Combining Linguistic Knowledge with Machine Learning for Domain-Specific Named Entity Recognition

Łukasz Kobyliński

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
lkobylinski@ipipan.waw.pl

Abstract

This paper discusses an approach to domain-specific named entity recognition in the context of limited availability of training data in the form of linguistic resources. We start from a small set of hand-tagged magazine articles, with identified mentions of named entities: titles of such types of creative work as books, magazines and theatrical plays. The annotated data is combined with linguistic knowledge in the form of rules and data collected from the Internet. This corpus is then used to train and evaluate machine learning methods in the task of automated named entity recognition.

1. Introduction

While many approaches to classical Natural Language Processing tasks, such as Named Entity Recognition, have already been proposed, most of them rely on large amounts of training data to create representative models, or are very general, as in the case of unsupervised methods.

In this paper we tackle the problem of recognizing a specific type of named entities: titles of books, articles, plays and other types of creative work. This problem, while important for linguists, for example in translation: to separate titles kept in their original form from other, translatable material, is not directly solvable with any of the available methods and tools. On one hand, there is not enough training data available (hand-tagged corpora, with identified instances of work titles), especially for languages other than English, to train well-known machine learning approaches to NER, on the other hand no simple rule may be devised to automatically recognize this type of entities, as it may be easily confused with general names and other types of named entities.

For this reason we propose to approach the problem using a bootstrapped-like process, combining several knowledge sources. We start from a small corpus of hand-tagged magazine articles, containing annotated mentions of work titles. Independently, we extract known titles from publicly available resources, such as Freebase and Wikidata. We then extract trigger rules from the original corpus and a general language corpus, within contexts, which contain the titles collected in the form of gazetteers. The trigger rules are a simple, but effective way of recognizing structures, such as “w filmie Pan Tadeusz” (“in [the movie] Pan Tadeusz”). Here, the trigger rule would be “w filmie” (“in [the movie]”). The gazetteers and trigger rules are then used as features in the final machine learning process.

Several knowledge sources are finally combined by a machine learning method: we experiment with the most promising approaches by training supervised learning methods on the seed dataset. Features, such as morphological tags, hypernyms from a WordNet and gazetters are used as features calculated on the training data and fed into the classifier to build a model used for tagging new, previously unseen data.

2. Previous Work

Several approaches to dealing with inadequate or not existent training data have been proposed in the past. Many methods have focused on extracting large NE-lists (gazetteers) from the Web. Such methods can accurately produce gazetteers consisting of such types of entities, as person names or city names. For example (Etzioni et al., 2005) proposed the KnowItAll system, which extracts entities from the Web, based on a set of user-defined predicates. (Nadeau et al., 2006) built upon the gazetteer generation of the KnowItAll system and included a named entity disambiguation layer, which helped to solve problems of ambiguous tokens, such as “It” (a city and a pronoun).

(Kozareva, 2006) used bootstrapping to induce a named entity classifier, based on a small set of seed annotated samples and data collected from the Web. The unlabeled data was used to extract named entities automatically, by using simple syntactic rules. Two classifiers were trained on the seed data and the most confident extracted entities were added to the training set. This process was repeated for a selected number of iterations, gradually improving its accuracy.

The recent work in the context of Named Entity Recognition in Polish includes two supervised systems, both based on Conditional Random Fields machine learning method, supplemented with dictionaries and external knowledge sources, such as WordNet. NERF (Waszczuk et al., 2013) has been trained on the manually annotated part of the National Corpus of Polish (Przepiórkowski et al., 2011) and consequently is able to identify several classes of named entities used within the type hierarchy of NCP. Liner2 (Marcińczuk et al., 2013) is distributed with three different pre-trained models, which may be used to identify general category of names in the text, 9 most common classes, such as first name, last name, city, and a more fine-grained model, which is able to identify 82 classes of named entities.

3. Our Approach

We have started with a small seed data set, consisting of 580 684 tokens, out of which 3 426 named entities had been annotated manually, by a qualified linguist. Titles

Feature	Description
orth	orthographic word form
base	base word form (lemma)
ctag	morphosyntactic tag
class	grammatical class
case	grammatical case
number	grammatical number
gender	grammatical gender
agr1	grammatical agreement with token on position -1
synonym	synonym of word from plWordNet
hypernym-1...hypernym-n	hypernym of word from plWordNet
prefix-1...prefix-n	word prefix of length n
suffix-1...suffix-n	word suffix of length n
starts_with_upper/lower_case	orthographic structure
starts_with_symbol/digit	orthographic structure
has_upper/lower_case	orthographic structure
has_symbol/digit	orthographic structure
name	recognized as *name* by NER model
dict_title	exists in the titles gazetteer
dict_trigger	exists in the trigger dictionary
trigger-1...trigger-n	has the trigger rule fired for this token?

Table 1: Features used for machine learning.

of written and musical work, as well as theatrical plays had been annotated. As most of the identified entities are phrases spanning more than one token, the total number of annotated tokens is 10 033 in this data set. An example annotated sentence has been presented below.

(1) *Pomiędzy <t>Nocą listopadową</t>a*
Between <t>Noc listopadowa</t>and
<t>Weselem</t>
<t>Wesele</t>

(2) *Grzegorzewski wystawił również <t>Nowe*
Grzegorzewski staged also <t>Nowe
Bloomusalem</t>.
Bloomusalem</t>.

