Προτεινόμενα θέματα πτυχιακών και μεταπτυχιακών διπλωματικών εργασιών

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

12 Μαΐου 2017

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 (ή να παρακολουθούν) το μεταπτυχιακό μάθημα «Γλωσσική Τεχνολογία» (των ΠΜΣ «Επιστήμη των Υπολογιστών» και «Πληροφοριακά Συστήματα») ή το μεταπτυχιακό μάθημα «Text Engineering and Analytics» (του ΠΜΣ «Επιστήμη των Δεδομένων»). Οι ενδιαφερόμενοι προπτυχιακοί φοιτητές θα πρέπει να έχουν περάσει το προπτυχιακό μάθημα «Τεχνητή Νοημοσύνη» με βαθμό τουλάχιστον 8.

Μπορείτε να μου προτείνετε και δικά σας θέματα εργασιών σχετικών με την Τεχνολογία Φυσικής Γλώσσας. Για περισσότερες πληροφορίες επικοινωνήστε μαζί μου μέσω ηλεκτρονικού ταχυδρομείου ή ελάτε να συζητήσουμε στο γραφείο μου, ώρες γραφείου.[2]

1. Σύστημα ανάλυσης συναισθήματος για μηνύματα κοινωνικών δικτύων

Η Ομάδα Επεξεργασίας Φυσικής Γλώσσας (ΕΦΓ) του ΟΠΑ έχει αναπτύξει μεθόδους μηχανικής μάθησης και συστήματα που κατατάσσουν μηνύματα κοινωνικών δικτύων (π.χ. tweets) σε κατηγορίες (π.χ. θετική, αρνητική, ουδέτερη) ανάλογα με το συναίσθημα που εκφράζουν (π.χ. για ένα προϊόν).[3] Η Ομάδα λαμβάνει επίσης μέρος σε σχετικούς διεθνείς διαγωνισμούς.[4] Σκοπός της εργασίας θα είναι η βελτίωση των σχετικών μεθόδων και του λογισμικού της Ομάδας, μεταξύ άλλων χρησιμοποιώντας βαθιά νευρωνικά δίκτυα, καθώς και η εκ νέου συμμετοχή της Ομάδας σε σχετικούς διεθνείς διαγωνισμούς. Η εργασία είναι δυνατόν να ανατεθεί σε ομάδα φοιτητών, που θα ασχοληθούν με ξεχωριστά τμήματα ενός συστήματος, με περισσότερες συνολικά απαιτήσεις.

2. Σύστημα εξόρυξης γνώμης από κριτικές χρηστών

Η Ομάδα ΕΦΓ του ΟΠΑ έχει αναπτύξει μεθόδους μηχανικής μάθησης που εξάγουν τη γνώμη των χρηστών ενός προϊόντος από κριτικές και σχόλιά τους (π.χ. σε ποια χαρακτηριστικά

---

[1] Βλ. http://nlp.cs.aueb.gr/.
[2] Βλ. http://www.aueb.gr/users/ion/contact_gr.html.
[3] Βλ. http://nlp.cs.aueb.gr/theses/karampatsis_msc_thesis.pdf.
[4] Βλ. http://alt.qcri.org/semeval2014/task9/, http://alt.qcri.org/semeval2015/task10/, http://alt.qcri.org/semeval2016/task4/, http://alt.qcri.org/semeval2017/task4/, http://nlp.cs.aueb.gr/pubs/semeval.2013.pdf, http://nlp.cs.aueb.gr/pubs/SemEval2014015.pdf, http://nlp.cs.aueb.gr/pubs/SemEval_2016_Task_4.pdf. Βλ. και το νεότερο διαγωνισμό http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html.

ενός προϊόντος αναφέρονται κυρίως οι χρήστες, ποια είναι κατά μέσο όρο η γνώμη τους για κάθε χαρακτηριστικό).[5] Επίσης, συνδιοργάνωσε (με το Ερευνητικό Κέντρο «Αθηνά» και άλλες ομάδες του εξωτερικού) και συμμετείχε σε σχετικούς διεθνείς διαγωνισμούς.[6] Σκοπός της εργασίας θα είναι η βελτίωση των σχετικών μεθόδων και του λογισμικού της Ομάδας, μεταξύ άλλων χρησιμοποιώντας βαθιά νευρωνικά δίκτυα, καθώς και η εκ νέου συμμετοχή της Ομάδας σε σχετικούς διεθνείς διαγωνισμούς, ενδεχομένως και στη διοργάνωσή τους. Η εργασία είναι δυνατόν να ανατεθεί σε ομάδα φοιτητών, που θα ασχοληθούν με ξεχωριστά τμήματα ενός συστήματος, με περισσότερες συνολικά απαιτήσεις.

