# Evaluation of Uryupina's coreference resolution features for Polish[1]

**Bartłomiej Nitoń**

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
bartek.niton@gmail.com

**Abstract**

Automatic coreference resolution is an extremely difficult and complex task. It can be approached in two different ways: using rule-based tools or machine learning. This article describes an evaluation of a set of surface, syntactic and anaphoric features proposed in Uryupina 2007 and their usefulness for coreference resolution in Polish texts.

**Keywords:** Uryupina, machine learning, coreference resolution, Polish language, surface features, syntactic features, anaphoricity and antecedenthood, BART

## 1. Introduction

Olga Uryupina's PhD thesis "Knowledge Acquisition for Coreference Resolution" (Uryupina 2007) describes over 350 linguistic features which can be used to recognize coreference. Since they are considered language-independent, we intend to verify this statement by checking the impact of a certain subset of features on coreference resolution for Polish.

Uryupina's classification of features is based on: surface similarity; syntactic knowledge; semantic compatibility; discourse structure and salience; anaphoricity and antecedenthood.

This paper concentrates on surface similarity, syntactic information, as well as anaphoricity- and antecedenthood-related features.

## 2. Features

This section describes features implemented and examined during research. They are grouped in accordance with Uryupina's classification. For example usage of presented configurations and more precise descriptions of them refer to (Uryupina 2007).

### 2.1. Surface similarity features

Co-referring descriptions frequently have similar surface form. Strings can be simplified, partially modified or kept intact. Surface similarity features are therefore based mostly on comparing mentions or their specified fragments.

In our study, we implemented and examined about 88 surface similarity features described in Uryupina's thesis. The thesis decomposes surface similarity problem into three sub-tasks: **normalization**, specific **substring selections** and **matching** proper.

**Normalization** covers different spellings of same name throughout a text, such as "MCDONALD'S" and "McDonald's", obviously referring to the same name. Uryupina describes three normalization functions: *no_case*, *no_punctuation* and *no_determiner*. The first function ignores case in strings, the second one strips off all punctuation marks and other auxiliary characters (like "-" or "#"), while the last one strips off determiners from text. During our research, the function *no_determiner* was ignored due to inapplicability of its direct definition in Polish (which lacks articles and displays a complex linguistic model of definiteness).

**Substring selection** covers the fact that some words in a mention are more informative and important than other ones. Therefore, instead of matching whole strings, one can compare only their most valuable, representative fragments. Uryupina describes four key words of a mention string: *head*, *last*, *first* and *rarest* word in a mention string[2].

The last sub-task, **matching**, is based on a string comparison function. Uryupina describes five string comparison algorithms:

- *exact match*, comparing whole mention strings;
- *approximate_match*, which is based on the minimum edit distance (MED) measure (Wagner and Fisher 1974); because MED does not take into consideration the length of a string, minimum edit distance is normalized by the length of anaphor or antecedent; length normalizations are marked as length_s or length_w;
- *matched_part*, counts overlap between strings in words or symbols;
- *abbreviation*, one of four abbreviation algorithms, in our experiment limited to two: *abbrev1* takes the

[1] The study was cofounded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00 and the *Computer-based methods for coreference resolution in Polish texts* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

[2] For the purpose of checking word rarity we used 1-grams extracted from the balanced subcorpus of the National Corpus of Polish (Przepiórkowski et al. 2012) to create two word frequency lists of orthographic word forms and their base forms.

initial letter of all the words in the mention string, produces an abbreviated word out of them and compares the created string with the head of the second mention; *abbrev2* algorithm works in the same way but ignores words starting with lowercase characters for building an abbreviated word (e.g. *abbrev1* would change the string "Federal Bureau of Investigations" into "FBoI" whereas *abbrev2* would change it into "FBI");

- *rarest(+contain)*, finds the rarest word in a mention string and checks if it occurs in some other mention.

In our experiment, we have implemented all surface features described in Uryupina's PhD thesis excluding ones using *no_determiner* normalization and using more complex abbreviation algorithms (*abbrev3* and *abbrev4*). All of them have been implemented in BART, a modular toolkit for coreference resolution (Versley et al. 2008), supplemented with a Polish language plugin.

