Εργαστήριο Βάσεων Δεδομένων

Θέματα Διπλωματικών Εργασιών 2013-2014

***Σημείωση*:** *Για να έχετε πρόσβαση σε κάποιες από τις βιβλιογραφικές αναφορές χρειάζεται να συνδέεστε από το υποδίκτυο του Πανεπιστημίου ή με VPN*

Περιεχόμενα

Θέμα 1

Εγκατάλειψη αποτελεσμάτων αναζήτησης: Ερμηνεία και εφαρμογή τεχνικών αναγνώρισης εγκαταλελειμμένων αναζητήσεων.

Κατά την αλληλεπίδραση ενός χρήστη με τη μηχανή αναζήτησης είναι αρκετά κοινό το σενάριο που περιλαμβάνει την υποβολή του ερωτήματος, την προβολή της λίστας των αποτελεσμάτων και ακολούθως της εγκατάλειψης της αναζήτησης χωρίς την πλοήγηση σε κάποιο αποτέλεσμα. Η συχνότητα του φαινομένου αλλά και το μεγάλο ενδιαφέρον που παρουσιάζει η ερμηνεία του για τη μηχανή αναζήτησης και τις προσφερόμενες υπηρεσίες, έχουν συγκεντρώσει αρκετό ενδιαφέρον από μέρους της ερευνητικής κοινότητας με στόχο την ανάλυσή του. Οι κύριες προκλήσεις που αντιμετωπίζουν οι ερευνητές του πεδίου αφορούν αρχικά στην αναγνώριση από server logs των ορίων του session μιας αναζήτησης, στην ανίχνευση αναζητήσεων που εγκαταλήφθηκαν και, τέλος, στην ερμηνεία της εγκατάλειψης.

Ένα session αναζήτησης μπορεί να περιλαμβάνει περισσότερες από μία υποβολές ερωτημάτων, εφόσον αφορούν στην ίδια πληροφοριακή ανάγκη και πραγματοποιούνται με τον ίδιο στόχο. Η ανίχνευση των ορίων ενός session σε μια online αναζήτηση αποτελεί αυτοτελές ερευνητικό θέμα το οποίο αφορά στην ανάλυση των logs επικοινωνίας των χρηστών με μια μηχανή αναζήτησης με στόχο την αναγνώριση των sessions και των ενεργειών που περιλαμβάνονται σε αυτά. Η αυτόματη ανάλυση της συμπεριφοράς των χρηστών στα πλαίσια ενός session παρουσιάζει εξαιρετικό ενδιαφέρον, εφόσον προσφέρει δυνατότητα εκτίμησης σε σχέση με το σε ποιό βαθμό οι χρήστες ικανοποιούν την πληροφοριακή τους ανάγκη και τις δυσκολίες που αντιμετωπίζουν για να διατυπώσουν το ερώτημά τους. Η ερμηνεία της εγκατάλειψης των αποτελεσμάτων αναζήτησης συγκεκριμένα είναι επίσης ανοιχτό ερευνητικό θέμα. Μια εγκαταλελειμμένη αναζήτηση δεν συνεπάγεται αυτόματα μη ικανοποίηση από τα αποτελέσματα, καθώς μπορεί ο χρήστης να κάλυψε την πληροφοριακή του ανάγκη βλέποντας μόνο τη λίστα των αποτελεσμάτων χωρίς να χρειάζεται να πλοηγηθεί σε αυτά, πχ στην περίπτωση ερωτημάτων που αφορούν σε έλεγχο ορθογραφίας.

Με βάση τα παραπάνω, στην παρούσα διπλωματική καλείστε να πραγματοποιήσετε μια ανασκόπηση της βιβλιογραφίας που αφορά στην ανίχνευση και ερμηνεία εγκαταλελειμμένων αναζητήσεων. Στη συνέχεια θα συλλέξετε ένα πειραματικό σώμα από search logs και θα υλοποιήσετε εφαρμογή ανίχνευσης εγκαταλελειμμένων αναζητήσεων χρησιμοποιώντας κάποιον από τους αλγορίθμους που προτείνονται στη βιβλιογραφία.

