Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems

Georgios Petasis †, Frantz Vichot §, Francis Wolinski §
Georgios Paliouras †, Vangelis Karkaletsis † and Constantine D. Spyropoulos †

† Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos",
153 10 Ag. Paraskevi, Athens, Greece

§ Informatique-CDC
4, rue Berthollet
94114 Arcueil, France

{petasis,paliourg,vangelis,costass}@iit.demokritos.gr
{frantz.vichot, francis.wolinski}@caissedesdepots.fr

Abstract

This paper presents a method that assists in maintaining a rule-based named-entity recognition and classification system. The underlying idea is to use a separate system, constructed with the use of machine learning, to monitor the performance of the rule-based system. The training data for the second system is generated with the use of the rule-based system, thus avoiding the need for manual tagging. The disagreement of the two systems acts as a signal for updating the rule-based system. The generality of the approach is illustrated by applying it to large corpora in two different languages: Greek and French. The results are very encouraging, showing that this alternative use of machine learning can assist significantly in the maintenance of rule-based systems.

1 Introduction

Machine learning has recently been proposed as a promising solution to a major problem in language engineering: the construction of lexical resources. Most of the real-world language engineering systems make use of a variety of lexical resources, in particular grammars and lexicons. The use of general-purpose resources is ineffective, since in most applications a specialised vocabulary is used, which is not supported by general-purpose lexicons and grammars. For this reason, significant effort is currently put into the construction of generic tools that can quickly adapt to a particular thematic domain. The adaptation of these tools mainly involves the adaptation of domain-specific semantic lexical resources.

Named-entity recognition and classification (NERC) is the identification of proper names in text and their classification as different types of named entity (NE), e.g. persons, organisations, locations, etc. This is an important subtask in most language engineering applications, in particular information retrieval and extraction. The lexical resources that are typically included in a NERC system are a lexicon, in the form of gazetteer lists, and a grammar, responsible for recognising the entities that are either not in the lexicon or appear in more than one gazetteer lists. The manual adaptation of those two resources to a particular domain is time-consuming and in some cases impossible, due to the lack of experts. The exploitation of learning techniques to support this adaptation task has attracted the attention of researchers in language engineering.

However, the adaptation of lexical resources to a specific domain at a certain point in time is not sufficient on its own. The performance of a NERC system degrades over time (Vichot et al., 1999; Wolinski et al., 2000) due to the introduction of new NEs or the change in the meaning of existing ones. We need to find ways that facilitate the maintenance of rule-based NERC systems. This paper presents such a method, exploiting machine learning in an innovative way. Our method controls rule-based NERC systems with NERC systems constructed by a machine learning algorithm. The method comprises two stages: the *training stage*, during which a super-

vised machine learning algorithm constructs a new system using data generated by the rule-based system, and the *deployment stage*, in which the results of the two systems are compared on new data and their disagreements are used as signals for change in the rule-based system. Note that, unlike most applications of supervised machine learning, the training data for the new system are not produced manually.

In order to illustrate the generality of this approach, we have tested it with two different NERC systems, one for Greek and another one for French. The results are very encouraging and show that machine learning techniques can be used for the maintenance of rule-based systems.

Section 2 presents existing work on the domain adaptation of NERC systems using machine learning (ML) techniques. Section 3 presents the two rule-based NERC systems for Greek and French. Section 4 explains our method and Section 5 describes the two experiments and presents the evaluation results. Finally, Section 6 concludes and presents our future plans.

2 Related Work

As mentioned above, the exploitation of learning techniques to support the domain adaptation of NERC systems has recently attracted the attention of several researchers. Some of these approaches are briefly discussed in this section.

Nymble (Bikel et al., 1997) uses statistical learning to acquire a Hidden Markov Model (HMM) that recognises NEs in text. Nymble did particularly well in the MUC-7 competition (DARPA, 1998), due mainly to the use of the correct features in the encoding of words, e.g. capitalisation, and the probabilistic modelling of the recognition system.