We have also created an initial gazetteer consisting of titles, by collecting the available data from Freebase and Wikidata. We have extracted available titles (in Polish) from Freebase from such categories as: *books*, *written work* and *creative work*. We have also queried Wikidata, to find all Polish *titles* or *original titles* available in this database, excluding titles of Wikinews articles. The gazetteers have been used in the feature generation phase, as binary features: sequence of word forms exists in the gazetteer, or not.

As we wanted to capture knowledge also from larger, more general corpora than the seed data, we have used the gazetteer entries as trigger words to mark contexts in National Corpus of Polish, in which they appeared. In such contexts we have mined trigger rules, using an associative classifier (Liu et al., 1998). The classifier was set to mine rules of the form: $orth_{-n} \ldots orth_{-2}\ orth_{-1} \Rightarrow class$, where $class$ is either beginning of a title, or a non-title word. For example, we have mined such rules as: *w filmie* (*in the movie*), *w powieści* (*in the novel*), *znanego jako* (*known as*).

Some of the triggers consisted of a single word, usually preceding a potential title. Such trigger words included: *autor* (*[the] author [of]*), *powieść* (*[the] novel*), *przedstawienie* (*[the] play*), etc.

Next, we have performed several pre-processing steps on the seed corpus, to enrich the annotation and calculate machine learning features. Pre-processing of the corpus included: morphological analysis, using the Maca framework (Radziszewski and Śniatowski, 2011) and Morfeusz analyzer (Woliński, 2006); morphosyntactic tagging, using the WCRFT2 tagger (Radziszewski, 2013) with a model trained on the National Corpus of Polish. General named entities have been identified using the Liner2 NER framework (Marcińczuk et al., 2013) and the *names* model.

Using the additional annotation layers described above, we have included in the set of generated features (see Table 1) some simple structural information, such as whether a word starts with an upper case letter, or a symbol, as well as grammatical features, as produced by the tagger. We have also used the Polish WordNet project (Piasecki et al., 2014) to identify synonyms and hypernyms of analyzed words. Simple language-domain knowledge in the form of a feature indicating grammatical agreement between neighboring tokens has also been incorporated in the feature set.

4. Experimental Results

We have divided the available seed data into development and test part. The machine learning methods have been tuned on the development part, using ten-fold cross-validation, while the final evaluation has been performed on the test part. The test part of the data consisted of 10% of the entire available data set, both in the case of total number of tokens and in the case of annotated named entities.

As for the evaluation of the proposed approach, we have experimented with several machine learning methods, including rule- and tree-based approaches, as well as Conditional Random Fields. Here, we present the best results, using the CRF method. The standard precision/recall statistics for the recognition of titles have been presented in Table 2.

class	precision (%)	recall (%)	F-measure (%)
title	64.68	39.90	49.35

Table 2: Experimental results.

5. Future Work

In this paper we have presented an approach to domain-specific named entity recognition, in the context of limited availability of training data that could be used for training supervised machine learning methods. We have combined ML methods with additional linguistic resources, such as gazetteers collected from the Web and trigger rules induced on the basis of the seed dataset.

There are many possibilities of improving the current result. As one of the improvements, we plan to include a shallow parser in the processing pipeline and extend the current set of machine learning features, by incorporating the information about syntactic words identified by the parser. This should improve the ability to detect the right boundary of named entities and, consequently, the overall performance of the system.

Secondly, we plan to employ a full bootstrapping approach to the learning process, by augmenting the original hand annotated corpus with contexts collected from general text corpora, the Web and publicly available data sources, such as Wikipedia. The contexts would be selected based on the existence of trigger phrases from the gazetteer and positive classification of the machine learning model in each iteration of the process.

6. References

Etzioni, Oren, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates, 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.*, 165(1):91–134.

Kozareva, Zornitsa, 2006. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics.

Liu, Bing, Wynne Hsu, and Yiming Ma, 1998. Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. New York, USA.

Marcińczuk, Michał, Jan Kocoń, and Maciej Janicki, 2013. Liner2 – a customizable framework for proper names recognition for Polish. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka (eds.), *Intelligent Tools for Building a Scientific Information Platform*. pages 231–253.

Nadeau, David, Peter D. Turney, and Stan Matwin, 2006. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Proceedings of the 19th International Conference on Advances in Artificial Intelligence: Canadian Society for Computational Studies of Intelligence*, AI'06. Berlin, Heidelberg: Springer-Verlag.

Piasecki, Maciej, Marek Maziarz, Stanisław Szpakowicz, and Ewa Rudnicka, 2014. PlWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proc. 7th International Global Wordnet Conference*.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik, 2011. National Corpus of Polish. In Zygmunt Vetulani (ed.), *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland.

Radziszewski, Adam, 2013. A tiered CRF tagger for Polish. In Robert Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka (eds.), *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag, page to appear.

Radziszewski, Adam and Tomasz Śniatowski, 2011. Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Waszczuk, Jakub, Katarzyna Głowińska, Agata Savary, Adam Przepiórkowski, and Michał Lenart, 2013. Annotation tools for syntax and named entities in the National Corpus of Polish. *International Journal of Data Mining, Modelling and Management*, 5(2):103–122.

Woliński, Marcin, 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski (eds.), *Intelligent Information Processing and Web Mining*, Advances in Soft Computing. Berlin: Springer-Verlag, pages 503–512.