Ενδεικτική Βιβλιογραφία:

- http://research.microsoft.com/en-us/um/people/sdumais/CIKM2012-fp085-diriye.pdf
- http://www.sciencedirect.com/science/article/pii/S002002550900053X
- http://www2012.wwwconference.org/proceedings/companion/p485.pdf
- http://www.dblab.upatras.gr/download/nlp/NLP-Group-Pubs/09-SIGIR_Queries_without_Clicks.pdf
- https://support.google.com/analytics/answer/1032402?hl=en

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Τζέκου Βιβή
Email επικοινωνίας: tzekou@ceid.upatras.gr

Θέμα 2

Αναγνώριση θυμού από σχόλια χρηστών στον Παγκόσμιο Ιστό: Μελέτη και πειραματική εφαρμογή.

Η αύξηση των μέσων κοινωνικής δικτύωσης, όπως τα ιστολόγια (blogs) και τα κοινωνικά δίκτυα (social networks) έχει στρέψει το ερευνητικό ενδιαφέρον στην ανάλυση συναισθήματος (sentiment analysis). Ο όρος ανάλυση συναισθήματος αναφέρεται στον αυτόματο εντοπισμό και εξαγωγή απόψεων, συναισθημάτων και διαθέσεων από έγγραφα κειμένου. Ο βασικός στόχος της ανάλυσης συναισθήματος είναι ο χαρακτηρισμός της πολικότητας ενός συγκεκριμένου κειμένου – αν η γνώμη που εκφράζεται σε αυτό ερμηνεύεται ως θετική, αρνητική, ή ουδέτερη. Έχει αποδειχθεί πως οι παραδοσιακές προσεγγίσεις ταξινόμησης κειμένου (text classification) μπορεί να είναι αρκετά αποτελεσματικές όταν εφαρμόζονται στο πρόβλημα της ανάλυσης συναισθήματος. Μοντέλα όπως Naïve Bayes (NB), Maximum Entropy (ME) και Support Vector Machines (SVM) μπορούν να προσδιορίσουν το συναίσθημα των κειμένων με υψηλή ακρίβεια.

Τέτοιου είδους παραδοσιακές δυαδικές προσεγγίσεις αποτυγχάνουν να αναγνωρίσουν πολλαπλά, συχνά αντικρουόμενα, συναισθήματα που συναντώνται σε ένα έγγραφο. Ο εντοπισμός συναισθηματικών καταστάσεων όπως θυμός, σαρκασμός, λύπη, ικανοποίηση κλπ μπορεί να αποδειχτεί πιο αντιπροσωπευτικός και χρήσιμος για την ανάλυση του συναισθήματος ενός εγγράφου από την απλή αναγνώριση της πολικότητας του εγγράφου αυτού. Η ποικιλομορφία στον τρόπο έκφρασης των προαναφερθέντων συναισθημάτων από τους χρήστες του Παγκόσμιου Ιστού καθιστά την αναγνώρισή τους ένα δύσκολο αλλά και ενδιαφέροντα στόχο της ανάλυση συναισθήματος.

Στην παρούσα διπλωματική καλείστε αρχικά να πραγματοποιήσετε μια βιβλιογραφική έρευνα πάνω στις προτεινόμενες τεχνικές αναγνώρισης συναισθημάτων από σχόλια χρηστών στον Παγκόσμιο Ιστό. Στη συνέχεια θα σχεδιάσετε και θα εκπαιδεύσετε έναν κατηγοριοποιητή ο οποίος θα επιτρέπει την αυτόματη αναγνώριση θυμού σε σχόλια χρηστών, χρησιμοποιώντας μετρικές που έχουν προταθεί από τη βιβλιογραφία καθώς και δικές σας μετρικές, ενώ θα πραγματοποιηθεί πειραματική μέτρηση της απόδοσης του προτεινόμενου κατηγοριοποιητή και θα παρουσιαστούν στην διπλωματική τα αποτελέσματα.

Ενδεικτική Βιβλιογραφία:

- http://en.wikipedia.org/wiki/Sentiment_analysis
- http://www.cse.unt.edu/~rada/papers/mihalcea.aaai06ss.pdf
- http://xldb.di.fc.ul.pt/xldb/publications/Carvalho09:Clues:Detecting:Irony_document.pdf

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Μεταπτυχιακός: Ντούρου Δήμητρα (ntourou@ceid.upatras.gr)
Υπεύθυνη Διδακτορικός: Κουμπούρη Αθανασία (koumpour@ceid.upatras.gr)

Θέμα 3

Μελέτη και εφαρμογή τεχνικών αυτόματης αναγνώρισης ιστολογίων.

