Evaluation of Coreference Resolution Tools for Polish from the Information Extraction Perspective

Adam Kaczmarek
Institute of Computer Science
University of Wrocław
Wrocław, Poland
akaczmarek@cs.uni.wroc.pl

Michał Marcińczuk
Department of Computational Intelligence
Wrocław University of Technology
Wrocław, Poland
michal.marcinczuk@pwr.edu.pl

Abstract

In this paper we discuss the performance of existing tools for coreference resolution for Polish from the perspective of information extraction tasks. We take into consideration the source of mentions, i.e., gold standard vs mentions recognized automatically. We evaluate three existing tools, i.e., IKAR, Ruler and Bartek on the KPWr corpus. We show that the widely used metrics for coreference evaluation (B^3, MUC, CEAF, BLANC) do not reflect the real performance when dealing with the task of semantic relations recognition between named entities. Thus, we propose a supplementary metric called PARENT, which measures the correctness of linking between referential mentions and named entities.

1 Introduction

In this paper we approach the problem of coreference resolution and its evaluation metrics. We consider this problem from a slightly different perspective—not as a simple clustering problem, but rather as a problem of extracting information from text. We make an observation that not every mention carries equal amount of information, e.g., when considering a pronoun resolution problem there are usually a few named entities that can be assigned to real world objects and relatively larger amount of pronouns that carry almost no information about the object they are referring to, without resolving the coreference with the named entity. Thus we do not want to treat named entities and pronouns equally as in the case below. We can imagine a document with two named entities, for simplicity each with equal count of n pronouns in gold coreferential clusters:

$$\{\text{Romeo}, \text{he}_1, \text{he}_2, \ldots, \text{he}_n\}$$
$$\{\text{Juliet}, \text{she}_1, \text{she}_2, \ldots, \text{she}_n\}$$

and two possible system responses, one with two pronouns interchanged between coreferential clusters:

$$\{\text{Romeo}, \mathbf{she_1}, \text{he}_2, \ldots, \text{he}_n\}$$
$$\{\text{Juliet}, \mathbf{he_1}, \text{she}_2, \ldots, \text{she}_n\}$$

and the second with the named entities interchanged:

$$\{\mathbf{Juliet}, \text{he}_1, \text{he}_2, \ldots, \text{he}_n\}$$
$$\{\mathbf{Romeo}, \text{she}_1, \text{she}_2, \ldots, \text{she}_n\}$$

According to the measures which do not distinguish between types of mentions and are based only on the similarity of clusters, these two responses are scored equally. However, from information extraction perspective the first answer is almost correct, while the second gives us totally incorrect information about both named entities. Thus we propose a supplementary method to score the performance of coreference resolution systems with respect to different types of mentions.

2 Related Work

We will present here work related to this topic in a two-way manner: first by introducing the coreference evaluation metrics and second describing current tools for coreference resolution for Polish.

2.1 Evaluation

Coreference evaluation is a widely studied problem in the literature. Starting from 1995 with the introduction of the MUC evaluation metric (Vilain et al., 1995) that calculates a score based on

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 24–33,
Hissar, Bulgaria, 10–11 September 2015.

