Multi-Pass Sieve Coreference Resolution System for Polish

Bartłomiej Nitoń and Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warsaw, Poland

bartek.niton@gmail.com
maciej.ogrodniczuk@ipipan.waw.pl

Abstract. This paper examines the portability of Stanford's multi-pass rule-based sieve coreference resolution system to inflectional language (Polish) with a different annotation scheme. The presented system is implemented in BART, a modular toolkit later adapted to the sieve architecture by Baumann et al. The sieves for Polish include processing of zero subjects and experimental knowledge-intensive sieve using the newly created database of periphrastic expressions. Evaluation shows that the results for Polish are higher than those seen on the CoNLL-2011/2012 data.

Keywords: coreference resolution, BART, the Stanford's multi-pass sieve architecture, Polish language, knowledge-based resources

1 Introduction

Coreference resolution, the task of grouping textual fragments that refer to the same entity in the discourse world, has been at the core of natural language understanding since the 1960s. Proper decoding of reference is important for various applications such as question answering, information extraction and retrieval, machine translation and text summarization.

Owing in large part to the public availability of several coreference-annotated corpora since the 1990s, such as MUC, ACE, and OntoNotes, significant progress has been made in the development of corpus-based approaches to coreference resolution. After a shift from heuristics to machine learning in the 2000s, recorded e.g. in Ng's survey paper [14], the beginning of the current decade brought reversal of these tendencies, with the most prominent multi-pass sieve approach [11], the winner of the CoNLL-2011 shared task on English coreference resolution, followed by several extensions such as Ratinov and Roth's learning-based sieves [25]. Application of this approach to other languages also showed considerable improvements in the resolution results [4,10].

Former coreference resolution systems for Polish [16,17] did not take into account the new advances brought to the field with multi-pass sieve models. In the current paper we adapt BART [27] and its Polish Language Plugin [9] to the sieve architecture following the approach of Baumann et al. [3] and investigate both how it improves coreference resolution score for Polish and how features specific to inflectional languages (lemmatization, zero pronouns and lack of definite articles) are reflected in a sieve-based resolver.

As a separate step we perform the experiment with a knowledge-intensive periphrastic sieve based on a newly created resource combining data from dictionary definitions, plWordNet, Wikidata and common clues found in crossword puzzles linked with potential answer words.

2 Polish Coreference Resolution Sieves

Sieve architecture relies on a sequence of hand-written rules (sieves), ordered from most to least precise. It is more or less a cascade of simple rule based coreference resolvers where the output of one is the input of the next. Thanks to that following sieves can use entity information gathered by previous ones, which makes sieve architecture entity-based. Decisions can be made, not about mentions in the text, but about entities–clusters of mentions in the system's model of the world — allowing the system to reason about the properties of entities as a whole. The system's precision ordering allows it to first link high-confidence mention-pairs, and only later consider lower-confidence sources of information. In our approach we are using both pair (antecedent-anaphora) and entity based features using those which perform better for specific sieve.

Our final system uses eight sieves for Polish coreference resolution. They are described in the next subsections in order of execution. We have also experimented with periphrastic sieve (see Section 4.3), but because of small score improvement, with high memory and time complexity we decided not to include it in our final system. Types of mentions matched by each sieve are shown in Table 1.

Sieve	**Matched mentions**	**Links**	**Correct links**	**Precision [%]**
1. ExactStringMatch	*nominal*	12930	11004	85.10
2. BaseStringMatch	*nominal*	7804	5873	75.26
3. PreciseConstructs	*nominal*	197	145	73.60
4. HeadMatchB	*nominal*	9273	5214	56.23
5. ZeroMatch	*zero*	13228	8632	65.26
6. PronounMatch	*pronominal*	3601	2153	59.79
7. ZeroToNP	*zero with nominal*	3595	1507	41.92
8. PronounToNP	*pronominal with nominal*	3508	1399	39.88

Table 1. Sieves precisions and matched types of mentions

Sieves are evaluated with *MUC* [28], B^3 [2], and *CEAFE* [12] metrics calculated using *Scoreference*[1], a mention detection and coreference resolution evaluation tool [16,

[1] http://zil.ipipan.waw.pl/Scoreference

Chapter 15]. Following i.a. CoNLL-2011 approach [20], for the final evaluation we used average score of the above metrics which tracked influence on different coreference dimensions (the B^3 measure being based on mentions, *MUC* on links, and *CEAFE* on entities).

