Lazy Low-Resource Coreference Resolution: a Study on Leveraging Black-Box Translation Tools

Semere Kiros Bitew, Johannes Deleu, Chris Develder and Thomas Demeester

IDLab, Ghent University - imec
{semerekiros.bitew, firstname.lastname}@ugent.be

Abstract

Large annotated corpora for coreference resolution are available for few languages. For machine translation, however, strong black-box systems exist for many languages. We empirically explore the appealing idea of leveraging such translation tools for bootstrapping coreference resolution in languages with limited resources. Two scenarios are analyzed, in which a large coreference corpus in a high-resource language is used for coreference predictions in a smaller language, i.e., by machine translating either the training corpus, or the test data. In our empirical evaluation of coreference resolution using the two scenarios on several medium-resource languages, we find no improvement over monolingual baseline models. Our analysis of the various sources of error inherent to the studied scenarios, reveals that in fact the quality of contemporary machine translation tools is the main limiting factor.

1 Introduction

End-to-end coreference resolution is the task of identifying and clustering all spans of text that refer to the same entity in a document. It serves as an important step for several downstream NLP tasks that involve natural language understanding, including question answering (Morton, 1999), information retrieval, and text summarization (Azzam et al., 1999; Baldwin and Morton, 1998). Recent advances in deep learning have resulted in state-of-the-art performance on coreference resolution (Lee et al., 2017; Fei et al., 2019; Kantor and Globerson, 2019; Joshi et al., 2020; Wu et al., 2020). The performance of these models, however, highly depends on the existence of large annotated datasets. Still, for many languages that lack large annotated coreference corpora, machine translation (MT) tools of an ever increasing quality are available. The idea studied in this work, is whether existing black-box translation tools can be readily leveraged for transferring the task of coreference resolution from one language to another.

We tackle the setting in which a large labeled corpus exists in a resource-rich language (i.e., the 'source' language) whereas only a smaller corpus exists in a smaller-resource language (called the 'target' language). Specifically, we consider two scenarios in which black-box MT tools can be integrated into a cross-lingual end-to-end coreference resolution system. The first scenario, `Translate-train`, uses an MT tool to translate the large source corpus into the target language, after which a coreference model is trained in the target language. In the second scenario, `Translate-test`, test examples in the target language are first machine translated to the source language, after which a pre-trained coreference model is used to predict the labels. The second scenario has the disadvantage that an MT tool is required at inference time.

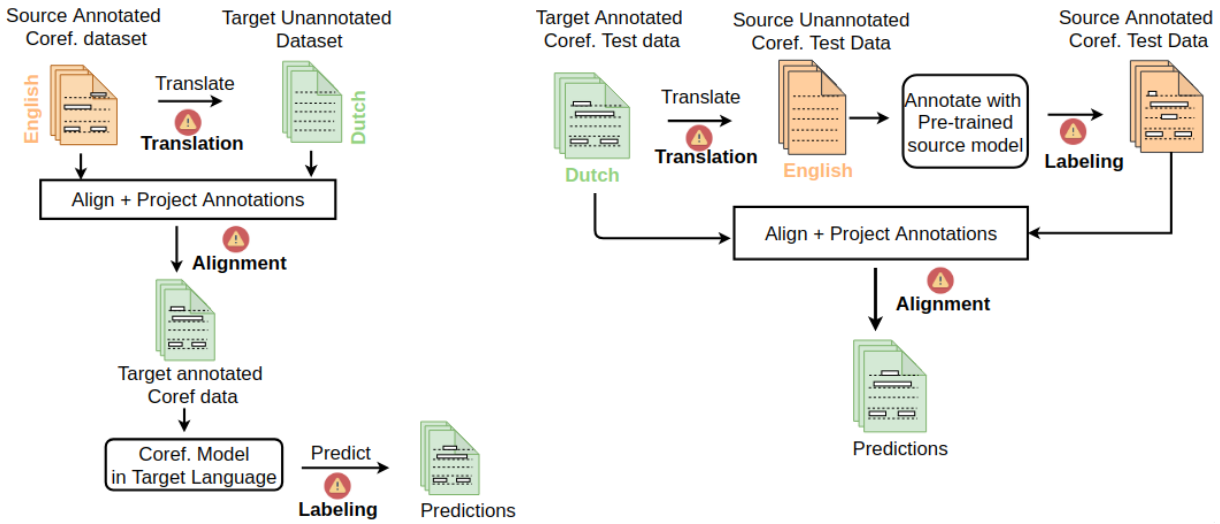
Similar transfer learning setups for basic sequence tagging tasks gave encouraging results (as discussed in Section 4), but we find this is no longer the case for the task of coreference resolution.

