CORBON 2017 Shared Task: Projection-Based Coreference Resolution

Yulia Grishina
Applied Computational Linguistics
FSP Cognitive Science
University of Potsdam
grishina@uni-potsdam.de

Abstract

The CORBON 2017 Shared Task, organised as part of the Coreference Resolution Beyond OntoNotes workshop at EACL 2017, presented a new challenge for multilingual coreference resolution: we offer a projection-based setting in which one is supposed to build a coreference resolver for a new language exploiting little or even no knowledge of it, with our languages of interest being German and Russian. We additionally offer a more traditional setting, targeting the development of a multilingual coreference resolver without any restrictions on the resources and methods used. In this paper, we describe the task setting and provide the results of one participant who successfully completed the task, comparing their results to the closely related previous research. Analysing the task setting and the results, we discuss the major challenges and make suggestions on the future directions of coreference evaluation.

1 Motivation

High-quality coreference resolution plays an important role in many NLP applications. However, developing a coreference resolver for a new language requires extensive world knowledge as well as annotated resources, which are usually expensive to create. Previous shared tasks on multilingual coreference resolution, such as the SemEval 2010 shared task on Coreference Resolution in Multiple Languages (Recasens et al., 2010) and the CoNLL 2012 shared task on Modeling Multilingual Unrestricted Coreference in OntoNotes (Pradhan et al., 2012), operated in a setting where a large amount of training data was provided to train coreference resolvers in a fully supervised manner. Our shared task has a different goal: We are primarily interested in a low-resource setting. In particular, we seek to investigate how well one can build a coreference resolver for a language for which there is no coreference-annotated data available for training.

With a rising interest in annotation projection, we focused on a projection-based task, which, in our opinion, could facilitate the application of existing coreference resolution algorithms to new languages. Annotation projection is a technique which allows us to automatically transfer annotations from a well-studied, typically resource-rich language to a low-resource language across parallel corpora. It was first introduced in the pioneering work of Yarowsky et al. (2001), who exploited annotation projection to induce POS taggers, NP chunkers and morphological analysers for several languages. Their approach is illustrated in Fig. 1, which shows automatic transfer of POS tags from English to French via word alignment. Thereafter, annotation projection was successfully applied for different NLP tasks, including coreference resolution (Postolache et al., 2006; Rahman and Ng, 2012; Grishina and Stede, 2015; Martins, 2015).

In the shared task, the participants were offered an automatically labelled source language corpus, which could be used to automatically transfer the annotations to the target side and subsequently train a new system. With English typically being the most well-studied and resource-rich language, we employed it as our source language. To verify the applicability of our projection-based approach to two different languages, we chose German and Russian as our target languages. We believe that, with this exciting setting, the shared task could help promote the development of coreference technologies that are applicable to a larger number of natural languages than is currently possible. In order to test the limitations of our approach and for a fair comparison, we also offered

Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017), co-located with EACL 2017, pages 51–55,
Valencia, Spain, April 4, 2017. ©2017 Association for Computational Linguistics

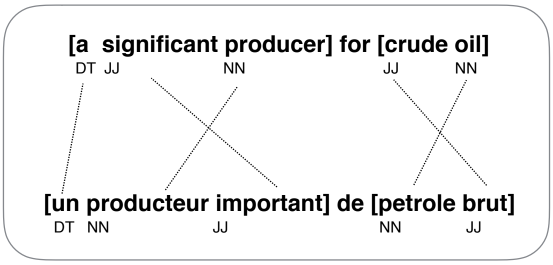
[a significant producer] for [crude oil]
DT JJ NN JJ NN
[un producteur important] de [petrole brut]
DT NN JJ NN JJ

Figure 1: Direct projection algorithm by Yarowsky et al. (2001)

the participants a more traditional setting, where one was supposed to develop a multilingual coreference resolver with no restriction on the resources and methods used.

The paper is structured as follows. Section 2 gives a detailed overview of the task setting. Section 3 describes the participating system and the evaluation results. In Section 4, we analyse the results and compare them to the related work on the topic. Finally, Section 5 presents conclusions.

