Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 145–151

Knowledge-based Named Entity Recognition in Polish

Aleksander Pohl
Jagiellonian University
ul. Łojasiewicza 4, 30-348 Kraków, Poland
Email: aleksander.pohl@uj.edu.pl

***Abstract*—This document describes an algorithm aimed at recognizing Named Entities in Polish text, which is powered by two knowledge sources: the Polish Wikipedia and the Cyc ontology. Besides providing the rough types for the recognized entities, the algorithm links them to the Wikipedia pages and assigns precise semantic types taken from Cyc. The algorithm is verified against manually identified Named Entities in the one-million sub-corpus of the National Corpus of Polish.**

I. Introduction

IN THE recent years the research conducted in the field of Information Extraction (IE) has brought many interesting results. First of all at the end of the twentieth century researchers have overcome the first major obstacle preventing wide-spread adoption of IE systems. Namely, by incorporating weakly supervised methods they were able to quickly adopt systems to new domains, by providing only a handful of training examples. Works of Brin [1] as well as Agichtein and Gravano [2] showed that large-scale information extraction is feasible, provided that we have access to large corpora, WWW in particular.

The ideas from the previous century were developed in two, slightly contradictory directions. First of all the term Open Information Extraction was introduced in works of Banko et al. [3]. This approach towards IE was pushed further in systems such as ReVerb [4] which does very well in the task of relation extraction. The goal of Open IE, much influenced by Information Retrieval is extraction of *all* information found in web-documents. Since it is very hard to build an ontology covering all the phenomena that might be described on the Web, these methods make very minimalistic assumptions about the world. As a result the extracted information is not transformed into some coherent semantic framework, but rather it is assumed that the user will find the relevant information by refining his/her query, the same as in traditional search engines.

The opposite approach towards IE comes from the researchers from the SemanticWeb camp. The primary difference comes from the fact that they are trying to utilize various knowledge-sources, especially taxonomies and ontologies, in order to extract the information available in the textual data and transform them into well established and broadly accepted schemas. In that form the information might be further automatically processed and consumed by intelligent systems. As a result, there are several important problems that have to be resolved, which are absent in Open IE systems. First of all, the concepts that are extracted must be identified within some ontology. The ambiguities also have to be resolved, especially among the relations. That approach is characteristic for the systems such as DBpedia Spotlight [5] and AIDA [6], both utilizing Wikipedia as the reference resource for concept identification. Recently Exner and Nugues [7] showed how relation extraction might be performed combining well known Natural Language Processing (NLP) techniques and DBpedia both as a reference conceptual scheme as well as source of examples used for training relation classifiers.

However if we look at the development of IE methods that are employed for Polish, we will discover that the state of research resembles the state of research in English in mid-nineties. To the best of our knowledge there are only two systems [8], [9] that perform template filling. Both of them employ manually constructed extraction grammars and operate in closed domains – the first one extracts information from mammogram reports and the second from patients' diabetic records. Regarding relation extraction there is only limited research in the field. [10] describes methods for automatic extraction of semantic relations aimed at development of the Polish Wordnet. The described method uses massive amounts of data in order to determine the best location for a new synset in a large semantic network. On the other hand [11] describes an ontology-based method for the extraction of relations form a bibliographical lexicon based on manually defined grammars. The only IE task that is well developed is Named Entity recognition[1], but the methods still rely on manual construction of grammars or tagging of large amounts of data.

As we know from the development of systems aimed at English these bottlenecks might be overcome if appropriate methods are employed. This work is a step in this direction. It presents an algorithm that does not require any manual construction of extraction grammars, nor tagging of large amounts of data. By utilizing Wikipedia link structure and classification of Wikipedia articles into the Cyc ontology [12] the system is able to precisely classify and link the entities to Wikipedia articles. Since there are knowledge-bases such as DBpedia [13] that are based-upon Wikipedia and Cyc is the largest known manually constructed ontology, this opens interesting possibilities for automatic processing of the extracted data by intelligent systems.

[1]The next section contains a detailed description of the available systems.

978-1-4673-4471-5/$25.00

TABLE I
EVALUATION OF SPROUT-BASED NER SYSTEM PRESENTED IN [19].

