Annotation projection for coreference resolution
STSM report

COST STSM Reference Number: COST-STSM-IS1312-060116-070265
COST Applicant: Yulia Grishina, University of Potsdam
Host: Anders Søgaard, University of Copenhagen
Period: 06.01.2016 - 10.02.2016

March 15, 2016

The goal of this short-term scientific mission was to investigate how linguistic annotations of discourse coherence can be used for annotation projection and development of linguistically annotated resources for multiple languages. In particular, we were dealing with annotations of coreference (which is necessary to maintain coherence in discourse).

Repeated references to entities are common for all languages, but they can vary considerably in their structure and usage. The emergence of parallel corpora made it possible to compare referring expressions in different languages based on the annotated data. However, linguistically annotated resources are unevenly distributed across the languages, and there are only few parallel coreference corpora available, mostly for major European languages. Our idea was to apply automatic annotation projection techniques, in order to alleviate the problem of resource scarcity.

Our work proposal was based on our earlier research where we applied a direct projection algorithm on a multilingual and multi-genre coreference corpus (Grishina and Stede, 2015), in order to see how well a projection algorithm works for a) different languages, b) different text genres. Therefore, we experimented with relatively similar languages (English-German) and less similar languages (English-Russian) and used three different genres of text (newswire, stories, medical instructions). At the same time, Agic and colleagues (2015) applied a projection algorithm to transfer POS annotations for 100 languages and used all possible combinations of them in order to implement a multi-sourced annotation transfer. The purpose of the proposed STSM was to collaborate on the implementation of a similar algorithm for coreference annotations and make it applicable to a wide range of languages.

Work carried out and results obtained were driven by the idea to explore the boundaries of large-scale annotation projection for coreference resolution. We collected the data that contains gold coreference annotations for multiple language from the recently published sources. We used the data with coreference annotations for as many languages as possible, in particular, from SemEval-2010[1] and CoNLL-2012[2] Shared Tasks. We used the data for the following languages: English, German, Italian, Spanish, Catalan, Dutch, Arabic and Chinese.

[1] http://stel.ub.edu/semeval2010-coref/

[2] http://conll.cemantix.org/2012/data.html

We developed the following three-step strategy to perform multilingual coreference resolution:

- A. Singleton detection: based on the work of (Recasens et al., 2013), we used the data for the languages described above to train singleton classifiers.
- B. Annotation projection: based on the work of (Agic et al., 2015), we reimplemented the projection algorithm to take multi-sourced coreference annotations and transfer them through word alignments to multiple target languages.
- C. Training of coreference resolvers: we used projected annotations as the training data for target coreference resolvers, comparing them to delexicalized and delicalized with cross-lingual word embeddings baselines.

Our first step was singleton detection. We compared coreference annotations in different languages and datasets and found out the following: while SemEval-2010 datasets contains both annotations (1) of singletons and (2) of coreference chains, CoNLL-2012 data only includes (2). We had to adapt the data to our task and automatically extract all the singleton mentions. We developed heuristics based on part-of-speech tagging and parse bits annotated in the gold data and augmented the existing annotations with singleton mentions. After that, we converted the data into the necessary format, in order to use it for training of an LSTM classifier for multiple languages. The purpose of singleton prediction was to get task-specific word representations for system training.

The second step was the adaptation of the projection algorithm for coreference. We adjusted the algorithm of (Agic et al., 2015) for the task of coreference resolution that combines multiple source languages and aggregates different source annotations (part-of-speech tags, syntactic trees) to transfer coreference chains from one language to another. We used IBM-1 word alignments and majority voting strategy for the projected annotations.

As for the projection data, we used the full version of the Bible (parallel translations) for the projection, which comprises around 100 different languages. We split the Bible by chapters and used these as separate documents to perform coreference resolution. We compared the most recent state-of-the-art coreference resolution systems and decided to use two of them, which are built on language-independent machine learning algorithms: Berkley entity resolution system (Durrett and Klein, 2014) to tag the English side of the Bible as well as HOTCoref system (Björkelund and Kuhn, 2014) to tag the German side. To train coreference resolvers for new languages, we introduced the following settings:

1. 'Projection only' approach: training coreference resolvers on the projected annotations only;
2. Delexicalized approach: training coreference resolvers on delexicalized data;
3. Delexicalized approach with cross-lingual clusters: training coreference resolvers using cross-lingual word embeddings clusters;
4. Combination of the above.

Our results include, firstly, the development of the preprocessing baseline for singleton detection and application of LSTM classifiers for this task, and, secondly, the development of the projection algorithm to transfer coreference annotations in multiple languages. For the time being, we obtained projected annotations for 10 European languages: German, French, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Spanish, Swedish. These languages were selected based on the availability of high-quality word alignments granted by the Copenhagen group.