All experiments were carried out on the Polish Coreference Corpus[2] [16] version 0.92 (all texts).

2.1 Polish Coreference Corpus

Polish Coreference Corpus (PCC) is a large corpus of Polish general nominal coreference built upon the National Corpus of Polish (NKJP)[3] [21]. Each text of the corpus is a 250–350-word sample consisting of full subsequent paragraphs extracted from longer texts. With its 1900 documents from 14 genres, containing about 540,000 tokens, 180,000 mentions and 128,000 coreference clusters, the PCC is among the largest manually annotated coreference corpora in the international community.

Mentions in PCC are understood as broadly as possible, with such complex components as relative clauses, coordinated phrases or prepositional-nominal phrases attached to semantic heads and included in respective nominal phrases. PCC also features annotation of zero anaphora, clitic pronouns attached to verbs, multi-level nested and discontinuous mentions. Appositions are attached (not linked) to respective mention and referential nominal groups are distinguished from attributive ones.

Coreference clusters group mentions with the same reference regardless of linguistic means used to invoke the referent in text.

2.2 Mention Types

During sieve preparation we decided to divide mentions into three types: *nominal*, *pronominal*, and *zero*. The idea was to match mentions within each group with high-precision sieves which would also have a positive impact on overall recall.

Nominal mentions are all nominal phrases whose syntactic head is a noun marked with a *subst* (general noun) or *ger* (gerund) tags[4] while pronominal mentions are first-, second- (annotated as *ppron12*) or third-person pronouns (*ppron3*).

The last group are zero mentions as defined in [8]. For Polish (also all Balto-Slavic languages and most Romance languages) it is possible for an independent clause to lack an explicit subject; its role is maintained by the predicate. Due to its rich morphology, value of person, number and/or gender category of the verb can be used to maintain agreement with referent, as in example below:

(1) Maria wróciła już z Francji. Ø Spędziła tam miesiąc.
'Maria came back from France. She had$_{sg:f}$ spent a month there.'

Zero mentions are marked with tags corresponding to verbal forms (*fin*, *praet*, *bedzie*, *winien* and *aglt*). Moreover, we take into account also verbs tagged as *impt*

[2] `http://zil.ipipan.waw.pl/PCC`
[3] `http://nkjp.pl`
[4] See `http://nkjp.pl/poliqarp/help/en.html` for a concise tag descriptions.

(imperative; it is omitted in [8]). We can clearly imagine texts representing dialogues or instructions using imperatives as zero mentions:

(2) Ø Upewnij się, że SZBD PostgreSQL został pomyślnie uruchomiony. Ø Zaloguj się na konto użytkownika postgres.
'Ø Make sure that PostgreSQL DBMS was successfully launched. Ø Log into postgres user account.'

2.3 Pass 1 and 2 – Exact and Base String Match

Exact String Match Sieve links two nominal mentions only if they contain exactly the same text, without any modification. As expected, this model is extremely precise, see Table 1.

Base String Match Sieve is working in the same way except that it is matching lemmatized forms of mention strings, obtained with Morfeusz morphological analyser[5] [30] and Pantera tagger[6] [1]. This sieve is also highly precise and is also working only on nominal mentions.

Surprisingly, the system is matching more mentions and gets better score when both *Exact* and *Base* sieves are used (see Table 2). We expected that *Base String Match Sieve* would cover all cases covered by *Exact String Match Sieve* but it seems that the setting can correct errors introduced by the tagger, supposedly assigning wrong base forms to some of analysed tokens. The configuration with both sieves obtains better score in every presented measure (see Table 2), so finally we decided to use both *Exact* and *Base* sieves in our system.

