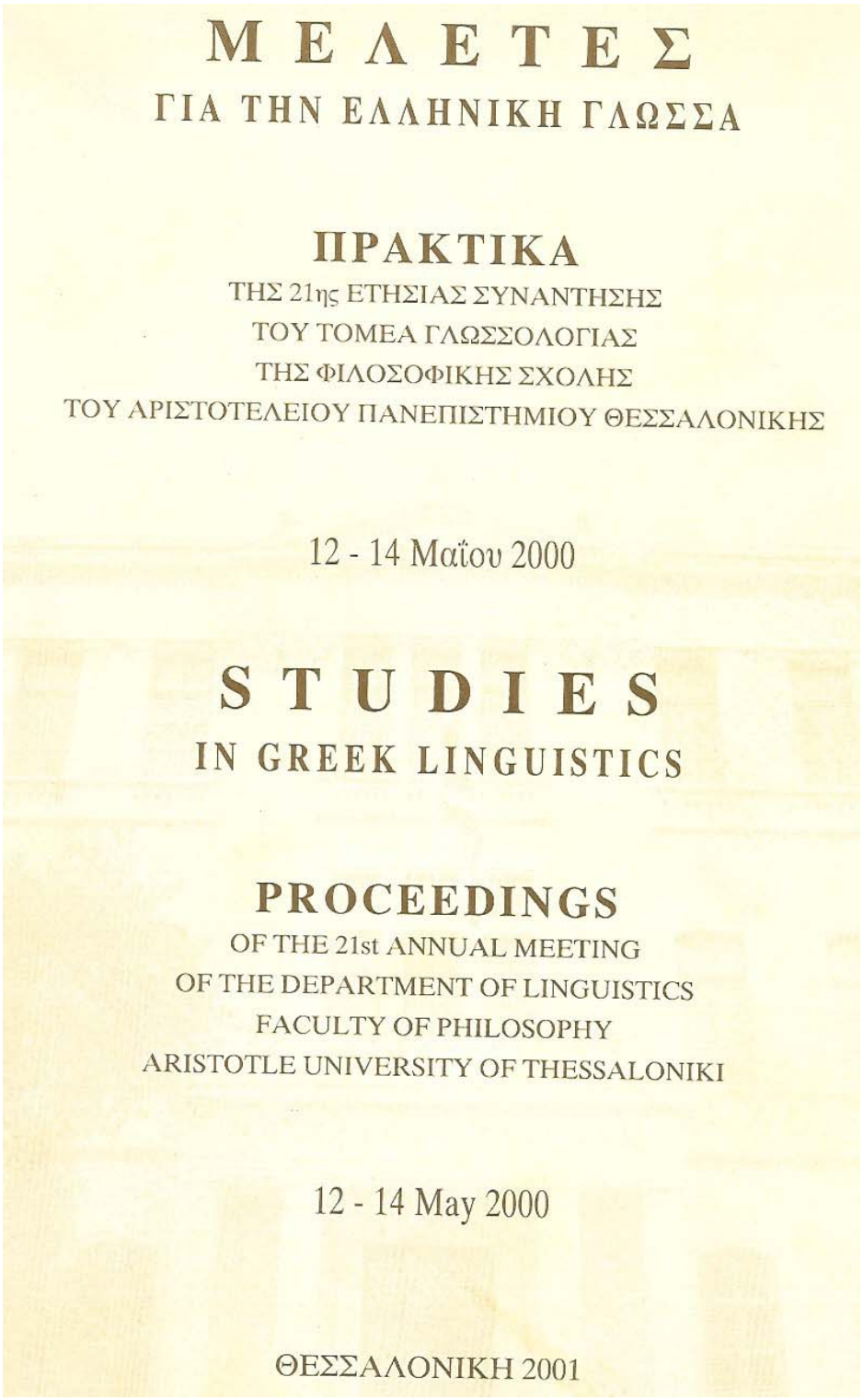
"Αυτόματη Αναγνώριση ονοματικών οντοτήτων για Εξαγωγή και Ανάκτηση Πληροφοριών"
(Σελίδες 12)
Π. Αρβανίτης - κ.ά (ομαδική εισήγηση), 2000.
Πρακτικά της 21ης Ετήσιας Συνάντησης Εργασίας του Τομέα Γλωσσολογίας του Τμήματος Φιλολογίας του Α.Π.Θ., σελ. 131-143,
Θεσσαλονίκη 2000.



ΜΕΛΕΤΕΣ
ΓΙΑ ΤΗΝ ΕΛΛΗΝΙΚΗ ΓΛΩΣΣΑ

ΠΡΑΚΤΙΚΑ
ΤΗΣ 21ης ΕΤΗΣΙΑΣ ΣΥΝΑΝΤΗΣΗΣ
ΤΟΥ ΤΟΜΕΑ ΓΛΩΣΣΟΛΟΓΙΑΣ
ΤΗΣ ΦΙΛΟΣΟΦΙΚΗΣ ΣΧΟΛΗΣ
ΤΟΥ ΑΡΙΣΤΟΤΕΛΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΘΕΣΣΑΛΟΝΙΚΗΣ

12 - 14 Μαΐου 2000

STUDIES
IN GREEK LINGUISTICS

PROCEEDINGS
OF THE 21st ANNUAL MEETING
OF THE DEPARTMENT OF LINGUISTICS
FACULTY OF PHILOSOPHY
ARISTOTLE UNIVERSITY OF THESSALONIKI

12 - 14 May 2000

ΘΕΣΣΑΛΟΝΙΚΗ 2001

Αυτόματη Αναγνώριση Ονοματικών Οντοτήτων για Εξαγωγή και Ανάκτηση Πληροφοριών

Παρασκευή Γιούλη[1] , Σωτήρης Μπούτσης[1], Ιάσων Δεμοίρος[1], Βασίλης Αντωνόπουλος[1], Χάρης Παπαγεωργίου[1], Στέλιος Πιπερίδης[1] , Παναγιώτης Αρβανίτης[2], Άλκηστις Χιδίρογλου[2], Μιχάλης Κοπιδάκης[3]

[1] Ινστιτούτο Επεξεργασίας Λόγου, [2]Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, [3]Εθνικό & Καποδιστριακό Πανεπιστήμιο Αθηνών

Abstract

In this paper, we describe work in progress for the development of a named entity recognizer for Greek. The system aims at information extraction applications where large scale text processing is needed. Speed of analysis, system robustness, and results accuracy have been the basic guidelines for the system's design. Our system is an automated pipeline of linguistic components for Greek text processing based on pattern matching techniques. Non-recursive regular expressions have been implemented on top of it in order to capture different types of named entities. For development and testing purposes, we collected a corpus of financial texts from several web sources and manually annotated part of it. Overall precision and recall are 86% and 81% respectively.