Named-entity recognition in Alembic (Vilain and Day, 1996) uses the transformation-based rule learning approach introduced in Brill's work on part-of-speech tagging (Brill, 1993). An important aspect of this approach is the fact that the system learns rules that can be freely intermixed with hand-engineered ones.

The RoboTag system presented in (Bennett et al., 1997) constructs decision trees that classify words as being start or end points of a particular named-entity type. A variant of this approach was used in the system presented by the New York University (NYU) in the Multilingual Entity Task (MET-2) of MUC-7 (Sekine, 1998).

The system developed for Italian in ECRAN (Cuchiarelli et al., 1998), uses unsupervised learning to expand a manually constructed system and improve its performance. The learning algorithm tries to supplement the manually constructed system by classifying recognised but unclassified NEs. In (Petasis et al., 2000) the manually constructed system was replaced by the supervised tree induction algorithm C4.5 (Quinlan, 1993), reaching very good performance on the MUC-6 corpora.

The partially supervised multi-level bootstrapping approach presented in (Riloff and Jones, 1999) induces a set of information extraction patterns, which can be used to identify and classify NEs. The system starts by generating exhaustively all candidate extraction patterns, using an earlier system called AutoSlog (Riloff, 1993). Given a small number of seed examples of NEs, the most useful patterns for recognising the seed examples are selected and used to expand the set of classified NEs. The end result is a dictionary of NEs and the extraction patterns that correspond to them.

Our method follows an alternative innovative approach to the use of learning for NERC. Instead of using ML to construct a NERC system that will be used autonomously, the system constructed by ML, according to our approach is used to monitor the performance of an existing rule-based NERC system. In this manner, the new system provides feedback on whether the rule-based system under control has become obsolete and needs to be updated. An important advantage of this approach is that no manual tagging of training data is needed, despite the use of a supervised learning algorithm.

Our method bears some similarities with systems based on active learning (Thompson et al., 1999). According to this technique, multiple classifiers performing the same task are used in order to actively create training data, through their disagreements. Usually, this involves an iterative procedure. First a few initial labelled examples are used to train the classifiers and then, unlabelled examples are presented to the classifiers. Examples that cause the classifiers to disagree are good candidates to retrain the classifiers on. The difference of active learning to our method is the use of a manually-constructed

rule-based NERC system as the basic system. The ML method is used only to identify when the rule-based NERC system should be updated, but not for creating new training instances. Another approach, which bears some similarity to ours, is presented in (Kushmerick, 1999) where a heuristic algorithm is used to monitor the performance of web-page wrappers.

3 Rule-based NERC Systems

A typical NERC system consists of a lexicon and a grammar. The lexicon is a set of NEs that are known beforehand and have been classified into semantic classes. The grammar is used to recognize and classify NEs that are not in the lexicon and to decide upon the final classes of NEs in ambiguous cases.

Manual construction of NERC systems is a complicated and time-consuming process, even for experts. The meaning of a single sentence may vary a lot according to which category a NE is assigned to. For example, the sentence "*Express group intends to sell Le Point for 700 MF*" indicates a sale of a newspaper company, if *"Le Point"* is classified as an organisation. Whereas the following sentence, which is grammatically identical to the previous one, *"Compagnie des Signaux intends to sell TVM430 for 700 MF"* gives only a price for an industrial product.

In order for a NERC system to be able to recognise and categorise correctly NEs, both the lexicon and the grammar have to be validated on large corpora, testing their efficiency and their robustness. However, this process does not ensure that the performance of the developed system will remain steady over time. Almost under all thematic domains, the introduction of new NEs or the change in the meaning of existing ones can increase the error rate of the system. Our approach tries to identify such cases, facilitating the maintenance of the NERC system.

The following subsections briefly describe the Greek and French rule-based NERC systems that have been used in our experiments.

