A Coreference-Annotated Corpus for Machine Translation Research

Ekaterina Lapshinova-Koltunski, Christian Hardmeier

Final report for the EAMT Sponsorship of Activities Call 2017

1 Formalities

1.1 First steps

The projected started with a delay, as we faced some formal problems at Saarland University which is the project coordinator and also the annotation location (as annotators were planned to be employed through Saarland University). After a third-part account for the project had been created (on the 23rd of March 2017), we were able to start the project with no further difficulties.

1.2 Annotator recruitment

The annotator selection process also took place during the first months. We were able to recruit an experienced annotator for our project. This annotator had been involved in the annotation of coreference and other discourse-related structures in the project GECCo, which was advantageous for us. At the same time, since this person already had a Master's degree in Translation and Interpreting and was therefore more expensive as an annotator, we reduced her working hours. However, since we saved time that we would have needed for training a less experienced annotator and this annotator works very fast, we achieved the planned aims at lower than budgeted cost despite the higher remuneration.

2 Annotation Process

2.1 Formulation of annotation categories

A detailed description of categories and disambiguation rules are needed to guarantee consistency throughout the whole process of annotation. During the first months of the project, we developed the annotation guidelines. They were based on the three existing ones described by Grishina and Stede [GS16], Guillou et al. [GHS+14] and Kunz [Kun12]. They concern the segmentation of nominal elements, the account of different antecedent and anaphora types and examples of various problematic cases [LKH17]. A copy of the annotation guidelines (which have been refined and enhanced until now) can be found in the project Overleaf `https://www.overleaf.com/read/qrrqwgxfzybn` (read-only copy). These guidelines also contain a discussion on how they differ from the existing and underlying annotation guidelines. We used the existing pre-annotated (or partly annotated) corpora and extended them with the complete annotation of full coreference chains, adding further referring expressions to the coreference chains.

Segmentation An annotated element is any pronoun, noun, nominal phrase or an elliptical construction that is a part of coreference (antecedent-anaphora) pair, as well as verbal phrases or sentences being antecedents of event anaphora.

Types of antecedents In our framework, we have different types of antecedents: nominal entities (represented by nouns, nominal phrase or pronouns), verbal phrases (event-vp) as in (1-a), fact sentences as in (1-b), split antecedent (multiple antecedents – all components of the antecedent are linked to the referring expression) or no explicit antecedent (in some cases, a referring expression is anaphoric, but no specific antecedent can be found in the text).

(1) a. *... you have to basically [combine everything you learned from project one and project two]. ultimately [that]'s the goal .*
b. *[We work for prosperity and opportunity] because they're right. [It]'s the right thing to do.*

Types of anaphors We include various types of referring expressions (anaphoras) into our analysis: proper names (*Herr Almeida Freire* in example (2-a)), nominal premodifiers as in (2-b), full nominal phrases (used with a definite article or a demonstrative modifier as in example (2-c)) and nominal phrases with quantifiers (*all people* in the meaning *all these people*).

(2) a. *In [seiner] EWSA-Stellungnahme zum "Bericht der Kommission zur Beobachtung des Handelsmarktes" schreibt [Herr Almeida Freire]...*
b. *The unionists used to be [[EU] supporters], but now they are questioning how [it] has developed...*
c. *This past spring, the U.S. Department of Education issued [a report, The Condition of Education 2000]. [The report] found that...*
d. *[Computers] are expensive. But [they] are useful. Computers cost a lot of money.*

Generic nouns can co-refer with definite full NPs or pronouns, but not with other generic nouns, see (2-d).

Different types of pronouns, such as personal, demonstrative, relative and reflexive ones can also serve as referring expressions. Demonstrative pronouns may also refer to locations and time (*there, here*). We also include the category of pronominal adverbs. Pronominal adverbs exist in both English and German, but are used very differently. In German, they are very common, but in English, they sound rather archaic and are generally avoided. They are formed by replacing a preposition and a pronoun, like *gegen+das* → *dagegen* in example (3). Pronominal adverbs are not considered in most coreference annotation schemes. However, they constitute around 8% of all referring expressions in the German language and are especially frequent in spoken and spoken-like language.

(3) *Viele Amerikaner haben Probleme mit [Rassismus]; doch wir sind [dagegen] immun.*

Linguistic chains may also include *substitution* and *ellipsis* in addition to referring expressions. These trigger a type reference relation between referents belonging to the same class [KS13, DBD81]. In substitution patterns, the referring expression is replaced with another element (example (4-a)). In ellipsis, it is completely left out, and the reference is implicit (example (4-b)).

