Οικονομικό Πανεπιστήμιο Αθηνών

ΤΜΗΜΑ
ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αποσαφήνιση της σημασίας λέξεων μέσω συνδυασμού Δικτύων Διάδοσης Ενεργοποίησης και του αλγορίθμου PageRank

ΕΥΑΓΓΕΛΙΑ ΗΛΙΟΠΟΥΛΟΥ

Επιβλέπων : Ι. ΑΝΔΡΟΥΤΣΟΠΟΥΛΟΣ

ΑΘΗΝΑ ΙΟΥΝΙΟΣ 2007

ΠΕΡΙΛΗΨΗ

Οι περισσότερες λέξεις των φυσικών γλωσσών είναι πολύσημες, δηλαδή έχουν διαφορετικές σημασίες ανάλογα με τα συμφραζόμενά τους. Η αποσαφήνιση λέξεων (*word sense disambiguation*) ασχολείται με τον εντοπισμό της σωστής σημασίας κάθε λέξης ενός κειμένου, συνήθως μεταξύ των δυνατών σημασιών που παραθέτει για κάθε λέξη ένα λεξικό. Στη διάρκεια αυτής της εργασίας, συνδυάσαμε τρεις μεθόδους αποσαφήνισης λέξεων που εκμεταλλεύονται σημασιολογικά λεξικά όπως το WordNet. Η πρώτη χρησιμοποιεί Δίκτυα Διάδοσης Ενεργοποίησης (Spreading Activation Networks), η δεύτερη τον αλγόριθμο PageRank, ενώ η τρίτη απλά επιστρέφει τη σημασία που το λεξικό αναφέρει ως συχνότερη. Ο συνδυασμός των τριών μεθόδων έγινε χρησιμοποιώντας επιβλεπόμενη μηχανική μάθηση, πιο συγκεκριμένα Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), μέσω των οποίων αναπτύχθηκαν ταξινομητές που αποφασίζουν πότε το συνολικό σύστημα θα πρέπει να εμπιστευεί κάθε μία από τις τρεις μεθόδους. Τα αποτελέσματα πειραμάτων που έγιναν στη διάρκεια της εργασίας με αγγλικά κείμενα δείχνουν ότι το συνολικό σύστημα επιτυγχάνει καλύτερα αποτελέσματα από τις τρεις μεμονωμένες αρχικές μεθόδους και ότι οι επιδόσεις του είναι συγκρίσιμες με εκείνες των καλύτερων διεθνώς συστημάτων.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω το Λέκτορα του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών κ. Ίωνα Ανδρουτσόπουλο, για την επίβλεψη της παρούσας πτυχιακής εργασίας και την πολύτιμη συμβολή και καθοδήγησή του. Επίσης, ευχαριστώ τον Υποψήφιο Διδάκτορα του Τμήματος Πληροφορικής του Οικονομικού Πανεπιστημίου Αθηνών κ. Γεώργιο Τσατσαρώνη, που με υπομονή και επιμονή με καθοδήγησε και με βοήθησε στην ανάπτυξη των μεθόδων και με τροφοδότησε με το απαραίτητο υλικό. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου, για την υποστήριξή της και κυρίως την υπομονή και κατανόηση που έδειξε σε όλη την περίοδο της εργασίας μου.

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

1.1 Αποσαφήνιση λέξεων

Με τον όρο *αποσαφήνιση λέξεων* (word sense disambiguation) αναφερόμαστε στο πρόβλημα της απόδοσης σε κάθε λέξη ενός κειμένου της σωστής έννοιάς της, συνήθως επιλέγοντας από ένα λεξικό μεταξύ των δυνατών εννοιών της συγκεκριμένης λέξης. Στα ακόλουθα δύο παραδείγματα αγγλικών προτάσεων:

*1. He cashed a check at the **bank**.*
*2. He sat on the **bank** of the river and watched the currents.*

η λέξη «bank» παρουσιάζεται με δύο διαφορετικές έννοιες, οι οποίες, σύμφωνα με το λεξικό WordNet 2.0 [16], μπορούν να περιγραφούν ως εξής:

1. *depository financial institution, **bank**, banking concern, banking company - - (a financial institution that accepts deposits and channels the money into lending activities.)*
2. ***bank** - - (sloping land (especially the slope beside a body of water))*

Μάλιστα, πάντα κατά το WordNet, η συγκεκριμένη λέξη έχει συνολικά δέκα έννοιες ως ουσιαστικό και οχτώ ως ρήμα.[1] Ανάλογη πολυσημία παρουσιάζουν οι περισσότερες λέξεις των φυσικών γλωσσών.

1.2 Σκοπός της εργασίας

Στη διάρκεια αυτής της εργασίας συνδυάσαμε τρεις μεθόδους αποσαφήνισης λέξεων που εκμεταλλεύονται σημασιολογικά λεξικά όπως το WordNet. Η πρώτη χρησιμοποιεί Δίκτυα Διάδοσης Ενεργοποίησης (Spreading Activation Networks), η δεύτερη τον αλγόριθμο PageRank, ενώ η τρίτη απλά επιστρέφει τη σημασία που το λεξικό αναφέρει ως συχνότερη. Ο συνδυασμός των τριών μεθόδων έγινε χρησιμοποιώντας επιβλεπόμενη μηχανική μάθηση, πιο συγκεκριμένα Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), μέσω των οποίων αναπτύχθηκαν ταξινομητές που αποφασίζουν πότε το συνολικό σύστημα θα πρέπει να εμπιστευτεί κάθε μία από τις τρεις μεθόδους. Τα αποτελέσματα πειραμάτων που έγιναν στη διάρκεια της εργασίας με αγγλικά κείμενα δείχνουν ότι το συνολικό σύστημα επιτυγχάνει καλύτερα αποτελέσματα από τις τρεις μεμονωμένες αρχικές

[1] Βλ. http://wordnet.princeton.edu/.

μεθόδους και ότι οι επιδόσεις του είναι συγκρίσιμες με εκείνες των καλύτερων διεθνώς συστημάτων, όπως, για παράδειγμα, τα συστήματα των [2] και [3].

1.3 Περιεχόμενα των επόμενων κεφαλαίων

Το Κεφάλαιο 2 περιλαμβάνει το υπόβαθρο της εργασίας. Περιγράφεται συνοπτικά το λεξικό WordNet, το οποίο και χρησιμοποιήσαμε, καθώς και τα σύνολα δεδομένων *Senseval 2*, *Senseval 3* και *Semcor 2*, από τα οποία αντλήσαμε τα δεδομένα των πειραμάτων μας. Στο ίδιο κεφάλαιο παρουσιάζεται λεπτομερέστερα το πρόβλημα της αποσαφήνισης λέξεων και μέθοδοι που έχουν αναπτυχθεί για την επίλυσή του. Στο Κεφάλαιο 3 περιγράφεται η συνδυασμένη μέθοδος που αναπτύχθηκε στη διάρκεια της εργασίας. Ακολουθούν, στο Κεφάλαιο 4, τα πειραματικά αποτελέσματα της εργασίας, τα οποία περιλαμβάνουν αποτελέσματα τόσο των τριών μεμονωμένων αρχικών μεθόδων όσο και του συνδυασμού τους. Τέλος, το Κεφάλαιο 5 συνοψίζει τα συμπεράσματα της εργασίας και προτείνει μελλοντικές μελέτες και βελτιώσεις.

2 ΥΠΟΒΑΘΡΟ

2.1 Το ηλεκτρονικό λεξικό WordNet

Το WordNet [16] είναι ένα ηλεκτρονικό λεξικό που περιέχει ουσιαστικά, ρήματα, επίθετα και επιρρήματα της αγγλικής γλώσσας, οργανωμένα σε σύνολα συνωνύμων (*synsets*).[2] Κάθε σύνολο συνωνύμων αντιστοιχεί σε μία σημασία, την οποία μπορούν να εκφράσουν οι λέξεις του συνόλου. Οι πολύσημες λέξεις συμμετέχουν σε τόσα σύνολα συνωνύμων όσα και οι δυνατές σημασίες τους. Για κάθε σύνολο συνωνύμων, το WordNet παρέχει και μια σύντομη επεξήγηση (*gloss*) της αντίστοιχης έννοιας. Για παράδειγμα, η λέξη «bank» ως ουσιαστικό συμμετέχει σε σύνολα συνωνύμων όπως τα ακόλουθα, όπου εντός παρενθέσεων φαίνονται οι επεξηγήσεις:

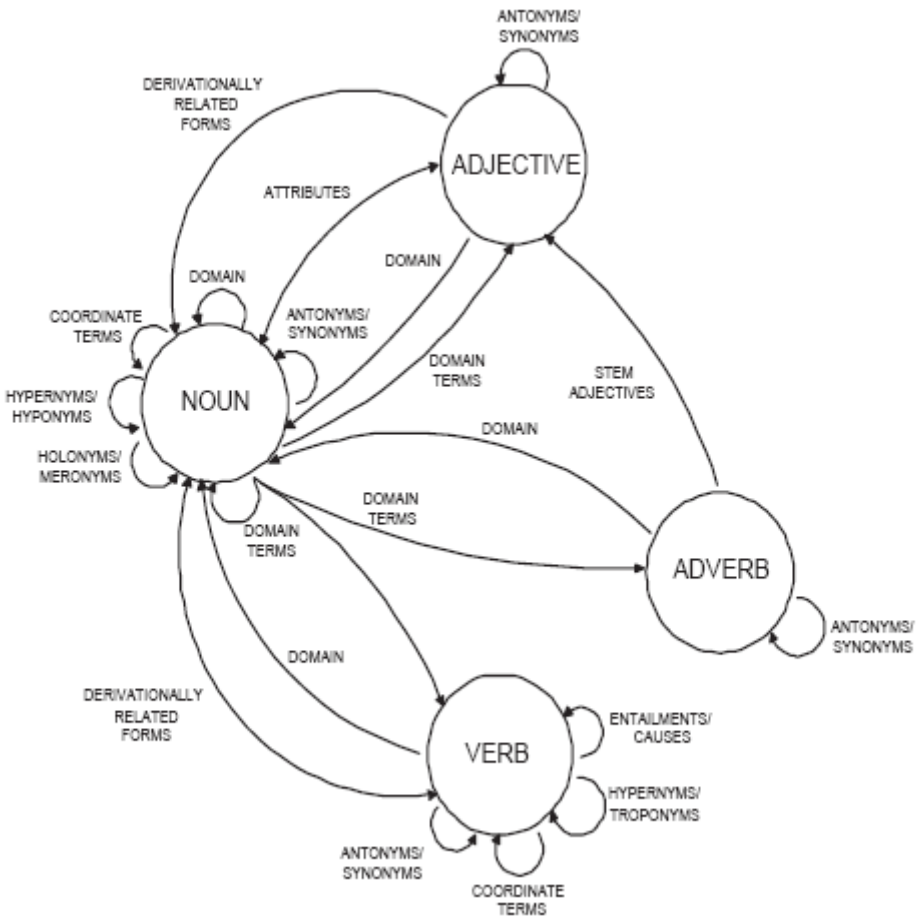
- {«bank»} (sloping land, especially the slope beside a body of water),
- {«depository financial institution», «bank», «banking concern», «banking company»} (a financial institution that accepts deposits and channels the money into lending activities),
- {«bank», «bank building»} (a building in which the business of banking is transacted),
- {«savings bank», «coin bank», «money box», «bank»} (a container, usually with a slot in the top, for keeping money at home),

ενώ ως ρήμα συμμετέχει σε σύνολα συνωνύμων όπως τα εξής:

- {«deposit», «bank»} (put into a bank account),
- {«trust», «swear», «rely», «bank»} (have confidence or faith in).

Τα σύνολα συνωνύμων συνδέονται με ακμές ποικίλων τύπων, απεικονίζοντας με αυτόν τον τρόπο τις σημασιολογικές σχέσεις των αντιστοίχων εννοιών. Περιλαμβάνονται, για παράδειγμα, ακμές που αντιστοιχούν σε σχέσεις υπερωνύμου–υπωνύμου (hypernym–hyponym), μερωνύμου-ολωνύμου (meronym– holonym), συνωνύμου–αντιθέτου (synonym–antonym) κλπ. Μια γραφική αναπαράσταση των σημασιολογικών σχέσεων του WordNet παρατίθεται στην Εικόνα 1. Η εικόνα προέρχεται από την εργασία [4], όπου και εξηγείται λεπτομερέστερα.

[2] Χρησιμοποιούμε τις εκδόσεις 2.0 και 2.1 του WordNet. Αντίστοιχα λεξικά έχουν κατασκευαστεί και για πολλές άλλες φυσικές γλώσσες, συμπεριλαμβανομένης της ελληνικής. Δυστυχώς η ελληνική μορφή του WordNet δεν διατίθεται ελεύθερα. Δείτε, επίσης, το EuroWordNet (http://www.illc.uva.nl/EuroWordNet/).



Εικόνα 1 : Σημασιολογικές σχέσεις του WordNet 2.0.

2.2 Πειραματικά σύνολα δεδομένων αποσαφήνισης λέξεων

Στα πειράματα της εργασίας χρησιμοποιήσαμε τρία καθιερωμένα σύνολα δεδομένων αποσαφήνισης λέξεων: τα *Senseval 2*[3], *Senseval 3*[4] και *Semcor 2*[5]. Πρόκειται για συλλογές κειμένων των οποίων οι λέξεις είναι σημειωμένες με έννοιες (σύνολα συνωνύμων) του WordNet 2.0 (για τα *Senseval 2* και *Senseval 3*) και του WordNet 2.1 (για το *Semcor 2*). Ο Πίνακας 1 δείχνει το πλήθος των διαφορετικών λέξεων και εμφανίσεων λέξεων των τριών συνόλων δεδομένων ανά μέρος του λόγου και συνολικά. Ο Πίνακας 2 δείχνει το πλήθος των πολύσημων και μονόσημων διαφορετικών λέξεων των τριών συνόλων δεδομένων ανά μέρος του λόγου, όπως επίσης και τη μέση πολυσημία, δηλαδή τον μέσο αριθμό εννοιών ανά λέξη. Πειραματιστήκαμε με την αποσαφήνιση ουσιαστικών, ρημάτων και επιθέτων. Δεν εξετάσαμε την αποσαφήνιση επιρρημάτων, των οποίων η πολυσημία είναι εν γένει χαμηλότερη, ιδιαίτερα στο *Senseval 3*. Όπως φαίνεται από τον τελευταίο πίνακα, η αποσαφήνιση των ρημάτων είναι ιδιαίτερα δύσκολη.