Category	Precision [%]	Recall [%]
People	90.6	85.3
Locations	88.0	43.4
Organizations	87.9	56.6

TABLE II
EVALUATION OF SPROUT-BASED NER SYSTEM PRESENTED IN [20]. ONLY THE BEST RESULTS ARE REPORTED.

Category	Precision [%]	Recall [%]
People	91.0	78.0
Locations	82.0	72.0
Organizations	92.0	52.0

II. RELATED WORK

Named entity recognition (NER) is a vast topic. There are many approaches towards it and systems that actually perform NER. To constrain the review of such systems we present only the work on NER in Polish.

The research on NER in Polish starts with the works of Piskorski [14], [15], [16]. These works describe the adaptation of the SProUT platform [17] for Polish and its application in the recognition of entities such as: time, percentage, money, organizations, locations and people.

Applying SProUT to NER requires manual definition of the rules, which in general consist of a pattern/action pairs built upon typed feature structures (TFS). The LHS of a rule matches segments with particular features and the RHS defines the TFS that is produced as the result of that match. Besides basic matching of words and features, the rules allow to use variables and call other rules. As a result the formalism is both expressive and transparent.

The adoption of SProUT for Polish required integration of a morphological analyzer for Polish – Morfeusz [18]. Also gazeteers were used and the morphological variants were produced automatically. The author elaborates on problems that are specific to Slavonic languages, Polish in particular, especially the rich morphology of names as well as problems with name segmentation. This causes serious problems for NER since there are cases where it is impossible to determine the base form of a NE without appealing to data such as verb frame subcategorization, which is usually out of scope of NER. The constructed rules were evaluated on a set of 100 news articles. The values are presented in Table I.

A more elaborate evaluation of this approach is presented in [20], where it is applied in the domain of cadastral information. The described system used a more elaborated entity classification taxonomy consisting of nearly 30 classes. It it also evaluated on a much bigger corpora counting more than 4 thousand entities. The evaluation methodology was more elaborate were exact and partial matches were counted separately. The results of the evaluation are presented in Table II.

A more lightweight approach is presented in works of Graliński et al. [21], [22] and Walas et al. [23]. The authors find SProUT formalism too complex for the tasks of NE translation and anonymization. [21] identifies errors in NE translation as one of the important problems that make the translation hard to understand. What is more 10% of errors in machine translation (MT) are due to invalid processing of named entities. [22] uses NER in task of anonymization of police reports that are used to improve MT. Since the researchers are not allowed to see the names of the suspects, they provide a set of Word macros that are used to remove sensitive data (such as names of people and companies) from the documents. [23] describes NER in the context of a question answering (QA) system.

Each of these systems uses a formalism based on Spejd [24]. The rules are less complex than in SProUT so the results are less precise. For instance in [21], depending on the evaluation scheme the precision varies between 76 to 88%. On the other hand [21] and [23] report that special handling of NEs improves both MT and QA.

Only recently the researchers of Polish NER systems started to use machine learning approach. Marcińczuk at al. [25] describes application of Hidden Markov Models (HMM) for this task, which was restricted to the recognition of names of people and organizations. The authors used HMM with 7 hidden states for each NE type. The transitions between states were modeled by the maximum likelihood over the training data, while tag emissions were generated by n-gram character language models with generalized form of Witten-Bell smoothing [26]. The authors used LingPipe [27] to compute the parameters of the HMM.

The system was trained on a corpus containing stock exchange reports with 670 person mentions and more than 3 thousand company mentions. It was evaluated on the same corpus (using 10-fold cross validation) and on a corpus containing police reports. The vanilla HMM reached 86% for people names and 80% for companies names (F_1 metric) on the stock exchange corpus. In order to improve the precision of the system the authors applied two heuristics: filtering and trimming. The first one imposed some restrictions on the form of the names, the second one stripped preceding and following words if they did not start with an upper case. The first heuristic improved the precision of the method from 63% to 89% with only small reduction of recall. The authors also report the evaluation of the system on the police report corpus. The results were significantly worse (55% F_1 metric), showing that this approach towards NER is highly domain-dependent.