In further experiments, we decided to use Uryupina's original configurations, i.e.:

- **all**: all 88 implemented surface features;
- **baseline1**: exact match for full names only, without use of normalization;
- **baseline2**: baseline1 features and head exact matching without normalization (triple: *no_normalization*, *head*, *exact_match*);
- **MED+head**: baseline1 and all approximate match features (triple: _, _, *approximate_match*);
- **MED-head**: baseline1 features and approximate match algorithms without substring selection (triple: _, *no_substring_selection*, *approximate_match*);
- **MED_w-head**: baseline1 and minimum edit distance (MED) measured in words features (*MED_w*, *MED_w_anaph*, *MED_w_ante* etc., *no_substring_selection*);
- **MED_s-head**: baseline1 and minimum edit distance (MED) measured in symbols features (*MED_s*, *MED_s_anaph*, *MED_s_ante* etc., *no_substring_selection*);
- **MED_bare-head**: baseline1 and all minimum edit distance without MED length normalizations and substring selection (*MED_s*, *MED_w* etc., *no_substring_selection*);
- **MED_ante-head**: baseline1 and all MED features with normalization by antecedent length and without substring selection (*MED_s_ante*, *MED_w_ante* etc., *no_substring_selection*);
- **MED_anaph-head**: baseline1 and all MED features with normalization by anaphor length and without substring selection (*MED_s_anaph*, *MED_w_anaph* etc., *no_substring_selection*);
- **Last**: baseline1, exact match for full names (triple: _, *no_substring_selection*, *exact_match*) and exact match for last word in mentions (triple: _, *last*, *exact_match*);
- **First**: baseline1, exact match for full names (triple: _, *no_substring_selection*, *exact_match*) and exact match for first word in mentions (triple: _, *first*, *exact_match*);
- **Rarest**: baseline1, exact match for full names (triple: _, *no_substring_selection*, *exact_match*) and rarest word-based features (triples: _, *rarest*, *exact_match* and _, *rarest*, *contain*), each rarest feature is implemented for base forms of words and text forms;
- **No_MED**: all implemented features without approximate match features;
- **No_abbrev**: all implemented features without *abbrev1* and *abbrev2*-based features;
- **No_rarest**: all features without rarest word-based ones;
- **No_rarest_parser**: all features without the rarest word-based ones and features using parsing (i.e., all types of matching except for abbreviation and head matching algorithms).

For each of the configurations presented above, different normalization strategies were used. We distinguished five types of possible normalization strategies: *no_normalization*, *no_case*, *no_punctuation*, *full normalization* and *all normalizations together*. *Full normalization* involves only *no_case+no_punctuation* features, while *all normalizations together* involves all specified features with normalization, e.g. all features with *no_case* and *no_punctuation* normalization and also features containing both normalizations (*no_case+no_punctuation*).

## 2.2. Syntactic knowledge features

Though none of the existing approaches rely solely on syntactic information, it is considered to be a valuable part of anaphora resolution algorithms. While Uryupina presents about 61 different syntactic features in her thesis, due to time constraints, we have taken into account only the 9 core syntactic features for the purpose of our research. Enlisted syntactic features implemented in BART coreference resolution system (Versley et al. 2008) during research for Polish coreference resolution are provided below:

- **Post-modification** (features: *postmodified(Mi)*, *postmodified(Mj)*): checks whether the markable is a syntactic construction where the head is not the last word.
- **Number** (features: *number(Mi)*, *number(Mj)*): checks the grammatical number of the anaphor or the antecedent.
- **Person** (features: *person(Mi)*, *person(Mj)*): checks the grammatical person of the anaphor or the antecedent.
- **Same number** (features: *same_number(Mi,Mj)*): checks if the anaphor and the antecedent share the same number.
- **Same person** (features: *same_person(Mi,Mj)*): checks if the anaphor and the antecedent share the same person.
- **Syntactic agreement** (features: *synt_agree(Mi,Mj)*): checks if the anaphor and the antecedent share the same number and person.

The last 5 configurations from the above list may also be considered morphological agreement features. By definition, all markables in a coreference chain refer to the same object, thus they should share the number and person categories.

### 2.4. Anaphoricity and antecedenthood

Anaphoricity- and antecedenthood-related features are responsible for discovering how likely it is that a given mention is an antecedent of another mention.

Features for discovering anaphoricity have been divided by Uryupina into six groups: surface, syntactic, semantic, salience, *same-head*, (Karttunen 1976)-motivated features (apposition, copula, negation, modal constructions, determiner, grammatical role and semantic class).

Surface, syntactic and salience features have already been presented in this paper while the evaluation of semantic features and Karttunen-motivated factors has been postponed for the time being. In current research, we have implemented *same-head* features. The *same-head* feature group consists of Uryupina's *same_head_exists(Mi)*, *same_head_exist(Mj)*, *same_head_distance(Mi)*, *same_head_distance(Mj)* features. They represent coreference knowledge on a very basic level. *Same_head_exist* checks if there is a mention with same head as given in the preceding text, *same_head_distance* describes distance between given markable and one with the same head in the preceding text.

