STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

NAMED ENTITY RECOGNITION FOR ROMANIAN

ADRIAN IFTENE(1), DIANA TRANDABĂŢ(1), MIHAI TOADER(2),
AND MARIUS CORÎCI(2)

Abstract. This paper presents a Named Entity Recognition system for Romanian, created using linguistic grammar-based techniques and a set of resources. Our system's architecture is based on two modules, the named entity identification and the named entity classification module. After the named entity candidates are marked for each input text, each candidate is classified into one of the considered categories, such as Person, Organization, Place, Country, etc. The system's Upper Bound and its performance in real context are evaluated for each of the two modules (identification and classification) and for each named entity type.

Named Entity Recognition (NER) is a common natural language processing task dedicated to the discovery of textual expressions such as the names of persons, organizations, locations, places, etc. Although a seemingly simple task, this task faces a number of challenges: entities may firstly be difficult to find, and once found, difficult to classify [3]. In this paper, we present the development of a NER system for Romanian. Even though the categories of named entities (NEs) are predefined, there are varying opinions on what categories should be regarded as NEs and how broad those categories should be. The categories chosen for a particular NER project may depend on the requirements of the project. The NER system for Romanian presented in this paper is intended to be part of a sentiment assessment system which monitors user feedback in rapport to an organization's brand or product. Therefore, we tried to refine the named entities types with regard to companies and products, so the categories we considered are: Person, Organization, Company, Region, Place, City, Country, Product, Brand, Model, and Publication.

NER systems use grammar-based techniques or statistical models (see for an overview [8]). Hand-crafted grammar-based systems typically obtain better

Received by the editors: April 7, 2011.

2010 *Mathematics Subject Classification.* 68T50, 68P20, 91F20.

1998 *CR Categories and Descriptors.* H.3.1. [**Information Systems**]: Content Analysis and Indexing – *Linguistic processing*; J.5. [**Computer Applications**]: Arts and Humanities – *Linguistics*.

Key words and phrases. named entity, information extraction.

precision, but at the cost of lower recall and months of work by experienced computational linguists. Statistical NER systems require a large amount of manually annotated training data. Machine learning techniques, such as the ones discussed in [6] or [7], allow systems-based adaptation two new domains, perform very well for coarse-grained classification, but require large training data. NER for Romanian has been attacked in [1], [4] and [5] (their advantages and drawbacks are discussed in the extended version of the paper). There is also a NER gazetteer for Romanian included in GATE [2]. The system presented in this paper obtains comparable results for most of the considered categories, and outperforms the existing approaches for Person recognition.

1. Our Solution

In the process of extracting named entities (NEs) we consider two steps: the first one is related to the identification of NEs and second one involves the classification of the identified NEs.

1.1. Named Entities Identification.

A rule-based approach was considered for the Named Entities Identification (NEI) task. The NEI module uses in a preprocessing step a text segmentator and a tokenizer. Given a text, we divide it into paragraphs, every paragraph is split into sentences, and every phrase is tokenized. Each token is annotated with two pieces of information: it's lemma and the normalized form (translated to the proper diacritics[1]). Every token written with a capital letter is then considered to be a NE candidate.

A special module was built for tokens with capital letters which are the first tokens in phrases, considering two situations:

(1) *when this first token of a phrase is in our stop word list* - we eliminate it from the named entities candidate list;
(2) *when the first token of a phrase is in our common word list* - in this case we have two possible situations: *a) when this common word is followed by lowercase words* - we check if it is a trigger word (cue words introducing NEs). If the first word of the sentence is in this list of trigger words, it is kept as NEs candidate. If the word is not in the trigger words list, it is eliminated from NEs candidates, as being just a common word written with capital letter due to its position. *b) when this common word is followed by uppercase words* - the first word of the sentence is kept in the NEs candidate list, and it will be subsequently decided if it will be combined with the following word in order to create a composed named entity.

[1] In Romanian online texts, two diacritics are commonly used, but only one is accepted by the official grammar.

