Detection of Nested Mentions for Coreference Resolution in Polish[⋆]

Maciej Ogrodniczuk[1], Alicja Wójcicka[1],
Katarzyna Głowińska[1,2], and Mateusz Kopeć[1]

[1] Institute of Computer Science, Polish Academy of Sciences
[2] Lingventa

Abstract. This paper describes the results of creating a shallow grammar of Polish capable of detecting multi-level nested nominal phrases, intended to be used as mentions in coreference resolution tasks. The work is based on existing grammar developed for the National Corpus of Polish and evaluated on manually annotated Polish Coreference Corpus.

1 Introduction

One of the numerous results of the National Corpus of Polish project[1] [1] was a formal shallow grammar of Polish, frequently referred to as *NKJP Grammar*, used by Spejd parser [2] to provide automated syntactic annotation [3] of the 1-billion-word corpus. The grammar was recently used by another project, CORE[2] for annotation of mentions — nominal groups referencing discourse-world objects in the Polish Coreference Corpus[3] [4], a 0.5-million-token manually annotated resource of general nominal coreference. Whereas in the former corpus the annotation of syntactic words and groups can be regarded as one of the target actions, in the latter one it is only the basis for subsequent identification of mentions (here: nominal constructs carrying reference to discourse-world objects). Therefore accuracy of this process and its compliance with mention representation (see Section 2) is crucial for the superior task of modelling coreference relations.

Nesting of nominal groups with disparate referents (see: *prezes firmy* '*CEO of a company*') has never been targeted by the NKJP grammar, therefore additional

[⋆] The work reported here was carried out within the *Computer-based methods for coreference resolution in Polish texts (CORE)* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40). The work was also co-financed by the Polish Ministry of Science and Higher Education as a Investment in CLARIN-PL Research Infrastructure and by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications".

[1] NKJP, Pol. Narodowy Korpus Języka Polskiego, see `http://www.nkjp.pl`.

[2] *Computer-based methods for coreference resolution in Polish texts*, see `http://zil.ipipan.waw.pl/CORE`.

[3] PCC, Pol. Polski Korpus Koreferencyjny, see `http://zil.ipipan.waw.pl/PolishCoreferenceCorpus`.

mechanisms have been implemented in the corpus to represent such inclusions (see Section 3). Sections 4 reports on the process of the incorporation of the new rules into grammar while Section 5 evaluates the usefulness of the result to coreference resolution by contrasting mentions detected automatically with the new version of the grammar against manual annotation of mentions in the Polish Coreference Corpus.

2 PCC Mention Model vs. NKJP Grammar

Mentions in PCC are all nominal phrases (NGs) — syntactic groups[4] with nominal or pronominal heads (syntactic and/or semantic). In semantic annotation it is vital to preserve the deep structure of such phrases, e.g. to distinguish *a song* from *the song which was played when we first met* (in Polish even more evident due to absence of articles). A nested nominal phrase is marked as separate from the superordinate phrase when it does not contain a finite verb form having syntactic/semantic head other than those of the superordinate phrase. Moreover, all potentially referential constructs are marked, because it is very difficult to define a clear-cut border between referentiality and non-referentiality, as in the following multi-word expression that usually is seen as non-referential:

Jedna jaskółka wiosny nie czyni. *'One swallow does not make a summer'.*

Tą jaskółką było zniesienie cenzury. Ale to nie znaczy, że wprowadzono demokrację. *'A censorship abolishment was this swallow. But it does not mean that democracy was established.'*

Since coreference resolution is a semantic task, the borderlines of nominal phrases are different from those in NKJP project, where, above all, syntactic criteria were taken into account. The PCC nominal phrase consists not only of adjectives, nouns, gerunds, conjunctions (coordinated groups) and subordinate numerals, but also of superordinate numerals (e.g., *trzy dziewczynki* *'three girls'*), relative subordinate clauses (e.g., *kwiaty, które dostałam wczoraj* *'the flowers, that I got yesterday'*), prepositional phrases, as well as adjectival participles. The complexity of the task is further increased by PP-attachment or by similar ambiguities involving potentially post-modifying adjectival participles.

The NKJP project was aiming for the creation of a 1-billion-word automatically annotated corpus of Polish, with a 1-million-word subcorpus annotated manually. Therefore, many decisions were influenced by the automatic annotation rules/process, and made in order to maintain a high level of consistency, whereas in the CORE project, the whole automatically pre-annotated corpus was verified and post-edited by the annotators. So some ambiguities could be solved by the linguists, e.g., PP-attachment ambiguities (*rozmowa o pogodzie* *'conversation about the weather'*, *rozmowa o piątej godzinie* *'conversation at 5 o'clock'*), potentially post-modifying adjectival participles (*wierzba płacząca* *'weeping willow'*, *dziecko płaczące z wściekłości* *'a child crying with rage'*).

