Machine Learning in Natural Language Processing

Georgios P. Petasis[*]

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
GR-153 10, P.O. BOX 60228, Aghia Paraskevi,
Athens, Greece
petasis@iit.demokritos.gr

Abstract. This thesis examines the use of machine learning techniques in various tasks of natural language processing, mainly for the task of information extraction from texts. The objectives are the improvement of adaptability of information extraction systems to new thematic domains (or even languages), and the improvement of their performance using as fewer resources (either linguistic or human) as possible. This thesis has examined two main axes: a) the research and assessment of existing algorithms of machine learning mainly in the stages of linguistic pre-processing (such as part of speech tagging) and named-entity recognition, and b) the creation of a new machine learning algorithm and its assessment on synthetic data, as well as in real world data for the task of relation extraction between named entities. This new algorithm belongs to the category of inductive grammar learning, and can infer context free grammars from only positive examples.

Keywords: information extraction, machine learning, grammatical inference.

1 Introduction

This doctoral thesis researches the possibility of exploiting machine learning techniques in the research area of natural language processing, aiming at the confrontation of the problems of upgrade as well as adaptation of natural language processing systems in new thematic domains or languages. The research is delimited in three important axes of information extraction systems:

- Part of speech recognition for the Greek language.
- Named entity recognition.
- Relation extraction between recognised named entities.

This thesis examines how machine learning methods and techniques can be exploited for the development of systems that support these tasks, which can be

[*] Dissertation Advisor: Constantine Halatsis, Professor

adapted more easily to new thematic domains and languages in contrast to the conventional systems that are rule based, manufactured often by experts. More specifically, this thesis researches techniques of machine learning along two main axes:

1. The application of existing techniques (both symbolic and statistical) in selected tasks of information extraction. These techniques are evaluated comparatively to each other in both the Greek and English languages. All existing machine learning algorithms that were examined require a vector of constant length as input. However the transformation of natural language into vectors of constant length is not always easy, without the use of arbitrary limits regarding the maximum number of words. This observation constituted the motivation for the creation of a new machine learning algorithm, which does not require vectors of constant length as input.
2. The development of a new machine learning algorithm, without the requirement for vectors of constant length as input. This new algorithm learns context free grammars from positive examples, with guidance via heuristics, such as minimum description length.

Regarding the first axis, named entity recognition systems were developed and evaluated, based on existing machine learning algorithms, such as decision trees and neural networks. The systems that were developed concern various thematic domains (management succession events, financial news, and juridical decisions) both in the Greek and English languages. These systems were evaluated in Greek texts, and they led to the recognition of the disadvantages and restrictions imposed by the examined algorithms, when applied on natural language data. From this analysis we concluded that one of the main problems when applying machine learning is the difficulty in managing data of variable length, as for example the information concerning all words of a sentence. On the contrary, a syntactic analyser can easily decide if a sentence (or part of a sentence) is described by a provided grammar. However, the manual development of grammars suitable for a specific task is a complex process, while the results frequently depend on the thematic domain and of course from the language. Consequently, if such a grammar can be automatically acquired with the use of machine learning, then the adaptation of systems that use such grammars to new thematic domains or languages can be considerably simplified.

The contribution of the developed systems is significant. The named entity recognition systems that were developed for the Greek language are among the first systems of their kind that have been reported in the bibliography. Simultaneously, the performance of the developed systems is satisfactory, and directly comparable to the performance of similar systems reported in the bibliography for the corresponding time period.

Regarding the second axis, and aiming at the confrontation of problems associated with the application of existing techniques, a new technique of machine learning has been developed. This new technique belongs to the category of inductive grammar learning. The main advantages of this method with respect to other machine learning methods are the ability to handle textual data, as well as the possibility of using learned grammars in existing systems, replacing manually developed grammars. The main objective of this new technique is the automatic grammar creation, which can be used with the plethora of available syntactic parsers that have been presented in the bibliography, replacing existing (and probably manually constructed) grammars for various tasks in information extraction systems.

