ΕΛΕΤΟ – 1° Συνέδριο «Ελληνική Γλώσσα και Ορολογία»
Αθήνα, 30, 31 Οκτωβρίου και 1 Νοεμβρίου 1997

ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΟΡΩΝ ΜΕ ΧΡΗΣΗ ΓΡΑΜΜΑΤΙΚΗΣ ΠΡΟΤΥΠΩΝ

Βύρων Γεωργαντόπουλος, Στέλιος Πιπερίδης

Περίληψη

Στο άρθρο αυτό παρουσιάζονται τα πρώτα αποτελέσματα μιας μεθόδου αυτόματης εξαγωγής όρων από σώματα κειμένων. Η μέθοδος στηρίζεται στην εφαρμογή μιας γραμματικής προτύπων που χρησιμοποιεί το φορμαλισμό ενοποίησης (feature-structure unification) και τελεστές κανονικών εκφράσεων-γραμματικών (regular expressions). Το σώμα κειμένων που χρησιμοποιήθηκε είναι ένα εγχειρίδιο οδηγιών της Hewlett-Packard μεγέθους περίπου 90000 λέξεων που περιελάμβανε έναν κατάλογο όρων έναντι του οποίου αξιολογήθηκαν τα αποτελέσματα της μεθόδου. Η μέθοδος εξήγαγε 124 από τους 214 όρους που είχαν εξαχθεί χειρωνακτικά, παρουσιάζοντας ποσοστό ανάκτησης (recall) 58%.

AUTOMATIC TERM EXTRACTION BASED ON PATTERN GRAMMARS

Byron Georgantopoulos, Stelios Piperidis

Abstract

In this paper, we present a method for the automatic extraction of terms from machine-readable text corpora. The method is based on a pattern grammar endowed with regular expressions and feature-structure unification capacity. The text corpus we have used consisted of a software manual by Hewlett-Packard extending to around 90000 wordforms, containing a term index against which the results of the method were evaluated. The method extracted 124 out of 214 manually coded terms, featuring a 58% recall.

ΑΥΤΟΜΑΤΗ ΕΞΑΓΩΓΗ ΟΡΩΝ ΜΕ ΧΡΗΣΗ ΓΡΑΜΜΑΤΙΚΗΣ ΠΡΟΤΥΠΩΝ

Βύρων Γεωργαντόπουλος, Στέλιος Πιπερίδης

1. Εισαγωγή

Στο άρθρο αυτό παρουσιάζονται τα πρώτα αποτελέσματα μιας μεθόδου αυτόματης εξαγωγής όρων από σώματα κειμένων. Η αυτόματη εξαγωγή όρων αποκτά ιδιαίτερο ενδιαφέρον σήμερα που μεγάλοι όγκοι κειμένων παράγονται πλέον ηλεκτρονικά, γεγονός που όδηγεί στην διατύπωση νέων απαιτήσεων για την διαχείριση και επεξεργασία τους (αυτόματη ταξινόμηση, ανάκτηση πληροφοριών, κλπ). Η εφαρμογή συστημάτων γλωσσικής τεχνολογίας για την ικανοποίηση των αναγκών αυτών απαιτεί την προσαρμογή (customisation) του συστήματος στην θεματική περιοχή, το γνωστικό πεδίο, των προς επεξεργασία κειμένων. Βασικό βήμα στην διαδικασία αυτή αποτελεί η βελτίωση και ο εμπλουτισμός των γλωσσικών πόρων (language resources) με την ενσωμάτωση της κατάλληλης ορολογίας. Η εφαρμογή μεθόδων αυτόματης εξαγωγής όρων προσφέρει μια έγκυρη, γρήγορη και χαμηλού κόστους λύση στην διαδικασία προσαρμογής.

