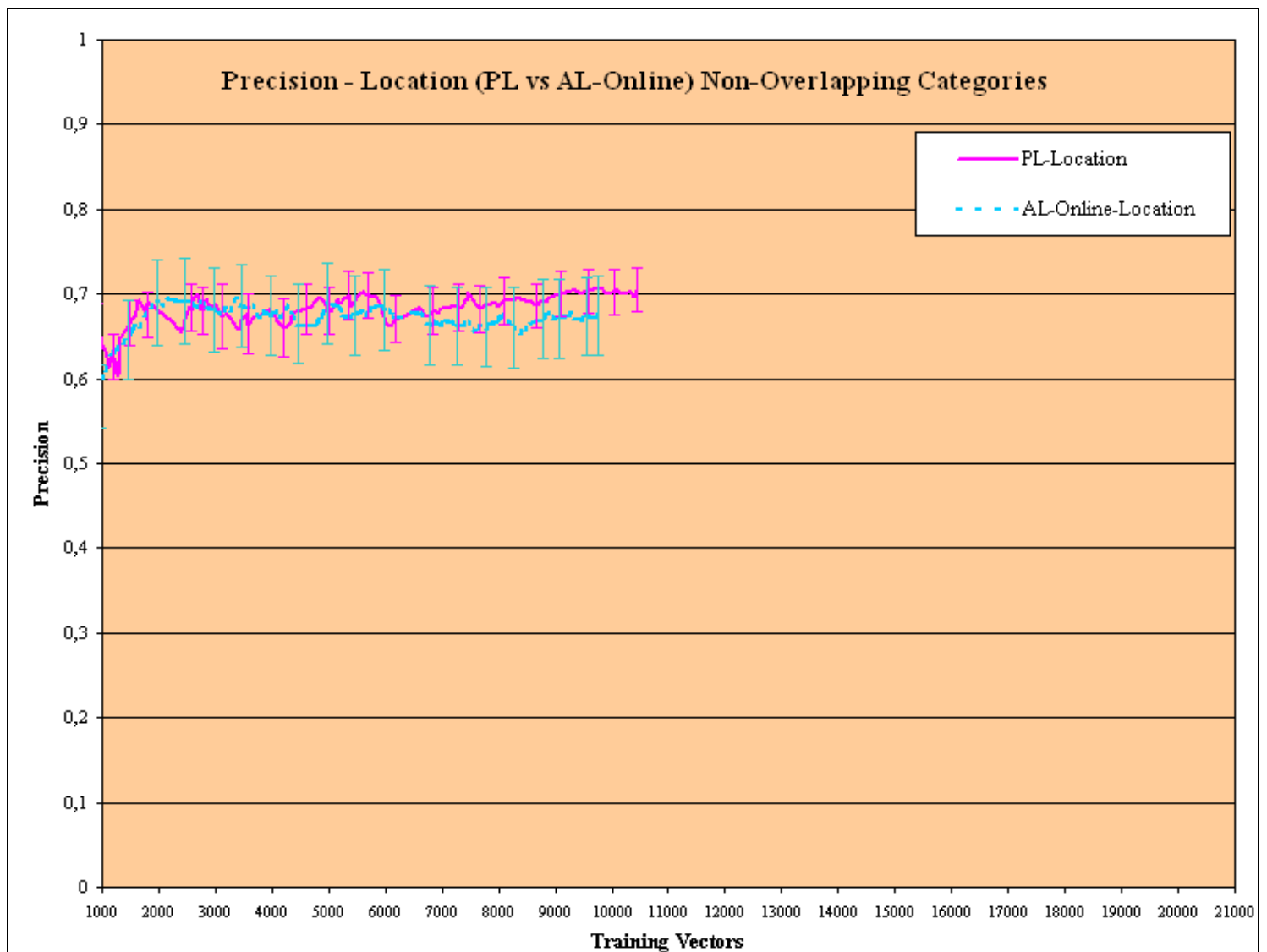
Στα ως τώρα πειράματα εξετάσαμε ξεχωριστά τις επιδόσεις των ΜΔΥ των τριών κατηγοριών ονομάτων. Στα πειράματα αυτής της υποενότητας, εξετάζουμε την περίπτωση όπου κάθε λεκτική μονάδα επιτρέπεται να ταξινομηθεί σε μία μόνο κατηγορία ονομάτων. Στην περίπτωση αυτή, οι τρεις ΜΔΥ εκπαιδεύονται και πάλι ανεξάρτητα, όπως στα προηγούμενα πειράματα, αλλά κατά την κατάταξη νέων λεκτικών μονάδων (δεδομένα ελέγχου) το σύστημα κατατάσσει κάθε λεκτική μονάδα στην κατηγορία ονομάτων της οποίας η αντίστοιχη ΜΔΥ είναι περισσότερο βέβαιη πως η λεκτική μονάδα ανήκει στη θετική της κατηγορία. Αν καμία ΜΔΥ δεν αποκριθεί ότι η λεκτική μονάδα ανήκει στη θετική της κατηγορία, τότε η λεκτική μονάδα δεν κατατάσσεται σε καμία κατηγορία ονομάτων.

Πραγματοποιήθηκαν δυο σειρές πειραμάτων, μία με παθητική μάθηση και μία με την επιγραμμική μέθοδο ενεργητικής μάθησης. Στην περίπτωση της παθητικής μάθησης, κάθε ΜΔΥ εκπαιδεύτηκε με τον ίδιο τρόπο όπως και στα προηγούμενα πειράματα παθητικής μάθησης για επικαλυπτόμενες κατηγορίες ονομάτων. Στα πειράματα της ενεργητικής μάθησης, εκπαιδεύσαμε σε κάθε βήμα (100 διανύσματα ανά δέσμη) την κάθε ΜΔΥ στα αντίστοιχα δεδομένα εκπαίδευσης που είχαν παραχθεί για αυτή τη ΜΔΥ στα προηγούμενα πειράματα με την επιγραμμική μέθοδο ενεργητικής μάθησης. Η αξιολόγηση της κάθε ΜΔΥ σε κάθε βήμα γινόταν όπως και πριν στα δεδομένα ελέγχου της πρώτης δεξαμενής. Ακολουθούν τα διάγραμμα ακρίβειας, ανάκλησης και f-measure για αυτές τις σειρές πειραμάτων.



