Κεφάλαιο 2

Ανίχνευση Ορίων Περιόδων

2.1 Εισαγωγή

Η συντριπτική πλειοψηφία των εφαρμογών επεξεργασίας κειμένου απαιτεί την αναγνώριση των ορίων των περιόδων που συνθέτουν το κείμενο, για να μπορέσει να προχωρήσει στα περαιτέρω στάδια ανάλυσης. Σύμφωνα με το συντακτικό της Νέας Ελληνικής [107] ο ορισμός της περιόδου έχει ως εξής:

> Ένας λόγος ολοκληρωμένος που αποτελείται από μία ή περισσότερες προτάσεις και καταλήγει, όταν είναι γραπτός σε τελεία ή βρίσκεται ανάμεσα σε δύο τελείες ονομάζεται **περίοδος**[1].

Βέβαια, εκτός από την τελεία και άλλα σημεία στίξης, όπως το ερωτηματικό, το θαυμαστικό και τα αποσιωπητικά, μπορεί να δηλώνουν το τέλος μιας περιόδου. Το πρόβλημα της αυτόματης ανίχνευσης ορίων περιόδων έχει να κάνει με την αμφισημία αυτών των σημείων στίξης.

[1] Στην Αγγλική βιβλιογραφία, η περίοδος αναφέρεται ως sentence ενώ οι προτάσεις (κύριες ή δευτερεύουσες) από τις οποίες αποτελείται καλούνται clauses.

Για παράδειγμα, εκτός από το τέλος μιας περιόδου, μία τελεία μπορεί να βρεθεί σε μία συντομογραφία (π.χ. *δρχ., Η.Π.Α.*) ή σε έναν δεκαδικό αριθμό (π.χ. *.02*). Επιπλέον, το είδος της αμφισημίας διαφέρει μεταξύ των σημείων στίξης που μπορεί να υποδηλώνουν το τέλος μιας περιόδου. Πιο κάτω φαίνονται μερικές χαρακτηριστικές περιπτώσεις:

Ο κ. Χ. Παπαδόπουλος μίλησε για το επενδυτικό πρόγραμμα του ΟΣΕ.

Οι χώρες της Κ.Α.Ε. θα προσθέσουν 100 εκατ. καταναλωτές στην Ε.Ε.

Και αυτοί που έχουν αναγκαστικά ελεύθερο χρόνο γιατί... δεν έχουν δουλειά;

«Και τα κανάλια;» θα ρωτήσετε.

Ο όμιλος (και όχι μόνο!) έχει πάρει άδεια από τον ΟΛΠ...

Η αμφισημία των σημείων στίξης ποικίλει ανάλογα με τον τύπο κειμένου ή/και με το σώμα κειμένων. Το 47% από τις τελείες που υπάρχουν στο σώμα κειμένων *Wall Street Journal*[1] δηλώνει συντομογραφίες ενώ το αντίστοιχο ποσοστό για το σώμα κειμένων *Brown*[2] είναι μόλις 10% [24]. Αυτό σημαίνει ότι αν δεν υπήρχε κανένα σύστημα αποσαφήνισης των σημείων στίξης και αν θεωρήσουμε οποιαδήποτε άλλη ασάφεια ασήμαντη, θα μπορούσε να ανιχνευτεί σωστά το 53% των περιόδων του σώματος *Wall Street Journal* και το 90% του σώματος *Brown*. Η απόδοση ενός συστήματος ανίχνευσης ορίων περιόδων, λοιπόν, πρέπει να είναι κατά πολύ μεγαλύτερη από αυτό το ποσοστό.

Ωστόσο, η έρευνα στην ανίχνευση ορίων περιόδων δεν είναι ανάλογη με αυτή άλλων περιοχών της επεξεργασίας κειμένου (π.χ. μορφολογική ανάλυση, συντακτική ανάλυση) με αποτέλεσμα το πρόβλημα να αντιμετωπίζεται συχνά με ανεπαρκή και επιφανειακό τρόπο. Οι τεχνικές που χρησιμοποιούνται (συνήθως απλές γραμματικές και εκτεταμένες λίστες συντομογραφιών) στοχεύουν στην επίλυση των πιο κοινών περιπτώσεων χωρίς να εμβαθύνουν στο πρόβλημα. Στην πλειοψηφία τους αυτές οι απλές τεχνικές είναι προσανατολισμένες σε ένα συγκεκριμένο τύπο κειμένου ή σε μία συγκεκριμένη φυσική γλώσσα και είναι δύσκολο, αν όχι αδύνατο, να προσαρμοστούν σε κάποιο άλλο τύπο ή σε μία άλλη γλώσσα χωρίς να πρέπει να σχεδιαστούν από την αρχή.

[1] Σώμα κειμένων αποτελούμενο από άρθρα της εφημερίδας *Wall Street Journal*.
[2] Σώμα κειμένων της Αγγλικής γλώσσας αποτελούμενο από μεγάλη ποικιλία ειδών κειμένων.

Ακόμη, εφόσον η ανίχνευση ορίων περιόδων είναι απλά ένα στάδιο της προεπεξεργασίας κειμένου, δεν πρέπει να απαιτεί υπέρογκους πόρους και σημαντικό υπολογιστικό κόστος. Έτσι η ανάπτυξη ενός συστήματος αναγνώρισης περιόδων που θα βασίζεται σε τεράστιες λίστες συντομογραφιών ή σε εξειδικευμένα λεξικά, με πληροφορία που δεν χρησιμοποιείται από τα επόμενα στάδια της επεξεργασίας, δεν είναι ασφαλώς η βέλτιστη λύση.

Η δική μας πρόταση για την ανίχνευση ορίων περιόδων σε κείμενα της Νέας Ελληνικής γλώσσας βασίζεται σε:

- **Απλές παραμέτρους** (π.χ. μήκος λέξεων) που δεν απαιτούν ιδιαίτερο υπολογιστικό κόστος για τον υπολογισμό τους. Έτσι, μεγάλος όγκος κειμένων μπορεί να αναλυθεί πολύ σύντομα.
- **Κανόνες αποσαφήνισης** των σημείων στίξης που **εξάγονται αυτόματα** από ένα σώμα εκπαίδευσης. Έτσι, η μέθοδός μας μπορεί να εφαρμοστεί, μετά από εκπαίδευση, σε οποιοδήποτε τύπο κειμένου ή σε κάποια άλλη γλώσσα με ιδιότητες παρόμοιες με τα Νέα Ελληνικά.

Στο επόμενο τμήμα παρουσιάζεται συνοπτικά η σχετική έρευνα ενώ στο τμήμα 2.3 περιγράφεται αναλυτικά η μέθοδός μας. Στο τμήμα 2.4 περιλαμβάνεται η αξιολόγηση της προτεινόμενης μεθόδου και στο τμήμα 2.5 η περίληψη του κεφαλαίου καθώς και τα συμπεράσματα που αποκομίστηκαν.