The latest achievement in Polish NER is described in [28]. The authors used Conditional Random Fields (CRF) model [29] to overcome the limitations of HMM model, e.g. ignorance of the right context of the name. They were detecting 5 types of entities: first names, surnames, country names, city names and road names. The developed model incorporated not only statistical data, but also knowlege, e.g. the list of names of peoples positions and titles taken from the Polish WordNet [10]. In general there were 5 types of features that were used to build the model: ortographic, binary-ortographic, WordNet-based, morphological and gazetteer-based.

The model was trained on one corpus (used in the previous works of Marcińczuk et al. [25]) and tested on that corpus following 10-fold cross-validation scheme and on two others corpora (police reports corpus and corpus of economic news). The results on the first corpus substantially outperformed the HMM model, reaching F_1 score of 92.5%. On the other hand the evaluation performed on the two other corpora showed that the drop in performance is significant – the model achieved 67.7% F_1 score on the police reports corpus and 72.3% F_1 on the economic news corpus, meaning that the model fits well to the training data, but causes problems when ported to different (even highly related, like in the case of stock exchange reports and economic news) domains.

The primary difference of the presented algorithm with the already described methods comes from the fact that this algorithm does not require any manual construction of NE recognition rules, nor tagging of the training data. As such it does not require any manual intervention, besides the *selection* of the name types that should be identified. The other difference comes from the fact that this algorithm is tested on the 1-million sub-corpus of the National Corpus of Polish [30], to the best of our knowledge the first such demonstration. This is particularly important, because the corpus contains a variety of language styles and text sources, contains a large number of NEs and allows for a comparison of different algorithms in a unified setting.

III. Entity recognition

The algorithm employed to recognize NEs is an adaptation of the Word Sense Disambiguation (WSD) algorithm described in [31]. That algorithm builds on the Wikipedia Miner disambiguation algorithm described in [32]. The original algorithm of Milne and Witten works as follows. First of all it recognizes occurrences of phrases used in Wikipedia to link to other Wikipedia articles. Since many of these phrases are ambiguous, e.g. *Washington* may refer to a state, a city, a person, a football club, a university and almost four hundred other concepts, the algorithm tries to disambiguate its meaning by employing machine-learning techniques. First of all it computes several features of each candidate target concept:

1) an average weighted *semantic relatedness* with other (unambiguous) concepts
2) a *probability* of the candidate concept
3) a *context goodness*, i.e. a measure showing if the context of the phrase is well defined

The *semantic relatedness* between two concepts[2] in the original algorithm is computed using the Normalized Google Distance (NGD) [33]:

$$sr_G(a,b) = 1 - \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))} \quad (1)$$

Where:

- $sr_G(a,b)$ – the measure of semantic relatedness between a and b,
- $|A|$ – the size of the set of articles that link to a,
- $|A \cap B|$ – the size of the set of articles that link both to a and b,
- $|W|$ – the number of all articles in Wikipedia.

The weights of the concepts are established according to their semantic relatedness with other concepts and the link probability of phrases they are identified by (e.g. the ratio of the number of phrase occurrences as link to the total number of phrase occurrences).

The *probability* of the candidate concept is computed as the ratio of the number of links with that phrase pointing to that concept to the total number of links containing that phrase.

When the features are computed, the algorithm tries to select the most probable meaning by employing a machine learning algorithm – C4.5 in this case [34]. The algorithm of Milne and Witten uses also Wikipedia link structure to train the classifier. Wikipedia links are used as positive examples for *phrase–concept* pairs. The negative examples are constructed from all the remaining concepts that could be linked via this phrase (i.e. there are links with the same anchor name that point to the other concepts). Since the internal link structure of Wikipedia is very dense, it is easy to generate millions of the training examples.

The primary improvement described in [31] is the usage of a Jaccard coefficient-inspired measure instead of the NGD:

$$sr_J(a,b) = \begin{cases} \frac{1}{1 - log\left(\frac{|A \cap B|}{|A \cup B|}\right)} & |A \cap B| > 0 \\ 0 & a \neq b \wedge |A \cap B| = 0 \\ 1 & a = b \wedge |A \cap B| = 0 \end{cases} \quad (2)$$

Where:

- $|A \cup B|$ – the size of the set of articles that link to a or b

The second improvement is the employment of several additional features that are already used in the original algorithm (but not in the classifier) or are trivial to compute:

- the *rank of the semantic relatedness* of the candidate concept among all the candidate concepts,
- the *rank of the probability* of the candidate concept among all the candidate concepts,
- the *link probability* of the concept.

