Automatic Named-Entity Recognition for Polish

Jakub Piskorski

DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
piskorsk@dfki.de

Abstract. In this paper, we present a rule-based named-entity recognition system for Polish built on top of SProUT, a novel multi-lingual NLP platform. We pinpoint the encountered difficulties and present some evaluation results.

1 Introduction

Named entities (NE) constitute a significant part of natural language texts and their proper recognition is crucial for intelligent content extraction systems. Although considerable work on named-entity recognition (NER) for a few major languages exists, research on this topic in the context of Slavonic languages has been almost neglected. Some NER systems for Bulgarian and Russian have been constructed by adapting the famous IE platform GATE, and were presented at a recent IESL workshop [1]. This paper presents a NER system for Polish, built on top of SProUT [2], a novel general purpose multi-lingual NLP platform[1]. Polish, analogously to other languages in the group, exhibits a highly inflectional character (e.g., nouns and adjectives decline in seven cases) and has a relatively-free word order. 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

SProUT is equipped with a set of reusable online processing components for basic linguistic operations. They can be 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 guarantee expressiveness and transparency. 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 a means of information transport into the

[1] The presented work has been funded by the German BMBF, under grant no. 524-40001-01-IMD01A (project SMARTWEB), and by non-financed personal effort of the author.

output descriptions. Functional operators provide a gateway to the outside world, and they are utilized for forming the output of a rule and for introducing complex constraints in the rules. Grammar rules can be recursively embedded, which in fact provides grammarians with a context-free formalism. The following rule for the recognition of committee or agency names gives an idea of the syntax of the formalism [2]:

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

Firstly, a word with the base form *agencja* (office) or *komitet* (committee) is matched. It is followed by a genitive NP (`@seek` statement – a call to a subgrammar). The introduction of the variables `#key`, `#stem`, `#rest` allows for transporting the corresponding values to the RHS of the rule, which creates a TFS of type `enamex`. The value of the slots `SURFACE` and `CONCEPT` (base form) are produced via an application of the functional operator `Conc` which concatenates its arguments. Note that, since only the keyword undergoes declension, the base form is reconstructed via concatenation of the stem of the keyword and the surface form of the genitive NP.

3 Adopting SProUT to the Processing of Polish

SProUT provides linguistic resources for the processing components for Germanic languages, which could be exploited for adapting SProUT to processing Polish texts w.r.t. NER (e.g., ca. 50,000 gazetteer entries). Subsequently, *Morfeusz*, a morphological analyzer for Polish [3] has been integrated (ca. 1,800,000 words).

Since extensive gazetteers constitute an essential resource in a rule-based NER system, some work focused on the acquisition of additional language-specific resources from various Web sources (ca. 7000 entries). Furthermore, we semi-automatically produced all orthographic and morphological variants for the subset of the acquired resources (e.g., a brute-force algorithm for generating full declension of first names and manual validation). The created entries were enriched with semantic and some basic morphological tags (e.g., CONCEPT). 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 is beneficial in case of single-word NEs covered by the *Morfeusz*, but becomes less important in the context of multi-word NEs due to the complex declension paradigm, and the fact that frequently some of the words they comprise of are unknown. Finally, we boosted the gazetteer by exploiting SProUT itself for the 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. Let us briefly address lemmatization of person names in Polish. In general, both first name and surname undergo declension. Lemmatization of first names is handled by the gazetteer which provides the base forms, whereas lemmatization of surnames is a more complex task. Firstly, we have implemented a bag of rough sure-fire rules, e.g., rules that convert suffixes like {*-skiego*, *-skim*} into the main-form suffix *-ski*, which covers a significant part of the surnames. Secondly, for surnames which do not match any of the such rules, more sophisticated rules are ap-

plied that take into account several factors, including the part-of-speech/gender of the surname, and contextual information, such as the gender of the first name. E.g., if the first name is feminine and the surname is a masculine noun, then the surname does not decline. In some contexts, e.g., in the phrase *Powiadomiono wczoraj G. Busha* ([They have informed] [yesterday] [G. Bush]), correctly inferring the main form of the surname *Busha* would at least involve a subcategorization frame for the verb *powiadomić* (to inform - takes NP_{acc} as argument). Therefore, such cases are not covered yet. The lemmatization component is integrated via a functional operator, and hence, any extensions w.r.t. lemmatization are straightforward. Lemmatization of organization names is done implicitly in the grammar rules as we will discuss in section 4.

