Recognition of Named Entities Boundaries in Polish Texts

Michał Marcińczuk and **Jan Kocoń**
Institute of Informatics, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
Wrocław, Poland
{michal.marcinczuk,jan.kocon}@pwr.wroc.pl

Abstract

In the paper we discuss the problem of low recall for the named entity (NE) recognition task for Polish. We discuss to what extent the recall of NE recognition can be improved by reducing the space of NE categories. We also present several extensions to the binary model which give an improvement of the recall. The extensions include: new features, application of external knowledge and post-processing. For the partial evaluation the final model obtained 90.02% recall with 91.30% precision on the corpus of economic news.

1 Introduction

Named entity recognition (NER) aims at identifying text fragments which refer to some objects and assigning a category of that object from a predefined set (for example: *person*, *location*, *organization*, *artifact*, *other*). According to the ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (LDC, 2008) there are several types of named entities, including: proper names, definite descriptions and noun phrases. In this paper we focus on recognition of proper names (PNs) in Polish texts.

For Polish there are only a few accessible models for PN recognition. Marcińczuk and Janicki (2012) presented a hybrid model (a statistical model combined with some heuristics) which obtained 70.53% recall with 91.44% precision for a limited set of PN categories (first names, last names, names of countries, cities and roads) tested on the CEN corpus[1] (Marcińczuk et al., 2013). A model for an extended set of PN categories (56 categories) presented by Marcińczuk et al. (2013) obtained much lower recall of 54% with 93% precision tested on the same corpus. Savary and Waszczuk (2012) presented a statistical model which obtained 76% recall with 83% precision for names of people, places, organizations, time expressions and name derivations tested on the National Corpus of Polish[2] (Przepiórkowski et al., 2012).

There are also several other works on PN recognition for Polish where a rule-based approach was used. Piskorski et al. (2004) constructed a set of rules and tested them on 100 news from the *Rzeczpospolita* newspaper. The rules obtained 90.6% precision and 85.3% recall for person names and 87.9% precision and 56.6% recall for company names. Urbańska and Mykowiecka (2005) also constructed a set of rules for recognition of person and organization names. The rules were tested on 100 short texts from the Internet. The rules obtained 98% precision and 89% recall for person names and 85% precision and 73% recall for organization names. Another rule-based approach for an extended set of proper names was presented by Abramowicz et al. (2006). The rules were tested on 156 news from the *Rzeczpospolita* newspaper, the *Tygodnik Powszechny* newspaper and the news web portals. The rules obtained 91% precision and 93% recall for country names, 55% precision and 73% recall for city names, 87% precision and 70% recall for road names and 82% precision and 66% recall for person names.

The accessible models for PN recognition for Polish obtain relatively good performance in terms of precision. However, in some NLP tasks like recognition of semantic relations between PNs (Marcińczuk and Ptak, 2012), coreference resolution (Kopeć and Ogrodniczuk, 2012; Broda et al., 2012a), machine translation (Graliński et al., 2009a) or sensitive data anonymization (Graliński et al., 2009b) the recall is much more important than the fine-grained categorization of PNs.

[1] Home page: http://nlp.pwr.wroc.pl/cen.

[2] Home page: http://nkjp.pl

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 94–99,
Sofia, Bulgaria, 8-9 August 2013.

Unfortunately, the only model recognising wide range of PN categories obtains only 54% recall. Therefore, our goal is to evaluate to what extent the recall for this model can be improved.

2 Evaluation methodology

In the evaluation we used two corpora annotated with 56 categories of proper names: KPWr[3] (Broda et al., 2012b) and CEN (already mentioned in Section 1). The KPWr corpus consists of 747 documents containing near 200K tokens and 16.5K NEs. The CEN corpus consists of 797 documents containing 148K tokens and 13.6K NEs. Both corpora were tagged using the morphological tagger WCRFT (Radziszewski, 2013).

We used a 10-fold cross validation on the KPWr corpus to select the optimal model. The CEN corpus was used for a cross-corpus evaluation of the selected model. In this case the model was trained on the KPWr corpus and evaluated on the CEN corpus. We presented results for strict and partial matching evaluation (Chinchor, 1992). The experiments were conducted using an open-source framework for named entity recognition called Liner2[4] (Marcińczuk et al., 2013).

