Control and Cybernetics

vol. **40** (2011) No. 2

Statistical proper name recognition in Polish economic texts*

by

Michał Marcińczuk and Maciej Piasecki

Wrocław University of Technology, Wrocław, Poland

Abstract: In the paper we present a Proper Name Recognition algorithm based on the Hidden Markov Model (HMM). Recognition of the Proper Names (PN) is treated as the basis for Named Entity Recognition problem in general. The proposed method is based on combining domain-dependent method based on HMM with domain independent methods based on gazetteers and hand-written rules for recognition and post-processing that capture the general properties of Polish PN structure. A large gazetteer with entries described morphologically was acquired from the web. The HMM re-scoring mechanism was applied as a basis for integration of different knowledge sources in PN recognition. Results of experiments on a domain corpus of Polish stock exchange reports, used for training and testing, are presented. A cross-domain evaluation on two other corpora is also presented. Adaptability of the method was analysed by applying the trained model to two other domain corpora.

Keywords: proper name recognition, named entity recognition, machine learning, hidden Markov model, rule-base approach, dictionary-base approach.

1. Introduction

Information Extraction is built around recognition of situation descriptions in the text. A typical description is anchored in the context of utterance interpretation with the help of so called *mentions of Named Entities* (LDC, 2008). From the linguistic point of view, a mention of a Named Entity is a language expression referring to an entity (or entities) of a selected type that is represented in the context of interpretation. The class of Named Entities is always delimited by the list of semantic types of their referents. Mentions encompass linguistically heterogeneous language expressions, but the core are Proper Names (PN), i.e. a basic, prototypical language means of referring to entities included in the interpretation context of the utterance. Other important classes

*Submitted: October 2010; Accepted: June 2011.

are pronouns and definite noun phrases in anaphoric and deictic functions. PNs deliver unambiguous anchoring to the interpretation context of an utterance and are a starting point for building the rest of factual information representation. Moreover, as usually the pragmatic information and analysis is limited in Information Extraction (IE), the proper recognition and classification of PN becomes especially important.

However, the problem of PN recognition is not trivial, as it will be shown on examples of simple methods applied to PN identification and classification. Both a dictionary-based recognition of PNs – so called *gazetteer based*[1] – and simple heuristics built around the use of big letters in PNs fail to achieve practically useful accuracy of recognition, see Section 5. One of the reasons for the limited accuracy of the gazetteer-based methods of PN recognition is the fact that the set of possible PNs is always unlimited in practical applications. In the case of languages with rich inflection, like Polish, not only basic forms of PNs but all their morphological forms must be properly recognised, including out-of-dictionary PNs. On the other hand, heuristics based on words starting with big letters suffer from text segmentation errors (e.g. into sentences) and domain specific practices of using big letters.

The appropriate identification of Named Entities, including PNs, and their classification with respect to types of their referents is very important for the following steps of IE, as well as for applications of the natural language technology. The combined identification and classification of Named Entities will be called Named Entity Recognition (NER), and such a process limited to PNs will be called Proper Name Recognition (PNR).

NER is a well studied task in the case of English (Marrero et al., 2009), but only few approaches have been proposed so far for Polish. Works for other Slavic languages are not numerous either: Czech (Kravalovà and Žabokrtský, 2009), Bulgarian (Osenova and Kolkovska, 2002) and Ukrainian (Katrenko and Adriaans, 2007). Dominant approaches to NER for Polish are based on manual construction of rules based on heuristics or grammars, see applications in Piskorski (2004a,b), Urbańska and Mykowiecka (2005), Mykowiecka et al. (2007), Abramowicz et al. (2006), Savary and Piskorski (2010), machine anonymization (Graliński et al., 2009b) and machine translation (Graliński et al., 2009a). Only few preliminary, alternative works were presented on the application of Machine Learning methods to NER, e.g. Memory Based Learning in Marcińczuk and Piasecki (2007), Decision Trees C4.5 and Naïve Bayes in Marcińczuk (2007).

Our general long term objective is to build a robust NER method of broad applicability. However, due to domain specific NE characteristics like different core dictionary or orthography rules, we want to achieve this objective in a two-level architecture. A domain-depended NER method, relatively easy to adapt to another domain, will be developed, and then a domain independent but closely

[1] *A gazetteer* (in Information Extraction) is a kind of dictionary of word form sequences known to be PNs. Gazetteers often deliver also information about a class to which a given PN belongs.

integrated NER method of limited accuracy will be constructed. The best result should be obtained by combining both levels.

In this paper we focus on a domain dependent but adaptable method. The goal of the work presented here is to develop a PNR method for a specific but important class of Polish texts and PNs occurring in them, namely for Polish texts related to economy, e.g. news from the economy domain or stock exchange reports – those reports must be produced regularly by all companies registered on the stock market and include important informations for investors. The selected domain is quite limited, but of practical importance and counts many thousands of documents per year.

The developed method should be relatively easy to adopt to another domain and it should provide balance between a workload required for adaptation and the achieved accuracy. As NER methods which are based on Machine Learning have potentially better ability to be more domain-independent and scalable, we follow the general setting of these approaches.

A solution presented here is limited to the core subset of NEs, namely PNs, as good accuracy of PNR seems to be a necessary condition for the good accuracy of the complete NER.

The research questions that need to be answered include:

- a type of the general framework of the adaptable method,
- a method of PN dictionary acquisition and learning of morphological forms for PNs,
- application of Machine or Statistical Learning keeping the workload required on training data preparation in acceptable limits,
- method of combining the detailed knowledge acquired automatically (e.g. with the help of statistical learning) and the general knowledge expressed in the form of hand-written rules and dictionaries.

In the paper we concentrate on recognition of selected types of PNs (discussed in the next section) and a selected model of machine learning, namely *Hidden Markov Model* (henceforth *HMM*).

We chose HMM due to its known successful application to NER for several languages, e.g. English (Bikel et al., 1997; Zhou and Su, 2002), Chinese (Carpenter, 2006), Dutch and Spanish (Malouf, 2002), and to the character of the task performed.

According to our best knowledge there were no earlier attempts to apply HMM in NER for any Slavonic languages, especially for Polish.

In the rest of the paper, we will first define the PNR problem considered. Then we will describe used corpora: the domain-dependent one (a training and testing corpus), and additional corpora used for cross-domain evaluation. Next, we will introduce two baseline measures. Finally, the proposed algorithm and its variations will be explained. Evaluation of the algorithm against baseline methods and other corpora will be discussed as well.

2. Task definition

The task of PN recognition is defined as identification and categorization of continuous expressions that are formal or informal names for entities belonging to a set of pre-defined categories.

Because an approach which is based on machine learning requires prior construction of an annotated training and testing corpus, mostly of a substantial size, we decided to limit the task of PN recognition to the categories of: people *first names*, *surnames*, names of *countries*, *cities* and *roads*. Other categories will be included in the future work.

