Information Extraction for Polish Using the SProUT Platform

Jakub Piskorski[1], Peter Homola[2], Małgorzata Marciniak[3], Agnieszka Mykowiecka[3], Adam Przepiórkowski[3], and Marcin Woliński[3]

[1] DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2] UFAL Charles Univ., Malostranské náměstí 25, CZ-118 00 Praha, Czech Rep.
[3] IPI PAN, Ordona 21, 01-237 Warszawa, Poland

Abstract. The aim of this article is to present the initial results of adapting SProUT, a multi-lingual Natural Language Processing platform developed at DFKI, Germany, to the processing of Polish. The article describes some of the problems posed by the integration of *Morfeusz*, an external morphological analyzer for Polish, and various solutions to the problem of the lack of extensive gazetteers for Polish. The main sections of the article report on some initial experiments in applying this adapted system to the Information Extraction task of identifying various classes of Named Entities in financial and medical texts, perhaps the first such Information Extraction effort for Polish.

1 Introduction

The task of Information Extraction (IE) is to identify instances of particular pre-specified classes of entities, events and relationships in natural language texts, and to extract relevant arguments of the so identified events or relationships [2]. Most of the research in IE has concentrated on English and a handful of other languages (e.g., [7], [8],[1], [5]), and relatively few efforts have been undertaken for constructing and adapting IE platforms to the processing of Slavonic languages. Initial attempts at the integration of research activities on IE for Slavonic languages were presented at a recent workshop held in conjunction with the RANLP 2003 conference, [9]. In particular, [6] provides a preliminary report on adapting SProUT,[1] a novel general purpose multi-lingual IE platform [3], to the processing of Czech, Polish, and Lithuanian.

The present paper describes in detail the utilization of SProUT for the construction of an IE system for Polish, and presents initial results on the deployment of this system to IE from financial and medical texts.

2 SProUT

An Achilles heel of earlier IE systems was the fact that they were either lacking efficiency or expressiveness of the underlying grammar formalism.

[1] **S**hallow Text **Pro**cessing with **U**nification and **T**yped Feature Structures

One of the major motivations for developing SProUT is to find a reasonable trade-off between these two crucial features. The grammar formalism deployed in SProUT is an amalgamation of very efficient finite-state techniques and unification-based formalisms which are known to guarantee transparency and expressiveness. A grammar consists of pattern/action rules, where the LHS of a rule is a regular expression over typed feature structures (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 provide a stronger expressiveness since they create dynamic value assignments and serve as means of information transport into the output descriptions. The following rule for the recognition of location Prepositional Phrasess (PPs) illustrates the syntax of SProUT rules:

```
loc-pp :>
 morph & [POS Prep, SURFACE #prep, INFL[CASE_PREP #c]]
 (morph & [POS Det, INFL [CASE_DET #c, NUMBER_DET #n, GENDER_DET #g]])?
 (morph & [POS Adj, INFL [CASE_ADJ #c, NUMBER_ADJ #n, GENDER_ADJ #g]])*
 gazetteer & [TYPE general_location,  SURFACE #loc]
->
 phrase & [CAT location-pp,PREP #prep, LOCATION #loc].
```

The first TFS matches a preposition. Then, one or zero determiners are matched. Subsequently, zero or more adjectives are consumed. Finally, the last TFS matches an item in a location gazetteer. The variables `#c`, `#n`, and `#g` establish coreferences expressing the agreement in case, number, and gender for all but the last matched item. The RHS triggers the creation of a TFS of type phrase, where the matched preposition and location are transported into the corresponding slots via the variables `#prep` and `#loc`. Grammar rules can be recursively embedded, which in fact provides grammarians with a context-free formalism. Nevertheless, grammars are compiled into extended finite-state networks with rich label descriptions. For the efficient processing of such networks, a bag of methods has been developed which go beyond standard finite-state techniques [10].