After we build the list with named entity candidates, we apply rules that unify adjacent candidates in order to obtain composed named entities. The most important rules are:

(1) *Rules related to a person's title* - in these cases, we unify words like *Doctor*, *Profesor* (En: Doctor, Professor) next to adjacent candidates;
(2) *Rules related to the Organization type* - we unify words like *Universitate*, *Partid* (En: University, Party) next to adjacent candidates;
(3) *Rules related to abbreviation words* - we unify abbreviations such as S.R.L., S.C., S.A. next to adjacent candidates;
(4) *Rules related to special punctuation signs* - in these cases we unify candidates separated by "&" or "-";
(5) *Rules related to candidates to named entities separated by stop words* - in these cases we unify candidates separated by specific stop words;
(6) *Rules for a specific model/product* - candidates are combined with numbers or with one or two letters, followed by digits.

Some of these rules are used also in the classification process, namely the rules related to Person, Organization and Model types. Beside uppercase words which are automatically NE candidates, we also consider as possible NE-trigger lowercase words expressing titles (e.g. profesor, avocat, doctor, etc. (En: professor, lawyer, doctor)).

1.2. **Named Entities Classification.** The NE resource for Romanian was build starting from the categories used in GATE [2]. Thus, we consider the following major categories: the "standard categories" of City, Organization, Company, Country, Person, and additional categories such as Brand, Product and Publication (for revues, newspapers, etc.), with a total of 572,730 NEs. For almost all major categories we consider subcategories. In the end, we have built a total of 14 main categories with 98 subcategories. After all NEs in the input text are identified and, if possible, compound NEs have been created, we apply the following classification rules:

(1) *contextual rules* - using contextual information, we are able to classify candidate NEs in one of the categories Organization, Company, Person, City and Country by considering a mix between regular expressions and trigger words. For example *oraş*, *capitală* (En: city, capital) are the triggers searched in order to classify a candidate NE as a City;
(2) *resource-based rules* - if no triggers were found to indicate what type of entity we have, we start searching our databases for the candidate entity. If the candidate NE is a compound one, we first try to find it as if (i.e. the complex NE) in our resources. If it cannot be found as a complex entity, we split it back and try to find the first entity and assign its type to the whole complex.

2. Evaluation

This section presents the performance of our NER system. Sections 2.1 and 2.2 discuss a first "development" evaluation step, where we wanted to evaluate the system's performance when all needed resources were available (i.e. all NE are can be found in our resources). The next sections, 2.3 and 2.4, present the evaluation of our system on a new corpus, for each module.

2.1. **Upper Bound Named Entities Identification Evaluation.** In the evaluation process, we manually annotated 48 files with a total of 24,244 words and with 1,638 NEs. Based on these development files, we incrementally built our rules, both for NE identification (NEI) and for NE classification (NEC). Also, we added all missing NEs in our resources and built special rules for the untreated cases. Partial matching represents the intersection between the gold NE and the NE identified by our system. The results show a F-measure of 95.76%.

The first main problem in NEI is related to the agreement between annotators when different types of NEs are adjacent. In these situations, some believe it would be a single entity, while others believe that two different entities should be considered, with different types. The second main problem in NEI is related to the cases when the first word of a sentence is a common word and is not followed by words with capitalized letter. In these cases, the system is trained to leave the first word of the sentence out of the candidate NE list. A total number of 4,346 common words appear 5,622 times in one or more resources as NEs. In other words, 1% of NEs is ambiguous with common words in our databases.

2.2. **Upper Bound Named Entities Classification Evaluation.** For correctly identified named entities, the percentage of the matched and partial matched NEs that have been properly categorized is 95.71%. The main problems in NEs classification (NEC) are related to the fact that there exist NEs that are in more than one list of NEs. A number of 5,243 NEs appears in more than two resources, summing up to 10,588 occurrences. The most important problem is due to the fact that we have products that have the same name as the company that produce them. Another problem is due to the fact that we have the same names for cities and places. A problem for the NEC module is related to cases when we have partial match on extracted NEs. This happens when in the initial text we have two gold entities, each with its type. In this case, due to our NE composition rules, our application extracts only one named entity which is not found in any class, and thus the system assign to this NE group the class of the first NE.