In case of person names we decided to recognize their sub-parts separately (i.e. *first names* and *surnames*) instead of whole names due to several reasons. Firstly, person names can be quite complex and comprise several elements, i.e. a first name, a second name, a surname, a maiden name, initials and a surname prefix (Marcińczuk and Piasecki, 2010b). Secondly, existing dictionaries of first names and surnames can be applied more effectively. Finally, in some cases, person names are not continuous expressions, for example "John and Mark Twain" are in fact two names of people: "John Twain" (discontinuous name) and "Mark Twain" (continuous name).

PNs express more constrained lexico-syntactic forms than referential noun phrases in general, i.e. constituents of the majority of PNs start with big letters[2] or express a specific pattern of small and big letters, e.g. *iPod*. The use of big letters simplifies PN recognition in comparison to the general NER. However, we can notice a significant number of problematic cases in which the big letters are not unambiguous PN markers, e.g., sequence of consecutive PNs or a PN at the sentence beginning that is especially error prone in combination with the usually imperfect text segmentation into sentences, see the discussion in Section 3.1. Recognition of the PN limits is much easier in the case of first names and surnames, as they are mostly one-word units, however, surname word forms are often ambiguous between a PN and a common noun.

We are aware that in general case PNs can form nested structures (see Fig. 1). However, we do not cover these phenomena here. This means that in the testing sets all embedded annotations were ignored, i.e. only the outer annotations were considered. These assumptions allowed us to represent every sentence as a sequence of labels in the IOB style (see Fig. 2).

3. Corpora

Concerning variety of texts containing PNs we used three available domain corpora as the basis for the preparation of the training and test data, namely: a corpus of stock exchange reports (henceforth CSER), a corpus of economic news (CEN) – representing the economy domain, and a corpus of police reports (CPR) – the public security domain. In comparison to our previous work, see

[2]With a notable difference of conjunctions in the institution names.

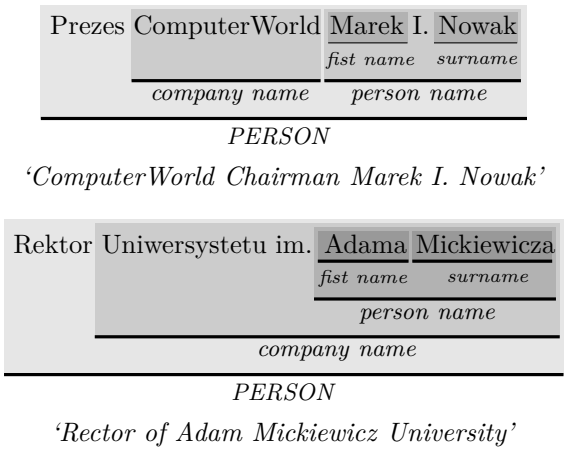

Figure 1. Examples of nested structures on PNs.

Marcińczuk and Piasecki (2010a,b), new types of annotations were added to the corpora, namely: names of countries, cities, and roads.

The following subsections contain a description of the corpus pre-processing and a brief description of all used corpora. A summary and direct comparison of corpora metrics is presented in Table 1. Detailed numbers for annotations in relation to the categories are presented in Table 3. The number of unique (distinct forms of names) and hapax legomena (annotated only once) words that were annotated is also presented.

3.1. Pre-processing

The corpora were pre-processed with the morphosyntactic tagger TaKIPI (Piasecki, 2007) extended with the *Guesser* module (Piasecki and Radziszewski, 2007). TaKIPI divides the text into sentences and tokens, assigning a morphosyntactic analysis to every token. However, due to the tagger segmentation errors, we annotated plain text firstly. After the annotation was completed, the annotated text was transformed back into plain text by stripping off NE annotations (marking begins and ends) and next tagged. In this way, we assured that the annotations span over whole tokens. Finally, the tagged text was automatically aligned with the annotations and converted into the IOB format, in which every token was assigned one of the labels: **B-TYPE** – a given token begins an annotation of type *TYPE*, **I-TYPE** – a token is inside an annotation of type *TYPE*, **O** – a token is outside an annotation (see Fig. 2).

Table 1. Corpora statistics

	Stock exchange reports	**Police reports**	**Economic news**
Documents	1 215	22	797
Sentences	10 097	1 527	7 305
Tokens	282 401	24 772	144 004
Annotations	4 011	1 004	4 997

Table 2. Statistics of proper name annotations

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**SUM**
Corpus of stock exchange reports (CSER)						
Annotations	686	689	1827	414	395	**4011**
Unique	227	378	239	62	98	**1004**
Hapax	99	247	114	33	57	**550**
% of hapax	7.97%	35.85%	6.24%	7.97%	14.43%	**13.71%**
Corpus of police reports (CPR)						
Annotations	333	411	191	27	42	**1 004**
Unique	167	174	103	13	31	**448**
Hapax	98	113	70	7	21	**309**
% of hapax	29.53%	27.49%	36.65%	25.93%	50.00%	**30.98%**
Corpus of economic news (CEN)						
Annotations	1097	1517	657	1695	31	**4997**
Unique	525	852	349	369	28	**2123**
Hapax	353	621	246	183	25	**1428**
% of hapax	32.18%	40.94%	37.44%	10.80%	80.65%	**28.58%**

The corpora were annotated only by one person with the use of Inforex[3], so that some annotations were missing (which turned out in later experiments, see Section 7.2). The added annotations were partially verified by another person.

TaKIPI provides robust text segmentation but leaves some inevitable errors which can negatively influence NER accuracy, e.g.:

- sentence level segmentation:
  - false sentence breaks, e.g. *ul. A. Krakowskiego 8 m 16* ('*A. Krakowskiego Street, 8/16*') is divided during segmentation into two sentences: *ul. A.* and *Krakowskiego 8 m 16*, while *A. Krakowskiego* is a full road name (*Krakowskiego* is a surname), or *Pan M. Marcisz podpisał, umowę.* ('*Mr. M. Marcisz signed an agreement*') is divided into *Pan M.* and *Marcisz podpisał umowę.* — in both cases the dot in the initial was identified as a sentence delimiter.
  - erroneous separation of company names into parts caused by dots occurring in the names, e.g.,
    - *Zarząd Z.O. "Bytom" S.A. podaje* is divided into <*Zarząd Z.O.*> and <*"Bytom" S.A. podaje*>;
    - *Z.Ch. "Permedia" S.A. w Lublinie* → <*Z.Ch.*> <*"Permedia" S.A. w Lublinie* >,
    - *Spółdzielczy Bank Ludowy im. Ks. P. Wawrzyniaka* → <*Spółdzielczy Bank Ludowy im.*> <*Ks.*> <*P. Wawrzyniaka*>
- token level segmentation
  - missing segmentation of words linked by a hyphen, such sequences occur often in multi-segment company names, e.g., *FEH "Ponar-Wadowice" S.A.* is divided into: <*FEH*> <"> <*Ponar-Wadowice*> <"> <*S.A.*>, while *Wadowice* (a city name) which is a PN will not be separated from the text and next recognised, or *Bank Austria-Creditanstalt Poland* is divided into [Bank] [Austria-Creditanstalt] [Poland], but *Austria* (a country name) is not recognised, as it is still non-separated,
  - abbreviated company name including extra signs, e.g. *Pekao S.A. I O/Tarnów* in which [O/Tarnów] is one token after segmentation and *Tarnów* is not separated.