Η λέξη "blog" είναι συντόμευση της λέξης "weblog", η οποία μεταφράζεται ως ιστολόγιο. Ο όρος χρησιμοποιείται για να περιγράψει ένα είδος ιστότοπων με δομή ημερολογίου. Τα βασικά χαρακτηριστικά των ιστολογίων περιλαμβάνουν τα εξής:

- Τα άρθρα που δημοσιεύονται είναι δομημένα σαν καταχωρίσεις σε ημερολόγιο.
- Τα άρθρα παρουσιάζονται συνήθως σε αντίστροφη χρονολογική σειρά.
- Οι επισκέπτες έχουν τη δυνατότητα να σχολιάζουν τα δημοσιευμένα άρθρα.

Τα blogs ξεκίνησαν από άτομα ή μικρές ομάδες και αφορούσαν ένα συγκεκριμένο θέμα ενδιαφέροντος για τους συγγραφείς και την κοινότητα στην οποία απευθύνονται. Τα τελευταία χρόνια έχουν αρχίσει να δημιουργούνται blogs από οργανισμούς ή εταιρείες, στα οποία συμμετέχουν πολλοί συγγραφείς, γεγονός το οποίο έχει συντελέσει στο αυξημένο ενδιαφέρον του κοινού για τα ιστολόγια.

Το ερευνητικό ενδιαφέρον για τα blogs είναι αρκετά αυξημένο καθώς αποτελούν φαινόμενο το οποίο γνωρίζει όλο και μεγαλύτερη άνθηση, ενώ έχει εμπορικές, κοινωνικές και πολιτικές προεκτάσεις. Αρκετές έρευνες στρέφονται στη μελέτη της χαρακτηριστικής δομής τους καθώς και των γλωσσολογικών ιδιαιτεροτήτων που παρουσιάζουν, με πιο χαρακτηριστικές τις προσπάθειες για αυτόματη αναγνώριση, συγκομιδή και δεικτοδότηση της blogosphere (το σύνολο των blogs στον Παγκόσμιο Ιστό). Σύμφωνα με τη βιβλιογραφία, είναι δυνατή η αυτόματη αναγνώριση των σελίδων που εντάσσονται σε ένα ιστολόγιο λαμβάνοντας υπόψη τα χαρακτηριστικά των ιστολογίων γενικά, ανεξαρτήτως από το θέμα το οποίο πραγματεύονται.

Στην παρούσα διπλωματική καλείστε αρχικά να πραγματοποιήσετε μια βιβλιογραφική έρευνα πάνω στις προτεινόμενες τεχνικές αυτόματης αναγνώρισης ιστολογίων. Στη συνέχεια θα σχεδιάσετε και θα εκπαιδεύσετε έναν κατηγοριοποιητή ο οποίος θα επιτρέπει την αυτόματη αναγνώριση ιστοσελίδων από blogs, χρησιμοποιώντας χαρακτηριστικά που έχουν προταθεί από τη βιβλιογραφία, ενώ θα πραγματοποιηθεί πειραματική μέτρηση της απόδοσης του προτεινόμενου κατηγοριοποιητή και θα παρουσιαστούν στην διπλωματική τα αποτελέσματα.

Ενδεικτική Βιβλιογραφία:

- http://en.wikipedia.org/wiki/Weblogs
- http://dl.acm.org/citation.cfm?id=1013455
- http://acl.ldc.upenn.edu/eacl2006/ws12_newtext.pdf#page=32
- https://www.era.lib.ed.ac.uk/handle/1842/1113
- http://dl.acm.org/citation.cfm?id=1459357
- http://dare.uva.nl/document/46517

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Τζέκου Βιβή
Email Επικοινωνίας: tzekou@ceid.upatras.gr

Θέμα 4

Μελέτη και εφαρμογή τεχνικών κατηγοριοποίησης συναισθήματος στο περιεχόμενο ιστολογίων

Η λέξη “blog” είναι συντόμευση της λέξης “weblog”, η οποία μεταφράζεται ως ιστολόγιο. Ο όρος χρησιμοποιείται για να περιγράψει ένα είδος ιστότοπων με δομή ημερολογίου. Τα blogs εξ' ορισμού εκφράζουν απόψεις ή/και συναισθήματα σε σχέση με το θέμα το οποίο πραγματεύονται. Οι συγγραφείς των άρθρων τοποθετούνται πάνω στο θέμα για το οποίο γράφουν, ενώ οι επισκέπτες σχολιάζουν την άποψη του γράφοντος. Με δεδομένο ότι τα blogs μπορεί να αφορούν σε διάφορα θέματα εμπορικού ή και κοινωνικού ενδιαφέροντος, εύλογα βγαίνει το συμπέρασμα ότι η ανίχνευση της άποψης των χρηστών του διαδικτύου για συγκεκριμένα θέματα, προϊόντα ή υπηρεσίες μπορεί να επιτευχθεί με μεγαλύτερη ευκολία μέσω της εκμετάλλευσης των ιδιαίτερων χαρακτηριστικών των blogs.

