Προτεινόμενα θέματα πτυχιακών και μεταπτυχιακών διπλωματικών εργασιών

Ίων Ανδρουτσόπουλος
Ομάδα Επεξεργασίας Φυσικής Γλώσσας[1]
Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών

12 Μαΐου 2018

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 «Γλωσσική Τεχνολογία» (του ΠΜΣ «Επιστήμη των Υπολογιστών») ή «Text Engineering and Analytics» (του ΠΜΣ «Επιστήμη Δεδομένων»). Οι ενδιαφερόμενοι προπτυχιακοί φοιτητές θα πρέπει να έχουν περάσει το προπτυχιακό μάθημα «Τεχνητή Νοημοσύνη» με βαθμό τουλάχιστον 8,5, να έχουν μέσο όρο βαθμολογίας τουλάχιστον 7 και να έχουν παρακολουθήσει (ή να μπορούν να παρακολουθήσουν) ατύπως το μεταπτυχιακό μάθημα «Γλωσσική Τεχνολογία». Για περισσότερες πληροφορίες επικοινωνήστε μαζί μου μέσω ηλεκτρονικού ταχυδρομείου ή από κοντά, ώρες γραφείου.[2]

1. Συμμετοχή στην ανάπτυξη συστήματος ερωταποκρίσεων

Η Ομάδα ΕΦΓ του ΟΠΑ έχει αναπτύξει σύστημα ερωταποκρίσεων για τον βιοϊατρικό διαγωνισμό BioASQ.[3] Το σύστημα περιλαμβάνει υποσυστήματα που:

- ανακτούν από συλλογές βιοϊατρικών εγγράφων τα έγγραφα που σχετίζονται περισσότερο με ένα ερώτημα διατυπωμένο σε φυσική γλώσσα,
- εντοπίζουν σε ανακτηθέντα βιοϊατρικά έγγραφα τα αποσπάσματα (π.χ. προτάσεις) που σχετίζονται περισσότερο με ένα ερώτημα,
- εντοπίζουν συγκεκριμένες απαντήσεις (π.χ. ζητούμενα ονόματα γονιδίων, ασθενειών, φαρμάκων) σε σχετικά αποσπάσματα βιοϊατρικών εγγράφων.

Υπάρχουν ευκαιρίες εκπόνησης εργασιών που θα επεκτείνουν το υπάρχον σύστημα ερωταποκρίσεων, για παράδειγμα:

- προσθέτοντας υποσύστημα που θα δέχεται ως είσοδο έγγραφα σχετικά με ένα ερώτημα και θα παράγει περίληψή τους, λαμβάνοντας υπόψιν και το ερώτημα,
- παράγοντας αυτόματα πολύ μεγάλα τεχνητά σύνολα δεδομένων εκπαίδευσης (π.χ. χρησιμοποιώντας τους τίτλους βιοϊατρικών εγγράφων ως ερωτήματα εκπαίδευσης, τα κυρίως μέρη των αντίστοιχων εγγράφων ως σχετικά έγγραφα, άλλα έγγραφα ως

[1] Βλ. http://nlp.cs.aueb.gr/.
[2] Βλ. http://www.aueb.gr/users/ion/contact_gr.html.
[3] Βλ. http://www.bioasq.org/.

μη σχετικά κατά την εκπαίδευση του υποσυστήματος ανάκτησης εγγράφων) και επανεκπαιδεύοντας το σύστημα με τα νέα σύνολα δεδομένων εκπαίδευσης,

- προσθέτοντας εύχρηστες διεπαφές χρήστη στο σύστημα, παράγοντας μια μορφή του συστήματος που θα μπορεί να χρησιμοποιηθεί στην πράξη (π.χ. ως εφαρμογή κινητής συσκευής) και αξιολογώντας την σε πραγματικές συνθήκες,
- τροποποιώντας κατάλληλα το σύστημα και αξιολογώντας το με άλλα, μη βιοϊατρικά σύνολα ανάκτησης πληροφοριών και ερωταποκρίσεων.[4]

Όλα τα τμήματα του συστήματος χρησιμοποιούν εκτενώς μεθόδους βαθιάς μηχανικής μάθησης (deep learning).

2. Μέθοδοι επαύξησης δεδομένων για συστήματα ΕΦΓ

Η εργασία αυτή θα μελετήσει μεθόδους που μπορούν να επαυξήσουν τεχνητά τα σύνολα δεδομένων εκπαίδευσης (data augmentation) συστημάτων ΕΦΓ, ιδιαίτερα συστημάτων που χρησιμοποιούν βαθιά μάθηση και χρειάζονται μεγάλα σύνολα δεδομένων εκπαίδευσης. Θα μελετηθούν υπάρχουσες μέθοδοι επαύξησης δεδομένων (π.χ. προσθήκη θορύβου, αντικατάσταση συνωνύμων, αυτόματη μετάφραση σε άλλη γλώσσα και πίσω, χρήση δεδομένων εκπαίδευσης συγγενικών προβλημάτων), θα προταθούν ενδεχομένως νέες μέθοδοι επαύξησης και θα μελετηθεί πειραματικά η επίπτωσή τους σε υπάρχοντα συστήματα που έχει αναπτύξει η ομάδα (π.χ. σύστημα ερωταποκρίσεων, εξόρυξης γνώμης, επεξεργασίας νομικών κειμένων, συντακτικής ανάλυσης) και άλλοι ερευνητές.