A problem that is similar to the problems on the token-level segmentation is the proper recognition of multi-segment surnames, e.g. *Machnik-Ochała*. The context sensitive morphological analysis is provided by Morfeusz – a morphological analyser (Woliński, 2006), and we assumed that segmentation of such sequences is also performed by Morfeusz and it is an integral part of the overall

[3]Inforex is a web browser based system for corpora management. The system is accessible at `http://nlp.pwr.wroc.pl/inforex`

1. Sample text:

Pani Iwona Nowak-Majewska mieszka w Zielonej Górze.

Ms. Iwona Nowak-Majewska lives in Zielonej Górze.

2. Annotated text:

Pani [Iwona] [Nowak]-[Majewska] mieszka w [Zielonej Górze].

Annotations:

Iwona	— first name
Nowak	— surname
Majewska	— surname
Zielonej Górze	— city name

3. Annotated text transformed to plain text:

Pani Iwona Nowak - Majewska mieszka w Zielonej Górze.

4. Segmentation:

[Pani] [Iwona] [Nowak] [-] [Majewska] [mieszka] [w] [Zielonej Górze] [.]

In comparison, the original text would be segmented as follows:

[Pani] [Iwona] [Nowak-Majewska] [mieszka] [w] [Zielonej Górze] [.]

5. Alignment and transformation to IOB format:

Pani	O
Iwona	B-PERSON_FIRST_NAM
Nowak	B-PERSON_LAST_NAM
-	O
Majewska	B-PERSON_LAST_NAM
mieszka	O
w	O
Zielonej	B-CITY_NAM
Górze	I-CITY_NAM

Figure 2. Pre-processing of a sample sentence.

segmentation process provided by TaKIPI (description of TaKIPI segmentation rules can be found in Broda et al., 2008).

Not only segmentation errors but also editor errors are a source of problems for PNR:

- missing space after a dot completing sentence causes merging two tokens into one and makes a PN a part of a larger token and unrecognised on the following stages of processing, e.g. in the sentence

  > *Beneficjentem gwarancji jest Zakład Wodociągów i Kanalizacji Sp. z o.o. z siedzibą w* ***Szczecinie.Prowizja*** *przysługująca (...)*
  >
  > '*The beneficiary of the guarantee is Water and Sewerage Service LPS with office in* ***Szczecin.Commission*** *it being entitled (...)*'

  tokeniser treats *Szczecinie.Prowizja* as one token, while in fact it consists of three elements including the city name *Szczecinie*.
- lack of space after an abbreviation, e.g., *st.Warszawy* results in treating the sequence as one token and thus 'hiding' a PN, here $Warszawy_{case=gen}$ inside it.

Because we assumed that our PNR program operates on the basis of segmentation produced by the tagger, we introduced a simple correction of errors, for the needs of experiments. On the basis of annotations introduced manually to the training corpus we have added always one space before and after each annotated token, thus separating them, even if they were not initially separated by the tagger.

3.2. Stock exchange reports

CSER consists of 1215 documents[4] collected from GPWInfoStrefa[5] – an official web page on which reports produced by all joint stock companies are collected. The corpus comprises all reports published by 185 different companies in the year 2004 (selected for harvesting).

The reports are written in the formal register. Expressions referring to persons mostly consist of a first name and a surname and are often preceded by a honorific *Pan*/*Pani* 'Mr./Ms.'. Expressions starting with an upper case character are quite numerous, i.e. (1) aliases (*Company*, *Agreement*, etc.), (2) offices (*Chairman*, *Executive*, etc.), (3) bodies in company structure (*Board of Directors*, *Supervisory Board*, *Board*, etc.). That makes PNR more difficult.

3.3. Police reports

CPR consists of statements produced by 12 witnesses and suspects which were provided by a local Police Department. Because of legal reasons the documents

[4]The corpus is available at `http://nlp.pwr.wroc.pl/inforex/?corpus=1&page=browse`
[5]Web page: `http://gpwinfostrefa.pl`

were manually anonymized beforehand. CPR is a sub-corpus of the corpus collected within the project on machine anonymization (Graliński et al., 2009b). As the CPR size is small in comparison to the whole corpus, a direct comparison with the results in Graliński et al. (2009b) was not possible. The documents are written in informal register and contain many one-word person names only, mostly pseudonyms and first names.

3.4. Economic news

CEN[6] consists of 797 news from the economic domain collected from Polish Wikinews[7]. The corpus comprises economic news published between 25^{th} February, 2005 and 10^{th} June, 2010 – the day of document harvesting. The documents are written in formal register. People are referred mostly by first name and surname. Most words starting with an upper-case letter are proper names. CEN includes also many other proper names, like names of products and events.

4. Evaluation

In the evaluation, a mention is *true positive* when both its type (first name, surname, country, city or road) and boundary are correct. When the type or the boundary is incorrect (must be exactly matched) the mention is *false positive*. A mention from the reference set that is not fully matched is *false negative*.

5. Base line

There is no commonly used reference corpus for NER, approaches known from literature were applied to various small corpora and differ in goals defined. Thus, it is difficult to refer to the known results as a baseline. Graliński et al. (2009b) obtained a precision of 93.53%, 88.66%, and 94.59% on a rule-based anonymizaton of the first names, surnames, and company names, respectively. 59 Interpol messages were used as a corpus. The recall was not measured because the authors could not evaluate the software on original texts, as they contained sensitive data. Another rule-based approach, presented in Urbańska and Mykowiecka (2005), achieved 98% of precision and 89% of recall for persons and 85% of precision and 73% of recall for organizations in tests on "*about one hundred short citations downloaded from Internet or made by testers*". Piskorski (2004b) obtained 90.6% of precision and 85.3% of recall in person recognition and 87.9% of precision and 56.6% of recall in company recognition using manually created grammars and gazetteers. The tests were performed on 100 financial news articles from the online version of *Rzeczpospolita*. Abramowicz et al. (2006) experimented with recognition of NE subtypes and obtained 55% of precision

[6]The corpus is available at `http://nlp.pwr.wroc.pl/inforex/?corpus=5&page=browse`
[7]Web page: `http://pl.wikinews.org`

and 73% of recall for cities, 91% of precision and 93% of recall for countries, 87% of precision and 70% of recall for roads, and 82% of precision and 66% of recall for persons. Their test corpus consisted of daily newspapers, an online financial magazine and different local news portals.