3.1 The Greek NERC System

The Greek NERC system (Farmakiotou et al., 2000) used for the purposes of this experiment forms part of a larger Greek information extraction system, being developed in the context of the R&D project MITOS.[1] The NERC component of this system mainly consists of three processing stages: linguistic pre-processing, NE identification and NE classification. The linguistic pre-processing stage involves some basic tasks: tokenisation, sentence splitting, part-of-speech tagging and stemming. Once the text has been annotated with part of speech tags, a stemmer is used. The aim of the stemmer is to reduce the size of the lexicon as well as the size and complexity of the NERC grammar.

The NE identification stage involves the detection of their boundaries, i.e., the start and the end of all the possible spans of tokens that are likely to belong to a NE. Identification consists of three sub-stages: initial delimitation, separation and exclusion. Initial delimitation involves the application of general patterns. These patterns are combinations of a limited number of words, selected types of tokens (e.g. tokens consisting of capital characters), special symbols and punctuation marks. At the separation sub-stage, possible NEs that are likely to contain more than one NE or a NE attached to a non-NE, are detected and attachment problems are resolved. Finally, at the exclusion sub-stage two types of criteria are used for exclusion from the possible NE list: the context of the phrase and being part of an exclusion list. Suggestive context for exclusion consists of common names that refer to products, services or artifacts. The exclusion list includes capitalized abbreviations of common nouns, financial terms, capitalized person titles, which are not ambiguous, and nouns commonly found in names of products, artifacts and services.

Once the possible NEs have been identified, the classification stage begins. Classification involves three sub-stages: application of classification rules, gazetteer-based classification, and partial matching of classified named-entities with unclassified ones. Classification rules take into account both internal and external evidence (McDonald, 1996), i.e., the words and symbols that comprise the possible name and the context in which it occurs. Gazetteer-based classification involves the look up of pre-stored lists of known proper names (gazetteers). The gazetteers contain stemmed forms and have been compiled from Web sites and an annotated train-

[1] http://www.iit.demokritos.gr/skel/mitos

ing corpus. The size of the gazetteers is rather small (3,059 names). At the partial matching sub-stage, classified names are matched against unclassified ones aiming at the recognition of the truncated or variable forms of names.

3.2 The French NERC System

The French NERC system has been implemented with the use of a rule-based inference engine (Wolinski et al., 1995). It is based on a large knowledge base (lexicon) including 8,000 proper names that share 10,000 forms and consist of 11,000 words. It has been used continuously since 1995 in several real-time document filtering applications (Wolinski et al., 2000). The uses of the NERC system in these applications are the following:

1. **Segmentation** of NEs, in order to improve the performance of the syntactic analyser, particularly in the case of long proper names which contain grammatical markers (e.g. prepositions, conjunctions, commas, full stops).
2. **Recognition** of known NEs in order to supply precise information to a document filtering module.
3. **Classification** of NEs in order to feed a document filtering module with information dealing with the very nature of the NEs quoted in the documents.

The NERC system tries to classify each NE in one of four different categories: association (non-commercial organisation), person, location or company.

For the classification of known entities, a crucial problem appears when several NEs share a single form. To deal with these cases, two sets of rules have been implemented:

1. **Local context**: For instance, "*Saint-Louis*" may be interpreted in one of the following ways: the capital of Missouri, a French group in the food production industry, a small industry "*les Cristalleries de Saint Louis*", a small town in France, a hospital in Paris, etc. Exploration of the local context using the proper name may enable, in certain cases, a choice to be made between these various interpretations. If the text speaks of "*St-Louis (Missouri)*", only the first interpretation should be adopted. In order to do this the knowledge base should contain information that "*Saint-Louis*" is in Missouri, and a rule should exist to interpret the affixing of a parenthesis.
2. **Global context**: Abbreviated NEs and acronyms are much more frequent sources of ambiguity and are almost always common to several NEs. In general, such ambiguous forms of NEs do not occur on their own in news but almost always together with non-ambiguous forms that enable the ambiguity to be removed. For instance, if the NEs "*Saint-Louis*" and "*Hôpital Saint-Louis*" appear in a single news item, the interpretation corresponding to the hospital is more likely to be the one that should be adopted.

