IKAR: An Improved Kit for Anaphora Resolution for Polish

Bartosz Broda *Łukasz Burdka* *Marek Maziarz*
Institute of Informatics, Wroclaw University of Technology, Poland
{bartosz.broda, marek.maziarz}@pwr.wroc.pl, luk.burdka@gmail.com

ABSTRACT
This paper presents Improved Kit for Anaphora resolution (IKAR) – a hybrid system for anaphora resolution for Polish that combines machine learning methods with hand written rules. We give an overview of anaphora types annotated in the corpus and inner workings of the system. The preliminary experiments evaluating IKAR resolution performance are discussed. We have achieved promising results using standard measures employed in evaluation of anaphora and coreference resolution systems.

KEYWORDS: Anaphora Resoulution, Coreference, Polish.

Proceedings of COLING 2012: Demonstration Papers, pages 25–32,
COLING 2012, Mumbai, December 2012.

1 Introduction

The basic definition of anaphora says that it links two expressions which point to the same entity in real world (Huang, 2010). Anaphora Resolution (AR) is a difficult and important problem for Natural Language Processing (NLP). The difficulty not only comes from the computational point of view, but also from wide range of linguistic issues For extensive review of both the theoretical aspect and approaches used see (Poesio et al., 2010).

In this paper we present IKAR (*Improved Kit for Anaphora Resolution*), a hybrid toolkit for AR for Polish. We combine Machine Learning (ML) and rule based methods. This approach is an extension of our preliminary experiments which was exclusively based on ML. As we need an AR for solving practical NLP problems (e.g., question answering), after initial ML experiments we observed that some phenomena can be easily tackled with rule based component. The combination of both in IKAR is guided by a ML classifier, which allows for further extensions of both typical features used by ML in the problem domain as well as adding more rules.

The work on AR is relatively sparse for Polish. There have been some early approaches, but they were limited in scope, e.g., (Marciniak, 2002; Matysiak, 2007). Recently, the field has been invigorated with baseline work on both rule-based and ML approaches to AR (Kopeć and Ogrodniczuk, 2012; Ogrodniczuk and Kopeć, 2011a,b). Their approach differs from ours in two important ways. First, the definition of anaphora is different and the dataset is smaller than employed in this work. See Sec. 2 for the subtypes of anaphora that we deal with in this work. Second, they evaluate rule-based system and statistical system independently.

2 Anaphora in KPWr

Coreference is a type of anaphora. It links two expressions which point to the same referent (or denote the same class of entieties) in real world (Huang, 2010; Stede, 2012). Out of different types of anaphora distinguished by semanticists, cf. (Halliday and Ruqaiya, 1976; Cornish, 1986; King, 2010), and computational linguists, (Mitkov et al., 2000; NP4, 2010; Poesio, 2004; ACE, 2010), we have chosen those phenomena which can be roughly named after (Mitkov, 2003) a *direct anaphora*. We distinguish: (a) coreference between nominal and pronominal expressions (Larson and Segal, 1996; King, 2010), (b) coreference between two nominal phrases (either based on identity of reference or on lexical synonymy/hyponymy/hypernymy (Mitkov, 2003)). We add to this group also (c) *zero anaphora* - i.e., anaphora of omitted subject. In order to further limit issues connected with coreference recognition we have decided to annotate only coreferential relations to proper nouns. The KPWr Corpus (Broda et al., 2012) was annotated by two annotators which worked on separated documents. The annotators were supported with precise guidelines (Maziarz et al., 2011a), during annotation process they were encouraged to ask a superior linguist anytime they needed and to modify this document. Because of the procedure the inter-annotator agreement could not be checked[1]. The KPWr Corpus is avaiable under the CC BY 3.0 licence (Broda et al., 2012).

(**1**) **Coreference between two proper names** (PN-PN type).

