ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΓΝΩΡΙΣΗ ΕΙΔΟΥΣ ΚΕΙΜΕΝΟΥ ΚΑΙ ΣΥΓΓΡΑΦΕΑ ΣΕ ΝΕΟΕΛΛΗΝΙΚΑ ΚΕΙΜΕΝΑ ΧΩΡΙΣ ΠΕΡΙΟΡΙΣΜΟΥΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΕΥΣΤΑΘΙΟΥ Σ. ΣΤΑΜΑΤΑΤΟΥ

ΔΙΠΛ. ΗΛΕΚΤΡΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΡΙΘΜΟΣ ΔΙΑΤΡΙΒΗΣ 87

ΑΠΡΙΛΙΟΣ 2000

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Συντομογραφιών

ΕΒΜ: Εκμάθηση Βάσει Μετασχηματισμών

ΕΛ: Επόμενη Λέξη

ΕΣΣ: Επόμενα Σημεία Στίξης

ΕΦ: Επιρρηματικές Φράσεις

ΟΦ: Ονοματικές Φράσεις

ΠΛ: Προηγούμενη Λέξη

ΠΣΣ: Προηγούμενα Σημεία Στίξης

ΠΦ: Προθετικές Φράσεις

ΡΦ: Ρηματικές Φράσεις

ΣΥΟΠ: Σύνολο Υποψηφίων Ορίων Περιόδων

ΣΦ: Συνδετικές Φράσεις

Περίληψη

Αντικείμενο αυτής της διατριβής είναι η στατιστική ανάλυση του ύφους Νεοελληνικών κειμένων χωρίς περιορισμούς, με στόχο την αυτόματη ταξινόμησή τους τόσο ως προς το είδος τους όσο και ως προς τον συγγραφέα τους. Η υφολογική πληροφορία εξάγεται μέσω της ανάλυσης του κειμένου από ένα υπολογιστικό εργαλείο ικανό να ανιχνεύει τα όρια των περιόδων και των φράσεων σε οποιοδήποτε κείμενο. Ο ανιχνευτής ορίων περιόδων και φράσεων επιτυγχάνει πολύ ικανοποιητικά αποτελέσματα ανάλυσης, αν και βασίζεται σε ελάχιστους πόρους. Πιο συγκεκριμένα, ο ανιχνευτής περιόδων βασίζεται σε πολύ απλή πληροφορία (όπως το μήκος των λέξεων) και σε κανόνες που εξάγονται αυτόματα από ένα σώμα εκπαίδευσης, σύμφωνα με μία νέα προσέγγιση μηχανικής εκμάθησης. Η ακρίβεια που επιτυγχάνει είναι της τάξης του 99,4% για ένα σώμα κειμένων 200.000 λέξεων. Απ' την άλλη, ο ανιχνευτής ορίων φράσεων βασίζεται σε ένα μικρό σύνολο λέξεων-κλειδιών και στις καταλήξεις των λέξεων για να εκτιμήσει την πιο πιθανή μορφολογική περιγραφή της κάθε λέξης. Η διαδικασία της ανάλυσης γίνεται μέσω της τεχνικής πολλαπλού περάσματος και στο σώμα που αναφέρθηκε παραπάνω επιτυγχάνει ανάκληση και ακρίβεια της τάξης του 90% και 95% αντίστοιχα. Έτσι, το εργαλείο αυτό προσφέρει γρήγορη και αξιόπιστη ανάλυση μεγάλων όγκων κειμένου, με ελάχιστο υπολογιστικό κόστος. Για την αναπαράσταση του ύφους χρησιμοποιούνται 22 υφολογικοί δείκτες που διακρίνονται σε τρία υφομετρικά επίπεδα: (i) *επίπεδο δείγματος*, που σχετίζεται με την έξοδο του ανιχνευτή ορίων περιόδων, (ii) *επίπεδο φράσης*, που σχετίζεται με την έξοδο του ανιχνευτή ορίων φράσεων και (iii) *επίπεδο ανάλυσης*, που αφορά στον τρόπο με τον οποίο έγινε η ανάλυση του κειμένου από τον ανιχνευτή ορίων φράσεων. Το τελευταίο επίπεδο αποτελεί ένα εναλλακτικό τρόπο σύλληψης της υφολογικής πληροφορίας και είναι η πρώτη φορά που χρησιμοποιείται. Επιπλέον, δεν