Η εξαγωγή όρων βρίσκει πολλές εφαρμογές στο χώρο της επεξεργασίας φυσικής γλώσσας και ειδικά με τον διαρκώς αυξανόμενο όγκο ηλεκτρονικών κειμένων σήμερα:

- **δεικτοδότηση κειμένων (text indexing)** - οι εξαγόμενοι όροι χρησιμοποιούνται απευθείας στον κατάλογο όρων του κειμένου
- **κατηγοριοποίηση-ταξινόμηση κειμένων (text classification)** - κείμενα με παρόμοιους όρους ταξινομούνται στην ίδια θεματική περιοχή
- **ανάκτηση/εξαγωγή πληροφορίας (information retrieval/extraction)** - ο χρήστης αναζητά κείμενα που τον ενδιαφέρουν με τη μορφή ερωτήσεων αποτελούμενων από όρους-κλειδιά. Από όλα τα διαθέσιμα κείμενα επιστρέφονται μόνο αυτά που περιέχουν αυτούς τους συγκεκριμένους όρους
- **κατασκευή περίληψης (text abstracting/summarisation)** - οι προτάσεις που περιέχουν όρους του κειμένου είναι κατά κανόνα και οι σημαντικότερες προτάσεις, αυτές που υποδηλώνουν σαφέστερα το περιεχόμενό του.

- **παραλληλοποίηση κειμένων (text alignment)** - όροι της μιας γλώσσας αντιστοιχούν συνήθως σε έναν μόνο όρο μιας άλλης γλώσσας

2. Μεθοδολογικές προσεγγίσεις

Σαν όρους ενός κειμένου ορίζουμε γενικά τις γλωσσικές πραγματώσεις των εννοιών ενός κειμένου. Δύο είναι οι βασικές μεθοδολογικές τάσεις στην εξαγωγή όρων σήμερα:

1. Με χρήση μιας ειδικά σχεδιασμένης γραμματικής όρων (συνήθως ελεύθερης συμφραζομένων), η οποία εφαρμόζεται σε κείμενα κατάλληλα γραμματικά σχολιασμένα και εξάγει όσες φράσεις αναγνωρίζονται από αυτή τη γραμματική [1] .
2. Με χρήση στατιστικών εργαλείων αντίστοιχων με αυτά που χρησιμοποιούνται για εφαρμογές ανάκτησης πληροφοριών και δεικτοδότησης κειμένων. Στα εργαλεία αυτά περιλαμβάνονται μετρήσεις συχνοτήτων, μετρικές από τη θεωρία πληροφορίας, μετρικές που υπολογίζουν τα συμφραζόμενα των λέξεων κ.α.[2], [9]

Αξίζει να σημειωθούν κάποιες διαφορές ανάμεσα στις δύο αυτές μεθόδους. Η γραμματική όρων περιγράφει τη συντακτική δομή που πρέπει να ικανοποιεί κάθε έγκυρος όρος, χωρίς να αποκλείεται το ενδεχόμενο αυτές οι συντακτικές δομές να ικανοποιούνται και από άλλες ακολουθίες λέξεων που δεν θεωρούνται σωστοί όροι. Αν, για παράδειγμα, ένας από τους κανόνες περιγράφει ότι ένα επίθετο και ένα ουσιαστικό συγκροτούν έναν όρο, η εφαρμογή της γραμματικής στην προηγούμενη πρόταση θα επιστρέψει ως αποτέλεσμα τις φράσεις "συντακτικές δομές", "έγκυρος όρος" και "σωστοί όροι". Για τη θεματική κατηγορία του παρόντος κειμένου, ο πρώτος όρος είναι αποδεκτός, ο δεύτερος αποδεκτός σε ευρύτερο πλαίσιο αλλά ο τρίτος όχι. Η αδυναμία της γραμματικής έγκειται στο ότι εφαρμόζει τους κανόνες της χωρίς διάκριση, περιγράφοντας την ικανή αλλά όχι και αναγκαία συνθήκη για να είναι μια ακολουθία λέξεων όρος. Επιπλέον μπορεί να εντοπίσει μόνο όρους με περισσότερες από μία λέξεις, μιας και μόνο σε αυτούς μπορεί να αποδοθεί συντακτική δομή. Συμπερασματικά, ο απώτερος στόχος μιας γραμματικής όρων είναι ο εντοπισμός σε ένα πρώτο στάδιο "υποψήφιων όρων".