By employing the different semantic relatedness measure and the additional features the automatically computed performance of the algorithm measured as the weighted precision and recall (F_1) jumped from 84 to 92.4% in the English dataset and from 83 to 91.7% in the Polish dataset. What is more by using the latest Polish Wikipedia dataset[3], it was measured that the performance of the algorithm was further increased to 93.5%.

Direct application of the WSD algorithm results in the identification of phrases that might be (unambiguously) linked to the Wikipedia articles. But its application in Named Entity Recognition requires several adjustments that are covered in the next section.

[2]Although this is not fully accurate, we identify the Wikipedia articles with concepts.

[3]Dump from the 2^{nd} of May 2013.

IV. ENTITY CLASSIFICATION

The regular definition of the Named Entity Recognition task [35] requires that the entities are recognized and *classified* into some well-defined conceptual scheme. Although in the time of MUC conferences [36] it was required that the classification scheme makes fine-grained distinctions allowing for e.g. distinguishing between civil and military objects[4], in recent years, e.g. in the tasks defined in the scope of ACE initiative [37], the conceptual scheme was substantially simplified. ACE 2008 defines 5 general types divided into 31 subtypes. The general types are as follows:

- facility
- geo-political entity
- location
- organization
- person

Similarly in the National Corpus of Polish [30] the following NEs are identified:

- personal names
- geographical names
- names of organizations and institutions
- words related to the above categories
- basic temporal expressions

This simplification of the conceptual scheme stems from the fact that these resources are domain-independent and as such, should make as few ontological assumptions as possible. Especially because in such domain-independent environment it is hard even for people to assign fine-grained classes, let alone force research teams to adopt one „golden" ontology.

But when it comes to the application of the results of IE in some specific domain, such sketchy schemas have limited value. If we wish to conduct intelligence analysis we are not interested in people in general, but those who have important roles in their societies. If we are considering financial investments, we are not interested in organizations „in general", but only those which operate on a specific market, e.g. mining companies. As a result, the more we are concerned with the application of the IE results, the more fine-grained conceptual scheme is needed.

In order to make the results of NEs recognition useful in practical settings, the Cyc ontology [38] is used as the taxonomy of the entities. Although there are alternatives such as YAGO [39] and DBpedia ontology [40], which are directly mapped to Wikipedia, both of them have certain deficiencies making them less useful for automatic processing of the extracted knowledge. Although YAGO's taxonomy is very dense, since its classes were extracted from Wikipedia's category scheme, it lacks features such as well defined disjointness relation. Although its authors have used heuristics to detect inconsistencies, they were error-prone. As a result there are contradictory facts in YAGO, e.g. *Gertrude Stein* is classified both as a *person* and as a *literary work*.

[4]It was assumed that a terrorism activity is targeted only at civil objects, such as churches, shopping centers and railway stations.

TABLE III
THE COVERAGE OF METHODS USED TO CLASSIFY THE ARTICLES IN THE POLISH WIKIPEDIA.

Method	Count (thousand)	Percentage
English Wikipedia [12]	283.9	25.8
Infobox mapping	534.9	48.6
People heuristic	213.1	19.4
Total	**688.6**	**62.6**

On the other hand the DBpedia ontology[5] is rather small (contains less than 400 classes) and lacks coverage. Since the entities are assigned to the classes according to the infoboxes attached to the articles and less than 2 million of articles (out of 4 million) have such infoboxes attached, many of the entities lack their type.

Cyc [38] on the other hand is a long running effort of building an ontology that would allow intelligent systems to perform common-sense inferences. As a result it contains thousands of well defined semantic classes. These classes are not only organized into subsumption relation, but they are also inter-related with disjointness relation. This feature of Cyc makes it particularly useful for ensuring high accuracy of the extracted informations, since inconsistencies in type assignment might be resolved automatically.

