Extraction of Polish Named-Entities

Jakub Piskorski

German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
piskorsk@dfki.de

Abstract

In this paper, we present some attempts towards constructing a named-entity recognition system for Polish on top of SProUT, a novel multi-lingual NLP platform and by deploying standard machine learning techniques.

1. Introduction

Named-entities (NE) constitute significant part of natural language texts and are widely exploited in various NLP applications. Although considerable work on named-entity recognition (NER) for few major languages exists, research on this topic in the context of Slavonic languages[1] has been almost neglected. Some NER systems for Bulgarian and Russian constructed by adapting the famous information extraction platform GATE (Cunningham et al., 2002) were presented at a recent IESL workshop (Cunningham et al. 2003). In this paper, we present some attempts towards constructing a NER system for Polish, built on top of SProUT[2] (Becker et al., 2002; Drożdżyński et al., 2004), a novel general purpose multi-lingual NLP platform and by deploying standard machine learning techniques. Polish is a West Slavonic language and, analogously to other languages in the group, it exhibits a highly inflectional character (e.g., nouns and adjectives decline in seven cases) and has a relatively free word-order (Świdziński and Saloni, 1998). Due to these specifics and general lack of linguistic resources for Polish, construction of a NER system for Polish is an intriguing task.

2. SProUT

2.1 System overview

Analogously to the widely-known GATE system, SProUT is equipped with a set of reusable Unicode-capable online processing components for basic linguistic operations, including tokenization, sentence splitting, morphological analysis, gazetteer lookup, and reference matching. Since typed feature structures (TFS) are used as a uniform data structure for representing the input and output by each of these processing resources, they can be flexibly combined into a pipeline that produces several streams of linguistically annotated structures, which serve as an input for the shallow grammar interpreter, applied at the next stage.

The grammar formalism in SProUT is a blend of very efficient finite-state techniques and unification-based formalisms which are known to guarantee transparency and expressiveness. To be more precise, a grammar in SProUT consists of pattern/action rules, where the LHS of a rule is a regular expression over TFSs with functional operators and coreferences, representing the recognition pattern, and the RHS of a rule is a TFS specification of the output structure. Coreferences express structural identity, create dynamic value assignments, and serve as means of information transport into the output descriptions. Functional operators[3] provide a gateway to the outside world, and they are primarily utilized for forming the output of a rule (e.g., concatenation of strings) and for introducing complex constraints in the rules (they can act as predicates that produce Boolean values). Furthermore, grammar rules can be recursively embedded, which in fact provides grammarians with a context-free formalism. The following rule for the recognition of prepositional phrases gives an idea of the syntax of the grammar formalism:

```
pp :>  morph &  [  POS Prep, SURFACE #prep,
                   INFL [CASE #c ] ]
       (morph & [  POS Adjective,
                   INFL [  CASE #c,
                           NUMBER #n,
                           GENDER #g ] ] ) *
       (morph & [  POS Noun, SURFACE #noun1,
                   INFL [  CASE #c,
                           NUMBER #n,
                           GENDER #g ] ] )
       (morph & [  POS Noun, SURFACE #noun2,
                   INFL [  CASE #c,
                           NUMBER #n,
                           GENDER #g ] ] ) ?
-> phrase & [  CAT pp, PREP #prep,
               AGR agr & [   CASE #c,
                             NUMBER #n,
                             GENDER #g]
               CORE_NP #core_np]],
where #core_np=Append(#noun1," ",#noun2).
```

The first TFS matches a preposition. It is followed by zero or more adjectives. Finally, one or two noun items are consumed. The variables `#c`, `#n`, `#g` establish coreferences expressing the agreement in case, number, and gender for all matched items (except for the initial preposition item which solely agrees in case with the other items). The RHS of the rule triggers the creation of a TFS

[1] Slavonic languages constitute a large group of the Indoeuropean language family and are further split into West, East and South Slavonic subgroups.