[3] Βλ. http://www.itri.brighton.ac.uk/events/senseval.
[4] Βλ. http://www.senseval.org/senseval3.
[5] Βλ. http://www.cs.unt.edu/~rada/downloads.html#semcor.

Σύνολο Δεδομένων	Ουσιαστικά	Ρήματα	Επίθετα	Επιρρήματα	Σύνολο
Senseval 2	1073	535	432	263	2303
Senseval 3	892	725	348	14	1979
Semcor 2	87422	47701	34835	19709	189667

Πίνακας 1 : Λέξεις στα Senseval 2, Senseval 3 και Semcor 2.

	Ουσιαστικά	Ρήματα	Επίθετα	Επιρρήματα	Σύνολο
Senseval 2					
Μονόσημες	260	33	80	91	464
Πολύσημες	813	502	352	172	1839
Μέση Πολυσημία	4	9	3	3	5
Senseval 3					
Μονόσημες	193	39	72	13	317
Πολύσημες	699	686	276	1	1662
Μέση Πολυσημία	5	11	4	1	7
Semcor 2					
Μονόσημες	16990	2584	9854	7831	37259
Πολύσημες	70432	45117	24981	11878	152408
Μέση Πολυσημία	4	10	4	2	5

Πίνακας 2 : Μονόσημες και πολύσημες λέξεις στα Senseval 2, Senseval 3 και Semcor 2.

2.3 Προηγούμενες μέθοδοι αποσαφήνισης λέξεων

Στις επόμενες υποενότητες περιγράφονται μερικές από τις μεθόδους αποσαφήνισης λέξεων που έχουν προταθεί, δίνοντας περισσότερη έμφαση στις μεθόδους που χρησιμοποιήθηκαν κατά τη διάρκεια της εργασίας ή που σχετίζονται με αυτές.

2.3.1 Η μέθοδος των Veronis και Ide

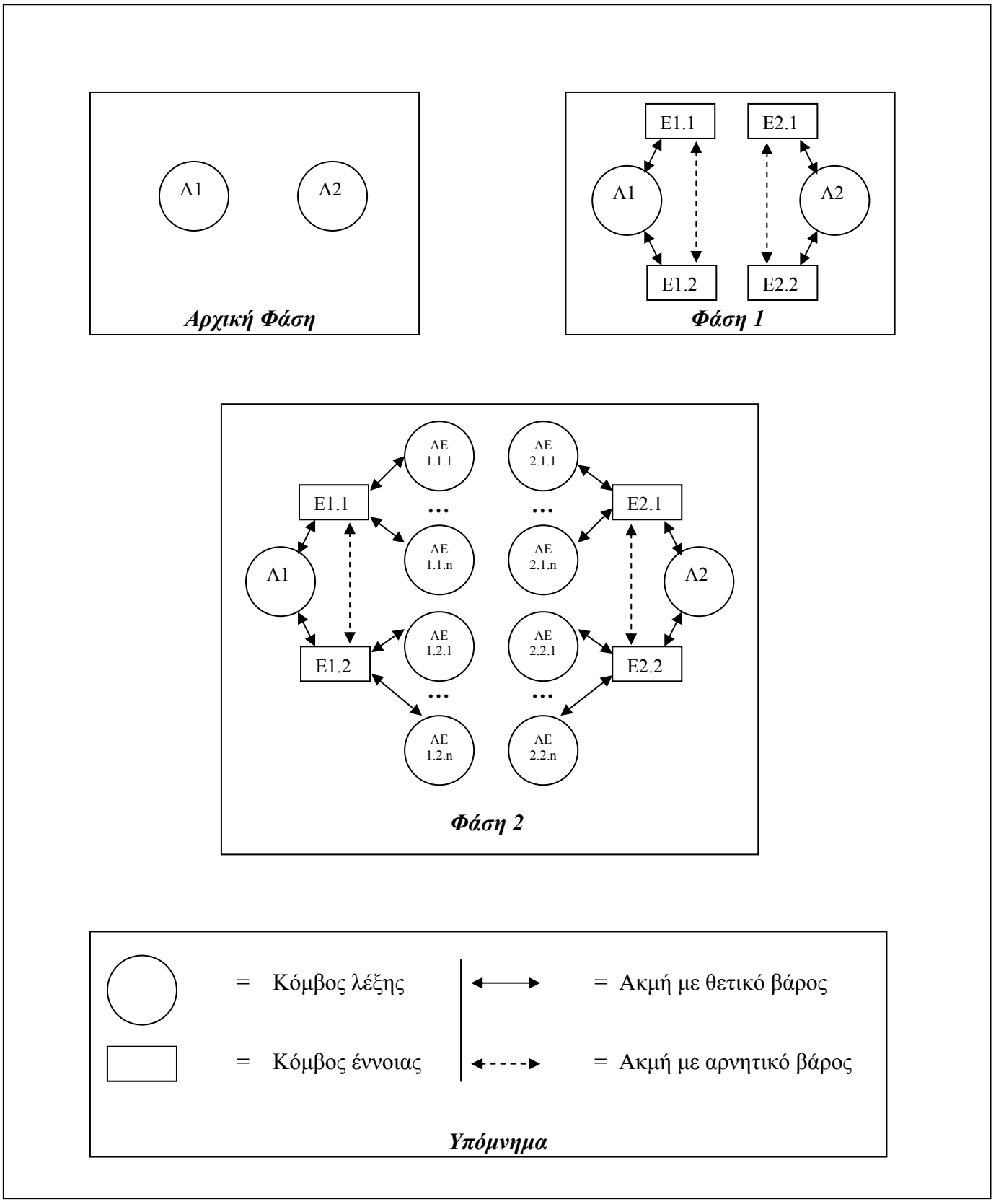
Οι Veronis και Ide [15, 17] ανέπτυξαν μία μέθοδο αποσαφήνισης λέξεων που χρησιμοποιεί *Δίκτυα Διάδοσης Ενεργοποίησης* (ΔΔΕ, Spreading Activation Networks – SAN), μια μορφή γράφου. Για κάθε πρόταση (ή γενικότερα τμήμα κειμένου) του οποίου τις λέξεις θέλουμε να αποσαφηνίσουμε, δημιουργούμε ένα διαφορετικό ΔΔΕ. Για κάθε λέξη προς αποσαφήνιση, κατασκευάζουμε έναν κόμβο, που αντιστοιχεί στη λέξη (κόμβοι Λ στο σχήμα 1). Στη συνέχεια, ανακτούμε τις έννοιες των λέξεων προς αποσαφήνιση από ένα ηλεκτρονικό λεξικό (π.χ. WordNet) και προσθέτουμε στο ΔΔΕ έναν νέο κόμβο για κάθε έννοια (κόμβοι Ε στο σχήμα 1). Κάθε αρχικός κόμβος λέξης συνδέεται άμεσα με όλους τους κόμβους που αντιστοιχούν σε έννοιες της λέξης μέσω

ακμών που έχουν θετικά βάρη. Κόμβοι που αντιστοιχούν σε έννοιες της ίδιας λέξης συνδέονται μεταξύ τους με ακμές που έχουν αρνητικά βάρη. Κατόπιν, η επεξήγηση (gloss) κάθε έννοιας αντλείται από το ίδιο λεξικό, χωρίζεται σε λεκτικές μονάδες (tokens) και οι λεκτικές μονάδες της επεξήγησης μετατρέπονται στα αντίστοιχα λήμματα (βασικές μορφές λέξεων) του λεξικού. Οι τερματικές λέξεις (stop-words) αφαιρούνται. Κάθε λήμμα (βασική μορφή λεκτικής μονάδας) που παράγεται με τον προαναφερθέντα τρόπο από την επεξήγηση μιας έννοιας προστίθεται ως κόμβος στο δίκτυο (κόμβοι ΛΕ στο σχήμα 1) και συνδέεται μέσω ακμών με θετικά βάρη με τον κόμβο της έννοιας. Το δίκτυο συνεχίζει να επεκτείνεται με τον ίδιο τρόπο, προσθέτοντας τώρα τους κόμβους εννοιών των λεκτικών μονάδων των επεξηγήσεων κ.ο.κ, μέχρι να έχουν προστεθεί σε αυτό κόμβοι για ένα μεγάλο μέρος ή για το σύνολο των λέξεων του λεξικού.

Αφού το δίκτυο κατασκευαστεί και αναπτυχθεί πλήρως, οι αρχικοί κόμβοι λέξεων ενεργοποιούνται, περίπου όπως ενεργοποιούνται οι κόμβοι εισόδου ενός νευρωνικού δικτύου. Η ενεργοποίηση επεκτείνεται σε ολόκληρο το δίκτυο, σύμφωνα μια στρατηγική διάδοσης ενεργοποίησης, η οποία εγγυάται ότι τελικά μόνο ένας κόμβος έννοιας για κάθε αρχικό κόμβο λέξης θα έχει θετική τιμή ενεργοποίησης. Αυτός ο κόμβος έννοιας θεωρείται πως παριστάνει την έννοια που έχει η αντίστοιχη λέξη στο τμήμα κειμένου που αποσαφηνίζουμε. Να σημειώσουμε ότι αυτή η προσέγγιση υποθέτει ότι όλες οι εμφανίσεις της ίδιας λέξης στο κείμενο που αποσαφηνίζουμε έχουν την ίδια έννοια, κάτι το οποίο είναι λογικό, τουλάχιστον για μικρά τμήματα κειμένου όπως προτάσεις ή παραγράφους.

Η παραπάνω περιγραφή της μεθόδου βασίζεται στην εργασία [4].[6] Σύμφωνα με τους συγγραφείς της εργασίας αυτής, η μέθοδος των Veronis και Ide επιτυγχάνει στα δεδομένα του *Senseval 2* ποσοστό ορθότητας (accuracy) ~44.72%. Το ποσοστό ορθότητας μετριέται ως το ποσοστό των εμφανίσεων λέξεων που αντιστοιχίζονται στη σωστή τους έννοια.

[6] Το σχήμα 1 περιελήφθη στην παρούσα εργασία κατόπιν αδείας των συγγραφέων της εργασίας [4].



Σχήμα 1 : Παράδειγμα επέκτασης του ΔΔΕ κατά τους Veronis και Ide.

2.3.2 Η μέθοδος των Τσατσαρώνη κ.ά.

Η μέθοδος των Τσατσαρώνη κ.ά. [4] χρησιμοποιεί και αυτή Δίκτυα Διάδοσης Ενεργοποίησης (ΔΔΕ). Τα ΔΔΕ κατασκευάζονται, όμως, με διαφορετική μέθοδο από αυτήν των Veronis και Ide (ενότητα 2.3.1), ενώ χρησιμοποιείται και ένα νέο σχήμα αντιστοίχισης βαρών στις ακμές του ΔΔΕ. Ο αλγόριθμος αποσαφηνίζει ένα κείμενο πρόταση προς πρόταση. Σε κάθε πρόταση αποσαφηνίζει μόνο τις λέξεις που υπάρχουν στο ηλεκτρονικό λεξικό που χρησιμοποιούμε (π.χ. WordNet). Επίσης, προϋποθέτει ότι κάθε λέξη του κειμένου έχει ήδη σημειωθεί με το μέρος του λόγου στο οποίο ανήκει (ουσιαστικό, ρήμα, επίθετο και επίρρημα), κάτι που μπορεί να γίνει με ένα σύστημα αναγνώρισης μερών του λόγου (part-of-speech tagger). Για κάθε πρόταση κατασκευάζεται ένα ΔΔΕ.

Για να κατασκευάσουμε το ΔΔΕ μιας πρότασης, αρχικά προστίθενται στο δίκτυο οι λέξεις της πρότασης (κόμβοι Λ) με όλες τις δυνατές, κατά το λεξικό, έννοιές τους (κόμβοι Ε), όπως στον αλγόριθμο των Veronis και Ide. Αυτή είναι η αρχική φάση (βλ. σχήμα 2). Στη συνέχεια, όλες οι έννοιες του λεξικού που συνδέονται άμεσα με τις έννοιες των κόμβων εννοιών της αρχικής φάσης, μέσω οποιασδήποτε σημασιολογικής σχέσης (π.χ. συνώνυμα, υπερώνυμα κλπ.), προστίθενται στο ΔΔΕ, μαζί με τις αντίστοιχες ακμές. Αυτό είναι το πρώτο βήμα επέκτασης (βλ. σχήμα 2). Καθώς όλες οι σχέσεις του WordNet, του λεξικού που χρησιμοποιούμε, είναι αμφίδρομες, αμφίδρομες θα είναι και οι ακμές που συνδέουν τους κόμβους εννοιών μεταξύ τους. Το δίκτυο συνεχίζει να αυξάνεται, επαναλαμβάνοντας το προηγούμενο βήμα, μέχρι να υπάρχει μονοπάτι μεταξύ κάθε ζεύγους αρχικών κόμβων λέξεων (κόμβοι Λ). Τότε το δίκτυο θεωρείται *συνδεδεμένο*. Εάν δεν υπάρχουν άλλες έννοιες για να επεκταθεί το ΔΔΕ και το ΔΔΕ που έχει προκύψει δεν είναι συνδεδεμένο, η αποσαφήνιση των λέξεων αποτυγχάνει. Για να μην παγιδευτούμε σε κύκλους, η επέκταση του δικτύου γίνεται με αναζήτηση πρώτα σε πλάτος, με χρήση κλειστού συνόλου.