2.2 Σχετική Έρευνα

Οι δημοσιευμένες εργασίες πάνω στην αναγνώριση ορίων περιόδων είναι ελάχιστες σε σύγκριση με άλλα στάδια της ανάλυσης (π.χ. συντακτική ανάλυση). Πολλοί ερευνητές αναφέρουν ότι ένας ανιχνευτής ορίων περιόδων περιλαμβάνεται στο σύστημά τους, αλλά δεν δίνουν πληροφορίες για τη σχεδίασή του και πολύ περισσότερο για την απόδοσή του [98]. Η πιο κοινή προσέγγιση είναι η χρήση απλών κανονικών γραμματικών (regular grammars) που προσπαθούν να αναγνωρίσουν είτε ακολουθίες χαρακτήρων (π.χ., τελεία-κενό-κεφαλαίο γράμμα) είτε ολόκληρες λέξεις και βασίζεται σε εκτεταμένες λίστες εξαιρέσεως (exception lists) για τον εντοπισμό των συντομογραφιών. Ως εναλλακτική λύση, ο Muller [67] προτείνει μία απλή

μορφολογική ανάλυση αντί για λίστες εξαιρέσεως, με στόχο το φιλτράρισμα των λέξεων με κοινές καταλήξεις που δεν είναι πιθανό να είναι συντομογραφίες. Τέτοιες προσεγγίσεις απαιτούν αρκετές ανθρωποώρες για την κατασκευή και την ανανέωση των λιστών εξαιρέσεων και των κανόνων της γραμματικής. Επιπλέον, είναι προσανατολισμένες σε ένα συγκεκριμένο τύπο κειμένου και πρέπει να φτιαχτούν από την αρχή σε περίπτωση που το σύστημα πρέπει να προσαρμοστεί σε κάποιο άλλο τύπο κειμένου.

Η πιο ακριβής προσέγγιση που έχει αναφερθεί [77] βασίζεται σε σώμα εκπαίδευσης 25 εκατομμυρίων λέξεων. Η ακρίβεια του για το σώμα *Brown* είναι της τάξης του 99,8%. Για κάθε λέξη του λεξικού, το μοντέλο αυτό απαιτεί τον υπολογισμό των πιθανοτήτων να είναι είτε η πρώτη είτε η τελευταία λέξη μιας περιόδου. Να σημειωθεί ότι αυτή η πληροφορία δεν είναι χρήσιμη σε επόμενα στάδια της ανάλυσης, όπως η μορφολογική ή η συντακτική επεξεργασία. Επιπλέον, δεν είναι σχεδιαστικά σωστό ένα στάδιο της προεπεξεργασίας κειμένου όπως η αναγνώριση περιόδων να απαιτεί υπέρογκο υπολογιστικό κόστος. Γι' αυτό το λόγο οι Reynar και Ratnaparkhi [75] προτείνουν ένα εκπαιδεύσιμο μοντέλο βασιζόμενο στην μέγιστη εντροπία το οποίο δεν απαιτεί υπολογιστικά περίπλοκη πληροφορία. Η πληροφορία που χρησιμοποιεί βασίζεται στο δείγμα (token) που περιλαμβάνει το σημείο στίξης που είναι υποψήφιο τέλος περιόδου, καθώς και τα δείγματα αμέσως πριν και μετά από αυτό. Για την κατασκευή μιας λίστας συντομογραφιών από το σώμα εκπαίδευσης χρησιμοποιείται ένας σχετικά απλός αλγόριθμος. Η προσέγγιση της μέγιστης εντροπίας επιτυγχάνει ακρίβεια της τάξης του 97,7% για το σώμα *Brown* με χρήση μιας χειροποίητης λίστας συντομογραφιών, προσφωνήσεων και εταιρικών αρχικών. Η ακρίβεια για το ίδιο σώμα κειμένων χωρίς τη χρήση αυτής της επιπλέον πληροφορίας είναι της τάξης του 97,5%.

Μια πρόσφατη εργασία, το σύστημα *SATZ* [71], φαίνεται να είναι μία πολύ εύρωστη προσέγγιση. Χρησιμοποιεί ένα νευρωνικό δίκτυο για την αποσαφήνιση των ορίων περιόδων και βασίζεται στις πιθανότητες πρωτεύοντος μέρους-του-λόγου (prior part-of-speech). Αυτές οι πιθανότητες αφορούν την ταξινόμηση των διάφορων μερών-του-λόγου στα οποία μπορεί να ανήκει μία λέξη (π.χ. η λέξη *μετρήσεις* μπορεί να είναι είτε ρήμα είτε ουσιαστικό αλλά είναι πιο πιθανό να είναι ουσιαστικό). Το σύστημα αυτό χρησιμοποιεί ένα λεξικό 30.000 λέξεων και πληροφορία που έχει να κάνει με

ένα περιβάλλον έξι δειγμάτων, δηλαδή τρία πριν το υποψήφιο τέλος περιόδου και τρία μετά από αυτό και επιτυγχάνει 98,5% ακρίβεια σε ένα σώμα από άρθρα της εφημερίδας *Wall Street Journal*. Επίσης, το *SATZ* μπορεί να εκπαιδευτεί και για άλλους τύπους κειμένου ή φυσικές γλώσσες. Το πρόβλημα είναι ότι για πολλούς τύπους κειμένου ή και για ολόκληρες γλώσσες ακόμα δεν υπάρχουν λεξικά που να περιέχουν πληροφορία σχετική με το πρωτεύον μέρος-του-λόγου ή σώματα κειμένων κατάλληλα σχολιασμένα ώστε να μπορεί να κατασκευαστεί αυτόματα ένα τέτοιο λεξικό. Ακόμα και στην περίπτωση όπου είναι διαθέσιμο ένα μορφολογικό λεξικό, η μετατροπή του σε ένα λεξικό πρωτεύοντος μέρος-του-λόγου (δηλ. ένα λεξικό που θα κατατάσσει τα πιθανά μέρη-του-λόγου μιας λέξης ανάλογα με τη συχνότητά τους) δεν είναι εύκολη αφού αυτή η πληροφορία συχνά σχετίζεται με τον τύπο κειμένου. Παρ' όλα αυτά, υποστηρίζεται ότι η απόδοση του συστήματος δεν επηρεάζεται σε μεγάλο βαθμό από τη μείωση του μεγέθους του λεξικού. Επιπλέον, αν και αυτή η πληροφορία μπορεί να είναι χρήσιμη σε ορισμένα στάδια περαιτέρω επεξεργασίας, ασφαλώς υπάρχουν εφαρμογές όπου δεν βρίσκει εφαρμογή (π.χ. αντιστοίχηση περιόδων (sentence alignment)).