[2] SProUT – Shallow Processing with Typed Feature Structures and Unification)

[3] SProUT comes with a set of circa 20 predefined functional operators.

of type phrase, where the surface form of the matched preposition is transported into the corresponding slot via the variable `#prep`. A value for the attribute `CORE_NP` is created through a concatenation of the matched nouns (variables `#noun1` and `#noun2`). This is realized via a call to a functional operator called `Append`. Grammars consisting of such rules are compiled into extended finite-state networks with rich label descriptions (TFSs). Since fully specified TFSs usually do not allow for minimization and efficient processing of such networks, a handful of methods going beyond standard finite-state techniques have been deployed to remedy this problem (Krieger & Piskorski, 2004).

SProUTs' shallow grammar interpreter comes with some additional functionalities, including rule prioritization, output merging mechanism (Busemann and Krieger, 2004), and reference matching tool, which can be activated on demand. The latter tool take as input the output structures generated by the interpreter, potentially containing user-defined information on variants of the recognized entities for certain NE classes, and performs an additional pass through the text, in order to discover mentions of previously recognized entities[4]. The variant specification is done explicitly by defining additional attributes, e.g., `VARIANT`, on the RHS of grammar rules, which contain a list of all variant forms (e.g., obtained by concatenating some of the constituents of the full name).

2.2 Adopting SProUT to the Processing of Polish

Since SProUT provides some linguistic resources for the processing components for Germanic and Roman languages, we could exploit these resources in the process of fine-tuning SProUT to processing Polish w.r.t. NER. Initially, the provided tokenizer resources could be easily adopted by extending the character set with some specific Polish characters and adjusting some of circa 30 predefined token classes. Subsequently, *Morfeusz*, a morphological analyzer for Polish which uses a rich tagset based on both morphological and syntactic criteria (Przepiórkowski and Woliński, 2003) has been integrated. It is capable of recognizing circa 1,800,000 Polish contemporary word forms. Some work has been accomplished in order to infer additional implicit information (e.g., tense) hidden in the tags generated by *Morfeusz*.

Extensive gazetteers constitute an essential resource in a rule-based NER system. Therefore, some work has focused on acquisition of such resources. Apart from adapting a subset of circa 50,000 gazetteer entries for Germanic languages (mainly first names, locations, organizations), which appear in Polish texts, we acquired additional language-specific resources from various Web sources. Further, we manually and semi-automatically produced all orthographic and morphological variants for the subset of the acquired gazetteer resources (e.g., we implemented a brute-force algorithm which generates full declension of first names). Since SProUT allows for associating gazetteer entries with a list of arbitrary attribute-value pairs, the created entries were additionally enriched with semantic tags and some basic morphological information, e.g., for the word form '*Argentyny*' (genitive form for *Argentyna*) the following entry has been created:

```
Argentyny | GTYPE:gaz_country | G_CASE:gen
          | CONCEPT:Argentyna
```

The created language-specific resources are summarized in the table in figure 1.

TYPE	AMOUNT
large companies	1211
federal government organizations	164
higher schools	68
Cities	2482
Countries	1727
Geographical regions	420
first names	1804

Figure 1. Language-specific gazetteer entries

Since producing all variant forms is a laborious job, and because the process of creating new names is very productive, a further way of establishing a better interplay between the gazetteer and the morphology module was achieved through an extension of the gazetteer processing module so as to accept lemmatized tokens as input. This solution is beneficial in case of single-word NEs covered by the morphological component. However, since declension of multi-word NEs in Polish is very complex, and frequently some of the words they comprise of are unknown, the next technique for boosting the gazetteer exploits the grammar formalism itself by introducing SProUT rules for the extraction, lemmatization and generation of diverse variants of the same NE from the available text corpora.