Η αύξηση των μέσων κοινωνικής δικτύωσης, όπως τα ιστολόγια (blogs) και τα κοινωνικά δίκτυα (social networks) έχει στρέψει το ερευνητικό ενδιαφέρον στην ανάλυση συναισθήματος (sentiment analysis). Ο όρος ανάλυση συναισθήματος αναφέρεται στον αυτόματο εντοπισμό και εξαγωγή απόψεων, συναισθημάτων και διαθέσεων από έγγραφα κειμένου. Μια υποκατηγορία αυτού του τομέα είναι η κατηγοριοποίηση συναισθήματος (sentiment classification). Δοθέντος ενός κειμένου είναι εφικτό με αυτοματοποιημένες μεθόδους να αποφανθούμε εάν το κείμενο έχει θετικό ή αρνητικό προσανατολισμό. Είναι ένα περίπλοκο και δύσκολο έργο για να επιτευχθεί, εάν λάβουμε υπόψη μας τις δυσκολίες που απαντώνται κατά την ανάπτυξη μιας μεθόδου για να αναγνωρίζει για παράδειγμα τον σαρκασμό. Έχει αποδειχθεί πως οι παραδοσιακές προσεγγίσεις ταξινόμησης κειμένου (text classification) μπορεί να είναι αρκετά αποτελεσματικές όταν εφαρμόζονται στο πρόβλημα της ανάλυσης συναισθήματος. Μοντέλα όπως Naïve Bayes (NB), Maximum Entropy (ME) και Support Vector Machines (SVM) μπορούν να προσδιορίσουν το συναίσθημα των κειμένων με υψηλή ακρίβεια.

Στην παρούσα διπλωματική καλείστε αρχικά να πραγματοποιήσετε μια επισκόπηση των τεχνικών κατηγοριοποίησης συναισθήματος που έχουν προταθεί στη βιβλιογραφία με έμφαση στις τεχνικές που εξειδικεύουν τις γενικές μεθόδους για τα ιστολόγια. Στη συνέχεια θα σχεδιάσετε και θα εκπαιδεύσετε έναν κατηγοριοποιητή ο οποίος θα επιτρέπει την αυτόματη αναγνώριση συναισθήματος (θετικό ή αρνητικό) από blogs, χρησιμοποιώντας χαρακτηριστικά που έχουν προταθεί από τη βιβλιογραφία, ενώ θα πραγματοποιηθεί πειραματική μέτρηση της απόδοσης του προτεινόμενου κατηγοριοποιητή και θα παρουσιαστούν στην διπλωματική τα αποτελέσματα.

Ενδεικτική Βιβλιογραφία:

- http://en.wikipedia.org/wiki/Weblogs
- http://dl.acm.org/citation.cfm?id=1013455
- https://www.era.lib.ed.ac.uk/handle/1842/1113
- http://dare.uva.nl/document/46517
- http://en.wikipedia.org/wiki/Sentiment_analysis
- http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf
- http://researcher.watson.ibm.com/researcher/files/us-kclang/pooling-multinomials-kdd09.pdf

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.

Υπεύθυνη Διδακτορικός: Κουμπούρη Αθανασία (koumpour@ceid.upatras.gr)

Θέμα 5

Κατηγοριοποίηση ιστοσελίδων με χρήση του URL: μελέτη και πειραματική εφαρμογή θεματικής κατηγοριοποίησης.