Currently, the system is equipped with a set of reusable online processing components for basic operations such as tokenization, morphological analysis, gazetteer lookup, or reference matching and provides the corresponding linguistic resources for the major Germanic, Romance and Asian languages.

3 Adapting SProUT to Polish

3.1 Morphological Analysis

The morphological analyser for Polish used in the current version of SProUT, *Morfeusz*, has been developed by M. Woliński on the basis of linguistic data provided by Z. Saloni, esp., his database of Polish verbs (cf. [13]) and the

stemming rules published as [14]. The program assigns all possible morphosyntactic interpretations to each input word, where each interpretation consists of a base form (a lemma) and a tag. The analyser is still under development. Currently, the analyser is capable of recognising about 1,800,000 wordforms that occur in contemporary Polish texts (the coverage of about 95% wordforms) and interpreting them as forms of about 110,000 lemmas.

Morfeusz uses the *IPI PAN tagset* [11,12] based on a homogeneous set of morphological (how a given form inflects; e.g., nouns inflect for case, but not for gender), morphosyntactic (in which categories it agrees with other forms; e.g., Polish nouns agree in gender with adjectives and verbs) and, as secondary, syntactic criteria. Some of the parts of speech (POSs), or flexemic classes, delimited on the basis of such criteria correspond rather directly to the traditional POSs, e.g., noun, adjective, adverb, preposition, conjunction, particle; other are more fine grained, e.g., various verbal classes such as infinitive, four classes for the four participial forms (two adjectival and two adverbial), non-past verb, impersonal *-no/-to* form, imperative, l-participle (forms like *przyszli* 'came' in *przyszliśmy* 'came-we'), gerund. The traditional classes of numerals and pronouns, usually defined on the basis of their semantics, are redefined here in purely morphological and morphosyntactic terms: the flexemic class of numerals only consists of cardinal numerals, and three specific classes for nominal pronouns are distinguished: non-3rd person pronoun, 3rd person pronoun, and the pronoun *siebie*. Other, more ephemeral flexemic classes are: depreciative noun, ad-adjectival adjective, post-prepositional adjective, future *być*, agglutinative *być* (forms like *-śmy* in *przyszliśmy* 'came we'), *winien*-like verb, predicative, alien nominal/other. Grammatical categories assumed in the tagset include traditional categories such as number, case, person, degree, aspect, gender, as well as more restricted categories, some of them first proposed in works by Z. Saloni and J.S. Bień, such as accentability, post-prepositionality, acommodability, agglutination, vocability and negation.

The following examples should give the impression of the granularity and the positional nature of the tagset:

- *myszami* 'mouse': `[mysz,subst:pl:inst:f]` — noun, plural number, instrumental case, feminine, with the base form *mysz*;
- *przyszedł* 'came': `[przyjść,praet:sg:m1.m2.m3:perf]` — l-participle, singular, one of the three masculine genders (i.e., in fact, this is an abbreviation for three different tags), perfective aspect;
- *jemu* 'him': `[on,ppron3:sg:dat:m1.m2.m3.n1.n2:ter:akc:npraep]` — 3rd person pronoun, singular, dative, one of three masculine or two neuter genders, 3rd person (some redundancy with respect to the information carried by the flexemic class makes formulating the rules of agreement easier), accented (strong) form of the pronoun, non-post-prepositional.

In order for the morhosyntactic informations provided by *Morfeusz* to be usable in SProUT, they must be converted into TFSs. The values of grammat-

ical categories are mapped into the corresponding types specified in the type hierarchy for Polish. This type hierarchy is fine-tuned to the results produced by *Morfeusz* in order to take into account the morphosyntactic particularities of Polish (such as a variety of genders) and in order to properly treat the fine-grained repertoire of parts of speech and the implicit information they provide about such categories as mood or tense. The result of the conversion of the tag `[mysz,subst:pl:inst:f]` (for the form *myszami*) is given below.