For applying inductive grammar learning, a new algorithm has been developed that learns grammars from positive examples only. This new algorithm can infer context free grammars, and it has been based on the existing algorithm GRIDS [1], improving both the used heuristic, as well as the search process in the space of possible grammars, increasing simultaneously the applicability of the new algorithm to bigger collections of data. The requirement for the algorithm to function only with positive examples emanates from the frequent absence of negative examples in the area of natural language processing. It should be noted that the presence of negative examples constitutes a necessary condition for the operation of most existing grammatical inference algorithms. The design of this new algorithm has been done in such a way that it can be used in classification tasks, such as named entity recognition. This kind of usage differs from the usual application of grammatical inference algorithms, as the verification or the syntactic analysis of sentences according to a grammar is not required. Instead, we are interested mainly in recognising sentence parts (phrases) and their classification in predefined semantic categories. The evaluation of this new algorithm has been performed on both synthetic languages, as well as on real world data for the task of relation extraction between named entities.

2 Information Extraction

Information extraction (IE) is the task of automatically extracting structured information from unstructured documents, mainly natural language texts. Due to the ambiguity of the term ”structured information”, information extraction covers a broad range of research, from simple data extraction from Web pages using patterns and regular grammars, to the semantic analysis of language for extracting meaning, such as the research areas of word sense disambiguation or sentiment analysis. The basic idea behind information extraction (the concentration of important information from a document into a structured format, mainly in the form of a table) is fairly old, with early approaches appearing in the 1950s, where the applicability of information extraction was proposed by the Zellig Harris for sub-languages, with the first practical systems appearing at the end of the 1970s, such as Roger Schank's systems [2,3], which exported "scripts" from newspaper articles. The ease of evaluation of information extraction systems in comparison to other natural language processing technologies such as machine translation or summarisation, where evaluation is still an open research issue, made IE systems quite popular and led to the Message Understanding Conferences (MUC) [4] that redefined this research field. Information extraction can be decomposed into several sub-tasks:

- Linguistic preprocessing, responsible for tasks such as token/sentence identification, part-of-speech tagging, morphological analysis, etc.
- Named-entity recognition, where domain specific entities, such as names of persons, organisations, and locations, monetary/time expressions, etc., are identified.
- Co-reference resolution, where named entity names or other mentions (such as pronouns) that refer to the same entity are grouped/related.
- Template element filling, a task responsible for grouping all properties of a real object into a single template element that represents the real task or event.
- Template relation, a task responsible for identifying relations among template elements.
- Scenario template, a task where related template elements are combined into a template that represents an event.

This thesis has investigated the use of machine learning in three key sub-tasks of information extraction: part of speech tagging, named entity recognition, and relation extraction. Part of speech tagging is an important sub-task of linguistic preprocessing, named entity recognition is an essential subtask of information extraction, and relation extraction is the main activity behind template element filling, template relation and scenario template.

2.1 Part of speech tagging

The term "part of speech tagging" refers to the process of assigning a unique tag to every word in a document, in a way that the part of speech of each word can be identified from its tag. Several approaches regarding this task for the Greek language have been presented in the literature, including the approach of Dermatas and Kokkinakis [5], where Hidden Markov Models were used, achieving an accuracy of 95%, when trained on a corpus of 110.000 words. Orphanos and Christodoulakis [6] combined decision trees with a morphological lexicon, achieving a performance of 93-95% regarding disambiguation of ambiguous words according to the lexicon, and an accuracy of 82-88% for words unknown to the lexicon. Papageorgiou et. al. [7] employed transformation-based error-driven learning (TBED) combined with a morphological lexicon, achieving an accuracy of 96% when trained on a corpus of 356.000 words. Finally, Malakasiotis [8] used active learning, achieving an accuracy of 80%, when trained on a corpus of 15.300 words.

This thesis examined the applicability of transformation-based error-driven learning (TBED) [9] to the Greek language [10], and the combination of TBED with a morphological lexicon [11], [12]. The contribution of this thesis to the task of part of speech tagging is three-fold:

- A tag set that extends the Penn Tree Bank tag set, in order to include information about gender, number and verb tenses.
- The first publicly available part of speech tagger for the Greek language.

- The accuracy of the proposed method approaches 98% when combined with a morphological lexicon, which is the higher reported accuracy of part of speech tagging for the Greek language.

