

Προτεινόμενα θέματα πτυχιακών και μεταπτυχιακών διπλωματικών εργασιών

Ίων Ανδρουτσόπουλος
Ομάδα Επεξεργασίας Φυσικής Γλώσσας¹
Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών

10 Οκτωβρίου 2011

Τα παρακάτω θέματα προσφέρονται τόσο για πτυχιακές όσο και για μεταπτυχιακές διπλωματικές εργασίες. Στην περίπτωση των μεταπτυχιακών διπλωματικών εργασιών, οι απαιτήσεις είναι περισσότερες. Οι ενδιαφερόμενοι μεταπτυχιακοί φοιτητές θα πρέπει να έχουν παρακολουθήσει επιτυχώς το μεταπτυχιακό μάθημα «Λογική και Τεχνητή Νοημοσύνη» του ΠΜΣ «Επιστήμη των Υπολογιστών» ή το μεταπτυχιακό μάθημα «Γλωσσική Τεχνολογία» του ΠΜΣ «Πληροφοριακά Συστήματα», κατά προτίμηση και τα δύο. Οι ενδιαφερόμενοι προπτυχιακοί φοιτητές θα πρέπει να έχουν παρακολουθήσει επιτυχώς το προπτυχιακό μάθημα «Τεχνητή Νοημοσύνη»· συνιστάται να έχουν παρακολουθήσει επίσης το μάθημα «Μηχανική Μάθηση».

Μπορείτε να μου προτείνετε και δικά σας θέματα εργασιών σχετικά με την Επεξεργασία Φυσικής Γλώσσας. Με ενδιαφέρουν ιδιαίτερα θέματα που αξιοποιούν αλγορίθμους από άλλους τομείς της Πληροφορικής ή των Μαθηματικών (π.χ. βελτιστοποίησης, θεωρίας πληροφορίας) σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας ή/και θέματα που αντιμετωπίζουν κάποιο σημαντικό πρόβλημα της καθημερινής ζωής.

Για περισσότερες πληροφορίες επικοινωνήστε μαζί μου μέσω ηλεκτρονικού ταχυδρομείου ή ελάτε να συζητήσουμε στο γραφείο μου, κατά προτίμηση ώρες γραφείου.²

1. Βελτίωση ελληνικού επισημειωτή μερών του λόγου

Η Ομάδα Επεξεργασίας Φυσικής Γλώσσας έχει αναπτύξει δύο εκδόσεις ενός ελληνικού επισημειωτή μερών του λόγου (part-of-speech tagger), δηλαδή ενός συστήματος που κατατάσσει αυτόματα τις λέξεις ελληνικών κειμένων στα μέρη του λόγου (ρήμα, ουσιαστικό, επίθετο κλπ.) αλλά και σε υποκατηγορίες αυτών (π.χ. ρήμα ενεργητικής φωνής αορίστου στο α' πρόσωπο ενικού, θηλυκό ουσιαστικό στη γενική πληθυντικού κλπ).³ Η πρώτη έκδοση του συστήματος χρησιμοποιεί έναν ταξινομητή k κοντινότερων γειτόνων και μεθόδους ενεργητικής μάθησης, που επιτρέπουν στο σύστημα να προτείνει το ίδιο παραδείγματα εκπαίδευσης. Η πρώτη έκδοση παρέχει, επίσης, γραφική διεπαφή χρήστη (GUI), αλλά όχι προγραμματιστική διεπαφή (API). Η δεύτερη, μεταγενέστερη έκδοση του συστήματος χρησιμοποιεί έναν ταξινομητή μεγίστης εντροπίας (maximum entropy classifier) και επιτυγχάνει καλύτερες επιδόσεις από την πρώτη έκδοση. Παρέχει

¹ Βλ. <http://nlp.cs.aueb.gr/>.

² Βλ. http://www.aueb.gr/users/ion/contact_gr.html.

³ Το λογισμικό διατίθεται από τη διεύθυνση <http://nlp.cs.aueb.gr/software.html>. Οι σχετικές προηγούμενες εργασίες βρίσκονται στη διεύθυνση <http://www.aueb.gr/users/ion/students.html>.

προγραμματιστική διεπαφή, αλλά όχι γραφική διεπαφή χρήστη, αντίθετα από την πρώτη έκδοση. Επίσης, η δεύτερη έκδοση δεν περιλαμβάνει μεθόδους ενεργητικής μάθησης.