This relation type links occurrences of coreferent proper names. Majority of the relation instances are instances of the same proper name:[2]

[1a] (...) chcę być tylko bliżej *Loty*, oto cały sekret. (...) moje i *Loty* serca rozumieją się tak doskonale. [‘(...) I only

[1] In near future we are aiming at annotating 10% of the annotated corpus in order to check the kappa

[2] All examples are taken from KPWr

wish to be closer to *Charlotte*, —that is the secret. (...) my heart and *Charlotte's* understand each other so perfectly.']

Seldom the same referent is named with different names:

[1b] Uniwersytet tworzy się z *Filii Uniwersytetu Warszawskiego w Białymstoku*. (...) zatrudnieni w *Filii w Białymstoku* stają się pracownikami Uniwersytetu. ['The University forms from *the Branch of Warsaw University in Białystok*. (...) employees hired at *the Branch in Białystok* become employees of the University.']

(2) Coreference between a proper name and an agreed noun phrase (PN-AgP type)

With PN-AgP link we capture coreference between a proper name and an agreed noun phrase based on hyponymy/hypernymy, i.e. common noun denoting a class of entities which includes a referent of a proper noun). Under *agreed noun phrase* we understand nominal phrase built on syntactic agreement on number, gender and case:[3]

[2a] (...) ataki na schorowanego generała *Jaruzelskiego*. To przecież Jarosław Kaczyński porównał *generała* do Adolfa Eichmanna ['(...) attacks on ailing general *Jaruzelski*. It was Jarosław Kaczyński who compared *the general* to Adolf Eichmann.']

Here *generał* 'general' is a class which includes general Jaruzelski.

We have annotated not only cases of anaphora, but also of cataphora. In [2b] a usual direction of anaphora is reversed: a common noun *człowiek* 'a man' – a head of AgP *człowiek... nieco otyły* 'a little bit obese man' – is a class of entities which *Napoleon* belongs to:

[2b] (...) jechał na białym koniu *człowiek-1* średniego wieku, *1-nieco otyły* (...). Pierwszym z tych jeźdźców był *Napoleon*, drugim byłem ja ['(...) *a little bit obese man* of the medium age rode a white horse (...). From these riders *Napoleon* was first, I was second.']

(3) Coreference between a pronoun and a proper name (PN-Pron type):

The main subclass of PN-Pron corefrence links a *personal* pronoun with a proper name. In [3a] a pronoun of the third person – *jej* (ona:ppron3:sg:dat:f)[4] – points to a name of a former female-treasurer of a municipal council – *Aniela T.*:

[3a] (...) wieloletnia była skarbnik gminy *Aniela T*. Wójt (...) kazał *jej* opuścić budynek ['*Aniela T.*, a long-standing treasurer (...). The borough leader (...) ordered *her* to leave the building.']

In KPWr we annotate also coreference between demonstrative pronouns and proper names. In [3b] the pronoun *tam* 'there' refers to the Internet:

[3b] (...) *internet* jest nie dla nich, że nie ma *tam* miejsc, które mogłyby ich zainteresować (...) ['(...) *the Internet* is not for them and *there* are not sites interesting for them']

(4) Zero anaphora - coreference between a proper noun and zero-subject (PN-ϕ type)

In Polish subject is often omitted. We wanted to link coreferent proper name and zero-subject; to avoid introducing into text artificial units, we have decided to establish links to verbs with zero-subjects, like in this example:

[4a] *Toronto Dominion Centre* - kompleks handlowo-kulturalny (...). *Składa się* z 3 czarnych budynków (...). ['*The Toronto-Dominion Centre* - is a cluster of buildings (...) of commercial and cultural function. (It) *consists* of three black buildings (...)']

[3] For further details and definitions please see (Radziszewski et al., 2012)

[4] The tags come from (Woliński, 2003)

3 An Improved Kit for Anaphora Resolution

The aim of our experiment with IKAR[5] (Improved Kit for Anaphora Resolution) is to mark pairs of annotations joined by the anaphora relation. Such a pair always consists of a mention and its antecedent. We recognize so far these relations that point backwards, i.e., pairs that consist of a mention and its antecedent (so cases of cataphora were excluded). A mention can be a proper name (PN), a pronoun or an AgP. The antecedent is always a PN. We leave recognizing zero-subjects for further works.