3. Συμμετοχή στην ανάπτυξη βιοϊατρικού συστήματος ερωταποκρίσεων

Η Ομάδα ΕΦΓ του ΟΠΑ συνδιοργανώνει (με το ΕΚΕΦΕ «Δημόκριτος») το διεθνή διαγωνισμό βιοϊατρικών συστημάτων ερωταποκρίσεων BioASQ.[7] Συμμετέχει επίσης στην ανάπτυξη του βιοϊατρικού συστήματος ερωταποκρίσεων Almosino, συνεργαζόμενη με το Ερευνητικό Κέντρο «Αθηνά». Βρίσκονται σε εξέλιξη ή έχουν ολοκληρωθεί εργασίες που εστιάζονται στην ανάπτυξη τμημάτων του συγκεκριμένου συστήματος, όπως:

- τμήμα που ανακτά από αποθετήρια βιοϊατρικών εγγράφων τα έγγραφα που σχετίζονται περισσότερο με ένα ερώτημα διατυπωμένο σε φυσική γλώσσα,[8]
- τμήμα που εντοπίζει σε βιοϊατρικά έγγραφα τα αποσπάσματα (π.χ. προτάσεις) που σχετίζονται περισσότερο με ένα ερώτημα,[9]
- τμήμα που εντοπίζει απαντήσεις (π.χ. ζητούμενα ονόματα γονιδίων, ασθενειών, φαρμάκων) σε σχετικά αποσπάσματα βιοϊατρικών εγγράφων,[10]
- τμήματα που παράγουν και αξιολογούν περιλήψεις βιοϊατρικών εγγράφων, λαμβάνοντας υπόψη ένα συγκεκριμένο ερώτημα.[11]

Υπάρχουν ευκαιρίες εκπόνησης πτυχιακών ή διπλωματικών εργασιών που θα επεκτείνουν προηγούμενες εργασίες ή θα εστιαστούν σε άλλα τμήματα του συστήματος, όπως π.χ.:

- αυτόματη κατάταξη ερωτημάτων σε κατηγορίες ανάλογα με τα είδη των απαντήσεων (π.χ. γονίδια, ασθένειες, φάρμακα) που απαιτούν,[12]

---

[5] Βλ. http://nlp.cs.aueb.gr/theses/ipavlopoulos-thesis.pdf, http://nlp.cs.aueb.gr/pubs/W14-1306.pdf, http://nlp.cs.aueb.gr/pubs/E14-1009.pdf, http://nlp.cs.aueb.gr/theses/ilazari_msc_thesis.pdf, http://nlp.cs.aueb.gr/theses/procopiou_msc_thesis.pdf.

[6] Βλ. http://alt.qcri.org/semeval2014/task4/, http://alt.qcri.org/semeval2015/task12/, http://alt.qcri.org/semeval2016/task5/, http://nlp.cs.aueb.gr/pubs/SemEval2014_ABSA_overview.pdf, http://nlp.cs.aueb.gr/pubs/SemEval2015_ABSA_overview.pdf, http://nlp.cs.aueb.gr/pubs/SemEval2016_ABSA_overview.pdf.

[7] Βλ. http://www.bioasq.org/, http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0564-6.

[8] Βλ. http://nlp.cs.aueb.gr/theses/dgkoumas_msc_thesis.pdf, http://nlp.cs.aueb.gr/theses/tryfona_msc_thesis.pdf.

[9] Βλ. http://nlp.cs.aueb.gr/theses/mgeorgiou_msc_thesis.pdf.

[10] Βλ. http://nlp.cs.aueb.gr/theses/andronis_msc_thesis.pdf.

[11] Βλ. http://nlp.cs.aueb.gr/theses/galanis_phd_thesis.pdf, http://nlp.cs.aueb.gr/pubs/coling2012.pdf, http://nlp.cs.aueb.gr/pubs/ucnlg2011.pdf, http://nlp.cs.aueb.gr/pubs/AUEB_ILSP_BioASQ_Task3b_CLEF2015.pdf.

[12] Βλ. http://nlp.cs.aueb.gr/theses/vrisagotis_final_report.pdf.