2 Task setting

The main goal of the CORBON 2017 Shared Task was the evaluation of multilingual coreference resolution in a low-resource scenario. Furthermore, we introduced an open setting in which we did not impose any restrictions on the resources and methods used by the participants in the development of their systems. In sum, the participants competed in two tracks:

- **Closed track:** coreference resolution on German and Russian using annotation projection. In this setting, the participants were allowed to use the English part of the OntoNotes corpus (Hovy et al., 2006) to train a source coreference resolver, or they could use any of the publicly-available coreference resolvers trained on the same data. They could then use whatever parallel corpus and method they prefer to project the English annotations into German/Russian and subsequently train a new coreference resolver on the projected annotations. As for additional linguistic information, the participants could use POS information provided by the parser of their choice.
- **Open track:** coreference resolution on German and Russian with no restriction on the kind of coreference-annotated data the participants can use for training. For instance, they could label their own German/Russian coreference data and use it to train a German/Russian coreference resolver, or adopt a heuristic-based approach where they employ knowledge of German/Russian to write coreference rules for these languages.

Since our main focus was on the low-resource setting, we did not provide any German or Russian manually coreference-annotated data to the participants. Instead, to facilitate system development in the closed setting, the shared task participants were provided an English-German and English-Russian parallel corpora as a training set. Specifically, we chose the English-German and English-Russian parts of the News-Commentary11 parallel corpus[1] taken from the OPUS collection of parallel corpora (Tiedemann, 2012).

The original sentence-aligned text files were split into documents and tokenised using EuroParl tools[2] (Koehn, 2005). The English side of the corpora was labelled automatically using the Berkeley Entity Resolution system (Durrett and Klein, 2014), which was trained on the English part of the OntoNotes corpus (Hovy et al., 2006).

Furthermore, in this setting, the participants were allowed to use other existing parallel texts processed in a similar manner. In the open track, there was no restriction on the data used for system training.

As for the test set, we chose the English-German-Russian parallel corpus described in Grishina and Stede (2015). The guidelines used for the annotation of the corpus are quite compatible with the OntoNotes guidelines for English (Version 6.0) in terms of the types of referring expressions that are annotated (Grishina and Stede, 2016). The exceptions are that they (a) handle only NPs and do not annotate verbs that are coreferent with NPs, (b) include appositions into the markable span and do not mark them as a separate relation, (c) mark relative pronouns as markables, and (d) annotate pronominal adverbs in German if they co-refer with an NP. A sample of the German and Russian annotations was provided to the participants to support their system development. The size of the training and test datasets are presented

[1] http://opus.lingfil.uu.se/News-Commentary11.php
[2] http://www.statmt.org/europarl/

	Training set			Test set		
	#docs	#sents	#tokens	#docs	#sents	#tokens
English	5749	221 844	5 341 828	–	–	–
German	5749	221 844	5 404 568	10	413	8753
English	4869	188 761	4 503 260	–	–	–
Russian	4869	188 761	4 290 891	10	413	8092

Table 1: Size of the training and test datasets

	closed track		official
System	German	Russian	score
CUNI	29.40	30.94	**30.17**

Table 2: Official CORBON 2017 Shared Task results

in Table 1.

The evaluation of the results was conducted in a similar way as in the CoNLL 2012 shared task (Pradhan et al., 2012). We employed three commonly-used scoring metrics | MUC, B-CUBED and CEAF$_e$ | and took the unweighted average of these scores (as computed by the official CoNLL 2012 scorer[3]) to determine the winning system. We did not evaluate singletons and therefore asked the participants to exclude them from their results prior to the submission.

3 CORBON 2017 systems and results

Out of several candidates, only one team successfully completed the task and submitted their results during the official evaluation period. This team consisted of Michal Novák, Anna Nedoluzhko and Zdenek Zabokrtskŷ from Charles University in Prague, Czech Republic. They submitted their results for the closed track, with the following system description:

- **CUNI:** The system submitted by Charles University (CUNI) is a projection-based coreference resolver for German and Russian. It is trained exclusively on coreference relations projected through a parallel corpus from English. The authors used the training corpus and automatic annotation of English coreference as provided by the shared task organizers. Their resolver makes use of multiple models, and each of them addresses a specific anaphoric mention type individually. Furthermore, it operates on the level of deep syntax. The original surface representation of coreference thus must be transferred to this level. Analogously, coreference relations found by their system must be in the end transformed back to the surface representation, in order to be evaluated in accordance with the task's requirements.

The system was assessed by computing the official CoNLL 2012 metric as described above, and the results of the shared task are presented in Table 2.

4 Discussion

The team from Charles University made an important contribution to the task of exploring annotation projection for multilingual coreference resolution. Of particular importance is the development of a projection-based coreference resolver for Russian, which is an under-resourced language in terms of coreference resolution.

The CUNI system achieved CoNLL scores of 29.40 and 30.94 for the German and Russian portions of the official evaluation dataset, respectively. As the authors themselves acknowledge, the model ablation analysis of their system showed that the models for third-person personal and possessive pronouns and NPs contributed the most to overall performance.