For unknown entities, three sets of rules have been implemented:

1. **Prototypes**: Many NEs are constructed according to some prototypes. These can be categorised using pattern matching rules. *Mr André Blavier*, *Kyocera Corp*, *Condé-sur-Huisne*, *Honda Motor*, *IBM-Asia*, *Bernard Tapie Finance*, *Siam Nissan Automobile Co Ltd* are good examples of such prototypes.
2. **Local context**: Many single-word unknown NEs (some known NEs as well) may also be categorised using the local context. For instance, the small sentences "*Peskine, director of the group*", "*the shareholders of Fibaly "* or "*the mayor of Gisenyi*" are used as categorisation rules.
3. **Global context**: After the first appearance of a NE in full, its head (e.g. family name, main company) is often used alone in the text instead of the full name. The company *Kyocera Corp*, for example, may be designated by the single word *Kyocera* in the remainder of the text. For each such unknown word, starting with a capital letter, a special rule examines whether it appears inside another NE in the text.

4 Controlling a Rule-based System Using Machine Learning

Machine learning has been used successfully to control a rule-based system that performs a different task, namely document filtering (Wolinski et al., 2000). The learning method used in that case was a neural network (Stricker et al., 2001).

In our present study, we control the rule-based NERC systems that have been presented in section 3, with NERC systems constructed by the C4.5 algorithm. Our method comprises two stages: the *training stage*, during which C4.5 constructs a new system using data generated by the rule-based system, and the *deployment stage*,

in which the results of the two systems are compared on new data and their disagreements are used as signals for change in the rule-based system. This section describes the basic principles of our control method.

4.1 Control method: training stage

The training stage of our method consists of the following processing steps (Figure 1):

Running the rule-based NERC system on a large training corpus (containing several thousands of NEs in our case). The aim of this process is to recognise and classify the NEs in the corpus. The end product is a set of NEs, associated with their class.

Constructing a separate NERC system by applying C4.5 on the data generated by the rule-based system. In this process, the classified NEs are used as training data by C4.5, in order to construct the second NERC system (trained NERC). For each classified NE a training example (vector) is created, containing information about the part of speech and gazetteer tags of the first and the last two words of the NE, as well as the two words preceding and the two following the NE. It is important to note that, unlike other uses of supervised machine learning methods, this approach does not require manual tagging of training data.

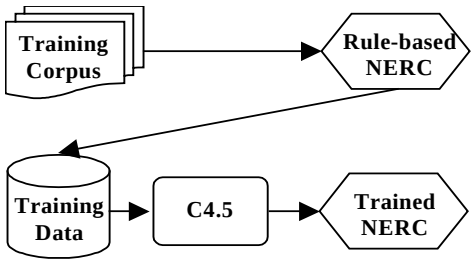

Figure 1: Training stage.

4.2 Control method: deployment stage

In the deployment stage, the two NERC systems are compared on a new corpus to identify disagreements. Despite the fact that the second method is trained on data generated by the first, the different nature of the NERC system generated by C4.5, i.e., a decision tree, leads to interesting disagreements between the two methods. The deployment stage consists of the following processing steps (Figure 2):

1. Running the rule-based NERC system on a new corpus. It should be stressed here that the documents in this corpus differ in some characteristic way from those in the training corpus. In our experiments the difference is chronological, i.e., the new corpus consists of recent news articles. The reason for adopting this approach is that we are interested in the maintenance of a rule-based system through time. An alternative approach might be for the new corpus to be from a slightly different thematic domain. In that case, the goal of the process would be the customisation of the rule-based system to a new domain.
2. Running the trained NERC system on the same corpus.
3. Comparing the results provided by both systems to identify cases of disagreement. The result is a set of data where the two systems disagree: in our case, disagreements deal with the different categories assigned by the NERC systems to NEs (see Section 5 for detailed results). These cases are then provided to the language engineer, who needs to evaluate them and decide on changes for the rule-based system.