Sieves	**Precision**			**F-score [%]**			
	Links	**Correct links**	**Precision [%]**	**MUC**	**B^3**	**CEAFE**	**CoNLL**
Exact	12930	11004	**85.10**	34.17	85.45	80.91	66.84
Base	19072	15462	81.07	44.18	86.51	82.27	70.99
Base+Exact	20734	16877	81.40	**44.43**	**86.54**	**82.31**	**71.09**

Table 2. *String Match* sieves comparison

2.4 Pass 3 – Precise Constructs Sieve

Initially this sieve was intended to mimic the original Precise Constructs Sieve described by [23] which linked two mentions if one of the following rules is fulfilled:

– the two nominal mentions are in an appositive construction

[5] http://sgjp.pl/morfeusz/
[6] http://zil.ipipan.waw.pl/PANTERA

- the two mentions are in copulative subject-object relation
- the candidate antecedent is headed by a noun and appears as a modifier in an NP whose head is the current mention
- the mention is a relative pronoun that modifies the head of the antecedent NP
- one mention is an acronym of the other
- one of the mentions is a demonym of the other.

Eventually we decided to link acronyms only due to decisions taken in the PCC annotation where appositive constructions are marked as single mention, copula constructions are not marked at all and intersecting mentions are never marked as being in the same cluster.

The acronym rule occurred to be highly precise even though it does not cover many cases in real text (only 197 matched links in full corpora, with 73.6% precision). For the rule we use a simple acronym detection algorithm which marks a mention as an acronym of another if its text equals the sequence of uppercase characters in the other mention.

Demonym case was also tested but it did not affect coreference score in a positive way.

2.5 Pass 4 – Strict Head Matching

Similarly to *Precise Constructs Sieve* this one was also inspired by [23]. It is responsible for matching nominal mentions and is the first one to use cluster information gathered by previous sieves.

Originally the sieve was passed three times with match rules relaxation after each pass. The most orthodox pass (*HeadMatchA*) links two mentions only if they match all of the following rules:

- *Cluster head match* – the mention head word matches any head word in the antecedent cluster
- *Word inclusion* – all the non-stopwords[7] in the mention cluster are included in the set of non-stopwords in the cluster of the antecedent candidate
- *Compatible modifiers only* — the mention's modifiers are all included in the modifiers of the antecedent candidate, with only nouns and adjectives taken into account
- *Not i-within-i* – the two mentions are not in an i-within-i construct [7].

The second (*HeadMatchB*) and more relaxed sieve removes *Compatible modifiers* rule, while the third one (*HeadMatchC*) removes also the *Word inclusion* constraint.

Because in *PCC* appositive phrases are marked as a single mention in *not i-within-i* rule, we are simply checking if one mention string is not embedded in another mention string.

In Table 3 we present *Strict Head Match* sieves configurations precision. Precision is calculated only for specified configuration, but all of the preceding sieves are used (*Exact String Match*, *Base String Match*, *Precise Constructs*).

[7] Polish stop-words list was taken from the Polish Wikipedia stop-words list `https://pl.wikipedia.org/wiki/Wikipedia:Stopwords`

Sieves: Base+Exact +Precise+...	**Precision** Links	Correct links	Precision [%]	**F-score [%]** MUC	B^3	CEAFE	CoNLL
HeadMatchA	1359	771	**56.73**	45.50	86.58	**82.39**	71.49
HeadMatchB	9273	5214	56.23	**48.99**	**86.61**	82.31	**72.64**
HeadMatchC	23049	7305	31.69	47.49	84.42	77.95	69.95
HeadMatchAB	9499	5339	56.21	48.97	**86.61**	82.30	72.63
HeadMatchAC	23621	7640	32.34	47.49	84.41	77.94	69.95
HeadMatchBC	27192	9655	35.51	47.49	84.41	77.94	69.95
HeadMatchABC	27247	9708	35.63	47.49	84.41	77.94	69.95

Table 3. Different *Head Match* sieves configurations comparison

As expected, the most precise configuration is the one using only *HeadMatchA* sieve. Unfortunately, it matches a small number of links as compared to other configurations. The biggest recall is acquired, as expected, by the *HeadMatchC* sieve, but in this case precision is very low.