the missing/wrong links between the coreference chains according to a minimal amount of such links needed to be added or removed to transform the system response into the key coreference chains. This approach leads to a counter-intuitive result in the case of merging large chains, when keeping the *recall* equal to 100% and dropping the *precision* only by a small amount independent from the size of improperly merged chains. This metric was followed by the B^3 score (Bagga and Baldwin, 1998) developed as an attempt to address some drawbacks of the MUC evaluation metric. In this metric *precision* and *recall* are calculated as an average score for every mention in the text. This metric, unlike MUC, takes into account singletons but is vulnerable to multiple singletons causing *precision* to increase. To overcome the disadvantages of MUC and B^3, Luo (2005) proposed a metric called CEAF. This metric uses an one-to-one mapping between the gold and the system coreference clusters mapping. The most important feature is that this metric can be considered as interpretable—the score reflects a percentage of mentions assigned to the correct clusters. However, it is still sensitive to the singletons and in some cases the correct links can be ignored. One of the latest metrics is BLANC—a metric based on the Rand index for clustering, which was introduced in the original form by Recasens and Hovy (2011). It focuses on the relations between every single pair of mentions—both coreferential and non-coreferential. The final values of *precision*, *recall* and F_1 are calculated as means of respective values for coreferential and non-coreferential links separately. This metric solves the problem of singletons and takes into account the size of the clusters. In the original form BLANC assumes that the mentions in the gold standard data and in the system response are the same. Luo et al. (2014) proposed a modified version of BLANC, called BLANC-SYS, which can handle imperfect mention recognition. This modification also introduced a joint way of scoring the mention detection in conjunction with the coreference resolution.

Twinless Mentions

Simultaneously to the development of the BLANC metric there were several observations made on the problematic nature of the *twinless* mentions[1] occurring due to imperfect mention detection in end-to-end coreference resolution systems. Cai and Strube (2010) addressed this problem for metrics considering only the coreferential relations between mentions. Additionally, they distinguished *twinless* singletons, which are not connected by any coreferential relation.

Evaluation from Applications Perspective

Holen (2013) made some critical observations on the nature of commonly used evaluation metrics, claiming that the loss of information value—an important factor in the perception of coreference resolution—is not addressed good enough in the current evaluation metrics. Some of the issues with different levels of informativeness of mentions were addressed by Chen and Ng (2013). The main idea was to extend the existing metrics with link weights that would reflect the informativeness of certain types of relations. These enhancements provided a more accurate way of scoring coreference results, however, making them less intuitive and harder to interpret. Tuggener (2014) presented an approach that considers coreference results as mention chains and scores every mention according to whether it has a correct direct antecedent. As an extension of this approach he proposed to consider the relations to the closest preceding nouns, e.g., two pronouns are not really useful for higher level applications of coreference resolution. The final proposition was to determine the so-called *anchor mentions* for each key coreference chain and to measure the score as the harmonic mean of the score for detection of these *anchor mentions* and the score for resolving mentions to *anchor mentions* that were found by the system.

2.2 Coreference Resolution for Polish

For Polish there were several approaches to coreference resolution—we took into consideration three tools implementing different approaches to this problem: a rule-based mention-pair system Ruler (Ogrodniczuk and Kopeć, 2011), a machine learning-based mention-pair system Bartek (Kopeć and Ogrodniczuk, 2012) based on the BART framework (Versley et al., 2008) and a machine learning-based entity-mention system IKAR (Broda et al., 2012a). However, these approaches were based on two different definitions of coreference: IKAR considers the coreference as a relation between a mention and a certain named entity. On

[1] A twinless mention is a mention which occurs only in the gold standard data or in the system response.

the other hand, Ruler and Bartek were designed to resolve the coreference relations between any two mentions.

3 IKAR with a Zero-Anaphora Baseline

The task of zero anaphora resolution in Polish was ignored in most of the studies as a non-trivial problem. To be able to fully compare these algorithms we needed first to implement a method for zero-anaphora resolution in IKAR. We made an approach to prepare a zero-anaphora resolution baseline based on the previous work made in IKAR. The main motivation for this baseline approach is the fact that, as stated by Kaczmarek and Marcińczuk (2015), Polish zero subjects carry at least the same amount of grammatical information as pronouns (gender, number and person), so we can approach the problem of zero-anaphora similarly to the pronoun coreference.