A direct comparison (evaluation on the same data sets) was not possible because of differences in problem statement and limited access to the data sets used in the mentioned works. Instead, two base line experiments were performed: the first one with a simple heuristics and the second one with *gazetteers* (constructed manually and collected automatically).

5.1. Simple heuristics

A simple heuristic applied during the first experiment was based on the assumption that a PN is a sequence of words in the upper case and the whole expression does not extend beyond the sentence limits. Intuitively, the heuristics should work for the majority of cases. The heuristics were implemented as a set of rules that were created on the basis of CSER and later tested on CPR and CEN. The rules recognize proper names that appear in a defined context. We tried to describe only the unambiguous context that clearly indicates the given name category. For instancem the expression *z siedzibą w X przy ul. Y* ('with residence in X, Y Street') should identify X as a *city* name and Y as a *road* name.

The rules took form of regular expressions. The complete list of rules consists of some 20 positions. Below, we present sample rules for every PN type:

- `<UCF> = \p{Lu}\p{Ll}+`
  — matches words starting with an upper case letter,
- `<UCF_SEQ> = <UCF>( <UCF>)*`
  — matches a sequence of words starting with an upper case letter,
- `<ROAD_IND> = ul \.?|al \.?|ulic[ayęą]|alei|alej[ąę]`
  — matches a road indicator, like *ul.*, *ulica*, *ulicy*, etc.,
- `<ROAD_IND> (<UCF_SEQ>)`
  — recognises *road names* after road indicator,
- `([Pp]an(i|a|u|ią)?|powiedział(a)?) (<UCF>) (<UCF>)`
  — recognises a sequence of *first name* and *surname*,
- `([Pp]an(i|a|u|ią)?|powiedział(a)?) (<UCF>) (<UCF>) (<UCF>)`
  — recognises a sequence of *first name*, *first name* and *surname*,
- `[0-9]{2}-[0-9]{3} (<UCF_SEQ>)`
  — recognises city name after postal code,
- `w (<UCF_SEQ>) , (<UCF_SEQ>) , <ROAD_IND> (<UCF_SEQ>)`
  — recognises a sequence of *country name*, *city name* and *road name*.

Table 3 presents the results obtained with the defined set of heuristics on CSER, CPR, and CEN. As expected, the highest F_1-measure of 70.00% was

Table 3. Performance of proper name recognition with simple heuristics

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**TOTAL**
			CSER			
Precision	95.51%	96.48%	96.60%	98.20%	96.61%	96.50%
Recall	58.89%	55.73%	48.22%	39.61%	93.67%	54.92%
F_1	72.86%	70.65%	64.33%	56.45%	95.12%	70.00%
			CPR			
Precision	0.00%	0.00%	100.00%	0.00%	97.37%	97.92%
Recall	0.00%	0.00%	5.24%	0.00%	88.10%	4.68%
F_1	0.00%	0.00%	9.95%	0.00%	92.50%	8.94%
			CEN			
Precision	96.36%	94.55%	37.50%	0.00%	100.00%	92.31%
Recall	4.83%	3.43%	0.46%	0.00%	38.71%	2.40%
F_1	9.20%	6.62%	0.90%	0.00%	55.81%	4.68%

obtained on the corpus that was used to define the heuristics. The F_1-measure for the other two corpora was very low, namely 8.94% for CPR and 4.68% for CEN. The main reason of the low results is low recall, which is caused by two factors. The first one is that PN patterns that occur in CSER do not match all possible structures of PN expressions. The other one is that in many cases it is impossible to distinguish between different PN types without additional knowledge (for example, in "He lives in X" *X* can be a *city name* or a *country name*). Plain heuristics are insufficient, even within the same domain. However, the level of precision is very high, namely 97.92% for CPR and 92.31% for CEN, making the heuristic useful as a supporting method. Basing on this observation we utilized the heuristics as an additional source of knowledge during the HMM post-processing (see Section 6.4).

5.2. Gazetteers

For the second baseline we used a typical *gazetteer-based approach*, in which only sequences of words matching *gazetteer* entries are recognised as annotations. Two gazetteers were applied, for the related statistics see Table 4. The first one was the gazetteer constructed manually by Piskorski (Piskorski, 2004b), henceforth PG, in which all entries were manually verified. The second gazetteer, henceforth IG (Internet-based gazetteer), was collected automatically from resources freely available on the web. It contains PNs of five types, see Table 4. PNs have been acquired from web pages, in which PNs were presented in some

Table 4. Statistics of gazetteers

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**SUM**
PG*						
Lemmas	118	16 997	29 370	201	0	**47 015**
Forms	1 166	44 608	29 699	1 761	0	**77 214**
Unambiguous	801	44 341	29 649	1 741	0	**76 532**
IG						
Lemmas	5 288	400 215	58 083	240	29 486	**493 312**
Forms	7 776	456 068	68 243	578	35 211	**567 876**
Unambiguous	3 904	432 250	50 726	393	20 530	**507 803**

* only selected categories are listed

structured way, e.g. as lists, tables or links to the predictable content. PNs were extracted with a help of basic tools available in Linux (wget, grep, cut, sed) and regular expressions. Person names were taken from web pages describing name meanings and calendars with names for a day; in the case of surnames – the PESEL database was used; for cities and road names we used the web page of the Central Statistical Office; and country names were acquired from Wikipedia.

IG, in comparison to PG, lacks many inflected forms of words. To fill this gap we implemented an automatic procedure of acquiring inflected PN forms from large un-annotated corpora, such as Polish Wikipedia. For every word form found in the Wikipedia but not present in our gazetteer we generated all potential lemmas by applying *Guesser* — a program which tries to guess the lemma on the basis of the word's suffix. If the guessed lemma for a given word is present in the gazetteer, then it is treated as an unknown form of known name and is added to the gazetteer. This procedure was applied to all gazetteers. In this way, we extended the IG by 14%.

Because we are interested in gazetteers as a resource to recognize PNs in the evaluation we counted all matched annotations. Only, within the same annotation type, the longest matching rule was applied. This model simulates the highest recall that could be achieved if always the correct annotation were selected.

We have tested several configurations of gazetteers on CSER and the best result is presented in Table 5. The combined PG + IG gazetteer yielded an average F_1-measure of 35.27%. The worst results were obtained for *road names* and *surnames* which are very ambiguous. The highest precision of 69.48% was obtained for *country names* probably due to the long list of possible values providing good coverage in comparison to *surnames*, *road names* and *city names*.