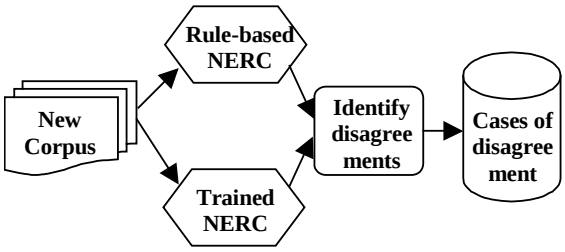

Figure 2: Deployment stage.

5 Results

In order to evaluate the proposed method, two different experiments were contacted, one for each language. The exact experimental settings as well as the evaluation results are presented in the following sections.

5.1 Results for the Greek System

For the experiment regarding the Greek language, we used three NE classes: organisations, persons and locations. For the purposes of the experiment, two corpora of financial news were used.[2] The first corpus that was used for training purposes, consisted of 5,000 news articles from the years 1996 and 1997, containing 10,010 instances of NEs (1,885 persons, 1,781 locations, 6,344 organisations). The second corpus

[2] The corpora were provided by the Greek publishing company Kapa-TEL.

that was used for evaluation purposes consisted of 5,779 news from the years 1999 and 2000 and contained 11,786 instances of NEs (1,137 persons, 810 locations, 9,839 organisations).

5.1.1 Aggregate Results

A good way to give an overview of the cases of disagreement of the two systems is through a contingency matrix, as shown in Table 1. The rows of this table correspond to the classification of the rule-based system, while the columns to the classification of the system constructed by C4.5.

Table 1: Overview of the results for Greek.

	organisation.	person	location
organisation	9,906	250	32
person	230	649	14
location	24	6	675

As we can see from Table 1, in 95% of the cases the two systems are in agreement. This means, that in order to update the rule-based NERC system, we have to examine only 5% of the cases, where the two systems disagree. Examining these cases gave us important insight regarding problems of the rule-based NERC system. Some examples are presented in the following sections.

5.2 Recognition problems

The examination of cases in disagreement revealed some interesting problems regarding NE recognition. These problems concern NEs that the rule-based system identified only partially and as a result classified them incorrectly.

For example, in the stage of initial delimitation, the general patterns fail to identify NEs that contain numbers in their names, like the organisation "Αθήνα 2004" (Athens 2004) representing the organising committee of 2004 Olympics.

In addition, during the separation phase some of the rules have not taken into account some inflexional endings, causing failures in separating some NEs. For example, in the phrase "ο υφ. Πολιτισμού Γ. Φλωρίδης" (the under-secretary of Culture Γ. Φλωρίδης) the recogniser failed to separate the person name from its title, due to the last accented character of the word "Πολιτισμού".

Finally, we were able to locate several stopwords and update our exclusion list. For instance, the phrase "γραμμών ISDN" (ISDN lines) was recognised as an organisation (as the word "γραμμών" is a frequent constituent of airline or shipping companies), but in reality the text was referring to ISDN telephone lines.

5.2.1 Classification problems

Except from the problems identified in the recognition phase, the examination of the cases of disagreement revealed various problems regarding mainly the classification grammar. In fact, some of our classification rules were found to be too general, leading to wrong classifications.

For example, according to one of the rules, a sequence of two words, starting with capital letters, constitutes a person name if it is preceded by a definite article and the endings of these two words belong in a specific set that usually denote person names. This rule caused the classification of various non-NEs as persons, including "του Ολυμπιακού Χωριού" (the Olympic Village).

Another example of an overly general rule is a rule that classifies a sequence of abbreviations or nouns starting with capital letter as an organisation, if this sequence is preceded by a comma that in turn is preceded by a NE already classified as an organisation. This rule caused the classification of few person names as organisations, such as "ο διοικητής της Εθνικής Τράπεζας, Θ.Καρατζάς" (the director of National Bank, Θ.Καρατζάς).