Το Uniform Resource Locator (URL) ενός διαδικτυακού πόρου αποτελεί τη διεύθυνσή του στο Παγκόσμιο Διαδίκτυο. Για έναν έμπειρο χρήστη το URL μπορεί να αποτελέσει ισχυρή ένδειξη που αφορά τόσο στο είδος του διαδικτυακού πόρου, όσο και στο θέμα ή τη λειτουργικότητά του. Το URL λόγω της φύσης του είναι πληροφορία συμπυκνωμένη και χαρακτηριστική του πόρου τον οποίο περιγράφει, ενώ συνήθως είναι αναγνώσιμο από άνθρωπο υπό την έννοια ότι περιέχει όρους κατανοητούς οι οποίοι φέρουν πληροφορία.

Το ερευνητικό ενδιαφέρον για το URL ενισχύεται από το γεγονός ότι σε συνηθισμένες συνθήκες επεξεργασίας είναι η πρώτη πληροφορία που λαμβάνουμε για μια ιστοσελίδα. Συνεπώς, αν το URL αποδειχτεί ικανό να αντιπροσωπεύσει με ακρίβεια τη σελίδα σ' ό,τι αφορά στο περιεχόμενό της ή σε άλλα χαρακτηριστικά της, αρκετές εργασίες όπως πχ η κατηγοριοποίηση ιστοσελίδων μπορούν να πραγματοποιηθούν αποδοτικά χωρίς επεξεργασία μεγάλου όγκου δεδομένων. Η χρήση του URL ως αντιπροσωπευτική πληροφορία μπορεί να συντελέσει στο να εκτελεστούν συγκεκριμένου τύπου εργασίες χωρίς να είναι απαραίτητο το κατέβασμα του περιεχομένου της ιστοσελίδας ή στις περιπτώσεις που αυτό δεν είναι διαθέσιμο, όπως πχ όταν το σύνολο της ιστοσελίδας είναι εικόνα. Αρκετές μέθοδοι έχουν προταθεί στη βιβλιογραφία για την επεξεργασία και την εξαγωγή χρήσιμης πληροφορίας από το URL, ενώ μια βασική πρόκληση στην εφαρμογή τους αποτελεί η αναγκαιότητα για κατάτμηση του URL σε όρους και η αναγνώριση λέξεων σε αυτό.

Στα πλαίσια της διπλωματικής θα κληθείτε να πραγματοποιήσετε μια βιβλιογραφική έρευνα πάνω στη χρησιμότητα του URL στο πεδίο της αυτόματης κατηγοριοποίησης ιστοσελίδων. Η έρευνα θα στοχεύει στην διερεύνηση της βιβλιογραφίας και στην καταγραφή των μεθόδων που χρησιμοποιούν μόνο το URL ή το URL σε συνδυασμό με άλλο τύπο πληροφορίας για να εξάγουν χαρακτηριστικά κατηγοριοποίησης. Επίσης θα περιλαμβάνει τις βασικές τεχνικές επεξεργασίας του URL που έχουν χρησιμοποιηθεί από τις μεθόδους που προτείνονται. Σημαντικό κομμάτι της παρούσας διπλωματικής αποτελεί η μελέτη της χρησιμότητας του URL για τη θεματική κατηγοριοποίηση, για την οποία θα υλοποιηθεί πειραματική εφαρμογή που θα επιχειρεί να επιδείξει τα βασικά συμπεράσματα των σχετικών ερευνών.

Ενδεικτική Βιβλιογραφία:

- http://dl.acm.org/citation.cfm?id=1013426
- http://dl.acm.org/citation.cfm?id=1099649
- http://www2009.org/proceedings/pdf/p1109.pdf

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Τζέκου Βιβή
Email επικοινωνίας: tzekou@ceid.upatras.gr

Θέμα 6

Ανάλυση χαρακτηριστικών της γλώσσας στα μέσα κοινωνικής δικτύωσης: Μελέτη και πειραματική μέτρηση ενδεικτικών γλωσσολογικών δεικτών.

Τα μέσα κοινωνικής δικτύωσης έχουν γίνει τα τελευταία χρόνια μέρος της καθημερινής ζωής της συντριπτικής πλειοψηφίας των χρηστών του Παγκόσμιου Ιστού. Η ερευνητική κοινότητα έχει δείξει μεγάλο ενδιαφέρον για το φαινόμενο, με έρευνες σε αρκετά πεδία να επιχειρούν να μελετήσουν τόσο τη δομή και τα χαρακτηριστικά όσο και τον κοινωνικό τους ρόλο. Στο πεδίο της επεξεργασίας φυσικής γλώσσας το ερευνητικό ενδιαφέρον είναι εξίσου μεγάλο, εφόσον για πρώτη φορά οι χρήστες του Παγκόσμιου Ιστού είναι σε θέση να συνδιαμορφώνουν σε τέτοια έκταση το περιεχόμενό του συνεισφέροντας με κείμενο και δεδομένα.