1 Εισαγωγή

Στόχος της παρούσας εργασίας είναι η παρουσίαση ενός εργαλείου αναγνώρισης και κατηγοριοποίησης Ονοματικών Οντοτήτων (ΟΟ) σε ελληνικά Κείμενα, το οποίο - μαζί με εργαλεία επιφανειακής συντακτικής ανάλυσης και επίλυσης συναναφορών - θα ενσωματωθεί σε συστήματα εξαγωγής και ανάκτησης πληροφοριών (Information Extraction & Retrieval). Το εν λόγω εργαλείο αναπτύσσεται στα πλαίσια του προγράμματος ΠΕΝΕΔ "oikONOMiA" στο Ινστιτούτο Επεξεργασίας Λόγου και σε συνεργασία με το Εθνικό και Καποδιστριακό Παν/μιο Αθηνών και το Αριστοτέλειο Παν/μιο Θεσσαλονίκης.

Η αναγνώριση και κατηγοριοποίηση ΟΟ αντιμετωπίζονται διεθνώς ως επιμέρους εργασία προκειμένου για την εξαγωγή πληροφορίας από κείμενο, κυρίως στα πλαίσια των Διεθνών Συνεδρίων Αξιολόγησης Τεχνολογίας Εξαγωγής Πληροφορίας (Message Understanding Conferences: MUC). Κατά την ανάπτυξη του εν λόγω συστήματος, ακολουθήσαμε τις προδιαφραφές του MUC-7 με τις αναγκαίες αλλαγές προκειμένου για την ελληνική. Σύμφωνα με τις προδιαγραφές αυτές, στοχεύουμε στην αναγνώριση ονομάτων προσώπων, οργανισμών, και τοπωνυμίων (ENAMEX), ημερομηνίες και εκφράσεις που δηλώνουν ώρα (TIMEX), καθώς επίσης και ποσοστώσεις και αριθμητικές εκφράσεις (NUMEX).

Αρχικά ένας επεξεργαστής πεπερασμένων καταστάσεων πραγματοποιεί αναγνώριση των βασικών δομικών συστατικών του κειμένου, δηλ. λεκτικές ενότητες και περίοδοι (tokenization και sentence boundary identification). Εν συνεχεία πραγματοποιείται μορφοσυντακτικός χαρακτηρισμός και λημματοποίηση των λέξεων του κειμένου με τη βοήθεια εργαλείων που επίσης αναπτύχθηκαν στο ΙΕΛ (part-of-speech Brill tagger, και lexicon-based lemmatizer). Ακολουθεί η αναγνώριση γνωστών ΟΟ με βάση λίστες ονομάτων προσώπων, οργανισμών τοποθεσιών κλπ, οι οποίες έχουν καταρτιστεί χειρωνακτικά. Επίσης, στο στάδιο αυτό (lookup module) λέξεις οι οποίες είναι ενδεικτικές της ύπαρξης ΟΟ στο περιβάλλον τους ή λέξεις που αποτελούν τμήμα πολυλεκτικών ΟΟ αναγνωρίζονται και χαρακτηρίζονται, με βάση αντίστοιχες λίστες λέξεων. Στο τελευταίο στάδιο, πραγματοποιείται η τελική αναγνώριση και κατηγοριοποίηση ΟΟ με χρήση μιας γραμματικής προτύπων (pattern grammar) που βασίζεται σε τεχνικές πεπερασμένων καταστάσεων. Για την ανάπτυξη του εργαλείου (κατάρτιση λιστών, εξαγωγή γραμματικών κανόνων) χρησιμοποιήθηκε ένα σώμα 120.000 περίπου λέξεων.

Τέλος, ένα σώμα κειμένων 30,000 λέξεων χρησιμοποιήθηκε για την αρχική αξιολόγηση του εργαλείου.
Η προσφορά του ΑΠΘ και του ΜΙΘΕ συνίσταται στην περεταίρω αξιολόγηση του εργαλείου με την κατάρτιση εκτεταμένου σώματος κειμένου και τον σχολιασμό των ΟΟ σε αυτό.
Ειδικότερα, συντελεστές στο έργο είναι:

- για τη γενική ευθύνη του συντονισμού των εργασιών των ερευνητικών ομάδων, ο κ. Στέλιος Πιπερίδης, Επιστημονικός υπεύθυνος του Έργου και Υπεύθυνος του Τμήματος Γλωσσικών Εφαρμογών

του ΙΕΛ, ο κ. Μ. Κοπιδάκης, Τακτικός καθηγητής στο Τμήμα Μεθοδολογίας, Ιστορίας και Θεωρίας της Επιστήμης του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών και η Άλκηστις Χιδίρογλου, Επίκουρη καθηγήτρια στο Τμήμα Γαλλικής Γλώσσας και Φιλολογίας του Α.Π.Θ..

- για την ευθύνη των υπολογιστικών εφαρμογών και την τεχνική επίβλεψη των εργασιών που πραγματοποιούνται στο Εργαστήριο Διδακτικής Ζωντανών Γλωσσών, την διεύθυνση του οποίου έχει η κ. Βάσω Τοκατλίδου, Τακτική καθηγήτρια του Τμήματος Γαλλικής Γλώσσας και Φιλολογίας, ο κ. Παν. Αρβανίτης διδάκτωρ του ΑΠΘ.
- το έργο πλαισιώνουν ακόμη νέοι ερευνητές και συγκεκριμένα στη Θεσσαλονίκη η κ. Έφη Παπαναστασίου και Μπέτυ Κακλαμανίδου, καθώς και φοιτητές του Τμήματος Γαλλικής Γλώσσας οι οποίοι εκπαιδεύονται στην χρήση των υπολογιστικών εργαλείων υποδομής του προγράμματος και αξιολογούν την λειτουργία τους στα πλαίσια του μαθήματος της Λεξικογραφίας.
-

2 Θεωρητική Θεμελίωση

Η επεξεργασία εξειδικευμένων μεθόδων, στα πλαίσια του έργου θα παρέχει την δυνατότητα αναζήτησης κειμένων με συγκεκριμένο θέμα, εφικτό με την ως τώρα σύγχρονη τεχνολογία, πέραν όμως αυτού θα παρέχει και την δυνατότητα εύρεσης τμημάτων των κειμένων που παρέχουν πληροφορία για το συγκεκριμένο θέμα και τέλος επιφανειακή κατανόηση ελεύθερου κειμένου.