Η στατιστική προσέγγιση στηρίζεται στην υπόθεση ότι οι όροι, ως λέξεις ή φράσεις που είναι χαρακτηριστικές της θεματικής περιοχής του κειμένου, έχουν την τάση να εμφανίζονται *συχνά*. Η συχνότητα επιδέχεται δύο διαφορετικές ερμηνείες: (1) συχνότερα από ότι σε ένα κείμενο που δεν ανήκει στη συγκεκριμένη θεματική περιοχη

και (2) απλά συχνότερα από τις άλλες λέξεις ή φράσεις του κειμένου. Με βάση αυτή τη συγκριτική αντίληψη, για κάθε φράση υπολογίζεται ένα βάρος που εκφράζει τη σημασία της για το κείμενο , εξαιρώντας τις γραμματικές λέξεις, άρθρα, αντωνυμίες, προθέσεις κλπ. οι οποίες εμφανίζουν αρκετά υψηλή συχνότητα σε οποιοδήποτε κείμενο αλλά δεν θεωρούνται όροι. Οι φράσεις για τις οποίες υπολογίζεται το μεγαλύτερο βάρος παρουσιάζουν την μεγαλύτερη πιθανότητα να είναι οι όροι του κειμένου. Στα χαρακτηριστικά της προσέγγισης αυτής είναι η δυνατότητα εντοπισμού μονολεκτικών όρων. Στα μειονεκτήματά της καταγράφεται η αδυναμία να εξάγει όρους που δεν ικανοποιούν τα στατιστικά κριτήρια, καθώς είναι πιθανό έγκυροι όροι να εμφανίζονται μόνο μία ή γενικά λίγες φορές στο κείμενο. Τέλος, η επιλογή της στατιστικής φόρμουλας επηρεάζει την αποδοτικότητα της προσέγγισης αυτής, με τρόπο ανάλογο με αυτόν που η καλυπτικότητα της γραμματικής επηρεάζει την προηγούμενη προσέγγιση.