2.2 Named entity recognition

The task of named entity recognition and classification (NERC) tries to identify names of "entities" in documents, and classify identified names into predefined semantic categories, which usually vary according to the thematic domain. A typical NERC system can be decomposed into three major subtasks: linguistic pre-processing, a lexicon, and a grammar. Linguistic pre-processing relates to tasks similar to tokenisation, sentence splitting, part-of-speech tagging, etc. The lexicon includes domain specific information, usually in the form of lists of known named-entities (gazetteer lists). Finally, the grammar is responsible for recognising the entities that are either not in the lexicon or appear in more than one gazetteer lists (disambiguation). The manual adaptation of those two resources to a particular domain is a time-consuming and in some cases impossible process, due to the lack of experts.

This thesis examined the applicability of machine learning as a solution to the problems of domain adaptation and performance tunning, by examining the following cases:

- The substitution of the grammar sub-system with machine learning.
- The adaptation/enrichment of the lexicon.
- The detection of when a NERC system is outdated and needs to be adapted to the (possibly changed) domain.

Regarding the substitution of grammar with machine learning, two approaches have been studied. In the first approach various machine learning algorithms have been examined, including symbolic ones (such as decision trees [13] and sub-symbolic ones (such as feed-forward multi-layered perceptrons [14]), using a representation proposed by this thesis for representing variable-length named-entities as vectors of constant size. The algorithms were compared to existing manually constructed systems on two languages (the MITOS [15,16] system for the Greek language, and the VIE [17] system for the English language). The results for the best performing algorithm (C4.5) are shown in figure 1, compared to the results of the two manually constructed NERC systems for the two languages. The results show that the proposed approach outperforms the VIE system for English, but does not outperform the MITOS system for the Greek language. In addition, experiments showed that maintaining word order is not important, as representations that ignore word order achieve comparable or better performance than representations that maintain it.

The second approach examines the *combination* of machine learning algorithms, which exploit different kinds of input information. A set of classification algorithms that try to detect whether a word is part (or not) of a named entity, are combined through a majority voter in order to classify all words in a

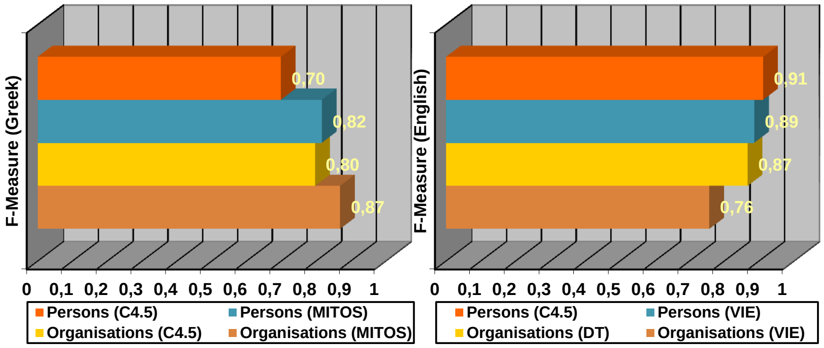


Fig. 1. Evaluation results of approach A, for the Greek and English languages.

document. In addition, noun phrases are identified, and classified as named entities through decision trees. The architecture of the second approach is shown in figure 2, where the evaluation results for Greek and English are shown in figure 3. Figure 4 shows the evaluation results of a NERC system developed by University of Edinburgh, UK, which is a hybrid system relying on manually constructed rules and machine learning [18,19] that has been trained and evaluated on the same English corpus as the ML-HNERC system (approach B). The results show that performance is higher for the Greek language compared to English, satisfying the objective for this system, which was motivated by the lower performance of approach A in Greek, compared to the English language. In addition, the performance of approach B on the English language is comparable to the performance of the system build by University of Edinburgh, which also build the top-scoring NERC system in the MUC-7 [20] competition. Maintaining the performance of a NERC system as a thematic domain timely evolves was also studied. The proposed approach makes an innovative use of machine learning, not to perform a task but rather to monitor the performance of another system. A machine learning based system (controller) is trained on the results of the system that will be monitored, and the deviation between the results of the two systems (the monitored and the control one) is recorded. As time passes and the thematic domain changes, the deviation between the two system is expected to change, as the two systems produce results based on possibly different input information. When the deviation exceeds a manually configured threshold, it is an indication that the monitored system is outdated, and needs to be updated. More details can be found in the thesis as well as in [21].