Αν καταλήξουμε σε συνδεδεμένο δίκτυο, εφαρμόζεται στη συνέχεια η στρατηγική ενεργοποίησης, η οποία αποτελείται και αυτή από επαναλήψεις. Αρχικά, όλοι οι κόμβοι του ΔΔΕ έχουν επίπεδο ενεργοποίησης 0, εκτός από τους αρχικούς κόμβους λέξεων, οι οποίοι έχουν επίπεδο ενεργοποίησης 1. Σε κάθε επανάληψη της στρατηγικής ενεργοποίησης, κάθε κόμβος μεταδίδει την τιμή ενεργοποίησής του σε όλους τους γείτονές του, ως συνάρτηση της τρέχουσας τιμής ενεργοποίησης που έχει και των βαρών των ακμών που τον συνδέουν με τους γείτονές του. Υιοθετείται η στρατηγική ενεργοποίησης που προτάθηκε από τους Berger κ.ά. [2004], με την τροποποίηση ότι χρησιμοποιείται ένα νέο σχήμα ανάθεσης βαρών στις ακμές. Συγκεκριμένα, σε κάθε επανάληψη ρ, κάθε κόμβος κ του ΔΔΕ έχει ένα επίπεδο ενεργοποίησης $\mathrm{A}_\kappa(\rho)$ και μία έξοδο $\mathrm{O}_\kappa(\rho)$, η οποία προκύπτει ως συνάρτηση του επιπέδου ενεργοποίησης:

$$\mathrm{O}_\kappa(\rho) = f\left(\mathrm{A}_\kappa(\rho)\right) \qquad (1)$$

Η έξοδος κάθε κόμβου κ, επηρεάζει το επίπεδο ενεργοποίησης της επόμενης επανάληψης κάθε κόμβου λ με τον οποίο συνδέεται άμεσα ο κόμβος κ μέσω ακμής $\alpha_{\kappa\lambda}$. Το επίπεδο ενεργοποίησης κάθε κόμβου λ του δικτύου στην επανάληψη ρ είναι συνάρτηση της εξόδου, της επανάληψης $\rho-1$, κάθε γείτονα κόμβου κ που συνδέεται άμεσα με τον λ μέσω ακμής $\alpha_{\kappa\lambda}$, καθώς και του βάρους $\mathrm{B}_{\kappa\lambda}$ της ακμής $\alpha_{\kappa\lambda}$, όπως φαίνεται στην ισότητα (2).

$$\mathrm{A}_{\lambda}(\rho)=\sum_{\kappa}\mathrm{O}_{\kappa}(\rho-1)\cdot\mathrm{B}_{\kappa\lambda} \tag{2}$$

Πιο συγκεκριμένα, αντί της ισότητας (1) χρησιμοποιείται η ισότητα (3):

$$\mathrm{O}_{\kappa}(\rho)=\begin{cases}0 & ,\text{ εάν } \mathrm{A}_{\kappa}(\rho)<\tau\\ \dfrac{F_{\kappa}}{\rho+1}\cdot\mathrm{A}_{\kappa}(\rho), & \text{αλλιώς}\end{cases} \tag{3}$$

Η ισότητα (3) δεν επιτρέπει σε κόμβους με χαμηλά επίπεδα ενεργοποίησης, πιο συγκεκριμένα κόμβους με επίπεδα ενεργοποίησης κάτω από ένα κατώφλι τ, να επηρεάσουν τους γειτονικούς τους κόμβους. Ο παράγοντας $\dfrac{1}{\rho+1}$ ελαττώνει την επίδραση κάθε κόμβου στους γείτονές του, όσο προχωρούμε σε επόμενες επαναλήψεις. Ο παράγοντας F_{κ}, που ορίζεται από την ισότητα (4), μειώνει την επίδραση των κόμβων που συνδέονται με πολλούς γείτονες.

$$F_{\kappa}=\left(1-\frac{\mathrm{K}_{\kappa}}{\mathrm{K}_{\mathrm{T}}}\right) \tag{4}$$

K_{T} είναι το συνολικό πλήθος των κόμβων του ΔΔΕ και K_{κ} είναι το πλήθος των κόμβων που συνδέονται άμεσα με τον κ μέσω ακμών που ξεκινούν από τον κ.

Στην ανάκτηση πληροφοριών, ένας συνήθης τρόπος για να υπολογίσουμε τη σημαντικότητα ενός όρου (term) σε ένα κείμενο είναι να πολλαπλασιάσουμε τη συχνότητα του όρου στο κείμενο (ΣΟ, token frequency - TF) με τον αντίστροφο (ή το λογάριθμο του αντίστροφου) του αριθμού των κειμένων στα οποία εμφανίζεται αυτός ο όρος (ΑΣΚ, inverse document frequency - IDF). Ένα ανάλογο σχήμα χρησιμοποιεί η μέθοδος των Τσατσαρώνη κ.ά. κατά την ανάθεση βαρών στις ακμές του ΔΔΕ.

Κατά την κατασκευή του δικτύου, κάθε ακμή του ΔΔΕ αρχικά λαμβάνει βάρος –1 εάν είναι αρνητική ακμή (ακμές που αναπαριστούν αντώνυμα και ανταγωνιστικές έννοιες της ίδιας λέξης), ή 1 εάν είναι ακμή με θετικό βάρος (όλες οι υπόλοιπες ακμές). Όταν η κατασκευή του δικτύου ολοκληρωθεί, πολλαπλασιάζουμε το αρχικό βάρος $\beta_{\lambda\kappa}$ κάθε ακμής $a_{\lambda\kappa}$ με την ακόλουθη ποσότητα:

$$\Sigma\mathrm{T}\mathrm{A}(\alpha_{\lambda\kappa})\cdot\mathrm{A}\Sigma\mathrm{K}\Delta(\alpha_{\lambda\kappa}) \tag{5}$$

Η ποσότητα ΣΤΑ ορίζεται στην ισότητα (6) και είναι η συχνότητα του τύπου της ακμής (edge type frequency – ETF), που αντιστοιχεί στην ποσότητα ΣΟ της ανάκτησης πληροφοριών. Αναπαριστά το ποσοστό των εξερχόμενων ακμών του κόμβου λ που είναι του ίδιου τύπου με την ακμή $\alpha_{\lambda\kappa}$. Κατά τον υπολογισμό των βαρών των ακμών, οι ακμές που αντιστοιχούν σε υπερώνυμα και υπώνυμα θεωρούνται του ίδιου τύπου, αφού είναι συμμετρικές. Η ιδέα πίσω από τον όρο ΣΤΑ, είναι να πριμοδοτεί ακμές των οποίων ο τύπος είναι συχνός ανάμεσα στις εξερχόμενες ακμές του κόμβου λ. Κι αυτό γιατί κόμβοι με πολλές εξερχόμενες ακμές του ιδίου τύπου είναι πιθανότερο να αποτελούν κομβικά σημεία (hubs) της σημασιολογικής σχέσης που αντιστοιχεί στις συχνές ακμές.

$$\text{ΣΤΑ}\left(\alpha_{\lambda\kappa}\right) = \frac{\left|\left\{\alpha_{\lambda\iota} \mid \tau\acute{\upsilon}\pi o\varsigma\left(\alpha_{\lambda\iota}\right) = \tau\acute{\upsilon}\pi o\varsigma\left(\alpha_{\lambda\kappa}\right)\right\}\right|}{\left|\left\{\alpha_{\lambda\iota}\right\}\right|} \tag{6}$$

Ο δεύτερος όρος της ισότητας (5), ορίζεται στην ισότητα (7) και είναι το ανάλογο του ΑΣΚ στην ανάκτηση πληροφοριών:

$$\text{ΑΣΚΔ}\left(\alpha_{\lambda\kappa}\right) = \log\frac{\text{N}+1}{\text{N}_{\tau\acute{\upsilon}\pi o\varsigma\left(\alpha_{\lambda\kappa}\right)}} \tag{7}$$

N είναι το πλήθος των κόμβων του ΔΔΕ, ενώ $\text{N}_{\tau\acute{\upsilon}\pi o\varsigma\left(\alpha_{\lambda\kappa}\right)}$ είναι το πλήθος των κόμβων που έχουν εξερχόμενες ακμές του ίδιου τύπου με την $\alpha_{\lambda\kappa}$. Όπως και με τον όρο ΑΣΚ, η ιδέα πίσω από τον όρο ΑΣΚΔ είναι ότι θέλουμε να πριμοδοτήσουμε ακμές με τύπους που σπανίζουν σε ολόκληρο το ΔΔΕ.

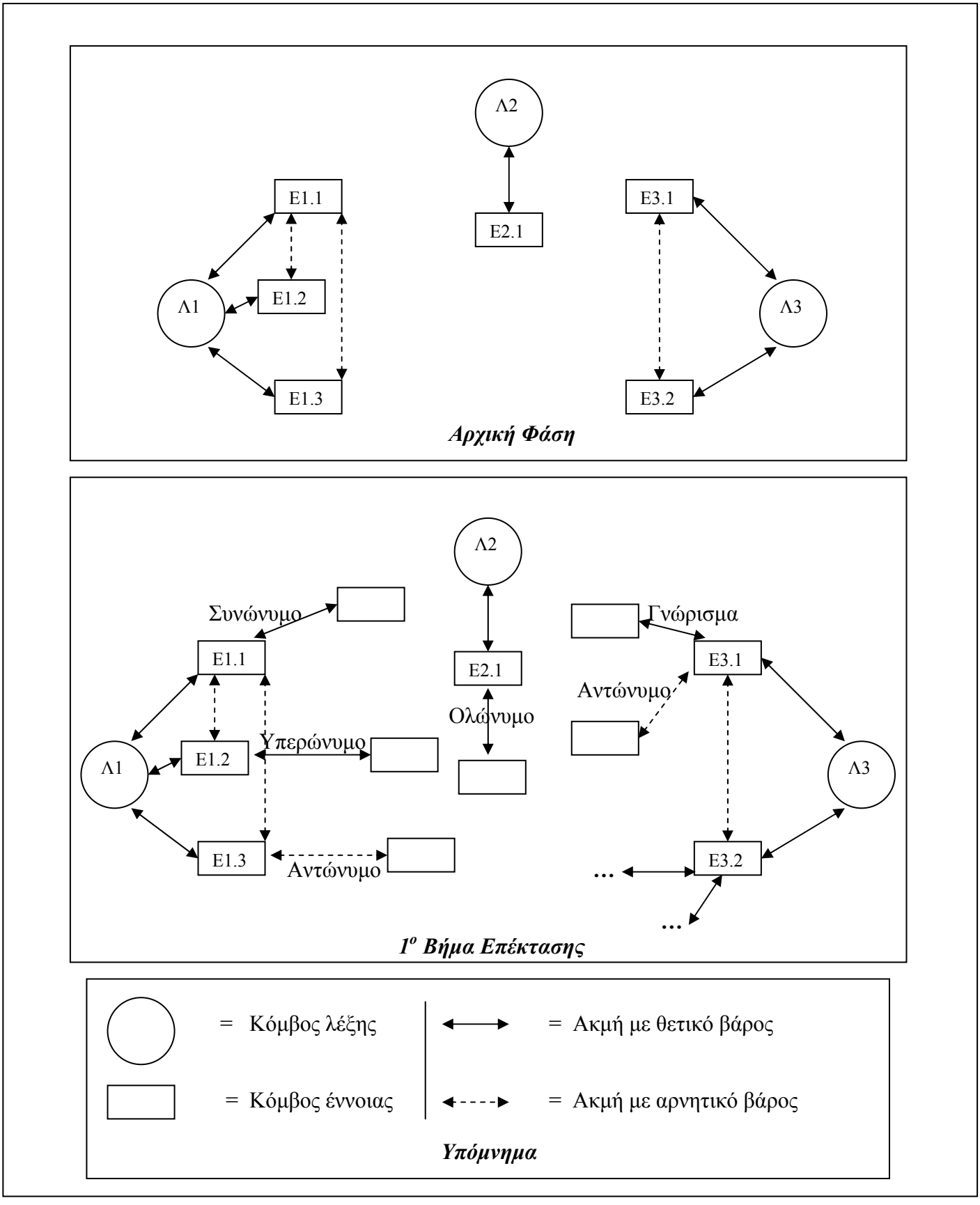
Ο αλγόριθμος αποτελείται από τέσσερα βήματα, που περιγράφονται παρακάτω:

Αλγόριθμος αποσαφήνισης Τσατσαρώνη κ.ά.
Βήμα 1: Χώρισε το κείμενο σε προτάσεις. Για κάθε μία πρόταση, εκτέλεσε τα βήματα 2 έως 4.
Βήμα 2: Κατασκεύασε ένα ΔΔΕ ξεκινώντας από τις λέξεις της πρότασης που ανήκουν σε μέρη του λόγου στα οποία εφαρμόζεται ο αλγόριθμος. Εάν το ΔΔΕ δεν είναι συνδεδεμένο, τερμάτισε τη διαδικασία της αποσαφήνισης.
Βήμα 3: Διάδωσε την ενεργοποίηση μέχρι όλοι οι κόμβοι να είναι ανενεργοί.[7] Για κάθε κόμβο λέξης, αποθήκευσε τον τελευταίο ενεργό κόμβο έννοιας. Αν υπάρχουν πολλοί, διάλεξε εκείνον με την υψηλότερη ενεργοποίηση.
Βήμα 4: Ανάθεσε σε κάθε λέξη την έννοια που αντιστοιχεί στον κόμβο που αποθηκεύτηκε στο προηγούμενο βήμα.

