Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-153

Coreference Resolution in Latvian

Artūrs ZNOTIŅŠ[1]
Institute of Mathematics and Computer Science, University of Latvia

Abstract. Coreference resolution (CR) is a current problem in natural language processing (NLP) research and it is a key task in applications such as question answering, text summarization and information extraction for which text understanding is of crucial importance. This paper describes a work in progress for improving Latvian coreference resolution that includes further experiments with the rule based *LVCoref* system, enlarging existing coreference corpus and the first efforts to adapt machine learning methods. *LVCoref* system now reaches 58.0% F-score using predicted mentions and 76.5% F-score if gold entity mentions are used.

Keywords. Coreference resolution, rule based, machine learning, corpus

Introduction

Coreference resolution is the task of grouping all the mentions of entities in a document into coreference chains so that all the mentions in a given chain refer to the same discourse entity.

While today most state-of-the-art coreference resolvers use machine learning, many coreference relations can be resolved using relatively simple rules. Recent work has shown that rule based approaches can even outperform machine learning models [1, 2]. It is not hard to create a simple coreference system that is based on simplistic surface level features. The situation gets more complicated with further improvements (external resource adaptations) to keep the system efficient and simple.

The Latvian language is currently under-resourced language, with a limited range of language processing tools, resources and limited earlier research on coreference resolution [3, 4]. The main aim of this paper is to further develop the author's created rule based system (*LVCoref)* [4] to create a baseline for further experiments including more linguistically specific aspects (e.g., zero anaphoras and event coreferences). To achieve this, the author inspects individual components (e.g., mention identification and scoring standardization) and optimizes the annotation guidelines to improve quality of the coreference corpus. This paper also describes first efforts to adapt statistical CR system for Latvian.

1. Data Set

The coreference corpus consists of 20 documents from broadcast news (statistics are given in Table 1) that were manually annotated using *MMAX* annotation tool [5]. The

[1] Corresponding Author: Artūrs Znotiņš, 8, Riga, LV-1002; E-mail: arturs.znotins@lumii.lv

analysis of the previous version of this corpus revealed that annotator disagreements were mainly caused by differently marked mentions, indefinite mentions that represent general concepts and inattention (missed mentions and coreference links). For this reason, the guidelines were clarified to reduce the influence of subjectivity. In addition, the corpus was annotated with information about enumerations, temporal expressions and some information about event coreferences (pronouns referring to events) and pleonastic *it*. Singletons (except named entities) are not annotated.

Approximately 30% of the whole data set was annotated twice by different annotators in order to measure inter-annotator agreement (see Table 2). Although inter-annotator agreement results (79.9% averaged F-score and 74.8% chance corrected Cohen's κ) are comparable to other research [6], there is room for improvement.

Table 1. Coreference data set statistics.

Number of documents	20
Number of sentences	958
Number of words	17,372
Number of mentions	1,263
Definite nominal mentions	556
Indefinite nominal mentions	480
Pronominals	227
Number of coreference chains	209
The average length of coreference chains	5.0

Table 2. Coreference data set inter-annotator agreement.

Measure	**F1**	**P**	**R**
MUC	87.5	84.7	90.4
B^3	76.7	66.7	90.1
Pairwise	75.6	69.4	82.8
AVG	79.9	73.6	87.8

2. Pre-processing

The coreference resolution system relies on morphosyntactic information produced by the following tools:

1. A statistical morphological tagger which achieves 97.9% accuracy for part of speech recognition and 93.6% for the full morphological feature tag set that includes case, gender, number, person and more fine grained information [7].

2. Syntactic parsing is done by a parser [8] based on *MaltParser* toolkit [9] and the hybrid dependency-based annotation model used in the Latvian Treebank [10]. The parser achieves 74.63% *UAS* (unlabeled attachment score).
3. Named entities are recognized with a CRF-based NER tool trained for Latvian that *provides* annotation of person names, geographic locations, organizations, media types and product names. NER currently reaches 84.6% F-score [4].

