Coreference Resolution for Latvian

Artūrs Znotiņš, Pēteris Paikens
University of Latvia, Faculty of Computing and Institute of Mathematics and Computer Science
Raina bulvaris 29, Riga, LV-1459, Latvia
E-mail: arturs.znotins@gmail.com, peteris@ailab.lv

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. We describe an implementation of coreference resolution tools for Latvian language, developed as a part of a tool chain for newswire text analysis but usable also as a separate, publicly available module. *LVCoref* is a rule based CR system that uses entity centric model that encourages the sharing of information across all mentions that point to the same real-world entity.
The system is developed to provide starting ground for further experiments and generate a reference baseline to be compared with more advanced rule-based and machine learning based future coreference resolvers. It now reaches 66.6 F-score using predicted mentions and 78.1% F-score using gold mentions.
This paper describes current efforts to create a CR system and to improve NER performance for Latvian. Task also includes creation of the corpus of manually annotated coreference relations.

Keywords: coreference resolution, rule based, entity centric model

1. 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 (van Deemter & Kibble, 1999). For example, given the following text (mention borders are marked with square brackets)

> *[Latvietis$_1$] [Jānis Bērziņš$_1$] ir [jauns zinātnieks$_1$] un [universitātes profesors$_1$]. [Profesors$_1$] ir veicis nozīmīgus pētījumus datorlingvistikā kopā ar [profesoru$_2$] [Pēteri Kalniņu$_2$]. [Viņš$_1$] kopā ar [savu$_1$] [līdzgaitnieku$_2$] [Kalniņu$_2$] uzstāsies konferencē Itālijā.*
>
> *[Latvian$_1$] [Jānis Bērziņš$_1$] is a [new scientist$_1$] and [professor at university$_1$]. The [professor$_1$], together with [professor$_2$] [Pēteris Kalniņš$_2$], have carried out important research in computer linguistics. [He$_1$], together with [his$_1$] [associate$_2$] [Kalniņš$_2$], will speak in the conference in Italy.*

the task is to group the mentions so that those referring to the same entity are placed together into a coreference chain (represented with same subscripted index).

Latvian is an under-resourced language, with a limited range of language processing tools and resources, and very limited earlier research on coreference resolution (Bārzdiņš et al., 2008). We believe that the described system is the first available implementation of coreference resolution for Latvian language.

Nowadays most coreference systems use knowledge rich features that require extra preprocessing. Typically coreference resolution requires following steps:

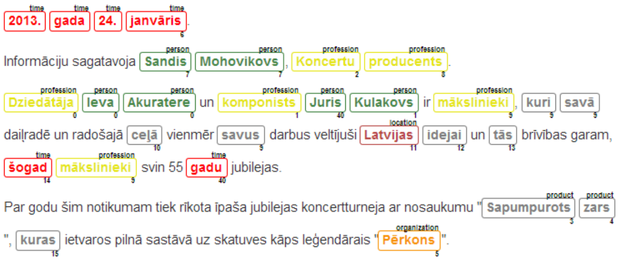
- identification of tokens and sentences;
- part of speech tagging;
- parsing;
- named entity recognition;
- mention identification;
- coreference resolution.

While today most state-of-the-art coreference resolvers use machine learning (Witten, Hall & Frank, 2011), many coreference relations can be resolved using relatively simple rules and recent work has shown that rule based approach can outperform machine learning models for coreference resolution (Haghighi & Klein, 2009; Lee et al., 2011). In this paper we have investigated these approaches and describe our implementation as adapted to Latvian language. In addition, we describe the changes to existing named entity recognition solutions aimed at better coreference resolution accuracy.

2. Proposed Solution

2.1. Entity Centric Model

LVCoref uses an entity centric model which allows each coreference decision to be globally informed by previously created coreference chains. It allows to use global constraints, e.g., linking two mentions is not allowed if it creates attribute disagreement ("*Jānis Bērziņš*" and "*Pēteris Bērziņš*" linked together by a common surname). It also diminishes distance between potential coreferent mentions if the closest same entity mention cannot be correctly linked with current mention based on local features.



2013. gada 24. janvāris.

Informāciju sagatavoja Sandis Mohovikovs, Koncertu producents.