Η παραπάνω περιγραφή της μεθόδου των Τσατσαρώνη κ.ά. βασίζεται στην αντίστοιχη περιγραφή της εργασίας [4]. Σύμφωνα με τους συγγραφείς της εργασίας, η μέθοδός τους επιτυγχάνει ποσοστό ορθότητας ~47.10% στα δεδομένα του *Senseval 2*. Η ίδια εργασία περιγράφει και μια παραλλαγή της μεθόδου, που χρησιμοποιεί και τις επεξηγήσεις (glosses) των σημασιών του WordNet, αλλά η οποία επιτυγχάνει

[7] Το ότι τελικά όλοι οι κόμβοι θα είναι ανενεργοί αποδεικνύεται στην εργασία [4].

χαμηλότερο ποσοστό ορθότητας (~45.68%) στο *Senseval 2*. Γενικά, τα πειραματικά αποτελέσματα της εργασίας [4] δείχνουν ότι η μέθοδος των Τσατσαρώνη κ.ά. επιτυγχάνει μεγαλύτερο ποσοστό ορθότητας από ό,τι η αρχική μέθοδος των Veronis και Ide, στην οποία βασίστηκε η μέθοδος των Τσατσαρώνη κ.ά.



Σχήμα 2 : Παράδειγμα επέκτασης του ΔΔΕ κατά τους Τσατσαρώνη κ.ά.

2.3.3 Η μέθοδος της Mihalcea

Ο αλγόριθμος της Mihalcea [6] ξεκινάει κατασκευάζοντας ένα γράφο που έχει αρχικά ως κόμβους τις λέξεις της πρότασης που θέλουμε να αποσαφηνίσουμε. Για κάθε δυνατή έννοια των λέξεων, ο αλγόριθμος προσθέτει στο γράφο έναν κόμβο έννοιας, ενώ, παράλληλα, προσθέτει ακμές μεταξύ κόμβων εννοιών ανάμεσα στις οποίες υπάρχει κάποια εξάρτηση. Οι εξαρτήσεις αυτές εκφράζονται ως βάρη των ακμών που συνδέουν τους αντίστοιχους κόμβους εννοιών και καθορίζονται μέσω της συνάρτησης «Εξαρτήσεις» του ψευδοκώδικα που ακολουθεί, της οποίας η υλοποίηση εξαρτάται από την εφαρμογή και τον τύπο των πηγών από τις οποίες αντλούμε τις έννοιες (π.χ. ηλεκτρονικό λεξικό). Καθορίζεται, επίσης, μία δυνατή μέγιστη απόσταση (ΜεγΑποστ), η οποία εισάγει έναν περιορισμό όσον αφορά την απόσταση σε ακμές μεταξύ των κόμβων των λέξεων και των κόμβων των εννοιών που μπορούν να τους αντιστοιχισθούν.

Στη συνέχεια, σε κάθε κόμβο αντιστοιχίζεται μία τιμή, με τη χρήση ενός αλγόριθμου αξιολόγησης γράφων. Η Mihalcea χρησιμοποίησε στα πειράματά της μια παραλλαγή του αλγορίθμου PageRank [18], που είναι γνωστός κυρίως από τη χρήση του στη μηχανή αναζήτησης Google, αλλά είναι δυνατή και η χρήση άλλων αλγόριθμων αξιολόγησης.

Τέλος, για κάθε κόμβο λέξης επιλέγουμε εκείνο τον κόμβο έννοιας με την υψηλότερη τιμή του προηγούμενου βήματος.

Αλγόριθμος αποσαφήνισης λέξεων της Mihalcea
Είσοδος: Σύνολο λέξεων $\Lambda = \{\lambda_\iota \mid \iota = 1...N\}$ μιας πρότασης.

Είσοδος: Σύνολο δυνατών εννοιών των λέξεων $\mathrm{E}_{\lambda_\iota} = \left\{\varepsilon_{\lambda_\iota}^{\tau} \mid \tau = 1...\mathrm{N}_{\lambda_\iota}\right\}, \iota = 1...\mathrm{N}$, όπου $\varepsilon_{\lambda_1}^{1}$ είναι η πρώτη έννοια της πρώτης λέξης κ.ο.κ.

Έξοδος: Σύνολο εννοιών $\mathrm{E} = \left\{\varepsilon_{\lambda_\iota} \mid \iota = 1...\mathrm{N}\right\}$, όπου η έννοια ε_{λ_i} αντιστοιχεί στη λέξη λ_i του συνόλου λέξεων της εισόδου.

Κατασκευή γράφου Γ με εξαρτήσεις εννοιών
```
//Ο αλγόριθμος εξετάζει τις λέξεις ανά δύο, προκειμένου να διαπιστώσει αν
//μεταξύ των εννοιών των δύο λέξεων υπάρχουν εξαρτήσεις.
Για ι = 1 έως Ν
    Για κ = ι+1 έως Ν
        //Έλεγχος εάν η απόσταση μεταξύ κόμβου λέξης και κόμβου
        //έννοιας υπερβαίνει τη μέγιστη.
        Εάν κ-ι > ΜεγΑποστ
            έξοδος από το βρόχο

        //Εξέταση ανά δύο κάθε έννοιας της λέξης ι με κάθε έννοια
        //της λέξης κ, για εντοπισμό εξάρτησης.
        Για τ = 1 έως N_λι
            Για σ = 1 έως N_λκ
```

βάρος_ακμής = Εξαρτήσεις ($\varepsilon^{\tau}_{\lambda_\iota}$, $\varepsilon^{\sigma}_{\lambda_\iota}$, λ_ι, λ_κ)
//Εάν υπάρχει εξάρτηση μεταξύ μιας έννοιας της
//λέξης ι με κάποια έννοια της λέξης κ, η τιμή
//που επιστρέφει η συνάρτηση «Εξαρτήσεις»
//θα είναι θετική.
Εάν βάρος_ακμής > 0
Πρόσθεσε_Ακμή(Γ, $\varepsilon^{\tau}_{\lambda_\iota}$, $\varepsilon^{\sigma}_{\lambda_\iota}$, βάρος_ακμής).

Υπολογισμός τιμών των κόμβων των εννοιών στο γράφο Γ
Επανάλαβε
Για κάθε κόμβο $\mathrm{K}_\alpha \in$ Κόμβοι(Γ)

$$\tau\iota\mu\acute{\eta}\left(\mathrm{K}_\alpha\right) = \left(1-\delta\right) + \delta \sum_{\mathrm{K}_\beta \in In\left(\mathrm{K}_\alpha\right)} \frac{\beta\acute{\alpha}\rho o\varsigma\_\alpha\kappa\mu\acute{\eta}\varsigma_{\beta\alpha}}{\sum_{\mathrm{K}_\gamma \in In\left(\mathrm{K}_\beta\right)} \beta\acute{\alpha}\rho o\varsigma\_\alpha\kappa\mu\acute{\eta}\varsigma_{\beta\gamma}} \tau\iota\mu\acute{\eta}\left(\mathrm{K}_\beta\right)$$

Μέχρι να υπάρχει σύγκλιση στις τιμές $\tau\iota\mu\acute{\eta}\left(\mathrm{K}_\alpha\right)$.

Ανάθεση εννοιών στις λέξεις
Για ι = 1 έως Ν

$$\varepsilon_{\lambda_\iota} = \max\left\{\mathrm{T}\tau\iota\mu\acute{\eta}\left(\varepsilon^{\tau}_{\lambda_\iota}\right) \mid \tau = 1...\mathrm{N}_{\lambda_\iota}\right\}$$

Στον παραπάνω ψευδοκώδικα, συμβολίζουμε με $In(K)$ το σύνολο των κόμβων που διαθέτουν ακμές οι οποίες οδηγούν άμεσα στον κόμβο K. Η τιμή του δ λαμβάνει τιμές στο διάστημα [0,1]. Η Mihalcea χρησιμοποίησε στην υλοποίησή της δ = 0.85. Στα πειράματα της παρούσας εργασίας, θέσαμε δ = 0.0001, για να οδηγούμαστε γρηγορότερα σε σύγκλιση.

2.3.4 Αποσαφήνιση λέξεων με επιβλεπόμενες μεθόδους μηχανικής μάθησης

Ως παράδειγμα μεθόδου αποσαφήνισης λέξεων που χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση αναφέρουμε τη μέθοδο της εργασίας [1], η οποία χρησιμοποιεί τον απλοϊκό ταξινομητή Bayes (Naïve Bayes), πιο συγκεκριμένα την πολυμεταβλητή μορφή Bernoulli του απλοϊκού ταξινομητή Bayes [19]. Στη μέθοδο αυτή, κάθε λέξη προς αποσαφήνιση παριστάνεται με ένα διάνυσμα $\left(F_1, F_2, ..., F_n\right)$. Τα F_i είναι δυαδικά χαρακτηριστικά (features) και δείχνουν ποιες λέξεις (από ένα συγκεκριμένο λεξιλόγιο) εμφανίζονται ή όχι σε ένα παράθυρο *k* λέξεων γύρω από τη λέξη που θέλουμε να αποσαφηνίσουμε. Αν, δηλαδή, η λέξη που αντιστοιχεί στο F_i εμφανίζεται στο παράθυρο της λέξης προς αποσαφήνιση, τότε F_i = 1. Διαφορετικά, F_i = 0. Στο παράθυρο περιλαμβάνονται όλες οι λέξεις, αφού πρώτα μετατραπούν στα αντίστοιχα λήμματα (βασικές μορφές των λέξεων). Εξαιρούνται τα προθέματα, ενώ αγνοούνται τα σημεία στίξης και τα κεφαλαία γράμματα. Στα πειράματα της εργασίας [1], το παράθυρο είχε μήκος 0, 1, 2, 3, 4, 5, 10, 25 ή 50 λέξεις, με τη θέση της λέξης προς αποσαφήνιση εντός του παραθύρου να ποικίλλει (π.χ. λιγότερες λέξεις αριστερά της λέξης προς αποσαφήνιση και περισσότερες δεξιά της).

Η μέθοδος της εργασίας [1] εκπαιδεύει έναν ξεχωριστό ταξινομητή για κάθε λέξη της οποίας τις εμφανίσεις θέλουμε να μπορούμε να αποσαφηνίσουμε (π.χ. έναν ταξινομητή για τις εμφανίσεις της λέξης «interest», έναν για τις εμφανίσεις της λέξης «line», κλπ.). Ο κάθε ταξινομητής εκπαιδεύεται σε εμφανίσεις της συγκεκριμένης λέξης στις οποίες έχει επισημειωθεί χειρωνακτικά η σωστή έννοια. Κατόπιν, χρησιμοποιείται για να αποφασίσει σε ποια έννοια της λέξης αντιστοιχούν νέες εμφανίσεις της. Αντίθετα από τις μεθόδους των ενοτήτων 2.3.1, 2.3.2 και 2.3.3, η μέθοδος αυτή δεν προσπαθεί να αποσαφηνίσει όλες τις λέξεις ενός κειμένου, αλλά μόνο τις εμφανίσεις συγκεκριμένων λέξεων. Οι λέξεις δε των οποίων οι εμφανίσεις αποσαφηνίζονται πρέπει να είναι σχετικά λίγες, αφού για κάθε μία λέξη απαιτείται η εκπαίδευση ενός ξεχωριστού ταξινομητή. Επομένως, τα πειραματικά αποτελέσματα που έχουν δημοσιευθεί για τη μέθοδο της εργασίας [1] δεν είναι άμεσα συγκρίσιμα με τα αποτελέσματα των προηγούμενων μεθόδων. Το ίδιο ισχύει για αρκετές άλλες μεθόδους αποσαφήνισης που χρησιμοποιούν επιβλεπόμενη μηχανική μάθηση [20, 21].

Ενδιαφέρον παρουσιάζει η εργασία [9], στην οποία χρησιμοποιείται επίσης επιβλεπόμενη μηχανική μάθηση, αλλά κατασκευάζεται ένας μόνο ταξινομητής, ο οποίος χρησιμοποιείται στη συνέχεια για την αποσαφήνιση των εμφανίσεων όλων των λέξεων. Μεταξύ άλλων, τα διανύσματα που παριστάνουν τις εμφανίσεις των λέξεων περιλαμβάνουν ένα χαρακτηριστικό που δείχνει τη λέξη (ακριβέστερα, το λήμμα, τη βασική μορφή της λέξης) για της οποίας εμφάνιση πρόκειται. Οι συγγραφείς της εργασίας [9] πειραματίστηκαν με πολλούς διαφορετικούς αλγορίθμους μάθησης, συμπεριλαμβανομένων των C4.5, PART και k-NN. Τα καλύτερα αποτελέσματα επιτεύχθηκαν με τον αλγόριθμο εκμάθησης δέντρων απόφασης C4.5. Η μελέτη των δέντρων απόφασης που παρήχθησαν, όμως, έδειξε ότι, με την εξαίρεση των σπανιότερων λέξεων, στην πραγματικότητα είχαν παραχθεί πολλά ξεχωριστά υπο-δέντρα, ένα για τις εμφανίσεις κάθε μίας λέξεως, οπότε στην ουσία είχαν και πάλι δημιουργηθεί ξεχωριστοί ταξινομητές για τις εμφανίσεις κάθε λέξης. Αυτό δείχνει ότι τα χρήσιμα χαρακτηριστικά ενδέχεται να είναι είναι πολύ διαφορετικά μεταξύ εμφανίσεων διαφορετικών λέξεων, κάτι που συνηγορεί υπέρ του να εκπαιδεύει κανείς ξεχωριστούς ταξινομητές για τις εμφανίσεις κάθε μίας λέξεως. Δεν αναφέρουμε τα πειραματικά αποτελέσματα της εργασίας [9], γιατί προέρχονται από την αποσαφήνιση κειμένων οικονομικού μόνο περιεχομένου (πρόκειται για το υποσύνολο των εγγράφων της συλλογής Semcor που αφορούν οικονομικά θέματα).