The essential information for creation of variants comes from the correct lemmatization of proper names, which is a challenging task with regard to Polish. Let us briefly address lemmatization of person names. In general, both first name and surname of a person undergo declension. Lemmatization of first names is handled by the gazetteer which provides the main forms (at least for the frequently used Polish first names), whereas lemmatization of surnames is a more complex task. Firstly, we have implemented a range of rough sure-fire rules, e.g., rules that convert suffixes like {*-skiego*, *-skim*, *-skiemu*} into the main-form suffix *-ski*, which covers a significant part of the surnames. Secondly, for surnames which do not match any of the sure-fire rules, slightly more sophisticated rules are applied that take into account several factors including: the part-of-speech of the surname (e.g., noun, adjective, or unknown), gender of the surname (in case it is provided by the morphology), and even contextual information, such as the gender of the preceding first name (possibly provided by the gazetteer). For instance, if the gender of the first name is feminine (e.g., *Stanisława*), and the surname is a masculine noun (e.g., *Grzyb 'mushroom'*), then the surname does not undergo declension (e.g. main form: *Stanisława Grzyb* vs. accusative form: *Stanisławę Grzyb*). If in the same context the first name is masculine (e.g., *Stanisław*), then the surname would undergo declension (e.g. nom: *Stanisław Grzyb* vs. acc:*Stanisława Grzyba*). On the other hand, if the surname is an adjective it always declines. No later

[4] The size of the contextual frame (e.g., a paragraph) for tracking entity mentions is parametrizable

than now, can we witness how useful the inflectional information for the first names provided by the gazetteer is. A maze of similar lemmatization rules was derived from the bizarre proper name declension paradigm presented in (Grzenia, 1998). Nevertheless, in sentences like, e.g., *Powiadomiono wczoraj wieczorem G. Busha o ataku* '[They have informed] [yesterday] [evening] [G. Bush] [about] [the attack]', correctly inferring the main form of the surname *Busha* would at least involve a subcategorization frame for the verb *powiadomić* 'to inform' (it takes accusative NP as argument). Since subcategorization lexica are not provided, such cases are not covered at the moment. The lemmatization component is integrated in SProUT simply via a functional operator. Hence, any extensions or adaptations to processing other languages w.r.t. lemmatization are straightforward. Lemmatization of organization names is done implicitly in the grammar rules as we will see in the next section.

2.3 NE-grammar for Polish

Within the highly declarative grammar paradigm of SProUT, we have developed grammars for recognition of MUC-like NE types (Chinchor and Robinson, 1998), including: persons, locations, organizations, etc.

In the first step, to avoiding starting from scratch, we recycled some of the existing NE-grammars for German and English via simply substituting crucial keywords with their Polish counterparts. As NEs mainly consists of nouns and adjectives, major changes focused on replacing the occurrences of the attribute `SURFACE` with the attribute `STEM` (main form) and specifying some additional constraints to control the inflection. Contrary to German and English, the role of morphological analysis in the process of NER for Polish is essential, as the following rule illustrates

```
org :> (morph & [  SURFACE #key,
                   STEM "urząd" & #stem,
                   INFL #infl]) |
       (morph & [  SURFACE #key,
                   STEM "komitet" & #stem,
                   INFL #infl])
       @seek(pl_np_gen) & [SURFACE #rest]
   -> enamex & [ SURFACE #surf,
                 TYPE organization,
                 SUBTYPE #stem,
                 CONCEPT #conc,
                 INFL #infl],
where  #surf=ConcWithBlanks(#key,#rest),
       #conc=ConcWithBlanks(#stem,#rest).
```

This rule identifies diverse morphological forms of keywords, such as *urząd* 'office', or *komitet* 'comitee' followed by a genitive NP (realized by the seek statement). The RHS of the rule generates a named-entity object, where the functional operator `ConcWithBlanks` simply concatenates all its arguments and inserts blanks between them. For instance, the above rule matches all variants of the phrase *Urząd Ubezpieczeń Zdrowotnych* (Health Insurance Office). It is important to notice that in this particular type of constructions, only the keyword undergoes declension (*urząd*), whereas the rest remains unchanged. Hence, the main form is reconstructed via concatenating the stem of the keyword and the surface forms of the remaining constituents (`CONCEPT` attribute). Actually, as soon as we had addressed the issue of lemmatization, the major part of the rules created so far for the particular NE classes had to be broken down into several rules, where each new rule covers different lemmatization phenomenon. Due to the fact that organization names are frequently built up of noun phrases, their lemmatization is complex and relies on proper recognition of their internal structure. The following fragment of the lemmatization schema for organization names visualizes the idea.