3. Mention Identification

The CR system identifies a set of predicted mentions from text automatically annotated with syntactical and named entity information. The system extracts three types of mentions: proper mentions from named entity chunks, pronominal and nominal mentions from the largest possible noun phrase for noun headword. Mentions can contain nested mentions.

The main aim of mention identification step is to identify as much of gold mentions as possible (currently recall of mentions is 91.2%). This generates a large number of spurious mentions, but these mentions can be rejected later.

The error analysis revealed that mention identification is one of the most important factors that affects CR performance. Parser and named entity tagger errors in earlier analysis stages are the main cause of incorrect mention identification.

Currently, the system uses a simple blacklist of idiomatic phrases to filter out certain non-mentions, e.g., pleonastic *it* in phrases like "*it means*".

4. Evaluation

The results were evaluated against an unweighted average of three coreference resolution metrics (MUC [6], B^3 [11] and CEAF [12]) in two settings (using gold mentions or predicted mentions) using version 7 of the official *CoNLL* scorer[2].

A naïve head match, where all mentions with the exact same head are linked together, is used as a baseline. The results (see Table 3) for this baseline are surprisingly good, showing string similarity is one of the most important features for coreference resolution.

Table 3. The evaluation results of the baseline.

	Gold mentions			Predicted mentions		
	F1	**P**	**R**	**F1**	**P**	**R**
MUC	66.6	94.2	51.5	58.3	75.9	47.3
B3	54.9	91.9	39.1	46.9	70.0	35.3
CEAF	54.5	66.7	46.0	45.5	55.5	38.5
AVG	58.6	84.3	45.5	50.2	67.1	40.4

[2] Available: https://code.google.com/p/reference-coreference-scorers

5. Rule Based System

LVCoref is a rule based CR system that uses a knowledge rich approach and entity centric model that encourages sharing of information across all mentions that point to the same real-world entity similarly to [2]. This method is based on a very simple idea: to apply rules one at a time from the highest to the lowest precision. In this way, rules that are further during resolution process are able to use information about already linked mentions that are more likely to be coreferent.

The system consists of many primitive rules (e.g., mention NER category, gender agreement) that considers two mentions and the coreference chains already linked with them. These primitive rules are used to construct larger rules (e.g., appositive or nominative predicative construction) that can be included in the rule set. For each of these rules, a filter function needs to be specified that goes through every mention and finds mention that should be linked to it (e.g., the closest mention in the syntax tree that satisfies the constraints of the rule). This architecture offers a simple way to experiment, as it provides a lot of freedom.

System uses 4 simple rule sets:

- exact string match (using mention normalization);
- precise constructions (includes appositives, nominal predicates, acronyms);
- head match (using entity level attribute agreement; also includes variants of person's name);
- pronoun anaphora (using mention compatibility and algorithm similar to Hobss' [13], considering 3 previous sentences).

Mention compatibility is based on the information about their represented coreference chain. Two mentions are acknowledged to be coreferent based on their morphological features (gender, number and case), syntactic constraints (one does not dominate another, *i-within-i* [1]), semantic category and the shared attributes of their represented mention chains.

Currently, if given gold mentions, *LVCoref* outperforms the baseline by 17.9 pp, but using predicted mentions by 7.8 pp (see Table 4).

Table 4. The evaluation results of the rule based system.

	Gold mentions			**Predicted mentions**		
	F1	**P**	**R**	**F1**	**P**	**R**
MUC	84.1	88.2	80.3	68.2	69.7	66.7
B3	82.9	90.6	76.4	76.0	79.4	72.8
CEAF	67.3	87.8	54.5	55.8	62.8	50.2
AVG	78.1	88.9	61.8	66.6	70.7	57.7

Each rule set increases performance by increasing recall and slightly decreasing precision (see Table 5). Surface string similarity (exact string match and strict head match) gives the largest increase in performance. Using gold mentions pronoun

anaphora resolution gives much larger improvement, but gold mention setup already includes information about anaphoricity of mentions that is hard task by itself.

Table 6 shows the effect of used components and features on performance of the system by testing performance with and without it. Exclusion of feature means that mentions are always compatible considering this feature (e.g., gender mismatch is allowed by removing gender information). Importance of syntactical information is tested by using simple baseline with flat dependency structure (arcs between all two proceeding word tokens). These results are similar in related research [2].