Επίσης, πρέπει να αναφερθεί ότι τόσο το σύστημα *SATZ* όσο και η προσέγγιση μέγιστης εντροπίας δεν κάνουν καμία διάκριση μεταξύ των διάφορων σημείων στίξης που μπορεί να δηλώνουν το τέλος της περιόδου και εφαρμόζουν τους ίδιους κανόνες παντού. Όμως, είναι σαφές ότι υπάρχουν σημαντικές διαφορές όσον αφορά την ασάφεια μεταξύ αυτών των σημείων στίξης. Για παράδειγμα, μία τελεία μπορεί να δηλώνει μία συντομογραφία ενώ ένα θαυμαστικό ή ένα ερωτηματικό όχι. Επομένως, είναι προφανές ότι για να επιτευχθεί η καλύτερη απόδοση, κάθε σημείο στίξης πρέπει να έχει δικούς του κανόνες αποσαφήνισης, που θα εκμεταλλεύονται στο μέγιστο βαθμό τα ιδιαίτερα χαρακτηριστικά του.

2.3 Μεθοδολογία

Πριν από την ανίχνευση των ορίων περιόδων σε ένα κείμενο, είναι απαραίτητος ο τεμαχισμός του κειμένου σε δείγματα (tokens), δηλ. αλφαριθμητικά που περιέχουν λέξεις, σημεία στίξης, αριθμούς κτλ. Σύμφωνα με την προσέγγισή μας η διαδικασία δειγματοποίησης πραγματοποιείται από τον ακόλουθο απλό κανόνα: δύο δείγματα

χωρίζονται από μία ακολουθία κενών. Ένα δείγμα που τελειώνει με ένα από τα ακόλουθα σημεία στίξης θεωρείται ως πιθανό τέλος περιόδου:

Τίτλος	Σημείο στίξης
τελεία	.
θαυμαστικό	!
ερωτηματικό	;
αποσιωπητικά	...

Επίσης, πιθανό τέλος περιόδου θεωρείται ένα δείγμα όταν μετά από ένα από τα παραπάνω σημεία στίξης περιέχει και μία ακολουθία άλλων σημείων στίξης όπως τα *),],},»* κτλ. (π.χ. το δείγμα *τέλειος!»*). Να σημειωθεί ότι στην Νέα Ελληνική γλώσσα υπάρχουν μερικές περιπτώσεις όπου και η διπλή τελεία (:) μπορεί να δηλώνει τέλος περιόδου, σε αυτή την εργασία, όμως, θεωρούμε αυτές τις περιπτώσεις ως αμελητέες.

2.3.1 Απλές παράμετροι

Η επιλογή των παραμέτρων που οδηγούν στην αναγνώριση των ορίων περιόδων είναι εμπειρική και τείνει να εκμεταλλευτεί στο έπακρο τα χαρακτηριστικά της Νέας Ελληνικής γλώσσας. Πιο συγκεκριμένα:

- Παρατηρήθηκε ότι η συντριπτική πλειοψηφία των λέξεων των Νέων Ελληνικών (γύρω στο 98%) τελειώνουν σε συγκεκριμένους χαρακτήρες: τα φωνήεντα το τελικό σίγμα (ς) και το νι (ν). Παρόμοιοι κανόνες διέπουν και άλλες φυσικές γλώσσες όπως τα Ιταλικά και τα Ισπανικά. Έτσι, σε πολλές περιπτώσεις είναι δυνατό να βρεθούν οι λέξεις που είναι πιθανόν να είναι συντομογραφίες απλά εξετάζοντας το τελευταίο τους γράμμα.
- Επιπλέον, παρατηρήθηκε ότι το μήκος (σε χαρακτήρες) της τελευταίας λέξης μιας περιόδου δεν είναι πιθανό να είναι πολύ μικρό (δηλ. 1 ή 2 χαρακτήρες), σε αντίθεση με την πρώτη λέξη μιας περιόδου.
- Τέλος, τα σημεία στίξης που πιθανόν να υπάρχουν ακριβώς δίπλα σε ένα υποψήφιο τέλος περιόδου είναι πολύ διαφωτιστικά και σε πάρα πολλές περιπτώσεις αρκούν για να επιλύσουν από μόνα τους την ασάφεια.

Τύπος σημείου στίξης	**Παραδείγματα**
Αρχικό	*(, [, «,* κτλ.
Τελικό	*),], »*, κτλ.
Κανένα	
Τύπος απλού χαρακτήρα	**Παραδείγματα**
Μικρός τελικός	*α, ε, η,* κτλ.
Κεφαλαίος τελικός	*Α, Ε, Η,* κτλ.
Μικρός μη-τελικός	*μ, κ, λ,* κτλ.
Κεφαλαίος μη-τελικός	*Μ, Κ, Λ,* κτλ.
Ψηφίο	*0, 1, 2,* κτλ.
Ειδικός	*%, #, $,* κτλ.
Άλλος	*a, b, c,* κτλ.

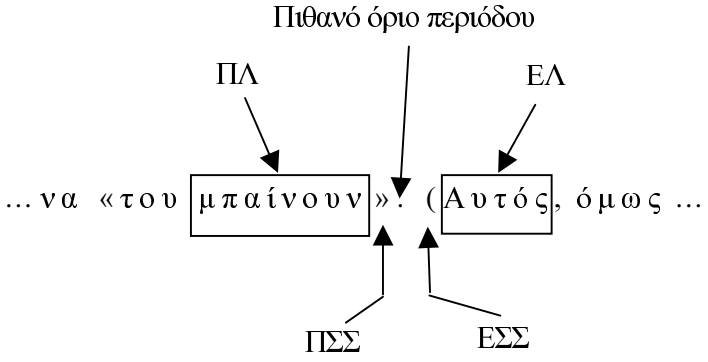
Πίνακας 2.1. Οι τύποι των χαρακτήρων.

Πριν προχωρήσουμε στην παράθεση των παραμέτρων που χρησιμοποιήθηκαν είναι χρήσιμο να ορίσουμε τον τύπο σημείου στίξης και τον τύπο χαρακτήρα όπως φαίνεται στον πίνακα 2.1. Να σημειωθεί ότι οι χαρακτήρες που συναντιούνται συχνά στο τέλος των λέξεων (π.χ. *α, ε, η, ς, ν,* κ.ά.) καλούνται *τελικοί* ενώ οι υπόλοιποι χαρακτήρες καλούνται *μη-τελικοί*. Επίσης, ο τύπος *ειδικός* καλύπτει συγκεκριμένους χαρακτήρες που συχνά βρίσκονται στο τέλος των δειγμάτων ενώ ο τύπος *άλλος* καλύπτει όλους τους χαρακτήρες που δεν καλύπτονται από τους προηγούμενους τύπους (π.χ. οι χαρακτήρες του λατινικού αλφάβητου).