Dziedātāja Ieva Akuratere un komponists Juris Kulakovs ir mākslinieki, kuri savā daiļradē un radošajā ceļā vienmēr savus darbus veltījuši Latvijas idejai un tās brīvības garam, šogad mākslinieki svin 55 gadu jubilejas.

Par godu šim notikumam tiek rīkota īpaša jubilejas koncertturneja ar nosaukumu "Sapumpurots zars", kuras ietvaros pilnā sastāvā uz skatuves kāps leģendārais "Pērkons".

Figure 1: Automatic coreference annotation with *LVCoref*.

2.2. System Description

LVCoref base module integrates other modules described here, and handles input/output formatting and used rules according to a configuration file. An evaluation module uses MMAX (Müller & Strube, 2006) format gold coreference links.

The rule module contains available rule sets. These rules are created by combining features from the feature module.

2.3. Pre-processing

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

1. The initial step is a statistical morphology tagger which achieves 97.9% accuracy for part of speech recognition and 93.6% for the full morphological feature tag set (Paikens, Rituma & Pretkalniņa, 2013).
2. Syntactic parsing is done by a parser (Pretkalniņa & Rituma, 2013) based on MaltParser toolkit (Lavelli et al., 2009) and the hybrid dependency-based annotation model used in the Latvian Treebank (Bārzdiņš et al., 2007). The parser is based on dependency grammar approach achieving 72% precision.
3. In addition, we identify mentions of named entities with a CRF-based NER tool trained for Latvian that provides annotation of person names, geographic locations and organizations, media types, product names. NER currently reaches 84.6% F-score.

2.4. Annotated Corpus

For evaluation purposes we manually annotated 6 interviews (see section 3.1 for used data set statistics) with coreference information. The evaluation data was encoded in MMAX format and featured 3 layers: the word layer, the sentence layer and the coreference layer.

5 mention categories were annotated in this corpus – person, location, media, organization and other (mentions that did not fit in other categories).

2.5. Named Entity Recognition

Before this project, there were two available NER systems for Latvian: *TildeNER* (Pinnis, 2012) and *LVTagger* (Paikens et al., 2012) both based upon the Stanford NER condition random field (CRF) classifier (Finkel, Grenager & Manning, 2005). For the purposes of this research we chose to adapt *LVTagger*, extending it with additional training data for modern news language.

Our chosen taxonomy consists of 7 types of NE (person, location, organization, product, media, sum and time). Nested expressions are not tagged as separate NEs, taking in account the longest NE. E.g., whole phrase "*Latvijas Republikas Finanšu un Veselības ministrijas*" (organization) is marked as one entity without marking "*Latvijas Republikas*" (organization) as another entity.

The named entity annotated corpus (45 000 words, 2 500 sentences) consists of manually annotated news articles (see Table 1). The corpus can be considered rather small when compared to CoNLL corpora which have over 300 000 tokens (Tjong Kim Sang & Meulder, 2003). While CoNLL corpora uses 4 NE types, *LVTagger* introduces 7 types, which makes the data sparser and therefore the NER task harder.

The standard CoNLL metric is used, where the output NE is considered correct only if its span and type is exactly the same as the span and type in the gold data.

We have improved gazetteer features for multiword expressions. We have increased F-score by about 4% by introducing distributional similarity features. For this we used unlabelled 83 million token corpus to create 200 similar world clusters as suggested by Faruqui and Pado (2010). Experiments with syntactic features (phrase head information) introduced only slight improvement. To atomically extract high quality gazetteers we use semantic database of NE's and frames which is constantly augmented with data from processed news articles.

Table 2 lists current results for NER accuracy.

Entity type	Count
location	910
media	63
organization	851
person	512
product	99
sum	245
time	301

Table 1: Named entity corpus statistics.

Entity type	F1	P	R
location	86.9	84.2	89.9
media	77.2	95.1	65.0
organization	74.0	77.5	70.9
person	86.8	89.1	84.6
product	14.0	39.3	8.5
sum	94.1	97.3	91.2
time	88.3	92.7	84.4
totals	84.6	91.0	79.1

Table 2: NER evaluation results.