3 Reduction of NE categories

In this section we investigate to what extent the recall of NE recognition can be improved by reducing the number of NE categories. As a reference model we used the statistical model presented by Marcińczuk and Janicki (2012). The model uses the Conditional Random Fields method and utilize four types of features, i.e. orthographic (18 features), morphological (6 features), wordnet (4 features) and lexicon (10 features) — 38 features in total. The model uses only local features from a window of two preceding and two following tokens. The detailed description of the features is presented in Marcińczuk et al. (2013). We did not used any post-processing methods described by Marcińczuk and Janicki (2012) (unambiguous gazetteer chunker, heuristic chunker) because they were tuned for the specific set of NE categories.

We have evaluated two schemas with a limited number of the NE categories. In the first more common (Finkel et al., 2005) schema, all PNs are divided into four MUC categories, i.e. *person*, *organization*, *location* and *other*. In the other schema, assuming a separate phases for PN recognition and classification (Al-Rfou' and Skiena, 2012), we mapped all the PN categories to a single category, namely NAM.

For the MUC schema we have tested two approaches. In the first approach we trained a single classifier for all the NE categories and in the second approach we trained four classifiers — one for each category. This way we have evaluated three models: **Multi-MUC** — a cascade of four classifiers, one classifier for every NE category; **One-MUC** — a single classifier for all MUC categories; **One-NAM** — a single classifier for NAM category.

Model	P	R	F
Multi-MUC	76.09%	57.41%	65.44%
One-MUC	70.66%	65.39%	67.92%
One-NAM	80.46%	78.59%	79.52%

Table 1: Strict evaluation of the three NE models

For each model we performed the 10-fold cross-validation on the KPWr corpus and the results are presented in Table 1. As we expected the highest performance was obtained for the One-NAM model where the problem of PN classification was ignored. The model obtained recall of 78% with 80% precision. The results also show that the local features used in the model are insufficient to predict the PN category.

4 Improving the binary model

In this section we present and evaluate several extensions which were introduced to the One-NAM model in order to increase its recall. The extensions include: new features, application of external resources and post processing.

4.1 Extensions

4.1.1 Extended gazetteer features

The reference model (Marcińczuk and Janicki, 2012) uses only five gazetteers of PNs (*first names*, *last names*, names of *countries*, *cities* and *roads*). To include the other categories of PNs we used two existing resources: a gazetteer of proper names called NELexicon[5] containing ca. 1.37 million of forms and a gazetteer of PNs extracted from the National Corpus of Polish[6] containing 153,477

[3]Home page: `http://nlp.pwr.wroc.pl/kpwr`.
[4]`http://nlp.pwr.wroc.pl/liner2`
[5]`http://nlp.pwr.wroc.pl/nelexicon`.
[6]`http://clip.ipipan.waw.pl/Gazetteer`

forms. The categories of PNs were mapped into four MUC categories: *person*, *location*, *organization* and *other*. The numbers of PNs for each category are presented in Table 2.

Category	Symbol	Form count
person	per	455,376
location	loc	156,886
organization	org	832,339
other	oth	13,612
TOTAL		1,441,634

Table 2: The statistics of the gazetteers.

We added four features, one for every category. The features were defined as following:

$$gaz(n,c) = \begin{cases} B & \text{if } n\text{-th token starts a sequence of words found in gazetteer } c \\ I & \text{if } n\text{-th token is part of a sequence of words found in gazetteer } c \text{ excluding the first token} \\ 0 & \text{otherwise} \end{cases}$$

where $c \in \{per, loc, org, oth\}$ and n is the token index in a sentence. If two or more PNs from the same gazetteer overlap, then the first and longest PN is taken into account.

4.1.2 Trigger features

A trigger is a word which can indicate presence of a proper name. Triggers can be divided into two groups: *external* (appear before or after PNs) and *internal* (are part of PNs). We used a lexicon of triggers called PNET (Polish Named Entity Triggers)[7]. The lexicon contains 28,000 inflected forms divided into 8 semantic categories (*bloc*, *country*, *district*, *geogName*, *orgName*, *persName*, *region* and *settlement*) semi-automatically extracted from Polish Wikipedia[8]. We divided the lexicon into 16 sets — two for every semantic category (with *internal* and *external* triggers). We defined one feature for every lexicon what gives 16 features in total. The feature were defined as following:

$$trigger(n,s) = \begin{cases} 1 & \text{if } n\text{-th token base is found in set } s \\ 0 & \text{otherwise} \end{cases}$$

[7] `http://zil.ipipan.waw.pl/PNET`.
[8] `http://pl.wikipedia.org`

4.1.3 Agreement feature

An agreement of the morphological attributes between two consecutive words can be an indicator of phrase continuity. This observation was used by Radziszewski and Pawlaczek (2012) to recognize noun phrases. This information can be also helpful in PN boundaries recognition. The feature was defined as following:

$$agr(n) = \begin{cases} 1 & \text{if } number[n] = number[n-1] \\ & \text{and } case[n] = case[n-1] \\ & \text{and } gender[n] = gender[n-1] \\ 0 & \text{otherwise} \end{cases}$$

The `agr(n)` feature for a token n has value 1 when the n-th and $n-1$-th words have the same case, gender and number. In other cases the value is 0. If one of the attributes is not set, the value is also 0.