As we can see in Table 3, the highest score is obtained for the first relaxation of *Head Match* sieve (B). Therefore, we choose *HeadMatchB* relaxation as our next sieve.

2.6 Pass 5 – Zero Mentions Match

This pass is matching zero mentions within their group. First of all it is checking whether both mentions are zero mentions and their numbers match. If both constraints are met, based on the part of speech tag we are then checking person (for *fin*, *bedzie*, *impt*, *aglt* tags) or gender (for *praet*, *winien*) match. If all conditions are met, mentions are marked as coreferent.

As we can see in Table 1, precision of this sieve is not very high (65.26%). It can be easily increased by matching only mentions in the same paragraph, which is in accordance with intuition: new object is brought into the discourse mostly at the beginning of the paragraph and then we are mentioning it with zero mentions. Bringing up the same paragraph constraint raises sieve precision to 79.72% (8409 total links, with 6704 out of them correct), but at the same time overall coreference score is decreasing (see Table 4). Because of that we decided not to use this constraint in the current version of the system.

2.7 Pass 6 – Pronoun Match Sieve

This pass is matching personal pronouns within their group. Pronouns are matched when their person, gender, and number agree.

Similarly to *Zero* sieve, precision of *Pronoun* sieve is not very high (59.79%). It can be also easily increased by matching only mentions in the same paragraph: a new object is brought into the discourse mostly at the beginning of the paragraph and then we are mentioning it by pronouns. Bringing up same paragraph constraint raises sieve precision to 78.9% (1673 total links, with 1320 out of them correct), but the correct

Configuration	F-score [%]			
	MUC	B^3	CEAFE	CoNLL
No SameP	**67.77**	86.58	**87.53**	**80.63**
Zero+SameP	65.33	86.81	86.67	79.60
Pronoun+SameP	67.00	**86.83**	87.28	80.37

Table 4. Overall system score with or without same paragraph constraint for *Zero* and *Pronoun* sieves

ones cover half of the links added without this constraint so at the same time the overall coreference score is decreasing (see Table 4). So as for *Zero Mentions Sieve* we decided not to use the same paragraph constraint at this time.

2.8 Pass 7 – Zero To Nominal Mention Sieve

This sieve is matching zero mentions against nominal mentions. This sieve and the next one are currently very simple and have low precision but at the same time offer a positive impact on overall *CoNLL* system score (see Table 5).

In this sieve we are simply checking if the antecedent is a nominal mention tagged as *subst* and is the first mention in the sentence. If previous constraints are met we are checking if the number and gender of mentions match.

System	F-score [%]			
	MUC	B^3	CEAFE	CoNLL
Exact	34.17	85.45	80.91	66.84
...+Base	44.43	86.54	82.31	71.09
...+Precise	44.59	86.56	82.32	71.16
...+HeadMatchB	48.99	86.61	82.31	72.64
...+Zero	63.60	**87.05**	86.19	78.95
...+Pronoun	66.14	86.90	87.32	80.12
...+ZeroToNP	66.78	86.78	87.31	80.29
...+PronounToNP	**67.77**	86.58	**87.53**	**80.63**

Table 5. Coreference resolution score changes after adding new sieves to the system

2.9 Pass 8 – Pronoun To Nominal Mention Sieve

This sieve is matching personal pronouns against nominal mentions. The link is created when the antecedent is a nominal mention and anaphor is a personal pronoun, they are in the same paragraph, and their number match. If previous constraints are met we are checking if pronoun gender and person are matching gender and person of any mention in the nominal mention cluster. Unknown gender and person is treated as a wildcard.

3 Results

Table 6 presents comparison of *Bartek-S1*, our sieve-based solution described in this article and two existing coreference resolution systems for Polish described in detail in [16]. *Ruler* is simple rule-based tool with design following [6] and *Bartek-3* is an adaptation of the BART system for Polish, being at the moment the best machine learning based system for coreference resolution for Polish.