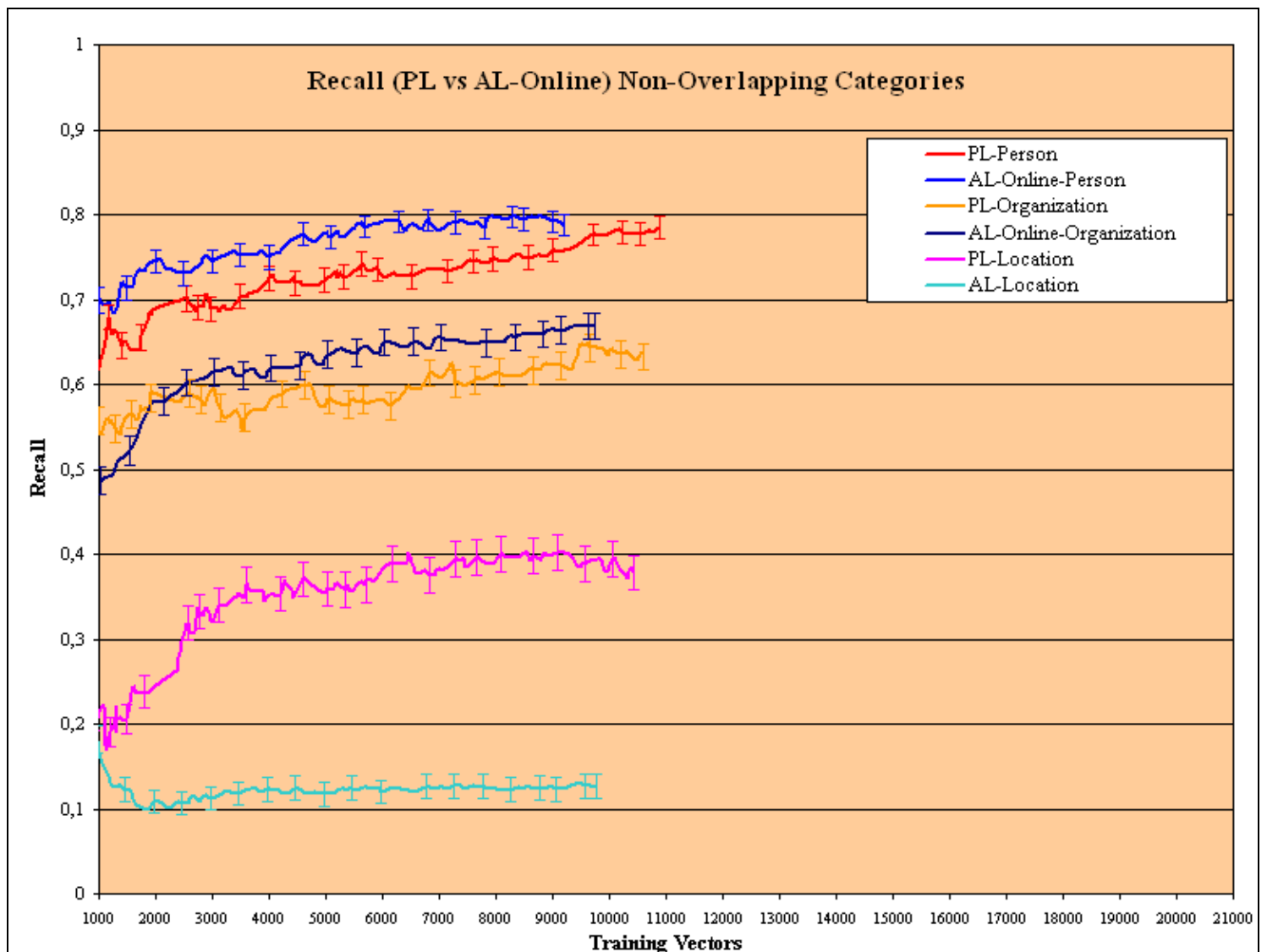
3.5.6 – Διάγραμμα ακρίβειας για τις κατηγορίες προσώπων και οργανισμών για μη επικαλυπτόμενες κατηγορίες

Στο διάγραμμα 3.5.6 βλέπουμε τα αποτελέσματα ακρίβειας των πειραμάτων παθητικής και ενεργητικής μάθησης για τις κατηγορίες των ονομάτων προσώπων και οργανισμών. Συγκρίνοντας με το αντίστοιχο διάγραμμα 3.5.1 των πειραμάτων επικαλυπτόμενων κατηγοριών, παρατηρούμε παρόμοια αποτελέσματα στην κατηγορία των προσώπων, με την παθητική μάθηση να υπερτερεί και πάλι. Στην κατηγορία των οργανισμών, ωστόσο, η ενεργητική μάθηση έχει τώρα σημαντικά υψηλότερες τιμές ακρίβειας σε σχέση με την παθητική μάθηση, ενώ στα πειράματα επικαλυπτόμενων κατηγοριών η διαφορά ήταν πολύ μικρή.