4.1.4 Unambiguous gazetteer look-up

There are many proper names which are well known and can be easily recognized using gazetteers. However, some of the proper names present in the gazetteers can be also common words. In order to avoid this problem we used an *unambiguous gazetteer look-up* (Marcińczuk and Janicki, 2012). We created one gazetteer containing all categories of PNs (see Section 4.1.1) and discarded all entries which were found in the SJP dictionary[9] in a lower case form.

4.1.5 Heuristics

We created several simple rules to recognize PNs on the basis of the orthographic features. The following phrases are recognized as proper names regardless the context:

- **a camel case word** — a single word containing one or more internal upper case letters and at least one lower case letter, for example *RoboRally* — a name of board game,
- **a sequence of words in the quotation marks** — the first word must be capitalised and shorter than 5 characters to avoid matching ironic or apologetic words and citations,
- **a sequence of all-uppercase words** — we discard words which are roman numbers and ignore all-uppercase sentences.

[9] `http://www.sjp.pl/slownik/ort`.

4.1.6 Names propagation

The reference model does not contain any document-based features. This can be a problem for documents where the proper names occur several times but only a few of its occurrences are recognised by the statistical model. The other may not be recognized because of the unseen or unambiguous contexts. In such cases the global information about the recognized occurrences could be used to recognize the other unrecognized names. However, a simple propagation of all recognized names might cause loss in the precision because of the common words which are also proper names. To handle this problem we defined a set of patterns and propagate only those proper names which match one of the following pattern: (1) a sequence of two or more capitalised words; (2) all-uppercase word ended with a number; or (3) all-uppercase word ended with hyphen and inflectional suffix.

4.2 Evaluation

Table 3 contains results of the 10-fold cross validation on the KPWr corpus for the One-NAM model, One-NAM with every single extension and a complete model with all extensions. The bold values indicate an improvement comparing to the base One-NAM model. To check the statistical significance of precision, recall and F-measure difference we used Student's t-test with a significance level $\alpha = 0.01$ (Dietterich, 1998). The asterisk indicates the statistically significant improvement.

Model	P	R	F
One-NAM	80.46%	78.59%	79.52%
Gazetteers	**80.60%**	**78.71%**	**79.64%**
Triggers	**80.60%**	78.58%	**79.58%**
Agreement	**80.73%**	**78.90%**	**79.80%**
Look-up	80.18%	**79.56%***	**79.87%**
Heuristics	79.98%	**79.20%***	**79.59%**
Propagate	80.46%	78.59%	79.52%
Complete	80.33%	**80.61%***	**80.47%***

Table 3: The 10-fold cross validation on the KPWr corpus for *One-NAM* model with different extensions.

Five out of six extensions improved the performance. Only for the name propagation we did not observe any improvement because the KPWr corpus contains only short documents (up to 300 words) and it is uncommon that a name will appear more than one time in the same fragment. However, tests on random documents from the Internet showed the usefulness of this extension.

For the unambiguous gazetteer look-up and the heuristics we obtained a statistically significant improvement of the recall. In the final model we included all the presented extensions. The final model achieved a statistically significant improvement of the recall and the F-measure.

To check the generality of the extensions, we performed the cross-domain evaluation on the CEN corpus (see Section 2). The results for the 56nam, the One-NAM and the Improved One-NAM models are presented in Table 4. For the strict evaluation, the recall was improved by almost 4 percentage points with a small precision improvement by almost 2 percentage points.

Evaluation	P	R	F
56nam model (Marcińczuk et al., 2013)			
Strict	93%	54%	68%
One-NAM model			
Strict	85.98%	81.31%	83.58%
Partial	91.12%	86.65%	88.83%
Improved One-NAM model			
Strict	86.61%	85.05%	85.82%
Partial	91.30%	90.02%	90.65%

Table 4: The cross-domain evaluation of the basic and improved One-NAM models on CEN.