[4] A syntactic group is the longest possible sequence of syntactic words that satisfies certain conditions, i.e., match a Spejd rule or a description in the annotation guidelines.

Syntactic annotation in the National Corpus of Polish was limited to joining words together into constituents. Spejd grammar used in the PCC annotation was the modified version of the NKJP grammar, but due to the fact that NKJP nominal groups were different from the CORE nominal phrases, some modifications were made, e.g., the numeral groups were changed into nominal phrases.

The nominal groups in the NKJP project were extensive — they consisted of as many elements as possible, for e.g. in a phrase composed of consecutive nouns in the genitive case such as *propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju*[5] *'proposal for setting the date of launching the process of introducing reform of the system'*, the whole phrase was the only detected nominal group despite the fact that seven other nested nominal phrases with distinct referents should have been detected.

3 Mention Detection Chain

MentionDetector (`http://zil.ipipan.waw.pl/MentionDetector`) is a tool that uses various information from several text processing applications to annotate Polish texts with mentions.

3.1 Preprocessing

The processing of a raw text begins with part-of-speech tagging with Pantera [6]. Then the text is shallow parsed with Spejd [2] and its morphological component Morfeusz SGJP [7]. The last step is to detect Named Entities, which is done by NER [8]. Information obtained from this step is then used to collect mention boundaries. Spejd has the biggest impact on mention detection, as it produces the largest number of noun groups and single-word nouns used as the mention candidates. With this respect, modifications of the Spejd grammar can bring the greatest benefit to the mention detection task.

3.2 Mention Detection Process

MentionDetector works in three steps:

1. It collects mention candidates from morphosyntactic, shallow parsing and/or named entity level (lack of any layer simply results in fewer mention candidates discovered) and also produces zero-anaphora candidates.
2. It removes redundant/unnecessary candidates.
3. It updates head information among mentions.

At the first stage of the process, mention candidates are extracted from the morphosyntactic level, taking all tokens with a noun (`subst|depr|ger`) or a personal pronoun (`ppron3|ppron12`) tags assigned by the parser. From the shallow parsing level, all syntactic noun groups (with `NG.*` type) and syntactic words

[5] Real NKJP example, see [5].

with noun or personal pronoun ctags (`Noun|Ppron.*`) are taken. Finally, from the named entity level, all named entities that contain at least one noun or pronoun token are also mention candidates. To enable zero subject processing, MentionDetector marks each verb in sentences that do not contain any noun/pronoun token in the nominative case[6], as a mention.

At the second stage redundant mentions are detected by removing one of any two mentions having exactly the same boundaries, exactly the same heads, when one mention is the head of another mention or when two mentions intersect, but not in any way described as previous cases. For such pairs, a "less important mention" is selected for removal, which basically means removing the shorter mention or any mention in case of ties. For example in the following sentence:

Największa zagadka lotnictwa cywilnego musi zostać rozwiązana.
'The greatest mystery of civil aviation must be solved.',

preprocessing may produce the following mention candidates: (semantic heads of multi-word mentions are underlined)

- *lotnictwa* *'aviation'* (based on a token tag or a syntactic word tag),
- *zagadka* *'mystery'* (based on a token tag or a syntactic word tag),
- *lotnictwa cywilnego* *'civil aviation'* (based on a syntactic noun group),
- *Największa zagadka lotnictwa cywilnego* *'The greatest mystery of civil aviation'* (based on a syntactic noun group).

The task of the second stage is then first to remove all duplicates (e.g. *zagadka* *'mystery'* could be found both as a token with a noun tag or a one-word noun group). Then finding mentions with the same heads will be followed by removing *lotnictwa* *'aviation'*, as there is a broader mention of *lotnictwa cywilnego* *'civil aviation'* with the same head. Similarly, *zagadka* *'mystery'* will be removed for the analogous reason.

At the third stage of the process the first token is simply marked as the head of each mention, which does not have one detected automatically.