Σκοπός της εργασίας είναι να βελτιώσει τη δεύτερη έκδοση του επισημειωτή μερών του λόγου, προσθέτοντας αρχικά γραφική διεπαφή παρόμοια με εκείνη της πρώτης έκδοσης. Κατόπιν, θα κατασκευαστούν και θα χρησιμοποιηθούν στη δεύτερη έκδοση περισσότερα δεδομένα εκπαίδευσης, ενδεχομένως χρησιμοποιώντας μεθόδους ενεργητικής μάθησης, μια που τα προηγούμενα πειράματα δείχνουν ότι αυτό θα βελτίωνε τις επιδόσεις του συστήματος. Θα διερευνηθούν, επίσης, μέθοδοι επιλογής ή εξαγωγής σύνθετων ιδιοτήτων (π.χ. με Principal Components Analysis). Ανάλογα με τα αποτελέσματα των νέων πειραμάτων της εργασίας, ενδέχεται να δοκιμαστούν και πρόσθετοι αλγόριθμοι μηχανικής μάθησης (π.χ. Conditional Random Fields) ή/και μέθοδοι ημι-επιβλεπόμενης μηχανικής μάθησης. Στην περίπτωση της ημι-επιβλεπόμενης μάθησης, το σύστημα εκπαιδεύεται αρχικά σε επισημειωμένα κείμενα (όπου κάθε λέξη έχει επισημειωθεί με το μέρος του λόγου στο οποίο ανήκει) και κατόπιν οι επιδόσεις του συστήματος βελτιώνονται χρησιμοποιώντας μη επισημειωμένα πρόσθετα κείμενα. Ενδέχεται, τέλος, να δοκιμαστούν συνδυασμοί των δύο υπαρχόντων εκδόσεων ή μέθοδοι συνεκπαίδευσης, όπου ένας ταξινομητής εκπαιδεύεται σε νέα παραδείγματα που έχουν επισημειωθεί από έναν άλλον ταξινομητή κ.ο.κ.

Η εργασία περιλαμβάνει: (α) αναζήτηση και μελέτη σχετικών εργασιών αναγνώρισης μερών του λόγου, ιδιαίτερα πρόσφατων εργασιών που εστιάζονται σε λιγότερο διαδοσμένες γλώσσες και χρησιμοποιούν ημι-επιβλεπόμενες ή μη επιβλεπόμενες μεθόδους μηχανικής μάθησης, (β) κατανόηση του υπάρχοντος και ανάπτυξη πρόσθετου λογισμικού, (γ) χειρωνακτική επισημείωση πρόσθετων κειμένων εκπαίδευσης και αξιολόγησης, (δ) διεξαγωγή εκτενών πειραμάτων, (ε) τεκμηρίωση των αποτελεσμάτων και του λογισμικού της εργασίας. Απαιτείται ευχέρεια προγραμματισμού.

2. Βελτίωση συστήματος αναγνώρισης ονομάτων οντοτήτων [ανατέθηκε]

Η αυτόματη αναγνώριση ονομάτων οντοτήτων (π.χ. ονόματα προσώπων, οργανισμών, τοποθεσιών, αλλά και ημερομηνίες, χρηματικά ποσά κλπ.) μέσα σε κείμενα και η κατάταξή τους σε κατηγορίες αποτελεί σημαντικό στάδιο σε πολλά συστήματα επεξεργασίας φυσικής γλώσσας (π.χ. συστήματα ερωταποκρίσεων, εξαγωγής πληροφοριών, παραγωγής περιλήψεων κλπ.). Στη διάρκεια προηγούμενων εργασιών, αναπτύχθηκε στην Ομάδα Επεξεργασίας Φυσικής Γλώσσας σύστημα αναγνώρισης και κατάταξης ονομάτων οντοτήτων για ελληνικά κείμενα, το οποίο χρησιμοποιεί Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines) και τεχνικές ενεργητικής μηχανικής μάθησης, με τις οποίες το ίδιο το σύστημα προτείνει στον εκπαιδευτή του παραδείγματα εκπαίδευσης. Το σύστημα έχει παρουσιαστεί σε διεθνή συνέδρια και περιοδικά και διατίθεται ελεύθερα.⁴ Έχει εκπαιδευτεί κυρίως σε κείμενα εφημερίδων και αναγνωρίζει ημερομηνίες, ονόματα προσώπων, οργανισμών και τοποθεσιών.

⁴ Το λογισμικό διατίθεται από τη διεύθυνση: <http://nlp.cs.aueb.gr/software.html>. Οι σχετικές εργασίες διατίθενται από στις διευθύνσεις: <http://www.aueb.gr/users/ion/students.html> και http://nlp.cs.aueb.gr/publications_gr.html.