Πιο συγκεκριμένα οι πληροφορίες που θα δίνονται, θα αφορούν τη γραμματική κατηγορία των λέξεων-κλειδιών, τις συντακτικές σχέσεις των λέξεων, την ύπαρξη ονοματικών οντοτήτων και των κατηγοριών τους και τέλος την ύπαρξη συναναφορών. Έτσι, ενώ ένα σύστημα ανάκτησης πληροφοριών απαντά στην ερώτηση του χρήστη με μια επίπεδη λίστα λέξεων-κλειδιών, το συγκεκριμένο σύστημα πραγματοποιεί μια πολύ πιο σύνθετη επεξεργασία της ερώτησης.

Ο σχεδιασμός του έργου έχει γίνει με βάση το γεγονός ότι οι εφαρμογές στις οποίες απευθύνεται είναι εφαρμογές πραγματικού χρόνου και έντασης δεδομένων. Σε αυτό το πλαίσιο, η ταχύτητα επεξεργασίας, και η σχετική ακρίβεια των αποτελεσμάτων είναι θεμελιώδους σημασίας. Συγκεκριμένα, προκειμένου το σύστημα να μπορεί να χρησιμοποιηθεί σε περιβάλλον πραγματικών εφαρμογών, πρέπει να είναι σε θέση να χειριστεί πραγματικά δεδομένα, δηλαδή ελεύθερα κείμενα του οικονομικού, συγκεκριμένα, πεδίου. Έτσι, στο έργο χρησιμοποιούνται κείμενα που προέρχονται από οικονομικές εφημερίδες και άλλες πηγές οικονομικής πληροφόρησης στο Διαδίκτυο, καθώς και κατεγραμμένο υλικό ραδιοφωνικών και τηλεοπτικών εκπομπών με οικονομικό περιεχόμενο.

Στην ανακοίνωση αυτή πρόθεσή μας είναι να παρουσιάσουμε ένα υποσύστημα αναγνώρισης ονοματικών οντοτήτων το οποίο αποτελεί μιαν από τις διαδικασίες του έργου.

Ως ονοματικές οντότητες καταγράφονται τα κύρια ονόματα προσώπων, εταιριών, οργανισμών, κ.ά., η μελέτη των οποίων εντάσσεται στο αντικείμενο της ονοματολογίας. Θα πρέπει να επισημανθεί εδώ οτι μπορεί οι μέχρι σήμερα έρευνες στον τομέα της ονοματολογίας να αφορούσαν κυρίως μελέτες ανθρωπωνυμίων και τοπωνυμίων, οι άπειροι ονοματικοί σχηματισμοί οι οποίοι καλύπτουν πολύπλευρες ανάγκες του τεχνολογικού πολιτισμού και της σύγχρονης καταναλωτικής κοινωνίας, απαιτούν μια νέα θεώρηση όπου η συγχρονία παίρνει την θέση της διαχρονίας σε τομείς όπως της τεχνολογίας, της οικονομίας και της βιομηχανίας. Έτσι η ονοματολογία αποσυνδέεται από τη γή και τον άνθρωπο για να προσεγγίσει το τεχνητό και το επιτηδευμένο. Άλλωστε σύμφωνα με τον Ch. Camproux : "*l' onomastique au sens large est la science du nom proper qu'il s'agisse du nom d' un avion, d'une pile electronque, d'un rasoir, d'un robot ….. ou qu'il s'agisse d'une localité ou d'une personne*".

Γιατί η αναγνώριση ονοματικών οντοτήτων;

Αντίθετα με την κλασσική θεωρία την οποία θεμελίωσε ο S. Mill για την απουσία σημασίας του κυρίου ονόματος, οτι δηλαδή το κύριο όνομα στερείται σημασίας και κατά συνέπεια βρίσκεται στο περιθώριο της σημασιολογίας, είναι δηλαδή μια απλή ετικέτα, θεωρία η οποία υποστηρίχθηκε στη συνέχεια από τους Gardiner, Togeby, Ullmann, Kripke, Lyons κ.ά.

Θεωρούμε σύμφωνα με τις νεώτερες θεωρίες οτι το κύριο όνομα δηλώνει μια αναφορά αλλά επίσης έχει και μια σημασία. Οι θεωρίες αυτές υποστηρίχθηκαν από τους Frege, Russel, Sorensen, Searle κ.ά., σύμφωνα με τους οποίους η χρήση του κύριου ονόματος είναι αδύνατη χωρίς να σημαίνει κάτι στον δέκτη, χωρίς δηλαδή ο δέκτης να παίρνει κάποιες πληροφορίες σχετικά με τον φορέα του ονόματος.

Αυτό είναι το απαύγασμα της γνώσης που έχουμε για την προβληματική του ονόματος ως προς το θεωρητικό του μέρος. Εννοείται οτι κάθε ένα από αυτά τα θεωρητικά δεδομένα χρήζει περαιτέρω ειδικής διερεύνησης.

Η μελέτη βέβαια της σημασίας του κύριου ονόματος ανάγεται στην αρχαιότητα και συγκεκριμένα στον Πλάτωνα, ο οποίος στο έργο του Κρατύλος θεωρεί οτι "τα ονόματα δείχνουν πως είναι το κάθε ον" και οτι "αυτός που γνωρίζει τα ονόματα, γνωρίζει και τα πράγματα".
Στο έργο αυτό θεωρούμε οτι τα κύρια ονόματα αποτελούν κατά κύριο λόγο τις σημασιολογικές κεφαλές που δυνάμει συμπληρώνουν τους θεματικούς ρόλους των γεγονότων που περιγράφονται σε ένα κείμενο.

3 Βιβλιογραφική Ανασκόπηση

Η ανασκόπηση της βιβλιογραφίας δείχνει ότι μπορεί κανείς να διακρίνει τρείς κατηγορίες συστημάτων με βάση την ακολουθούμενη μεθοδολογία: *νομοθετικά συστήματα*, κύριο χαρακτηριστικό των οποίων είναι η ύπαρξη γραμματικών κανόνων (rule based), συστήματα που χρησιμοποιούν *μεθόδους μηχανικής μάθησης* (learning machines), και, τέλος, *υβριδικά συστήματα*, που συνδυάζουν τη χρήση κανόνων και μηχανών μάθησης.

Ενα τυπικό νομοθετικό σύστημα αναγνώρισης ονοματικών οντοτήτων περιλαμβάνει εργαλεία για τη λεκτική επεξεργασία των κειμένων (διαχωρισμός λέξεων και προτάσεων), αναγνώριση μέρους του λόγου και μορφολογική ανάλυση κάθε λέξης (POS tagging). Ακολουθεί η συντακτική ανάλυση των κειμένων σε δύο στάδια: αρχικά πραγματοποιείται αναγνώριση γνωστών ονοματικών οντοτήτων με χρήση ονοματικών καταλόγων (gazetteer lists), και ακολουθεί η αναγνώριση ονοματικών φράσεων και των υπόλοιπων ονοματικών οντοτήτων με τη βοήθεια σχετικών γραμματικών κανόνων που περιέχουν και λεκτικές πληροφορίες (π.χ. μέρος του λόγου, λέξεις-κλειδιά). Τέλος, πραγματοποιείται η αναγνώριση των εναλλακτικών διατυπώσεων μιας ονοματικής οντότητας (aliases) (π.χ. "Ολυμπιακή Αεροπορία", "Ολυμπιακή", "Ο.Α.", "Olympic Airways").