3 Η ΜΕΘΟΔΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ

3.1 Εισαγωγή

Όπως αναφέρθηκε στο τέλος του προηγούμενου κεφαλαίου, οι περισσότερες μέθοδοι αποσαφήνισης λέξεων που χρησιμοποιούν επιβλεπόμενη μηχανική μάθηση εκπαιδεύουν έναν ξεχωριστό ταξινομητή για τις εμφανίσεις κάθε λέξεως. Αυτό οφείλεται στο ότι τα χαρακτηριστικά που είναι χρήσιμα για την αποσαφήνιση των εμφανίσεων μίας λέξεως συχνά είναι ελάχιστα χρήσιμα κατά την αποσαφήνιση των εμφανίσεων άλλης λέξεως. Για παράδειγμα, η ύπαρξη της λέξης «financial» στα συμφραζόμενα μιας εμφάνισης της λέξης «bank» είναι μια καλή ένδειξη ότι πρόκειται για εμφάνιση της «bank» με την έννοια της τράπεζας. Η ύπαρξη (ή μη) της λέξης «financial» στα συμφραζόμενα μιας εμφάνισης άλλης λέξεως, όμως, ενδέχεται να προσφέρει ελάχιστες πληροφορίες ως προς την έννοια αυτής της εμφάνισης. Εκτός αυτού, οι κατηγορίες των ταξινομητών αντιστοιχούν σε δυνατές έννοιες, και οι δυνατές έννοιες μίας λέξης δεν έχουν σχέση με τις δυνατές έννοιες μιας άλλης. Η εκπαίδευση ενός ξεχωριστού ταξινομητή για την αποσαφήνιση των εμφανίσεων κάθε λέξης κάνει δύσκολη την εφαρμογή αυτών των μεθόδων σε περιπτώσεις όπου θέλουμε να μπορούμε να αποσαφηνίζουμε όλες τις λέξεις των κειμένων και όχι μόνο τις εμφανίσεις συγκεκριμένων λέξεων.

Η μέθοδος που υλοποιήσαμε στη διάρκεια της παρούσας εργασίας χρησιμοποιεί και αυτή επιβλεπόμενη μηχανική μάθηση, αλλά με διαφορετικό τρόπο. Αντί να προσπαθεί να εκπαιδεύσει έναν ή περισσότερους ταξινομητές που να αποφασίζουν απευθείας για τις έννοιες των λέξεων, εκπαιδεύει έναν ταξινομητή που αποφασίζει πότε να εμπιστεύεται άλλες μεθόδους, που δεν χρησιμοποιούν επιβλεπόμενη μηχανική μάθηση. Πιο συγκεκριμένα, κάθε φορά που θέλουμε να αποσαφηνίσουμε τις λέξεις μιας πρότασης, καλούμε πρώτα τις μεθόδους (α) των Τσατσαρώνη κ.ά. (ενότητα 2.3.2), (β) της Mihalcea (ενότητα 2.3.3) και (γ) μια απλοϊκή μέθοδο (baseline) που επιστρέφει πάντα τη συχνότερη (σύμφωνα με το WordNet) έννοια κάθε λέξης. Κατόπιν χρησιμοποιούμε έναν ταξινομητή υψηλότερου επιπέδου, που έχει εκπαιδευθεί με επιβλεπόμενη μηχανική μάθηση, ο οποίος αποφασίζει ποια από τις τρεις μεόδους πρέπει να εμπιστευθούμε. Πρόκειται ουσιαστικά για μια μορφή «stacking» [22, 10].

Για την εκπαίδευση του ταξινομητή του ανώτερου επιπέδου, χρησιμοποιούμε Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, Support Vector Machines – SVMs) [23, 24, 25]. Για την ακρίβεια, εκπαιδεύουμε τρεις ΜΔΥ, μία για κάθε μία από τις μεθόδους (α), (β) και (γ). Κάθε ΜΔΥ μαθαίνει να διαχωρίζει τις περιπτώσεις όπου η αντίστοιχη μέθοδος κάνει λάθος ή όχι. Η συγκεκριμένη υλοποίηση ΜΔΥ που χρησιμοποιούμε, η *SVM Light* [8] [26], έχει τη δυνατότητα να επιστρέφει και ένα βαθμό βεβαιότητας (έναν πραγματικό αριθμό) για κάθε απόφασή της. Το πρόσημο του βαθμού βεβαιότητας δείχνει αν η ΜΔΥ θεωρεί ότι πρόκειται για θετική (ορθή απόφαση της αντίστοιχης μεθόδου) ή αρνητική (λανθασμένη απόφαση) περίπτωση, ενώ η απόλυτη τιμή του δείχνει χονδρικά πόσο βέβαια είναι η ΜΔΥ για την απόφασή

[8] Βλ. http://svmlight.joachims.org/.

της.[9] Η τελική απόφαση, δηλαδή το ποια από τις τρεις μεθόδους (α), (β) ή (γ) θα εμπιστευτούμε σε κάθε μία εμφάνιση λέξης, είναι συνάρτηση των τριών βαθμών βεβαιότητας που επιστρέφουν οι ΜΔΥ για τη συγκεκριμένη εμφάνιση λέξεως. Πειραματιστήκαμε με τρεις διαφορετικές συναρτήσεις αυτού του είδους, που τις ονομάζουμε «στρατηγικές». Οι στρατηγικές περιγράφονται σε επόμενες ενότητες.

3.2 Χαρακτηριστικά του ταξινομητή υψηλότερου επιπέδου

Και οι τρεις ΜΔΥ της προηγούμενης ενότητας παριστάνουν κάθε εμφάνιση λέξεως προς αποσαφήνιση με ένα διάνυσμα χαρακτηριστικών (feature vector), αλλά τα χρησιμοποιούμενα χαρακτηριστικά διαφέρουν ελαφρά μεταξύ των τριών ΜΔΥ. Τα χαρακτηριστικά που χρησιμοποιήσαμε είναι τα ακόλουθα.[10] Με λ παριστάνουμε τη λέξη της οποίας εμφάνιση παριστάνει το διάνυσμα.

1. **Part Of Speech (POS):** Το μέρος του λόγου της προς αποσαφήνιση εμφάνισης λέξης (ουσιαστικό, ρήμα κλπ.). [11] Στα δεδομένα των πειραμάτων μας, τα μέρη του λόγου έχουν σημειωθεί χειρωνακτικά. Στην πράξη, αυτή η πληροφορία θα μπορούσε να προστεθεί από ένα σύστημα επισημείωσης μερών του λόγου (part-of-speech tagger) [7, 11]. Το χαρακτηριστικό αυτό χρησιμοποιείται, επειδή ενδέχεται, για παράδειγμα, μία μέθοδος να επιτυγχάνει υψηλό ποσοστό ορθότητας σε εμφανίσεις λέξεων ενός μέρους του λόγου και χαμηλό ποσοστό ορθότητας σε εμφανίσεις λέξεων ενός άλλου μέρους του λόγου. Οπότε θα πρέπει ο ταξινομητής υψηλότερου επιπέδου να μάθει να εμπιστεύεται τη συγκεκριμένη μέθοδο στην πρώτη περίπτωση αλλά όχι στη δεύτερη. Αυτό το χαρακτηριστικό είναι το ίδιο και στις τρεις ΜΔΥ.
2. **Number of Senses (NS):** Το πλήθος των δυνατών εννοιών της λ, σύμφωνα με το χρησιμοποιούμενο λεξικό. Ενδέχεται, για παράδειγμα, μία μέθοδος να επιτυγχάνει υψηλό ποσοστό ορθότητας μόνο στην περίπτωση λέξεων με χαμηλή πολυσημία. Και αυτό το χαρακτηριστικό είναι το ίδιο και στις τρεις ΜΔΥ.
3. **Correct Word Assignment (CWA):** Το χαρακτηριστικό αυτό δείχνει πόσες εμφανίσεις της λ στα δεδομένα εκπαίδευσης αποσαφήνισε σωστά η μέθοδος (α, β, ή γ της προηγούμενης ενότητας) που αντιστοιχεί στη συγκεκριμένη ΜΔΥ.[12] Αν, για παράδειγμα, μια μέθοδος αποσαφήνισε σωστά μεγάλο αριθμό εμφανίσεων της λ στα δεδομένα εκπαίδευσης, εύλογο είναι να την εμπιστευόμαστε περισσότερο και σε νέες εμφανίσεις της λ. Το

[9] Η απόλυτη τιμή είναι ουσιαστικά η απόσταση από το υπερ-επίπεδο διαχωρισμού της ΜΔΥ, οπότε απαιτείται κανονικοποίηση προκειμένου οι βαθμοί βεβαιότητας να είναι φραγμένοι σε συγκεκριμένο διάστημα (π.χ. στο [-1, +1]). Επιστρέφουμε σε αυτό το θέμα στα επόμενα κεφάλαια.

[10] Θα ήταν καλύτερα οι τιμές των αριθμητικών χαρακτηριστικών να ήταν κανονικοποιημένες στο ίδιο διάστημα (π.χ. [-1,+1] ή [0,+1]).

[11] Στα πειράματα της εργασίας, το χαρακτηριστικό παίρνει τις τιμές 1, 2, 3 ή 4, ανάλογα με το αν η λ είναι ουσιαστικό, ρήμα, επίθετο ή επίρρημα, αντίστοιχα. Αυτή η προσέγγιση έχει το μειονέκτημα ότι κάνει τη ΜΔΥ να θεωρεί ότι π.χ. το μέρος του λόγου που αντιστοιχεί στην τιμή 1 είναι πιο όμοιο με εκείνο που αντιστοιχεί στην τιμή 2 από ό,τι με εκείνο που αντιστοιχεί στην τιμή 3, κάτι το οποίο δεν ισχύει. Θα ήταν προτιμότερο να είχαν χρησιμοποιηθεί τέσσερα ξεχωριστά δυαδικά χαρακτηριστικά, ένα για κάθε μέρος του λόγου, τα οποία να έδειχναν το καθένα αν η λ ανήκει ή όχι στο αντίστοιχο μέρος του λόγου.

[12] Αν η λ δεν εμφανίζεται καθόλου στα κείμενα εκπαίδευσης, το χαρακτηριστικό αυτό έχει τιμή 0. Θα ήταν καλύτερα το χαρακτηριστικό αυτό να δείχνει το ποσοστό (και όχι το πλήθος) εμφανίσεων της λ στα δεδομένα εκπαίδευσης που αποσαφήνισε σωστά η μέθοδος, ώστε οι τιμές του χαρακτηριστικού να μην εξαρτώνται από το μέγεθος του σώματος εκπαίδευσης.

χαρακτηριστικό αυτό είναι διαφορετικό σε κάθε ΜΔΥ, αφού αλλάζει η μέθοδος αποσαφήνισης στην οποία αντιστοιχεί η ΜΔΥ.[13]

4. **Correct POS Assignment (CPA):** Το χαρακτηριστικό αυτό δείχνει τι ποσοστό εμφανίσεων λέξεων του μέρους του λόγου της *λ* στα δεδομένα εκπαίδευσης αποσαφήνισε σωστά η μέθοδος στην οποία αντιστοιχεί η συγκεκριμένη ΜΔΥ. Και αυτό το χαρακτηριστικό είναι διαφορετικό σε κάθε ΜΔΥ, αφού αλλάζει η μέθοδος αποσαφήνισης στην οποία αντιστοιχεί η ΜΔΥ.
5. **Collocate Words (CW):** Υπάρχει η πιθανότητα να υπάρχουν συγκεκριμένα συμφραζόμενα (γειτονιές λέξεων) με τα οποία όποτε εμφανίζεται η *λ*, μία από τις τρεις μεθόδους που συνδυάζουμε να πετυχαίνει είτε σχεδόν πάντα είτε σχεδόν ποτέ να αποσαφηνίσει σωστά τη *λ*. Προκειμένου να σηματοδοτούμε το κατά πόσον η *λ* βρίσκεται σε ένα τέτοιο περιβάλλον συμφραζομένων που εγγυάται σχεδόν σίγουρη επιτυχία ή αποτυχία, χρησιμοποιούμε ένα χαρακτηριστικό, διαφορετικό για κάθε ΜΔΥ, ανάλογα με τη μέθοδο στην οποία αντιστοιχεί η ΜΔΥ, που προκύπτει ως ακολούθως. Θεωρούμε ως συμφραζόμενα κάθε εμφάνισης της *λ* το προηγούμενο και το επόμενο λήμμα.[14] Η τιμή του χαρακτηριστικού είναι το πόσες φορές στα δεδομένα εκπαίδευσης η μέθοδος της ΜΔΥ ερμήνευσε σωστά τις εμφανίσεις της *λ*, μετρώντας μόνο εμφανίσεις της *λ* που είχαν τα ίδια συμφραζόμενα με αυτά που έχει η συγκεκριμένη εμφάνιση της *λ* που έχουμε να αποσαφηνίσουμε.[15].
6. **Collocate POS (CP):** Το χαρακτηριστικό αυτό, ένα για κάθε ΜΔΥ, προκύπτει όπως ακριβώς στην προηγούμενη περίπτωση (CW), με τη διαφορά ότι αντικαθιστούμε τόσο τα συμφραζόμενα (προηγούμενο και επόμενο λήμμα) όσο και την ίδια την εμφάνιση της *λ* που αποσαφηνίζουμε με τα μέρη του λόγου τους.
7. **Word Frequency – Total (WFT):** Το χαρακτηριστικό αυτό είναι ο λόγος του πλήθους των εμφανίσεων της *λ* σε όλα τα κείμενα εκπαίδευσης προς το συνολικό πλήθος εμφανίσεων λέξεων των κειμένων εκπαίδευσης. Ενδέχεται, για παράδειγμα, μια μέθοδος να επιτυγχάνει υψηλά ποσοστά ορθότητας μόνο για συχνά εμφανιζόμενες λέξεις *λ*, για τις οποίες ενδέχεται να υπάρχουν περισσότερες ή/και ακριβέστερες πληροφορίες στο λεξικό.
8. **Word Frequency in Document (WFD):** Είναι παρόμοιο με το προηγούμενο χαρακτηριστικό, αλλά μετράμε μόνο τις εμφανίσεις της *λ* στο κείμενο από το οποίο προέρχεται η προς αποσαφήνιση εμφάνιση λέξης και διαιρούμε με το συνολικό πλήθος των εμφανίσεων λέξεων του κειμένου αυτού.