5.3 Results for the French System

The corpus used for the French experiment contained dispatches from the Agence France-Presse from April 1998 until January 2001. The thematic domain of the corpus was shareholding events. This corpus contained six thousand documents, including 180,983 instances of NEs with the following distribution: companies (45%), locations (45%), persons (7%) and associations (non commercial organisations) (3%). For the purposes of this experiment, the corpus was chronologically split in two parts. The part containing the chronologically earlier messages was used for training purposes while the second part, containing the most recent messages, was used in order to evaluate our approach. In this experiment, we mainly focused on four NE categories, instead of the three categories used for the Greek experiment. This differentiation

originates in the fact that the French NERC system further categorises organisations into associations (non-profit organisations) and companies.

5.3.1 Aggregate Results

The contingency matrix giving an overview of the cases of disagreement of the two systems is shown in Table 2. It appears that in 91% of the cases the two systems are in agreement.

Table 2: Overview of the results for French.

	associat.	person	location	company
associat.	808	6	31	618
person	3	4,498	46	509
location	11	51	6,870	2,526
company	296	67	534	34,946

Examining the disagreement cases gave us important insight regarding problems of the rule-based system. The following sections present some interesting examples.

5.3.2 Recognition problems

Similarly to the Greek experiment, the examination of disagreements revealed some interesting problems in the recognition of NEs. For instance, “Europe 1” is a well-known French radio station, also written sometimes as “Europe Un” (Europe One). The rule-based system failed to identify “Europe Un” and only identified “Europe” as a location. The source of the problem is the lack of a mapping between fully written numbers and numerical figures.

Another example is the phrase “Le Mans Re”, which is a shortened version of the company name “Les mutuelles du Mans Reassurance” (a Reinsurance company). The rule-based system recognised only “Le Mans” as a location, due to the well-known French city. What is needed here is an extension of the segmentation rules to include “Re” as a “company designator”, such as “Motor”, “Bank” or “Telecom”.

5.3.3 Classification problems

Most of the classification problems that were identified concerned NEs already known to the system that meanwhile have acquired new meanings. For example, “Ariane II rachète” (Ariane II buys) is classified as a person, due to the word “Ariane” contained in the lexicon as a person forename. In reality, “Ariane II” is a new company that should also be included in the lexicon database. Another example is “Orange” already included in the lexicon as an old French city. In the meanwhile, a new French company has been created having the same name, as in the example “Orange, valorisée par les analystes” (Orange, estimated by analysts). Also in this case, the lexicon must be updated with a second entry for this entity, categorised as a company.

Besides lexicon omissions, some problems regarding the classification grammar were also revealed. First, overly general rules were identified, such as the one that classifies entities starting from “A” and followed by numbers as French highway names. This rule wrongly classified the NE “A3XX” as a highway, while the text was referring to an airplane model: “L’A3XX, un avion” (The A3XX, an air plane).

Our approach also succeeded in locating well-known NEs used in a new context. For example, the rule-based NERC system recognises “Taittinger” as a company while the system learned by C4.5 disagrees with this classification in the sentence “la famille Taittinger” (the family Taittinger). In this case, the grammar should be updated with a rule saying that the word “family” in front of a proper name suggests a person name.

6 Conclusions

In this paper, we have proposed an alternative use of machine learning in named-entity recognition and classification. Instead of constructing an autonomous NERC system, the system constructed with the use of machine learning assists in the maintenance of a rule-based NERC system. An important feature of the approach is the use of a supervised learning method, without the need for manual tagging of training data. The proposed approach was evaluated with success for two different languages: Greek and French.

On-going work aims at reducing the number of disagreements between the two systems down to those that are essential for the improvement of the system. Currently, there are many cases where the two systems disagree, but the rule-based system is correct.

Another extension that we are examining is to train a NERC system to not only classify, but also recognise NEs. We believe that this exten-

sion will lead to the identification of more problematic cases in the recognition phase.