2.3 Relation extraction between recognised named entities

Relation extraction is the task of identifying semantic relations that hold between interesting entities in text data, and classify them into proper semantic categories. Being a challenging subtask of information extraction, it extracts the knowledge required to move from named entity recognition to data interpreta-

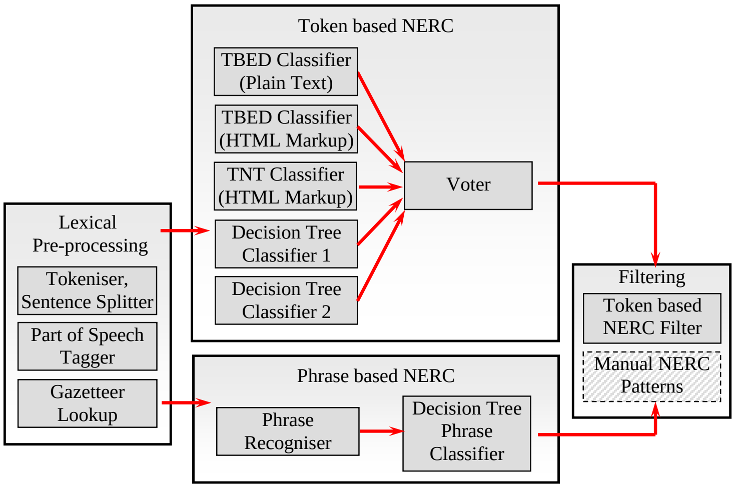

Token based NERC
TBED Classifier (Plain Text)
TBED Classifier (HTML Markup)
TNT Classifier (HTML Markup)
Decision Tree Classifier 1
Decision Tree Classifier 2
Voter
Lexical Pre-processing
Tokeniser, Sentence Splitter
Part of Speech Tagger
Gazetteer Lookup
Phrase based NERC
Phrase Recogniser
Decision Tree Phrase Classifier
Filtering
Token based NERC Filter
Manual NERC Patterns

Fig. 2. The architecture of the ML-HNERC system (approach B).

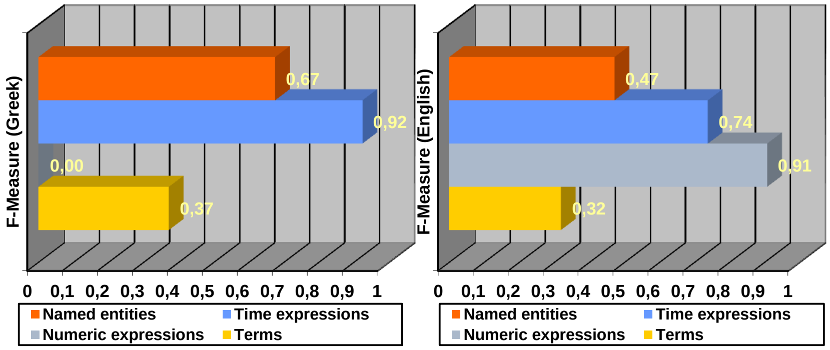


Fig. 3. Evaluation results of approach B, for the Greek and English languages.

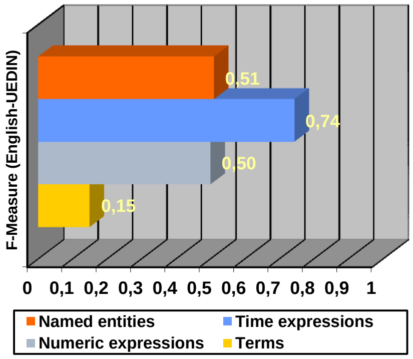


Fig. 4. Evaluation results of the system developed by University of Edinburgh, UK.

tion and understanding. The motivation behind the proposed approach is the simplification of the representations used in order to apply machine learning on natural language processing tasks, while the main objective is to examine the suitability of *grammatical inference* for the task of relation extraction. In this thesis, a supervised machine learning approach is proposed: assuming the existence of a named entity recogniser (NERC), the proposed approach extracts binary relations between named entities already identified in texts. Operating at the sentence level, a context-free grammar (CFG), which captures the patterns connecting related entities, is inferred from positive examples only. A new grammatical inference algorithm, eg-GRIDS+ ([22,23]), has been developed in order to infer a CFG only from positive examples. The need for negative feedback to control overgeneralisation, is eliminated through the use of minimum description length (MDL) [24].