Τα διανύσματα εκπαίδευσης είναι, επίσης, σημειωμένα με την ορθή τους κατηγορία: +1 για περιπτώσεις (εμφανίσεις λέξεων) όπου η μέθοδος αποσαφήνισης

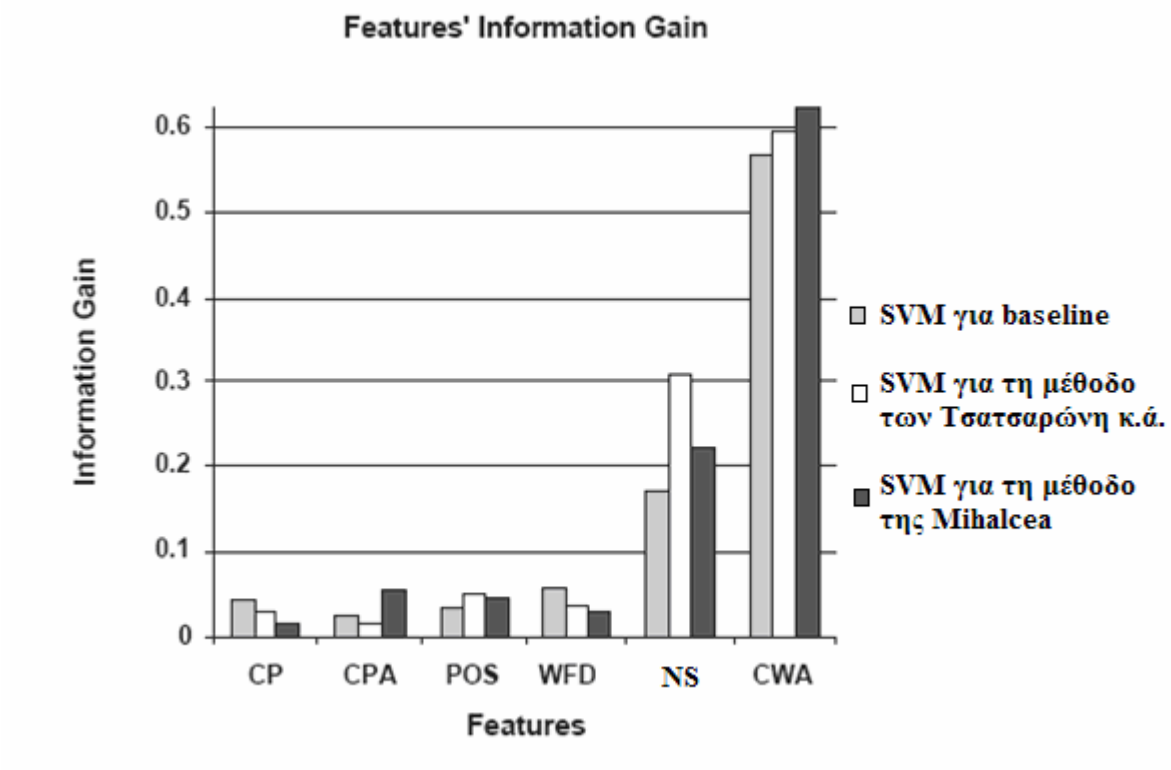
[13] Εναλλακτικά, θα μπορούσαν τα διανύσματα και των τριών ΜΔΥ να συμπεριλαμβάνουν τιμές CWA και για τις τρεις μεθόδους, κάτι που ενδεχομένως θα βελτίωνε τα αποτελέσματα, αφού θα μπορούσε, για παράδειγμα, μια ΜΔΥ να μάθει να μην εμπιστεύεται τη «δική» της μέθοδο όποτε πρόκειται για λέξη *λ* για την οποία μια άλλη μέθοδος έχει πολύ υψηλό CWA. Αντίστοιχες ενέργειες μπορούν να γίνουν και για άλλα χαρακτηριστικά.

[14] Αν η εμφάνιση της *λ* που αποσαφηνίζουμε είναι η πρώτη λέξη της πρότασης, ως προηγούμενο λήμμα χρησιμοποιούμε την ψευδο-λέξη «*start_of_text*», ενώ αν είναι η τελευταία λέξη της πρότασης, ως επόμενο λήμμα χρησιμοποιούμε την ψευδο-λέξη «*end_of_text*».

[15] Θα ήταν προτιμότερο οι τιμές και αυτού του χαρακτηριστικού να ήταν ποσοστά και όχι αριθμοί εμφανίσεων. Αν η *λ* δεν εμφανίζεται στα δεδομένα εκπαίδευσης με τα συγκεκριμένα συμφραζόμενα, η τιμή του χαρακτηριστικού ειναι μηδέν.

που αντιστοιχεί στη ΜΔΥ είχε παραγάγει τη σωστή απάντηση (έννοια) και – 1 στις άλλες περιπτώσεις (λάθος έννοια).

Η Εικόνα 2 παρουσιάζει τις μετρήσεις του πληροφοριακού κέρδους (information gain) κάθε χαρακτηριστικού για κάθε ΜΔΥ. Οι μετρήσεις έγιναν με 6-πλή διασταυρωμένη επικύρωση (6-fold cross-validation) στα δεδομένα εκπαίδευσης.[16] Τα χαρακτηριστικά που δεν φαίνονται στην εικόνα, δηλαδή τα Collocate Words (CW) και Word Frequency – Total (WFT), είχαν αμελητέο (πρακτικά μηδενικό) πληροφοριακό κέρδος και δεν χρησιμοποιήθηκαν στη συνέχεια. Όπως φαίνεται στην Εικόνα 2, τα γνωρίσματα Number of Senses (NS) και Correct Word Assignment (CWA) είναι τα κορυφαία γνωρίσματα και για τις τρεις ΜΔΥ.



Εικόνα 2 : Πληροφοριακό κέρδος των χαρακτηριστικών των τριών ΜΔΥ.

3.3 Στρατηγικές συνδυασμού των αποκρίσεων των τριών ΜΔΥ

Όπως προαναφέρθηκε, η τελική απόφαση για την έννοια που θα αποδοθεί σε κάθε εμφάνιση λέξεως προκύπτει ως συνάρτηση των βαθμών βεβαιότητας των τριών ΜΔΥ. Υπενθυμίζεται ότι το πρόσημο του κάθε βαθμού βεβαιότητας δείχνει αν η ΜΔΥ που επέστρεψε το βαθμό θεωρεί πως η μέθοδος αποσαφήνισης που της αντιστοιχεί αποσαφήνισε σωστά ή λανθασμένα τη συγκεκριμένη εμφάνιση λέξεως,

[16] Ο υπολογισμός του πληροφοριακού κέρδους έγινε χρησιμοποιώντας το Weka (βλ. http://www.cs.waikato.ac.nz/ml/weka/). Η τάξη που χρησιμοποιήθηκε είναι η InfoGainAttributeEval, με τις εξής επιλογές: -M: True, -B: False.

ενώ η απόλυτη τιμή του βαθμού δείχνει πόσο βέβαια είναι η ΜΔΥ για την απόφασή της. Πειραματιστήκαμε με τρεις διαφορετικές συναρτήσεις αυτού του είδους, τις οποίες ονομάζουμε «στρατηγικές»:

1. Στην περίπτωση αυτή ακολουθούμε την απόφαση της μεθόδου αποσαφήνισης της οποίας η ΜΔΥ παρήγαγε τον πιο θετικό (ή το λιγότερο αρνητικό, αν δεν υπάρχει θετικός) βαθμό βεβαιότητας.
2. Η δεύτερη στρατηγική είναι παρόμοια με την πρώτη. Η μόνη διαφορά είναι ότι κανονικοποιούνται πρώτα οι βαθμοί βεβαιότητας (*BB*) κάθε μίας ΜΔΥ ως ακολούθως, όπου *ελάχιστοςBB* είναι ο ελάχιστος *BB* που έχει παραγάγει η συγκεκριμένη ΜΔΥ στα δεδομένα αξιολόγησης και ομοίως *μέγιστοςBB* ο μέγιστος *BB* που έχει παραγάγει η ΜΔΥ[17].

   $$\text{ΚανονικοποιημένοςBB} = \frac{\text{BB - ελάχιστοςBB}}{\text{μέγιστοςBB - ελάχιστοςBB}}$$

3. Σύμφωνα με την τρίτη στρατηγική, αν η ΜΔΥ μιας μεθόδου προτείνει μία έννοια με θετικό βαθμό βεβαιότητας, τότε αυτή η έννοια παίρνει μία ψήφο. Αν προτείνει μία έννοια με αρνητικό βαθμό βεβαιότητας, τότε χάνει μία ψήφο. Γενικά, είναι ένας απλός μηχανισμός «ψηφοφορίας», όπου κάθε ΜΔΥ υπερψηφίζει με +1 (όταν είναι θετικός ο βαθμός βεβαιότητας) ή καταψηφίζει με -1 (όταν είναι αρνητικός ο βαθμός βεβαιότητας) μία έννοια μιας λέξης. Στο τέλος επιλέγεται η έννοια με τις περισσότερες ψήφους. Αν υπάρχει ισοψηφία, δεν επιλέγεται καμία. Για το λόγο αυτό, κρίθηκε απαραίτητο, εκτός από τον υπολογισμό της ορθότητας, να υπολογίζονται επίσης η ακρίβεια (precision) και η ανάκληση (recall).

[17] Οι τιμές που επιστρέφει η ΜΔΥ δεν είναι κανονικοποιημένες σε ένα συγκεκριμένο διάστημα (π.χ. από -1 ως 1). Για το λόγο αυτό, στη δεύτερη στρατηγική κανονικοποιούμε τις τιμές με βάση τον προαναφερθέντα τύπο. Σίγουρα δεν είναι η πιο μελετημένη κανονικοποίηση. Για παράδειγμα, η κανονικοποίηση θα μπορούσε να γίνεται εφαρμόζοντας μια σιγμοειδή συνάρτηση στις αποκρίσεις της ΜΔΥ, επιλέγοντας τις παραμέτρους της σιγμοειδούς μέσω διασταυρωμένης επικύρωσης, όπως γίνεται σε άλλες υλοποιήσεις ΜΔΥ. Παρ' όλα αυτά, η κανονικοποίηση που χρησιμοποιήσαμε φαίνεται να λειτουργεί ικανοποιητικά.

4 ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

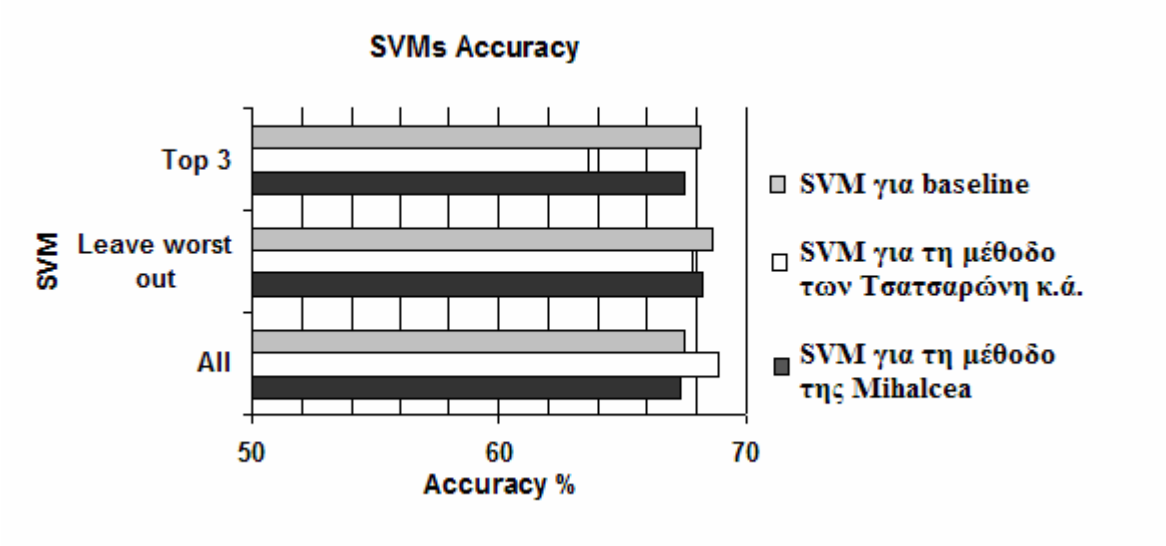
Στο κεφάλαιο αυτό θα παρουσιάσουμε τα πειραματικά αποτελέσματα της εργασίας. Τα αποτελέσματα έχουν υπολογιστεί στα σύνολα δεδομένων *Senseval 2* και *Senseval 3* με 6-πλή διασταυρωμένη επικύρωση (6-fold cross-validation). Σε όλα τα πειράματα χρησιμοποιήσαμε πολυωνυμικό πυρήνα στις ΜΔΥ, με τις προεπιλεγμένες (από την υλοποίηση της ΜΔΥ) τιμές των παραμέτρων του πολυωνυμικού πυρήνα και της παραμέτρου C.[18] (Η παράμετρος αυτή ρυθμίζει τη βαρύτητα που δίδεται στη μεγιστοποίηση του περιθωρίου της ΜΔΥ έναντι της ελαχιστοποίησης των σφαλμάτων κατά την εκπαίδευση).