## 3. Evaluation

Following i.a. CONLL-2011 (Pradhan et al. 2011), for evaluation, we used an average score of MUC (Vilain et al. 1995), $B^3$ (Bagga and Baldwin 1998) and CEAFE

| Configuration | no | no_case | no_punctuation | full | all |
|---|---|---|---|---|---|
| **all** | **0.72** | **0.72** | **0.72** | **0.72** | **0.72** |
| baseline1 | 0.69 | **0.70** | 0.69 | **0.70** | **0.70** |
| baseline2 | **0.69** | **0.69** | **0.69** | **0.69** | **0.69** |
| MED+head | **0.70** | **0.70** | **0.70** | **0.70** | **0.70** |
| MED-head | **0.71** | **0.71** | **0.71** | **0.71** | **0.71** |
| MED_w-head | 0.69 | **0.70** | 0.69 | **0.70** | 0.69 |
| **MED_s-head** | **0.72** | **0.72** | **0.72** | **0.72** | **0.72** |
| MED_bare-head | 0.70 | 0.70 | 0.70 | 0.70 | **0.71** |
| **MED_ante-head** | **0.72** | **0.72** | **0.72** | **0.72** | **0.72** |
| **MED_anaph-head** | **0.72** | **0.72** | 0.71 | **0.72** | **0.72** |
| last | 0.69 | **0.70** | 0.69 | **0.70** | **0.70** |
| first | 0.69 | **0.70** | 0.69 | **0.70** | **0.70** |
| **rarest** | **0.72** | **0.72** | **0.72** | **0.72** | **0.72** |
| no_MED | **0.71** | **0.71** | **0.71** | **0.71** | 0.70 |
| **no_abbrev** | **0.72** | **0.72** | **0.72** | 0.71 | 0.71 |
| no_rarest | **0.70** | **0.70** | **0.70** | **0.70** | **0.70** |
| no_rarest_parser | **0.70** | **0.70** | **0.70** | **0.70** | **0.70** |

Table 1: Different surface similarity configurations, the classifier's performance (average F-score for $B^3$, MUC and CEAFE measures) in 10 fold cross-validation on the 390 files sample from Polish Coreference Corpus.

| Configuration | CEAFM | CEAFE | MUC | B3 | average |
|---|---|---|---|---|---|
| **all** | 0.75 | 0.80 | **0.52** | 0.83 | **0.72** |
| baseline1 | 0.77 | **0.82** | 0.43 | **0.86** | 0.70 |
| baseline2 | 0.75 | 0.80 | 0.44 | 0.84 | 0.69 |
| MED+head | 0.74 | 0.78 | 0.49 | 0.83 | 0.70 |
| MED-head | 0.76 | 0.81 | 0.48 | 0.85 | 0.71 |
| MED_w-head | 0.77 | **0.82** | 0.43 | **0.86** | 0.70 |
| **MED_s-head** | 0.77 | **0.82** | 0.49 | 0.85 | **0.72** |
| MED_bare-head | 0.77 | **0.82** | 0.44 | **0.86** | 0.70 |
| **MED_ante-head** | 0.77 | **0.82** | 0.49 | **0.86** | **0.72** |
| **MED_anaph-head** | **0.78** | **0.82** | 0.49 | **0.86** | **0.72** |
| last | 0.77 | **0.82** | 0.43 | **0.86** | 0.70 |
| first | 0.77 | 0.81 | 0.43 | **0.86** | 0.70 |
| **rarest** | 0.76 | **0.82** | 0.50 | 0.84 | **0.72** |
| no_MED | 0.74 | 0.80 | 0.50 | 0.83 | 0.71 |
| **no_abbrev** | 0.75 | 0.80 | **0.52** | 0.83 | **0.72** |
| no_rarest | 0.74 | 0.78 | 0.50 | 0.83 | 0.70 |
| no_rarest_parser | 0.75 | 0.80 | 0.48 | 0.83 | 0.70 |

Table 2: F-scores for different classifiers, different variants of configuration and *no_case* normalization, the *average* column describes average value of B3, MUC and CEAFE metrics, best results and configurations are marked with bold font.