3.1 IKAR Approach to Coreference Resolution

In the current approach IKAR divides the coreference resolution problem into four subcategories of coreferential relations, each pointing to a named entity, but originating from different types of mentions, namely: named entities, agreed noun phrases, personal pronouns and zero subjects. The coreference resolution mechanism for each type (except zero subjects) was originally implemented in IKAR as a C4.5 decision tree classifier[2] (Quinlan, 1993) utilizing different sets of features. The coreference is resolved in entity-mention manner, where discourse entities are introduced by named entities, what means that for each mention we perform a binary classification of pairs consisting of the considered mention and a preceding named entity. In the final step the relations are disambiguated to avoid assigning one mention to many different entities. The disambiguation is based on the number of mentions assigned to given entity and on the distance to the mention.

3.2 Naïve Zero-Anaphora in IKAR

The classifier for recognition of pronoun and zero-anaphora links uses the *pronounlink* features which take into consideration the grammatical agreement (person, number, gender) and consider either a direct coreference relation from the pronoun/zero subject to a named entity or a coreference relation to an agreed phrase that is semantically similar to the named entity. This semantic similarity is calculated using a wordnet[3] distance between the phrase's head and a synset inferred from the type of named entity.

4 PARENT Metric

To address the problem with non-intuitive results from an information extraction perspective, we propose a supplementary measure called PARENT (Performance of Anaphora Resolution to ENTities) that will reflect the amount of correct information returned by a coreference resolution system.

4.1 Defining and Referring Mentions

For the purpose of our scoring metric we introduce concepts of *defining* and *referring* mentions. The *defining* mentions are mentions which we consider as self-defining, i.e., carrying enough information to be identified as real-world objects. The *referring* mentions are those mentions which do not hold this property. All mentions in a document can be divided into two disjoint subsets: *defining* mentions and *non-defining* mentions.

$$M_{all} = M_{defining} \cup M_{non\text{-}defining}$$

$$M_{defining} \cap M_{non\text{-}defining} = \varnothing$$

The *non-defining* mention subset is then defined as a union of *referring* mentions that we are particularly interested in and *ignored* mentions which we do not want to consider in the scoring procedure, for the purpose of scoring different variants of coreference resolution (e.g., pronoun resolution or zero subject coreference resolution in a isolation).

$$M_{non\text{-}defining} = M_{referring} \cup M_{ignored}$$

$$M_{referring} \cap M_{ignored} = \varnothing$$

The split into $M_{defining}$, $M_{referring}$ and $M_{ignored}$ should be made on the basis of some criteria which will be taken as a parameter for the scoring algorithm. The split criteria must be also independent from the gold mention annotation, as it can be applied to the system response as well. For example,

[2] IKAR uses an implementation from the Weka software (Hall et al., 2009).

[3] A wordnet for Polish called Słowosieć (Maziarz et al., 2012) was used.

$\{\underbrace{\text{Romeo}}_{defining}, \underbrace{\text{he}_1, \ldots \text{he}_n}_{referring}, \underbrace{\text{boy, young man} \ldots}_{ignored}\}$

(a) Mention split 1 – noun phrases are ignored.

$\{\underbrace{\text{Romeo}}_{defining}, \underbrace{\text{he}_1, \ldots \text{he}_n, \text{boy, young man} \ldots}_{referring}\}$

(b) Mention split 2 – no mentions are ignored.

Figure 1: Examples of mentions split.

if one want to evaluate the performance of linking pronouns with proper names, then the *defining* set will contain proper names, the *referring* set will contain pronouns and the *ignored* set will contain the remaining mentions (i.e., noun phrases) (see Figure 1a). In another scenario (see Figure 1b) one may want to evaluate the performance of linking non-proper names with proper names. Then, the *defining* set will contain proper names (as in the first example) and the *referring* set will contain all the remaining mentions (i.e., pronouns, noun phrases). The *ignored* set will remain empty.

4.2 Precision and Recall

The existing cluster-based metrics for coreference evaluation do not make distinction between the *defining* and *referring* mentions. However, as shown in section 1, from the perspective of information extraction the links between *referring* and *defining* mentions are much more important than the links between *referring* mentions only. Taking into account this assumption we will define precision and recall as follows.