2.6. Identification of entity mentions

To resolve coreferences, one must first detect the mentions that are going to be linked in coreference chains. Mention identification finds pronouns, common nouns, and named entity mentions. In general, we choose the largest possible noun phrase for the particular head word, based on the sentence syntactic analysis. For example, in phrase "*Kultūras ministrija*" ("*the Ministry of Culture*") only the whole phrase "*Kultūras ministrija*" is marked as a mention and not "*ministrija*". Mentions can be nested, e.g., *"[[Latvijas Nacionālā teātra] direktors]*" (*"the [director of the [Latvian National Theatre]]").*

We also use a whitelist of known entity head words and acronyms, and a blacklist of idiomatic phrases to filter out certain non-mentions, e.g., pleonastic *"tas"* (*"it"*) in phrases like *"tas nozīmē"* ("*it means*").

2.7. Coreference Module

The method is based on applying rules one at a time from the highest to lowest priority, thus in deciding whether two mentions refer to the same real entity. In this way, system can also consider information about the related mentions joined together in previous steps.

2.7.1. Exact string match

This rule links two named entity mentions only if they contain exactly the same text by comparing lemmatized phrases.

2.7.2. Precise constructions

This rule set links two mentions if any of the conditions below is satisfied:

- Appositive. Standard Haghighi and Klein (2009) definition to detect appositives is used: one mention is dependent on another, e.g., *"[profesors$_1$] [Jānis Bērziņš$_1$]"* (*"[professor$_1$] [Jānis Bērziņš$_1$]"*).
- Predicative nominative are in a subject-object relation being dependent on same verb "*būt*" ("to *be*"), e.g., "[*Jānis Bērziņš$_1$] ir [pasniedzējs$_1$*]" (*"[Jānis Bērziņš$_1$] is a [professor$_1$]"*).
- Acronym – mentions are linked if one of them equals the sequence of upper case characters in the other mention, e.g., "*Ekonomikas ministrija*" and "*EM*".

2.7.3. Strict head match

Two mentions are linked based on naive matching of their head words, if the second mention does not introduce new entity attributes, e.g., "*Latvijas Republikas Augstākā tiesa*" ("*the Supreme Court of the Republic of Latvia*") and *"Latvijas Augstākā tiesa"* ("*the Supreme Court of Latvia*") are considered coreferent but "*Latvijas Augstākā tiesa*" and *"Krievijas Augstākā tiesa" ("the Supreme Court of Russia"*) are not because the latter one introduces new attribute "*Russia*".

2.7.4. Pronoun anaphora

Pronoun antecedents are searched in three previous sentences using Hobbs' algorithm (1976).

Mention compatibility is based on the information about their represented coreference chain. Two mentions are acknowledged as coreferent based on their morphological features (gender, number and case), syntactic constraints (one does not dominate another, i-within-i (Haghighi & Klein, 2009)), semantic category and their represented mention chains shared attributes.

3. Results and Evaluation

3.1. Data Set

Evaluation data came from created coreference corpus using all 6 annotated interviews. Data statistics are listed in Table 4.

Two of six interviews (~30% of whole data set) were annotated twice by different annotators in order to measure inter-annotator agreement. We measured it in the same way as *LVCoref* was evaluated against gold standard (see Table 3).

Analysis of inter-annotator disagreements showed that majority of errors resulted from accidental annotation mistakes and completely missed coreference chains and mentions. Although inter-annotator agreement results (76.2 % averaged F-score) are comparable to other research (Vilain et al., 1995), there is room for improvement by optimizing annotation guidelines.

	F1	P	R
MUC	84.3	84.3	84.3
B^3	73.7	84.5	66.0
Pairwise	70.7	83.1	61.4
AVG	76.2	84.0	70.6

Table 3: Coreference data set inter-annotator agreement

Number of documents	6
Number of sentences	778
Number of words	13 768
Number of mentions	1 088
Number of coreference chains	333
Number of singleton mentions	180
The average length of the coreference chain	3.27

Table 4: Coreference data set statistics.

3.2. Baseline

As a baseline for evaluation we use the naive head match method, linking only mentions with the exact same head. More sophisticated resolution models have been suggested, but they are rarely compared with this baseline, admitting that it performs better than expected. For the MUC-7 test data Soon's system (Soon, Ng & Lim, 2001) outperforms head match only by 5%, while Uryupina's system (2007) outperforms baseline by 15%.