- αυτόματη επισημείωση επιστημονικών άρθρων (ή/και ερωτημάτων) με έννοιες βιοϊατρικών οντολογιών,[13]
- επέκταση πλατφόρμας ανάπτυξης συστημάτων ερωταποκρίσεων.[14]

Όλα τα τμήματα του συστήματος χρησιμοποιούν εκτενώς μεθόδους μηχανικής μάθησης. Τα πιο πρόσφατα τμήματα χρησιμοποιούν και βαθιά νευρωνικά δίκτυα.

4. Επισημείωση μερών του λόγου και ονομάτων οντοτήτων

Η Ομάδα ΕΦΓ του ΟΠΑ έχει αναπτύξει επισημειωτές μερών του λόγου (part-of-speech taggers) για ελληνικά κείμενα, χρησιμοποιώντας μεθόδους μηχανικής μάθησης.[15] Ο πιο πρόσφατος επισημειωτής μερών του λόγου χρησιμοποιεί διανυσματικές παραστάσεις λέξεων (word embeddings)[16] και επιτυγχάνει τα ίδια ή καλύτερα αποτελέσματα με προηγούμενους επισημειωτές οι οποίοι χρησιμοποιούσαν χειρωνακτικά κατασκευασμένα χαρακτηριστικά (features). Η προτεινόμενη εργασία θα διερευνήσει πρώτα αν οι επιδόσεις του πιο πρόσφατου επισημειωτή μερών του λόγου μπορούν να βελτιωθούν περαιτέρω χρησιμοποιώντας διανυσματικές παραστάσεις λέξεων περισσότερων διαστάσεων, περισσότερα δεδομένα εκπαίδευσης ή/και πιο πρόσφατες μεθόδους βασισμένες σε βαθιά νευρωνικά δίκτυα. Κατόπιν θα διερευνήσει πώς ο επισημειωτής μερών του λόγου μπορεί να μετατραπεί, ώστε να επισημειώνει και ονόματα οντοτήτων (π.χ. ονόματα προσώπων, εταιρειών, τοποθεσιών), λαμβάνοντας υπόψη και συγκρίνοντας με προηγούμενες σχετικές εργασίες.[17] Σε περίπτωση που επαρκεί ο χρόνος, θα αναπτυχθεί επίσης μια ειδική μορφή του επισημειωτή ειδικά για νομικά ή βιοϊατρικά κείμενα ή μηνύματα κοινωνικών δικτύων.[18] Θα επιδιωχθεί η ενσωμάτωση του επισημειωτή σε διαδεδομένη συλλογή εργαλείων επεξεργασίας φυσικής γλώσσας.[19]

5. Διεπαφές γραπτών προφορικών διαλόγων

Πρόσφατα εκδηλώνεται μεγάλο ενδιαφέρον για συστήματα που επιτρέπουν στους χρήστες τους να κλείνουν εισιτήρια, να παραγγέλνουν φαγητό, να ζητούν πληροφορίες κ.λπ.

---

[13] Βλ. http://nlp.cs.aueb.gr/pubs/jbms_dense_vectors.pdf, http://link.springer.com/article/10.1007/s10618-014-0382-x.
[14] Βλ. http://nlp.cs.aueb.gr/theses/andriopoulos_msc_thesis.pdf.
[15] Βλ. http://nlp.cs.aueb.gr/theses/ekoleli_final_report.pdf, http://nlp.cs.aueb.gr/theses/malakasiotis_final_msc_report.pdf, http://nlp.cs.aueb.gr/theses/pappas_final_report.pdf, http://nlp.cs.aueb.gr/theses/asikis_msc_thesis.pdf.
[16] Βλ. https://code.google.com/archive/p/word2vec/, http://nlp.stanford.edu/projects/glove/.
[17] Βλ. http://nlp.cs.aueb.gr/pubs/ijait_greek_nerc.pdf, http://nlp.cs.aueb.gr/theses/vassilakos_final_report.pdf, http://nlp.cs.aueb.gr/pubs/setn2006_paper.pdf, http://nlp.cs.aueb.gr/theses/lucarelli_msc_final_report.pdf, http://nlp.cs.aueb.gr/theses/konstas_final_report.pdf, http://nlp.cs.aueb.gr/theses/antonellos_msc_thesis.pdf
[18] Βλ. http://nlp.cs.aueb.gr/theses/mKarampatsisBScThesis.pdf, http://aclweb.org/anthology-new/D/D11/D11-1141.pdf, http://www.ark.cs.cmu.edu/TweetNLP/.
[19] Βλ. π.χ. http://www.nltk.org/, https://opennlp.apache.org/, http://gate.ac.uk/, http://www.ellogon.org/.