5 Conclusions

In the paper we discussed the problem of low recall of models for recognition of a wide range of PNs for Polish. We tested to what extent the reduction of the PN categories can improve the recall. As we expected the model without PN classification obtained the best results in terms of precision and recall.

Then we presented a set of extensions to the One-NAM model, including new features (morphological agreement, triggers, gazetteers), application of external knowledge (a set of heuristics and a gazetteer-based recogniser) and post-processing (proper names propagation). The final model obtained 90.02% recall with 91.30% precision on the CEN corpus for the partial evaluation what is a good start of further NE categorization phase.

Acknowledgments

This work was financed by Innovative Economy Programme project POIG.01.01.02-14-013/09.

References

Witold Abramowicz, Agata Filipowska, Jakub Piskorski, Krzysztof Węcel, and Karol Wieloch. 2006. Linguistic Suite for Polish Cadastral System. In *Proceedings of the LREC'06*, pages 53–58, Genoa, Italy.

Rami Al-Rfou' and Steven Skiena. 2012. SpeedRead: A fast named entity recognition pipeline. In *Proceedings of COLING 2012*, pages 51–66, Mumbai, India, December. The COLING 2012 Organizing Committee.

Bartosz Broda, Lukasz Burdka, and Marek Maziarz. 2012a. Ikar: An improved kit for anaphora resolution for polish. In Martin Kay and Christian Boitet, editors, *COLING (Demos)*, pages 25–32. Indian Institute of Technology Bombay.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012b. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA.

Nancy Chinchor. 1992. MUC-4 Evaluation Metrics. In *Proceedings of the Fourth Message Understanding Conference*, pages 22–29.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In The Association for Computer Linguistics, editor, *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.

Filip Graliński, Krzysztof Jassem, and Michał Marcińczuk. 2009a. An Environment for Named Entity Recognition and Translation. In L Màrquez and H Somers, editors, *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 88–95, Barcelona, Spain.

Filip Graliński, Krzysztof Jassem, Michał Marcińczuk, and Paweł Wawrzyniak. 2009b. Named Entity Recognition in Machine Anonymization. In M A Kłopotek, A Przepiórkowski, A T Wierzchoń, and K Trojanowski, editors, *Recent Advances in Intelligent Information Systems.*, pages 247–260. Academic Pub. House Exit.

Mateusz Kopeć and Maciej Ogrodniczuk. 2012. Creating a coreference resolution system for polish. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

LDC. 2008. ACE (Automatic Content Extraction) English Annotation Guidelines for Relations (Version 6.2).

Michał Marcińczuk and Maciej Janicki. 2012. Optimizing CRF-Based Model for Proper Name Recognition in Polish Texts. In Alexander F. Gelbukh, editor, *CICLing (1)*, volume 7181 of *Lecture Notes in Computer Science*, pages 258–269. Springer.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2 - A Customizable Framework for Proper Names Recognition for Polish. In Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, Marzena Kryszkiewicz, and Marek Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 231–253. Springer.

Michał Marcińczuk and Marcin Ptak. 2012. Preliminary study on automatic induction of rules for recognition of semantic relations between proper names in polish texts. In Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue — 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, volume 7499 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer-Verlag, September.

Jakub Piskorski, Peter Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliński. 2004. Information Extraction for Polish Using the SProUT Platform. In Mieczyslaw A. Kłopotek, Slawomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference*, Advances in Soft Computing, Zakopane. Springer-Verlag.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw.

Adam Radziszewski and Adam Pawlaczek. 2012. Large-Scale Experiments with NP Chunking of Polish. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *TSD*, volume 7499 of *Lecture Notes in Computer Science*, pages 143–149. Springer Berlin Heidelberg.

Adam Radziszewski. 2013. A Tiered CRF Tagger for Polish. In Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, Marzena Kryszkiewicz, and Marek Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 215–230. Springer Berlin Heidelberg.

Agata Savary and Jakub Waszczuk. 2012. Narzędzia do anotacji jednostek nazewniczych. In Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN. Creative Commons Uznanie Autorstwa 3.0 Polska.

Dominika Urbańska and Agnieszka Mykowiecka. 2005. Multi-words Named Entity Recognition in Polish texts. In Radovan Grabík, editor, *SLOVKO 2005 – Third International Seminar Computer Treatment of Slavic and East European Languages, Bratislava, Slovakia*, pages 208–215. VEDA.