Στην Εικόνα 3 παρουσιάζουμε το ποσοστό ορθότητας (accuracy) των τριών ΜΔΥ, που κρίνουν κατά πόσον οι μέθοδοι (α) των Τσατσαρώνη κ.ά., (β) της Mihalcea και (γ) η απλοϊκή (baseline) μέθοδος αναθέτουν σε κάθε εμφάνιση λέξης τη σωστή ή λανθασμένη έννοια. Τα πειράματα αναφέρονται στα δεδομένα των Senseval 2 και Senseval 3. Παρουσιάζουμε τα ποσοστά ορθότητας των τριών ΜΔΥ όταν χρησιμοποιούμε όλα τα χαρακτηριστικά (*All*), όλα εκτός από το χειρότερο (*Leave worst out*) ή τα κορυφαία τρία από πλευράς πληροφοριακού κέρδους. Σε κάθε περίπτωση εξαιρούμε τα χαρακτηριστικά που είχαν αμελητέο πληροφοριακό κέρδος, δηλαδή εξαιρούμε πάντα τα Collocate Words (CW) και Word FOC TS (βλ. ενότητα 3.2). Παρατηρούμε ότι η ΜΔΥ για τη μέθοδο των Τσατσαρώνη κ.ά. επιτυγχάνει υψηλότερο ποσοστό ορθότητας όταν χρησιμοποιούνται όλα τα χαρακτηριστικά. Αντιθέτως, οι άλλες δύο μέθοδοι επιτυγχάνουν υψηλότερο ποσοστό ορθότητας όταν χρησιμοποιούνται όλα τα χαρακτηριστικά εκτός από το χειρότερο, αλλά οι διαφορές είναι μικρές.

Στον Πίνακα 3 παρουσιάζουμε το πλήθος των θετικών και των αρνητικών παραδειγμάτων εκπαίδευσης για κάθε ΜΔΥ (μέσοι όροι για τις έξι επαναλήψεις της διασταυρωμένης επικύρωσης). Παρατηρούμε ότι δεν υπάρχει ιδιαίτερα μεγάλη ανισοκατανομή, η οποία θα μπορούσε να κάνει μια ΜΔΥ να κατατάσσει πάντα τα διανύσματα στην συχνότερη κατηγορία, με την εξαίρεση της μεθόδου Baseline, όπου τα θετικά παραδείγματα εκπαίδευσης είναι αρκετά περισσότερα από τα αρνητικά.[19]

[18] Θα ήταν καλύτερα οι τιμές των παραμέτρων να είχαν προκύψει μέσω ρύθμισης παραμέτρων (parameter tuning) στα δεδομένα εκπαίδευσης κάθε επανάληψης της διασταυρωμένης επικύρωσης.

[19] Στην περίπτωση αυτή, θα είχε ενδιαφέρον να εξεταστεί το ενδεχόμενο να δοθούν διαφορετικά βάρη στα παραδείγματα εκπαίδευσης των δύο κατηγοριών, χρησιμοποιώντας την παράμετρο *–j* του SVM^{Light}.



Εικόνα 3 : Ποσοστά ορθότητας των τριών ΜΔΥ με διαφορετικά σύνολα χαρακτηριστικών.

	Θετικά	Αρνητικά
Tsatsaronis et al. (SANs)	1727	1842
Mihalcea (PageRank)	1959	1610
Baseline (πιο συχνή έννοια)	2328	1241

Πίνακας 3 : Αρνητικά και θετικά παραδείγματα εκπαίδευσης των τριών ΜΔΥ στα δεδομένα του Senseval 2.

Στους Πίνακες 4, 5 και 6, που ακολουθούν, παραθέτουμε τα ποσοστά ορθότητας (accuracy) των τριών μεθόδων αποσαφήνισης και του συνδυασμού τους. Κάθε πίνακας αφορά τους τρεις συνδυασμούς χαρακτηριστικών που αναφέραμε προηγουμένως (τρία κορυφαία, πέντε κορυφαία, όλα) για μία συγκεκριμένη στρατηγική. Παρατηρούμε ότι με τις Στρατηγικές 1 και 2, η συνδυασμένη μέθοδος της παρούσας εργασίας επιτυγχάνει εν γένει σαφώς καλύτερα αποτελέσματα από τις αρχικές τρεις μεθόδους, με την εξαίρεση της συλλογής Senseval 3, όπου τα αποτελέσματα της συνδυασμένης μεθόδου με τη Στρατηγική 1 (και σε μικρότερο βαθμό με τη Στρατηγική 2) δεν διαφέρουν ουσιαστικά από τα αποτελέσματα της απλοϊκής μεθόδου (Baseline). Αξίζει να σημειωθεί, επίσης, ότι από μόνες τους οι μέθοδοι των Τσατσαρώνη κ.ά. και της Mihalcea δεν ξεπερνούν την απλοϊκή μέθοδο. Αντιθέτως, με τη Στρατηγική 3 η συνδυασμένη μέθοδος αποδίδει αισθητά χειρότερα, κάτι που δείχνει πως η Στρατηγική 3 χρειάζεται βελτιώσεις.

Στους Πίνακες 7, 8 και 9 παρουσιάζονται τα αντίστοιχα αποτελέσματα λαμβάνοντας υπόψη μόνο τις πολύσημες λέξεις. Βλέπουμε ότι και στην περίπτωση αυτή ο συνδυασμός μεθόδων της παρούσας εργασίας εν γένει υπερέχει, αλλά σε μικρότερο βαθμό από ό,τι στην περίπτωση όπου λαμβάνουμε υπόψη όλες τις λέξεις (πολύσημες και μη).

Στις Εικόνες 4 και 5, παραθέτουμε τα διαστήματα εμπιστοσύνης 95% των τριών βασικών μεθόδων αποσαφήνισης και του συνδυασμού τους. Κάθε εικόνα αφορά τους τρεις συνδυασμούς των χαρακτηριστικών και για τις τρεις στρατηγικές με τη διαφορά ότι στην Εικόνα 5 λάβαμε υπόψη στις μετρήσεις μόνο τις πολύσημες λέξεις. Στην Εικόνα 4, παρατηρούμε ότι οι δύο πρώτες στρατηγικές, με οποιονδήποτε συνδυασμό χαρακτηριστικών, παρουσιάζουν στατιστικά σημαντική υπεροχή έναντι της απλοϊκής μεθόδου, η οποία υπερέχει των μεθόδων Τσατσαρώνη κ.ά. και Mihalcea, ενώ η τρίτη στρατηγική δεν παρουσιάζει στατιστικά σημαντική διαφορά από τα αποτελέσματα της απλοϊκής μεθόδου. Αντίθετα, από την Εικόνα 5 προκύπτει ότι οι δύο πρώτες στρατηγικές δεν διαφέρουν στατιστικά σημαντικά από την απλοϊκή μέθοδο όταν λαμβάνονται υπόψη μόνο οι πολύσημες λέξεις.

Τέλος, στους Πίνακες 10 και 11, παραθέτουμε τα αποτελέσματα των μετρήσεων της ακρίβειας (precision) και της ανάκλησης (recall) για την τρίτη στρατηγική, για όλες τις λέξεις και μόνο για τις πολύσημες λέξεις αντίστοιχα.

Σύνολο Δεδομένων	Λέξεις		Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 1**		
	Μονό-σημες	Πολύ-σημες				Κορυφαία 3 χαρακ/κά	Κορυφαία 5 χαρακ/κά	Όλα τα χαρακ/κά
Senseval 2	464	1839	0.471	0.567	0.613	0.6799	0.6791	0.6699
Senseval 3	317	1662	0.458	0.505	0.634	0.6361	0.6392	0.6351
Senseval 2 και 3 μαζί	781	3501	0.465	0.537	0.623	0.6597	0.6607	0.6539

Πίνακας 4 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης.

Σύνολο Δεδομένων	Λέξεις		Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 2**		
	Μονό-σημες	Πολύ-σημες				Κορυφαία 3 χαρακ/κά	Κορυφαία 5 χαρακ/κά	Όλα τα χαρακ/κά
Senseval 2	464	1839	0.471	0.567	0.613	0.6834	0.6834	0.6765
Senseval 3	317	1662	0.458	0.505	0.634	0.6412	0.6437	0.6392
Senseval 2 και 3 μαζί	781	3501	0.465	0.537	0.623	0.6639	0.6651	0.6593

Πίνακας 5 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης.

Σύνολο Δεδομένων	Λέξεις		Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 3**		
	Μονό-σημες	Πολύ-σημες				Κορυφαία 3 χαρακ/κά	Κορυφαία 5 χαρακ/κά	Όλα τα χαρακ/κά
Senseval 2	464	1839	0.471	0.567	0.613	0.6180	0.6341	0.6290
Senseval 3	317	1662	0.458	0.505	0.634	0.5723	0.5831	0.6068
Senseval 2 και 3 μαζί	781	3501	0.465	0.537	0.623	0.6029	0.6187	0.6233

Πίνακας 6 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης.

Σύνολο Δεδομένων	Πολύσημες Λέξεις	Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 1**		
					Κορυφαία 3 χαρακ/κά	Κορυφαία 5 χαρακ/κά	Όλα τα χαρακ/κά
Senseval 2	1839	0.3670	0.4719	0.5758	0.5992	0.5981	0.5867
Senseval 3	1662	0.3706	0.4217	0.5734	0.5667	0.5703	0.5655
Senseval 2 και 3 μαζί	3501	0.3688	0.4482	0.5747	0.5838	0.5850	0.5767

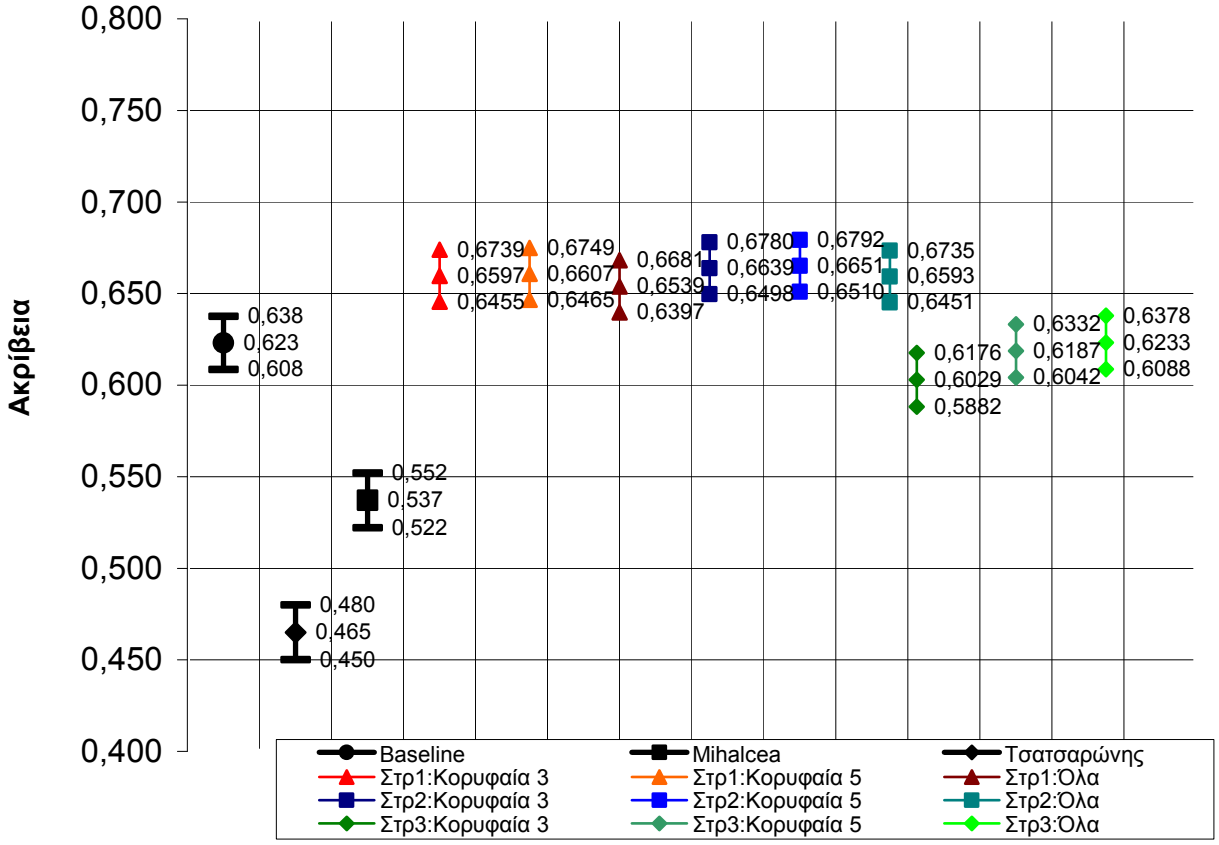
Πίνακας 7 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης (μόνο πολύσημες λέξεις).