First, we want to relate the *recall* to finding a relation between a *referring* mention m_r and at least one *defining* mention m_d from the same gold coreferential cluster. We are interested in connecting the *referring* mentions to the proper discourse entities introduced by the *defining* mentions which are coreferential with the *referring* mentions in the gold standard data. This way we infer additional information about the entities based on the context of the *referring* mentions. For that purpose it is sufficient for each *referring* mention m_r to have a coreferential link with only one m_d from its gold standard cluster.

Second, we want the *precision* to reflect the ambiguity of information extracted from the coreference resolution system response, i.e., for a *referring* mention m_r we want to penalize situations when m_r is assigned to a *defining* mention from a cluster which does not contain the m_r. The applied penalty is meant to be proportional to the distinct number of entities (represented by their *defining* mentions) assigned to each *referring* entity. This will address situations when the system returns non-existent coreferential links between either two *defining* mentions or between a *defining* and a *referring* mention. We also want the *precision* and the *recall* to be interpretable in following way:

- *Precision* should indicate the ratio of correct relations between *referring* mentions and entities to all relations between *referring* mentions and entities returned by the system
- *Recall* should indicate the ratio of correct relations between *referring* mentions and entities to all relations between *referring* mentions and entities that are expected to be found basing on the gold standard data.

4.3 Description

Intuitively this metric works on links between two predefined groups of mentions. Additionally we map all the *defining* mentions occurring in the same gold cluster into one entity about which we will extract information based on the coreferential relations with *referring* mentions. We do not want to penalize missing some of the defining mention links in cases when a gold cluster contains multiple defining mentions and relate the score to ambiguity of found links. We also do not want to consider the correctness of links between the *non-defining* mention pairs, because these relations do not give us any valuable information.

A *true positive* (tp) will be a correct relation between a *referring* mention and a *defining* mention from the same gold coreference cluster (redundant relations between the *referring* mention and other *defining* mentions from the same gold cluster will be ignored).

A *false positive* (fp) will be an incorrect relation between a *referring* mention and a *defining* mention from different gold coreference clusters (redundant relations between the *referring* mention and other *defining* mentions from the same gold cluster will be ignored).

A *false negative* (fn) will be a pair of a *referring* mention m_r and a *defining* mention, such that no *defining* mention from the gold coreferential cluster containing m_r are found.

4.4 Formal Definition

Formally we will denote the gold set of clusters as C^{key} and i-th gold cluster C_i^{key} will be defined as follows:

$$C_i^{key} = \{\underbrace{m^i_{d_1} \dots m^i_{d_l}}_{defining}, \underbrace{\underbrace{m^i_{r_1} \dots m^i_{r_n}}_{referring}, \underbrace{m^i_{z_l} \dots m^i_{z_k}}_{ignored}}_{non-defining}\}$$

The gold cluster constitutes an *entity*. We will introduce the notation of equivalence classes with respect to the coreference relations to denote the *entity* that given mention belongs to according to the gold standard clusters:

$$[\![m]\!]^{key} = C_i^{key} \text{ such that } m \in C_i^{key}$$

We will define a gold relation set G as follows:

$$G = \{(m^i_{r_l}, C_i^{key}) | \forall C_i^{key} \in C^{key} \forall m^i_{r_l} \in C_i^{key}$$

This set contains pairs of a *referring* mention and the gold cluster it belongs to, one for each *referring* mention, defining mapping from the mentions to the entities they should indicate.

The system set of clusters will be denoted by C^{sys} and the relation set based on the system response will be defined as follows:

$$\begin{aligned} S \;=\; \{ \;& (m^i_{r_l}, [\![m^i_{d_k}]\!]^{key}) | \; \forall C_i^{sys} \in C^{sys} \\ & \forall m^i_{r_l} \in C_i^{sys} \quad \forall m^i_{d_k} \in C_i^{sys}\} \end{aligned}$$

This set contains pairs of a *referring* mention and an *entity* it indicates, represented by the gold clusters containing *defining* mentions that are marked by the system as coreferential with the *referring* mention.