Table 5. The cumulative performance of the rule based system as rule sets are incrementally added.

	MUC			B3			CEAF			AVG		
	F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
Predicted mentions												
Exact match	48.1	85.9	33.4	35.3	84.4	22.3	36.7	52.4	28.2	40.0	74.2	28.0
+ Precise construction	52.4	84.1	38.0	40.9	82.1	27.2	45.3	57.1	37.5	46.2	74.5	34.3
+ Strict head match	62.2	75.4	52.9	52.3	68.1	42.5	54.0	64.5	46.4	56.2	69.4	47.3
+ Pronouns	65.2	72.3	59.4	54.5	64.2	47.3	54.2	58.9	50.2	58.0	65.1	52.3
Gold mentions												
Exact match	50.8	98.9	34.2	37.4	99.0	23.0	40.4	60.1	30.4	42.8	86.0	29.2
+ Precise construction	57.0	98.6	40.1	45.4	98.4	29.5	50.5	64.1	41.6	45.4	98.4	29.5
+ Strict head match	72.2	95.5	58.1	62.6	93.3	47.1	65.5	76.9	57.1	66.8	88.5	54.1
+ Pronouns	84.4	94.4	76.3	71.9	91.1	59.4	73.1	75.5	70.9	76.5	87.0	68.9

Table 6. Effect of individual components on performance of the system.

	$\Delta F1_{avg}$	ΔP_{avg}	ΔR_{avg}
Gender	0.3	1.0	-0.1
Number	0.9	3.1	-0.6
Type	0.4	4.2	-2.3
Syntactical constraints	0.1	0.4	-0.1
Entity level constraints	0.7	3.2	-1.1
Named entity labels	1.1	0.6	1.4
Syntactical information	2.3	0.3	3.5

6. Machine Learning

Durret and Klein report that a high-performance coreference system is attainable with a small number of feature templates that use only surface-level information [14]. This coreference system uses a simple set of features in a discriminative learning framework. System places a distribution over possible choices of antecedents (one of the previous mentions) or to introduce new entity with a log-linear model. During learning, system

optimizes for conditional likelihood augmented with a parameterized loss function weighting different kinds of errors differently. The surface feature set only considers properties of mentions and mention pairs (mention type, complete, semantic head, the first and last word, the word immediately preceding and the word immediately following a mention, mention length, number of sentences and mentions between mentions) and conjunctions of these features. These large numbers of lexicalized and data-driven features implicitly model linguistic phenomena such as definiteness and centering.

In order to successfully adapt this system for Latvian there were several modifications that included:

- addition of information about token lemmas;
- integration of rule based mention identification model;
- addition of mention normalization;
- integration of information produced by preprocessing tools.

For training of the coreference model, the whole data set is divided into training and test sets, applying 5-fold cross-validation. The results of the adapted system (see Table 7) are only slightly better than baseline, but learning curve (see Figure 1) shows that larger training set should improve performance of the system.

Table 7. The evaluation results of the adapted system using machine learning approach.

	Gold mentions			**Predicted mentions**		
	F1	**P**	**R**	**F1**	**P**	**R**
MUC	83.4	80.3	86.8	60.8	62.7	60.6
B3	50.5	40.5	67.7	48.0	45.9	52.2
CEAF	47.0	62.6	38.0	42.1	48.7	38.1
AVG	62.6	61.1	64.2	51.3	52.4	50.3

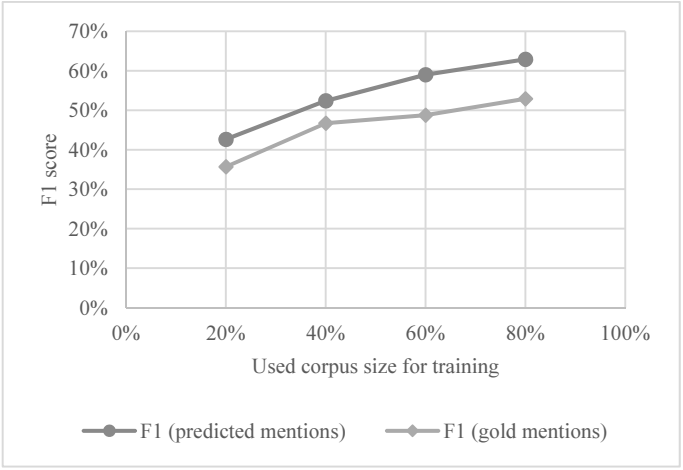

Figure 1. Learning curve.