Σύνολο Δεδομένων	Πολύσημες Λέξεις	Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 2**		
					Κορυφαία 3 γνωρίσματα	Κορυφαία 5 γνωρίσματα	Όλα τα γνωρίσματα
Senseval 2	1839	0.3670	0.4719	0.5758	0.6035	0.6035	0.5948
Senseval 3	1662	0.3706	0.4217	0.5734	0.5728	0.5758	0.5703
Senseval 2 και 3 μαζί	3501	0.3688	0.4482	0.5747	0.5890	0.5904	0.5833

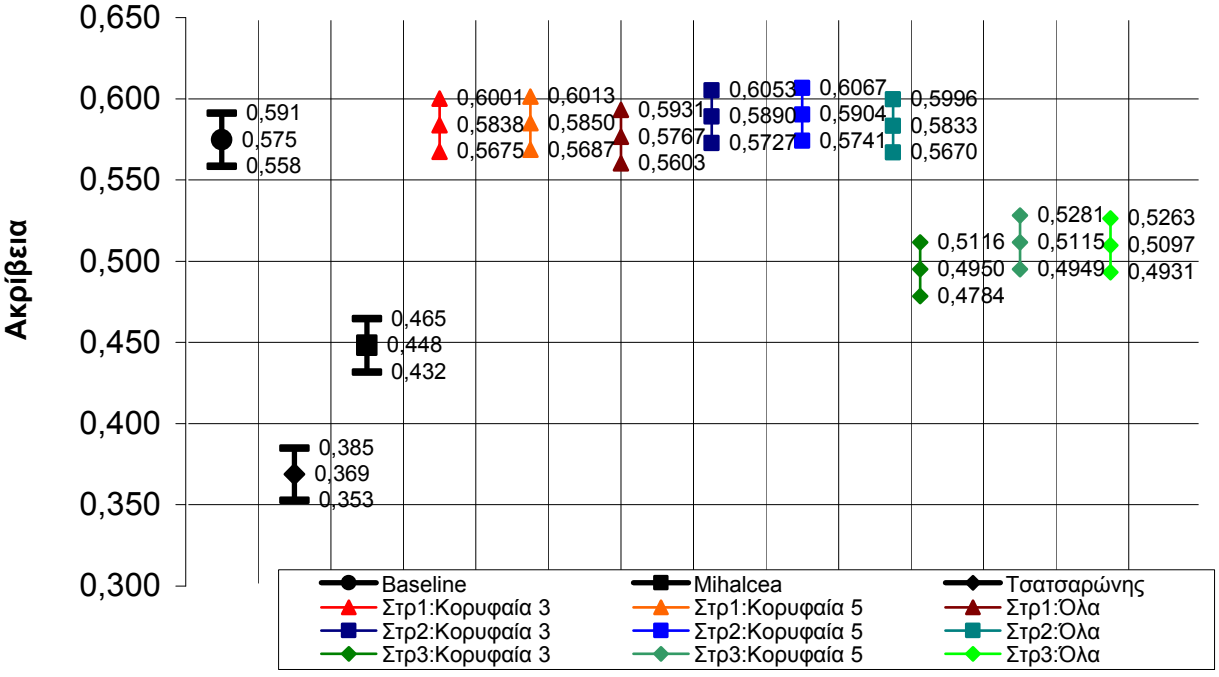
Πίνακας 8 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης (μόνο πολύσημες λέξεις).

Σύνολο Δεδομένων	Πολύσημες Λέξεις	Τσατσαρώνη κ.ά. (SANs)	Mihalcea (PageRank)	Baseline (πιο συχνή έννοια)	**Στρατηγική 3**		
					Κορυφαία 3 γνωρίσματα	Κορυφαία 5 γνωρίσματα	Όλα τα γνωρίσματα
Senseval 2	1839	0.3670	0.4719	0.5758	0.5097	0.5299	0.5215
Senseval 3	1662	0.3706	0.4217	0.5734	0.4781	0.4905	0.4963
Senseval 2 και 3 μαζί	3501	0.3688	0.4482	0.5747	0.4950	0.5115	0.5097

Πίνακας 9 : Ποσοστά ορθότητας (accuracy) των μεθόδων αποσαφήνισης (μόνο πολύσημες λέξεις).



Εικόνα 4 : Διαστήματα εμπιστοσύνης 95%.



Εικόνα 5 : Διαστήματα εμπιστοσύνης 95% (μόνο πολύσημες λέξεις).

Σύνολο Δεδομένων	Κορυφαία 3 γνωρίσματα		Κορυφαία 5 γνωρίσματα		Όλα τα γνωρίσματα	
	Ακρίβεια	Ανάκληση	Ακρίβεια	Ανάκληση	Ακρίβεια	Ανάκληση
Senseval 2	0.6842	0.5636	0.7051	0.5762	0.7110	0.5640
Senseval 3	0.6555	0.5078	0.6739	0.5138	0.6785	0.5184
Σύνολο	0.6714	0.5378	0.6912	0.5474	0.6963	0.5430

Πίνακας 10 : Ακρίβεια (precision) και ανάκληση (recall) τρίτης στρατηγικής.

Σύνολο Δεδομένων	Κορυφαία 3 γνωρίσματα		Κορυφαία 5 γνωρίσματα		Όλα τα γνωρίσματα	
	Ακρίβεια	Ανάκληση	Ακρίβεια	Ανάκληση	Ακρίβεια	Ανάκληση
Senseval 2	0.5819	0.4535	0.6086	0.4692	0.6126	0.4540
Senseval 3	0.5657	0.4139	0.5872	0.4211	0.5933	0.4265
Σύνολο	0.5746	0.4347	0.5989	0.4464	0.6036	0.4410

Πίνακας 11 : Ακρίβεια (precision) και ανάκληση (recall) τρίτης στρατηγικής (μόνο πολύσημες λέξεις).

5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΗ ΜΕΛΕΤΗ

5.1 Συμπεράσματα

Ξεκινώντας από τρεις μεθόδους που δεν χρησιμοποιούν μηχανική μάθηση (μια μέθοδο που χρησιμοποιεί Δίκτυα Διάδοσης Ενεργοποίησης, μια μέθοδο που χρησιμοποιεί τον αλγόριθμο PageRank και μια απλοϊκή μέθοδο που επιστρέφει πάντα τη συχνότερη έννοια κάθε λέξης), αναπτύξαμε μία συνδυασμένη μέθοδο αποσαφήνισης λέξεων, η οποία χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση για να μάθει πότε να εμπιστεύεται κάθε μία από τις αρχικές τρεις μεθόδους. Πιο συγκεκριμένα, εκπαιδεύσαμε τρεις Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ), μία για κάθε μία από τις αρχικές μεθόδους. Κάθε ΜΔΥ μαθαίνει να προβλέπει πότε η αντίστοιχη αρχική μέθοδος εντοπίζει τη σωστή ή λανθασμένη έννοια. Η τελική απόκριση (έννοια) της συνδυασμένης μεθόδου προκύπτει ως συνάρτηση των βαθμών βεβαιότητας που επιστρέφουν οι τρεις ΜΔΥ. Τα καλύτερα αποτελέσματα της συνδυασμένης μεθόδου προκύπτουν όταν ακολουθούμε την απόκριση της αρχικής μεθόδου για την οποία η αντίστοιχη ΜΔΥ είναι περισσότερο βέβαια πως πρόκειται για ορθή απόκριση, αφού οι βαθμοί βεβαιότητας της κάθε ΜΔΥ κανονικοποιηθούν λαμβάνοντας υπόψη το πραγματικό εύρος τους. Η συνδυασμένη μέθοδος αξιολογήθηκε στα σύνολα δεδομένων *Senseval 2* και *Senseval 3*, επιτυγχάνοντας εν γένει καλύτερα αποτελέσματα από τις τρεις αρχικές μεθόδους. Η δεύτερη καλύτερη μέθοδος ήταν η απλοϊκή, που επιστρέφει πάντα τη συχνότερη έννοια, ενώ οι υπόλοιπες δύο αρχικές μέθοδοι από μόνες τους ήταν αισθητά χειρότερες.

5.2 Προτάσεις μελλοντικής μελέτης

Τα πειράματα της παρούσας εργασίας εκτελέσθηκαν με ΜΔΥ πολυωνυμικού πυρήνα. Θα είχε ενδιαφέρον να εξετασθεί η χρήση άλλων πυρήνων στις ΜΔΥ (π.χ. Radial Basis Function), που οδηγούν συχνά σε καλύτερα αποτελέσματα. Επίσης, θα είχε ενδιαφέρον να εξετασθεί κατά πόσον τα αποτελέσματα βελτιώνονται με προσεκτική ρύθμιση (tuning) των παραμέτρων των ΜΔΥ και των πυρήνων τους. Ακόμη, θα είχε ενδιαφέρον να προστεθούν περισσότερα χαρακτηριστικά στις ΜΔΥ και να βελτιωθούν τα υπάρχοντα. Μερικές ιδέες για πρόσθετα χαρακτηριστικά και βελτιώσεις των υπαρχόντων αναφέρθηκαν στην ενότητα 3.2.

Τα πειράματα της εργασίας έδειξαν, επίσης, τη σημαντική επίδραση της στρατηγικής συνδυασμού των βαθμών βεβαιότητας των ΜΔΥ, καθώς και την ανάγκη διερεύνησης παραλλαγών των στρατηγικών που χρησιμοποιήσαμε. Τέλος, θα ήταν σκόπιμο να αξιολογηθεί η συνδυασμένη μέθοδος και σε άλλες συλλογές δεδομένων, ιδιαίτερα τη Semcor 2, που είναι πολύ μεγαλύτερη από τις συλλογές που χρησιμοποιήσαμε (βλ. ενότητα 2.2), ώστε τα αποτελέσματα να είναι πιο αξιόπιστα.

ΑΝΑΦΟΡΕΣ

[1] T. Pedersen, *"A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation"*. Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 63-69, Seattle, Washington, 2000.

[2] R. Florian, S. Cucerzan, C. Schafer and D. Yarowsky, *"Combining classifiers for word sense disambiguation"*. Natural Language Engineering, 8(4):327–341, 2002.

[3] A. Montoyo, A. Suarez, G. Rigau and M. Palomar, *"Combining knowledge- and corpus-based word sense disambiguation methods"*. Journal of Artificial Intelligence Research, 23:299–330, 2005.

[4] G. Tsatsaronis, M. Vazirgiannis and I. Androutsopoulos, *"Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri"*. Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1725-1730, Hyderabad, India, 2007.

[5] M. Barthélemy, E. Chow and T. Eliassi-Rad, *"Knowledge Representation Issues in Semantic Graphs for Relationship Detection"*. Proceedings of the Spring AAAI Symposium, AAAI Press, pp. 91-98, 2005.

[6] R. Mihalcea, *"Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling"*. Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 411 – 418, Vancouver, British Columbia, Canada, 2005.

[7] Ν. Τσίχλας, *«Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με Τεχνικές Μηχανικής Μάθησης»*. Μεταπτυχιακή διπλωματική εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2003.
http://www.aueb.gr/users/ion/docs/tsichlas_final_report.pdf

[8] J. Preiss, *"A detailed comparison of WSD systems: an analysis of the system answers for the Senseval-2 English all words task"*. Natural Language Engineering, 12(3): 209 – 228, 2006.

[9] G. Paliouras, V. Karkaletsis, I. Androutsopoulos, and C. D. Spyropoulos, *"Learning Rules for Large-Vocabulary Word Sense Disambiguation: a comparison of various classifiers"*. Proceedings of the 2nd International Conference on Natural Language Processing, pp. 383 – 394, Patras, Greece, 2000.

[10] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, P. Stamatopoulos, *"Stacking classifiers for anti-spam filtering of e-mail"*. Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 44-50, Carnegie Mellon University, Pittsburgh, USA, 2001.

[11] Π. Μαλακασιώτης, *«Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές ενεργητικής μάθησης»*. Μεταπτυχιακή διπλωματική εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών, 2005.
http://www.aueb.gr/users/ion/docs/malakasiotis_final_msc_report.pdf

[12] C. J.C. Burges, *"A Tutorial on Support Vector Machines for Pattern Recognition"*. Proceedings of the 4th Conference on Data Mining and Knowledge Discovery, pp. 121-167, New York, 1998.

[13] P. Tarau, R. Mihalcea and E. Figa, *"Semantic Document Engineering with WordNet and PageRank"*. Symposium on Applied Computing, Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 782-786, 2005.

[14] F. Crestani, *"Application of Spreading Activation Techniques in Information Retrieval"*. Artificial Intelligence Review, 11(6):453–482, 1997.

[15] N. M. Ide and J. Veronis, *"Word sense disambiguation: the state of the art"*. Computational Linguistics, 24(1):1–40, 1998.

[16] C. Fellbaum, *WordNet – an electronic lexical database*. MIT Press, 1998.

[17] J. Veronis and N. M. Ide, *"Word sense disambiguation with very large neural networks extracted from machine readable dictionaries"*. International Conference On Computational Linguistics, Proceedings of 13th conference on Computational Linguistics (COLING 1990), pp. 389–394, Helsinki, Finland, 1990.

[18] R. Mihalcea, P. Tarau and E. Figa, *"PageRank on semantic networks, with application to word sense disambiguation"*. International Conference On Computational Linguistics, Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Article No. 1126, Geneva, Switzerland, 2004.

[19] V. Metsis, I. Androutsopoulos and G. Paliouras, *"Spam Filtering with Naïve Bayes – Which Naive Bayes?"*. Proceedings of the 3rd Conference on E-mail and Anti-Spam (CEAS 2006), Mountain View, CA, USA, 2006.

[20] D. Yarowski, *"Word-sense disambiguation using statistical models of roget's categories trained on large corpora"*. Proceedings of the 14th International Conference on Compuitational Linguistics, pp. 454–460, Nantes, France, 1992.

[21] C. Leacock, G. A. Miller and M. Chodorow, *"Using corpus statistics and wordnet relations for sense identification"*. Computational Linguistics, 24(1):147–165, 1998.

[22] D. Wolpert, *"Stacked Generalization"*. Neural Networks, 5(2):241–260, 1992.

[23] C. Cortes and V. P. Vapnik, *"Support-vector networks"*, Machine Learning, 20(3):273-297, 1995.

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

[25] V. P. Vapnik, *"Statistical Learning Theory"*, John Wiley and Sons, Inc., New York, 1998.

[26] T. Joachims, *"Making large-scale SVM learning practical. Advances in Kernel methods - support vector learning"*. B. Scholkopf, C. Burges and A. Smola (ed.), MIT Press, 1999.