$$\begin{bmatrix} \text{SURFACE} & \text{'myszami'} \\ \text{STEM} & \text{'mysz'} \\ \text{POS} & \textit{noun} \\ \text{INFL} & \begin{bmatrix} \text{GENDER} & \textit{feminine} \\ \text{CASE} & \textit{inst} \\ \text{NUMBER} & \textit{plural} \end{bmatrix} \end{bmatrix}$$

Unlike the other two East European languages integrated with SProUT, Czech and Lithuanian, Polish uses agglutination in its verbal system to express past tense and subjunctive. Such agglutinated verbal forms consist of two or more segments associated with independent tags. This information is recorded in the values of a parochial attribute for Polish, AGGL(utination), containing a list of FSs for each segment of such a word, e.g., for *przyszedłem*:

$$\begin{bmatrix} \text{SURFACE} & \text{'przyszedłem'} \\ \text{STEM} & \text{'przyjść'} \\ \text{POS} & \textit{verb} \\ \text{INFL} & \begin{bmatrix} \text{VERBAL\_FORM} & \textit{finite} \\ \text{GENDER} & \textit{masculine} \\ \text{PERSON} & \textit{first} \\ \text{NUMBER} & \textit{singular} \\ \text{TENSE} & \textit{past} \end{bmatrix} \\ \text{AGGL} & \left\langle \begin{bmatrix} \text{SURFACE} & \text{'przyszedł'} \\ \text{STEM} & \text{'przyjść'} \\ \text{INFL} & [\ldots] \end{bmatrix}, \begin{bmatrix} \text{SURFACE} & \text{'em'} \\ \text{STEM} & \text{'być'} \\ \text{INFL} & [\ldots] \end{bmatrix} \right\rangle \end{bmatrix}$$

This solution reflects the principle that every word form should be represented by exactly one FS at the morphological and surface syntactic level. On the other hand, segments forming such multi-segment words have specific (morpho)syntactic functions (e.g., the l-participle controls the valence of the verb, while the agglutinative form encodes grammatical person), so they should also be clearly separated in the resulting FS.

3.2 Gazetteer

Gazetteer lookup is usually considered to be an independent part of linguistic analysis, in which the input stream of characters or tokens is matched against a list of proper names (e.g., locations) and keywords (e.g., company designators). Some of the gazetteer resources for English and German, consisting of

about 50,000 named entities (NE), mainly first names, locations, organizations, positions and titles, could be reused for Polish, since they also appear in Polish texts. Further, we acquired from various Web sources some language specific resources, i.e., dictionary forms of about 1,200 names of large companies and federal government organizations in Poland, around 500 locations including geographical regions, and 350 frequently used Polish first names. Due to the highly inflectional nature of Polish, this basic pool of gazetteer entries had to be extended by adding, for all entries, their morphological variants, together with basic morphological information (SProUT allows for associating gazetteer entries with a list of arbitrary attribute-value pairs). For some named-entity types this task was acomplished manually (e.g., countries), whereas for others (e.g., common first names in Polish) some brute-force algorithms for generation of all variants have been implemented. Since generating all variants is a time-consuming task, and because the process of creating new names is very productive, additional ways of establishing a better connection between the gazetteer and the morphological component had to be found. Firstly, the gazetteer component was modified so as to accept morphologically annotated tokens as input. This way, it is possible to search the gazetteer for the values of the attribute STEM containing the lemmatized form of the word. This works well for single words, provided they are correctly analyzed by the morphological component. However, the declension of multi-word NEs in Polish is very complex, and many of them are composed of words which are not necessarily covered by the morphological component. This observation led us to the creation of specific rules for extracting diverse variants of the same NEs from large text corpora. The following rule illustrates the idea.

```
pl_org :>
 (morph & [SURFACE #key, STEM "bank" & #main, INFL #infl] |
  morph & [SURFACE #key, STEM "agencja" & #main, INFL #infl] |
  morph & [SURFACE #key, STEM "komisja" & #main, INFL #infl])
  @seek(pl_np_gen) & [SURFACE #rest]
->
  gazetteer_entry & [ENTRY #entry, GTYPE gaz_organization,
                     GSUBTYPE #main, CONCEPT #mainForm, INFL #infl],
  where #entry=ConcWithBlanks(#key,#rest),
        #mainForm=ConcWithBlanks(#main,#rest).
```