Η τεράστια αύξηση του περιεχομένου που παράγεται από χρήστες καθώς και οι δυνατότητες για αλληλεπίδραση μεταξύ των χρηστών δημιουργούν ιδιαίτερες συνθήκες, οι οποίες επιδρούν επίσης στη μορφή και τα χαρακτηριστικά της γλώσσας που χρησιμοποιείται στα online κειμενικά δεδομένα. Χαρακτηριστικό παράδειγμα αποτελούν τα σχόλια χρηστών, τα οποία σαν είδος λόγου εντοπίζονται ανάμεσα στον προφορικό και τον γραπτό λόγο, υιοθετώντας χαρακτηριστικά και από τα δύο είδη. Οι ιδιαιτερότητες του κειμένου που παράγεται από χρήστες με στόχο τη δημοσίευση στον παγκόσμιο ιστό παρουσιάζουν μεγάλο ενδιαφέρον, αφενός λόγω των προκλήσεων που παρουσιάζει η μελέτη τους και αφετέρου λόγω του ευρύτατου φάσματος εφαρμογής των τεχνικών ανάλυσής τους.

Στην παρούσα διπλωματική θα κληθείτε να πραγματοποιήσετε αρχικά μια βιβλιογραφική μελέτη των ερευνητικών εξελίξεων σε σχέση με το περιεχόμενο των μέσων κοινωνικής δικτύωσης, ιδωμένο από την οπτική γωνία της γλωσσολογικής ανάλυσής του. Στη συνέχεια θα συλλέξετε σώμα κειμένων από μέσα κοινωνικής δικτύωσης, το οποίο θα επεξεργαστείτε και θα χρησιμοποιήσετε για να επιβεβαιώσετε πειραματικά τους κυριότερους από τους γλωσσολογικούς δείκτες που αναφέρονται στη βιβλιογραφία.

Ενδεικτική Βιβλιογραφία:

- http://www.mpi-sws.org/~cristian/LASM_2013_files/LASM/index.html

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Τζέκου Βιβή
Email επικοινωνίας: tzekou@ceid.upatras.gr

Θέμα 7

Γλωσσολογικοί πόροι για τεχνικές αυτόματης αντιστοίχισης(μετάφρασης) παράλληλων κειμένων.

Ένα σημαντικό πεδίο που απασχολεί την ερευνητική κοινότητα παγκοσμίως είναι η μετάφραση μεγάλων όγκων δεδομένων στις διάφορες φυσικές γλώσσες. Για αυτόν το σκοπό έχουν προταθεί πολλές τεχνικές και συστήματα. Σε αυτόν το σκοπό έρχεται να προστεθεί η αντιστοίχιση παράλληλου κειμένου, το οποίο είναι ένα σύνολο κειμένων του ίδιου περιεχομένου σε διαφορετικές γλώσσες. Με άλλα λόγια το ένα κείμενο (source language) είναι η μετάφραση του άλλου (target language). Σε ένα bitext (παράλληλο κείμενο σε δύο γλώσσες) μπορεί με διάφορες τεχνικές να γίνει αντιστοίχιση είτε σε προτασιακό επίπεδο είτε σε επίπεδο όρων της πρότασης και φράσεων. Μια τέτοια διαδικασία είναι ιδιαίτερα σημαντική και χρήσιμη καθώς μπορεί να βελτιώσει την απόδοση της μετάφρασης ανά ζευγάρια γλωσσών και να εμπλουτίσει τα δεδομένα και τους υπάρχοντες πόρους για τις γλώσσες αυτές.

Η παρούσα διπλωματική προτείνει την επισκόπηση της βιβλιογραφικής έρευνας που έχει πραγματοποιηθεί στο πεδίο, και την αξιολόγηση τεχνικών κι εργαλείων αυτόματης αντιστοίχισης παράλληλου κειμένου. Εν συνεχεία, θα προχωρήσουμε με την έρευνα διαθέσιμων γλωσσολογικών πόρων παράλληλων κειμένων και την επεξεργασία τους ώστε να είναι δυνατή η μετέπειτα χρήση τους. Θα χρησιμοποιήσουμε αυτά τα δεδομένα για την αξιολόγηση τεχνικών αντιστοίχισης με σκοπό να συγκρίνουμε την απόδοσή των συστημάτων αυτών ανάλογα με το ζευγάρι γλωσσών και με τις δυνατότητες που μας προσφέρουν. Ένα ακόμα βήμα θα μπορούσε να είναι, αναλόγως των αποτελεσμάτων που θα προκύψουν, η πρόταση κανόνων που θα στοχεύει στην βελτίωση της απόδοσης της αντιστοίχισης.