System	F-score [%]			
	MUC	B^3	CEAFE	CoNLL
Ruler	58.21	81.94	80.04	73.40
Bartek-3	64.68	85.31	85.24	78.41
Bartek-S1	**67.16**	**86.66**	**87.57**	**80.47**

Table 6. Coreference resolution systems for Polish; scores for *Bartek-S1* were counted on the same subset of 530 texts from *PCC 0.92* as scores of *Ruler* and *Bartek-3* taken from [16].

The comparison shows that even without using complex statistical mechanisms our system performs slightly better than previous systems for Polish (more than 2% *CoNLL* score increase over state-of-the-art *Bartek-3* system). The reason for that is twofold: firstly we explicitly divide mentions by types and match them within each group, which was not present in previous systems (specially for zero mentions, not treated as separate problem at all); secondly using sieve architecture provide us with the whole entity information. In conclusion, sieve architecture outperforms previous systems because it gives us a mechanism to divide coreference resolution into subproblems making information flow very natural: use highly precise general sieves first, match mentions within each mention type, try to match mentions of different types using cluster (entity) information.

4 Experiments with a Periphrastic Sieve

After completion of the sieve system additional experiment was performed to verify whether knowledge-intensive resources could be used as input for a high-precision sieve. Even though the results did not meet our expecations, we present them below.

4.1 Related Work

Ponzetto and Strube [18,19] describe use of Wikipedia, WordNet and semantic role tagging in computing semantic relatedness between anaphor and antecedent to achieve 2.7 points MUC F_1 score improvement on ACE 2003 data.

Rahman and Ng [24] labelled nominal phrases with FrameNet semantic roles achieving 0.5 points B^3 and CEAF F_1 score improvement and used YAGO type and means relations achieving 0.7 to 2.8 points improvement on OntoNotes-2 and ACE 2004/2005 data.

Durrett and Klein [5] incorporated in their system shallow semantics by using WordNet hypernymy and synonymy, number and gender data for nominals and propers, named entity types and latent clusters computer from English Gigaword corpus, reaching 1.6 points improvement on gold data and 0.36 points on system data.

For Polish, WordNet and Wikipedia-related features were used to improve verification of semantic compatibility for common nouns and named entities in BARTEK-3 coreference resolution system [16, Section 12.3] resulting in improvement of approx. 0.5 points MUC F_1 score. Experiments with integration of external vocabulary resources coming from websites registering the newest linguistic trends in Polish, fresh loan words and neologisms not yet covered by traditional dictionaries have been also performed showing low coverage of new constructs in evaluation data [15].

All these results showed challenges regarding knowledge-based resources, mainly concerning the memory and time complexity of the task as well as low coverage of complex features in the test data, but at the same time brought some (sometimes tiny) improvements to coreference resolution scores. In this article we describe if this 'tiny' improvements can be also acquired using the new knowledge database for Polish *Periphraser*.

4.2 Periphraser

Periphraser is a newly created knowledge base of conventionalized periphrastic nominal expressions (i.e. phrases headed by a noun) together with their textually attested realizations. For instance, the database entry for the phrase "Lewandowski" will include the phrase "the Polish international" while "pediatrics" will be featured as "medical care for children". The database is still expanding and at this moment contains over:

- 78,000 meanings and 193,000 expressions from *SJP*[8], a community-built dictionary of Polish
- 72,000 meanings and 183,000 expressions from *plWordNet*[9] [13], the largest WordNet of Polish
- 157,000 meanings and 384,000 expressions from *Wikidata*[10]
- 239,000 meanings and 497,000 expressions from the crosswords portal *Szarada.net*[11].

4.3 Periphrastic Sieve

The periphrastic sieve is intended to link mentions which are hard to match using syntactic features only but which are attested by the knowledge sources included in *Periphraser*. The match can be achieved by:

- *heads matching* – checking if mentions heads are connected in *Periphraser*
- *whole expressions matching* – checking if the whole mentions strings are connected in *Periphraser*

[8] http://sjp.pl/
[9] http://plwordnet.pwr.wroc.pl/
[10] https://www.wikidata.org/
[11] http://szarada.net/

– *head to expression matching* – checking if the head of one mention is connected to the other mention string in *Periphraser*.