Table 5. Performance of recognition PNs with gazetteers on **CSER**

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**TOTAL**
CSER						
Precision	44.95%	11.73%	32.99%	69.48%	7.25%	23.49%
Recall	88.19%	43.54%	72.30%	83.57%	67.59%	70.78%
F_1	59.55%	18.48%	45.31%	75.88%	13.10%	35.27%

Table 6. Performance of recognition PNs with gazetteers on **CPR** and **CEN**

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**TOTAL**
CPR						
Precision	83.54%	34.65%	30.73%	100.00%	8.20%	42.19%
Recall	82.28%	42.58%	35.08%	92.59%	50.00%	55.98%
F_1	82.90%	38.21%	32.76%	96.15%	14.09%	48.12%
CEN						
Precision	52.91%	28.75%	10.00%	90.43%	0.45%	31.65%
Recall	64.54%	31.84%	36.83%	95.87%	35.48%	61.42%
F_1	58.15%	30.22%	15.72%	93.07%	0.88%	41.77%

Nevertheless, the precision is surprisingly low. One of the main reasons is incorrect recognition of adjective *Polski* '*Polish*' as a country name, which has the same form as a genitive form of *Poland*. The adjective as a part of an organization name is written from an upper case. The other reason is incorrect recognition of country name boundaries, e.g., in *Stany Zjednoczone Ameryki Północnej* '*United States of North America*' only *Stany Zjednoczone Ameryki* '*United States of America*' were recognized — this shows that the dictionaries are incomplete. For CPR and CEN the gazetteer obtained 48.12% and 41.77% F_1-measure, respectively (see Table 6).

The gazetteers achieved lower precision than the method based on heuristics, but better recall, especially for corpora different than the one on which the rules were developed. The precision of the gazetteer-based recognition seems to be surprisingly low at first glance. There are several reasons for the low precision. Firstly, many False Positives (FPs) were caused by names that are also common words occurring at the sentence beginnings. In some cases the common word is a part of a multiword expression, for example *Zysk operacyjny* '*operating profit*' or *Przychody netto* '*net revenue*'. Moreover, some general names of positions, groups, and events are written starting with an upper case letter, e.g., in *Walne Zgromadzenie* '*general meeting*' *Walne* is incorrectly recognized

as a city name. Recognition of multi-word expressions can reduce these types of errors. Some expressions match only partially gazetteer's entries and the corresponding sub-sequences are recognised and annotated as complete PNs, e.g., *Centrum* from *Polskiego Centrum Akredytacji* '*Polish centre of certification*' is incorrectly recognized as a city name. Finally, the precision can be decreased by a potential type ambiguity of entries in the gazetteers (i.e., different PN types are assigned to the same expression across different gazetteers), e.g. 7.89% names in extended IG, 2.27% in IG and 0.37% in PG are ambiguous (see Table 4). Deficiences in recall of the gazetteers are caused by: missing lemmas and word forms, erroneous inflected forms (e.g. automatically inflected) and occurrences of alternative forms of a PN.

We used two baselines from the two sources of knowledge, respectively: heuristic rules and gazetteers. Since both approaches were developed independently and the heuristics utilize only the contextual information, there was no straightforward way of combining them. We could only simply merge results obtained with both methods, but the effect is predictable: increased completeness at the expense of decreased precision. Any other solution than a simple union of recognized mentions loses simplicity, required of a baseline to facilitate its comparison with the proposed methods. More sophisticated utilization of gazetteers and heuristic rules with HMM was proposed in Sections 6.3 and 6.4, respectively.

6. Hidden Markov model

We applied an existing HMM implementation called *LingPipe2008* (Alias-i., 2008), which is based on the first-order HMM.

6.1. HMM internal states

Following Carpenter (2006), for each annotation of a type T seven hidden states are defined in the HMM architecture, namely: B-T for the first token in a multi-token entity of type T, M-T for the internal token in a multi-token entity of type T, E-T for the last token in a multi-token entity of type T, W-T for a single token entity of type T, B-O-T for a token not in an entity following an entity of type T, E-O-T for a token not in an entity preceding a token of type T and W-O-T for a single token between two entities and following an entity of type T. There are also 3 additional states defined: BOS for the begin-of-sentence tag, W-O for any middle token not in an entity and EOS for the end-of-sentence tag. An example of a sentence encoded in the IOB format and its description in terms of HMM internal states is presented in Fig. 3.

The probability of emitting a state for an observed word is calculated on the basis of the *n-gram language model* — n was a priori set to 10, the default LingPipe value. Transitions between states are modelled by the Maximum Likelihood Estimate over training data. The n-gram occurence and tran-

Word	Eng.	Token IOB label	HMM internal state
			BOS
Pani	*Ms*	O	W-O-BOS
Iwona	*Iwona*	B-PERSON_FIRST_NAM	W-PERON_FIRST_NAM
Nowak	*Nowak*	B-PERSON_LAST_NAM	W-PERON_LAST_NAM
-	-	O	W-O-PERSON_LAST_NAM
Majewska	*Majewska*	B-PERSON_LAST_NAM	W-PERON_LAST_NAM
mieszka	*lives*	O	B-O-PERON_LAST_NAM
w	*in*	O	E-O-CITY_NAM
Zielonej	*Zielona*	B-CITY_NAM	B-CITY_NAM
Górze	*Góra*	I-CITY_NAM	E-CITY_NAM
			EOS

Figure 3. Representation of IOB labels as HMM internal states

sition probability are smoothed with the Witten-Bell algorithm, that has one hyper-parameter Θ modifying the interpolation ratio (higher interpolation ratios favour precision over recall). In *LingPipe* Θ is set to n by default.

The HMM uses a *first-best decoder* to search for the most probable sequence of states over a given text according to the state encoding presented above. The *first-best decoder* is implemented using Viterbi's algorithm.

6.2. Language model based re-scoring

For a given sentence, represented as a sequence of words, the re-scoring algorithm takes the m most probable sequences of states produced by the HMM *m-best decoder* (one state for every word in the sequence). Here, m was a prori set to 64. The sequence of states is transformed to a sequence of annotations and a non-annotated text. The sentence is sliced according to the sequence of annotations. For every text slice the probability that the slice represents a given annotation type is calculated on the basis of the *n-gram language model* trained on the level of character sequences. The new joint probability is assigned to every sequence of states and the most probable one is taken.

The *n-gram language model* for an annotation type T counts occurrences of character sequences of length n in text slices that are annotated as T. A separate *n-gram language model* is built for every annotation type and an additional one called an *outer language model* is built for text slices that are not annotated.

6.3. Re-scoring based on gazetteers

In order to utilize the gazetteers we changed the re-scoring algorithm to favor annotations present in the gazetteers. We introduced the following rule:

> If a sequence of words is annotated as *TYPE* and is present in the gazetteer of *TYPEs* then set the probability to 1 instead of using the *language model* to determine the probability.

The other steps of re-scoring remain unchanged. In this way, we increase the probability for those state sequences that contain annotations present in the gazetteers.