The eg-GRIDS+ algorithm: A bias towards "simple" grammars As eg-GRIDS+ uses no negative evidence, an additional criterion is needed to direct the search through the space of context-free grammars and avoid overly general grammars. The approach of *minimum description length (MDL)* has been adopted in eg-GRIDS+, which directs the search process towards grammars that are compact, i.e., ones that require few bits to be encoded, while at the same time they encode the example set in a compact way, i.e. few bits are required to encode the examples using the grammar. Assuming a context-free grammar G and a set of examples (sentences) T that can be recognised (parsed) by the grammar G, the total description length of a grammar, henceforth *model description length* abbreviated as ML, is the sum of two independent lengths:

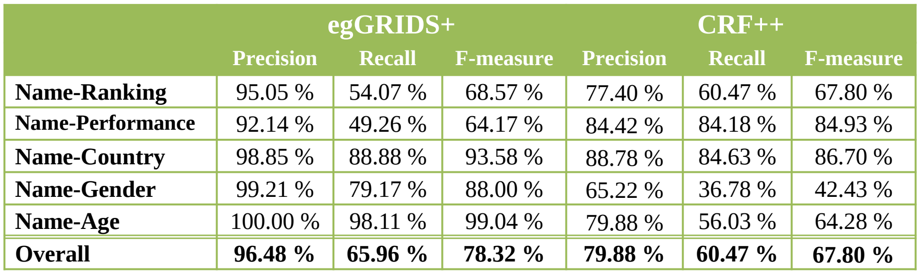
- The grammar description length (GDL), i.e. the bits required to encode the grammar rules and transmit them to a recipient who has minimal knowledge of the grammar representation, and
- The derivations description length (DDL), i.e. the bits required to encode and transmit all examples in the set T as encoded by grammar G, provided that the recipient already knows G.

The first component of the ML directs the search away from the sort of trivial grammar that has a separate rule for each training sentence, as this grammar will have a large GDL. However, the same component leads to another sort of trivial grammar, a grammar that accepts all sentences. In order to avoid this, the second component estimates the *derivation power* of the grammar, by measuring the way the *training examples* are generated by the grammar, and helps to avoid overgeneralisation by penalising general grammars. The higher the derivation power of the language, the higher its DDL is expected to be. The initial overly specific grammar is trivially best in terms of DDL, as usually there is a one-to-one correspondence between the examples and the grammar rules, i.e. its derivation power is low. On the other hand, the most general grammar has the worst score, as it involves several rules in the derivation of a single sentence, requiring substantial effort to track all the rules involved in the generation of the sentence.

Architecture of eg-GRIDS+ and the learning operators eg-GRIDS+ uses the training sentences in order to construct an *initial, "flat" grammar*. Then eg-GRIDS+ generalises this initial grammar, using one of the two available iterative search processes: beam or genetic search. Both search strategies utilise the same search operators in order to produce more general grammars:

- Merge NT: merges two non-terminal symbols into a single symbol, thereby replacing all their occurrences in all rules with the new symbol.
- Create NT: creates a new non-terminal symbol X, which is defined as a sequence of two or more existing non-terminal symbols. X is defined as a new production rule that decomposes X into its constituent symbols.
- Create Optional NT: duplicates a rule created by the "Create NT" operator and appends an existing non-terminal symbol at the end of the body of the rule, thus making this symbol optional.
- Detect Center Embedding: aims to capture the center embedding phenomenon. This operator tries to locate the most frequent four-gram of the form "A A B B", which is replaced by a new non-terminal symbol X and a new rule of the form "$X \rightarrow A\,A\,X\,B\,B$". All symbol sequences that match the pattern "$A+\ X?\,B+$" are replaced with X.
- Rule Body Substitution: examines whether the body of a production rule R is contained in bodies of other production rules. In such a case, every occurrence of the body of rule R in other rule bodies is replaced by the head of rule R.