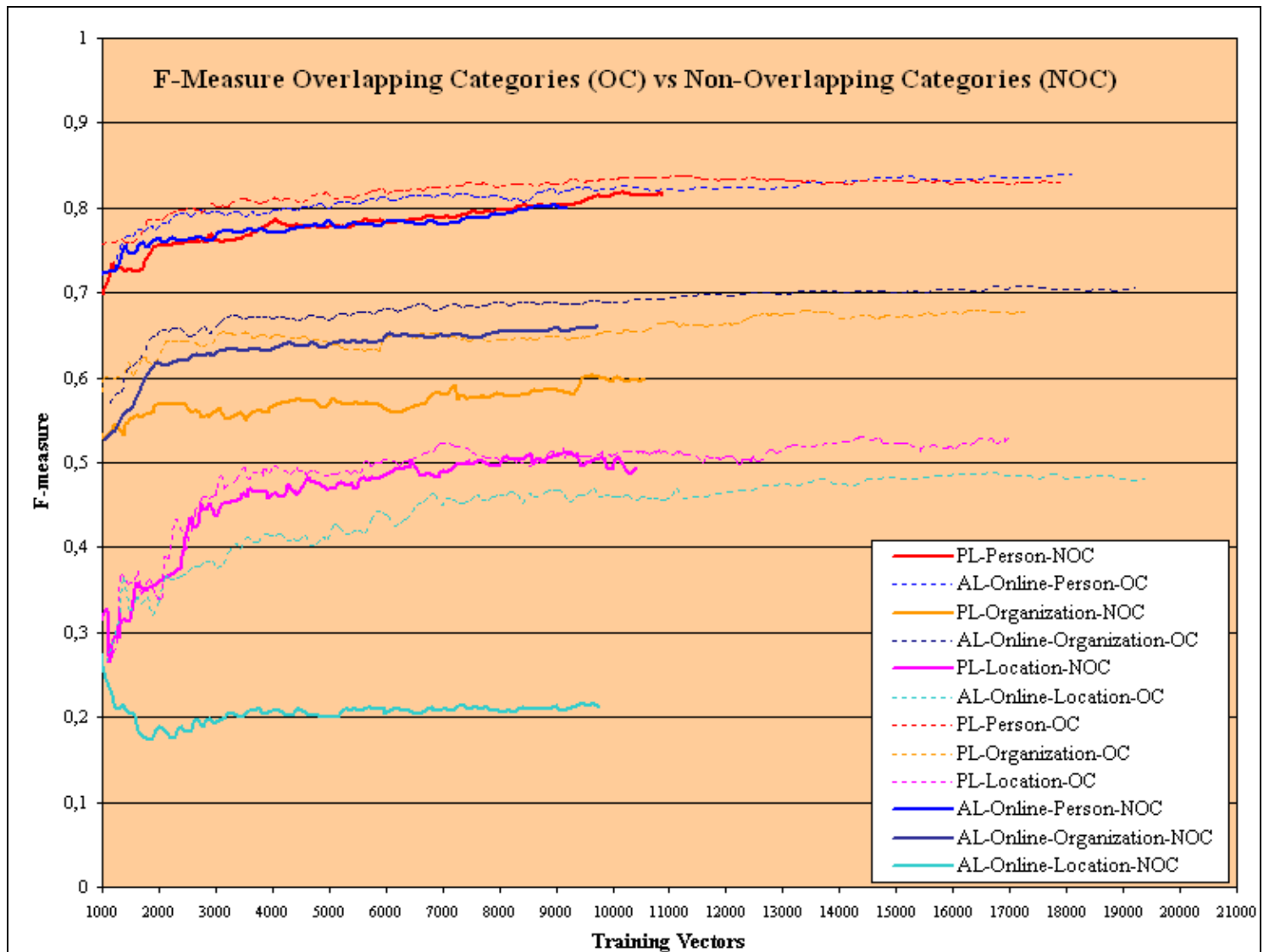
3.5.7 – Διάγραμμα ακρίβειας για την κατηγορία των τοποθεσιών για μη επικαλυπτόμενες κατηγορίες

Στο διάγραμμα 3.5.7 έχουμε τα αποτελέσματα της ακρίβειας για την κατηγορία των τοποθεσιών. Η επικάλυψη μεταξύ παθητικής και ενεργητικής μάθησης είναι ιδιαίτερα μεγάλη, ενώ οι ράβδοι λάθους έχουν μεγάλο εύρος λόγω μικρού αριθμού διανυσμάτων ελέγχου, όπως ακριβώς συνέβη και στα αντίστοιχα πειράματα επικαλυπτόμενων κατηγοριών (διάγραμμα 3.5.2).