6.4. Re-scoring on the basis of heuristics

As already described in Section 5.1, we have defined a set of heuristics that reached high precision (ca. 95%) but very low recall (below 10%). The set of rules was divided into two subsets: *top precision rules* — containing the best rules that reached almost 100% precision and *lower precision rules* — including the rest of rules. The *top precision rules* recognise annotations on the basis of strong text evidence that could be missed by HMM. Those rules are applied directly to the text after re-scoring. All recognized annotations are immediately added to the results without any other processing.

The direct application of the *lower precision rules* might significantly decrease the final precision, so we decided to use it as an additional source of knowledge. These rules are applied to the text before re-scoring, since some incorrect annotations might be discarded in the re-scoring phase.

7. Proper name recognition

During the evaluation we followed the tenfold cross validation scheme on CSER. CSER sentences were randomly divided into 10 sets with nearly the same number of annotations in every set.

7.1. Preliminary experiment on CSER

Text passed to the HMM can be represented as a sequence of word forms or lemmas. In the preliminary experiment we have tested which of the two settings provides better results. For both types of text representation four configurations were tested, namely:

- **hmm** – HMM itself,
- **re-score** – HMM with re-scoring as described in Section 6.2,
- **gaze** – HMM with modified re-scoring as described in Section 6.3,
- **heur** – HMM with modified re-scoring and applied heuristics as described in Section 6.4.

All tested configurations achieved F_1-measure above 80%, which is higher than the best result for the gazetteers, i.e. 35.27% (see Table 5). By applying the re-scoring to the HMM results we achieved a small improvement by ca. 0.50% for word forms and by ca. 1.50% for lemmas. The best result was obtained for *heur* — HMM on word forms with modified re-scoring (the detailed results for every annotation type are presented in Table 8).

Table 7. Result of PNs recognition on CSER (tenfold CV)

	hmm	*re-score*	*gaze*	*heur*
Word form				
Precision	82.16%	83.55%	83.55%	83.56%
Recall	**90.00%**	89.50%	89.50%	89.70%
F_1	85.90%	86.42%	86.42%	**86.52%**
Lemma				
Precision	78.03%	81.01%	79.08%	*n/a*
Recall	89.25%	88.36%	91.30%	*n/a*
F_1	83.27%	84.52%	84.75%	*n/a*

Table 8. Detailed result of PNs recognition based on *word forms* with *re-scoring*, *gazetters* and *heuristics* on CSER (10-fold CV)

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**TOTAL**
Precision	81.78%	80.72%	80.20%	61.51%	76.27%	83.56%
Recall	97.87%	92.39%	81.46%	72.15%	84.38%	89.70%
F_1	89.10%	86.16%	80.82%	66.41%	80.12%	86.52%

7.2. Error analysis

We analysed the results of the selected five-folds from the ten-fold cross validation on CSER in order to determine the most frequent causes of *false positives*. We identified 9 groups of errors, listed in Table 9. The majority of *false positives*, 114 cases, were caused by incorrect annotations of words that are a part of another annotation such as an *institution name* and a *product name*. The second group of errors were missing annotations (i.e. correctly recognized annotations that were not annotated in the tested corpus), 51 cases – 1.3% of all annotations.

The next four groups of errors were related to the PN orthography. In 47 cases expressions recognized as names started with a lower case letter, while all PNs we consider here start with an upper case letter. The next 20 cases were single letters and 12 cases contained a symbol that cannot be a part of PN. The last 6 cases did not contain a vowel, while first names and surnames should include at least one in the case of Polish. This groups of errors can be fixed by applying simple filtering rules, see the next section.

We also identified two groups of errors that were related to syntactic properties of PNs. In 5 cases the first word in a sentence was incorrectly recognized as a name, despite its inappropriate syntactic category. In some cases, the first word was followed by the reflexive marker *się* which occurs more often after a verb than a PN.

Table 9. Error analysis on first five-folds of CSER ten-fold cross validation.

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	SUM
Part of other proper name	28	33	27	16	10	**114**
Missing	4	11	16	16	4	**51**
Lower case	7	12	10	12	6	**47**
Incorrect type	16	5	7	5	0	**33**
Initial	8	10	2	0	0	**20**
With symbol	2	2	3	5	0	**12**
Beginning of sentence	5	1	0	1	1	**8**
No vowel	4	2	0	0	0	**6**
Incomplete	1	0	0	0	4	**5**

7.3. Post processing rules

On the basis of error analysis the following correction rules were constructed:

1. **BeforeSie** – removes an annotation if the corresponding text is followed by *się*. Phrases like *Upoważnia się* (Eng. *is allowed*) are incorrectly marked as a name. In expressions of the scheme '*X się*', *X* is not a PN in the majority of cases.
   Applied to: all PN types
2. **CutRoadPrefix** – excludes road indicators (ex. *ul.* 'st.', *ulica* 'street' etc.) from text annotated as road name. Road indicators are not considered to be a part of road names unless the indicator is written from an upper case and is a part the official road name, ex. *Aleje Jerozolimskie*.
   Applied to: roads
3. **FirstNotLowerCase** – removes an annotation if the corresponding text starts with a lower case character. Polish proper names do not start with a lower case character. However, some foreign surnames can, for example *d'Arc*.
   Applied to: all types
4. **HasAlphanumeric** – removes an annotation if the corresponding text does not contain any alpha-numeric characters. This rule discards sequences of symbols, like "—", "_____", ". . ." etc..
   Applied to: all types
5. **HasVowel** – removes an annotation if the corresponding text does not contain any vowel. *First names* and *surnames* must contain at least one vowel in Polish. Other names do not obey this rule as they can be abbre-

viations in fact, for example *W-Z* is a road name and *NRD* is a country name.
Applied to: first names and surnames

6. **Length** – removes an annotation if the corresponding text is one character long. All names but road names should contain at least two characters.
Applied to: first names, surnames, cities and countries
7. **NoSymbol** – removes an annotation if the corresponding text contains one of the following symbols: +, /, *, =
Applied to: first names, surnames, countries and cities
8. **NoDot** – removes an annotation if the corresponding text contains a dot. Initials that are part of person names are not considered as first names nor surnames.
Applied to: first and surnames
9. **NoPatternAaA** – removes an annotation if the first and last letter of the corresponding text are upper case and the other are lower case.
Applied to: first and surnames
10. **NoHyphen** – removes an annotation if the corresponding text contains a hyphen.
Applied to: first and surnames
11. **NoUnderline** – removes an annotation if the corresponding text contains an underline sign "_"
Applied to: all types
12. **Trim** – excludes words starting with a lower case letter appearing at the beginning and at the end of the annotated text., e.g., "i Marek" (*and Mark*) is transformed to "Marek" (*Mark*)
Applied to: all types

7.4. Evaluation of post processing

The influence of the post-processing rules was tested on the remaining five-folds that had not been examined during the construction of the rules. The results are presented in Table 10. As we can notice, the application of post-processing improved the results by ca. 2% for F_1. The best result of 89.67% F_1-measure was obtained for re-scoring with the gazetteers, heuristics and post-processing, i.e., in the case in which all three kinds of knowledge: lexical (gazetteers), structural (heuristics), and statistical (HMMs) were applied together.