(Luo 2005) metrics which track influence of different coreference dimensions (the $B^3$ measure being based on mentions, MUC on links and CEAFE based on entities); we will also present CEAFM (Luo 2005) metric for consideration. In order to train coreference decisions, tests were performed with J48, WEKA's (Witten and Frank 2005) implementation of the C4.5 decision tree learning algorithm (Quinlan 1993) and weka classifier, which uses WEKA machine learning toolkit for classification. As data for learning, we used a fragment of the Polish coreference corpus built within the CORE project (Computer-based methods for coreference resolution in Polish texts). As training data, we used 390 texts from the Polish coreference corpus (see http://zil.ipipan.waw.pl/PolishCoreferenceCorpus and Ogrodniczuk et al. 2013).

### 3.1. Surface similarity features

Table 1 presents results of coreference resolution for Polish using various definitions of surface feature configurations, with each configuration evaluated using all presented types of normalizations. Table 2 presents a set of coreference scores for different measure methods and *no_case* normalization.

The finding here is that using normalizations for Polish coreference resolution can result in slight, but not very noticeable increase. Best score is obtained for *no_case* normalization. An interesting conclusion results from the worst setting, *no_punctuation* (worse even than the score for no normalization), which can indicate that in Polish, punctuation can in some cases help resolve coreferring pairs of markables. But it is apparent that proper normalization does not significantly increase coreference resolution results in Polish.

From the configuration point of view, the best score is acquired for **all**, **MED_s-head**, **MED_ante-head**, **MED_anaph-head**, **rarest** and **no_abbrev** configurations. The average F-score approaches 0.72 and

| Configuration (F-score) | CEAFM | CEAFE | MUC | B3 | average |
|---|---|---|---|---|---|
| all | 0.75 | 0.80 | 0.52 | 0.83 | 0.72 |
| MED_s-head | 0.77 | **0.82** | 0.49 | 0.85 | 0.72 |
| MED_ante-head | 0.77 | **0.82** | 0.49 | **0.86** | 0.72 |
| MED_anaph-head | **0.78** | **0.82** | 0.49 | **0.86** | 0.72 |
| rarest | 0.76 | **0.82** | 0.50 | 0.84 | 0.72 |
| no_abbrev | 0.75 | 0.80 | 0.52 | 0.83 | 0.72 |
| syntaktyczne | 0.71 | 0.77 | 0.00 | 0.83 | 0.53 |
| all + synt | 0.75 | 0.80 | **0.53+** | 0.84+ | 0.72 |
| MED_s-head + synt | 0.76- | 0.80- | 0.48- | 0.84- | 0.71- |
| MED_ante-head + synt | 0.77 | **0.82** | 0.49 | 0.85- | 0.72 |
| MED_anaph-head + synt | 0.77- | 0.81- | 0.49 | 0.85- | 0.72 |
| **rarest + synt** | 0.77+ | **0.82** | 0.51+ | 0.85+ | **0.73+** |
| no_abbrev + synt | 0.74- | 0.79- | 0.52 | 0.83 | 0.72 |
| same_head | 0.71 | 0.77 | 0.00 | 0.83 | 0.53 |
| all + same_head | 0.61- | 0.66- | 0.45- | 0.72- | 0.61- |
| MED_s-head + same_head | 0.71- | 0.77- | 0.44- | 0.81- | 0.67- |
| MED_ante-head + same_head | 0.71- | 0.76- | 0.44- | 0.81- | 0.67- |
| MED_anaph-head + same_head | 0.73- | 0.78- | 0.45- | 0.82- | 0.68- |
| rarest + same_head | 0.76 | **0.82** | 0.50 | 0.84 | 0.72 |
| no_abbrev + same_head | 0.61- | 0.66- | 0.45- | 0.72- | 0.61- |
| synt + same_head | 0.72 | 0.78 | 0.07 | 0.83 | 0.56 |
| all + synt + same_head | 0.57- | 0.62- | 0.45- | 0.68- | 0.58- |
| MED_s-head + synt + same_head | 0.68- | 0.74- | 0.44- | 0.78- | 0.65- |
| MED_ante-head + synt + same_head | 0.70- | 0.76- | 0.45- | 0.80- | 0.67- |
| MED_anaph-head + synt + same_head | 0.69- | 0.75- | 0.44- | 0.79- | 0.66- |
| rarest + synt +,same_head | 0.74- | 0.80- | 0.49- | 0.83- | 0.71- |
| no_abbrev + synt + same_head | 0.57- | 0.61- | 0.45- | 0.68- | 0.58- |

Table 3: F-score for different coreference resolution metrics and best surface features configurations alone or combined with syntactic (*synt*) and *same-head* features. The *average* column describes average value of B3, MUC and CEAFE metrics. Best results and configurations are marked with bold font, minus and plus signs are marking whether selected configuration is better or worse than the one using surface features only (used normalization is *no_case* normalization).