This rule identifies variant forms of keywords such as *agencja* 'agency' followed by a genitive NP (realized by the seek statement). The RHS of the rule generates a gazetteer entry, 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 *Agencja Restrukturyzacji i Modernizacji Rolnictwa* 'Agency for Restructuring and Modernisation of Agriculture'. Note that in this particular construction, only the keyword undergoes declension, so even if the morphological unit fails to rec-

ognize some of the constituents, we could relax the rule by replacing a call to the rule for genitive NPs with a rule which maps a sequence of capitalized words and conjunctions. This general automatic lemmatisation of unknown multi-words turned out to further boost the power of the gazetteer.

4 Information Extraction

4.1 IE from Financial Texts

The first IE task reported here focuses on the identification of typical NEs (e.g., time expressions, quantities, proper names) from financial texts. Obviously, some of the grammar fragments for German and English NEs could be straightforwardly adopted to Polish by substituting crucial keywords with their Polish counterparts. However, major changes centered around replacing the occurrences of the attribute SURFACE with the attribute STEM (main form) and specifying additional constraints to control the inflection — NEs mainly consist of nouns and adjectives, which exhibit highly inflectional character in Polish. Morphological analysis plays an essential role here, since even rules for identifying such simple entities as time spans involve morphological information. This is illustrated with the following rule for matching expressions like *Od stycznia do lutego 2003* ('from January till February 2003'), where genitive forms of month names are required.

```
pl_time_span :>
 token & [SURFACE "od"]
 @seek(pl_month) & [STEM #start, INFL [CASE_NOUN gen, NUMBER_NOUN sg]]
 token & [SURFACE "do"]
 @seek(pl_month) & [STEM #end, INFL [CASE_NOUN gen, NUMBER_NOUN sg]]
 gazetteer & [GTYPE gaz_year, CONCEPT #year]
->
 timex & [FROM [MONTH #start, YEAR #year], TO [MONTH #end,YEAR #year]].
```

Some essential information for coreference resolution comes from the correct lemmatization of proper names, which is a challenging task in Polish, esp., in case of multi-word names. We tackled this by defining rules which, depending on the type of a NE and its internal structure, specify the construction of the main form from the surface form. The following fragment of the schema for lemmatization of organization names with the corresponding examples visualize the idea. N-key represents nominal keywords such as *ministerstwo* ('ministry'). The constituents which undergo declension are bracketed.

- ORG: [Adj] [N-key] NP-gen, e.g., *Naczelnej Izby Kontroli*,
- ORG: [N-key] [Adj] NP-gen, e.g., *Komisji Europejskiej Praw Człowieka*,
- PERSON: [First-Name] [Last-Name], e.g., *Aleksandra Kwaśniewskiego*.

For each rule in such schema a corresponding NE-rule has been defined. However, the situation can get even more complicated, since NEs may potentially have more than one internal syntactic structure. 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]',

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

This poses a problem in the context of lemmatisation, not to mention singular–plural ambiguity of the word *biblioteki* (singular–genitive vs. plural–nominative–accusative). Introducing multi-word keywords in the NE-rules would potentially solve the problem (e.g., *Biblioteka Główna* in the example above).

There are still other issues which complicate lemmatization of proper names in SProUT. For instance, even if we identify a part of an organization name which undergoes declension (e.g., *Komisji Europejskiej* 'of the European Commision' in *Komisji Europejskiej Praw Czlowieka* 'of the European Commision for Human Rights'), we cannot simply lemmatize such an organization name via a concatenation of the main forms of the words which undergo declension with the rest. 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* in the example is a feminine noun.