8. Cross-domain evaluation

The collected CSER corpus represents one domain — economy. As the post-processing rules were constructed on the basis of observation collected from the analysis of errors performed on CSER, we have to assume that the post-processing rules are biased not only in favour of the particular corpus but also

of the particular domain. Only gazetteers were built for the general language. Thus, one must expect decreased results in applying the PN recogniser to a different corpus.

In order to evaluate the cross-domain portability of the constructed PN recognition method, we retrained the HMM model on the corpus of stock exchange reports and next evaluated it on the corpus of police reports and economic news. We have evaluated all configurations previously tested on CSER, e.g., **hmm** — pure HMM, **re-score** — HMM extended with re-scoring, **re-score post** — HMM extended with re-scoring and post-processing, **gaze post** — HMM extended with modified re-scoring (gazetteers) and post-processing, and **heur post** — HMM extended with modified re-scoring (gazetteers, heuristics) and post-processing.

The summary of achieved results is presented in Table 11. As we could expect, the results of the cross-domain application are much worse than on the source corpus (CSER, compare with Table 10), e.g., the best result on CSER is 89.67% F_1-measure, but only 74.62% on CPR and 64.57% on CEN (the detailed results for the best configuration are presented in Table 12). Since CEN includes texts of richer variety, the performance of our HMM-based method is much worse than for CPR, but still all variants give an improvement in comparison to the gazetteer-based recognition alone (the top baseline for every corpus is presented in the first column). We observed similar tendency between following configurations as for CSER. The modifications introduced in CSER that improved the results also improved results in the other two corpora. We observed only one exception, in CPR re-scoring applied on pure HMM decreased precision. However, the re-scoring combined with post-processing gives better improvement than post-processing of pure HMM.

9. Summary

In the paper we presented a Proper Name Recognition algorithm based on the Hidden Markov Model (HMM), where recognition of Proper Names is treated as the core of the general Named Entity Recognition problem. Ambiguity of Proper Names on the level of their forms and semantics makes an approach based exclusively on hand-written rules difficult to apply on a large scale covering various texts and domains (e.g., cadastral domain in Abramowicz et al., 2006). Hand-written rules were successfully applied only for heterogeneous text (e.g., police reports in Graliński et al., 2009b, medical reports in Mykowiecka et al., 2007). Methods based on Machine Learning (ML) depend heavily on the used training corpus. No large corpus of Polish annotated with NEs is available yet. Thus, we aimed at a combination of a domain-depended method based on ML with domain independent methods based on gazetteers and rules that capture a general structure and contextual use of Polish PNs. The HMM was selected as an ML method due to its earlier promising applications in Named Entity Recognition for other languages. Domain independent knowledge was

Table 10. Comparison of PN recognition results on the CSER folds 6–10 without (*before*) and with post-processing (*post*)

	re-score		**gaze**		**heur**	
	before	*post*	*before*	*post*	*before*	*post*
Precision	*86.15%*	89.78%	*85.16%*	88.67%	*85.19%*	88.69%
Recall	*89.28%*	89.33%	*90.43%*	90.48%	*90.63%*	90.68%
F_1	*87.69%*	89.56%	*87.72%*	89.56%	*87.83%*	89.67%

Table 11. Summary of cross-domain evaluation on CPR and CEN

	baseline	**hmm**	**re-score**	**re-score post**	**gaze post**	**heur post**
			CPR			
Precision	42.19%	49.09%	42.22%	66.08%	67.65%	**67.86%**
Recall	55.98%	64.44%	78.80%	78.98%	82.47%	**82.87%**
F_1	48.12%	55.73%	55.00%	71.96%	74.33%	**74.62%**
			CEN			
Precision	31.65%	47.91%	40.86%	54.88%	57.14%	**57.46%**
Recall	61.42%	47.51%	65.88%	66.04%	72.96%	**73.70%**
F_1	41.77%	47.71%	50.44%	59.95%	64.09%	**64.57%**

Table 12. Best results of cross-domain evaluation on CPR and CEN

Names:	*first names*	*surnames*	*cities*	*countries*	*roads*	**TOTAL**
			CPR			
Precision	66.50%	70.12%	79.55%	60.00%	73.33%	**67.86%**
Recall	83.79%	86.91%	76.09%	54.55%	84.62%	**82.87%**
F_1	74.15%	77.62%	77.78%	57.14%	78.57%	**74.62%**
			CEN			
Precision	53.02%	54.43%	48.18%	76.29%	12.96%	**57.46%**
Recall	86.33%	79.76%	66.97%	63.60%	22.58%	**73.70%**
F_1	65.70%	64.71%	56.05%	69.37%	16.47%	**64.57%**

expressed by gazetteers that had been constructed manually, but also collected from freely available sources on the web. Recognition and post-processing rules express intermediate characteristics: they were developed by the analysis of a particular domain-dependent corpus, but they reflect more general properties of PNs as Polish language expressions. The HMM re-scoring mechanism was applied to integrate different knowledge sources in the PN recognition. The proposed gazetteer acquisition method solves partly the problem of the recognition of different morphological forms of Polish PNs.

The algorithm was tested on three domain corpora and compared with two simple baseline algorithms: one based on the heuristic of the first upper case letter in all PN elements (Section 5.1) and another based on a gazetteer only (Section 5.2). The algorithm was trained on the domain corpus of stock exchange reports and yielded promising results when evaluated according to the ten-fold cross validation scheme (Table 10): recall of 90.68%, precision of 88.69% and F_1 measure of 89.67%. The best results were achieved with the gazetteers and heuristics integrated in the re-scoring phase and post-processing applied as the last step.

Results achieved in recognition of person names and surnames are better than for other sub-classes. The result achieved for country names is surprisingly low. It is caused by a few frequent expressions that are ambiguous between a PN interpretation and an adjective interpretation, e.g., *Polski* can be interpreted as a noun in the genitive case or as an adjective in the nominative case. As our post-processing rules do not refer to morphosyntactic information, errors of this type cannot be eliminated yet. A suitable extension of the rule format is our main goal for the work in progress.

The results achieved on the domain corpus are promising given the modest efforts required for preparing the training corpus and applying the ready-to-use LingPipe tool for HMMs. As other knowledge sources are domain independent, the only cost of the adaptation of the method to another domain is preparation of the training corpus. The annotation of the corpus of stock exchange reports (CSER) took about 26 person hours. In that time 4 011 annotations of five types were added, which gives the average speed of text processing of ca. 181 tokens/minute and ca. 2.6 annotations/minute. The high speed of annotation was possible due to the fact that only words starting with an upper case letter had to be considered by the annotator.