Evaluation For the purposes of the evaluation, annotated corpus from the BOEMIE research project was used, which contained 800 HTML pages, retrieved from various sites of athletics associations, containing pages with news, results and athlete's biographies. Evaluation was performed through 10 fold-cross validation, and performance was measured in terms of precision, recall and F-measure. The obtained performance results are shown in table 1. Evaluation results suggest that the proposed approach performs well in comparison to the state of the art, despite the difficulties of comparing results obtained on different



	egGRIDS+			CRF++		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Name-Ranking	95.05 %	54.07 %	68.57 %	77.40 %	60.47 %	67.80 %
Name-Performance	92.14 %	49.26 %	64.17 %	84.42 %	84.18 %	84.93 %
Name-Country	98.85 %	88.88 %	93.58 %	88.78 %	84.63 %	86.70 %
Name-Gender	99.21 %	79.17 %	88.00 %	65.22 %	36.78 %	42.43 %
Name-Age	100.00 %	98.11 %	99.04 %	79.88 %	56.03 %	64.28 %
Overall	**96.48 %**	**65.96 %**	**78.32 %**	**79.88 %**	**60.47 %**	**67.80 %**

Table 1. Relation extraction performance results.

corpora. For example, in [25], the presented approach, expanding on a basis of 55 manually constructed seed rules, exhibits precision around 88% with 43% recall on 1032 news reports on Nobel prizes from New York Times, BBC and CNN. In addition, Conditional Random Fields (CRF++) were applied on the same corpus, achieving lower results than the approach based on eg-GRIDS+.

3 Conclusions

This thesis proposes the exploitation of machine learning in nodal points of a typical information extraction system, having as aim the assistance adapting the system into new thematic domains and perhaps languages. This first research topic of this thesis involves part of speech tagging for the Greek language. Transformation-based error-driven learning (TBED) has been applied for the first time in the Greek language, achieving high performance, directly comparable with corresponding systems for the Greek language, requiring at the same time considerably less training data. Simultaneously, the approach that is described in this thesis constituted the first Greek part-of-speech tagger that has been distributed freely, as an open source application, with important acceptance from the scientific community, as denoted by the number of citations in the relative publications.

The second research topic concerns the area of named entity recognition. Three machine learning algorithms were examined for this task, both symbolic and stochastic ones, achieving satisfactory results. The algorithms were examined in various thematic domains, in both the English and Greek languages, confirming not only the ability of machine learning to support the task of named entity recognition, but also the adaptability of the machine learning based approaches not only in new thematic domains, but also in languages. The research that has been contacted in the context of this thesis is included among the first information extraction systems for the Greek language that have been reported in the bibliography. In addition, it has been observed that, at least for the task of named entity recognition, the order of words in a sentence is not important. Despite the fact that initially this observation seemed surprising, the widespread use of the "bag-of-words" representation - which also ignores the word order - not only for named entity recognition, but also for other natural language processing tasks, shows the correctness of this initial observation.

The third research topic concerns the development of new machine learning algorithm, able to infer context free grammars from positive only examples. An important characteristic of this new algorithm is its ability to processes large volumes of data, a consequence of the observation that it is computationally cheaper to predict the result of applying a learning operator, than to apply the operator and evaluate the produced grammar. The results achieved by the new algorithm on the "Omphalos" [26] competition were also significant, where it solved the first problem without human intervention within the competition time period, while it has been successfully combined with the winning algorithm, removing the need of the winning algorithm for human intervention, in cases where its heuristic could not drive further the learning process.

References

1. Langley, P., Stromsten, S.: Learning context-free grammars with a simplicity bias. In: Proceedings of the 11th European Conference on Machine Learning. ECML '00, London, UK, UK, Springer-Verlag (2000) 220–228
2. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures. L. Erlbaum, Hillsdale, NJ (1977)
3. Schank, R.C., Kolodner, J.L., DeJong, G.: Conceptual information retrieval. In: Proceedings of the 3rd annual ACM conference on Research and development in information retrieval (SIGIR '80), Cambridge, UK (1980) 94–116
4. Marsh, E., Perzanowski, D.: Muc-7 evaluation of ie technology: Overview of results. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html` (1998)
5. Dermatas, E., Kokkinakis, G.K.: Automatic stochastic tagging of natural language texts. Computational Linguistics **21**(2) (1995) 137–163
6. Orphanos, G.S., Christodoulakis, D.N.: Pos disambiguation and unknown word guessing with decision trees. Computer Engineering (1999) 134–141
7. Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S.: A unified pos tagging architecture and its application to greek. In: Proceedings of the 2nd Language Resources and Evaluation Conference, Athens, European Language Resources Association (June 2000) 1455–1462
8. Malakasiotis, P.: Αναγνώριση μερών του λόγου σε ελληνικά κείμενα με τεχνικές ενεργητικής μάθησης (Part-of-speech tagging in Greek texts using active learning techniques). Master's thesis, Department of Informatics, Athens University of Economics and Business (2005)
9. Brill, E.: Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput. Linguist. **21**(4) (December 1995) 543–565
10. Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D.: Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques. In: Proceedings of the ECCAI Advanced Course on Artificial Intelligence (ACAI '99), Chania, Greece (July 5–16 1999)
11. Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, I., Spyropoulos, C.D. In: A Greek Morphological Lexicon and Its Exploitation by Natural Language Processing Applications. Volume 2563 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2003) 401–419 http://www.springerlink.com/content/hcdjrlvj5nlybf5c/.
12. Spiliotopoulos, D., Petasis, G., Kouroupetroglou, G.: Prosodically Enriched Text Annotation for High Quality Speech Synthesis. In: Proceedings of the 10th International Conference on Speech and Computer (SPECOM-2005), Patras, Greece (October 17–19 2005) 313–316
13. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
14. Perantonis, S.J., Ampazis, N., Virvilis, V.: A learning framework for neural networks using constrained optimization methods. Annals of Operations Research **99**(1) (2000) 385–401