Η πληροφορία στην οποία βασίστηκε η διαδικασία αποσαφήνισης περιέχεται σε δύο δείγματα: το δείγμα που περιλαμβάνει το υποψήφιο τέλος περιόδου (το οποίο καλούμε *προηγούμενο-δείγμα*) και το ακριβώς επόμενο (που καλείται *επόμενο-δείγμα*). Αυτά τα δύο δείγματα περιέχουν τα εξής χαρακτηριστικά:

- **Προηγούμενη Λέξη (ΠΛ):** το αλφαριθμητικό που απομένει μετά την απομάκρυνση όλων των σημείων στίξης από την αρχή και το τέλος *του προηγούμενου-δείγματος*.
- **Επόμενη Λέξη (ΕΛ):** το αλφαριθμητικό που απομένει μετά την απομάκρυνση όλων των σημείων στίξης από την αρχή και το τέλος του *επόμενου-δείγματος*.
- **Προηγούμενα Σημεία Στίξης (ΠΣΣ):** μία ακολουθία από σημεία στίξης που μπορεί να βρίσκονται μεταξύ του υποψήφιου τέλους περιόδου και της *προηγούμενης λέξης*.

- **Επόμενα Σημεία Στίξης (ΕΣΣ):** μία ακολουθία από σημεία στίξης που μπορεί να βρίσκονται μεταξύ του υποψήφιου τέλους περιόδου και της *επόμενης λέξης*.



Σχήμα 2.1. Ένα παράδειγμα των χαρακτηριστικών που χρησιμοποιούμε.

Ένα παράδειγμα αυτών των χαρακτηριστικών δίνεται στο σχήμα 2.1. Τελικά, η χρήσιμη πληροφορία που εξάγεται από αυτά τα χαρακτηριστικά αποτελείται από τις ακόλουθες παραμέτρους:

ΠΛ: μήκος (σε χαρακτήρες), τύπος πρώτου χαρακτήρα, τύπος τελευταίου χαρακτήρα, αν περιέχει ή όχι τελεία.

ΕΛ: μήκος (σε χαρακτήρες), τύπος πρώτου χαρακτήρα, τύπος τελευταίου χαρακτήρα, αν περιέχει ή όχι τελεία.

ΠΣΣ: τύπος τελευταίου σημείου στίξης

ΕΣΣ: τύπος πρώτου σημείου στίξης

ΠΛ: μήκος = *8*
τύπος πρώτου χαρακτήρα = *μικρός μη-τελικός*
τύπος τελευταίου χαρακτήρα = *μικρός τελικός*
περιέχεται τελεία = *όχι*
ΕΛ: μήκος = *5*
τύπος πρώτου χαρακτήρα = *κεφαλαίος τελικός*
τύπος τελευταίου χαρακτήρα = *μικρός τελικός*
περιέχεται τελεία = *όχι*
ΠΣΣ: τύπος τελευταίου σημείου στίξης = *τελικός*
ΕΣΣ: τύπος πρώτου σημείου στίξης = *αρχικός*

Σχήμα 2.2. Ένα παράδειγμα των παραμέτρων.

Οι αντίστοιχες παράμετροι για το παράδειγμα του σχήματος 2.1 φαίνονται στο σχήμα 2.2.

2.3.2 Εκμάθηση Βάσει Μετασχηματισμών

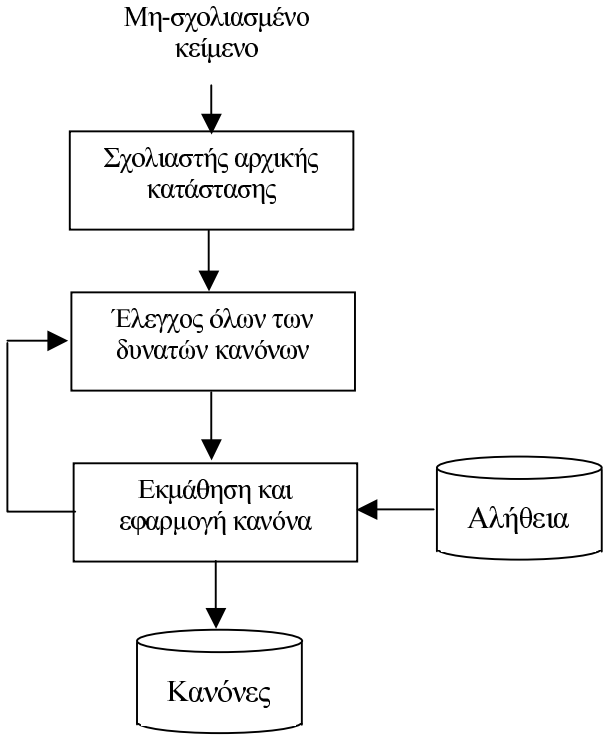
Η διαδικασία αποσαφήνισης των ορίων περιόδου και η αυτόματη εξαγωγή των κανόνων που θα περιγραφούν στα επόμενα τμήματα είναι μία παραλλαγή της θεωρίας *Εκμάθησης Βάσει Μετασχηματισμών* (EBM) (transformation-based learning) [13]. Η θεωρία EBM είναι ανεξάρτητη εφαρμογής και γλώσσας και έχει εφαρμοστεί σε ευρύ φάσμα εφαρμογών, όπως σχολιασμός μέρους-του-λόγου (part-of-speech tagging) [13], αναγνώριση ορίων φράσεων (text chunking) [73] και σχολιασμός πράξεων διαλόγου (dialog act tagging) [79], επιτυγχάνοντας πολύ αξιόλογα αποτελέσματα. Παρ' όλα αυτά, τα ιδιαίτερα χαρακτηριστικά μιας εφαρμογής, εφόσον ληφθούν υπ' όψιν, μπορούν να οδηγήσουν στη βέλτιστη απόδοση και αυτό επιχειρήθηκε με την παρούσα εργασία. Πριν προχωρήσουμε λοιπόν στην αναλυτική περιγραφή της μεθόδου μας, κρίνουμε σκόπιμο να περιγράψουμε εν συντομία την EBM, για να γίνουν πιο κατανοητές οι διαφορές και οι ομοιότητες μεταξύ των δύο μεθόδων.

Η EBM εξάγει αυτόματα γλωσσολογική γνώση σε μορφή κανόνων χρησιμοποιώντας σώματα κειμένων που είναι ήδη σχολιασμένα (annotated corpora). Αρχικά, το σώμα εκπαίδευσης σχολιάζεται μέσω ενός *σχολιαστή αρχικής κατάστασης* (initial-state annotator) και το σώμα που προκύπτει συγκρίνεται με την *αλήθεια* (truth), δηλ. το σώμα κειμένων που έχει σχολιαστεί χειρονακτικά. Στην συνέχεια, εξάγεται μία διατεταγμένη λίστα μετασχηματισμών ακολουθώντας την εξής διαδικασία: Κάθε δυνατός κανόνας της ακόλουθης μορφής εφαρμόζεται στο σχολιασμένο σώμα:

ΑΝ *περιβάλλον δράσης* ΤΟΤΕ *μετασχηματισμός*

όπου ο *μετασχηματισμός* αλλάζει την κατάσταση μιας ετικέτας (tag) εφόσον ισχύει η συνθήκη που περιγράφεται από το *περιβάλλον δράσης*. Επιπλέον, μέσω μιας αντικειμενικής συνάρτησης υπολογίζεται ο βαθμός στον οποίον το σώμα που προκύπτει μετά την εφαρμογή του κανόνα μοιάζει με την *αλήθεια*. Ο κανόνας με τη μεγαλύτερη τιμή αυτής της συνάρτησης επιλέγεται ως ο καλύτερος και εφαρμόζεται στο σχολιασμένο σώμα, με αποτέλεσμα να παράγει ένα νέο σώμα. Η εκμάθηση συνεχίζεται με την εφαρμογή όλων των δυνατών κανόνων σε αυτό το νέο σχολιασμένο σώμα. Έτσι, κάθε κανόνας που εξάγεται βελτιώνει την ακρίβεια του σχολιασμένου σώματος. Η εκμάθηση τελειώνει όταν κανένας κανόνας δεν βελτιώνει

την ακρίβεια πέρα από κάποιο προκαθορισμένο κατώφλι. Η διαδικασία εκμάθησης φαίνεται στο σχήμα 2.3.