We also tested the portability of the method to another domain represented by two corpora: the corpus of police reports (CPR) and the corpus of economic news (CEN). The algorithm was trained on the CSER corpus and next applied to both CPR and CEN corpora. Rules were developed only by the analysis of the CSER corpus. The obtained results were significantly worse for both test corpora, 74.62% and 64.57% of F_1 measure for CPR and CEN, respectively. However, it is worth emphasising that in the case of both corpora the achieved results were significantly better than the baseline gazetteer approach. Thus, the knowledge acquired by machine learning from the training corpus can be partly

transferred to other domain. Also the component of simple hand-written rules improved significantly the precision.

Concerning a possible future NER algorithm of a wide applicability, three types of knowledge should be combined in its construction: (a) domain-independent knowledge extracted by means of machine learning from a general training corpus, (b) domain-specific knowledge extracted from a domain-specific corpus by means of ML, and (c) hand-written rules focused on post-processing. Thus, we plan to investigate the applicability of different types of machine learning methods to different types of knowledge. Methods focused on probability, like HMMs, seem better suited for building domain-specific components, but for capturing the domain-independent aspects we need a method which allows for easier generalisation even for the sake of coverage.

Acknowledgements The work was co-funded by the European Union Innovative Economy Programme project POIG.01.01.02-14-013/09.

References

Abramowicz, W., Filipowska, A., Piskorski, J., Wecel, K. and Wieloch, K. (2006) Linguistic Suite for Polish Cadastral System. In: *Proceedings of the LREC'06*. ELRA, Genoa, Italy, 53-58.

Alias-i. (2008) LingPipe 3.9.0., http://alias-i.com/lingpipe, (October 1, 2008).

Broda, B., Piasecki, M. and Radziszewski, A. (2008) Towards a Set of General Purpose Morphosyntactic Tools for Polish. In: *Proceedings of the 16th International Conference Intelligent Information Systems*. Academic Publishing House Exit, 441–450.

Carpenter, B. (2006) Character language models for Chinese word segmentation and named entity recogntion. In: *Proceedings of the 5th ACL Chinese Special Interest Group (SIGHan)*, Sydney, Australia. ACL, 169–172.

Graliński, F., Jassem, K. and Marcińczuk, M. (2009a) An Environment for Named Entity Recognition and Translation. In: L. Màrquez and H. Somers, eds., *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. EAMT, 88–95.

Graliński, F., Jassem, K. and Marcińczuk, M. (2009b) Named Entity Recognition in Machine Anonymization. Kłopotek In: M. A., Przepiorkowski A. A., Wierzchoń T. and Trojanowski K. , eds., *Recent Advances in Intelligent Information Systems.* Academic Publishing House Exit, 247–260.

Katrenko, S. and Adriaans, P. (2007) Named Entity Recognition for Ukrainian: A Resource-Light Approach. In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. ACL, Prague, Czech Republic, 88–93.

Kravalová J. and Žabokrtský, Z. (2009) Czech named entity corpus and SVM-based recognizer. In: *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Suntec, Singapore. ACL, 194–201.

LDC (2008) ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (Version 6.6). Technical report, Linguistic Data Consortium.

Malouf, R. (2002) Markov models for language-independent named entity recognition. In: *Proceedings of the Sixth Conf. on Natural Language Learning (CoNLL-2002)*. ACL, 183–186.

Marcińczuk, M. (2007) *Pattern Acquisition Methods for Information Extraction Systems*. Master's thesis, Blekinge Tekniska Högskola, Sweden.

Marcińczuk, M. and Piasecki, M. (2007) Pattern Extraction for Event Recognition in the Reports of Polish Stockholders. In: *Proceedings of the International Multiconference on Computer Science and Information Technology*, Wisła, Poland. IMCSIT, **2**, 275-284.

Marcińczuk, M. and Piasecki, M. (2010a) Named Entity Recognition in the Domain of Polish Stock Exchange Reports. In: M.A. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek and S.T. Wierzchoń, eds., *Proceedings of the 18th International Conference Intelligent Information Systems*. Wydawnictwo Akademii Podlaskiej, Siedlce, 127-140.

Marcińczuk, M. and Piasecki, M. (2010b) Study on Named Entity Recognition for Polish Based on Hidden Markov Models. In: P. Sojka, A. Horák, I. Kopecek and K. Pala, eds., *Proceedings of Text, Speech and Dialogue: 13th International Conference, TSD 2010*. **LNCS 6231**, 142-149.

Marrero, M., Sánchez-Cuadrado, S., Lara, J.M. and Andreadakis, G. (2009) Evaluation of Named Entity Extraction Systems. *Research in Computing Science*, 41:47–58.

Mykowiecka, A., Kupść, A., Marciniak, M. and Piskorski, J. (2007) Resources for Information Extraction from Polish texts. In: *Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Wydawnictwo Poznańskie, Poznań, 99–103.

Osenova, P. and Kolkovska, S. (2002) Combining the Named-entity Recognition Task and NP Chunking Strategy for Robust Pre-processing. In: *Proc. of The 1st Workshop on Treebanks and Linguistic Theories, Sozopol, Bulgaria*. Bulgarian Academy of Sciences, 167–182.

Piasecki, M. (2007) Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, **11**(1–2):151–167.

Piasecki, M. and Radziszewski, A. (2007) Polish Morphological Guesser Based on a Statistical A Tergo Index. In: *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications*. IMCSIT, 247–256.

Piskorski, J. (2004a) Extraction of Polish named entities. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (ELR, 2004)*. ACL, 313–316.

Piskorski, J. (2004b) Named-Entity Recognition for Polish with SProUT. In: L. Bolc, Z. Michalewicz and T. Nishida, eds., *Intelligent Media Technology for Communicative Intelligence*. **LNCS 3490**, Springer, 122–133.

Savary, A. and Piskorski, J. (2010) Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish. In: M.A. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek and S.T. Wierchnoń, eds., *Intelligent Information Systems*. Wydawnictwo Akademii Podlaskiej, Siedlce, 141–154. http://iis.ipipan.waw.pl/2010/proceedings/iis10-14.pdf.

Urbańska, D. and Mykowiecka, A. (2005) Multi-words Named Entity Recognition in Polish texts. In: *SLOVKO 2005 – Third International Seminar on Computer Treatment of Slavic and East European Languages, Bratislava, Slovakia*. VEDA Vydavatel'stvo Slovenskej akademie vied, 208–215.

Woliński, M. (2006) Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In: *Proceedings of IIS:IIPWM'6*. Springer, 503–512.

Zhou, G. and Su, J. (2002) Named Entity Recognition using an HMM-based Chunk Tagger. In: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL, 473–480.