Finally, 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 (e.g., sequences of capitalized words and keywords). Consequently, a mechanism for rule prioritisation has been deployed in order to give higher preference to rules performing lemmatisation, i.e., to filter the matches found. The whole grammar consists of 67 rules.

A small corpus consisting of 100 financial news articles from the online version of *Rzeczpospolita* — a leading Polish newspaper (`http://www.rzeczpospolita.pl`), has been selected for the analysis and evaluation purposes. It consists of about 25,000 tokens. The obtained precision–recall metrics are depicted in the following table.

TYPE	PRECISION	RECALL
time	81.3%	85.9%
percentage	100.0%	100.0%
money	97.8%	93.8%
organizations	87.9%	56.6%
locations	88.4%	43.4%
persons	90.6%	85.3%

The somewhat worse results obtained for persons, locations and organizations are due to the problems discussed throughout the paper. Further, 79.6% of the identified NEs were lemmatized correctly. The state-of-the-art results with respect to the precision and recall of NE Systems for English and a few other less inflective languages vary between 90% and 95%, which indicates that there is some space for improvement. In particular, we expect to increase recall via providing additional gazetteer resources and utilization of a component for lemmatization of unknown multi words. Accessing such module could be simply realized as a call to a dedicated functional operator with appropriate arguments specifying the type of the argument and method of lemmatization.

4.2 IE from Medical Texts

The second IE task reported here concerns the extraction of data about the size of pathological changes from a medical corpus containing descriptions of mammographical examinations collected in several Warsaw hospitals.[2] In order to obtain this goal, some domain specific resources had to be created. In particular, we defined special gazetteer entries for all types of semantic information needed, e.g., we define tumor and cyst concepts by means of the following description:

```
guz | GTYPE:g_med_change | CONCEPT:c_tumor | G_CASE:nom |
      G_NUMBER:sg | G_GENDER:msc3
torbiel | GTYPE:g_med_change | CONCEPT:c_cyst | G_CASE:nom |
          G_NUMBER:sg | G_GENDER:f
```

As already mentioned above, in case of Polish, gazetteer entries cannot be treated as pure strings, with no morphological information. As we cannot combine the information from a gazetteer and a morphological module, we have included in the gazetteer all necessary inflectional forms of keywords with appropriate morphologic information. In the IE process, we distinguish noun phrases containing one of defined keywords (i.e., a word as being a name of a type of a pathological change). If we encounter any size designation within such a nominal phrase, we output a structure consisting of the type of the change and its size.

[2] The corpus was collected by Teresa Podsiadły, IBiIB PAN, Warsaw.

Sizes of changes are represented in texts in different ways, e.g.: *guz –15 mm* 'tumor –15 mm', *...o wymiarach 2 cm × 3 cm* '... with dimensions 2 cm × 3 cm', *...o śr. ok. 1,5 cm* '...about 1.5 cm in diameter'. We represent them all by two attributes representing dimension. If only the diameter of a change is known, we put it as the value of both attributes.

In case of phrases like: *dwie spikularne zmiany o śr. ok. 15 mm i 20 mm* 'two spicular changes of diameters about 15 mm and 20 mm', where the keyword describing a change is in plural and two sizes are coordinated, we produce two descriptions of changes.

The next two phrases show that very similar texts can describe completely different information. In *drobne guzki o śr. od 5 –10 mm* 'small tumors 5 –10 mm in diameter', numbers represent the minimal and the maximal size of several tumors. In *guzek na godz. 5 –10 mm* 'small tumor at 5 –10 mm', 5 describes the place of a tumor of size 10 mm. In this instance, the information about grammatical number of a keyword is helpful to distinguish these cases.

No precise evaluation figures are currently available, but the initial results of the work reported here seem to be very promising.