Στα περισσότερα από τα συστήματα της παραπάνω κατηγορίας, οι γραμματικοί κανόνες και οι κατάλογοι ονομάτων κατασκευάζονται χειρωνακτικά από ειδικούς. Ακολουθούν παραδείγματα τέτοιων συστημάτων από τα πιο πρόσφατα MUC.

Το σύστημα PET του New York University στο MUC-6 [11], βασίζεται σε κανόνες *προτύπων* (patterns). Τα πρότυπα αυτά περιλαμβάνουν πληροφορία όπως η ύπαρξη κεφαλαίου στην αρχή μιας λέξης, η ύπαρξη λέξεων-κλειδιών στα συμφραζόμενα, όπως τίτλοι προσώπων ("Mr.", "Esq."), προσδιοσριστικά εταιριών ("Co.","Inc.") κ.ά.

Στο σύστημα LaSIE του University of Sheffield [9] οι γραμματικοί κανόνες έχουν επίσης προέλθει χειρωνακτικά ενώ για την προσαρμογή του συστήματος σε νέα θεματική περιοχή απαιτείται ο εμπλουτισμός και τροποποίηση των καταλόγων ονομάτων του. Η συνολική ανάκτηση και ακρίβεια για το LaSIE ήταν 85,3%.

Το σύστημα του UMIST, FACILE [3], χρησιμοποιεί και αυτό κανόνες που λαμβάνουν υπόψη τα συμφραζόμενα, μόνο που ένας κανόνας δεν περιλαμβάνει πρότυπα, αλλά κάθε μέλος του έχει τη μορφή συνόλου ζευγών χαρακτηριστικών-τιμών (attribute-value pairs), ενώ η εφαρμογή ενός κανόνα αποφασίζεται με βάση έναν αλγόριθμο προτίμησης. Η ακρίβεια για το σύστημα αυτό ήταν 87% ενώ η ανάκτηση 78%.

Καθώς η κατασκευή γραμματικών κανόνων είναι ιδιαίτερα χρονοβόρα και μάλιστα απαιτείται επανάληψη της διαδικασίας αυτής για αναπροσαρμογή του συστήματος σε καινούρια θεματική περιοχή, κάποια συστήματα κάνουν χρήση μεθόδων μηχανικής μάθησης για την αυτόματη εξαγωγή γραμματικών κανόνων από ειδικά σχολιασμένα κείμενα. Οι μέθοδοι μηχανικής μάθησης που υιοθετούνται διαφέρουν από σύστημα σε σύστημα.

Το Nymble [2] χρησιμοποιεί μάθηση με στατιστικά μοντέλα (HMM) για την κατασκευή γραμματικών κανόνων και στο MUC-7 πέτυχε ανάκτηση 89% και ακρίβεια 92%, χωρίς μάλιστα τη χρήση καταλόγων ονομάτων που θα μπορούσαν να βελτιώσουν επιπλέον την απόδοση του συστήματος.

΄Ενα σύστημα που βαθίζεται στη μάθηση κανόνων με μετασχηματισμό (transfromation-based) είναι το Alembic [1], που ξεκινά με περιορισμένο σύνολο κανόνων, τους οποίους μετασχηματίζει φτιάχντοντας

νέους κανόνες. Το σύστημα αυτό έδωσε 85% ανάκληση και 86%, ένω εμπλουτισμένο με σύνολο κανόνων κατασκευασμένων στο χέρι έδωσε αντίστοιχα 91% ανάκληση και 92% ακρίβεια.

Άλλα συστήματα που χρησιμοποιούν μηχανική μάθηση για την κατασκευή μοντέλου αναγνώρισης ονοματικών οντοτήτων στοχεύουν στον εντοπισμό των ορίων μιας ονοματικής οντότητας εκπαιδεύοντας τη μηχανή μάθησης σε κατάλληλα σχολιασμένα δεδομένα.Έτσι, το Autolearn [7] κατασκευάζει δένδρα αποφάσεων από δεδομένα εκμάθησης. Το συγκέκριμένο σύστημα δεν είχε καλές επιδόσεις (ανάκτηση: 47%, ακρίβεια 81%) καθώς δεν χρησιμοποιούσε καθόλου γλωσσολογικούς πόρους.

Αντίθετα, το MENE του New York University [4] εκπαιδεύεται με τον υπολογισμό δεσμευμένων πιθανότητων εμφάνισης του ορίου κάποιας ονοματικής οντότητας με δεδομένα συγκεκριμένα συμφραζόμενα. Τα συμφραζόμενα αυτά εκφράζονται με τη μορφή χαρακτηριστικών που εξάγονται αυτόματα από κατάλληλα χαρακτηρισμένα κείμενα, λεξικά ή αποτελέσματα άλλων συστημάτων. Κάθε φορά, επιλέγεται το ενδεχόμενο που μεγιστοποιεί την εντροπία (Maximum Entropy). Το σύστημα αυτό σημείωσε συνολική ανάκτηση και ακρίβεια 88.8 %.

Τέλος, το σύστημα που είχε την καλύτερη επίδοση στο MUC-7 ήταν ένα υβριδικό σύστημα, το LTG του University of Edinburgh [16] . Αυτό δεν βασίζεται στην ύπαρξη πλούσιων καταλόγων ονομάτων αλλά συνδυάζει τη χρήση γλωσσικής πληροφορίας, υπό μορφή χειρωνακτικών κανόνων, με το στατιστικό μοντέλο μεγιστοποίησης εντροπίας.