Then we can define *precision* and *recall* as follows:

$$precision = \frac{TP}{TP + FP} = \frac{|G \cap S|}{|S|}$$

$$recall = \frac{TP}{TP + FN} = \frac{|G \cap S|}{|G|}$$

Twinless Mentions

The PARENT metric is also designed to jointly score mention detection with coreference resolution. The problem of *twinless* mentions is treated like Cai and Strube (2010) did—the *twinless* singletons are removed from both gold and system clusters, as we consider only coreferential links between mentions. *Defining* mentions produced by the system, which are not present in the gold data but they were linked with other mentions, are added to the gold data as singletons. This is done because they can produce *false positives* and they must be added to the gold data in order to be included in the evaluation. In other case, those *false positives* would be ignored. The rest of *twinless* non-singleton mentions are left as they are.

4.5 Specific Case Analysis

Here we discuss some specific cases to illustrate the methodology of PARENT scoring:

- a missing link between *defining* mentions—as long as we can correctly connect referring mention with one *defining* mention it is enough, so these missing links should not have negative impact on neither *precision* nor *recall*;
- a missing link between a *referring* mention and a *defining* mention will decrease the *recall* by a unit value;
- an incorrect link between *defining* mentions referring to different entities (clusters) in the gold standard data—this type of error will decrease the *precision* proportionally to the number of entities represented by *defining* mentions in the system cluster and to the number of *referring* mentions.

  Given a system response cluster C_j^{sys}, for each *referring* mention m_r in this cluster we will increase the *true positives* value for this cluster (tp_j) by one if there is a *defining mention* in this cluster that is coreferential with m_r in the gold standard data and we will increase the value of *true and false positives* for this cluster ($tp_j + fp_j$) by the number of entities. So the final precision for such cluster will be equal to:

  $$\frac{\sum_{m^j_{r_l} \in C_j^{sys}} \mathbb{1}_{\exists m^j_{d_k} \in C_j^{sys}, (m^j_{r_l}, [\![m^j_{d_k}]\!]^{key}) \in G}}{entities_j \times referring_j}$$

  where

  $$entities_j = |\{[\![m^j_{d_k}]\!]^{key} : \forall m^j_{d_k} \in C_j^{sys}\}|$$

  and

  $$referring_j = |\{m^j_{r_l} : \forall m^j_{r_l} \in C_j^{sys}\}|$$

- an incorrect link between a *referring* mention and a *defining* mention will decrease *precision* by a value proportional to the number of entities assigned to this mention by the system being scored (analogously to the previous case);

- an one-cluster solution with only gold mentions will be scored with $recall = 100\%$ (all relations between referring mentions and entities are found) and precision inversely proportional to the number of entities, i.e.:

$$precision = \frac{1}{\#entities} = \frac{1}{|C^{key}|}$$

- an one-cluster solution with invented mentions $I = \{i_1, \ldots i_m\}$ will have lower precision calculated as:

$$precision = \frac{|R|}{(|R| + |I|) \times |C^{key}|}$$

where R is a set of all *referring* mentions from the gold clusters;

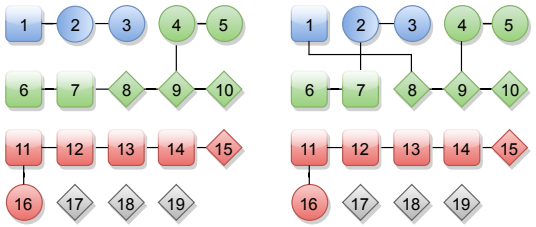
- an all-singleton solution will have both *precision* and *recall* equal to 0.