Although Wikipedia is not aligned with Cyc in its entirety, in our recent effort [12] we created a classification of more than 2.2 million of the English Wikipedia entities into the Cyc ontology with precision reaching 93%. That mapping was made available in the N-triple format, with links to the English DBpedia[6] and OpenCyc[7]. In order to use that mapping in the Polish Wikipedia the following procedure was employed:

1) The interlingual links from the latest DBpedia[8] were downloaded.
2) The interlingual links were used to establish a mapping between the English and the Polish DBpedia entities.
3) The Cyc types were assigned to the corresponding Polish DBpedia entities.

There were 626 thousand of links in the English DBpedia that were employed in the second step. But since the type assignment covered only 2.2 million of articles (out of 4 million) the number of classified articles in the Polish Wikipedia was 284 thousand. To extend the coverage of the method infoboxes from the Polish Wikipedia were mapped to terms in Cyc and a simple heuristic for people marking all articles with *Urodzeni w* (Eng. *Born in*) and *Zmarli w* (Eng. *Died in*) categories as instances of `Person` class was applied. The first heuristic yielded 535 thousand of classifications and the second 213 thousand. Since the methods were overlapping, the final result was a classification of 689 thousand of Polish Wikipedia articles (out of 1.1 million). The results are summarized in Table III.

[5]`http://dbpedia.org/Ontology`
[6]`http://klon.wzks.uj.edu.pl/wiki-types/`
[7]`http://www.cyc.com/platform/opencyc`
[8]`http://wiki.dbpedia.org/Downloads38`

The second adjustment which was applied to the Wikipedia-based disambiguation algorithm concerned filtering of the entities that were provided as the NER result. Since contents of Wikipedia is encyclopedic, it covers descriptions both of classes and individuals. The WSD algorithm does not make distinction between these types of entities. However, the result of NER should be a discovery of *names*, i.e. proper names (usually) of individual things. We are not expecting from the NER system to annotate occurrences of words such as *woman*, *child* or *man*, because they are instances of common nouns. Yet, the algorithm is able to properly disambiguate these words and its original version produces such results. So the filtering was concerned with rejecting the instances of common nouns and similar expressions from the final result.

In general, making the distinction between classes and individuals is not a trivial task. Since it was not the primary concern of this research, a simple, yet powerful heuristic was employed: for each Wikipedia article the total number of links starting with an uppercase letter and a lowercase letter was counted. If the number of links starting with an uppercase letter was greater or equal to the number of links starting with a lowercase latter, the article was considered as describing an individual. A small test performed on a random sample of 200 entities showed that this heuristic is able to produce distinctions with precision above 90%. This result is very good compared to the results of a similar task of distinguishing between types and instances in the Wikipedia category system [41], where several heuristics were employed, yielding similar precision.

V. Evaluation

The evaluation of the algorithm was performed on the one-million sub-corpus (ONEM) of the National Corpus of Polish [42] with manually annotated named entities [30], [43]. ONEM contains excerpts from many sources: literature, newspapers, the Internet, speech transcripts and other. It contains 3888 excerpts which amount to more than 50 thousand of NEs.

The annotation covered identification of names referring to people, locations, organizations and temporal expressions. Some of these types were subdivided into subtypes. Table IV summarizes the taxonomy of the entities. It should be noted that not only the nominal expressions were annotated, but also all other types of expressions that refer to NEs, adjectives in particular. So in the following expressions *restauracja w **Warszawie*** (Eng. *restaurant in Warsaw*) and ***warszawska** restauracja* (Eng. *Warsaw restaurant*), the boldfaced fragments were annotated as referring to *Warsaw*. These occurrences of NEs are called *relational expressions*.

Although, in general proper names do not follow the principle of compositionality, since their contents might be arbitrary, it was decided that the annotation covers also subordinate names, which are parts of larger units. E.g. in the expression *Dom dziecka w Oruni* (Eng. *Orphanage in Orunia*), the whole expression is annotated as a name of an organization, while *Oruni* is annotated as a name of a location. From the point of view of the evaluation of NER this gives some added value, since algorithms might be rewarded for accurate partial matches within longer expressions.

TABLE IV
The types of entities annotated in ONEM [30].

Type	Subtypes	Description
persName	forename surname addName	given name family name nickname
orgName		an organization
geogName		a geographical entity (excludes geopolitical entities)
placeName	district settlement region country bloc	a district within a city a city or a village a country region a country a bloc of countries
date		at least partially determined date
time		at least partially determined time

TABLE V
The counts of entities annotated in ONEM. Time and date expressions are skipped.