3.3. Evaluation

	Gold mentions			Predicted mentions		
	F1	P	R	F1	P	R
Baseline						
MUC	62.6	65.1	60.3	56.6	56.5	56.7
B^3	74.2	80.3	69.0	71.2	74.9	67.7
Pairwise	59.3	79.5	47.3	53.6	63.8	46.2
AVG	65.4	75.0	54.5	60.5	65.1	53.4
LVCoref						
MUC	84.1	88.2	80.3	68.2	69.7	66.7
B^3	82.9	90.6	76.4	76.0	79.4	72.8
Pairwise	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

Table 5: Coreference resolution evaluation results.

System was evaluated against three coreference resolution metrics: pairwise, MUC (Vilain et al., 1995) and B^3 (Bagga & Baldwin, 1998) in two settings (using gold mentions or predicted mentions).

MUC is a link based metric which measures how many predicted and gold mention chains need to be merged to cover gold and predicted clusters respectively.

B^3 is a mention based metric which measures the proportion of overlap between predicted and gold mention clusters for a given mention.

	MUC			B^3			Pairwise			AVG		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Predicted mentions												
Exact match	52.0	87.3	37.0	73.7	96.9	59.4	47.4	97.1	31.4	57.7	93.8	40.7
+ Precise construction	55.0	86.7	40.3	74.7	96.3	61.0	47.8	94.8	31.9	59.2	92.6	41.6
+ Strict head match	65.1	70.6	60.4	75.7	81.9	70.3	53.8	63.7	46.5	64.8	72.1	54.4
+ Pronouns	68.2	69.7	66.7	76.0	79.4	72.8	55.8	62.8	50.2	66.6	70.7	57.7
Gold mentions												
Exact match	53.2	92.8	37.3	74.1	98.1	59.5	47.6	98.7	31.4	58.3	96.5	40.8
+ Precise construction	56.4	91.9	40.7	75.2	97.5	61.2	48.6	96.4	32.5	60.1	95.3	42.1
+ Strict head match	74.2	88.4	64.0	80.6	92.8	71.2	61.8	88.8	47.4	72.2	90.0	55.3
+ Pronouns	84.1	88.2	80.3	82.9	90.6	76.4	67.3	87.8	54.5	78.1	88.9	61.8

Table 6: Cumulative performance as rule sets are added.

The output has the results for each of the three metrics mentioned earlier, both in terms of precision and recall, as well as F-score.

Table 5 lists the performance of the system. Given gold mentions *LVCoref* outperforms baseline by 12.7%, but using predicted mentions by 6.2%.

Table 6 illustrates the performance of the system as the 4 rule sets are incrementally added. Each successive rule set increases system performance by increasing recall and slightly decreasing precision. With respect to individual contributions, this analysis highlights two significant performance increases: exact string match and strict head match. It illustrates that a large percentage of mentions in text are repetitions of previously mentioned entities based on string similarity. Precise constructions give only a slight performance increase because they are relatively infrequent.

3.4. Error Analysis

To understand the errors in the system, we analyzed two documents from evaluation set and categorized them into the following distinct groups:

1. Non-anaphoric constructions. Identifying whether noun phrase is nested mention or part of the stable construction is not a trivial task, e.g., "*Aktieru zāle*" is stable construction and "*Aktieru*" is a non-anaphoric construction.
2. Indefinite noun phrases. Latvian does not explicitly distinguish definite and indefinite nouns, so it is unclear if mention with same head introduces a new entity or refers to a previous mention, e.g., "*Privatizācijas aģentūra*" and "*aģentūra*".
3. Morphological tagging/disambiguation errors. E.g., singular mentions "*šuvēja*" and "*šuvējas*" ("*tailor*") are not linked together because of incorrect grammatical number identification (equal singular genitive and plural nominative forms).
4. Syntactic errors make it difficult to find appositive and predicative nominative constructions.
5. Pronoun anaphora resolution. Demonstrative pronoun "*tas*" ("*it*") often refers to event mention, e.g., "*plānot*" ("*to plan*"). This system currently does not mark event mentions, thus missing all mentions that are verbal phrases.

Another considerable source of errors is caused by insufficient semantic information, e.g., when the plural personal pronoun "*mēs*" ("*we*") is used to refer to an organization in an interview.

4. Conclusion

The presented approach offers a useful yet easy to implement baseline for further work and is currently the only available coreference resolution system for Latvian. The implementation is currently used as a part of a larger system for newswire text analysis and fact extraction. We also plan to make an evaluation of the impact of coreference resolution precision on the precision of final fact extraction by the end of this year.