Ενδεικτική Βιβλιογραφία:

- http://en.wikipedia.org/wiki/Parallel_text_alignment
- http://en.wikipedia.org/wiki/Bitext_word_alignment
- Veronis, J. 2000. From the Rosetta stone to the information society. 1–24

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Σιμάκη Βασιλική
Email επικοινωνίας: simaki@ceid.upatras.gr

Θέμα 8

Μελέτη τεχνικών ταξινόμησης συναισθήματος και πειραματική αξιολόγηση με έμφαση στην χρήση emoticons.

Τα τελευταία χρόνια παρατηρείται ένας αυξανόμενος αριθμός ερευνών στον τομέα της κατανόησης του συναισθήματος (sentiment) σε γραπτές πηγές κειμένων. Μια υποκατηγορία αυτού του τομέα είναι η ταξινόμηση συναισθήματος (sentiment classification). Δοθέντος ενός κειμένου είναι εφικτό με αυτοματοποιημένες μεθόδους να αποφανθούμε εάν το κείμενο έχει θετικό προσανατολισμό ή αν έχει αρνητικό? Είναι ένα περίπλοκο και δύσκολο έργο για να επιτευχθεί, εάν λάβουμε υπόψη μας τις δυσκολίες που απαντώνται κατά την ανάπτυξη μιας μεθόδου για να αναγνωρίζει για παράδειγμα τον σαρκασμό. Έχει αποδειχθεί πως οι παραδοσιακές προσεγγίσεις ταξινόμησης κειμένου (text classification) μπορεί να είναι αρκετά αποτελεσματικές όταν εφαρμόζονται στο πρόβλημα της ανάλυσης συναισθήματος. Μοντέλα όπως Naïve Bayes (NB), Maximum Entropy (ME) και Support Vector Machines (SVM) μπορούν να προσδιορίσουν το συναίσθημα των κειμένων με υψηλή ακρίβεια.

Πρόσφατες έρευνες έχουν δείξει ότι τα emoticons μπορούν να χρησιμοποιηθούν ως δείκτες συναισθήματος επειδή εκφράζουν συναισθηματική κατάσταση. Παρόλα αυτά η χρήση των emoticons ποικίλλει σε μεγάλο βαθμό μεταξύ των χρηστών. Μερικοί χρήστες χρησιμοποιούν τα emoticons πιο συχνά και εκφράζουν αδύναμο συναίσθημα, ενώ άλλοι με πολύ μικρότερη συχνότητα χρήσης emoticons εκφράζουν έντονο συναίσθημα. Έτσι γεννάται το ερώτημα σε τι βαθμό μπορούν να αξιοποιηθούν τα emoticons στην ταξινόμηση συναισθήματος.

Στην παρούσα διπλωματική καλείστε να πραγματοποιήσετε μια επισκόπηση των μεθόδων ταξινόμησης συναισθήματος που έχουν προταθεί καθώς και των εφαρμογών τους. Επιπλέον θα πραγματοποιηθεί συγκριτική μελέτη των κυριότερων μεθόδων αξιολόγησης της ταξινόμησης συναισθήματος. Τέλος, καλείστε να υλοποιήσετε μέθοδο αξιολόγησης για το εάν και κατά πόσο η χρήση emoticons μπορεί να χρησιμοποιηθεί αποτελεσματικά/βοηθητικά στην ταξινόμηση συναισθήματος.

Ενδεικτική Βιβλιογραφία:

- http://en.wikipedia.org/wiki/Sentiment_analysis
- http://www.mn.uio.no/ifi/english/people/aca/jread/Read2005.pdf
- http://pranjalv.com/sentiment/

Προτεινόμενα Μαθήματα:

- Γλωσσική Τεχνολογία

Επιβλέπων: Χριστοδουλάκης Δ.
Υπεύθυνη Διδακτορικός: Κουμπούρη Αθανασία (koumpour@ceid.upatras.gr)