χρησιμοποιείται καμία λεξιλογική πληροφορία σε αντίθεση με τις προηγούμενες προσεγγίσεις. Αυτό το σύνολο των υφολογικών δεικτών σε συνδυασμό με τις στατιστικές τεχνικές της *πολλαπλής παλινδρόμησης* και της *διαχωριστικής ανάλυσης* οδηγούν στην αυτόματη αναγνώριση του είδους κειμένου (π.χ. επιστημονικά άρθρα, ρεπορτάζ εφημερίδων, μαγειρικές συνταγές, κ.ά.). Αυτές οι τεχνικές της πολυπαραγοντικής στατιστικής διακρίνονται για την εύκολη εκπαίδευσή τους και τη γρήγορη απόκρισή τους καθώς βασίζονται στον υπολογισμό απλών γραμμικών συναρτήσεων. Τα πειράματα ελέγχου της προτεινόμενης μεθοδολογίας πραγματοποιήθηκαν σε ένα σώμα που δημιουργήθηκε από κείμενα που βρέθηκαν σε σελίδες του Διαδικτύου και δεν υπέστησαν καμία χειρονακτική προεπεξεργασία. Τα αποτελέσματα κρίνονται ως πολύ ικανοποιητικά καθώς επιτυγχάνεται ακρίβεια της τάξης του 82-85%, που είναι πολύ υψηλότερη από αντίστοιχα συστήματα της Αγγλικής γλώσσας. Αντίστοιχα πειράματα πραγματοποιήθηκαν για την αυτόματη ταξινόμηση ενός σώματος κειμένων ως προς το συγγραφέα. Η προτεινόμενη μεθοδολογία εφαρμόστηκε σε δύο ομάδες τυχαία επιλεγμένων συγγραφέων μίας εβδομαδιαίας εφημερίδας. Η ακρίβεια που επιτυγχάνεται από το σύστημα αναγνώρισης συγγραφέα (αξιολόγηση κλειστού συνόλου) ήταν στην καλύτερη περίπτωση 81%, σαφώς καλύτερη από την αντίστοιχη απόδοση της πιο σύγχρονης λεξιλογικής μεθόδου (74%). Ωστόσο, τα καλύτερα αποτελέσματα επιτεύχθηκαν από τον συνδυασμό των δύο προσεγγίσεων (87%). Επίσης, παρατηρήθηκε ότι τόσο το μήκος του κείμενου όσο και το μέγεθος του σώματος εκπαίδευσης παίζουν πολύ σημαντικό ρόλο για την αξιόπιστη αναπαράσταση του προσωπικού ύφους ενός συγγραφέα. Εκτός από την αναγνώριση συγγραφέα, πραγματοποιήθηκαν και πειράματα επιβεβαίωσης συγγραφέα καθώς και στατιστικοί έλεγχοι της σημαντικότητας των προτεινόμενων υφολογικών δεικτών.

STATISTICAL IDENTIFICATION OF GENRE AND AUTHOR IN UNRESTRICTED MODERN GREEK TEXT

Summary

This dissertation deals with the statistical analysis of the style of unrestricted Modern Greek text, aiming at the automatic text categorization in terms of genre and author. The stylistic information is extracted via the analysis of the input text by a computational tool able to detect sentence and chunk boundaries in unrestricted text. Although the sentence and chunk boundaries detector is based on minimal resources, it achieves very high accuracy results. In more detail, the sentence boundary detector is based on simple metrics (such as word-length) and disambiguation rules that are extracted automatically from a training corpus, according to a new machine learning methodology. This tool achieves 99.4% accuracy for a 200,000 word corpus. On the other hand, the chunk boundary detector is based on a small lexicon of keywords and common word-suffixes in order to assign the most likely morphological description to every word. Multiple-pass parsing is, then, applied and the recall and precision results on the aforementioned corpus are 90% and 95% respectively. This tool, therefore, provides a reliable solution for the rapid analysis of large volumes of text, requiring minimal computational cost. A set of 22 style markers is used for the representation of the style and they may be distinguished in three stylometric levels: (i) *token-level*, that concerns the output of the sentence boundary detector, (ii) *phrase-level*, that deals with the output of the chunk boundary detector, and (iii) *analysis-level*, that concerns the way in which the text has been analyzed by the chunk boundary detector. The latter may be considered as an alternative way for capturing the stylistic information and is used for first time. Additionally, in contrast to other approaches, lexically-based measures are excluded. This set of style markers in combination with the statistical techniques of *multiple regression* and *discriminant analysis* lead to the automatic detection of text-genre (e.g. academic prose, reportage, recipes, etc.). These

techniques of multivariate statistics are characterized by minimal training and response time cost since they are based on the calculation of simple linear functions. Experiments for testing the proposed methodology were conducted on a corpus that has been constructed by texts downloaded from the Internet, without any manual preprocessing. The classification results are quite satisfying since the proposed approach achieves 82-85% accuracy, which is far higher than the ones that have been referred for other systems (proposed for English). Similar experiments were conducted for the automatic categorization of a corpus in terms of author. The proposed methodology was applied to two groups of randomly-selected authors of a weekly newspaper. The classification accuracy achieved by the author identification system (closed set evaluation) is, in the best case, 81%, which is considerably higher than the corresponding performance of the most modern lexically-based approach (74%). However, the best results are achieved by the combination of these two approaches (87%). Moreover, the text-length and the size of the training corpus play an important role for the reliable representation of the personal style of a particular author. In addition to the author identification, we conducted experiments on author verification as well as statistical tests of importance of the proposed set of style markers.