Μνημοσύνη

Δευ, 18/10/2010 - 16:23 — webmaster

«Μνημοσύνη»: Περιβάλλον Επεξεργασίας Συλλογών Κειμένων

Το **ΜΝΗΜΟΣΥΝΗ** αποτελεί ένα ολοκληρωμένο σύστημα επεξεργασίας φυσικής γλώσσας, το οποίο ενσωματώνει υψηλής ποιότητας γλωσσικούς πόρους και υπολογιστικά εργαλεία με στόχο την αυτόματη εξαγωγή δομημένης πληροφορίας από μη δομημένα ηλεκτρονικά έγγραφα. Χρησιμοποιείται κυρίως για την αυτόματη επεξεργασία ελεύθερων κειμενικών εγγράφων. Διασφαλίζει:

1. την επεξεργασία μεγάλου όγκου πληροφοριών,
2. υψηλή ακρίβεια στην αναγνώριση των ονοματικών οντοτήτων (Named Entities) και γεγονότων (Events),
3. δυνατότητα προσθήκης νέων πηγών δεδομένων με χαμηλό κόστος.

Κύρια χαρακτηριστικά

Δεδομένα εισόδου: Δυνατότητα επεξεργασίας συλλογών κειμένων, διαφόρων μορφών (HTML, PDF, TXT) και αποθηκευμένων σε ποικίλα μέσα (αρχεία, σελίδες ιστοτόπων, βάσεις δεδομένων).

Γλωσσικοί πόροι: Το σύστημα χρησιμοποιεί ποικίλους γλωσσικούς πόρους, όπως λεξιλόγια διαφόρων γλωσσών, ορθογραφικά λεξικά, μορφολογικά λεξικά, ορολογικά λεξικά, θησαυρούς κ.ά.

Επισημειώσεις: Επιτρέπει διαφορετικές αναλύσεις ενός κειμένου, οι οποίες εκφράζονται ως επισημειώσεις διαφορετικών επιπέδων. Η ύπαρξη διαφορετικών επιπέδων επισημειώσεων παρέχει μεγάλη ευελιξία στο σύστημα και δυνατότητα διασύνδεσης μεταξύ των αναλύσεων.

Αναλυτές και ροή παραγωγής: Οι αναλυτές αποτελούν τους μηχανισμούς παραγωγής των αναλύσεων. Συνδέονται μεταξύ τους μέσω διαφορετικών ροών παραγωγής, κατά τέτοιο τρόπο ώστε το αποτέλεσμα επεξεργασίας ενός αναλυτή να μπορεί να αποτελεί είσοδο για τον επόμενο. Οι ροές λειτουργούν παράλληλα επιτυγχάνοντας μεγαλύτερη ταχύτητα στην επεξεργασία.

Γλώσσα περιγραφής κανόνων: Η γλώσσα περιγραφής «Κανών» επιτρέπει τη δημιουργία σημασιολογικών επισημειώσεων που βασίζονται στον προσδιορισμό συντακτικών κανόνων.

Φιλτράρισμα: Τα αποτελέσματα μπορούν να φιλτράρονται προτού δοθούν προς επεξεργασία σε άλλους αναλυτές. Με τον τρόπο αυτό, εξασφαλίζεται η ελαχιστοποίηση της πληροφορίας που μεταφέρεται στην επόμενη φάση, απλοποιώντας κατά συνέπεια τους κανόνες των επόμενων αναλυτών.

Μηχανισμοί ασαφούς ταιριάσματος: Το σύστημα, αφού αναγνωρίσει μια ονοματισμένη οντότητα (π.χ. πρόσωπο, οργανισμό, εταιρεία κτλ.), παρέχει τη δυνατότητα ταιριάσματος της οντότητας αυτής με τις οντότητες που είναι καταχωρισμένες σε μια υπάρχουσα βάση δεδομένων. Οι μηχανισμοί που παρέχονται χωρίζονται σε δύο κατηγορίες: στους λεξικογραφικούς, που χρησιμοποιούν τεχνικές ορθογραφικής διόρθωσης (π.χ. αποστάσεις λεκτικών), και στους στατιστικούς, που μετράνε την ομοιότητα του λεκτικού με τα λεκτικά της βάσης λαμβάνοντας υπόψη τον αριθμό και τη βαρύτητα των τριγραμμάτων (ή qγραμμάτων).

Εξαγωγείς: Εξειδικευμένοι αναλυτές αναλαμβάνουν τη μεταφορά της εξαγόμενης πληροφορίας στους επιθυμητούς προορισμούς και μορφές (π.χ. XML, Database tables κ.ά.).

Ειδικός μηχανισμός εποπτείας: Επιτρέπει την παρακολούθηση της όλης διαδικασίας, καθώς και την καταγραφή της σε διαφορετικά μέσα (αρχεία, βάση δεδομένων).

Έλεγχος εξαγόμενης πληροφορίας: Παρέχεται η δυνατότητα παρακολούθησης της εξαγόμενης πληροφορίας σε κάθε στάδιο της επεξεργασίας. Ο μηχανισμός αυτός επιτρέπει την αποσφαλματοποίηση της διαδικασίας βήμα προς βήμα.

Παραθυρική εφαρμογή για την επιβεβαίωση, διόρθωση και συμπλήρωση της εξαγόμενης πληροφορίας. Ο χρήστης της εφαρμογής μπορεί με εύχρηστο και ευέλικτο τρόπο να προσθέτει, να διορθώνει και να καταργεί επισημειώσεις της αυτόματης διαδικασίας, μεγιστοποιώντας έτσι την ποιότητα και την ακρίβεια της πληροφορίας.

Εφαρμογή ιστού για την παρουσίαση των αποτελεσμάτων με ποικίλους τρόπους, όπως έγχρωμες επισημειώσεις πάνω στο κείμενο, πίνακες ανά κατηγορία πληροφορίας, φιλτράρισμα σε όλα τα χαρακτηριστικά της πληροφορίας, συγκεντρωτικά αποτελέσματα και συγκρίσεις με αποτελέσματα άλλων μεθόδων.

Source URL: http://www.neurolingo.gr/el/technology/mnemosyne.html