Σχήμα 2.3. Η διαδικασία εκμάθησης σύμφωνα με την θεωρία EBM.

Η εφαρμογή της γνώσης, που εξήχθηκε με τη μορφή κανόνων, σε ένα νέο κείμενο (εκτός του σώματος εκπαίδευσης) γίνεται ως εξής: το κείμενο αρχικά σχολιάζεται μέσω του σχολιαστή αρχικής κατάστασης και στην συνέχεια εφαρμόζεται η διατεταγμένη λίστα των μετασχηματισμών. Πρέπει να σημειωθεί ότι ένας κανόνας εφαρμόζεται σε ολόκληρο το κείμενο πριν εφαρμοστεί ο επόμενος κανόνας.

Η EBM είναι μία θεωρία ανεξάρτητη-εφαρμογής. Για να προσαρμοστεί σε μία συγκεκριμένη εφαρμογή πρέπει να καθοριστούν τα ακόλουθα:

- Ο σχολιαστής αρχικής κατάστασης
- Το φάσμα των επιτρεπόμενων μετασχηματισμών
- Η αντικειμενική συνάρτηση για τη σύγκριση του σώματος με την αλήθεια.

Ο ορισμός κάθε δυνατού μετασχηματισμού είναι μία επίπονη διαδικασία. Γι' αυτό το λόγο, συνήθως χρησιμοποιούνται αλγόριθμοι καθοδηγημένοι από τα δεδομένα (data-

driven), με στόχο τον αποκλεισμό περιπτώσεων που δεν είναι πιθανό να εντοπιστούν σε ένα κείμενο. Επίσης, η ΕΒΜ θεωρητικά είναι ανεξάρτητη της πολυπλοκότητας και της ακρίβειας του σχολιαστή αρχικής κατάστασης. Παρ' όλα αυτά, όσο πιο ακριβής είναι ο σχολιαστής αρχικής κατάστασης τόσο μικρότερο είναι το χρονικό κόστος εκπαίδευσης καθώς εξάγονται λιγότεροι κανόνες.

2.3.3 Διαδικασία αποσαφήνισης

Σύμφωνα με τη μέθοδό μας, η αναγνώριση των ορίων περιόδου ενός κειμένου μπορεί να διακριθεί σε τρία στάδια. Αρχικά, όλα τα υποψήφια όρια περιόδου θεωρείται ότι δηλώνουν τέλος περιόδου. Η ακρίβεια που επιτυγχάνεται από αυτή την αρχικοποίηση είναι το **κάτω όριο** της απόδοσης του συστήματος, δηλ. τα ακόλουθα στάδια πρέπει να επιτύχουν ακρίβεια σημαντικά βελτιωμένη σε σχέση με αυτό το όριο. Στην συνέχεια εφαρμόζεται στο κείμενο ένα σύνολο από κανόνες της ακόλουθης μορφής:

Σύνολο Κανόνων 1: ΑΝ *περιβάλλον*
ΤΟΤΕ *απομάκρυνε το όριο περιόδου*

όπου το *περιβάλλον* είναι η συνθήκη που ενεργοποιεί τον κανόνα και αποτελείται είτε από το συνδυασμό ΠΛ-ΕΛ είτε από το συνδυασμό ΠΣΣ-ΕΣΣ του υποψήφιου τέλους περιόδου. Ένα παράδειγμα περιβάλλοντος που αναφέρεται στην περίπτωση των σχημάτων 2.1 και 2.2 φαίνεται στο σχήμα 2.4.

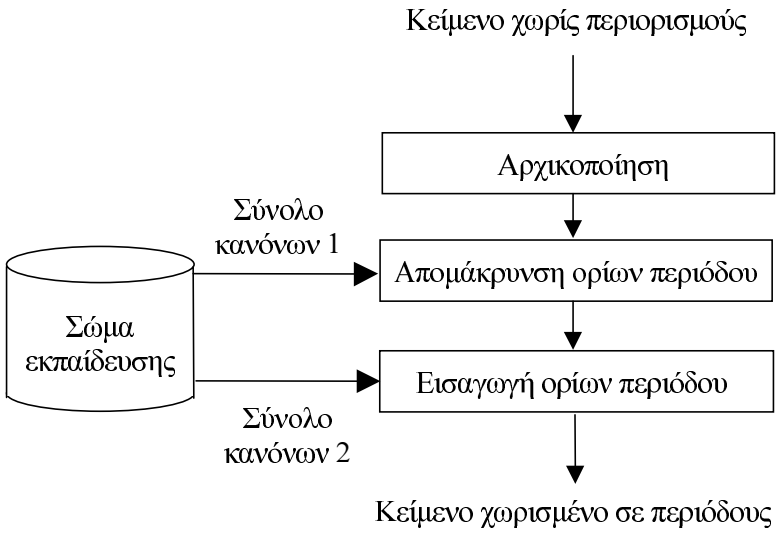
((ΠΛ: μήκος = *8*
τύπος πρώτου χαρακτήρα = *μικρός μη-τελικός*
τύπος τελευταίου χαρακτήρα = *μικρός τελικός*
περιέχεται τελεία = *όχι*)
ΚΑΙ
(ΕΛ: μήκος = *5*
τύπος πρώτου χαρακτήρα = *κεφαλαίος τελικός*
τύπος τελευταίου χαρακτήρα = *μικρός τελικός*
περιέχεται τελεία = *όχι*))
Ή
((ΠΣΣ: τύπος τελευταίου σημείου στίξης = *τελικός*)
ΚΑΙ
(ΕΣΣ: τύπος πρώτου σημείου στίξης = *αρχικός*))

Σχήμα 2.4. Ένα παράδειγμα περιβάλλοντος.