4 Ορισμός – Είδη Ονοματικών Οντοτήτων κατά τις προδιαγραφές του MUC-7

Για τον προσδιορισμό ενός σχήματος σχολιασμού (annotation schema) ΟΟ σε ελληνικά κείμενα, ακολουθήσαμε τις οδηγίες – προδιαγραφές του MUC-7 [6]. Συγκεκριμένα, αναγνωρίζουμε κύρια ονόματα (ENAMEX) του τύπου PERSON, ORGANISATION και LOCATION, χρονικές εκφράσεις (TIMEX) οι οποίες διακρίνονται σε DATE και TIME, και τέλος αριθμητικές εκφράσεις (NUMEX) οι οποίες διακρίνονται σε MONEY και PERCENT. Ακολουθεί συνοπτικά αναφορά των προδιαγραφών του MUC-7 για τα είδη και τις υποκατηγορίες των ΟΟ προσαρμοσμένες στα ελληνικά δεδομένα:

Person: Ως ΟΟ του τύπου person αναγνωρίζονται ονόματα φυσικών προσώπων ή οικογενειών, όταν δεν χρησιμοποιούνται ως εταιρικές επωνυμίες. Πχ.:
κ. [person Κώστας Λιάσκας /person]
οικογένεια [person Κέννεντυ /person]
κυβέρνηση [person Κ. Σημίτη /person]
η [org Μυτιληναίος /org]
Επίσης, σύμφωνα με τις προδιαγραφές του MUC-7, τίτλοι – ονοματικές κεφαλές των φράσεων στις οποίες περιέχονται τα ονόματα, όπως "κ., κος, κον.", "κα", "πρόεδρος", "διευθύνων σύμβουλος", κλπ. δεν μετέχουν της ΟΟ.

Organization: Ως ΟΟ του τύπου organisation χαρακτηρίζονται ονόματα σωματείων, εταιριών, κυβερνητικών φορέων, συλλόγων, διεθνών και άλλων οργανισμών, χρηματιστηρίων, πολιτικών κομμάτων, ενώσεων, ορχηστρών, αθλητικών ομάδων, στρατών, εκκλησιών, πρεσβειών, εργοστασίων, νοσοκομείων, ξενοδοχείων, μουσείων, πανεπιστημίων. ΟΟ του τύπου location περιλαμβάνονται μέσα στα όρια ενός organisation μόνον όταν βρίσκονται σε θέση ΟΦ-προσδιορισμού σε γενική (Τράπεζα της Ελλάδος). Αντιθέτως, δεν περιλαμβάνονται μέσα στα όρια της ΟΟ όταν αποτελούν συμπλήρωμα προθετικής φράσης η οποία, με τη σειρά της προσδιορίζει την ΟΦ-οργανισμό (Ελληνική Πρεσβεία στα Τίρανα):
[org ΧΑΑ /org]
[org Χρηματιστήριον Αξιών Αθηνών /org]
της χρηματιστηριακής [org Α. Σαρρής /org]
[org Πλαστικά Θράκης ΑΕ /org]
[org Νομική Σχολή /org] του [org Πανεπιστημίου Αθηνών /org]
[org Ελληνική Πρεσβεία /org] στα [loc Τίραννα /loc]
[org Ελληνική Πρεσβεία της Αλβανίας /org]
Επίσης, λέξεις όπως εταιρεία, οργανισμός, κλπ, οι οποίες είναι ενδεικτικές της παρουσίας ΟΟ του τύπου organisation, περιλαμβάνονται στην χαρακτηρίσιμη οντότητα μόνον όταν γράφονται με κεφαλαίο το αρχικό γράμμα:

εκδόσεις [org Σάκουλα /org]
[org Εκδόσεις Ερμής /org]
Από τον κανόνα αυτό εξαιρούνται οι λέξεις υπουργείο και χρηματιστήριο οι οποίες ακόμα και όταν απαντούν με μικρό αρχικό γράμμα, αποτεούν μέρος της ΟΟ
[org υπουργείο Εξωτερικών /org]
[org χρηματιστήριο της Φρανκφούρτης /org]
Τέλος, συντομογραφίες πριν ή μετά τα ονόματα εταιρειών λαμβάνονται ως προσφύματα ή επιθήματα εταιρικών ονομάτων και μαρκάρονται μέσα στην ΟΟ: "Αφοί", "Α.Ε.", "Α.Χ.Ε.".

Location: Ονόματα πολιτικά ή γεωγραφικά, όπως ήπειροι, χώρες, νομοί, πόλεις, κοινότητες, επαρχίες, γειτονιές, αεροδρόμια, λεωφόροι, δρόμοι, νησιά, εθνικοί δρυμοί, υδάτινοι όγκοι, βουνά, ουράνια σώματα, φανταστικοί ή μυθικοί τόποι, μνημειακές τοποθεσίες ή κτήρια χαρακτηρίζονται ως location:
[loc Παράδεισος /loc] [loc Αμαρουσίου /loc]
[loc Λ. Αμαλίας 35 /loc]
στο βόρειο [loc Αιγαίο /loc]
[loc ΒΙΠΕ Κομοτηνής /loc]
[loc Ε.Ο. Πατρών - Αθηνών - Θεσσαλονίκης - Ευζώνων /loc]
Σύμφωνα με τις προδιαγραφές του MUC-7, ονόματα κρατών, πόλεων κλπ, τα οποία χρησιμοποιούνται μετωνυμικά αντί ονομάτων οργανισμών, κυβερνητικών φορέων κλπ χαρακτηρίζονται ως location:
Η [loc Ιταλία /loc] νίκησε τη [loc Βραζιλία /loc].

Date and Time: Διακρίνονται σε απόλυτες (absolute temporal expressions) και σχετικές χρονικές εκφράσεις (relative temporal expressions) που εκφράζουν ημερομηνία ή ώρα.
[date Παρασκευή 23 Ιουλίου /date]
από [date Παρασκευή 12 έως Κυριακή 14 Μαίου 2000 /date]
στις [time 10.00 μ.μ. /time]
[date χθες /date]
Επίσης, δεκαετίες και αιώνες, ονόματα εποχών, μηνών, ημερών, εορτών κλπ ανήκουν στην κατηγορία αυτή.
[date δεκαετία του '80 /date]
πριν την [date Πρωτοχρονιά του 2000 /date]
το [date οικονομικό έτος 2000 /date]
το [date σχολικό έτος 2000 /date]
[time 10πμ ώρα Ελλάδας /time]

Money and Percent: Στην κατηγορία αυτή ανήκουν εκφράσεις που δηλώνουν χρηματικό ποσό (monetary expressions) και εκφράσεις που δηλώνουν ποσοστό είτε είναι γραμμένες αριθμητικά ή ολογράφως:
[money 1 εκ. δρχ /money]
[percent 30% /percent]
Οι εκφράσεις που δηλώνουν χρηματικό ποσό πρέπει απαραιτήτως να συνοδεύονται από το νόμισμα και προαιρετικά από το όνομα της χώρας
[money 10 εκατ. δολάρια ΗΠΑ /money]