In conclusion, the method presented in this paper proposes a simple and effective use of machine learning for the maintenance of rule-based systems. The scope of this approach is clearly wider than that examined here, i.e., named-entity recognition.

Acknowledgements

This research has been carried out thanks to the Hellenic – French scientific cooperation project ADIET (PLATON no. 00521 TH). It also used results of the Greek R&D project MITOS (EPET II – 1.3 – 102).

References

Bennett S.W., Aone C. and Lovell C., 1997. Learning to Tag Multilingual Texts through Observation. *Proc. of the Second Conference on Empirical Methods in NLP*, pp. 109-116.

Bikel D., Miller S., Schwartz R. and Weischedel R., 1997. Nymble: a High-Performance Learning Name-finder. *Proc. of 5th Conference on Applied natural Language Processing*, Washington.

Defense Advanced Research Projects Agency, 1998. *Proc. of the Seventh Message Understanding Conference (MUC-7)*, Morgan Kaufmann.

Brill E., 1993. A corpus-based approach to language learning. *PhD Dissertation*, Univ. of Pennsylvania.

Cuchiarelli A., Luzi D., and Velardi P., 1998. Automatic Semantic Tagging of Unknown Proper Names. *Proc. of COLING-98*, Montreal.

Farmakiotou D., Karkaletsis V., Koutsias J., Sigletos G., Spyropoulos C.D. and Stamatopoulos P., 2000. Rule-based Named Entity Recognition for Greek Financial Texts. *Proc. of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pp. 75-78.

Kushmerick N., 1999. Regression testing for wrapper maintenance. *Proc. of National Conference on Artificial Intelligence*, pp. 74-79.

McDonald D., 1996. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In *B. Boguraev & J. Pustejovski (eds.) Corpus Processing for Lexical Acquisition*, MIT Press, pp 21–39.

Petasis G., Cucchiarelli A., Velardi P., Paliouras G., Karkaletsis V., Spyropoulos C.D., 2000. Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods. *Proc. of ACM SIGIR-2000, Ath*ens, Greece.

Quinlan J. R., 1993. C4.5: Programs for machine learning. Morgan-Kaufmann, San Mateo, CA.

Riloff E., 1993. Automatically Constructing a Dictionary for Information Extraction Tasks. *Proc. of the National Conference on Artificial Intelligence*, pp. 811-816.

Riloff E. and Jones R., 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proc. of the National Conference on Artificial Intelligence*, pp. 474-479.

Sekine, S., 1998. NYU: Description of the Japanese NE System used for MET-2. *Proc. of the Seventh Message Understanding Conference (MUC-7).*

Stricker M., Vichot F., Dreyfus G., Wolinski F., 2001. Training Context-sensitive Neural Networks with few Relevant Examples for TREC-9 Routing. In *Text Retrieval Conference, TREC-9*, NIST Special Publication, Gaithersburg, USA, to appear.

Thompson C., Califf M., Mooney R., 1999. Active Learning for Natural Language Parsing and Information Extraction. *Proc. of the International Conference on Machine Learning*, pp. 406-414.

Vichot F., Wolinski F., Ferri H. C., Urbani D., 1999. Using Information Extraction for Knowledge Entering, In *Advances in Intelligent Systems - Concepts, Tools and Applications*, S. G. Tzafestas (Ed.), Kluwer academic publishers, Dordrecht, The Netherlands, pp. 191-200.

Vilain M., and Day D., 1996. Finite-state phrase parsing by rule sequences. *Proc. of COLING-96*, vol. 1, pp. 274-279.

Wolinski F., Vichot F., Dillet B., 1995. Automatic Processing of Proper Names in Texts. In *European Chapter of the Association for Computer Linguistics, EACL*, Dublin, Ireland, pp.23-30.

Wolinski F., Vichot F., Stricker M., 2000. Using Learning-based Filters to Detect Rule-based Filtering Obsolescence. In *Recherche d' Information Assistée par Ordinateur, RIAO*, Paris, France, pp.1208-1220.