In our experiments we used lemmatized forms of strings, both on the side of *Periphraser* with strings tagged by *Concraft-pl* [29] and on the *PCC* side with texts tagged during preanotation by *Pantera* [1] tagger. Scores are counted on the subcorpus of 1250 short PCC texts using *Scoreference* tool. *Periphraser* sieve follows *Strict Head Matching* sieve (see Section 2.5) in our experiments.

Matching	Precision			F-score [%]			
	Links	Correct links	Precision [%]	MUC	B^3	CEAFE	CoNLL
Heads	**7905**	**641**	8.1	63.59	84.28	83.47	77.11
Expressions	1606	292	**18.2**	**67.51**	**86.53**	**87.21**	**80.42**
Head to expression	4102	472	11.5	65.98	85.67	85.86	79.17
Heads (descr ana)	**4733**	**219**	4.6	64.64	85.13	84.84	78.20
Expressions (descr ana)	107	48	**44.9**	**68.06**	**86.92**	**87.82**	**80.93**
Head to expression (descr ana)	1540	115	7.5	67.03	86.36	86.97	80.12

Table 7. *Periphraser* – possible ways of matching (head or expression)

From the previous experiments we know that matching mentions by their heads without any additional constraints is not sufficiently effective. In fact, matching them using *Periphraser* is even more error prone because we are using, more or less, synonyms (cf. first part of Table 7). In *Periphraser* the mentioned *whole expression* can be in fact a single mention, in which case again mention matching will be brought to synonymous head matching problem.

Because in the text (usually) we are using simple entity name before we are writing about it, to avoid repetition, in a more descriptive way, thus it will be natural that anaphor should consist of more than one significant word (see the second part of Table 7).

As we can see in Table 7, matching periphrastic expressions without any constraints is very imprecise. It is getting more promising when we assume that anaphora should consist of more than one word, but it is still far from being satisfying and does not cover many cases in real texts. One more conclusion from this experiment is that it is best to match whole mentions instead of using their heads only.

Another problem is whether it is better to use pair or entity information while linking possibly periphrastic mentions. In general, entity information should help, but in our system nominal sieves preceding periphrastic ones are not very precise (see Section 2.5), match small number of mentions (see Section 2.4) and exact/base strings (which is not bringing too much new data to the entity). In Table 8 we can see that sieve based on the entity and mention pair features gives more or less the same results.

As we can see simple matching is not very precise, especially when we are matching mentions by heads. Thus we must add some constraints to minimize error rate.

Matching	Precision			F-score [%]			
	Links	Correct links	Precision [%]	MUC	B3	CEAFE	CoNLL
Pairs	**1606**	292	18.2	**67.51**	**86.53**	**87.21**	**80.42**
Entities	1582	**422**	**26.7**	**67.51**	86.52	**87.21**	80.41
Pairs (descr ana)	107	**48**	**44.9**	**68.06**	**86.92**	**87.82**	**80.93**
Entities (descr ana)	**111**	**48**	43.2	**68.06**	**86.92**	**87.82**	**80.93**

Table 8. *Periphraser* – possible ways of matching ’whole expressions’ (pair or entity)

4.4 Error Analysis

Analyzing sieve errors we discovered that a lot of them can be filtered out by grammatical number match rule, e.g.:

- *ślimaka ’snail’ – winniczków ’pomatia snails’*,
- *słońcem ’sun’ – gwiazdy ’stars’*,
- *Szczecinie – miasta ’cities’*.

Other common errors come from imperfections of *Periphraser* database. It is still under construction and requires some data cleaning which is especially visible while matching mentions by heads coming from crossword-related part of the data, e.g.:

- *minus ’minus’ – plus ’plus’*, crossword answers can be based on antonyms
- *tajemnica ’secret’ – film ’movie’*, because there is a French movie titled ’Un Secret’
- *liga ’league’ – mistrza ’champion’*, crossword answers can be also based on associations, here with *Liga Mistrzów ’the Champions League’*.