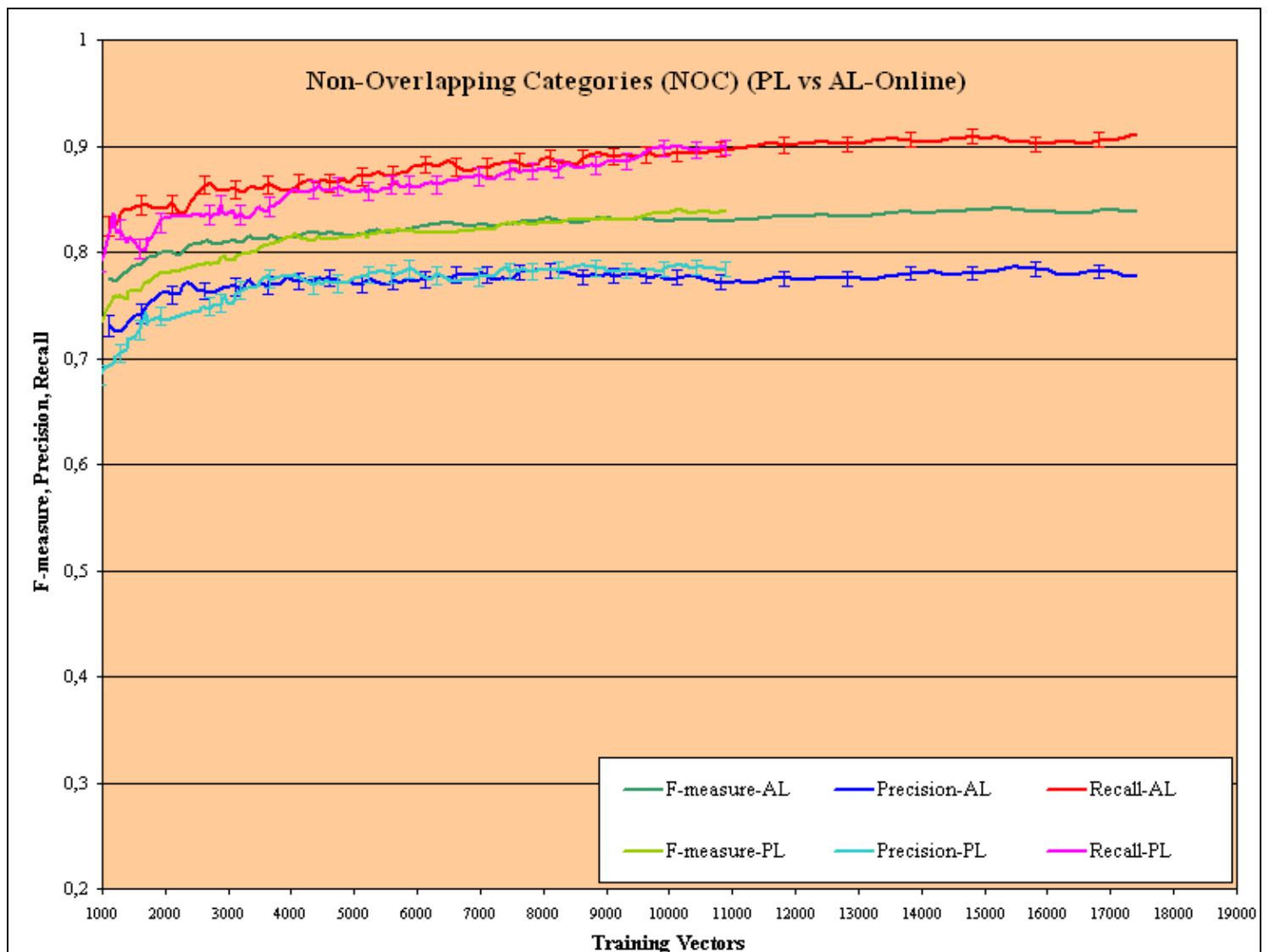
3.5.8 – Διάγραμμα ανάκλησης για τις κατηγορίες προσώπων, οργανισμών και τοποθεσιών για μη επικαλυπτόμενες κατηγορίες

Το διάγραμμα 3.5.8 παρουσιάζει τα αποτελέσματα της ανάκλησης και για τις τρεις κατηγορίες ονομάτων. Συγκρίνοντας με το αντίστοιχο διάγραμμα 3.5.3, η διαφορά μεταξύ ενεργητικής και παθητικής μάθησης είναι τώρα πολύ μεγαλύτερη, με την ενεργητική μάθηση να υπερτερεί σαφώς. Για τους οργανισμούς, πάλι η ενεργητική υπερτερεί της παθητικής μάθησης, αλλά η διαφορά είναι ελαφρά μικρότερη από εκείνη του διαγράμματος 3.5.3. Τέλος, στην περίπτωση των τοποθεσιών, η παθητική μάθηση υπερτερεί, όπως στο αντίστοιχο διάγραμμα 3.5.4, αλλά η διαφορά από την ενεργητική μάθηση είναι τώρα πολύ μεγαλύτερη.



3.5.9 – Διάγραμμα *F-measure* για επικαλυπτόμενες και μη επικαλυπτόμενες κατηγορίες

Στο διάγραμμα 3.5.9 συγκρίνουμε το *F-measure* των πειραμάτων των επικαλυπτόμενων κατηγοριών με τα αντίστοιχα αποτελέσματα των πειραμάτων των μη επικαλυπτόμενων κατηγοριών. Η γενική εικόνα είναι ότι η κατάταξη των λεκτικών μονάδων σε μία μόνο κατηγορία δε δείχνει να βελτιώνει τις επιδόσεις του συστήματος. Αντιθέτως τις χειροτερεύει.