15. Petasis, G., Petridis, S., Paliouras, G., Karkaletsis, V., Perantonis, S.J., Spyropoulos, C.D.: Symbolic and Neural Learning for Named-Entity Recognition. In: Proceedings of European Best Practice Workshops and Symposium on Computational Intelligence and Learning (COIL 2000), Chios, Greece (June 19–23 2000) 58–66
16. Petasis, G., Petridis, S., Paliouras, G., Karkaletsis, V., Perantonis, S.J., Spyropoulos, C.D. In: Symbolic and Neural Learning of Named-Entity Recognition and Classification Systems in Two Languages. Volume 18 of International Series in Intelligent Technologies. Springer Berlin / Heidelberg (January 2002) 193–210 http://www.springer.com/mathematics/book/978-0-7923-7645-3.
17. Humphreys, K., Gaizauskas, R., Cunningham, H., Azzam, S.: GATE: VIE technical specifications. Technical report, ILASH, University of Sheffield (1997) Included in the documentation of GATE 1.0.0.
18. Curran, J.R., Clark, S.: Investigating gis and smoothing for maximum entropy taggers. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1. EACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 91–98
19. Curran, J.R., Clark, S.: Language independent ner using a maximum entropy tagger. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 164–167
20. Chinchor, N.A.: Proceedings of the Seventh Message Understanding Conference (MUC-7) named entity task definition. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, VA (April 1998) 21 pages version 3.5, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.
21. Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D.: Using Machine Learning to Maintain Rule-based Named - Entity Recognition and Classification Systems. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. ACL '01, Toulouse, France, Association for Computational Linguistics (July 9–11 2001) 426–433
22. Petasis, G., Paliouras, G., Karkaletsis, V., Halatsis, C., Spyropoulos, C.D.: E-GRIDS: Computationally Efficient Grammatical Inference from Positive Examples. GRAMMARS **7** (2004) 69–110 Technical Report referenced in the paper: http://www.ellogon.org/petasis/bibliography/GRAMMARS/GRAMMARS2004-SpecialIssue-Petasis-TechnicalReport.pdf.
23. Petasis, G., Paliouras, G., Spyropoulos, C.D., Halatsis, C.: Eg-GRIDS: Context-Free Grammatical Inference from Positive Examples Using Genetic Search. In Paliouras, G., Sakakibara, Y., eds.: Grammatical Inference: Algorithms and Applications, Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI 2004). Volume 3264 of Lecture Notes in Computer Science., Athens, Greece, Springer Berlin / Heidelberg (October 11–13 2004) 223–234
24. Rissanen, J.: Stochastic Complexity in Statistical Inquiry Theory. World Scientific Publishing Co., Inc., River Edge, NJ, USA (1989)
25. Xu, F., Uszkoreit, H., Li, H.: A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (June 2007) 584–591
26. Starkie, B., Coste, F., van Zaanen, M.: The omphalos context-free grammar learning competition. In Paliouras, G., Sakakibara, Y., eds.: Grammatical Inference: Algorithms and Applications, Proceedings of the 7^{th} International Colloquium, ICGI 2004, Athens, Greece, October 11-13, 2004. Volume 3264 of Lecture Notes in Computer Science., Springer (2004) 16–27