`[Adj] [N-key] NP-gen`
(e.g., [*Naczelnej*] [*Izby*] *Kontrolii*)

`[Adj] [N-key] [Adj] NP-gen`
(e.g., [*Okręgowy*] [*Komitet*] [*Organizacyjny*] *Budowy Autostrady*)

`N-key` represents nominal keywords such as *ministerstwo* (ministry). The constituents which undergo declension are bracketed. For each rule in such schema a corresponding NER rule has been defined. However, the situation can get even more complicated, since NEs may have potentially more than one internal syntactical structure, which is typical for Polish, since adjectives may either stand before a noun, or they can follow a noun. For instance, the phrase *Biblioteki Głównej Wyższej Szkoły Handlowej* has at least three possible internal structures:

(1) [*Biblioteki Głównej*] [*Wyższej Szkoły Handlowej*]
`[of the main library] [of the Higher School of Economics]',
(*2) [*Biblioteki Głównej Wyższej*] [*Szkoły Handlowej*]
`[of the main higher library] [of the School of Economics]', and
(*3) [*Biblioteki*] [*Głównej Wyższej Szkoły Handlowej*] `[of the library] [of the Main Higher School of Economics]'.

This poses a serious complicacy in the context of lemmatization, not to mention singular-plural ambiguity of the word *biblioteki* (singular-gen vs. plural-nom-acc), etc. In order to tackle this problem, some experiments proved that an introduction of multiple keywords (e.g., '*Biblioteka Główna*' in the example above) would potentially reduce the number of ambiguities.

Last but not least, there exists another issue which complicates lemmatization of proper names in SProUT. We might easily identify the structure of organization names such as *Komisji Europejskiej Praw Człowieka* (of the European Commission for Human Rights), but the part which undergoes declension, viz. *Komisji Europejskiej* (of the European Commission) can not be simply lemmatized via a concatenation of the main forms of these two words. This is because *Morfeusz* returns the nominal masculine form as the main form for an adjective, which generally differs in the ending from the corresponding feminine form (masc: *Europejski* vs. fem: *Europejska*), whereas the word *Komisja* is a feminine noun. Once again, functional operators were utilized to find a rough workaround and minimize the problem.

Ultimately, somewhat 'more relaxed' rules have been introduced in order to capture entities which could not have been captured by the ones based on morphological features and ones which perform lemmatization. For example, such rules cover sequences of capitalized words and some keywords. For the rule presented at the beginning of section 2.3 (and for similar rules), a relaxed variant has been introduced, where the call to the sub-grammar for genitive NPs was replaced with a call to a rule which maps a sequence of capitalized words and

conjunctions. Consequently, SProUTs' mechanism for rule prioritization has been deployed in order to give higher preference to rules capable of performing lemmatization, i.e., to filter the matches found by the interpreter and rules which potentially instantiate higher number of slots in the output structures. The current grammar consists of 143 rules.

2.4 NE-grammar evaluation

A corpus consisting of 100 financial news articles from an online version of leading Polish newspaper has been selected for analysis and evaluation purposes. The precision-recall metrics for time, money and percentage expressions are 81.3%-85.9%, 97.8-93.8%, and 100-100%, respectively. Somewhat worse results were obtained for persons (90.6-85.3%), locations (88-43.4%), and organizations (87.9-56.6%) due to the problems outlined in the previous sections. 79.6% of the detected NEs were lemmatized correctly. The peculiarities of Polish pinpointed in this article reveal the indispensability of integrating additional nice-to-have components including lemmatizer for unknown multi-words (Erjavec and Džeroski, 2003), valence dictionary (Przepiórkowski, 2004), morphosyntactic tagger (Dębowski, 2004), and morphological generation module, in order to gain recall and improve the overall performance of the presented grammar-based approach.