Ο στόχος της προτεινόμενης εργασίας είναι η βελτίωση του υπάρχοντος συστήματος (ή η ανάπτυξη νέου), ώστε να είναι ευκολότερο να προσαρμοστεί σε νέες συλλογές εγγράφων (π.χ. ιατρικά ή νομικά άρθρα, αντί για άρθρα εφημερίδων) και νέες κατηγορίες ονομάτων (π.χ. ονόματα πρωτεϊνών, ονόματα διοικητικών οργάνων κλπ.), αλλά και να μπορεί ευκολότερα να ενσωματωθεί σε μεγαλύτερα συστήματα. Η βελτίωση της ταχύτητας του υπάρχοντος συστήματος είναι επίσης σημαντική, προκειμένου να είναι δυνατόν να χρησιμοποιηθεί ευκολότερα σε πρακτικές εφαρμογές. Θα διερευνηθεί, επίσης, αν οι επιδόσεις του είναι δυνατόν να βελτιωθούν χρησιμοποιώντας έναν υπάρχοντα επισημειωτή μερών του λόγου. Αν επαρκέσει ο χρόνος, θα διερευνηθεί επίσης η επέκταση του συστήματος, ώστε να εντοπίζει περιπτώσεις κυρίων ονομάτων που αναφέρονται στις ίδιες οντότητες (π.χ. «Ο.Τ.Ε.» και «Οργανισμός Τηλεπικοινωνιών της Ελλάδος», «ο κ. Γ. Παπαδημητρίου» και «ο Παπαδημητρίου»).

Η εργασία περιλαμβάνει: (α) μελέτη της σχετικής βιβλιογραφίας, (β) ενδεχομένως δημιουργία νέων κατάλληλα επισημειωμένων σωμάτων κειμένων, (γ) ανάπτυξη του νέου συστήματος (ή βελτίωση του υπάρχοντος), συμπεριλαμβανομένης προγραμματιστικής και γραφικής διεπαφής, (δ) αξιολόγησή του και (ε) τεκμηρίωση του λογισμικού και των αποτελεσμάτων της εργασίας. Απαιτείται ευχέρεια προγραμματισμού.

3. Ανάπτυξη πειραματικού ελληνικού συντακτικού αναλυτή εξαρτήσεων

Η συντακτική ανάλυση είναι χρήσιμο στάδιο στις περισσότερες εφαρμογές επεξεργασίας φυσικής γλώσσας. Ιδιαίτερα χρήσιμα σε πρακτικές εφαρμογές (π.χ. συμπίεση προτάσεων, εξαγωγή πληροφοριών, συστήματα ερωταποκρίσεων) έχουν αποδειχθεί τα τελευταία χρόνια τα δέντρα εξαρτήσεων (dependency trees). Οι κόμβοι των δέντρων αυτών παριστάνουν όλοι μεμονωμένες λέξεις (π.χ. δεν υπάρχουν κόμβοι για ονοματικά σύνολα, ρηματικά σύνολα ή άλλες φράσεις) και οι ακμές των δέντρων παριστάνουν εξαρτήσεις μεταξύ λέξεων (π.χ. άρθρο-ουσιαστικό, επίθετο-ουσιαστικό, ρήμα-αντικείμενο). Διατίθενται ελεύθερα αξιόπιστοι συντακτικοί αναλυτές αυτού του είδους για τα αγγλικά (κυρίως, αλλά και άλλες γλώσσες), αλλά όχι για τα ελληνικά.⁵

Στη διάρκεια αυτής της εργασίας θα αναπτυχθεί μια πρώτη, πειραματική έκδοση ενός αναλυτή συντακτικών εξαρτήσεων για ελληνικά κείμενα, που θα βελτιωθεί κατόπιν στη διάρκεια μελλοντικών εργασιών. Η εργασία περιλαμβάνει: (α) μελέτη της εκτενούς σχετικής βιβλιογραφίας, (β) δημιουργία μιας πρώτης (μικρής κλίμακας) συλλογής κειμένων επισημειωμένων με τα συντακτικά δέντρα των προτάσεων (treebank), (γ) ανάπτυξη εργαλείων υποβοήθησης της επισημείωσης, (δ) ανάπτυξη του συντακτικού αναλυτή, (ε) αξιολόγησή του και (στ) τεκμηρίωση του λογισμικού και των αποτελεσμάτων της εργασίας. Απαιτείται ευχέρεια προγραμματισμού και παρακολούθηση του μαθήματος «Γλωσσική Τεχνολογία». Η εργασία αυτή είναι δυνατόν να ανατεθεί σε ζεύγος φοιτητών, αλλά με περισσότερες απαιτήσεις.