7. Conclusion and Further Work

In this work the performance of the existing rule based system (used as a part of a larger system for newswire text analysis and fact extraction) is increased by approximately 3 percentage points (58% F1-score using predicted mentions and 76.5% using gold mentions). The author has improved the quality and the size of Latvian coreference corpus by improving annotation guidelines and adding extra information. Results of rule based coreference system are comparable with ones recently achieved for linguistically similar languages [15, 16, 17] and other languages [18]. Additionally, the author adapts statistical CR system for Latvian, but the size of the training set currently limits the achieved results.

Further planned work includes increasing the size of Latvian coreference corpus and experiments with event coreferences.

LVCoref along with annotated data is publicly available at *github.com/chaosfoal/LVCoref.*

8. Acknowledgements

The research leading to these results has received funding from the research project "Information and Communication Technology Competence Center" of EU Structural funds, contract nr. L-KC-11-0003, signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.7 "Creation of the New Information Archive Access Product based on Advanced NLP".

References

[1] D. Klein and A. Haghighi. Simple coreference resolution with rich syntactic and semantic features, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[2] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu and D. Jurafsky. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 2011.

[3] G. Bārzdiņš, N. Grūzītis, G. Nešpore, B. Saulīte, I. Auziņa and K. Levāne–Petrova. Multidimensional Ontologies: Integration of Frame Semantics and Ontological Semantics, *Proceedings of the XIII Euralex Internacional Congress*, 2008.

[4] A. Znotins and P. Paikens. Coreference Resolution for Latvian, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014*), 2014.

[5] C. Müller and M. Strube. Multi-level annotation of linguistic data with MMAX2, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 2011.

[6] M. Vilain, J. Burger, J. Aberdeen, D. Connolly and L. Hirschman. A model-theoretic coreference scoring scheme, *MUC*, 1995.

[7] P. Paikens, L. Rituma and L. Pretkalniņa. Morphological analysis with limited resources: Latvian example, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 2013.

[8] L. Pretkalniņa and L. Rituma. Statistical syntactic parsing for Latvian, *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 2013.

[9] A. Lavelli, J. Hall, J. Nillson and J. Nivre. MaltParser at the EVALITA 2009 Dependency Parsing Task, *Proceedings of EVALITA 2009*, 2009.

[10] G. Bārzdiņš, Grūzītis, N. G. N. and B. Saulīte, Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order, *Proceedings of the 16th Nordic Conference of Computational Linguistics*, 2007.

[11] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains, *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference,* 1998.

[12] X. Luo. On coreference resolution performance metrics, *Proceedings of HLT-EMNLP 2005*, 2005.
[13] J. R. Hobbs. Resolving Pronoun References. *Readings in Natural Language*, 1976.
[14] G. Durret and D. Klein. Easy Victories and Uphill Battles in Coreference Resolution, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013.
[15] I. Goenaga, O. Arregi, K. Ceberio, A. D. de Ilarraza and A. Jimeno. Automatic Coreference Annotation in Basque, *11th International Workshop on Treebanks and Linguistic Theories,* 2012.
[16] M. Kopeć and M. Ogrodniczuk. Creating a Coreference Resolution System for Polish, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.
[17] M. Novák and Z. Žabokrtský. Resolving Noun Phrase Coreference in Czech, *8th Discourse Anaphora and Anaphor Resolution Colloquium*, 2011.
[18] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio and Y. Versley. SemEval-2010 Task 1: Coreference Resolution in Multiple Languages, *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010.