We analyze the different sources of error related to integrating the MT tool in the pipeline. As it turns out, translation errors have the strongest impact on the effectiveness of the proposed methods, followed by prediction errors and alignment issues.

2 Approach

2.1 Translate-train

The goal of the `Translate-train` approach (visualized in Fig. 1a) is to create a dataset in the target language, on which a model can be trained. We follow the approach used by Jain et al. (2019) for NER, but we now apply it for coreference resolution.



(a) Translate-train procedure

(b) Translate-test procedure

Figure 1: Annotation projection approaches, with indication of the main sources of error through the ⚠ icon.

We assume access to labeled training data in the source language, a MT tool, an alignment tool, and a test set in the target language. First, we use the MT tool to translate the entire training set from the source to the target language. This results in an unannotated dataset in the target language. Second, we identify and label all mentions of entities in the translated target document by aligning the source and target documents using the alignment tool. Finally, a competitive monolingual method is used to train a coreference model directly in the target language, which is then evaluated on the test data.

2.2 Translate-test

The `Translate-test` approach (see Fig. 1b) follows Shah et al. (2010), and assumes access to a large training corpus in the source language, an off-the-shelf MT system, an alignment tool, and a test set in the target language. First, the test set is translated into the source language. A competitive model trained on the source language training corpus is used to annotate the translated test set (i.e., identify and cluster mentions into groups). With the alignment tool, the translated documents in the source language are aligned with the original ones in the target language, after which the predicted labels are projected onto them for evaluation.

3 Experimental Evaluation

Data — Our evaluation set was created for the SemEval-2010 (Recasens et al., 2010) shared task, and contains coreference annotations for six languages (see Table 1 for dataset statistics). We use Dutch, Spanish, Italian, and Catalan as our target languages, and the corresponding SemEval-2010 datasets are used to train and test the respective monolingual coreference models. As our large and high-quality source dataset, we use the English OntoNotes 5.0 coreference dataset from the CoNLL 2012 shared task (Pradhan et al., 2012).

Coreference Models — For the `Translate-train` scenario, we use the end-to-end neural coreference resolution method from Lee et al. (2017) to train and evaluate on the target languages. This model considers all spans of text as potential mentions and finds the most probable antecedents for each span. For each span, a span ranking model is used to decide which of the previous spans are good antecedents, whereby a trained pruner eliminates less likely mentions. During training, the marginal log-likelihood of all correct antecedents in the gold clusters is optimized. In our `Translate-test` experiments, we use `SpanBert` (Joshi et al., 2020), an English end-to-end coreference resolver, trained on the OntoNotes corpus.

Translation Tool — In both scenarios, we use Google Translate[1] as our publicly available MT tool of choice.

Alignment — For the alignment step, we compare Fast-Align from Dyer et al. (2013), a simple unsupervised statistical word alignment model, with the `Heuristics` method from the work of Jain et al. (2019).

Baselines — We compare our `translate-train`

[1]The translation of documents using Google Translate was done on 02-12-2020.

	Training			Development			Test		
	#docs	**#sents**	**#tokens**	**#docs**	**#sents**	**#tokens**	**#docs**	**#sents**	**#tokens**
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
Dutch	145	2,544	46,894	23	496	9,165	72	2,410	48,007
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
Italian	80	2,951	81,400	17	551	16,904	46	1,494	41,586
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

Table 1: SemEval-2010 Dataset Statistics

Method	**Alignment tool**	**Dutch**	**Spanish**	**Catalan**	**Italian**
Translate-train	Fast-Align	0.280	0.410	0.410	0.340
Translate-train	Heuristics	0.260	0.390	0.370	0.307
Translate-test	Fast-Align	0.365	0.461	0.480	0.362
Translate-test	Heuristics	0.358	0.438	0.453	0.347
End2end Coref	-	**0.380**	**0.516**	**0.533**	0.430
Sucre or Tan-l*	-	0.191	0.490	0.482*	**0.607**

Table 2: Monolingual and Cross-Lingual results in terms of Average Coreference F1

and translate-test approaches with a model trained on annotated data in the target language (i.e., End2end Coref). We also consider two alternative baseline systems for which end-to-end coreference results were reported on the SemEval 2010 shared task data: Sucre and Tan-l. The Sucre system (Kobdani and Schütze, 2010) uses engineered features for words, mentions and mention pairs and uses classical machine learning classifiers to cluster mentions. It reports the best results for Spanish, Italian and Dutch. The Tan-l system (Attardi et al., 2010) uses dependency parse trees to detect mentions and trains a binary classifier to decide the pairwise relationship between the extracted mentions and reports the best result for Catalan. Works such as van Cranenburgh (2019) and Rahman and Ng (2012) are not used as baselines because they make use of external resources (mention detectors, NER, Alpino parse trees[2], etc.).