Entity type	Test corpus	Tuning corpus
persName	19619	336
orgName	10914	217
geogName	4001	69
placeName	15432	311
total	**49966**	**933**

Regarding the scope of the evaluation – it did not cover date and time expressions, since the WSD algorithm is not aimed at theses types of expressions. What is more, it is apparent that these expressions are domain independent and can be captured by a well developed grammar or some other NER technique. On the other hand the evaluation included the relational expressions, which are quite hard to spot using traditional NER methods.

During the developement of the algorithm, the ONEM corpus was split into two parts – one containing 100 randomly selected text excerpts, used for the tuning of the algorithm and the other containing 3788 text excerpts. The counts of entities in both of the corpora are given in Table V. The results reported in the next section are based only on the second part of the corpus.

The tuning corpus was used solely for the recognition of the Cyc types that were present in the corpus and for their mapping to the corresponding name types in ONEM. The mapping is summarized in Table VI. The mapping covers only the most general Cyc types. If a more specific Cyc type was assigned to the entity, the appropriate type was selected using the generalization relation (in Cyc called `genls`). Since some of the specific types generalize to more than one of the general types (e.g. a *Country* generalizes both to a *GeopoliticalEntity* and an *Organization*) the first matching type was selected (according to the priority presented in Table VI).

TABLE VI
THE MAPPING BETWEEN ONEM TYPES AND CYC TYPES.

Priority	ONEM type	Cyc type
1	persName	Person HumanGivenName HumanSurname Saint God FictionalCharacter PersonTypeByEthnicity PersonTypeByOccupation
2	placeName	GeopoliticalEntity PopulatedPlace
3	geogName	GeographicalPlace Territory AstronomicalBody Place
4	orgName	Organization

We should noted however that the precise Cyc type was not discarded in that procedure. The mapping was provided in order to compare the results of the algorithm with the coarse-grained annotation from the ONEM corpus. In the applications of the NER algorithm the Cyc types are easily accessible since they are attached to the Wikipedia articles, indentified by the disambiguation algorithm.

VI. RESULTS

The results of the evaluation are given in Table VII. The system was evaluated only against the whole NE units. So if the system recognized a geographical name inside a name of an organization, the result was counted as invalid, even though there was an annotation that captured the geographical name as a subordinate of the organization name. It is because we believed that the non-compositionality of NEs is their primary characteristic and usually only the whole lexical unit should be recognized and annotated. The internal lexical structure of a given NE has only anecdotal value for information extraction. What is more, if more date (such as the location of a given place or an architectural structure) regarding the entity is needed, it can be found in the DBpedia knowledge base.

The table contains the results for exact matches as well as partial matches. An *exact match* is counted as a match which is exactly the same as in the annotation, so any characters such as quotation marks have to be present in the system provided annotation, even though they do not carry much information. A *partial match* is counted for a system-provided annotation which is completely covered by the reference annotation. So this excludes matches that only partly overlap.

The direct comparison with the performance of the other NER systems shows that the primary deficiency of the algorithm is low recall. Although such comparisons should be made with care, since none of the systems described in section II was tested against ONEM corpus.

Overall recall of 41.8% for the exact matches and 48.6% for the partial matches is below expectations for a NER system. It should be noted however that the coverage of the classification of the entities was not complete (only 62.5%),

TABLE VII
THE RESULTS OF THE EVALUATION OF THE KNOWLEDGE-BASED NER ALGORITHM. P - PRECISION, R - RECALL, F_1 - $2 * P * R/(P + R)$

	Exact match			Partial match		
Entity type	**P** [%]	**R** [%]	$\mathbf{F_1}$ [%]	**P** [%]	**R** [%]	$\mathbf{F_1}$ [%]
persName	**95.2**	34.6	50.8	**93.7**	46.6	62.2
orgName	82.7	35.8	50.0	73.3	42.5	53.8
geogName	78.2	34.4	47.8	68.3	37.7	48.6
placeName	91.8	**57.0**	**70.3**	91.5	**58.3**	**71.2**
overall	**90.0**	**41.8**	**57.1**	**86.3**	**48.6**	**62.2**

so we could expect better results if all Wikipedia entities had a type assigned. What is more – the system tries not only to detect the type of the entity, but also to disambiguate it against the Wikipedia articles. As a result, only the entities that have a corresponding entry in Wikipedia might be recognized.