3.1 Experimental Settings

The idea of the whole experiment is as follows: we create a list of annotation pairs on the basis of the annotated corpus, extract features from these pairs and classify if they should be joined by the anaphora relation. Then we compare the outcome with real relation instances.

The learning sequence has to contain positive and negative examples. The selection of positive examples is straightforward, i.e., they consists of coreference annotation pairs. The selection of negative examples needs more attention. We use different approaches and features for each one of the three recognized relation types (PN-PN, PN-AgP, PN-Pron).

(1) **Coreference between two proper names** (PN-PN type). For each entity referred to by a proper name a chain of references is created. Then each PN is linked to the nearest PN referring to the same entity that occurred in the text before. These pairs constitute positive PN-PN examples. For each mention, negative pairs are created by that mention and its false 'antecedents' from certain range between original mention and its real antecedent. This procedure guarantees that they will not point to actually the same entity. We produced 3006 positive and 14676 negetive examples using this approach.

For each relation type a distinct set of features is established for classification purposes. The PN-PN recognition appeared to be the easiest one. The PN-PN classifier is based mostly on similarity of both PN phrases. Following features are extracted in order to determine if two annotations should be joined. *CosSimilarity*: it measures how much both phrases are formed by the same set of words. Base forms of each token are compared. *TokenCountDiff*: difference in number of tokens forming each PN. *SameChannName*: feature indicating if both PNs share the same type of proper name. *Number*: feature indicating if both PNs share the same grammatical number. *Gender*: feature indicating if both PNs share the same gender. We employ C4.5 decision trees (Quinlan, 1993) in the experiments.

(2) **Coreference between an agreed noun phrase and a proper name** (PN-AgP type). Similarly to PN-PN case, all AgPs in KPWr are related to the first occurrence of a given entity. It is more natural to see the anaphora between an AgP and the nearest PN (of course, if it points to the very same entity). We generate positive PN-AgP examples taking an AgP and its antecedent PN from the same coreference chain. Negative examples are generated in a different way than for PN-PN. For each mention, we choose any proper name that occurred in the text earlier not further than 50 tokens. We take just up to two negative 'antecedents' for each mention. This way we have obtained 1077 positive and 1833 negative examples.

Unlike in a PN-PN case, both annotations (i.e., a PN and an AgP) does not need to sound similar or even share the same gender. Thus, to tell whether an AgP refers to a given PN we

[5] Will be released on GPL `http://nlp.pwr.wroc.pl/en/tools-and-resources/ikar`

need to focus on semantic similarity of AgP's head and a semantic category of a particular PN. We use only one feature called SemanticLink to determine if the relation is present. It is more complex than PN-PN set of features so it needs a closer look. SemanticLink takes advantage of the Polish WordNet (Piasecki et al., 2009) to rate the semantic similarity.

Semantic Link algorithm (For a pair of AgP head name category) For each name category a representative synset from the wordnet was selected manually. Then the procedure is following: First, find matching synset for AgP's head. Search is capped up to 10 similarity, hypernym and holonym edges. If it cannot be found, switch places of category's synset and head's synset and search again. (In case the head's synset is more general than that of category's.) If it cannot be found, the distance is minimal number of edges separating head and the category synset. (Note that a head usually gets more than one synset, because its meaning is not disambiguated.) The score: 1/distance can be interpreted as how well AgP's head can refer to PN from a given category. If there is no better antecedent candidate between AgP and a given PN then it is a positive match.

(**3**) **Coreference between a pronoun and a proper name** (PN-Pron type). When recognizing pronoun-PN relations the important thing we have to focus on is the distance, sharing the mention and its antecedent. In Polish language a pronoun often refers to the latest entity that shares number and gender with it (this observation is supported by the results of our experiments). However, it can happen that a pronoun refers directly to an AgP instead of a PN. Then, we check if a given AgP refers to the same PN. If so we should assume that this PN is in fact coreferent to the pronoun. Again, we use a single binary feature called Pronoun Link. Negative examples were generated like in PN-AgP case. We have obtained 450 positive and 596 negative examples.