4.6 The Problem of Split

The PARENT metric is parametrized with the definitions of *defining* and *referring* mentions. This task may occur to be not as easy as it seems due to the fact that it may not be exactly clear how to conclusively describe mentions that are informative enough. Therefore, we left these definitions to be introduced as a parameter to the PARENT metric to allow an introduction of custom definitions of *defining* and *referring* mentions. That possibility is also important for testing only certain parts of coreference resolution systems.

4.7 A Case Study for Metric Comparison

Here we present a case study, which show the advantage of the PARENT metric over other cluster-based metrics. Figure 2 contains a visualization of a gold standard (Figure 2a) and a response returned by a system (Figure 2b). The squares represent *defining* mentions (which are named entities in this case) and the remaining shapes represent *referring* mentions (the circles—pronouns and diamonds—nouns). The blue, red and green color represents groups of mentions referring to the same entity. The system response contains



(a) Gold standard (b) System response.

Figure 2: Examples of mentions split.

incorrect links between $\{2, 3\}$ and $\{6, 7\}$, and $\{4, 5, 8, 9, 10\}$ and $\{1\}$. As can be seen in Table 1 the cluster-based metrics (MUC, B^3, CEAFE, CEAFM, BLANC) scored the response over 70% of F-measure. However, from the perspective of information extraction, the response is not so useful, as most of the *referring* mentions were incorrectly linked with the *defining* mentions—only two (15 and 16) out of nine *referring* mentions were correctly linked with their *defining* mentions. The linguistically aware metrics presented by Chen and Ng (2013) (LMUC, LB^3, LCEAFE, LCEAFM)[4] scored between 30% and 70%—the values are a bit more accurate than their counterparts. According to PARENT the response was scored only 22.2% and the value is much more accurate.

5 Evaluation

We evaluated the following tools for Polish coreference resolution: IKAR, Bartek and Ruler. The results for IKAR were obtained for several different configuration settings. We tested it on the gold standard mentions and on the mentions that were automatically added by simply annotating all the agreed phrases and pronouns, and by using Minos (Kaczmarek and Marcińczuk, 2015) for the detection of zero subject verbs. Bartek and Ruler were tested on the same corpus but with system mentions annotated by their own system for automatic mention annotation, i.e., MentionDetector (Kopeć, 2014). For the evaluation we used 10-fold cross validation on the KPWr corpus (see next section). IKAR was trained for each fold on the training part. For Bartek and Ruler we used the pre-trained models distributed with the tools.

[4]We used the following weights: $w_{nam} = 1$, $w_{nom} = 0$, $w_{pron} = 0$ and $w_{sing} = 10^{-38}$—the weights are set to 0 except for the relations between named entities and other mentions and a small weight of 10^{-38} for singletons. This is the closest configuration to PARENT.

	MUC	B^3	CEAFE	CEAFM	LMUC	LB^3	LCEAFE	LCEAFM	BLANC	PARENT
F_1	80.0%	74.5%	80.8%	78.9%	30.8%	47.4%	66.7%	30.8%	76.6%	22.2%

Table 1: Comparison of different metrics for a sample system response.

5.1 KPWr Corpus

We used a subcorpus of the KPWr corpus (Broda et al., 2012b) version 1.1. It contains 689 documents with a total of 27 452 links (14 141 of them are links other than zero-anaphora). The links were manually annotated between four types of mentions: named entities, agreed nominal phrases, pronouns and zero subjects.

5.2 PARENT Configuration

We used a split, where the set of *defining* mentions contains named entities and the set of *referring* mentions contains nouns, pronouns and zero subjects.

5.3 Impact of Automatic Mention Detection

In the previous study the results of coreference resolution of IKAR were only measured on the gold set of mentions. Here we want to present the impact of the automatic mention detection on the performance of this tool. To simulate the environment with automatically detected mentions we considered as mentions all the hand-annotated agreed noun phrases and all words tagged as personal pronouns using WCRFT tagger (Radziszewski, 2013) and used Minos (Kaczmarek and Marcińczuk, 2015) to annotate potential zero subjects. The results shown in Table 2 indicate a decrease of precision for coreference resolution with automatically detected mentions—particularly significant is the loss of precision for PARENT metric that is several times higher than for BLANC.