Regarding precision, the algorithm achieves similar results as the other described systems and in some cases (names of people and places) even outperforms the already known methods. Still the performance for organization and geographical names is below our expectations. A manual inspection of the results showed that the low precision might be at least partly a result of the organization – place name distinction used by the annotators.

It is reported [30] e.g. that a name of a country might be treated both as a name of a place and an organization, depending on the context. This causes serious problems for the algorithm, since each entity has only one type assigned. But this seems to cause problems not only for the algorithm, but also for the annotators. E.g. in the sentence ***Amerykańska*** *prasa twierdzi, że mimo oficjalnego poparcia kampanii* ***USA*** *przez rząd* ***Pakistanu***, ... (Eng. ***American*** *press says that despite the official support of the* ***US*** *campaign by the government of* ***Pakistan***, ...), *American* and *Pakistan* are marked as names of places, while *US* is marked as a name of an organization. Such annotation does not seem to be coherent. Suffice it to say that the creators of ACE NER tasks [37] introduced the geopolitical entity type to resolve this kind of problem.

VII. CONCLUSIONS

We presented a knowledge-based algorithm for named entity recognition in Polish texts. It was tested on the one-million manually annotated sub-corpus of the National Corpus of Polish. The results obtained by the algorithm show that this method, although achieving reasonably good precision, has low recall. Definitely a combination of the presented method with grammar-based method would give much better results, since the coverage of this algorithm is limited by the contents found in Wikipedia. What is more, the classification of Wikipedia entities is not complete and further worsens the recall.

On the other hand it should be stressed that the algorithm does not require any manual work, neither in the construction of multi-word grammars, nor in tagging of large amounts of textual data. What is more, the classification scheme based

on the Cyc ontology allows for a direct consumption of the results of the algorithm by intelligent systems. This feature of the algorithm together with the moderately good results show that further research in this direction should be pursued.

REFERENCES

[1] S. Brin, "Extracting Patterns and Relations from the World Wide Web," in *The World Wide Web and Databases*. Springer, 1999, pp. 172–183.

[2] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*. ACM, 2000, pp. 85–94.

[3] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open Information Extraction from the Web," in *IN IJCAI*, 2007, pp. 2670–2676.

[4] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[5] P. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia Spotlight: shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 2011, pp. 1–8.

[6] M. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum, "Aida: An online tool for accurate disambiguation of named entities in text and tables," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, 2011.

[7] P. Exner and P. Nugues, "Entity Extraction: From Unstructured Text to DBpedia RDF Triples," in *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, G. Rizzo, P. Mendes, E. Charton, S. Hellmann, and A. Kalyanpur, Eds., 2012, pp. 58–69.

[8] A. Mykowiecka, A. Kupść, and M. Marciniak, "Rule-based medical content extraction and classification," *Intelligent Information Processing and Web Mining*, pp. 237–245, 2005.

[9] M. Marciniak and A. Mykowiecka, "Automatic processing of diabetic patients' hospital documentation," in *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. Association for Computational Linguistics, 2007, pp. 35–42.

[10] M. Piasecki, S. Szpakowicz, and B. Broda, *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2009.

[11] W. Jaworski, "Ontology-based content extraction from polish biobibliographical lexicon," in *Recent Advances in Intelligent Information Systems*. EXIT, 2009, pp. 27–40.

[12] A. Pohl, "Classifying the Wikipedia Articles into the OpenCyc Taxonomy," in *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference*, G. Rizzo, P. Mendes, E. Charton, S. Hellmann, and A. Kalyanpur, Eds., 2012, pp. 5–16.

[13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," *The Semantic Web*, pp. 722–735, 2007.

[14] J. Piskorski, "Automatic named-entity recognition for Polish," in *Proceedings of the International International Workshop on Intelligent Media Technology for Communicative Intelligence, Warsaw, Poland*, 2004.

[15] J. Piskorski, P. Homola, M. Marciniak, A. Mykowiecka, A. Przepiórkowski, and M. Woliński, "Information extraction for Polish using the SProUT platform," *Intelligent Information Processing and Web Mining, Advances in Soft Computing*, pp. 227–236, 2004.