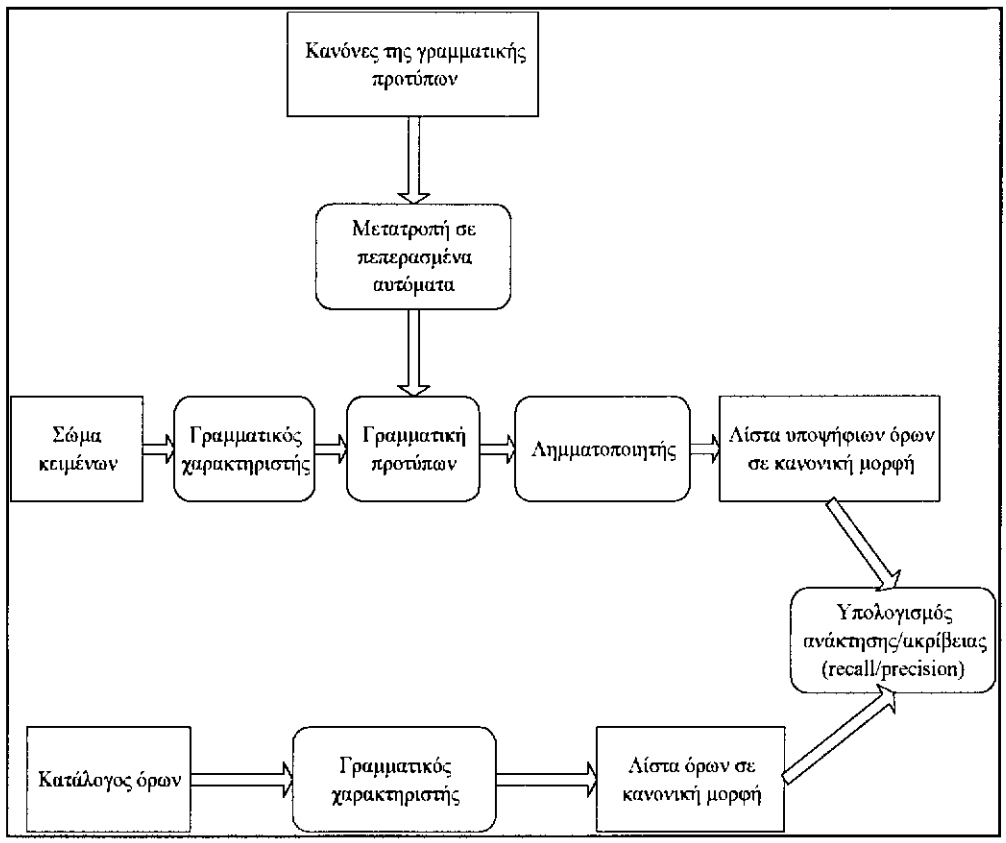
Αλλες προσεγγίσεις συνδυάζουν την στατιστική επεξεργασία με την γλωσσολογική μοντελοποίηση [3], [4], [5], [6]. Πρόκειται για υβριδικά συστήματα που αρχικά δημιουργούν μια λίστα υποψήφιων όρων με τη βοήθεια γραμματικών και στη συνέχεια "φιλτράρουν" αυτούς τους όρους με στατιστικά εργαλεία ώστε να απομακρύνουν τους όρους εκείνους που ικανοποιούν μεν τη γραμματική, αλλά δεν είναι χαρακτηριστικοί της θεματικής περιοχής του κειμένου ώστε να αποτελούν έγκυρους όρους.

3. Περιγραφή της μεθόδου

Η μέθοδος που περιγράφεται στο άρθρο αυτό έχει στόχο την εξαγωγή υποψήφιων όρων, η εγκυρότητα των οποίων θα ελεγχεί χειρωνακτικά. Τα βασικά στάδια της μεθόδου συνίστανται σε :

α. γραμματικό χαρακτηρισμό με βάση ένα μορφολογικό λεξικό και ένα σύστημα κανόνων για επίλυση μορφολογικών αμφισημιών
β. συντακτική ανάλυση με βάση μια γραμματική προτύπων
γ. λημματοποίηση με βάση το μορφολογικό λεξικό και την γραμματική κατηγορία που προκύπτει από τον γραμματικό χαρακτηρισμό.

Το διάγραμμα ροής της μεθόδου απεικονίζεται στο παρακάτω σχήμα:



Η **γραμματική** που χρησιμοποιήθηκε για την συντακτική ανάλυση είναι ένα υποσύνολο της γραμματικής προτύπων που παρουσιάστηκε στο [8]. Πρόκειται για μια γραμματική που χρησιμοποιεί το φορμαλισμό ενοποίησης (feature-structure unification) και τελεστές κανονικών εκφράσεων-γραμματικών (regular expressions). Για παράδειγμα, το πρότυπο που περιγράφει όρους της μορφής ΟΥΣΙΑΣΤΙΚΟ ΠΡΟΘΕΣΗ ΟΥΣΙΑΣΤΙΚΟ έχει την παρακάτω διατύπωση:

```
term_ pattern :   (cat = No
                   term = Tt;Tc),
                  ^(cat = Pn
                  type = Cl),
                  [[ (cat = Pp
                      type = Sp);
                       ^ (cat = At
```

```
        gender = G
        number = N
        case = C) ] ;
    (cat = Pp
     type = Pa
     gender = G
     number = N
     case = C)],
    (cat = No
     term = Tt;Tc
     gender = G
     number = N
     case = C).
```

Το σύμβολο '^' υποδηλώνει προαιρετικότητα (0 ή 1 εμφάνιση) και το σύμβολο ';' είναι ο διαζευκτικός τελεστής. Ο βασικός περιορισμός που εκφράζεται από το παραπάνω πρότυπο είναι η συμφωνία αριθμού, γένους και πτώσης για τα επιμέρους στοιχεία του όρου (ουσιαστικά, άρθρο, κλπ).