Να σημειωθεί, ότι χαρακτηρίζονται ως ΟΟ κύρια ονόματα οντοτήτων (*Εθνική Τράπεζα της Ελλάδος),* συντμήσεις ονομάτων και ακρωνύμια (*ΕΤΕ),* συντομευμένες εκφορές ονομάτων (*Εθνική)* καθώς επίσης και μετωνυμικές εκφορές της ονομασίας οντοτήτων (*Σοφοκλέους = Χρηματιστήριο Αξιών Αθηνών)*. Δεν αποτελούν όμως ΟΟ ονομασίες υποστηρικτών πολιτικών κομμάτων, ομάδων, κλπ., ονομασίες προσφερομένων υπηρεσιών, προϊόντων, ή άλλων κατασκευασμάτων, ονομασίες αθλητικών συναντήσεων, μεγάλων γεγονότων κλπ. (Ολυμπιακοί Αγώνες), ονόματα ασθενειών, βραβείων, κλπ (βραβείο Νόμπελ, νόσος Αλτσχάιμερ), και, τέλος, ονοματικές και αντωνυμικές συν-αναφορές ονομάτων:
ο πρόεδρος της [org Intrasoft /org] ανακοίνωσε ότι η εταιρεία
ανώτατο στέλεχος της Επιτροπής υποχρεώθηκε

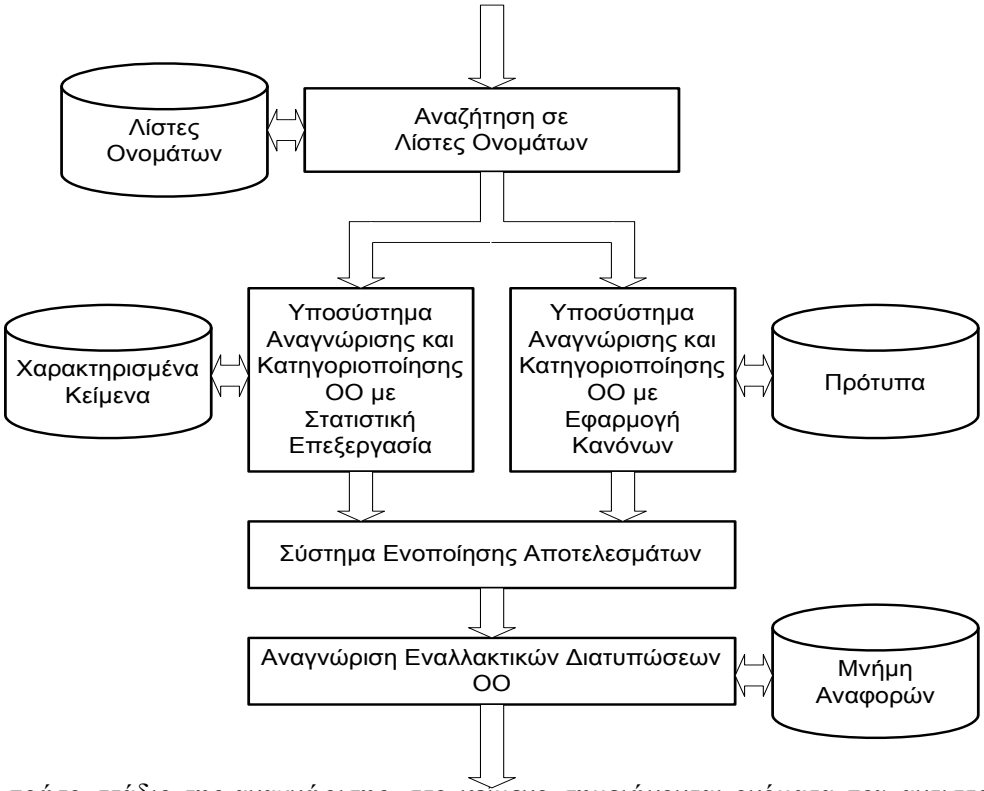
5 Μεθοδολογία

Για την ανάπτυξη του εργαλείου χρησιμοποιήθηκε ένα σώμα κειμένων 12,000,000 περίπου λέξεων, προερχόμενα από το διαδίκτυο. Καθώς τα υπό ανάπτυξιν συστήματα βρίσκουν πεδίο εφαρμογής κυρίως σε κείμενα οικονομικού περιεχομένου, τα άρθρα τα οποία συνελέγησαν προέρχονται κυρίως από

οικονομικές εφημερίδες και περιοδικά (Express, Ναυτεμπορική, Ισοτιμία, Οικονομικός Ταχθδρόμος και ΒΗΜΑ). Επίσης, το ζητούμενο ήταν τα κείμενα να περιλαμβάνουν όσο το δυνατόν περισσότερα Κύρια Ονόματα, και επομένως, πραγματοποιήθηκε μια δεύτερη επιλογή των κειμένων με βάση των αριθμό των λέξεων που αρχίζουν με κεφαλαίο γράμμα. Το τελικό σώμα κειμένων αποτελείται από 150,000 λέξεις περίπου, και αυτό σχολιάστηκε με βάση τις ανωτέρω προδιαγραφές. Για το σχολιασμό ΟΟ στο σώμα κειμένων χρησιμοποιήθηκε ένα ειδικό εργαλείο σχολιασμού (TclTk graphical user interface). Το σχολιασμένο σώμα κειμένων χρησιμοποιήθηκε τόσο για την ανάπτυξη του εργαλείου όσο και για την αξιολόγηση: 120,000 λέξεις χρησιμοποιήθηκαν για την εξαγωγή γραμματικών κανόνων προτύπων, κατάρτιση λιστών λέξεων οι οποίες είναι ενδεικτικές της ύπαρξης ΟΟ ή μετέχουν σε ΟΟ, ενώ οι υπόλοιπες 30,000 λέξεις χρησιμοποιήθηκαν για την αξιολόγηση του συστήματος.

6 Αρχιτεκτονική του Συστήματος

Σκοπός του έργου οικΟΝΟΜία είναι να αναπτυχθεί ένα υβριδικό σύστημα αναγνώρισης ΟΟ το οποίο θα χρησιμοποιεί δύο υποσυστήματα: το πρώτο θα βασίζεται σε κανόνες ενώ το δεύτερο σε στατιστική επεξεργασία (Εικόνα 1). Προς το παρόν έχει αναπτυχθεί το πρώτο υποσύστημα. Το σύστημα στην είσοδό του δέχεται κείμενο στο οποίο έχουν προηγηθεί λεκτική ανάλυση (tokenisation), μορφοσυντακτική ανάλυση και λημματοποίηση.