⁵ Βλ. π.χ. <http://nlp.stanford.edu/software/index.shtml>.

4. Αυτόματος εντοπισμός περιπτώσεων λογοκλοπής σε κείμενα [ανατέθηκε]

Ως «λογοκλοπή» (plagiarism) εννοείται η αντιγραφή κειμένων ή τμημάτων κειμένων (π.χ. πτυχιακών εργασιών, επιστημονικών άρθρων) και η εμφάνισή τους ως πρωτότυπων κειμένων ενός άλλου συγγραφέα. Το φαινόμενο αυτό λαμβάνει τα τελευταία χρόνια εκρηκτικές διαστάσεις, εν μέρει λόγω της εξάπλωσης του Παγκόσμιου Ιστού και της εύκολης πρόσβασης σε μεγάλες συλλογές κειμένων (π.χ. ηλεκτρονικές εκδόσεις επιστημονικών περιοδικών, ψηφιακές βιβλιοθήκες).

Η εργασία περιλαμβάνει: (α) μελέτη των μεθόδων εντοπισμού περιπτώσεων λογοκλοπής που έχουν προταθεί στη βιβλιογραφία⁶, (β) συγκέντρωση ή/και δημιουργία συνόλων δεδομένων που θα είναι δυνατόν να χρησιμοποιηθούν κατά την πειραματική αξιολόγηση μεθόδων εντοπισμού λογοκλοπής, (γ) υλοποίηση και πειραματική αξιολόγηση μεθόδων εντοπισμού λογοκλοπής, (δ) βελτιώσεις υπάρχουσών μεθόδων εντοπισμού λογοκλοπής, ενδεχομένως αξιοποιώντας και μεθόδους εντοπισμού παραφράσεων⁷, (ε) τεκμηρίωση του λογισμικού και των αποτελεσμάτων της εργασίας. Απαιτείται ευχέρεια προγραμματισμού.

5. Αυτόματη αποσαφήνιση εννοιών λέξεων

Οι περισσότερες λέξεις των φυσικών γλωσσών έχουν πολλές δυνατές έννοιες (π.χ. «γράμμα» του ταχυδρομείου ή «γράμμα» της αλφαβήτου, «άπειρος» με τη μαθηματική έννοια ή αυτός που δεν έχει πείρα). Οι μέθοδοι αυτόματης αποσαφήνισης εννοιών λέξεων (word sense disambiguation) επιχειρούν να μαντέψουν την έννοια με την οποία χρησιμοποιείται κάθε λέξη σε ένα κείμενο, συχνά θεωρώντας δεδομένο ένα υπολογιστικό λεξικό που παραθέτει τις δυνατές έννοιες των λέξεων.⁸ Η Ομάδα Επεξεργασίας Φυσικής Γλώσσας, σε συνεργασία με ερευνητές άλλων πανεπιστημίων, έχει αναπτύξει μια μέθοδο αποσαφήνισης εννοιών λέξεων που χρησιμοποιεί μεθόδους μαθηματικής βελτιστοποίησης και μέτρα σημασιολογικής ομοιότητας εννοιών λέξεων. Η μέθοδος, που έχουν δοκιμαστεί κυρίως σε κείμενα ειδήσεων, επιτυγχάνει ήδη πολύ καλά αποτελέσματα συγκρινόμενη με το διεθνή ανταγωνισμό. Σκοπός της προτεινόμενης εργασίας θα είναι η περαιτέρω βελτίωσή της αλλά και η δοκιμή της σε βιοϊατρικά επιστημονικά άρθρα (όπου εμφανίζονται συχνά αμφίσημοι όροι).

Η εργασία περιλαμβάνει: (α) μελέτη των μεθόδων αποσαφήνισης εννοιών λέξεων που έχουν προταθεί στη βιβλιογραφία,⁹ (β) υλοποίηση βελτιώσεων της υπάρχουσας μεθόδου και αξιολόγησή τους σε υπάρχοντα σύνολα δεδομένων, (γ) αξιολόγηση της μεθόδου και των βελτιώσεών της σε νέα σύνολα δεδομένων που θα προέρχονται από βιοϊατρικά κείμενα, (δ) τεκμηρίωση του λογισμικού και των αποτελεσμάτων της εργασίας. Απαιτείται ευχέρεια προγραμματισμού.

⁶ Βλ. π.χ. <http://www.springerlink.com/content/1574-020x/45/1/>.

⁷ Βλ. <http://www.jair.org/papers/paper2985.html>.

⁸ Βλ. π.χ. <http://wordnet.princeton.edu/>.

⁹ Βλ. π.χ. <http://dl.acm.org/citation.cfm?id=1459355>.