Χειρισμός ερωτήσεων ορισμού σε συστήματα ερωταποκρίσεων για συλλογές εγγράφων

Ίων Ανδρουτσόπουλος, Δημήτρης Γαλάνης και Ίρις Μηλιαράκη

Περίληψη:

Τα συστήματα ερωταποκρίσεων φυσικής γλώσσας για συλλογές εγγράφων φιλοδοξούν να επαυξήσουν τις δυνατότητες των σημερινών μηχανών αναζήτησης και ανάκτησης πληροφοριών. Επιτρέπουν στους χρήστες τους να θέτουν ερωτήσεις σε φυσική γλώσσα (π.χ. «Ποιος είναι ο πρωθυπουργός της Ινδονησίας;», «Πόσα χρήματα ξόδεψε για διαφήμιση η ΔΕΗ το 2002;»), στις οποίες επιχειρούν να αποκριθούν με ακριβείς απαντήσεις (π.χ. ένα όνομα προσώπου, ένα χρηματικό ποσό) ή αποσπάσματα κειμένων που εξάγονται από τα διαθέσιμα έγγραφα μιας συλλογής ή τον Παγκόσμιο Ιστό. Η ομιλία εστιάζεται στο χειρισμό ερωτήσεων ορισμού (π.χ. «Τι είναι η θαλασσαιμία;», «Ποιος ήταν ο Βολταίρος;»), μια ιδιαίτερα συχνή κατηγορία ερωτήσεων που πολλά συστήματα ερωταποκρίσεων δυσκολεύονται να χειριστούν. Στην πρώτη ενότητα της ομιλίας θα παρουσιαστεί μια μέθοδος ευρέσεως απαντήσεων σε ερωτήσεις ορισμού, η οποία επεκτείνει με τεχνικές επιβλεπόμενης μηχανικής μάθησης και αυτόματα παραγόμενα πρότυπα προηγούμενες προσεγγίσεις που βασίζονταν στη χρήση του WordNet και χειρωνακτικά κατασκευασμένων προτύπων. Η μέθοδος αυτή αξιολογήθηκε με ερωτήσεις και συλλογές εγγράφων των διαγωνισμών TREC, με σημαντικά καλύτερα αποτελέσματα από εκείνα των προηγούμενων προσεγγίσεων. Στη δεύτερη ενότητα της ομιλίας θα παρουσιαστεί μια τεχνική που παράγει με αυτόματο τρόπο παραδείγματα εκπαίδευσης για τη μέθοδο της πρώτης ενότητας, αξιοποιώντας ηλεκτρονικές εγκυκλοπαίδειες. Με την τεχνική αυτή η μέθοδος της πρώτης ενότητας μετατρέπεται ουσιαστικά σε μη επιβλεπόμενη. Πειραματικά αποτελέσματα δείχνουν πως η μη επιβλεπόμενη μορφή της μεθόδου είναι καταλληλότερη, όταν χρησιμοποιείται ως συμπλήρωμα μηχανών αναζήτησης του Παγκοσμίου Ιστού.