COME-ALPs: Coreference Annotation with MErging Heuristics using ALignment-based Projection in Parallel Corpora

Gabriela Gonzalez-Saez[*], **Mariam Nakhlé**[*,†], **Illia Kholosha**[*]
Rachel Atherly[*], **Marco Dinarelli**[*]
(*) Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, 38000 Grenoble, France
(†) Dragon LLM, 75008, Paris, France
contact: gabriela-nicole.gonzalez-saez@grenoble-inp.fr

Abstract

Multi-lingual, parallel datasets annotated with discourse phenomena like coreferences are a rare resource. These datasets are useful and informative to evaluate models for NLP tasks taking long contextual information into account, as proved by the large literature published in the last couple of years on e.g. Context-Aware Neural Machine Translation (CA-NMT). Inspired by resources published in previous work (Lapshinova-Koltunski et al., 2022), in this paper we propose an automated procedure to annotate multi-lingual, parallel data with coreferences. Through the use of accurate alignment and coreference annotation tools, we project the annotation from English data, where tools are most often more accurate, to one or more target languages. We apply some consistency constraints to obtain more accurate annotations on both source and target side. Using our procedure we generated two new resources that can be used for evaluating CA-NMT models. One starting from the well-known TED Talk's data released for the IWSLT17 shared task (Cettolo et al., 2017), where we project the annotation from English to target languages as diverse as French, German and Chinese. The second resource is derived from the WMT24 shared task (Kocmi et al., 2024), consisting of news domain data in the same set of target languages. We release these resources, as well as the code framework for applying our annotation procedure, to the community.

Keywords: coreference resolution, neural machine translation, alignment

1. Introduction

Coreference resolution is essential for many NLP applications, including Neural Machine Translation (NMT). Coreference-based evaluation has proven more appropriate for CA-NMT, as witnessed by the abundant literature published in the last couple of years (Bawden et al., 2017; Müller et al., 2018; Voita et al., 2019). As discussed by Lupo et al. (2022a), traditional quantitative metrics for NMT, even recent neural metrics, can't adequately capture translation improvements yielded by correct translations of words involved in discourse phenomena since they are relatively rare. Still these words are crucial for a correct understanding of texts in a long context.

[1] Even when traditional sentence-level metrics are applied to a larger span of segments or even entire documents, they may fail to fully or informatively capture the kinds of document level phenomena that still separate human translations from high-quality sentence-level NMT (Castilho and Knowles, 2024). When a test set requires coreference resolution annotation, research teams often need to develop an entire pipeline, covering coreference resolution and alignment, which also includes testing different methods and their implementation compatibilities, which is both time consuming and not standardized. Meanwhile, existing coreference annotated datasets remain limited in both size and language coverage, restricting comprehensive evaluation and cross-lingual analysis. For example, parallel annotations are particularly important for assessing how NMT models handle coreference at the document level, yet such resources remain scarce.

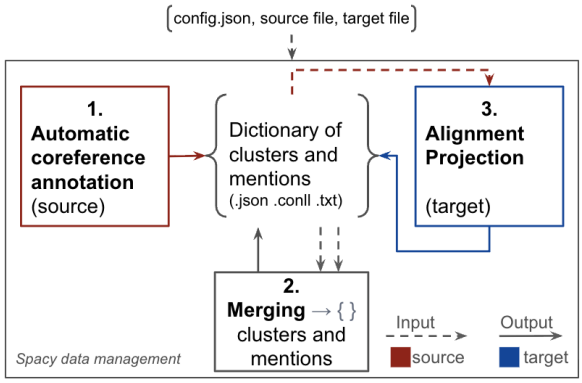

Figure 1: Modules of the COME-ALPs framework.

To address this gap, we describe and propose a procedure for automating the process of i) coreference annotation, ii) alignment in parallel corpora, and iii) annotation projection across languages. Our approach integrates multiple coreference models and employs heuristic-based merging techniques to improve the annotation quality. By providing a flexible and scalable solution, we aim to facilitate the creation of new coreference-annotated

[1] Institute of Engineering Univ. Grenoble Alpes

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1688–1695
11–16 May 2026.

datasets and enhance the study of coreference phenomena in multilingual contexts.

Our contributions are the following:

- We present a framework for document-level coreference annotation.[2] This can be used in turn to perform coreference resolution in multiple languages via coreference projection from English, where tools have most often better performances, to one or more target languages. We release the code for applying our annotation procedure.
- We analyse the impact of merging heuristics on our automatic annotations on the final annotation quality.
- We release two dataset resources annotated using our framework: both based on the IWSLT and WMT test sets in three language pairs, namely English-{French, German, Chinese};
- We provide CA-NMT evaluation of several models using the above-mentioned resources.