The analysis of the resolver's stages showed that while for Russian the resolver trained on the annotations projected from English achieves 66% of the quality achieved by the English resolver (CoNLL score), this number drops to 46% for German (Novák et al., 2017).

A more detailed analysis of Precision and Recall scores showed that, on one hand, the system was able to achieve relatively high average Precision scores[4] (62.5 and 59.56 for German and Russian, respectively). On the other hand, average Re-

[3] https://github.com/conll/reference-coreference-scorers

[4] Average Precision and Recall scores are computed as an unweighted average of MUC, B-CUBED and CEAF Precision and Recall respectively.

call numbers for both languages are considerably lower: 20.3 for German and 21.2 for Russian.

Since it was not possible to compare their results to those obtained in a similar setting, we briefly compare them to the most closely related work on annotation projection for coreference resolution. Firstly, we consider the experimental evaluation of projection method quality conducted by Grishina and Stede (2015) on the same dataset using gold annotations (without system training). Grishina and Stede's results exhibited a similar balance between Precision and Recall scores, where a higher Precision was accompanied by a comparatively lower Recall (P=68.0/82.1 and R=45.8/62.6 for German/Russian). Furthermore, we look at two related studies by Souza and Orăsan (2011) and Martins (2015), who also experimented with cross-lingual training on different languages and datasets, but in a similar projection-based setting. While the former fails to beat a simple baseline that clusters together mentions with the same head[5], the latter achieves F1 scores of 38.82 for Spanish and 37.23 for Portuguese. These performance numbers are slightly higher than the corresponding results for German and Russian.

In sum, the results of the shared task show that a projection-based approach applied to coreference resolution can support creating coreference resolvers even if no manually annotated data is available. In particular, this approach is already able to achieve promising Precision scores, thus providing coreference-annotated data of fair quality. However, the coverage of the projected annotations still requires improvement, which, in our view, could be achieved by using, for instance, a bilingual dictionary or automatically induced paraphrases in order to retrieve missing coreference mentions on the target side.

Another way to improve Recall could be to increase the robustness of mention detection by using multiple source annotations. Specifically, if a coreference mention is absent in the first source language and therefore cannot be projected, it could still be recovered by another source language.[6] Furthermore, the choice of the source language(s) in respect to the target language is also an interesting factor that influences the projection results; however, this issue needs to be investigated further by comparing the quality of a projection approach for different languages in the same setting.

5 Conclusions

In this paper, we presented the CORBON 2017 Shared Task, the first evaluation task on projection-based coreference resolution. The novelty of this task is that it did not provide any manually annotated gold data as the training set, but relied solely upon the automatic annotations obtained by using a state-of-the-art English coreference resolver. The results of the task show that, in this low-resource setting, it is possible to build a new resolver for two different languages with reasonably high Precision scores. Therefore, we conclude that this task can be seen as a fair starting point for projection-based multilingual coreference resolution.

Overall, we believe that this task has successfully continued the important tradition of evaluating state-of-the art coreference systems. Moreover, we hope that it will bring more interest to the task of cross-lingual coreference resolution and will hopefully contribute to the future progress of our field.

The complete data package for the shared task was made available via `https://github.com/yuliagrishina/CORBON-2017-Shared-Task`.

Acknowledgements

The shared task coordinator would like to thank CORBON 2017 workshop organisers – Vincent Ng and Maciej Ogrodniczuk – for providing continuous support and sharing their insights throughout the whole organisation process.

References

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *Transactions of the Association for Computational Linguistics*.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora, Beijing, China*, page 14. Association for Computational Linguistics.

[5] Probably due to erroneous annotations on the source side, as the authors themselves acknowledge.

[6] However, combining coreference annotations coming from several sources is not a trivial task, as shown in Grishina and Stede (2017).

Yulia Grishina and Manfred Stede, 2016. *Parallel coreference annotation guidelines*. Unpublished manuscript.

Yulia Grishina and Manfred Stede. 2017. Multi-source annotation projection of coreference chains: assessing strategies and testing opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, Valencia, Spain, April. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1427–1437, Beijing, China, July. Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdenek Zabokrtskỳ. 2017. Projection-based coreference resolution using deep syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON), Valencia, Spain*. Association for Computational Linguistics.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *Proceedings of 5th international conference on Language Resources and Evaluation (LREC)*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730. Association for Computational Linguistics.

Marta Recasens, Lluís Màrquez, Emili Sapena, M Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

José Guilherme Camargo Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Discourse Anaphora and Anaphor Resolution Colloquium*, pages 59–69. Springer.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of 8th international conference on Language Resources and Evaluation (LREC)*, pages 2214–2218.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.