2.3. **Named Entities Identification Evaluation in Real Context.** For testing the system in real context we created a new test corpus, unseen by our system, containing 38 files manually annotated with a total of 19,509 words and 1,215 NEs. The evaluation of the system with this test corpus shows a F-measure of 90.72%. Besides the problems discussed in the upper bound evaluation, we found additional problems related to the extraction of entities of the type Title (which are usually written with lowercase letters) and are very dependent to our resource list. The problems related to Title account for 3.70% of the total number of NEs error in this corpus (i.e. 45 from 144 titles weren't extracted) and comes from the fact that we don't have enough entities in our resources.

2.4. **Named Entities Classification Evaluation in Real Context.** Our system correctly classified (total or partial match) 66.73% of the NEs in our test corpus. Interesting is the case of Undecided entities, entities which are not classified in any of our types by human annotator in the test corpus. In 13 of these cases, our system was not able to classify the extracted entity, similar to the gold annotation. For Companies, Organization and Person types, the errors appear because the NEs were not found in our resources and no contextual rules could be applied. For Publication and Product types, the errors occurred because they frequently are marked interchangeable in the test corpus, since it is difficult to distinguish between them. For Region type, the major cause of errors is due to the fact that respective NE exists also in resources for other type, such as City, Place, and Country. An interesting example is the case of PNL, which does not exist in any of our resources. In some cases, when it is proceeded by the word "partid" (En: party), it is correctly classified as Organization, but in all other cases, the system does not identify any type for it. Thus is a clear example where anaphora resolution would greatly increase the system performance.

3. Conclusions

This paper presents a Named Entity Recognition system for Romanian, created using linguistic grammar-based techniques and a set of resources. The architecture of our system involves two modules, named entity identification and named entity classification module, successively applied. The goal of the described system is to recognize named entities for Romanian, distinguishing between 14 NE types. Even if we consider so many categories, we still manage to have comparable results (and even better for specific categories) with existing systems for Romanian, which identify less NE types.

Future work will be related to the elimination of problems related to common words that are at the beginning of sentences. To fix these problems, we intend to use statistical information about common words obtained from a large corpus, such as the Romanian Wikipedia. Another envisaged future direction is related to anaphora, which could be of great benefit in order to transfer the type of one classified entity to all its referees.

Acknowledgment

The research presented in this paper was funded by the Sectoral Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944. The authors of this paper thank the colleagues Alexandru Gînscă, Emanuela Boroş, Augusto Perez, Dan Cristea from Faculty of Computer Science Iasi.

References

1. S. Cucerzan and D. Yarowsky, *Language independent named entity recognition combining morphological and contextual evidence*, In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC, 1999, pp. 90–99.
2. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, *Gate: A framework and graphical development environment for robust nlp tools and applications*, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
3. A. Iftene, D. Trandabăţ, I. Pistol, M. Moruz, M. Husarciuc, and D. Cristea, *Uaic participation at qa@clef2008*, Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, Lecture notes in Computer Science, vol. 5706, 2009, pp. 448–451.
4. R. Ion, *Word sense disambiguation methods applied to english and romanian*, PhD Thesis, 2007.
5. L. M. Machison, *Named entity recognition for romanian (roner)*, Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2009, 2009, pp. 53–56.
6. Y. Mehdad, V. Scurtu, and E. Stepanov, *Italian named entity recognizer participation in ner task @ evalita 09*, Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 2009.
7. D. Nadeau, *Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision*, PhD Thesis, 2007.
8. D. Nadeau and S. Sekine, *A survey of named entity recognition and classification*, Linguisticae Investigationes **30** (2007), no. 1, 3–26, Publisher: John Benjamins Publishing Company.

[1] "Al. I. Cuza" University of Iasi, Faculty of Computer Science, Romania
E-mail address: adiftene@info.uaic.ro,dtrandabat@info.uaic.ro

[2] Intelligentics, Cluj-Napoca, Romania
E-mail address: marius@intelligentics.ro,mtoader@gmail.com