Εφόσον εφαρμοστούν όλοι οι κανόνες του συνόλου κανόνων 1, εφαρμόζεται ένα δεύτερο σύνολο κανόνων της ακόλουθης μορφής:

Σύνολο Κανόνων 2: ΑΝ *περιβάλλον*
ΤΟΤΕ *εισήγαγε όριο περιόδου*

όπου το *περιβάλλον* είναι η συνθήκη που ενεργοποιεί τον κανόνα και ορίζεται όπως πιο πάνω. Η διαδικασία αποσαφήνισης φαίνεται στο σχήμα 2.5. Παρατηρούμε ότι η διάταξη στην εφαρμογή των κανόνων είναι διαφορετική από την EBM. Σύμφωνα με την EBM, η διάταξη εφαρμογής των κανόνων έχει να κάνει με το κατά πόσο βελτιώνουν την απόδοση του μοντέλου. Αντίθετα, σύμφωνα με την προτεινόμενη μέθοδο, η διάταξη εφαρμογής των κανόνων έχει να κάνει με το μετασχηματισμό που επιφέρουν. Έτσι, πρώτα εφαρμόζονται όλοι οι κανόνες που μετασχηματίζουν ένα όριο περιόδου σε ένα απλό σημείο στίξης και στην συνέχεια όλοι οι κανόνες που μετασχηματίζουν ένα απλό σημείο στίξης σε όριο περιόδου.



Σχήμα 2.5. Δομή της διαδικασίας αποσαφήνισης.

2.3.4 Αυτόματη εξαγωγή κανόνων

Τα σύνολα κανόνων 1 και 2 που περιγράφηκαν στο προηγούμενο τμήμα εξάγονται αυτόματα βάσει του σώματος εκπαίδευσης. Πιο συγκεκριμένα, για κάθε πιθανό συνδυασμό περιβάλλοντος και σημείου στίξης που μπορεί να δηλώνει τέλος περιόδου εξάγεται από το σώμα εκπαίδευσης η ακόλουθη πληροφορία (συνάρτηση Prolog):

rule(ΣΗΜΕΙΟ_ΣΤΙΞΗΣ, ΠΕΡΙΒΑΛΛΟΝ, N1, N2)

όπου

- Το *ΣΗΜΕΙΟ_ΣΤΙΞΗΣ* είναι ένα από τα τέσσερα πιθανά σημεία στίξης που θεωρούμε ότι δηλώνουν τέλος περιόδου (δηλ. τελεία, θαυμαστικό, ερωτηματικό και αποσιωπητικά).
- Το *ΠΕΡΙΒΑΛΛΟΝ* είναι είτε ένας συνδυασμός ΠΛ-ΕΛ είτε ένας συνδυασμός ΠΣΣ-ΕΣΣ.
- Το *N1* είναι ένας ακέραιος που δείχνει πόσες φορές ανιχνεύτηκε στο σώμα εκπαίδευσης το συγκεκριμένο *ΣΗΜΕΙΟ_ΣΤΙΞΗΣ* στο συγκεκριμένο *ΠΕΡΙΒΑΛΛΟΝ* να μη δηλώνει τέλος περιόδου.
- Το *N2* είναι ένας ακέραιος που δείχνει πόσες φορές ανιχνεύτηκε στο σώμα εκπαίδευσης το συγκεκριμένο *ΣΗΜΕΙΟ_ΣΤΙΞΗΣ* στο συγκεκριμένο *ΠΕΡΙΒΑΛΛΟΝ* να δηλώνει τέλος περιόδου.

Στην συνέχεια, για να συμπεριληφθεί κάποιος από αυτούς τους κανόνες στο σύνολο κανόνων 1 ενός συγκεκριμένου σημείου στίξης πρέπει να συμφωνεί με το ακόλουθο κριτήριο:

$$N1 > N2$$

ΚΑΙ

$$N2 < ΣΥΟΠ * Επιείκεια$$

όπου το *ΣΥΟΠ* είναι το Σύνολο Υποψηφίων Ορίων Περιόδων σε ολόκληρο το σώμα εκπαίδευσης και η *Επιείκεια* είναι ένας αριθμός μεταξύ 0 και 1, που υποδεικνύει το βαθμό αξιοπιστίας των παραγόμενων κανόνων. Όσο μικρότερη είναι η *Επιείκεια* τόσο πιο ακριβές είναι το σύνολο των κανόνων. Αντίθετα, όσο πιο υψηλή είναι η *Επιείκεια* τόσο πιο μεγάλο γίνεται το παραγόμενο σύνολο των κανόνων. Στα πειράματα που περιγράφονται στο τμήμα 2.4 η *Επιείκεια* τέθηκε ίση με 0.01 (ή αλλιώς 99% αξιοπιστία). Το αντίστοιχο κριτήριο για το σύνολο κανόνων 2 είναι:

$$N1=0$$

ΚΑΙ

$$N2>0$$

Η εξαγωγή αυτών των κριτηρίων έγινε εμπειρικά. Πρέπει να σημειωθεί ότι αρχικά τα κριτήρια για τα δύο σύνολα κανόνων ήταν συμμετρικά. Από τα πειράματα όμως φάνηκε ότι το κριτήριο για το δεύτερο σύνολο κανόνων έπρεπε να είναι σαφώς πιο δύσκολο για να επιτευχθεί πολύ υψηλή ακρίβεια.

2.4 Αξιολόγηση

Ένα σύστημα προσδιορισμού των ορίων περιόδων μπορεί να παράγει δύο ειδών σφάλματα [71]:

- **Θετικό σφάλμα:** προκύπτει σε περίπτωση που ένα σημείο στίξης θεωρηθεί λανθασμένα ως τέλος περιόδου.
- **Αρνητικό σφάλμα:** προκύπτει σε περίπτωση που ένα σημείο στίξης που δηλώνει τέλος περιόδου δεν αναγνωριστεί από το σύστημα.

Πριν περάσουμε στα αναλυτικά αποτελέσματα είναι απαραίτητο να περιγραφούν τα σώματα κειμένων πάνω στα οποία εφαρμόστηκε η μέθοδός μας.

2.4.1 Σώματα κειμένων

Για να ελεγχθεί η απόδοση του ανιχνευτή ορίων περιόδων κατασκευάσαμε δύο σώματα κειμένων:

- Ένα σώμα κειμένων 366.294 λέξεων που αποτελείται από κείμενα της εβδομαδιαίας εφημερίδας *Το Βήμα*. Τα κείμενα που περιλαμβάνονται ανήκουν σε όλους σχεδόν τους τύπους κειμένου που μπορούν να βρεθούν σε μία εφημερίδα [11], όπως ρεπορτάζ, άρθρα, επιστολές αναγνωστών, αθλητικά ρεπορτάζ, συνεντεύξεις κ.ά. Το σώμα αυτό κατασκευάστηκε με κείμενα που βρέθηκαν στην ηλεκτρονική έκδοση της εφημερίδας στο Διαδίκτυο[1].
- Ένα σώμα 81.212 λέξεων αποτελούμενο από αποφάσεις του Ανωτάτου Δικαστηρίου (*Άρειος Πάγος* και *Συμβούλιο της Επικρατείας*). Το σώμα αυτό κατασκευάστηκε με κείμενα που βρέθηκαν στο Διαδίκτυο[2].