5.4 Modifications due to BLANC-SYS

We performed the evaluation using the reference implementation of the coreference scorer (Pradhan et al., 2014). However, due to the fact that we wanted to measure how these systems are capable of recognizing proper coreferential relations even with imperfect mentions detected—for IKAR we mostly recognize much more mentions than are needed and we can omit only some zero subjects—we use a specific evaluation setting. Namely we compare the system results with the gold standard corpus that is modified by adding all system-invented mentions as singletons. This is done due to the fact that in the most recent version of BLANC-SYS metric we are penalized for finding incorrect *non-coreferential* links either between *twinless* singleton mentions in the system response or connecting them to the gold standard mentions. So basically we are penalized for not finding coreference relations where they do not occur. This is due to the fact that BLANC-SYS is intended to jointly score coreference resolution with mention detection. However for information extraction tasks we do not infer any information from singleton mentions and we are basically focused on relations between phrases, so such an approach is not suitable from this perspective.

5.5 PARENT and BLANC Result Comparison

In Table 2 we present results of IKAR with different settings of the mention detection and scores for PARENT and BLANC. We show results for IKAR without zero anaphora baseline (*NonZero*) and with it (*All*). The results for *All* and *NonZero* mention settings are however not directly comparable due to the evaluation setting for *NonZero* mentions that scores only the coreferential relations between named entities, agreed noun phrases and pronouns, excluding zero-anaphora. We can also observe that for each configuration we got much lower scores for PARENT than BLANC. That indicates that although the coreference resolution system can recognize partial coreference clusters quite well it does not necessarily mean that the information extracted from its result is as reliable as the BLANC score would indicate. In a real-world scenario, where the mentions must be automatically recognized beforehand, IKAR does not resolve the links between defining mentions and referring mentions properly. Only 11% of the those links are correct. Also the recall drops by more than half from 66% to only 32%. In the context of information extraction for named entities it is a very low result.

5.6 Algorithm Comparison

In Table 3 we present results of these three systems measured with the BLANC and PARENT metrics. The configuration of PARENT metric was similar

Mentions	Metric	Precision	Recall	F_1
Gold NZ	BLANC	69.02%	71.38%	70.11%
Gold NZ	PARENT	34.94%	30.78%	32.73%
NonZero	BLANC	56.12%	70.18%	58.62%
NonZero	PARENT	7.15%	30.26%	11.57%
Gold All	BLANC	69.94%	67.71%	68.73%
Gold All	PARENT	31.10%	33.95%	32.46%
All	BLANC	57.99%	66.35%	60.39%
All	PARENT	11.09%	32.26%	16.50%

Table 2: IKAR results for different settings.

Algorithm	Mentions	Precision	Recall	F_1
IKAR	NonZero	7.15%	30.26%	11.57%
IKAR	All	11.09%	32.26%	16.50%
Bartek	NonZero	13.49%	5.29%	7.60%
Bartek	All	17.67%	4.89%	7.66%
Ruler	NonZero	14.77%	5.10%	7.59%
Ruler	All	14.07%	3.00%	4.95%

Table 3: Evaluation of the tools for coreference resolution for Polish with the PARENT metric.

to this presented in section 5.2 for the *All* mention setting. For the *NonZero* mention setting we excluded from *referring* mentions all zero-anaphora similarly to what was done for BLANC evaluation in section 5.5. The lower results for Bartek and Ruler can be explained by the fact that these algorithms were not tuned to recognize relations to named entities.