Στο πρώτο στάδιο της αναγνώρισης, στο κείμενο σημειώνονται ονόματα που αντιστοιχούν σε γνωστές καταχωρήσεις καταλόγων ονομάτων. Οι κατάλογοι αυτοί έχουν καταρτιστεί χειρωνακτικά και προέρχονται από διάφορες πηγές: από το Χρυσό Οδηγό, το Χρηματιστήριο Αξιών Αθηνών, τον Οργανισμό Τηλεπικοινωνιών Ελλάδος, το Τεχνικό Επιμελητήριο Ελλάδας, την Εθνική Στατιστική Υπηρεσία, κλπ. Από τις ανωτέρω πηγές προέκυψαν οι λίστες προσώπων, οργανισμών και τοπωνυμίων, οι οποίες εμπλουτίστηκαν περεταίρω με ονόματα από το σώμα κειμένων που χρησιμοποιήθηκε για ανάπτυξη (130,000 λέξεις). Συνολικά, έχουν περιληφθεί 1.059 ονόματα επιχειρήσεων, 793 ονόματα τοωνυμίων, και 1.496 ονόματα προσώπων.

Επιπλέον, σημειώνονται και οι λέξεις οι οποίες είναι ενδεικτικές της ύπαρξης ή μετέχουν ΟΟ και οι οποίες θα αξιοποιηθούν στη συνέχεια από τους κανόνες. Για το σκοπό αυτό έχουν καταρτιστεί κατάλογοι λέξεων - μονολεκτικών ή πολυλεκτικών – καθώς επίσης και κανονικών εκφράσεων (regular expressions) οι οποίες είτε απαντούν στο περιβάλλον ΟΟ και είναι, επομένως, ενδεικτικές της ύπαρξης ΟΟ συγκεκριμένης κατηγορίας, είτε μετέχουν των ΟΟ, όπως είναι επαγγέλματα, τίτλοι προσώπων, νομίσματα κρατών, κλπ. Οι λέξεις αυτές εξήχθησαν αυτομάτως από το σώμα κειμένων με χρήση στατιστικών μεθόδων (mutual information statistics), και εν συνεχεία κατηγοριοποιήθηκαν χειρωνακτικά

ανάλογα με τη λειτουργία τους και το σημασιολογικό τους περιεχόμενο. Οι τελικοί κατάλογοι εμπλουτίστηκαν περεταίρω κατά την ανάπτυξη της γραμματικής. Στο σύστημα έχουν ενσωματωθεί 57 τέτοιοι κατάλογοι οι οποίοι περιλαμβάνουν 920 λέξεις και κανονικές εκφράσεις.

Στο δεύτερο στάδιο, το μερικώς χαρακτηρισμένο κείμενο θα διοχετεύεται ταυτόχρονα σε δύο υποσυστήματα αναγνώρισης και κατηγοριοποίησης ονομάτων που λειτουργούν παράλληλα: το ένα υποσύστημα θα χρησιμοποιεί ένα στατιστικό μοντέλο και το άλλο κανόνες προτύπων. Ένα κατάλληλα σχολιασμένο σώμα κειμένων, το οποίο βρίσκεται υπό ανάπτυξη θα χρησιμοποιηθεί για την εκπαίδευση του στατιστικού μοντέλου. Το υποσύστημα κανόνων το οποίο παρουσιάζουμε εν συνεχεία, χρησιμοποιεί κανόνες προτύπων οι οποίοι έχουν κατασκευαστεί χειρωνακτικά και οι οποίοι λειτουργούν στη βάση των ήδη αναγνωρισμένων ΟΟ και των ενδεικτικών λέξεων, ώστε να αναγνωριστούν σύνθετες ΟΟ ή να επιλυθούν αμφισημίες. Μορφοσυντακτική πληροφορία (για το μέρος του λόγου, γένος, αριθμό, πτώση, πρόσωπο κλπ) καθώς επίσης και πληροφορία για ύπαρξη κεφαλαίου ή μη λαμβάνεται επίσης υπόψιν. Για παράδειγμα, η 'Τράπεζα της Ελλάδος' θα αναγνωριστεί στο στάδιο αυτό ως organisation με λειτουργία κανόνα ο οποίος ενώνει λέξεις όπως τράπεζα, Χρηματιστήριο κλπ. με μία ΟΟ του τύπου location της οποίας προηγείται οριστικό άρθρο σε γενική, εφόσον το 'Τράπεζα' έχει αναγνωριστεί ως orgdeignator και το 'Ελλάδος' ως location.

Η γραμματική περιλαμβάνει κανόνες υπό μορφήν κανονικών εκφράσεων (regular expressions) [14], που μετετρέπονται σε μεταγραφείς πεπερασμένων καταστάσεων (finite state transducers) οι οποίοι εφαρμόζονται στις λέξεις του κειμένου σειριακά ώσπου να ικανοποιηθεί η αρχή του μεγαλύτερου ταιριάσματος (longest match). Η γραμματική αποτελείται από 110 συνολικά κανόνες: 17 για την αναγνώριση ΟΟ του τύπου person, 19 για ΟΟ του τυπου location, 37 για organization, 23 για date, 5 για time, 7 για money and 2 για percent. Ένα παράδειγμα κανόνα δίνεται στη συνέχεια:

```
markup (
[geosign+, atdf_ge^, {cap_aj, locadj_cap}^, abbr^, {const({'[person', '[loc'}, {'/person]', '/loc]'}), cap_word, cap_rg}+, dig^],
'[loc', '/loc]')

o

conditional_markup_upward(
[{cap_rg, cap_word}+ ], '[loc', '/loc]',
[{geosign, indiclocverb}, as_se],
[] )

<EOR>
```

Το τρίτο στάδιο, θα αφορά ενοποίηση των αποτελεσμάτων των δύο υποσυστημάτων του δεύτερου σταδίου. Η διαδικασία της ενοποίησης θα πραγματοποιείται από ένα τρίτο υποσύστημα, που θα εκπαιδευτεί να επιλέγει μεταξύ των αποτελεσμάτων των δύο υποσυστημάτων, μεγιστοποιώντας την ακρίβεια του τελικού αποτελέσματος. Οι εναλλακτικές διατυπώσεις κάθε χαρακτηρισμένης οντότητας αναγνωρίζονται στο τέταρτο στάδιο. Για το σκοπό αυτό, θα χρησιμοποιείται λίστα εναλλακτικών διατυπώσεων παράλληλα με μεθόδους συνδυαστικής επεξεργασίας. Μία μνήμη αναφορών θα χρησιμεύει για την αποθήκευση ονομάτων, των γνωστών μέχρι κάθε στιγμή εναλλακτικών διατυπώσεών τους, γραμματικών και συντακτικών χαρακτηριστικών, κλπ.

7 Αξιολόγηση του συστήματος

Έχει ήδη αναφερθεί ότι η αξιολόγηση του συστήματος αναγνώρισης και κατηγοριοποίησης ΟΟ πραγματοποιήθηκε σε τμήμα του σχολιασμένου σώματος κειμένων (30,000 λέξεις). Τα αποτελέσματα της αξιολόγησης παρουσιάζονται κατωτέρω (Πίνακας 1).