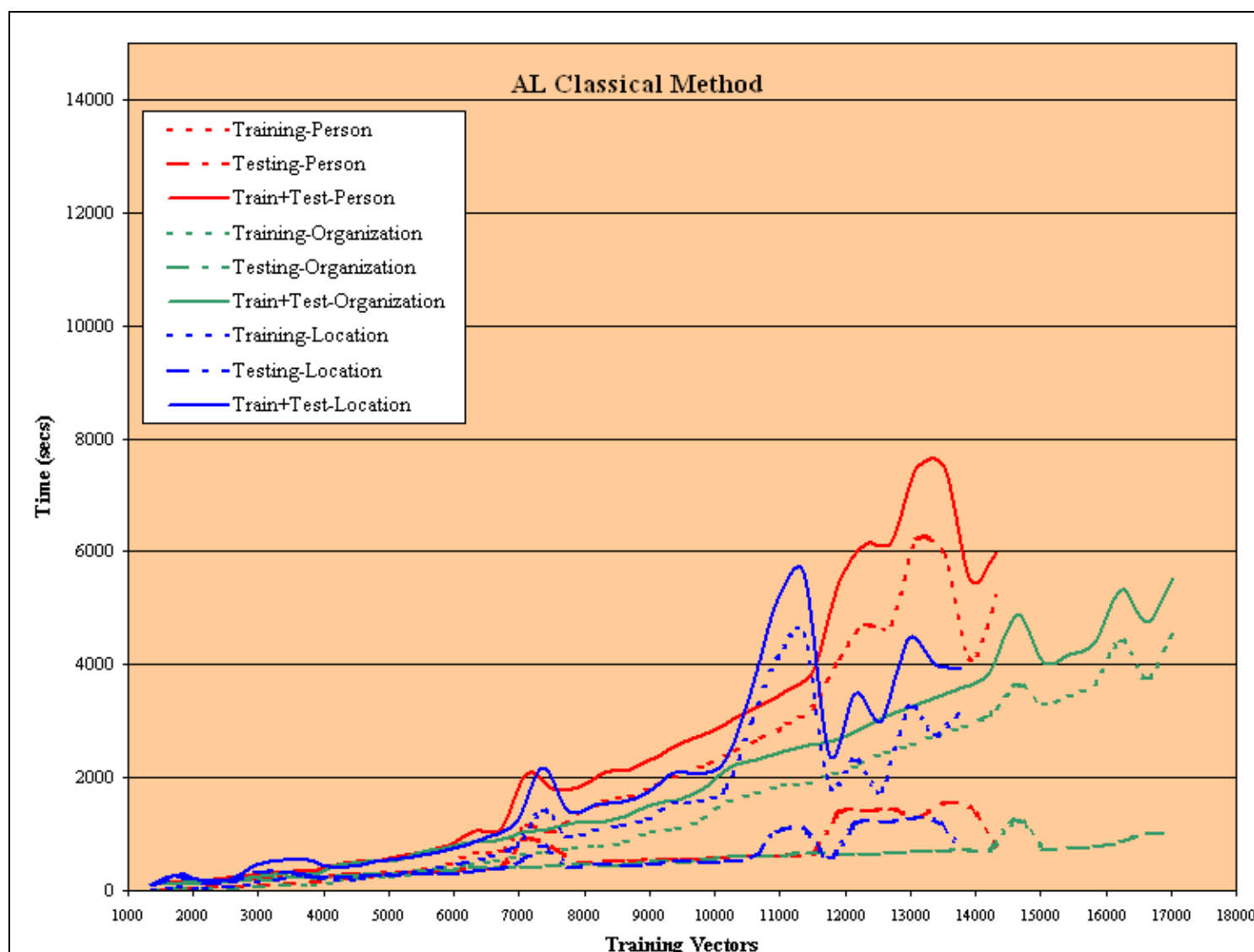


3.5.10 – Διάγραμμα ακρίβειας, ανάκλησης και *F-measure* για μη επικαλυπτόμενες κατηγορίες ονομάτων παθητικής μάθησης και ενεργητικής μάθησης με επιγραμμική μέθοδο

Το διάγραμμα 3.5.10 παρουσιάζει τα συνολικά αποτελέσματα ακρίβειας, ανάκλησης και *F-measure* για μη επικαλυπτόμενες κατηγορίες ονομάτων για την παθητική και την ενεργητική μάθηση. Η γενική εικόνα είναι ότι υπάρχει μεγάλη επικάλυψη μεταξύ ενεργητικής και παθητικής μάθησης, η οποία ίσως οφείλεται εν μέρει στη χαμηλή επίδοση της ενεργητικής μάθησης στην κατηγορία των τοποθεσιών. Η μελλοντική βελτίωση της ενεργητικής μάθησης σε αυτή την κατηγορία ενδεχομένως θα άλλαζε τη συνολική εικόνα υπέρ της ενεργητικής μάθησης.

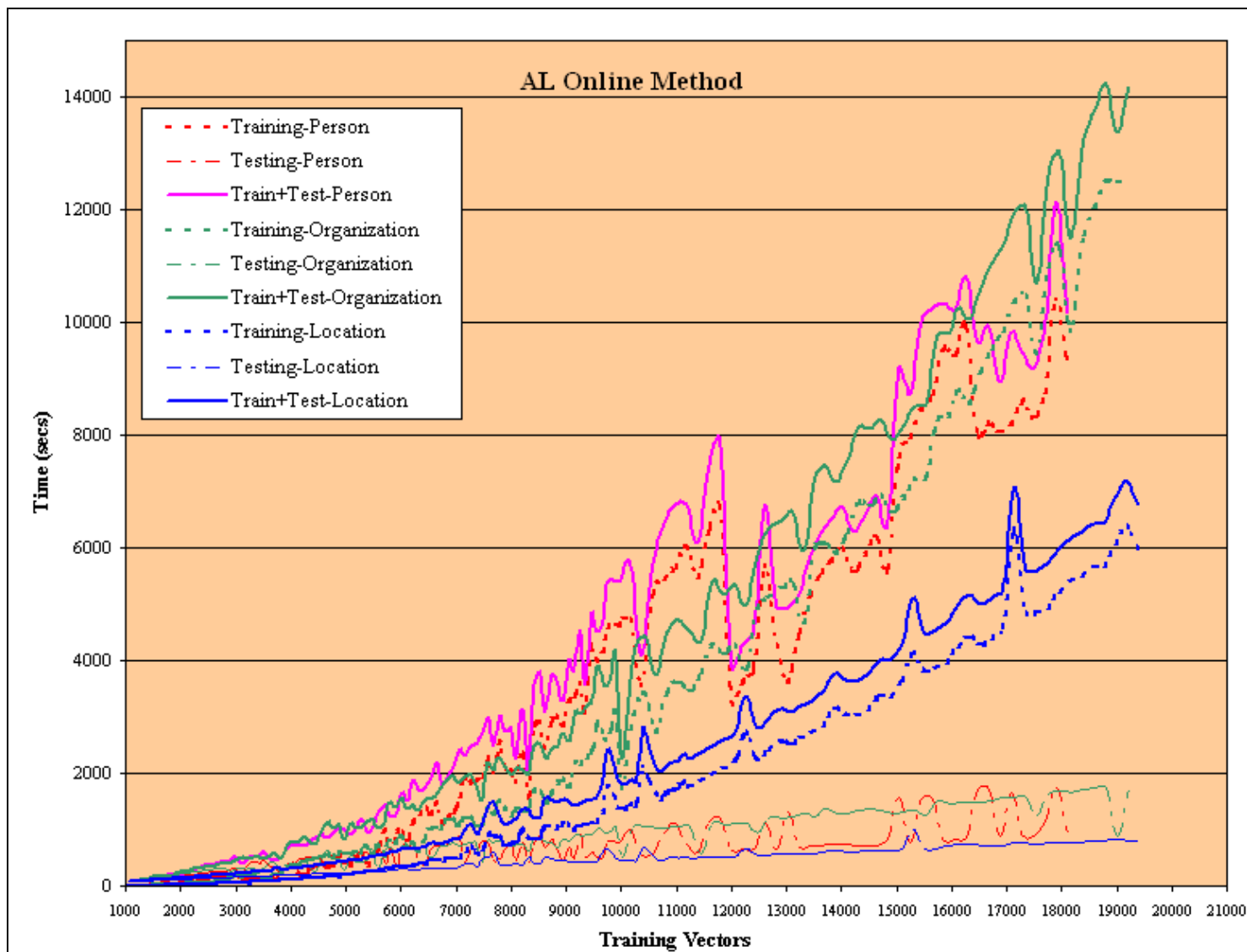
3.6 Χρονικά Διαγράμματα Ενεργητικής Μάθησης