The number of sentences in the resulting multilingual projected corpus varies between 19 489 and 28 457 per language. Our evaluation has shown that, on average, we were able to transfer 86 000 coreference mentions per language, depending on the number of available parallel sentences. We obtained the highest number of transferred mentions for Portuguese (99 016 mentions) and the lowest number for Romanian (68 163).

To evaluate the quality of our projections at this point, we intended to compare our projections against automatic annotations produced by language-specific systems (if available). This turned out to be a challenging task, as most of the known systems either adopt different conventions regarding coreference mentions (e.g. including NPs of different size or allowing verbs to participate in coreference chains), or exhibit low-quality results and therefore could not be used as a 'gold standard' for our evaluation. At the current stage, we were able to perform extensive evaluation of our method for German in two settings[3]: (a) evaluation of the projected annotations against automatic annotations produced by one of the state-of-the-art coreference resolvers[4] (these were used as a 'gold standard') and (b) evaluation of the projected annotations using a small sample of data manually annotated by the author of this report according to our multilingual annotation guidelines from (Grishina and Stede, 2015). In the first setting, we achieved a considerably high F-score of 54.41 for the detection of coreference mentions in our corpus, and a fair F-1 of 33.89 for coreference chains which we attribute to the differences in the types of extracted mentions produced by automatic annotations on the source side, word alignments and mistakes that come from the system used for evaluation. In the second setting, we obtained the F-score of 63.87 for mention detection and the F-score of 46.4 for coreference links using a small portion of high-quality manually annotated data. It is worth pointing out that the state-of-the-art systems in question show the following resuts: the German system exhibited F-1 of 54.44 on SemEval-2010 dataset (Klenner and Tuggener, 2011), while Berkeley entity linking system achieved F-score of 61.71 on CoNLL-2012 dataset (Durrett and Klein, 2013). These results are considered high for this task.

In sum, we developed a new strategy on performing large-scale coreference projection which could be applicable to as many languages as possible, including truly low-resourced languages. The obtained results demonstrate the convenience of our approach for coreference resolution. Our future work includes the investigation of how the projected data could be used to train coreference resolvers for multiple languages and its comparison to the known baselines. Yulia Grishina, Zeljko Agic and Anders Søgaard plan to summarize the results of the STSM in a more elaborated form in a paper on coreference projection. We will present the evaluation of our algorithm as compared to the available language-specific coreference resolvers, language-independent systems trained on the gold standard Shared Tasks data and the baselines presented above.

The developed language-independent algorithm supports creating coreference corpora in new languages, thus contributing to the TextLink goals regarding the enhancement of interoperability

[3] To compute the scores we used coreference scorer provided by the Shared Task: http://conll.github.io/reference-coreference-scorers/. To evaluate coreference chains, we computed the average of the standard coreference metrics MUC, CEAF, B-cubed and BLANC.

[4] http://www.cl.uzh.ch/de/research/coreferenceresolution.html

of linguistic resources on discourse structure. The divergences in the analysis and representation of coreference across different languages impair multilingual language processing. Gaining knowledge about the distribution of these devices in different languages will support the standardization of the annotation schema and mapping from one language to another.

In particular, this STSM contributed to the following TextLink objectives:

- **Contribution to WG1 Resources**: it contributed a projected corpus with projected coreference chains for 10 European langauges, which could potentially be extended to other languages.
- **Contribution to WG3 Tools**: the developed pipeline and the projection algorithm can serve as an automatic pre-annotation step prior to the manual annotation of coreference, and the projected annotations could be used to train coreference resolvers for new languages.

The developed algorithm and the projected corpus will be made available after the publication of results and accessible via http://angcl.ling.uni-potsdam.de/resources.html.

Overall, this STSM contributed to the development of more discourse-annotated resources for under-resourced languages and domains, which thereby helps to promote language diversity and equality.

References

[1] Agic, Zeljko; Hovy, Dirk; Søgaard, Anders. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. The 53rd Annual Meeting of the Association for Computational Linguistics (ACL). Beijing, China.

[2] Björkelund, Anders, and Jonas Kuhn. 2014. Learning Structured Perceptrons for Coreference Resolution with Latent Antecedents and Non-local Features. The 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, USA.

[3] Durrett, Greg, and Dan Klein. A joint model for entity analysis: Coreference, typing, and linking. Transactions of the Association for Computational Linguistics 2 (2014): 477-490.

[4] Grishina, Yulia, and Stede, Manfred. 2015. Knowledge-lean projection of coreference chains across languages. The 8th Workshop on Building and Using Comparable Corpora (BUCC). The 53rd Annual Meeting of the Association for Computational Linguistics (ACL). Beijing, China.

[5] Recasens, Marta, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The Life and Death of Discourse Entities: Identifying Singleton Mentions. HLT-NAACL.

[6] Klenner, Manfred, and Tuggener, Don. 2011. An incremental entity-mention model for coreference resolution with restrictive antecedent accessibility. In: Recent Advances in Natural Language Processing (RANLP 2011), Hissar, Bulgaria, 12 September 2011 - 14 September 2011, 178-185.