The currently achieved F-score – 66.6% using predicted mentions and 78.1% using gold mentions – was satisfactory for use in our text analysis problem and is comparable with results recently achieved for linguistically similar languages (Goenaga et al., 2012; Kopeć & Ogrodniczuk, 2012; Novák & Žabokrtský, 2011) and other languages (Recasens et al., 2010), although their research shows options for future work in improving accuracy. Morphological, syntactic, semantic information and the entity centric model all provide noticeable contribution to coreference resolution performance.

Precision of mention identification is one the most important factors that affects the performance of the end-to-end coreference system. Error analysis revealed that the main problems of coreference resolution are related to non-anaphoric constructions, indefinite noun phrases and pronoun coreference resolution.

Currently planned future work includes machine learning experiments for coreference resolution and incorporating available semantic database knowledge (facts about popular entities) to support high quality gazetteer maintenance for named entity recognition and to help resolve coreferences using global semantic information.

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

5. 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".

This work has been supported by the European Social Fund with the project “Support for Doctoral Studies at University of Latvia” and University of Latvia Faculty of Computing.

We are grateful to the reviewers of this paper for comments which helped us to improve the paper.

6. References

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

Bārzdiņš, G., Grūzītis, N., Nešpore, G., Saulīte, B. (2007). Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, pp. 13--20.

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

Faruqui, M., Pado, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of Konvens 2010*, Saarbrucken, Germany.

Finkel, J. R., Grenager, T., and Manning, C. (2005). *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05), Stroudsburg, PA, USA: Association for Computational Linguistics pp. 363--370.

Goenaga, I., Arregi, O., Ceberio, K., de Ilarraza, A. D., Jimeno, A. (2012). Automatic Coreference Annotation in Basque. In *11th International Workshop on Treebanks and Linguistic Theories*, pp. 115--126.

Haghighi, A., Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1152--1161.

Hobbs, J. R. (1976). Resolving Pronoun References. In *Readings in Natural Language*. Los Altos, California: Morgan Kaufman Publishers, pp. 339--352.

Kopeć, M., Ogrodniczuk, M. (2012). Creating a Coreference Resolution System for Polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 192--195.

Lavelli, A., Hall, J., Nillson, J., Nivre, J. (2009). MaltParser at the EVALITA 2009 Dependency Parsing Task. *In Proceedings of EVALITA 2009*.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 73--79.

Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt a. M., Germany: Peter Lang, pp. 197--214.

Novák, M., Žabokrtský, Z. (2011), Resolving Noun Phrase Coreference in Czech. In I. Hendrickx, S. L. Devi, A. H. Branco & R. Mitkov (Eds.), *8th Discourse Anaphora and Anaphor Resolution Colloquium*, pp 24--34.

Paikens, P., Auziņa, I., Garkāje, G., Paegle, M. (2012). Towards named entity annotation of Latvian National Library corpus. In Human Language Technologies - The Baltic Perspective: *Proceedings of the Fifth International Conference Baltic HLT 2012*, pp. 169--175.

Paikens, P., Rituma, L., Pretkalniņa, L. (2013). Morphological analysis with limited resources: Latvian example. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pp. 267--277.

Pinnis, M. (2012). Latvian and Lithuanian Named Entiy Recognition with TildeNER. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1258--1265.

Pretkalniņa, L., Rituma, L. (2013). Statistical syntactic parsing for Latvian. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 279--289.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., Versley, Y. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 1--8.

Soon, W. M., Ng, H. T., Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. In Computional Linguistics, 27, pp. 521--544.

Tjong Kim Sang, E. F., De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pp. 142--147.

Urypina, O. (2007). *Knowledge Acquisition for Coreference Resolution. PhD thesis.* Saarland University, Saarbrücken, Germany.

van Deemter, K., Kibble, R. (1999). What is coreference, and what should coreference annotation be? In *Proceedings of ACL workshop on Coreference and Its Application*s, pp. 90--96.

Vilain, M. Burger, J., Aberdeen, J., Connolly, D., Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *MUC*, pp. 45--52.

Witten, I. H., Frank, E., Hall, M. A. (2011). *Practical Machine Learning Tools and Techniques.* Amsterdam: Morgan Kaufmann.