Στην υποενότητα αυτή επικεντρωνόμαστε στους χρόνους εκπαίδευσης και κατάταξης των τριών ΜΔΥ όταν χρησιμοποιούνται οι δύο μέθοδοι ενεργητικής μάθησης. Από τα παρακάτω διαγράμματα (3.6.1 και 3.6.2) φαίνεται ότι και στις δύο μεθόδους ενεργητικής μάθησης η ΜΔΥ των προσώπων απαιτεί τον περισσότερο χρόνο εκπαίδευσης από τις τρεις ΜΔΥ. Ακολουθούν η ΜΔΥ των οργανισμών και μετά εκείνη των τοποθεσιών. Αυτό ενδέχεται να οφείλεται στο ότι η ΜΔΥ των προσώπων χρησιμοποιεί τις περισσότερες ιδιότητες (150), ενώ οι ΜΔΥ των οργανισμών και των τοποθεσιών χρησιμοποιούν λιγότερες (142 και 133 αντίστοιχα).



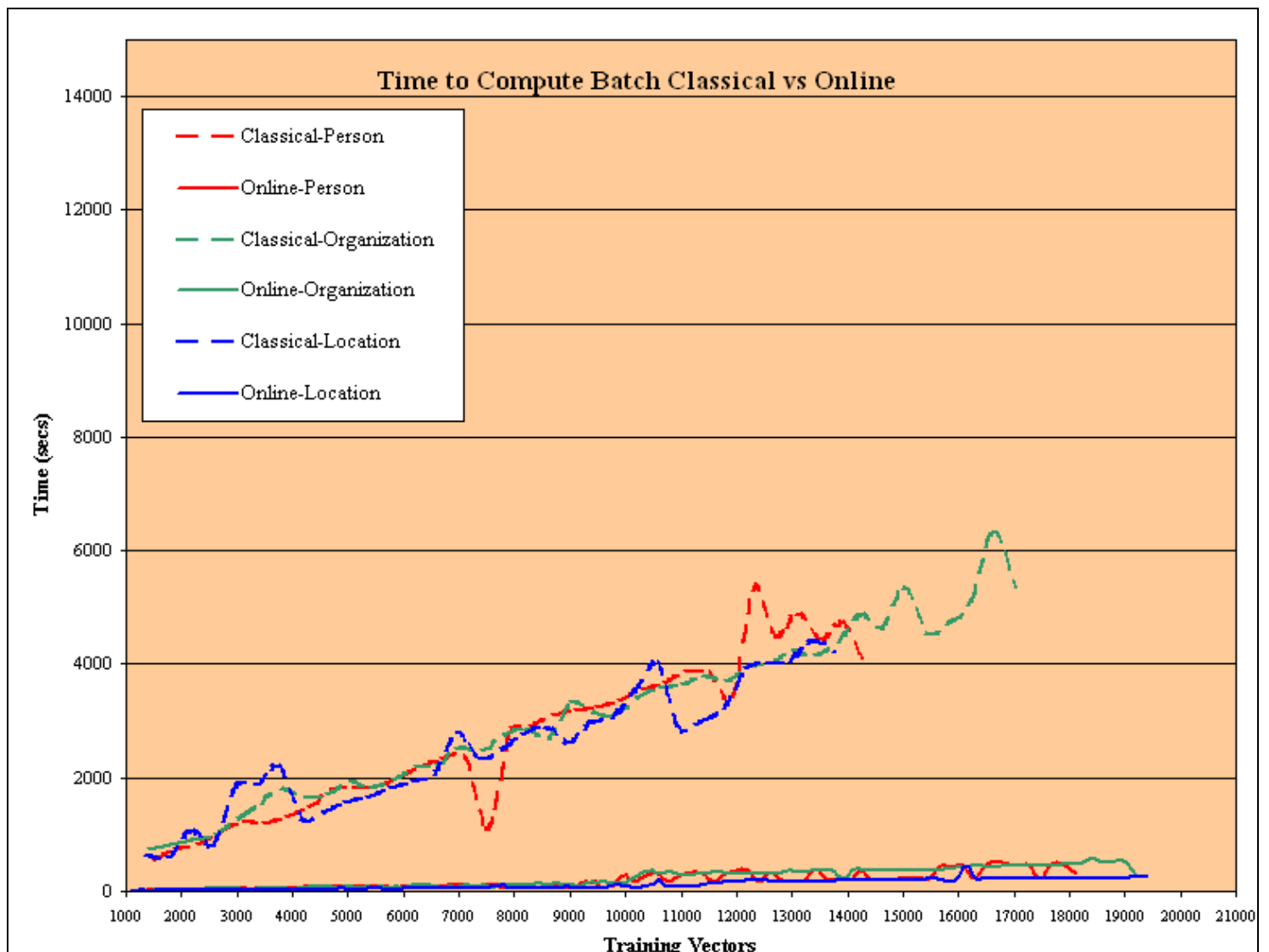
3.6.1 – Διάγραμμα χρόνου για την κλασική μέθοδο ενεργητικής μάθησης