Pronoun Link algorithm Check if there is an AgP between a pronoun and a PN that meets Semantic Link criteria for a given PN and gender and number for a given pronoun and there in no closer AgP which meets these criteria. If the condition is fulfilled there is a link between that pronoun and a PN. Otherwise, check if a pronoun and a PN share the same gender and number and if there is no closer PN that meets these criteria. If the condition is fulfilled there is a Pronoun Link.

3.2 Resolution process in IKAR

When given a plain text the process of anaphora resolution requires a few additional steps. The text needs to be divided into tokens and sentences. Next, we need to perform morphologicall analysis (Radziszewski and Śniatowski, 2011) and a morpho-syntactic disambiguation (Radziszewski and Śniatowski, 2011). We find proper names using Liner2 (Marcińczuk et al., 2011). Finally, the text is chunked (Maziarz et al., 2011b).

All possible mentions have to be annotated in the text. All pronouns are considered to be mentions and those AgPs which heads are on the list of possible mention keywords. Such list is created by IKAR during the learning process. Finally, the resolution process can be initiated.

PN-PN Resolution For PN possible antecedents are PNs that appeared previously. After the classification is done it is possible that one mention was classified to have more than one antecedent (which in fact may be the same entity).

All mentions that are classified to have an antecedent are being processed in order starting from the beginning of the text. If a mention refers to only one antecedent then it is checked if that antecedent refers to any other word. If so then the relation is rerouted to that word which is thought to be the first occurrence of this entity in the text. Now, any already processed mention points to the first occurrence of the entity.If a mention refers to more than one antecedent it is checked if it refers to the same entity. If there are more than one entities - an entity with greater number of references is chosen. If all of them are referenced by the same number of relations the one that occurred in a text closer to the mention is chosen. At the end of the process every PN is matched with the same entity.

PN-AgP Resolution When PNs are already matched with certain entities the possible PN-AgP relations can be determined. If there are a lot of PNs referring to the same entity the possible antecedent is the one closest to the mention. It is also possible that for one mention more than one PN were classified as antecedents. The Semantic Link score is calculated for each of them and the one with the best score is chosen as an antecedent. If there are two candidates with the same score the one closer to the mention is chosen.

PN-Pronoun Resolution Possible relations between pronouns and PNs are determined the same way as PN-AgP relations. If there is more than one antecedent for a given mention the closest is chosen. However, we allow only one relation for each pronoun. Also PN-AgP relations are already resolved at this point so if a pronoun refers directly to an AgP it is clear to which PN it really refers to.

4 IKAR Evaluation

There are three classifiers dedicated for each relation type. Therefore we evaluate each of them separately. We also use SemEval Scorer for calculating B^3, BLANC and MUC measures. The scorer compares a classified file with a source file. We employ 10-fold cross-validation in both evaluation settings. The ZeroR classifier was used as a baseline.

The F-measure of Weka-based evaluation for C.45 are on average higher by 0.12 pp. than the baseline (we omit detailed results for brevity). The results of Scorer evaluation are shown in the Tab. 1. Also, the results are higher than the baseline. The achieved results are higher than other contemporary systems presented for Polish (Kopeć and Ogrodniczuk, 2012; Ogrodniczuk and Kopeć, 2011a,b). Alas, those results are not directly comparable as the guidelines for annotation of corpora differ and the size of the dataset used in this paper is larger.

Measure	Classifier	Precision	Recall	F-measure
B^3	ZeroR	99.98%	71.34%	83.27%
B^3	C4.5	98.37%	89.81%	93.89%
MUC	ZeroR	0.00%	0.00%	0.00%
MUC	C4.5	95.16%	74.65%	83.67%
BLANC	ZeroR	47.67%	49.99%	48.81%
BLANC	C4.5	94.34%	77.32%	83.61%

Table 1: SemEval evaluation

5 Conclusions and Further Works

In this paper we have presented an Improved Kit for Anaphora Resolution (IKAR) for Polish. The system was evaluated on the data annotated in the KPWr Corpus. The types of anaphora annotated in the KPWr were also described. The evaluation was performed using two independent methodologies. Its outcome indicates that described approaches are promising for the anaphora resolution. We are planning to compare the outcome of our work to GATE's ANNIE IE and other applications developed for Polish.