3. Σύστημα αναγνώρισης ονομάτων οντοτήτων

Σκοπός αυτής της εργασίας θα είναι να αναπτύξει ένα δημόσια διαθέσιμο σύστημα αναγνώρισης ονομάτων οντοτήτων (named entity recognizer) για ελληνικά κείμενα, βασισμένο σε βαθιά μάθηση, το οποίο θα αντικαταστήσει παλαιότερο σύστημα που παρέχει ήδη η ομάδα. Θα χρησιμοποιηθούν υπάρχουσες μέθοδοι αναγνώρισης ονομάτων οντοτήτων που βασίζονται σε βαθιά μάθηση, καθώς και υπάρχοντα ελληνικά σύνολα δεδομένων αναγνώρισης ονομάτων οντοτήτων, τα οποία θα επεκταθούν χειρωνακτικά ή/και αυτόματα (βλ. και προηγούμενη εργασία). Στο βαθμό που θα το επιτρέψει ο διαθέσιμος χρόνος, το σύστημα θα δοκιμαστεί και σε βιοϊατρικά ή/και νομικά κείμενα.

4. Εργαλειοθήκη ΕΦΓ για ελληνικά κείμενα

Σκοπός αυτής της εργασίας θα είναι να αναπτύξει μια δημόσια διαθέσιμη εργαλειοθήκη επεξεργασίας ελληνικών κειμένων, που θα περιλαμβάνει διαχωριστή λεκτικών μονάδων (tokenizer), διαχωριστή προτάσεων (sentence splitter), επισημειωτή μερών του λόγου (POS tagger), σύστημα αναγνώρισης ονομάτων οντοτήτων και συντακτικό αναλυτή εξαρτήσεων (dependency parsing). Έχει σχεδόν ολοκληρωθεί, με πολύ καλά αποτελέσματα, εργασία που αναπτύσσει ελληνικό επισημειωτή μερών του λόγου και συντακτικό αναλυτή εξαρτήσεων βασισμένους σε βαθιά μάθηση, ενώ θα αναπτυχθεί παράλληλα (βλ. προηγούμενη εργασία) νέο σύστημα αναγνώρισης ονομάτων οντοτήτων για ελληνικά κείμενα. Η παρούσα εργασία θα αναπτύξει ευέλικτο (που να μπορεί να προσαρμοστεί εύκολα σε νέα είδη κειμένων) διαχωριστή λεκτικών μονάδων και προτάσεων, τους οποίους θα συνδυάσει με τον

[4] Βλ. π.χ. http://www.msmarco.org/ και http://data.allenai.org/arc/.

επισημειωτή μερών του λόγου και τον συντακτικό αναλυτή που θα έχουν εντωμεταξύ ολοκληρωθεί, καθώς ενδεχομένως και με το νέο σύστημα αναγνώρισης ελληνικών ονομάτων οντοτήτων. Θα μελετηθούν υπάρχοντες εργαλειοθήκες επεξεργασίας φυσικής γλώσσας (π.χ. NLTK, spaCy) και θα επιδιωχθεί η δημιουργία αντίστοιχης εργαλειοθήκης για ελληνικά κείμενα, που θα περιλαμβάνει τα παραπάνω εργαλεία και μελλοντικά περισσότερα. Θα δοθεί ιδιαίτερη έμφαση στην αξιοπιστία, τη δυνατότητα επέκτασης και την τεκμηρίωση της εργαλειοθήκης.

5. Σύστημα ανάλυσης συναισθήματος και εξαγωγή γνώμης για ελληνικά κείμενα

Η Ομάδα ΕΦΓ του ΟΠΑ έχει αναπτύξει μεθόδους βαθιάς μάθησης που αναλύουν το συναίσθημα (sentiment analysis) και γενικότερα τη γνώμη των χρηστών ενός προϊόντος, υπηρεσίας κ.λπ., όπως εκφράζεται σε σχόλια κοινωνικών δικτύων (π.χ. tweets) ή κριτικές χρηστών (π.χ. σε ποια χαρακτηριστικά ενός προϊόντος αναφέρονται κυρίως οι χρήστες, ποια είναι κατά μέσο όρο η γνώμη τους για κάθε χαρακτηριστικό). Επίσης, συνδιοργάνωσε (με το Ερευνητικό Κέντρο «Αθηνά» και άλλες ομάδες του εξωτερικού) και συμμετείχε σε σχετικούς διεθνείς διαγωνισμούς.[5] Σκοπός της εργασίας θα είναι η βελτίωση των σχετικών μεθόδων και του λογισμικού της Ομάδας, η προσαρμογή τους στα Ελληνικά (συμπεριλαμβανομένης της δημιουργίας κατάλληλων δεδομένων εκπαίδευσης και αξιολόγησης) και η δημιουργία συστήματος ανάλυσης συναισθήματος και εξαγωγής γνώμης για ελληνικά κείμενα (ενδεχομένως σε συνδυασμό με την προηγούμενη εργασία). Θα δοθεί ιδιαίτερη έμφαση στην αξιοπιστία, τη δυνατότητα επέκτασης και την τεκμηρίωση του συστήματος. Η εργασία είναι δυνατόν να ανατεθεί σε ομάδα φοιτητών, με περισσότερες απαιτήσεις.

[5] Βλ. http://alt.qcri.org/semeval2014/task4/, http://alt.qcri.org/semeval2015/task12/, http://alt.qcri.org/semeval2016/task5/.