[1] http://tovima.dolnet.gr
[2] http://senanet.com/cgi-bin/dsearch/form

Όπως γίνεται αντιληπτό το πρώτο σώμα αποτελείται από ένα μωσαϊκό τύπων κειμένων ενώ το δεύτερο είναι πιο ομοιογενές. Έτσι, μπορούν να εξαχθούν συμπεράσματα για τη συμπεριφορά της προτεινόμενης μεθοδολογίας σε συνθήκες ομοιογενούς ή όχι σώματος εκπαίδευσης. Και τα δύο αυτά σώματα περιλαμβάνουν κείμενα που ήταν ήδη σε ηλεκτρονική μορφή. Ακόμη, δεν έγινε καμία προεπεξεργασία εκτός από την αφαίρεση κάποιων τίτλων (χειρονακτικά). Στη συνέχεια, το κάθε σώμα χωρίστηκε σε σώμα εκπαίδευσης και σώμα ελέγχου, όπως φαίνεται στον πίνακα 2.2. Να σημειωθεί ότι το **κάτω όριο** είναι η ακρίβεια που προκύπτει αν θεωρήσουμε ότι όλες οι τελείες και όλα τα θαυμαστικά, τα ερωτηματικά και τα αποσιωπητικά δηλώνουν τέλος περιόδου. Όπως φαίνεται το σώμα των αποφάσεων δικαστηρίου είναι σημαντικά πιο ασαφές από αυτό του *Βήματος*.

Πηγή	***Το Βήμα***	**Αποφάσεις ανωτάτου δικαστηρίου**
Λέξεις	366.294	81.212
Περίοδοι	16.010	2.024
Υποψήφια όρια περιόδου	20.113	3.407
Κάτω όριο (%)	79,6	59,4
Σώμα εκπαίδευσης (λέξεις)	165.465	39.335
Σώμα ελέγχου (λέξεις)	200.829	41.877

Πίνακας 2.2. Τα σώματα κειμένων.

2.4.2 Απόδοση

Ο ανιχνευτής ορίων περιόδων εκπαιδεύτηκε με βάση το σώμα εκπαίδευσης του *Βήματος* και στην συνέχεια ελέγχθηκε με βάση το αντίστοιχο σώμα ελέγχου. Η ίδια διαδικασία ακολουθήθηκε για την περίπτωση του σώματος των αποφάσεων δικαστηρίου. Τα αποτελέσματα δίνονται στον πίνακα 2.3. Η απόδοση του συστήματος είναι σαφώς καλύτερη στην περίπτωση του σώματος του *Βήματος*. Πρέπει πάντως να επισημανθεί η διαφορά στο κάτω όριο μεταξύ των δύο σωμάτων.

Ο πίνακας 2.4 δείχνει αναλυτικά αποτελέσματα για κάθε σημείο στίξης καθώς και τον αριθμό των παραγόμενων κανόνων για την περίπτωση του σώματος του *Βήματος*. Επίσης, η χρησιμότητα καθεμίας από τις προτεινόμενες παραμέτρους φαίνεται στον πίνακα 2.5. Πιο συγκεκριμένα, αυτά τα αποτελέσματα προέκυψαν αγνοώντας μία

παράμετρο κάθε φορά και υπολογίζοντας την νέα απόδοση του συστήματος, χωρίς τη μετεκπαίδευσή του. Η τελευταία στήλη αυτού του πίνακα, υπό τον τίτλο *Απόκλιση,* δείχνει την απόκλιση της ακρίβειας του συστήματος όταν αγνοείται η αντίστοιχη παράμετρος από την αρχική ακρίβεια (όταν λαμβάνονται υπ' όψιν όλες οι παράμετροι) που είναι ίση με 10.911. Παρατηρούμε ότι οι πιο σημαντικές παράμετροι είναι ο τύπος του τελευταίου χαρακτήρα της προηγούμενης και της επόμενης λέξης ενώ η πληροφορία αν η επόμενη λέξη περιέχει ή όχι τελεία δεν συνεισφέρει καθόλου στην ακρίβεια του συστήματος.

Σώμα εκπαίδευσης	***Το Βήμα***	**Αποφάσεις ανωτάτου δικαστηρίου**
Λέξεις	200.829	41.877
Περίοδοι	8.736	1.033
Υποψήφια όρια περιόδου	10.977	1.747
Κάτω όριο (%)	79,6	59,1
Θετικά σφάλματα	40	22
Αρνητικά σφάλματα	26	4
Ακρίβεια (%)	99,4	98,5

Πίνακας 2.3. Η απόδοση του ανιχνευτή ορίων περιόδων.

Σημείο στίξης	**Αριθμός κανόνων**	**Σωστά**	**Θετικά σφάλματα**	**Αρνητικά σφάλματα**	**Κάτω όριο (%)**	**Ακρίβεια (%)**
Τελεία	190	9.796	29	17	78,8	99,5
Θαυμαστικό	32	270	0	3	94,9	98,9
Ερωτηματικό	46	522	0	5	95,6	99,1
Αποσιωπητικά	44	323	11	1	51,1	96,4
Σύνολο	312	10.911	40	26	79,6	99,4

Πίνακας 2.4. Αναλυτικά αποτελέσματα για το σώμα κειμένων του *Βήματος*.

Αγνοούμενη παράμετρος	**Σωστά**	**Θετικά σφάλματα**	**Αρνητικά σφάλματα**	**Απόκλιση (%)**
ΠΛ: μήκος	9.702	1.266	9	0,11
ΠΛ: τύπος πρώτου χαρακτήρα	9.697	1.273	7	0,11
ΠΛ: τύπος τελευταίου χαρακτήρα	9.382	1.586	9	0,14
ΠΛ: περιέχει τελεία	10.622	329	26	0,03
ΕΛ: μήκος	10.907	53	17	0,01
ΕΛ: τύπος πρώτου χαρακτήρα	10.496	459	22	0,04
ΕΛ: τύπος τελευταίου χαρακτήρα	10.101	865	11	0,07
ΕΛ: περιέχει τελεία	10.911	40	26	0,00
ΠΣΣ: τύπος τελευταίου χαρακτήρα	8.915	2.061	1	0,18
ΕΣΣ: τύπος επόμενου χαρακτήρα	10.735	16	226	0,02

Πίνακας 2.5. Ανάλυση της χρησιμότητας των προτεινόμενων παραμέτρων.

Σχετικά με το χρονικό κόστος, πρέπει να αναφερθεί ότι η ανάλυση ολόκληρου του σώματος ελέγχου του *Βήματος* (200.829 λέξεις) διάρκεσε 25,71 δευτερόλεπτα, δηλ. αναλύθηκαν 7.811 λέξεις ανά δευτερόλεπτο. Ο υπολογιστής που χρησιμοποιήθηκε ήταν ένας Pentium στα 133 MHz.