4 NE-grammar for Polish

Within the declarative grammar paradigm of SProUT, we have developed grammars for recognition of MUC-like NE types [4], including persons, locations, organizations, etc. Fore each NE type, fine-grained output structures were defined (e.g., slots for person names, including given name, surname, title, position, sex, variant, etc.). As a matter of fact, our recognition task resembles more the Template Element (TE) extraction task rather than the simpler NE-boundary recognition problem.

In first step, to avoid 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 consist 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, since even recognition of simple entities such as time spans involves morphological information.

As we have already seen in section 2, the lemmatization of the NEs is partially encoded within the grammars. Due to some particularities of Polish, such as the adjective position in NPs (they can either stand before a noun, or they can follow a noun), 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. For example, organization names are frequently built up of NPs and their lemmatization relies on proper recognition of their internal structure. The following fragment of the lemmatization schema for organizations visualizes the idea.

`[A] [N-key]` NP_{gen} (e.g., [*Naczelną*] [*Izbą*] *Kontrolii*)
supreme$_{ins}$ chamber$_{ins}$ audit$_{gen}$

`[A] [N-key] [A]` NP_{gen} (e.g., [*Okręgowemu*] [*Komitetowi*] [*Organizacyjnemu*] *Budowy Autostrady*)
local$_{acc}$ committee$_{acc}$ organizing$_{acc}$ building-the-highway'$_{gen}$

`N-key` represents nominal keywords. The constituents which undergo declension are bracketed. For each rule in such schema a corresponding NER rule has been defined. However, since NEs may have potentially more than one internal syntactical structure, the task can become more challenging. For instance, the phrase *Biblioteki Głównej Wyższej Szkoły Handlowej* has three possible segmentations:

(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],
(3) [*Biblioteki*] [*Głównej Wyższej Szkoły Handlowej*]
[of the library (of the libraries)] [of the Main Higher School of Economics]

In order to tackle such complicacies, an introduction of collocations (e.g., '*Biblioteka Główna*' in the example above) reduced the number of ambiguities.

Additionally, somewhat more 'relaxed' rules have been introduced in order to capture NEs which could not have been captured by those based on morphological features and those which perform lemmatization. Such rules cover sequences of capitalized words, conjunctions, and some keywords. Consequently, SproUT's mechanism for rule prioritization has been deployed in order to give higher preference to rules capable of performing lemmatization and rules which potentially instantiate higher number of slots in the output structures. The current grammar consists of 143 rules.

In an additional pass through the text, mentions of previously recognized entities are discovered, based on the variants in the output structures returned by the grammar interpreter. Since the 'variant' slots solely include main forms (and no morphological generation functionality is available), only nominative mentions could be identified. However the major part of the mentions appeared to be nominative.

5 Evaluation and Conclusions

A corpus consisting of 100 financial news articles from an online version of *Rzeczpospolita* (a 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. Although the recall values are still far away from the state-of-the-art results obtained for the more studied languages, the initial results are promising. The peculiarities of Polish pinpointed in this article reveal the indispensability of integrating additional language-specific resources and components including lemmatizer for unknown multi-words, subcategorization dictionary, morphlogical generation, and a morphosyntactic tagger, in order to gain recall and improve the overall performance of the presented grammar-based approach. A machine learning approach is envisaged once sufficient amount of annotated data is available.

References

1. Cunningham, H., Paskaleva, E., Bontcheva, K., Angelova, G. Proceedings of the Workshop IESL, Borovets, Bulgaria, (2003).
2. Drożdżyński, W., Krieger, H.-U., Piskorski, J., Schäfer, U., Xu, F. *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., (2004).
3. Przepiórkowski, A., Woliński, M.. *A flexemic tagset for Polish*. Proceedings of Morph logical Processing of Slavic Languages, EACL-2003, Budapest, Hungary, (2003).
4. Chinchor, N., Robinson, P. *MUC-7 Named Entity Task Definition (version 3.5)*. Proceedings of the MUC-7, Fairfax, Virginia, USA, (1998).