(4) a. *Do you prefer the blue shirt or [the red shirt]? – I would like the red [one].*
b. *...if I take any one of these balls... and I count how many [neighboring balls] that there are around it, the answer's always twelve [].*

Substitution and ellipsis are mostly analysed within separate chains in other studies. We included them into our framework, since they are often used in similar contexts as coreference if used cross-lingually.

Another category that is considered here but is excluded from most analyses is that of comparative reference, which does not trigger co-reference in the strict sense. Together with other cases (substitution and ellipsis) it rather involves type-reference, co-classification or "sloppy identity", see [KS12]. The linguistic means signaling comparative reference include such word as *same, equal, identical* or particular adjectives in the comparative form.

2.2 Data preparation

Our aim was to use the existing resources and extend them with: (1) complete annotation of full coreference chains; (2) additional referring expressions to achieve full coreference chains. We planned to use the following resources: ParCor corpus, the dataset used for the DiscoMT workshop shared task and the English-German component of the data described by Grishina and Stede [GS15]. We started with the annotation of ParCor and DiscoMT data, as the texts in these dataset come from the same source, they are TED talks. It is easier for an annotator to work with a similar register before switching to the next register. We were not able to receive the data from Grishina and Stede, as some of the copyright issues have not yet been clarified by the authors.

Therefore, we decided to find additional texts from a resource that has not been annotated for coreference yet. We decided for the WMT data, as this data type has been extensively used for other machine translation-oriented tasks. Besides that, it represents a different register and genre than included into DiscoMT data and ParCor. Table 1 presents an overview of the total number of tokens included into our corpus.

language	ParCor	DiscoMT	WMT news	total
English	31,971	39,764	10,644	82,379
German	30,305	37,452	10,593	78,350
total	62,276	77,216	21,237	160,729

Table 1: Corpus data

For the DiscoMT dataset, the existing annotations cover English only. So, for the part of the German data (DiscoMT), we had to find the corresponding translations into German (DiscoMT 2015 shared tasks were English-French only). Then, we had to prepare the data (tokenise and sentence split it accordingly). In the end, this dataset could be extended to a parallel corpus.

Overall, we extended annotations (to add all types of referring expressions) for 71,735 tokens of the English data and 30,305 of the German data. 10,644 tokens of the English data and 48,045 tokens of the German data was annotated from scratch. The annotation of the WMT news (21,237) was a different, more laborious process, as neither the English, nor the German texts were pre-annotated for any coreference-related structures. The total amount of tokens in the annotated corpus comprises ca. 160,000.

The annotated resource that we have created represents a reasonably-sized data set for training coreference resolution components that can be used for MT or other cross-lingual applications. It is comparable in size (with a larger amount of text, but fewer annotated mentions) to the ARRAU corpus [PA08], which features a similarly rich coreference annotation and covers a greater variety of genres, but does not include multilingual parallel text. Moreover, although the amount of data is not enough to train an MT system, this dataset will be large enough for MT tuning, testing and evaluation, which is an important improvement over the existing data situation.

3 Annotation Results

At this stage, all the texts have been annotated. The annotations of ca. 110,000 tokens are completely finalised. About 50,000 remaining tokens still need a final check for inconsistencies, which will be done within the next month. We present an overview of the annotated structures (absolute numbers) in Table 2 below.

	English	German	total
pronoun	6,577	4,241	10,818
np	2,344	2,469	4,813
vp	126	130	256
clause	316	304	620
total mention	9,363	7,144	16,507

Table 2: Annotated mentions and their subcategories

Overall, we have ca. 16,500 annotated mentions at the moment. The annotated mentions are classified according to their morpho-syntactic type: pronouns (pronoun), nominal phrases (np), verbal phrases (vp) and clauses (clause). This differentiation was introduced for a practical reason, as it allows to further classify mentions according to their function or the role in a coreference chain.

Currently, we have annotated 4,787 full coreference chains in our data, see Table 3. The number of chains is estimated according to the number of antecedents assigned. We also calculate the average chain length (total number of mentions/total number of chains). The German translations contain more chains than their English sources. At the same time, these chains seem to be shorter in comparison to the English ones.

	English	**German**	**total**
chain number	2,332	2,455	4,787
average chain length	2.82	2.69	2.75

Table 3: Annotated chains

To evaluate the reliability of the annotated coreference chains, we created a second annotation of two files in each language. As a measure of inter-annotator agreement, we computed the mention overlap and entity-based CEAF scores [Luo05] between the two annotations, treating our regular annotator as the hypothesis to be evaluated and the second annotator as the reference. The scores were calculated with the CoNLL reference scorer implementation [PLR+14] and are shown in Table 4 as a macro-average over the two documents.