5 Concluding Remarks and Future Work

SProUT, a flexible multi-lingual NLP system first applied to Germanic languages, turned out to be readily adaptable to Polish. The main problem encountered when implementing simple IE applications for financial and medical domains was the lack of an extensive gazetteer for Polish. This problem was alleviated by developing small in-house dedicated gazetteers and by the use of a bunch of rough techniques for generating additional gazetteer entries. Furthermore, we partially addressed the issue of named-entity lemmatization which is not trivial in Polish due to several specific phenomena, e.g., adjective position in nominal phrases.

It is clear that further improvements could be achieved by the integration of additional processing resources, mainly, for the lemmatisation of multiword names (e.g., by deploying a morphological generator), and by the addition of a co-reference resolution component; due to the highly inflectional character of Polish, the former of these possible improvements is particularly important. Moreover, the construction and integration of a subcategorisation lexicon is envisaged for more complex IE tasks, involving aspects of deep parsing. In our opinion, the success of the initial experiments described here justifies our intent to further pursue this line of research.

6 Acknowledgements

We are indebted to Witold Drożdżyński for his contribution to the task of adapting SProUT to Polish. The work reported in this article has been partially funded by the grant from the German Ministry for Education, Science,

Research, and Technology (BMBF) to the DFKI project COLLATE (no. 01 IN A0), and by the EU project MEMPHIS under grant no. IST-2000-25045 and by additional non-financed personal effort of the authors.

References

1. Aone, C. and Ramos-Santacruz, M. (2000). RESS: A large-scale relation and event extraction system. *In Proceedings of ANLP 2000, Seattle, USA.*
2. Appelt, D. and Israel, D. (1999). An introduction to information extraction technology. *A Tutorial prepared for IJCAI Conference.*
3. Becker, M., Drożdżyński, W., Krieger, H., Piskorski, J., Schaefer, U., and Xu, F. (2002). SProUT — Shallow Processing with Typed Feature Structures and Unification. *In Proceedings of ICON 2002, Mumbai, India.*
4. Bering, C., Drożdżyński, W., Erbach, G., Guasch, C., Homola, P., Lehmann, S., Li, H., Krieger, H.-U., Piskorski, J., Schaefer, U., Shimada, A., Siegel, M., Xu, F., and Ziegler-Eisele, D. (2003). Corpora and evaluation tools for multilingual named entity grammar development. *Proceedings of the International Workshop: Multilingual Corpora — Linguistic Requirements and Technical Perspectives, Lancaster, UK.*
5. Ciravegna, F., Lavelli, A., and Satta., G. (2000). Bringing information extraction out of the labs: The Pinocchio environment. *In Proceedings of ECAI 2000, Berlin, Germany.*
6. Drożdżyński, W., Homola, P., Piskorski, J., and Zinkevičius, V. (2003). Adopting SProUT to processing Baltic and Slavonic languages. In *Proceedings of the IESL Workshop in conjunction with the RANLP2003 Conference, Bulgaria.*
7. Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M. (1997). FASTUS — A cascaded finite-state transducer for extracting information from natural language text. *In Finite-State Language Processing, E. Roche and Y. Schabes, MIT Press, Cambridge, MA.*
8. Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. (1998). University of Sheffield: Description of the LaSIE-II system as used for MUC-7. *In Proceedings of MUC-7.*
9. IESL (2003). Information Extraction for Slavonic Languages. *Proceedings of the IESL Workshop in conjunction with the RANLP2003 Conference, Bulgaria.*
10. Krieger, H. and Piskorski, J. (2003). Speed-up methods for complex annotated finite state grammars. *DFKI Report.*
11. Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003.*
12. Przepiórkowski, A. and Woliński, M. (2003b). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), *EACL 2003.*
13. Saloni, Z. (2001). *Czasownik polski. Odmiana, słownik.* Wiedza Powszechna, Warsaw.
14. Tokarski, J. (1993). *Schematyczny indeks* a tergo *polskich form wyrazowych.* Wydawnictwo Naukowe PWN, Warsaw. Elaborated and edited by Zygmunt Saloni.