2.4.3 Σύγκριση με τη θεωρία EBM

Όπως αναφέρθηκε στην παράγραφο 2.3.2 η μέθοδος αυτόματης εκμάθησης των κανόνων και η διαδικασία αποσαφήνισης είναι παραλλαγή της θεωρίας EBM. Για να μπορέσουμε να συγκρίνουμε την αποτελεσματικότητα των δύο μεθόδων εφαρμόσαμε τη θεωρία EBM στο πρόβλημα ανίχνευσης ορίων περιόδων. Όπως αναφέρθηκε και στο τμήμα 2.3.2 για να εφαρμοστεί η EBM σε μία συγκεκριμένη εφαρμογή πρέπει να οριστούν τα ακόλουθα [13]:

- *Ένας σχολιαστής αρχικής κατάστασης:* στην περίπτωσή της ανίχνευσης ορίων περιόδου θεωρήσαμε ότι όλα τα υποψήφια όρια περιόδου είναι αληθή (απόδοση κάτω ορίου).
- *Το φάσμα των επιτρεπόμενων μετασχηματισμών:* για την ανίχνευση ορίων περιόδων υπάρχουν μόνο δύο δυνατοί μετασχηματισμοί

  Μετασχηματισμός 1: απλό σημείο στίξης → όριο περιόδου

  Μετασχηματισμός 2: όριο περιόδου → απλό σημείο στίξης
- *Μία αντικειμενική συνάρτηση για την εξαγωγή του πιο κατάλληλου κανόνα:* ως τέτοια συνάρτηση ορίστηκε ο υπολογισμός των σφαλμάτων, δηλ. το άθροισμα των θετικών και των αρνητικών σφαλμάτων. Ο κανόνας της μορφής «ΑΝ *περιβάλλον δράσης* ΤΟΤΕ *μετασχηματισμός*» που ελαχιστοποιεί το σύνολο των σφαλμάτων επιλέγεται ως ο καλύτερος.

Η EBM εφαρμόστηκε στο πρόβλημα ανίχνευσης ορίων περιόδου στο σώμα του *Βήματος* για την περίπτωση των τελειών που είναι η πιο πολυπληθής κατηγορία. Συγκριτικά αποτελέσματα σε σχέση με τη δική μας μέθοδο δίνονται στον πίνακα 2.6. Όπως φαίνεται, η μέθοδός μας πέτυχε πιο ακριβή αποτελέσματα. Πιστεύουμε ότι αυτό συνέβη γιατί η μέθοδός μας βασίστηκε στα ιδιαίτερα χαρακτηριστικά του προβλήματος (μη-επικαλυπτόμενο περιβάλλον δράσης, μικρός αριθμός

μετασχηματισμών κ.ά.) για να επιτύχει την υψηλότερη δυνατή ακρίβεια. Άλλωστε, η θεωρία EBM ουσιαστικά πραγματοποιεί μία αναζήτηση σύμφωνα με τον αλγόριθμο *αναρρίχησης λόφου* (hill climbing) ο οποίος δεν εγγυάται ότι θα βρεθεί η βέλτιστη λύση.

Αλγόριθμος εκμάθησης	**Συνολικές περιπτώσεις**	**Θετικά σφάλματα**	**Αρνητικά σφάλματα**	**Ακρίβεια (%)**
EBM	9.842	389	24	95,8
Ο δικός μας	9.842	29	17	99,5

Πίνακας 2.6. Σύγκριση των δύο αλγορίθμων εκμάθησης.

Σχετικά με το χρονικό κόστος της εκπαίδευσης του συστήματός μας, αν *t* είναι ο χρόνος που απαιτείται από τη θεωρία EBM για να εξαχθούν *n* κανόνες, τότε η μέθοδός μας απαιτεί προσεγγιστικά:

$$\textit{κόστος εκπαίδευσης} = t / (n-1)$$

αφού όλοι οι κανόνες εξάγονται συγκρίνοντας μόνο μία φορά το σώμα εκπαίδευσης με την *αλήθεια* (δηλ. το σωστά χωρισμένο κείμενο). Να σημειωθεί ότι η EBM εξάγει ένα κανόνα κάθε φορά που συγκρίνει το σώμα εκπαίδευσης με την *αλήθεια*.

2.5 Περίληψη - Συμπεράσματα

Σε αυτό το κεφάλαιο παρουσιάσαμε έναν ανιχνευτή ορίων περιόδων ικανό να χειριστεί κείμενα της Νέας Ελληνικής γλώσσας. Σε αντίθεση με τις σύγχρονες μεθόδους, το σύστημα αυτό βασίζεται σε απλή πληροφορία που δεν χρειάζεται ιδιαίτερο υπολογιστικό κόστος και είναι εύκολα διαθέσιμη. Δεν χρησιμοποιούνται περίπλοκοι πόροι, όπως εκτεταμένες λίστες συντομογραφιών. Αν αναλογιστούμε ότι στην πλειοψηφία των εφαρμογών επεξεργασίας κειμένου, η ανίχνευση ορίων περιόδων είναι απλά ένα στάδιο προεπεξεργασίας, τότε γίνεται αντιληπτό πόσο σημαντικές είναι αυτές οι ιδιότητες. Η απόδοση του ανιχνευτή ορίων περιόδων είναι πολύ υψηλή και συγκρίσιμη με συστήματα που βασίζονται σε πολύ πιο περίπλοκους πόρους.

Η εξαγωγή της γνώσης γίνεται αυτόματα μέσω ενός σώματος εκπαίδευσης. Έτσι, το σύστημά μας μπορεί να προσαρμοστεί εύκολα σε κάποιον συγκεκριμένο τύπο

κειμένου ή ακόμα και σε κάποια άλλη γλώσσα με ιδιότητες παρόμοιες με αυτές της Νεοελληνικής. Η μέθοδος εκμάθησης, μία παραλλαγή της θεωρίας EBM, αποδείχτηκε ότι στο συγκεκριμένο πρόβλημα συμπεριφέρεται καλύτερα από την παραδοσιακή EBM τόσο στην ακρίβεια όσο και στο χρονικό κόστος εκπαίδευσης.

Η απόδοση του συστήματος ελέγχθηκε σε δύο σώματα κειμένων, ένα αποτελούμενο από μωσαϊκό τύπων κειμένων και ένα ομοιογενές υφολογικά. Τα αποτελέσματα δείχνουν πως ο ανιχνευτής ορίων περιόδων επιτυγχάνει καλύτερα αποτελέσματα στην πρώτη περίπτωση (99,4% έναντι 98,5% ακρίβεια). Ωστόσο, η διαφορά των δύο σωμάτων κειμένων όσον αφορά το κάτω όριο είναι αρκετά σημαντική. Έτσι, ο ανιχνευτής μας πέτυχε να ανεβάσει την απόδοση κατά 29,8% σε σχέση με το κάτω όριο στην περίπτωση του σώματος του *Βήματος* και κατά 39,4% στην περίπτωση του σώματος των αποφάσεων δικαστηρίου.