Στα δύο διαγράμματα (3.6.1 και 3.6.2) φαίνεται επίσης καθαρά ότι η φάση του ελέγχου (training) είναι περίπου γραμμική με πολύ μικρή κλίση και στις δύο μεθόδους ενεργητικής μάθησης.



3.6.2 – Διάγραμμα χρόνου για την επιγραμμική μέθοδο ενεργητικής μάθησης

Είναι ενδιαφέρον ότι ο χρόνος επανεκπαίδευσης των ΜΔΥ αυξάνει πολύ πιο απότομα όταν χρησιμοποιείται η επιγραμμική μέθοδος, ενώ αντίθετα η αύξηση είναι λιγότερο απότομη όταν χρησιμοποιείται η κλασική μέθοδος. Ενδέχεται στην κλασική μέθοδο να καταλήγουμε να επιλέγουμε διαρκώς τα ίδια (ή πολλά ίδια) παραδείγματα εκπαίδευσης από ένα σημείο και πέρα (βλ. και ενότητα 3.4.2), τα οποία ουσιαστικά αγνοούνται από τις ΜΔΥ, ενώ στην επιγραμμική μέθοδο τα επιλεγόμενα παραδείγματα ενδέχεται να έχουν μεγαλύτερη ποικιλία, οπότε δεν είναι δυνατόν να αγνοηθούν.



3.6.3 – Διάγραμμα χρόνου επιλογής διανυσμάτων εκπαίδευσης

Το διάγραμμα 3.6.3 δείχνει ότι η φάση επιλογής διανυσμάτων εκπαίδευσης της επιγραμμικής μεθόδου απαιτεί περίπου σταθερό και σημαντικά λιγότερο χρόνο από την αντίστοιχη φάση της κλασικής μεθόδου ενεργητικής μάθησης.

4. Συμπεράσματα και Μελλοντικές Επεκτάσεις

Συνοψίζοντας, σε αυτή την εργασία μελετήθηκε η επέκταση του συστήματος Αναγνώρισης και Κατάταξης Ονομάτων Οντοτήτων των Λουκαρέλλι και Βασιλάκου. Εκτός από χρονικές εκφράσεις, που αναγνωρίζονται όπως στο προηγούμενο σύστημα με τη χρήση προτύπων, το νέο σύστημα αναγνωρίζει και κατατάσσει ονόματα προσώπων, οργανισμών και τοποθεσιών χρησιμοποιώντας τρεις ανεξάρτητες ΜΔΥ και μία μόνο σάρωση των κειμένων. Οι τρεις ΜΔΥ είναι δυνατόν να εκπαιδευθούν τόσο με παθητική όσο και με ενεργητική μάθηση. Στην ενεργητική μάθηση μελετήθηκαν δυο διαφορετικές μέθοδοι επιλογής διανυσμάτων προς επισημείωση, η μέθοδος που είχαν χρησιμοποιήσει οι Λουκαρέλλι και Βασιλάκος, την οποία ονομάζουμε «κλασική», και μια υπολογιστικά πολύ φθηνότερη μέθοδος, καλούμενη «επιγραμμική», που επιτυγχάνει εξίσου καλά ή και καλύτερα αποτελέσματα με την κλασική.

Αφήνοντας κατά μέρος την αναγνώριση ημερομηνιών, με την οποία δεν ασχοληθήκαμε ιδιαίτερα σε αυτή την εργασία, τα καλύτερα αποτελέσματα επιτεύχθηκαν εν γένει στην αναγνώριση ονομάτων προσώπων, ακολουθούμενα από την αναγνώριση ονομάτων οργανισμών, ενώ αντιθέτως τα αποτελέσματα της αναγνώρισης ονομάτων τοποθεσιών ήταν πολύ χαμηλότερα. Για την κατηγορία των **προσώπων**, το καλύτερο f-measure (**83,93%**) επιτεύχθηκε με την επιγραμμική μέθοδο στα 18.111 διανύσματα εκπαίδευσης. Για την κατηγορία των **οργανισμών**, το καλύτερο f-measure (**70,87%**) επιτεύχθηκε πάλι με την επιγραμμική μέθοδο στα 17.349 διανύσματα εκπαίδευσης. Τέλος, για την κατηγορία των **τοποθεσιών** το καλύτερο f-measure (**52,99%**) επιτεύχθηκε με την παθητική μάθηση στα 14.368 διανύσματα εκπαίδευσης. Αντίθετα από τα αποτελέσματα των Λουκαρέλλι και Βασιλάκου, τα αποτελέσματα της ενεργητικής μάθησης σε αυτή την εργασία ήταν σε αρκετές περιπτώσεις χειρότερα από εκείνα της παθητικής, αν και τα αποτελέσματα δεν είναι άμεσα συγκρίσιμα, επειδή οι τρεις ΜΔΥ εκπαιδεύονται τώρα ανεξάρτητα, αντίθετα από το σύστημα των Λουκαρέλλι και Βασιλάκου.