Από την γραμματική του [8] που αριθμούσε 77 κανόνες κωδικοποιήθηκε ένα υποσύνολο που αναγνωρίζει δίλεκτους και τρίλεκτους όρους. Κάθε κανόνας μετατράπηκε σε ένα **πεπερασμένο αυτόματο** (finite-state automaton) ενισχυμένο (1) με δυνατότητες ενοποίησης συντακτικών χαρακτηριστικών και (2) με τελεστές κανονικών εκφράσεων. Τα χαρακτηριστικά αυτά, όπως φαίνεται από το παράδειγμα, μπορεί να είναι η γραμματική κατηγορία (ουσιαστικό, άρθρο, επίρρημα, κλπ.) ή χαρακτηριστικά υποκατηγοριοποίησης όπως γένος, πτώση, αριθμός, έγκλιση, φωνή κλπ. Οι τελεστές κανονικών εκφράσεων περιλαμβάνουν τελεστές όπως προεραιτικότητα, επανάληψη, διάζευξη κλπ.

Το **σώμα κειμένων** που χρησιμοποιήθηκε για την εφαρμογή της μεθόδου είναι ένα εγχειρίδιο οδηγιών της Hewlett-Packard μεγέθους περίπου 90000 λέξεων. Το κείμενο αυτό επιλέχτηκε επειδή συμπεριλάμβανε έναν **κατάλογο όρων** έναντι του οποίου αξιολογούνται τα αποτελέσματα της μεθόδου. Κατά την αξιολόγηση χρησιμοποιείται η **κανονική μορφή** των όρων στην οποία κάθε λέξη αντικαθίσταται από το λήμμα της.

4. Αποτελέσματα - εκτιμήσεις

Η αξιολόγηση των αποτελεσμάτων βασίστηκε στην σύγκριση των όρων που εξάγει η μέθοδος με τους όρους που απαρτίζουν τον κατάλογο όρων που συνόδευε το κείμενο. Προηγουμένως όλοι οι όροι μετασχηματίστηκαν σε μια κανονικοποιημένη μορφή η οποία περιλαμβάνει μόνο τα λήμματα των λέξεων. Με αυτόν τον τρόπο ταυτίστηκαν όροι που περιείχαν τις ίδιες λέξεις ελάχιστα διαφοροποιημένες, π.χ. στην πτώση. Για παράδειγμα, ο όρος *δείκτης επιλογής* του καταλόγου όρων απαντάται στο κείμενο μόνο ως *δείκτη επιλογής*.

Εξαιρώντας τους μονολεκτικούς όρους, το κατάλογος όρων του κειμένου περιείχε συνολικά 214 όρους. Η μέθοδος εξήγαγε 4729 όρους από τους οποίους 124 περιλαμβάνονταν στους 214 σωστούς όρους. Υπολογίστηκαν έτσι :
ποσοστό ανάκτησης (recall) 124/214 = 58%
ποσοστό ακρίβειας (precision) 124/4729= 2,6%.
Το ποσοστό ανάκτησης κρίνεται ικανοποιητικό. Μελέτη των όρων που δεν εντοπίστηκαν έδειξε ότι το 17% από αυτούς περιείχε μη ελληνικές λέξεις, λέξεις που δεν περιέχονταν στο λεξικό του γραμματικού χαρακτηριστή ή λέξεις για τις οποίες ο γραμματικός χαρακτηριστής απέδιδε λανθασμένη γραμματική κατηγορία. Ποσοστό 8,8% ήταν όροι αποτελούμενοι από 4 λέξεις, ενώ η γραμματική περιελάμβανε κανόνες κάλυψης όρων μέχρι 3 λέξεων. Αντίθετα, το ποσοστό ακρίβειας είναι χαμηλό, γεγονός αναμενόμενο που αποδίδεται στην εγγενή ιδιότητα των γραμματικών να παράγουν περισσότερες υποψήφιες φράσεις επειδή οι κανόνες τους είναι γενικοί και παραμένουν πάντα στο συντακτικό επίπεδο.