References

(2010). Annotation of cross-document coreference: A pilot study.

(2010). Automatic content extraction.

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., and Wardyński, A. (2012). Kpwr: Towards a free corpus of polish. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Cornish, F. (1986). *Anaphoric Relations in English and French: A Discourse Perspective*. Croom Helm Ltd.

Halliday, M. A. K. and Ruqaiya, H. (1976). *Cohesion in English*. Longman, London.

Huang, Y. (2010). Coreference: Identity and similarity. In Brown, K., editor, *Concise Encyclopedia of Philosophy of Language and Linguistics*. Elsevier.

King, J. (2010). Anaphora: Philosophical aspects. In Barber, A. and Stainton, R. J., editors, *Concise Encyclopedia of Philosophy of Language and Linguistics*. Elsevier.

Kopeć, M. and Ogrodniczuk, M. (2012). Creating a coreference resolution system for polish. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Larson, R. and Segal, G. (1996). *Knowledge of Meaning: An Introduction to Semantic Theory*. The MIT Press.

Marciniak, M. (2002). Anaphora binding in Polish. Theory and implementation. In *Proceedings of DAARC2002,* Lisbon.

Marcińczuk, M., Stanek, M., Piasecki, M., and Musiał, A. (2011). Rich Set of Features for Proper Name Recognition in Polish Texts. In *Proceedings of the International Joint Conference Security and Intelligent Information Systems, 2011*, Lecture Notes in Computer Science (to be published). Springer.

Matysiak, I. (2007). Information extraction systems and nominal anaphora analysis needs. In *Proceedings of the International Multiconference on Computer Science and Information Technology*.

Maziarz, M., Marcińczuk, M., Piasecki, M., Radziszewski, A., Nowak, J., Wardyński, A., and Wieczorek, J. (2011a). Wytyczne do znakowania koreferencji [guideliness for coreference annotation].

Maziarz, M., Radziszewski, A., and Wieczorek, J. (2011b). Chunking of Polish: guidelines, discussion and experiments with Machine Learning. In *Proceedings of the LTC 2011*.

Mitkov, R. (2003). Anaphora resolution. In *The Oxford Handbook of Computational Linguistics*, chapter 14. Oxford University Press.

Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., and Sotirova, V. (2000). Coreference and anaphora: developing annotating tools, annotated resources and annotation stages. In *Proceedings of DAARC2000*, pages 49–58.

Ogrodniczuk, M. and Kopeć, M. (2011a). End-to-end coreference resolution baseline system for Polish. In Vetulani, Z., editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 167–171, Poznań, Poland.

Ogrodniczuk, M. and Kopeć, M. (2011b). Rule-based coreference resolution module for Polish. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 191–200, Faro, Portugal.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A wordnet from the ground up*. Oficyna wydawnicza Politechniki Wroclawskiej.

Poesio, M. (2004). The mate/gnome proposals for anaphoric annotation (revisited). In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162, Boston.

Poesio, M., Ponzetto, S. P., and Versley, Y. (2010). Computational models of anaphora resolution: A survey.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann.

Radziszewski, A., Maziarz, M., and Wieczorek, J. (2012). Shallow syntactic annotation in the Corpus of Wrocław University of Technolog. *Cognitive Studies*, 12.

Radziszewski, A. and Śniatowski, T. (2011). Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Radziszewski, A. and Śniatowski, T. (2011). A Memory-Based Tagger for Polish. In *Proceedings of the LTC 2011*. Tagger available at http://nlp.pwr.wroc.pl/redmine/projects/wmbt/wiki/.

Stede, M. (2012). *Discourse Processing*. Morgan & Claypool Publishers.

Wroliński, M. (2003). System znaczników morfosyntaktycznych w korpusie ipi pan. *Polonica*, 12:39–55.