4 Towards the New Grammar

4.1 Change of Perspective

The original NKJP grammar detects nominal groups, but does not always reveal properly their internal structure. This is due to the order and structure of rules which are designed to detect the longest possible sequence irrespective of the fact whether the group is nested or not. For example the old version of the grammar detects the group: *bardzo małym druczkiem* *'in very small print'*, consisting of two parts: adjectival group *bardzo małym* *'very small'* and noun *druczkiem* *'print'*; the structure of the group can be shown in this way: *bardzo małym druczkiem.* This division is not entirely correct, as the whole group should be interpreted as

[6] Marking verbs instead of adding empty tokens representing zero subjects is just a technical measure implemented in PCC to maintain the original text unchanged.

a group without children: *bardzo małym druczkiem*. The second interpretation, without nesting, is obtained by constructing a new version of grammar.

On the other hand, a nested group *usług firmy* *'services of the company' (gen)* is interpreted as a group without children: *usług firmy* by the old version of grammar. The new version provides another interpretation; it detects the whole phrase 'usług firmy' and additionally preserves the information about the two smaller groups, which make up this group: *usług* (which is marked as syntactic and semantic head of the group) and *firmy*.

4.2 Rule Modification

In order to obtain such a result the structure of the section of rules detecting syntactical groups was modified.

First of all, rules for syntactic groups without nesting are in the new version of the grammar separated from rules for groups with nesting and are placed before them. The internal order of the first part of rules is based on two principles: the type of the group and length of the group. Generally speaking, more specialized rules (e.g. rule detecting addresses or dates) appear earlier in the grammar while the most frequent groups, nominal-adjective groups, are processed at the end. Within types, the rules are ordered from the broadest to the narrowest. The last group of rules corresponds to the creation of syntactic groups out of single nouns, adjectives and numerals.

Groups without nesting should contain only syntactic words (any syntactic group can be an element of such a group). In order to achieve such a result, rules describing groups without nesting are constructed in different ways from rules for groups with nesting. The main problem related to this part of grammar consists in the fact that even groups with complicated structure, containing e.g. adjectives and particles or numerals (as in a group: *kilka kolejnych filii szkolnych* *'a few other school branches'*) have to be built only from syntactic words. While designing rules, the recursiveness of adjective-nominal constructs has to be taken into consideration.

The most problematic group of rules in this part of the grammar is constituted by rules detecting nominal-nominal groups without nesting. Nominal-nominal groups in most cases are nested, but there are some exceptions, e.g. proper names of persons (*Jan Kowalski*) or appositions (*malarz pejzażysta* *'landscape painter'*). The rules for these groups are quite restrictive in order to avoid for example a situation, where a nested group in the genitive is interpreted as an apposition in the genitive (in Polish the text *malarza pejzażysty* has two interpretations: *'a landscape painter (gen)'* or *'a painter of a landscapist (gen)'*, the first is not nested, unlike the second). Our solution consists in making only nested groups from two subsequent nouns, if both are in the genitive and their orthographical forms begin with a small letter.

The second part of rules detecting syntactical groups — the part responsible for nested groups — is built in another manner. The only elements of these groups are other syntactical groups, nested or not nested. Recursiveness of such constructions cannot be achieved by a single rule with regular expressions; all parts of the grammar must be repeated. For example, if we have a group *przedłużenie terminu złożenia projektu budżetu* *'prolonging of the date of submitting the project of the budget'*, our aim is to detect the following structure: *przedłużenie terminu złożenia projektu budżetu*. In the first step the grammar detects a group *projektu budżetu*, in the second — *złożenia projektu budżetu*, in the third — *terminu złożenia projektu budżetu* and so on.

4.3 Nested Groups

There are four main types of nested groups: case-governed groups, prepositional groups, coordinated groups (conjunction governed groups) and relative clauses. Prepositional groups are excluded from this attempt since they are often very hard to distinguish — not only by parsers, but also by native speakers — between the two groups: the group with a preposition that is governed by a verb and a group governed by another nominal group. For example the text *Jaś obserwuje Marysię przy jedzeniu* can be interpreted as *'John is watching Mary while eating'* or *'John is watching how Mary eats'*.

Other types of groups are recognized by the new version of grammar. As mentioned above, in this part of the grammar, the proper order of repeated groups of rules is crucial. The problem arises that different types of groups with nesting can be embedded in all other types of groups (e.g., a coordinated group in a case-governed group and vice versa; a relative clause in a coordinated group and vice versa). Therefore the rules detecting various types of groups must be placed alternately. For example, the group *bandy partyzantów i terrorystów* *'gangs of partisans and terrorists'* is made out of two smaller groups: the one-element group *bandy* *'gangs'* and the coordinated group *partyzantów i terrorystów* *'partisans and terrorists'*. If the rules detecting coordinated groups were placed first, the grammar would find the group *partyzantów i terrorystów* and in the second step the group *bandy partyzantów i terrorystów* would be created, which is the desirable result. However, there also exist groups such as: *naszego państwa oraz sposobu realizacji* *'(of) our state and way of realisation'*. The internal structure of the group is: *naszego państwa oraz sposobu realizacji*, so there is a group with nesting within the coordinated group. If the rules for coordinated groups where at the beginning of this part of the grammar, an incorrect group such as *państwa oraz sposobu* *'our state and way'* would be created. Therefore the order of the rules is as follows:

1. the group of rules detecting case-governed groups, restricted only to the context without comma or conjuction on the right side of the given string (the group *bandy partyzantów* from *bandy partyzantów i terrorystów* is not found in the first step; on the other hand, the group *sposobu realizacji* being a part of *naszego państwa oraz sposobu realizacji* is detected)
2. the rules responsible for coordinated groups (the groups *partyzantów i terrorystów* and *naszego państwa oraz sposobu realizacji* are found)

3. the rules detecting case-governed groups, without the restriction mentioned above (the whole group *bandy partyzantów i terrorystów* is found)

The whole procedure is repeated by detecting longer groups and should be applied also to relative clauses (in the recent version of the grammar this method is used only by case-governed and coordinated groups).

5 Evaluation

Tables 1 and 2 present results of evaluation of the new grammar in two settings: setting 1 corresponds to real-life conditions, with best to-date mention detection, compensating potential grammar deficiencies with named entity recognition and zero-anaphora detection. Setting 2 intends to better illustrate gains resulting directly only from grammar improvements by including in the evaluation only groups detected by the grammar (without named entities etc.), i.e. `NG`, `Noun` and `Ppron` syntactic groups.

The evaluation has been carried out on a test set comprising of 530 texts (out of approx. 1,800) randomly selected from the Polish Coreference Corpus.

Table 1. Evaluation results, setting 1

		NKJP Grammar	**New version**
Mention statistics	**Total gold mentions**	53,407	53,407
	Total system mentions	51,217	51,750
	Total common mentions	33,839	34,176
Mention detection results	**Precision**	66.07%	66.04%
	Recall	63.36%	63.99%
	F1	64.69%	65.00%

Table 2. Evaluation results, setting 2

		NKJP Grammar	**New version**
Mention statistics	**Total gold mentions**	53,407	53,407
	Total system mentions	65,853	69,475
	Total common mentions	31,582	33,122
Mention detection results	**Precision**	47.96%	47.67%
	Recall	59.13%	62.02%
	F1	52.96%	53.91%

The difference in the number of system mentions between settings is a result of the second step of the mention detection algorithm, removing unnecessary mentions using simple heuristics.

Both settings show improvement of recall at the expense of precision (with F1 improved). Relatively low scores (in 50s–60s) results from the strict definition of mention match (exact boundaries) and the mention model itself, e.g. heavily dependent on relative clauses (difficult to access algorithmically).

6 Conclusions

The experiment showed slight improvement in absolute figures as far as mention detection is concerned, but should be regarded as the first step towards further reconstruction of NKJP grammar to enable nesting of different types of syntactic groups, not only the nominal ones. The feasibility of such a process has been confirmed.

In the mention detection chain some actions were taken in order to compensate grammar deficiencies. Now, with use of the new grammar, some of these deficiencies have been overcome.

References

1. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., eds.: Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish, in Polish]. Wydawnictwo Naukowe PWN, Warsaw (2012)
2. Przepiórkowski, A., Buczyński, A.: Spejd: Shallow Parsing and Disambiguation Engine. In Vetulani, Z., ed.: Proceedings of the 3rd Language & Technology Conference, Poznań, Poland (2007) 340–344
3. Głowińska, K.: Anotacja składniowa. [1] 107–127
4. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Polish Coreference Corpus. In Vetulani, Z., ed.: Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland, Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza (2013) 494–498
5. Górski, R.L., Przepiórkowski, A., Łaziński, M., Lewandowska-Tomaszczyk, B.: Polski korpus. In: Academia. Polska Akademia Nauk 4–7
6. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In Loftsson, H., Rögnvaldsson, E., Helgadóttir, S., eds.: Advances in Natural Language Processing. Volume 6233 of Lecture Notes in Computer Science., Springer (2010) 3–14
7. Woliński, M.: Morfeusz – a practical tool for the morphological analysis of Polish. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds.: Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference, Wisła, Poland (June 2006) 511–520
8. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A.: Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish. In: Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10), Wisła, Poland (2010) 531–539 PTI.