Η παρούσα γραμματική προτύπων μπορεί να εμπλουτιστεί με επιπλέον χαρακτηριστικά που θα βελτιώσουν την αποδοτικότητά της. Σε αυτά περιλαμβάνονται:

- Η στατιστική επεξεργασία (με μεθόδους όπως: μετρήσεις συχνοτήτων, υπολογισμός βάρους με TFIDF [11], NC-value [7], log-likelihood, mutual information [2]) των όρων που εξάγει η γραμματική ώστε να προκριθούν οι έγκυροι όροι του κειμένου.
- Η κωδικοποίηση στο πεπερασμένο αυτόματο κανόνων που αναγνωρίζουν όρους μεγαλύτερου μήκους.
- Η χρήση μόνο του μέγιστου σε κάλυψη όρου, σε περίπτωση που αυτός εμπεριέχει μικρότερους σε μήκος όρους. Κατ΄αυτόν τον τρόπο οι ανακτώμενοι όροι μειώνονται σημαντικά.

- Ο αποκλεισμός των λειτουργικών λέξεων (functional words) από τους όρους κατά τη διαδικασία αξιολόγησης.
- Η χρησιμοποίηση επιπλέον συντακτικής πληροφορίας (όπως η κεφαλή στις ονοματικές φράσεις) ώστε να ταυτίζονται ονοματικές φράσεις με το ίδιο περιεχόμενο αλλά με διαφορετική σειρά λέξεων (π.χ. *εταιρίες κατασκευών, κατασκευαστικές εταιρίες*).

5. Αναφορές

[1] Bourigault D. (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. Proceedings of the 14th International Conference on Computational Linguistics.

[2] Church K. W. and Hunks P. (1990) Word Association, Norms, Mutual Information, And Lexicography Computational Linguistics, Vol 16, Number 1.

[3] Dagan I. and Church K. W. (1994) Termight: Identifying and Translating Techical Terminology. Proceedings of the EACL 1994.

[4] Daille B., Gaussier E., Lange J. M.,(1994) Towards automatic extraction of monolingual and bilingual terminology, Proceedings of COLING 94, pp 515-521.

[5] Daille B. (1994), Study and implementation of combined techniques for automatic extraction of Terminology. in The Balancing Act: Combining Symbolic and Statistical Approaches to Languages, Workshop at the 32nd Annual Meeting of ACL, Las Cruces, Nouveau Mexique.

[6] Frantzi K. and Ananiadou S.,(1996) Extracting nested collocations, Proceedings of COLING 96, pp 41-46.

[7] Frantzi, K.T. and Ananiadou, S. (1997) Automatic term recognition using contextual clues, Proceedings of Mulsaic 97, IJCAI, Japan

[8] Gavriilidou M, Lambropoulou P. Report on the Constituent Grammar, RENOS project, LREI- 62-048, Athens, 1994

[9] Hatcher A.J. (1960) An introduction to the analysis of English noun compounds. In Word, 16, 356-373.

[10] Smadja F. A. and McKeown K. R. (1990) Automatically Extracting and Representing Collocations For Language Generation, Proceedings of the 28th annual Meeting of the ACL.

[11] Salton, G. (1989), Automatic text processing : the transformation, analysis, and retrieval of information by computer, Reading, Mass. Wokingham : Addison-Wesley.

Βύρων Γεωργαντόπουλος

Στέλιος Πιπερίδης

Ινστιτούτο Επεξεργασίας Λόγου

Μάργαρη 22, 115 25 Αθήνα

{byron, spip}@ilsp.gr