Επιπλέον, αναζητήθηκαν και ταξινομήθηκαν τα λάθη ανάλογα με την πηγή τους (Πίνακας 2). Σημαντικό ποσοστό των λαθών (18.4%) σε προηγούμενα στάδια επεξεργασίας του κειμένου: αναγνώριση λέξεων και περιόδων, μορφοσυντακτικός χαρακτηρισμός, λημματοποίηση (tokenization, part-of-speech tagging,

lemmatization). Επίσης, παρατηρείται αμφισημία, κυρίως μεταξύ προσώπων και εταιρικών επωνυμιών, ιδίως σε περιπτώσεις όπου δεν υπάρχουν σαφείς ενδείξεις μέσα στο κείμενο.
Η [person Γερμανός / person] εξέδωσε 1.000.000 νέες μετοχές.

Τα ορθογραφικά ή τυπογραφικά λάθη καθώς επίσης και η μείξη των χαρακτήρων του ελληνικού με το λατινικό αλφάβητο ευθύνονται για το 11.5% των αποτυχιών .

Τέλος, θέλοντας να δούμε πόσο οι λίστες γνωστών ΟΟ συνεισφέρουν στην τελική αναγνώριση και καατηγοριοποίηση, τις παραλείψαμε από το σύστημα και κάναμε ένα ακόμα πείραμα με τα ίδια κείμενα. Τα ποσοστά, πολύ χαμηλότερα από το προηγούμενο πείραμα, παρουσιάζονται στον Πινακα 3

NE Type	Precision	Recall	F-Measure
Person	0.71	0.71	0,71
Loc	0.85	0.82	0,83
Org	0.80	0.72	0,76
Money	0.99	0.95	0,97
Percent	1.00	0.98	0,99
Date	0.89	0.84	0,86
Time	0	0	0
Total	0.86	0.81	0,83

Πίνακας 1: Αποτελέσματα

NE Type	Error Distribution			
	Λάθη (%) προ-επεξεργασίας	Λάθη (%) ορθογραφικά	Λάθη (%) αμφισημίας	Λοιπά
Person	46.0%	00.0%	12.6%	41.4%
Loc	12.5%	00.0%	09.7%	77.8%
Org	09.2%	15.7%	06.1%	69.0%
Date	15.1%	23.3%	00.0%	61.6%
Money	81.9%	00.0%	00.0%	18.1%
Total	18.4%	11.5%	06.6%	63.5%

Πίνακας 2: Κατανομή Λαθών

NE Type	Precision	Recall	F-Measure
Person	0.80	0.34	0.47
Org	0.77	0.36	0.49
Loc	0.82	0.14	0.23
Date	0.89	0.84	0.86
Money	0.99	0.95	0.96
Percent	1.00	0.98	0.98
Total	0.75	0.45	0.56

Πίνακας 3:

8 Επίλογος

Παρουσιάστηκε ένα εργαλείο αναγνώρισης και κατηγοριοποίησης ΟΟ σε ελληνικά κείμενα το οποίο προορίζεται να ενσωματωθεί σε σύστημα εξαγωγής πληροφορίας. We implemented finite state techniques favoring efficient text processing and adopted a modular design allowing fast customization to

the needs and particularities of specific applications. We also carried out an elaborate evaluation of the system's output and identified the design and implementation aspects we should enhance. Since work is still in progress, we expect that benchmarks will further improve; the system, however, has already reached a level of performance (F=83%) which is satisfying for many real-world applications.

Βιβλιογραφία

1. Aberdeen J., Burger J., Day D., Hirschman L., Robinson P., Vilain M. 1995. Mitre: description of the Alembic system used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
2. Bikel D., Miller S., Schwartz R., Weischedel R.. Nymble: a high-performance learning name-finder, Conference on Applied Natural Language Processing (1997)
3. Black W., Rinaldi F., Mowatt D. Facile: description of the NE system used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
4. Borthwick A., Sterling J., Agichtein E., Grishman R. 1997. Description of the MENE Named Entity System as used in MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
5. Brill E. A corpus-based approach to language learning. Doctoral Dissertation, Univ. of Pennsylvania (1993)
6. Chinchor N., MUC-7 Named Entity Task Definition, Version 3.5 (1997)
7. Cowie J. 1995. Description of the CLR/NMSU systems used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
8. Di Christo, P., S. Harie, C. De Loupy, N. Ide, and J. Veronis. Set of programs for segmentation and lexical look up, MULTEXT LRE 62-050 project Deliverable 2.2.1 (1995)
9. Gaizauskas R., Wakao T., Humphreys K., Cunningham H., Wilks Y. 1995. University of Sheffield: Description of the LaSIE system as used for MUC-6. Proceedings of Sixth Message Understanding Conference (1995)
10. Gallippi A., Learning to recognize names across languages. Proceedings of the 16th International Conference on Computational Linguistics (1996)
11. Grishman R. 1995. The NYU system for MUC-6 or where's the syntax. Proceedings of Sixth Message Understanding Conference (1995)
12. Grishman R., Tipster architecture design document version 2.3. Technical report, DARPA (1997)
13. Karkaletsis V., Spyropoulos C., Petasis G. Named entity recognition from Greek texts: the GIE project (1999)
14. Karttunnen L., The Replace Operator. In Finite State Language Processing, ed. Roche Em. and Schabes Yv., MIT Press (1997)
15. Krupka G., Hausman K. IsoQuest: description of the NetOwl extractor system as used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
16. Mikheev A., Grover C., Moens M. 1997. Description of the LTG System used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
17. Neumann G., Backofen R., Baur J., Becker M., Braun C. 1997. An information extraction core system for real world German text processing. ACL (1997)
18. Sekine S., Grishman R., Shinnou H.. A decision tree method for finding and classifying names in Japanese texts, Sixth Workshop on Very Large Corpora (1998)
19. Sekine S. NYU: description of the Japanese NE system used for MET-2. Proceedings of Seventh Message Understanding Conference (1998)
20. Yu S., Bai S., Wu P. Description of the Kent Ridge Digital Labs system used for MUC-7. Proceedings of Seventh Message Understanding Conference (1998)
21. Van Noord Gertjan and Dale Gerdemann. An Extendible Regular Expression Compiler for Finite-state Approaches in Natural Language Processing. WIA, Potsdam, Germany (1999)
22. Pazienza M.T. Ed., *Information Extraction: Multidisciplinary contributions to an emerging Information Technology*, Lecture Notes in Artificial Intelligence 1299, Springer-Verlag, Berlin Heidelberg, 1997