3. ML-based NE Detection

The second approach deploys standard machine learning techniques (C4.5) for generating a classifier in form of a decision tree used for determining whether a token is a part of a NE. The scope was limited to recognition of organizations, locations and person names. We combine several sources of statistical evidence of the tokens in the training corpus. In particular, we use features such as token type (e.g., number, capitalization, punctuation, bracket), and a feature which corresponds to the observation that an absence in the dictionary is a weak evidence of being a name. Additionally, certain part-of-speech tags were used as features (e.g., adjective, noun, verb). In order to increase the number of effective rules that are learned, we collapsed adjacent syntactically similar tokens into a single token. For learning the features which strongly correlate with names we utilized the same corpus selected for the evaluation of the rule-based approach. The preliminary results are promising and reveal that combining various information sources results in better accuracy of the classifier (76,2 %). The conversion of the training corpus into a set of features was realized via the application of a pipeline of SProUTs' processing resources. Further contextual sources of information for improving the discrimination power of the classifier are under investigation (in particular various sizes of the contextual window).

4. Conclusions and Acknowledgements

We have presented a preliminary attempt towards constructing a NER system for Polish via fine-tuning SProUT, a flexible multi-lingual NLP platform, by introducing some languuage-specific components which could be easily integrated via functional operators, and by deploying standard machine learning techniques. Although the recall values are still far away from the state-of-the-art results obtained for the more studied languages, the initial evaluation results are very promising. Proximate work will center around integration of further language-specific resources and components in order to better tackle the lemmatization task.

I am greatly indebted to Witold Drożdżyński, Anna Drożdżyńska and Marcin Rzepa for their contribution. The work reported here was supported by the EU-funded project MEMPHIS under grant no. IST-2000-25045 and by additional non-financed personal effort of the author and the persons mentioned above.

5. References

M. Becker, W. Drożdżyński, H-U. Krieger, J. Piskorski, U. Schäfer, and F. Xu. 2002. *SProUT - Shallow Processing with Typed Feature Structures and Unification*. In Proceedings of ICON 2002, Mumbai, India.

S. Busemann, H.-U. Krieger. 2004. *Resources and Techniques for Multilingual Information Extraction*. In Proceedings of LREC 2004, Lissabon, Portugal.

N. Chinchor and P. Robinson. 1998. *MUC-7 Named Entity Task Definition (version 3.5)*. In Proceedings of the MUC-7, Fairfax, Virginia, USA.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. 2002. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. In Proceedings of the ACL'02, Philadelphia, USA.

H. Cunningham, E. Paskaleva, K. Bontcheva, G. Angelova. 2003. Proceedings of the Workshop on *Information Extraction for Slavonic and Other Centrtal and Eastern European Languages*, Borovets, Bulgaria.

Ł. Dębowski. 2004. Trigram morphosyntactic tagger for Polish. In Proceedings of IIS 2004, Zakopane, Poland.

W. Drożdżyński, H-U. Krieger, J. Piskorski, U. Schäfer, F. Xu. 2004. *Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications*. In German AI Journal *KI-Zeitschrift*, Vol. 01/04, Gesellschaft für Informatik e.V.

T. Erjavec and S. Džeroski. 2003. *Lemmatising Unknown Words in Highly Inflective Languages*. In Proceedings of the IESL 2003, Borovets, Bulgaria.

J. Grzenia. 1998. *Słownik nazw własnych - ortografia, wymowa, słowotwórstwo i odmiana*. Publisher: PWN, Seria: Słowniki Języka Polskiego, ISBN: 83-01-12500-4.

H-U. Krieger, J. Piskorski. 2004. *Speed-up methods for complex annotated finite-state grammars*. DFKI Report.

A. Przepiórkowski. 2004. Towards the design of a Syntactico-Semantic Lexicon for Polish. In Proceedings of IIS 2004, Zakopane, Poland.

A. Przepiórkowski, and M. Woliński. *A flexemic tagset for Polish*. In Proceedings of Morphological Processing of Slavic Languages, EACL-2003, Budapest, Hungary.

M. Świdziński and Z. Saloni. 1998. *Składnia współczesnego języka polskiego*. Publisher: PWN, ISBN: 83-01-12712-0.