Δεδομένων των παραπάνω αποτελεσμάτων, προτείνονται οι εξής μελλοντικές βελτιώσεις του συστήματος:

- Η χρήση ενός επισημειωτή μερών του λόγου (*part of speech tagger*), που θα επιτρέψει να ενσωματωθούν στα διανύσματα ιδιότητες οι οποίες θα δείχνουν σε ποια μέρη του λόγου ανήκει η προς κατάταξη λεκτική μονάδα και οι γειτονικές της. Η αναγνώριση συγκεκριμένων μερών του λόγου (π.χ. δεικτικές αντωνυμίες, προθέσεις) που συναντώνται πολύ συχνά πριν από ονόματα τοποθεσιών ενδέχεται να βελτιώσει τις επιδόσεις του συστήματος σε αυτή την κατηγορία ονομάτων.
- Η περαιτέρω διερεύνηση της κατηγορίας των τοποθεσιών και ενδεχομένως η υιοθέτηση διαφορετικής προσέγγισης (π.χ. διαφορετικές ιδιότητες). Ιδιαίτερη σημασία πρέπει να δοθεί στο γεγονός ότι το Πληροφοριακό Κέρδος των ιδιοτήτων που χρησιμοποιήθηκαν σε αυτή την κατηγορία ονομάτων ήταν πολύ χαμηλό (η πλειοψηφία των ιδιοτήτων είχε $0 \leq IG < 0.01$). Επιπλέον, κρίνεται απαραίτητο να επισημειωθούν επιπλέον παραδείγματα ελέγχου για την κατηγορία των τοποθεσιών.

- Η προσθήκη δεύτερης σάρωσης των κειμένων, όπως στις εργασίες των Λουκαρέλλι και Βασιλάκου, όπου η προσθήκη δεύτερης σάρωσης είχε βελτιώσει αρκετά τις επιδόσεις του συστήματος.
- Η διερεύνηση του γεγονότος ότι η ενεργητική μάθηση είχε σε αρκετές περιπτώσεις χειρότερα αποτελέσματα από την παθητική στα πειράματα αυτής της εργασίας.
- Η βελτίωση της ταχύτητας του συστήματος, ιδιαίτερα κατά τη φάση της εκπαίδευσης, ενδεχομένως με τη χρήση κάποιας άλλης υλοποίησης ΜΔΥ (π.χ. SVMlight [11]). Αυτό με τη σειρά του θα κάνει δυνατό να χρησιμοποιηθούν περισσότερα διανύσματα εκπαίδευσης.

Βιβλιογραφικές Αναφορές

- [1] *Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων σε Ελληνικά Κείμενα*, Λουκαρέλλι Γεώργιος, Διπλωματική Εργασία Μεταπτυχιακού Διπλώματος Ειδίκευσης, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.
http://www.aueb.gr/users/ion/docs/lucarelli_msc_final_report.pdf
- [2] *A Greek Named-Entity Recognizer That Uses Support Vector Machines and Active Learning*, Georgios Lucarelli and Ion Androutsopoulos, Πρακτικά του 4ου Πανελληνίου Συνεδρίου Τεχνητής Νοημοσύνης (ΣΕΤΝ 2006), Ηράκλειο Κρήτης, 2006.
http://www.aueb.gr/users/ion/docs/setn2006_paper.pdf
- [3] *Αναγνώριση και Κατάταξη Ονομάτων Οντοτήτων σε Ελληνικά Κείμενα με Χρήση Μηχανών Διανυσμάτων Υποστήριξης*, Πτυχιακή Εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2006.
http://www.aueb.gr/users/ion/docs/vassilakos_final_report.pdf
- [4] *Named Entity Recognition in Greek Texts With an Ensemble of Support Vector Machines and Active Learning*, Georgios Lucarelli, Xenophon Vassilakos and Ion Androutsopoulos, International Journal on Artificial Intelligence Tools, World Scientific (υπό δημοσίευση).
http://www.aueb.gr/users/ion/docs/ijait_greek_nerc.pdf
- [5] *An Introduction to Support Vector Machines*, N. Cristianini and J. Shawe-Taylor, Cambridge University Press, 2000.
- [6] *Active Learning With Support Vector Machines*, Andreas Vlachos, MSc Thesis, School of Informatics, University of Edinburgh, 2004.
<http://www.inf.ed.ac.uk/publications/thesis/online/IM040138.pdf>
- [7] *Proceedings of the Seventh Message Understanding Conference*, MUC-7, 1998.
- [8] *LibSVM – A Library for Support Vector Machines*, Chih-Chung Chang and Chih-Jen Lin, 2001.
<http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [9] *Fast Uncertainty Sampling for Labelling Large E-mail Corpora*, Richard Segal, Ted Markowitz, William Arnold, CEAS 2006, Third Conference on E-mail and Anti-Spam, Mountain View, California, ΗΠΑ, July 27-28, 2006.
<http://www.ceas.cc/2006/20.pdf>
- [10] *Online Active Learning Methods for Fast Label Efficient Spam Filtering*, D. Sculley, CEAS 2007, Fourth Conference on E-mail and Anti-Spam, Mountain View, California, ΗΠΑ, August 2-3, 2007.
<http://www.ceas.cc/2007/papers/paper-61.pdf>

[11] *SVMlight – Making Large-Scale SVM Learning Practical*, T. Joachims, περιλαμβάνεται στο βιβλίο *Advances in Kernel Methods - Support Vector Learning* των B. Schölkopf, C. Burges και A. Smola (Επιμ.), MIT Press, 1999.
http://www.cs.cornell.edu/People/tj/publications/joachims_99a.pdf