this result corresponds to proper normalization for **MED_anaph-head** and **no_abbrev**, while for the rest of configurations it is reached despite of normalization.
All implemented surface features are used by the **all** profile while **no_abbrev** uses most of them – which can point to the reason why they obtain the highest score.
As can be seen, a slight score increase (by 0.03) is obtained when rarest words are used. Most of MED-based features specially with normalization usage (in this case normalization is understood as division by anaphor or antecedent length in signs or words) also works very well so as minimum edit distance based on signs which is better than one based on words.
What is interesting, configurations using head words obtain slightly lower scores than those not using it. This may be caused by a large number of different orthographic forms in Polish. In further research, it should also be checked how those features would work when they take into consideration base forms of words Those forms can be received using a morphological analyzer called Morfeusz (Woliński 2006, see also http://sgjp.pl/morfeusz/).

### 3.2. Syntactic and same-head features

Table 3 presents coreference scores for configurations using syntactic and *same_head* features combined with surface ones. As can be seen, the only configuration with a score higher than 0.72 (obtained using only surface features) is the one based on rarest words and syntax (**rarest + synt** configuration). It can be said that rarest features are very good predictors of coreference in the Polish language (rarest configuration gives satisfying score even with *same_head* features) and it cooperates very well with syntactic features. For other configurations, syntactic features do not provide any advantage, or even lower the coreference resolution score. Configurations using *same_head* affect coreference in a very negative way. Also, using only syntactic information, *same_head* or even both of those feature groups at the same time does not produce satisfying results. The conclusion is that surface similarity features are indispensable in coreference resolution for Polish and no sufficient score is likely to be obtained with higher-level features only.

## 4. Conclusions

Other groups of features presented by Uryupina, especially discourse-based ones, are currently being revised. Because all features are implemented and evaluated in BART, coreference resolution toolkit, the endpoint would be resolving the best feature configuration for Polish coreference resolution and, based on that, creating an end-to-end Polish coreference resolution system getting raw Polish text as input.
For now, the best score was obtained for a combination of rarest word-based surface features and a couple of implemented syntactic ones.
Even at this point it can be said that Uryupina's features are mostly language-independent (as she claimed in her PhD thesis) and the ones described in this paper excluding *same_head* work very well also for Polish.

## References


Bagga A., Baldwin B. (1998). *Algorithms for scoring coreference chains. The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566.

Karttunen, L. (1976). Discourse referents. In: J. McKawley (Eds.), *Syntax and Semantics*, Volume 7, pp. 361–385. Academic Press.

Luo X. (2005). *On coreference Resolution Performance Metrics. HLT '05 Proceedings*. pp. 25– 32.

Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., Zawisławska M. (2013). *Interesting Linguistic Features in Coreference Annotation of an Inflectional Language*. M. Sun, M. Zhang, D. Lin, H. Wang (Eds.): *12th China National Conference on Computational Linguistics (12th CCL) and the 1st International Symposium on Natural Language Processing based on Naturally Annotated Big Data (1st NLP-NABD)*, LNCS 8202, pp. 97–108. Springer, Berlin–Heidelberg.

Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., Xue N. (2011). CoNLL-2011 Shared Task: *Modeling Unrestricted Coreference in OntoNotes. Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL 2011),* Portland, Oregon.

Przepiórkowski A., Bańko M., Górski R. L., Lewandowska-Tomaszczyk B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Quinlan J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA.

Uryupina O. (2007). *Knowledge Acquisition for Coreference Resolution*. PhD thesis.

Versley Y., Ponzetto S. P., Poesio M., Eidelman V., Jern A., Smith J., Yang X., Moschitti A. (2008). *BART: Modular Toolkit for Coreference Resolution. Proceedings of the ACL-08: HLT Demo Session (Companion Volume)*, pp. 9–12.

Vilain M., Burger J., Aberdeen J., Connolly D., Hirschman L. (1995). *A model theoretic coreference scoring scheme. Proceedings of the 6th conference on Message understanding, MUC6 '95*, pp. 45–52, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wagner R. A., Fisher M. J. (1974). *The string-to-string correction problem*. Journal of the ACM 21(1), pp. 168–173.

Witten I. H., Eibe F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, Cal., 2nd edition.

Woliński M. (2006). *Morfeusz — a practical tool for the morphological analysis of Polish*. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (Eds.) *Proceedings of the Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pp. 503–512. Springer-Verlag, Berlin, 2006.