6 Conclusions

We faced the fact that the current state-of-the-art coreference metrics do not take into account various level of mention informativeness. To deal with this problem we introduced a new metric called PARENT[5] that is designed to measure the ability of coreference resolution system to retrieve information about entities in the text. In contrast to the enhanced metrics presented by Chen and Ng (2013), PARENT is not as generic, however, it gives intuitive and interpretable results for given kinds of coreference relations. PARENT is also independent from the number of the correct/incorrect *defining* mentions and from the size of clusters, while these metrics are influenced by size of clusters as well as by counts of the *defining* mentions. In comparison to the approach presented by Tuggener (2014), PARENT is not constrained by the assumption that the coreferential relations must be interpreted either as relations to the closest preceding noun or to a single *anchor mention* for a cluster what makes it more robust in case of imperfect mention detection. PARENT also seems to be more generic by allowing a flexible definition of *defining* and *referring* mentions. The main difference between PARENT and the other metrics is that PARENT treats all *defining* mentions from a *gold* cluster as one object and does not require more than one relation between a *referring* mention and such an object that can be as set of *defining* mentions. Being aware of some drawbacks of PARENT method (e.g., the score does not reflect reliably the coreference resolution quality between defining mentions) we will advise to use it as a complementary score for one of state-of-the-art metrics for scoring coreference systems.

The results for coreference resolution for Polish reported in the literature were optimistic. However, when dealing with an information extraction task, where the linking between defining mentions and referring mentions is much more important than between referring mentions only, the performance drops significantly. The best results we obtained were 17.67% of precision for the Bartek system and 32.26% of recall for IKAR measured using the proposed metric PARENT. This shows, that for information extraction tasks oriented on named entities, like recognition of semantic relations between named entities (Marcińczuk and Ptak, 2012), the performance of coreference resolution systems for Polish needs a significant improvement.

Acknowledgement Work financed by the Polish Ministry of Science and Higher Education, a program in support of scientific units involved in the development of a European research infrastructure for the humanities and social sciences in the scope of the consortia CLARIN ERIC and ESS-ERIC, 2015-2016. One of the authors is receiving Scholarship financed by European Union within European Social Fund.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Bartosz Broda, Łukasz Burdka, and Marek Maziarz. 2012a. IKAR: An Improved Kit for Anaphora Resolution for Polish. In Martin Kay and Christian

[5]The PARENT metric evaluation was implemented as a part of Liner2 toolkit (Marcińczuk et al., 2013).

Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 25–32. Indian Institute of Technology Bombay.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012b. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen Chen and Vincent Ng. 2013. Linguistically aware coreference evaluation metrics. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1366–1374.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Gordana Ilic Holen. 2013. Critical reflections on evaluation practices in coreference resolution. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *HLT-NAACL*, pages 1–7. The Association for Computational Linguistics.

Adam Kaczmarek and Michał Marcińczuk. 2015. Heuristic algorihtm for zero subject detection in Polish (to be published). In *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence. Springer Berlin / Heidelberg.

Mateusz Kopeć and Maciej Ogrodniczuk. 2012. Creating a Coreference Resolution System for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 192–195, Istanbul, Turkey. ELRA.

Mateusz Kopeć. 2014. 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*, pages 221–225, Gothenburg, Sweden. Association for Computational Linguistics.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard H. Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 24–29.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *In Proc. of HLT/EMNLP*, pages 25–32. URL.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2—A Customizable Framework for Proper Names Recognition for Polish. In Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, Marzena Kryszkiewicz, and Marek Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 231–253. Springer.

Michał Marcińczuk and Marcin Ptak. 2012. Preliminary Study on Automatic Induction of Rules for Recognition of Semantic Relations between Proper Names in Polish Texts. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 7499 of *Lecture Notes in Computer Science*, pages 264–271. Springer Berlin Heidelberg.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference*, Matsue, Japan, January.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011. 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.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard H. Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 30–35.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In H. Rybiński M. Kryszkiewicz M. Niezgódka R. Bembenik, Ł. Skonieczny, editor, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

Marta Recasens and Eduard H. Hovy. 2011. BLANC: implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 231–235, Gothenburg, Sweden, April. Association for Computational Linguistics.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 9–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.