	Precision	**Recall**	**F-score**
English			
mentions	89.20%	73.89%	80.71%
CEAF*e*	82.90%	67.13%	74.13%
German			
mentions	84.80%	69.76%	76.54%
CEAF*e*	72.53%	60.36%	65.88%

Table 4: Inter-annotator metrics for coreference chains

As seen from the table, we observe a better agreement for the English texts. We suppose that the reason for more disagreement for German texts is the complexity of the linguistic structures triggering coreference in this language. However, a more detailed analysis of the agreement results is needed to understand the reasons, which is a part of our future work.

We also performed automatic inconsistency checks to prove if the annotated data contains any (1) marked mentions outside of chains; (2) antecedents of chains that are not marked as first elements of chains; (3) some other error types. The detected errors have then been being corrected by the annotators. Besides that, the annotator have been adding the following categories (that were not included into the annotation scheme at the very beginning): (a) bare nouns; (b) indefinite nouns and (c) quantifiers (both) as demonstratives.

While the annotation of the corpus is essentially finished, we are still performing some final data cleanup. We expect that we can release the final version of the corpus on Github and the LINDAT repository in March 2018. Until the official release is made, we are happy to share the preliminary version on request. Already now, the corpus data is being used for the preparation of psycholinguistic experiments on the English pronoun ‘it’. The corpus was described in a research paper that will be published in the LREC-2018 proceedings [LKHK18].

4 Conclusion

The differences in coreference realisation in multiple languages present a huge challenge to machine translation and are of interest for contrastive linguists and researchers in translation studies. A parallel corpus with full annotation of coreference is a valuable resource with a variety of uses. The corpus will help us study the mechanisms involved in coreference translation in order to develop a better understanding of the phenomenon. It will serve as a resource for creating and evaluating coreference-aware MT systems without having to rely on notoriously inaccurate automatic coreference resolvers. Finally, it can also be used as a training and development resource for the creation of multilingual or monolingual coreference resolution systems. Moreover, we address the demand for better approaches to evaluate complex linguistic phenomena that are not covered by existing annotation schemes.

5 Cost breakdown estimated for the entire activity or event

The project was granted with 5600 Euro. The sum of money spent until the end of the annotation process (ca. 4300 Euro) entirely consists of personnel costs for the annotator, who will have worked for 9.5 months at a rate of 37.5% full-time equivalents (with the contract starting on the 8th of May and ending on the 23rd of January 2018). The relatively high remaining amount (ca. 1300 Euro) is due to the fact that we were able to recruit a very experienced and already trained annotator who needed significantly less time than anticipated to create the annotations. We have received permission from the EAMT to use the remaining funds for a test suite-based semi-automatic evaluation of coreference translation that will be carried out and presented in connection with WMT 2018.

References

[DBD81] R.-A. De Beaugrande and W. U. Dressler. *Einführung in die Textlinguistik.* Niemeyer, Tübingen, 1981.

[GHS$^+$14] Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. *ParCor 1.0: Pronoun Coreference Annotation Guidelines.* Edinburgh, Uppsala, March 21 2014.

[GS15] Yulia Grishina and Manfred Stede. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora, Beijing, China*, page 14, 2015.

[GS16] Yulia Grishina and Manfred Stede. *Parallel coreference annotation guidelines.*, November 2016.

[KS12] Kerstin Kunz and Erich Steiner. Towards a comparison of cohesive reference in english and german: System and text. In M. Taboada, S. Doval Suárez, and E. González Álvarez, editors, *Contrastive Discourse Analysis. Functional and Corpus Perspectives.* Equinox, London, 2012.

[KS13] K. Kunz and E. Steiner. Cohesive substitution in english and german: A contrastive and corpus-based perspectivet. In Karin Aijmer and Bengt Altenberg, editors, *Advances in Corpus-Based Contrastive Linguistics. Studies in honour of Stig Johansson*, pages 201–232. John Benjamins, Amsterdam, 2013.

[Kun12] Kerstin Kunz. *Richtlinien für die Korrektur von kohäsiven Referenzmitteln*, December 2012.

[LKH17] Ekaterina Lapshinova-Koltunski and Christian Hardmeier. *Coreference Corpus Annotation Guidelines*, December 2017.

[LKHK18] Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Marie-Pauline Krielke. Parcorfull: a Parallel Corpus Annotated with Full Coreference. In *Proceedings of LREC-2018*, Miyazaki, Japan, 7-12 May 2018. ELRA.

[Luo05] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[PA08] Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[PLR^{+}14] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June 2014. Association for Computational Linguistics.