Metrics — For evaluation, we report the average F1 of the MUC, B3, and CEAF4 coreference resolution metrics, as proposed in Denis and Baldridge (2007).

3.1 Results

Our end-to-end monolingual baseline outperforms the Sucre and Tan-l systems on Catalan, Spanish and Dutch, as shown in Table 2. For Italian, our baseline shows inferior performance, possibly due to the small number of training examples (i.e., only 80 documents). Interestingly, our cross-lingual models remain unable to surpass the effectiveness of their monolingual counterparts, although the former leverage a much larger coreference corpus than the latter. The Translate-test is consistently better than Translate-train, which we hypothesize is due to the superior quality of the English SpanBert model,especially in comparison with the End2end Coref models trained on the translated (i.e., noisy) source corpus. The Fast-Align alignment strategy consistently outperforms the Heuristics based alignment method in both the Translate-train and Translate-test approaches. Jain et al. (2019) showed that the Heuristics improved on their Fast-align and indicated the reason to be that named entities are low-frequency words. To improve its performance, we trained Fast-Align on the additional parallel corpus Europarl (Koehn, 2005).

3.2 Error Analysis

In this section we discuss the contributing factors to the low performance of the Translate-test setup (being the better of both scenarios). From 10 randomly sampled test documents, which contain a total of 424 mentions and 127 mention clusters, we quantify three particular sources of error (see Fig. 1b):

Translation Error — To measure the impact of the imperfect translation step, we annotate the Dutch-to-English translated documents with coreference labels (i.e., perfect annotation on the noisy translations). We also manually align the noisy English documents with the original Dutch

[2] http://www.let.rug.nl/vannoord/alp/Alpino/

	Source text	Google Translate	Correct translation
1.	Het gesprek ging onder meer over *[Punt].*	The conversation was about *[Dot].*	The conversation, amongst others, was about *[Punt]*
2.	*[Mark Grammens]* . . .	*[Mark Grams man]* . . .	*[Mark Grammens]* . . .
3.	. . . en *[zijn]* stelling is bekend	. . . and *[its]* position is well known	. . . [his] position is well known
4.	*[Die]* nam daar genoegen mee.	*[Which]* was content with that.	*[He]* was content with that

Table 3: Literal translation error (1 & 2) and pronoun mistranslation (3 & 4) examples

Model	F1
`Translate-test`	0.415
only *translation* error	0.490
only *labeling* error	0.613
only *alignment* error	0.896

Table 4: Error breakdown for a random sample of 10 Dutch SemEval-2010 documents.

documents (i.e., to simulate perfect alignment).

Automatic Labeling Error — To see the impact of the prediction model, we use `SpanBert` to annotate the manually translated documents (i.e., assuming perfect translation), again followed by a manual alignment step (i.e., to avoid alignment errors).

Alignment Error — To quantify the noise induced by the alignment step, we manually translate the documents to English and manually assign the coreference labels, after which `Fast-Align` is applied for alignment with the original Dutch documents for evaluation.

Our analysis on the error breakdown is shown in Table 4. The largest source of error for the `translate-test` model appears to be the MT step followed by the labeling error, whereas the impact of the alignment error is rather small. We looked into the translation errors, and observed that the coreference results are most degraded due to incorrectly translated pronouns, and literal translations of (parts of) named entities.

The labeling error leads to a hypothetical F1 (i.e., in the absence of other errors) of 0.613 on the selected documents. This is considerably below the reported `SpanBert` performance of 0.796 on the Ontonotes test set (Joshi et al., 2020). We hypothesize this is partly due to the shift in domain between the English Ontonotes data and the SemEval data in Dutch, as well as some differences in coreference annotation guidelines between both datastets. For example, coreference relations with verbs are annotated in Ontonotes but not in SemEval.