ανταλλάσσοντας γραπτά μηνύματα (SMS, chat) με τεχνητούς πράκτορες (chatbots).[20] Η προτεινόμενη εργασία θα διερευνήσει υπάρχοντα εργαλεία που διευκολύνουν την ανάπτυξη συστημάτων αυτού του είδους, ενδεχομένως και σχετικούς αλγορίθμους μηχανικής μάθησης.[21] Επίσης, θα αναπτύξει πειραματικά συστήματα αυτού του είδους (π.χ. για κλείσιμο αιθουσών ή παροχή πληροφοριών σε φοιτητές του ΟΠΑ) τα οποία και θα αξιολογήσει. Η εργασία είναι δυνατόν να ανατεθεί σε ομάδα φοιτητών, που θα ασχοληθούν με ξεχωριστά πειραματικά συστήματα, με περισσότερες συνολικά απαιτήσεις.

6. Ανάπτυξη πειραματικού συντακτικού αναλυτή εξαρτήσεων για τα Ελληνικά

Η εργασία αυτή προσφέρεται μόνο ως μεταπτυχιακή διπλωματική εργασία. Στη διάρκειά της θα αναπτυχθεί ένας πειραματικός συντακτικός αναλυτής εξαρτήσεων (dependency parser) για τα Ελληνικά, ο οποίος θα βασίζεται σε διανυσματικές παραστάσεις λέξεων (word embeddings) και βαθιά νευρωνικά δίκτυα. Η εργασία θα χρησιμοποιήσει υπάρχοντες αλγορίθμους και μοντέλα νευρωνικών δικτύων για συντακτική ανάλυση εξαρτήσεων, καθώς και υπάρχοντα σώματα συντακτικών δέντρων (treebanks) για τα Ελληνικά και άλλες γλώσσες.[22] Θα επιδιωχθεί η βελτίωση των υπαρχόντων αλγορίθμων και μοντέλων, ώστε να λειτουργούν ικανοποιητικότερα στα Ελληνικά, καθώς και η συμμετοχή σε διεθνείς διαγωνισμούς.[23]

7. Εξαγωγή πληροφοριών από νομικά κείμενα

Η Ομάδα Επεξεργασίας Φυσικής Γλώσσας του ΟΠΑ έχει αναπτύξει σε συνεργασία με εταιρεία του εξωτερικού συστήματα που εξάγουν πληροφορίες από νομικά κείμενα (π.χ. πρόσωπα, οργανισμούς, αρμόδια δικαστήρια, ποσά που αναφέρονται σε συμβόλαια). Υπάρχουν ευκαιρίες εκπόνησης πτυχιακών ή διπλωματικών εργασιών, ενδεχομένως αμειβόμενων, που θα επεκτείνουν τα υπάρχοντα συστήματα, για παράδειγμα:

- εξετάζοντας τρόπους αυτόματης επαύξησης των δεδομένων εκπαίδευσης (data augmentation),
- συμμετέχοντας στη βελτίωση των ήδη χρησιμοποιούμενων μεθόδων μηχανικής μάθησης (κυρίως ανατροφοδοτούμενα νευρωνικά δίκτυα),
- συμμετέχοντας στην ανάπτυξη μεθόδων για την εξαγωγή πρόσθετων πληροφοριών (π.χ. εντοπισμός νέων τύπων οντοτήτων ή/και εξαγωγή σχέσεων οντοτήτων) ενδεχομένως από νέα είδη νομικών κειμένων.

---

[20] Βλ. π.χ. http://www.bloomberg.com/features/2016-microsoft-future-ai-chatbots/, https://medium.com/@ben8128/the-messaging-landscape-in-2016-13b25cdf2f6e#.703rkwvf3.
[21] Βλ. π.χ. https://dev.botframework.com/, http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/, https://github.com/facebookresearch/ParlAI/.
[22] Βλ. π.χ. http://www.aclweb.org/anthology/D/D14/D14-1082.pdf, https://transacl.org/ojs/index.php/tacl/article/viewFile/885/198, https://transacl.org/ojs/index.php/tacl/article/viewFile/798/208, http://www.aclweb.org/anthology/P09-1039, http://universaldependencies.org/.
[23] Βλ. π.χ. http://universaldependencies.org/conll17/.