[16] J. Piskorski, "Named-entity recognition for Polish with SProUT," in *Intelligent Media Technology for Communicative Intelligence*. Springer, 2005, pp. 122–133.

[17] W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu, "Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications," *Künstliche Intelligenz*, vol. 1, pp. 17–23, 2004.

[18] M. Woliński, "Morfeusz—a practical tool for the morphological analysis of Polish," *Intelligent information processing and web mining*, pp. 511–520, 2006.

[19] J. Piskorski, "Extraction of Polish Named-Entities," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC*, 2004, pp. 313–316.

[20] W. Abramowicz, A. Filipowska, J. Piskorski, K. Węcel, and K. Wieloch, "Linguistic Suite for Polish Cadastral System," in *Proceedings of the LREC*, vol. 6, 2006, pp. 53–58.

[21] F. Graliński, K. Jassem, and M. Marcińczuk, "An environment for named entity recognition and translation," in *Proceedings of the 13th Annual Conference of the European Association for Machine Translation, Barcelona, Spain*, 2009, pp. 88–95.

[22] F. Graliński, K. Jassem, M. Marcińczuk, and P. Wawrzyniak, "Named Entity Recognition in Machine Anonymization," *Recent Advances in Intelligent Information Systems*, pp. 247–260, 2009.

[23] M. Walas and K. Jassem, "Named entity recognition in a Polish question answering system," *Intelligent Information Systems*, pp. 181–192, 2010.

[24] A. Buczyński and A. Przepiórkowski, "Spejd: A shallow processing and morphological disambiguation tool," in *Human Language Technology. Challenges of the Information Society*. Springer, 2009, pp. 131–141.

[25] M. Marcińczuk and M. Piasecki, "Named Entity Recognition in the Domain of Polish Stock Exchange Reports," *Intelligent Information Systems, Siedlce*, pp. 127–140, 2010.

[26] I. Witten and T. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *Information Theory, IEEE Transactions on*, vol. 37, no. 4, pp. 1085–1094, 1991.

[27] B. Carpenter and B. Baldwin, *Text Analysis with LingPipe 4*. LingPipe Inc, 2011.

[28] M. Marcińczuk, M. Stanek, M. Piasecki, and A. Musiał, "Rich Set of Features for Proper Name Recognition in Polish Texts," in *Security and Intelligent Information Systems*. Springer, 2012, pp. 332–344.

[29] J. Lafferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." Morgan Kaufmann, 2001, pp. 282–289.

[30] A. Savary, J. Waszczuk, and A. Przepiórkowski, "Towards the Annotation of Named Entities in the National Corpus of Polish," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, 2010.

[31] A. Pohl, "Improving the Wikipedia Miner Word Sense Disambiguation Algorithm," in *Proceedings of Federated Conference on Computer Science and Information Systems 2012*. IEEE, to appear.

[32] D. Milne and I. Witten, "Learning to link with Wikipedia," in *Proceeding of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 509–518.

[33] R. Cilibrasi and P. Vitanyi, "The Google similarity distance," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, no. 3, pp. 370–383, 2007.

[34] J. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[35] M. Moens, *Information extraction: algorithms and prospects in a retrieval context*. Springer-Verlag New York Inc, 2006, vol. 21.

[36] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proceedings of COLING*, vol. 96, 1996, pp. 466–471.

[37] NIST, "Automatic Content Extraction 2008 Evaluation Plan (ACE08)," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/ace/-2008/doc/ace08-evalplan.v1.2d.pdf

[38] D. B. Lenat, "CYC: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[39] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[40] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia-A crystallization point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, 2009.

[41] C. Zirn, V. Nastase, and M. Strube, "Distinguishing between instances and classes in the wikipedia taxonomy," *The Semantic Web: Research and Applications*, pp. 376–387, 2008.

[42] A. Przepiórkowski, M. Bańko, R. L. Górski, and B. Lewandowska-Tomaszczyk, *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, 2012.

[43] A. Savary and J. Piskorski, "Lexicons and grammars for named entity annotation in the National corpus of Polish," *Intelligent Information Systems, Siedlce, Poland*, pp. 141–154, 2010.