4 Related Work

The key concept used in the presented transfer learning scenarios, is *annotation projection*, as originally proposed by Yarowsky et al. (2001) for part-of-speech tagging. It relies on the transfer of annotations from the source language to the target language. Most annotation projection methods depend on parallel corpora in which the source data is labeled using a trained model before projecting the labels onto the data in the target language (Hwa et al., 2005; Postolache et al., 2006; Zeman and Resnik, 2008; Ehrmann et al., 2011; de Souza and Orăsan, 2011; Fu et al., 2014; Ni et al., 2017; Grishina, 2019).

Alternatively, other works relied on the use of bilingual dictionaries for annotation projection (Mayhew et al., 2017; Xie et al., 2018). The `Translate-train` idea of creating a noisy translated corpus with projected annotations has been proposed as well (Tiedemann et al., 2014; Jain et al., 2019), for the task of dependency parsing and NER, respectively. Shah et al. (2010) used MT in the other direction (`Translate-test`) for the task of NER.

A common problem in both annotation projection scenarios is the alignment of text spans between languages, for which unsupervised statistical alignment models can be used (Shah et al., 2010; Ni et al., 2017), such as the IBM models 1-6 (Brown et al., 1993; Och and Ney, 2000). A few recent works (Mayhew et al., 2017; Xie et al., 2018) perform translation on a word or span level to avoid the alignment problem. Others explored alignment heuristics such as matching words based on their surface forms and translations (Ehrmann et al., 2011; Jain et al., 2019), or using external information such as Wikipedia links (Nothman et al., 2013; Al-Rfou et al., 2015).

The prior works applied annotation projection to the tasks of NER, POS, or dependency parsing, and proved relatively successful (i.e., close to monolingual models in the target language). Postolache et al. (2006); de Souza and Orăsan (2011) are notable prior works that applied the idea of annotation projection to the task of coreference resolution. Unlike our work, they depend on the existence of parallel corpora and are focused on a single language pair to test their ideas. Moreover, they have a pipeline that extracts mentions using external annotation tools, or even manually, before clustering them into coreference chains. We, however, perform both the mention identification and clustering in a span-based end-to-end fashion.

For the task of end-to-end coreference resolution, we explore using machine translation for annotation projection, especially with medium-resource languages for which strong MT systems exist. We investigate if MT systems can be used for transferring coreference knowledge (model, dataset) without having to rely on parallel corpora.

5 Conclusion and Future work

While the idea of leveraging MT to improve NLP task performance for low resource languages is not new, this idea to the best of our knowledge has not been pursued for coreference resolution. We contribute by comparing two conceptually different methods; the `Translate-train` and `Translate-test` approaches. We further present a rigorous quantitative error analysis. From our work, we conclude that (i) for coreference resolution the MT approaches are not very successful. (ii) our error analysis suggests this is mainly due by translation errors followed by labeling and alignment errors.

We believe MT models can be still leveraged in cross-lingual transfer learning for coreference resolution, but we speculate that access to the internals of the models, such as attention weights, will be needed. Moreover, future work will need to investigate hybrid strategies, combining transfer learning from other languages with the available data in the target language, to override issues due to MT uncertainty or differences in annotation guidelines.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Giuseppe Attardi, Maria Simi, and Stefano Dei Rossi. 2010. Tanl-1: coreference resolution by parse analysis and similarity clustering. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 108–111.

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

Breck Baldwin and Thomas S Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 1–6.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

José Guilherme Camargo de 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.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Hongliang Fei, Xu Li, Dingcheng Li, and Ping Li. 2019. End-to-end deep reinforcement learning based coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 660–665.

Ruiji Fu, Bing Qin, and Ting Liu. 2014. Generating chinese named entity data from parallel corpora. *Frontiers of Computer Science*, 8(4):629–641.

Yulia Grishina. 2019. *Assessing the applicability of annotation projection methods for coreference relations*. Ph.D. thesis, Universität Potsdam.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(3):311–326.

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine-translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677.

Hamidreza Kobdani and Hinrich Schütze. 2010. Sucre: A modular system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92–95. Association for Computational Linguistics.

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

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

Thomas S Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *arXiv preprint arXiv:1707.02483*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Oana Postolache, Dan Cristea, and Constantin Orasan. 2006. Transferring coreference chains through word alignment. In *LREC*, pages 889–892.

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.

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.

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.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Eighteenth Conference on Computational Natural Language Learning (CoNLL 2014)*.

Andreas van Cranenburgh. 2019. A dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. *arXiv preprint arXiv:1808.09861*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. Technical report, JOHNS HOPKINS UNIV BALTIMORE MD DEPT OF COMPUTER SCIENCE.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.