Other possible constraints are: full grammatical agreement (person, gender, number match) or simple semantic class agreement. Table 9 presents best results acquired by *Periphraser* sieve using various types of constraints:

- *Base* – the system scores without using *Periphraser* sieve
- *Exp1* – a pair based matching of whole expressions with grammatical number match rule and descriptive anaphora rule
- *Exp2* – an entity based matching of whole expressions using grammatical agreement, basic semantic classes agreement and descriptive anaphora rule.

To conclude, using *Periphraser* in coreference resolution system for Polish does not bring significant improvements. To make periphrastic sieve precision satisfying we must use a lot of constraints and at the same time recall is dropping very quickly. Even then we get some errors hard to recognise without knowing wider text context, e.g.:

- *prezydenta ’president’ – głowę państwa ’head of state’*
- *telewizją kablową ’cable television’ – sieci kablowej ’cable network’*
- *narkotyków ’drugs’ – środków odurzających ’intoxicants’*
- *prezydent Rosji ’president of Russia’ – Borys Jelcyn ’Boris Yeltsin’*.

Matching	Precision			F-score [%]			
	Links	Correct links	Precision [%]	MUC	B3	CEAFE	CoNLL
Base	N/A	N/A	N/A	68.03	**86.93**	**87.83**	80.93
Exp1	**94**	**47**	50	**68.07**	**86.93**	**87.83**	**80.94**
Exp2	24	17	**70.8**	68.04	**86.93**	**87.83**	80.93

Table 9. The most precise *Periphraser* sieve configurations

5 Conclusion and Future Work

In this article we described adaptation of a multi-pass sieve approach to coreference resolution for Polish showing its good adaptability and advantage of using inflectional properties, reflected in the resolution score, higher than those seen on the CoNLL-2011/2012 data.

On the other hand, we were testing *Periphraser* database usage for coreference resolution. Summarizing *Periphraser* sieve is matching what it meant to match, but in real texts it is not always correct match. For example, in text about Russian politicians, both Boris Yeltsin and Vladimir Putin can be referred as president or prime minister of Russia. Moreover, such knowledge has low coverage in the test data. *Periphraser* is still under development so one of the next steps will be data verification. More complex machine learning algorithms can also be applied (initial experiments with *C4.5* algorithm used to generate a decision tree [22] resulted in insignificant improvement of 0.03% *CoNLL* score).

Currently the system offers rule-based sieves only, so the most immediate step would be testing whether combination of rule-based and machine-learning sieves could improve resolution results as seen in the work of [25] and [26]. The most promising candidates for adoption of statistical methods seem to be sieves matching mentions of different types (*nominal*, *pronominal*, and *zero*), preceded by rule-based highly precise sieves clustering mentions within a single type.

Acknowledgements

The work reported here was carried out within the research project financed by the Polish National Science Centre (contract number 2014/15/B/HS2/03435) and as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

1. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) Advances in Natural Language Processing: Proceedings of the 7th International Conference on NLP (IceTAL 2010), pp. 3–14. Springer Berlin Heidelberg, Berlin, Heidelberg (2010), `http://dx.doi.org/10.1007/978-3-642-14770-8_3`

2. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: Proceedings of The 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. pp. 563–566 (1998)
3. Baumann, J., Kühling, X., Ruder, S.: Rule-based Coreference Resolution with BART (2014)
4. Chen, C., Ng, V.: Combining the Best of Two Worlds: A Hybrid Approach to Multilingual Coreference Resolution. In: Joint Conference on EMNLP and CoNLL: Proceedings of the Shared Task. pp. 56–63 (2012)
5. Durrett, G., Klein, D.: Easy Victories and Uphill Battles in Coreference Resolution. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1971–1982. Association for Computational Linguistics, Seattle, Washington, USA (October 2013), http://www.aclweb.org/anthology/D13-1203
6. Haghighi, A., Klein, D.: Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. In: Carroll, J.A., van den Bosch, A., Zaenen, A. (eds.) Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 848–855. The Association for Computational Linguistics (2007)
7. Haghighi, A., Klein, D.: Simple coreference resolution with rich syntactic and semantic features. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009): Volume 3. pp. 1152–1161. Association for Computational Linguistics (2009)
8. Kopeć, M.: Zero subject detection for Polish. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers. pp. 221–225. Association for Computational Linguistics, Gothenburg, Sweden (2014)
9. Kopeć, M., Ogrodniczuk, M.: Creating a Coreference Resolution System for Polish. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012. pp. 192–195. ELRA, Istanbul, Turkey (2012)
10. Krug, M., Puppe, F., Jannidis, F., Macharowsky, L., Reger, I., Weimar, L.: Rule-based Coreference Resolution in German Historic Novels. In: Proceedings of the 4th Workshop on Computational Linguistics for Literature. pp. 98–104. Association for Computational Linguistics, Denver, Colorado, USA (June 2015), http://www.aclweb.org/anthology/W15-0711
11. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task. pp. 28–34. Association for Computational Linguistics (2011)
12. Luo, X.: On Coreference Resolution Performance Metrics. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 25–32. HLT '05, Association for Computational Linguistics, Vancouver, Canada (2005), http://dx.doi.org/10.3115/1220575.1220579
13. Maziarz, M., Piasecki, M., Szpakowicz, S.: Approaching plWordNet 2.0. In: Proceedings of the 6th Global Wordnet Conference. Matsue, Japan (January 2012)
14. Ng, V.: Supervised Noun Phrase Coreference Research: The First Fifteen Years. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 1396–1411. ACL 2010, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
15. Ogrodniczuk, M.: Discovery of Common Nominal Facts for Coreference Resolution: Proof of concept. In: Prasath, R., Kathirvalavakumar, T. (eds.) Mining Intelligence and Knowledge Exploration (MIKE 2013), Lecture Notes in Artificial Intelligence, vol. 8284, pp. 709–716. Springer-Verlag, Berlin, Heidelberg (2013)
16. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Coreference in Polish: Annotation, Resolution and Evaluation. Walter De Gruyter (2015), http://www.degruyter.com/view/product/428667

17. Ogrodniczuk, M., Kopeć, M.: Rule-based coreference resolution module for Polish. In: Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011). pp. 191–200. Faro, Portugal (2011)
18. Ponzetto, S.P., Strube, M.: Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In: Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. pp. 192–199. Association for Computational Linguistics, New York City, USA (June 2006), http://www.aclweb.org/anthology/N06-1025
19. Ponzetto, S.P., Strube, M.: Knowledge Derived from Wikipedia for Computing Semantic Relatedness. Journal of Artificial Intelligence Research 30(1), 181–212 (2007)
20. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: CoNLL-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task. pp. 1–27. CONLL Shared Task '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2132936.2132937
21. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. Wydawnictwo Naukowe PWN, Warsaw (2012)
22. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
23. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.: A multi-pass sieve for coreference resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 492–501. Association for Computational Linguistics (2010)
24. Rahman, A., Ng, V.: Coreference Resolution with World Knowledge. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 814–824 (2011)
25. Ratinov, L., Roth, D.: Learning-based Multi-sieve Co-reference Resolution with Knowledge. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1234–1244. EMNLP-CoNLL'12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2390948.2391088
26. Stoyanov, V., Eisner, J.: Easy-first Coreference Resolution. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). pp. 2519–2534. Mumbai, India (2012)
27. Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A.: BART: A modular toolkit for coreference resolution. In: Association for Computational Linguistics (ACL) Demo Session (2008)
28. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A Model-Theoretic Coreference Scoring Scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6). pp. 45–52 (1995)
29. Waszczuk, J.: Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). pp. 2789–2804. Mumbai, India (2012)
30. Woliński, M.: Morfeusz reloaded. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 1106–1111. ELRA, Reykjavík, Iceland (2014), http://www.lrec-conf.org/proceedings/lrec2014/index.html