2. State of the art

A number of studies have addressed the annotation of context-aware phenomena, with particular attention to datasets, annotation methods, and tools for handling discourse-level information. In this section, we review the main existing resources for coreference annotation and related work.

2.1. Coreference datasets

Several datasets have been developed to capture coreference phenomena across languages. Jiang et al. (2023) introduced a Chinese–English parallel corpus annotated for discourse-level features, including Named Entities, Terminology, Coreference, and Quotations. The annotation was performed by professional translators, resulting in 8,585 annotated mentions on the English side and 6,070 on the Chinese side. In contrast to such manually curated datasets, our goal is to provide an automated framework for coreference annotation that does not rely on professional translators. This approach makes it possible to efficiently generate new coreference annotated data and enables bilingual analyses of discourse structures in both source and target languages.

ParCorFul2.0 (Lapshinova-Koltunski et al., 2022) is a multilingual test suite with coreference annotation. The parallel data are in English, German, French, and Portuguese. However, while the corpus is multilingual and parallel, the coreference annotations themselves are language-specific. Mentions are annotated independently in each language without explicit cross-lingual alignment. Building on ParCorFull2.0, the more recent CorefUD corpus (Nedoluzhko et al., 2022) expands coreference annotation to twelve additional languages. However, these annotations remain monolingual rather than parallel, limiting their use for cross-lingual comparison. In contrast, our annotation procedure projects coreference chains directly from a source language to one or more target languages, assigning consistent identifiers to aligned mentions. This allows cross-lingual coreference mappings and facilitates multilingual evaluation.

Finally, several tools have been developed for automatic coreference resolution, such as NeuralCoref[3] from HuggingFace, and more recently FastCoref (Otmazgin et al., 2022) and Maverick (Martinelli et al., 2024). These tools provide pipelines for mention detection and coreference chain construction, and they form the foundation for developing and extending multilingual coreference resources such as those discussed above.

2.2. Alignment Tools

A key step in building parallel annotated datasets is word or phrase alignment, which establishes correspondences between linguistic units in source and target languages. Several tools have been developed for this purpose, including FastAlign (Wang and Lepage, 2016), SimAlign (Sabet et al., 2020), and Awesome-Align (Dou and Neubig, 2021). Such neural alignment tools, have shown improvement in the analysis of several specific phenomena and specific languages (Zheng et al., 2023).

To enable the projection of annotations of coreference chains across languages we use alignment tools, that helps to obtaining annotated parallel corpora from multilingual data. This approach has been successfully applied to the creation of silver annotated datasets for low-resource languages (Kuchmiichuk, 2023), and to the construction of test suites evaluating the ability of neural translation models to handle discourse-level context (Bawden et al., 2017; Müller et al., 2018).

A number of neural alignment models have also been proposed in recent years. For instance, Zenkel et al. (2020) introduced BAO-Guided Alignment, which incorporates a dedicated alignment layer into the Transformer decoder. Similarly, Chen et al. (2020b) proposed SHIFT-ATT and SHIFT-AET, which extract alignments from attention weights and from an Alignment-Enhanced Transformer (AET),

[2] https://github.com/gabrielanicole/come-alps

[3] https://github.com/huggingface/neuralcoref

respectively, integrating an additional alignment module into the Transformer architecture. Other approaches, such as the one presented in (Chen et al., 2020a), introduce a "garbage collector" mechanism to filter noisy attention alignments and improve the overall quality of alignment extraction.

In this work, we propose to unify alignment and coreference resolution tools within a single, compatible framework. By integrating these two components, we aim to build a scalable and accurate pipeline for multilingual and parallel coreference annotation.

3. COME-ALPs: Annotation Framework

Our annotation framework consists of three main modules : 1) **Coreference** annotation on a source language (always English in this paper), using one or more annotation tools; 2) **Merging** the annotations from the different tools applying some consistency heuristics, to obtain possibly a more accurate annotation; and 3) Alignment-based coreference **projection** to a target language.

Like in many other annotated corpora, annotated mentions and mention sets, i.e. clusters, have a unique identifier across the corpus. In order to make the annotation really parallel, we use the same identifiers for source and target languages, that is between original and projected mentions. This makes the retrieval of corresponding mentions in source and target languages, and e.g. quantitative coreference-based evaluation of CA-NMT models straightforward, even with languages the evaluators are not proficient with.

Coreferent mentions are projected from source to target language using word-level alignment. In order to obtain an annotation on the target language the more accurate as possible, we apply a *continuity* consistency constraint: target words instantiating a mention must be adjacent in the sentence. When projected mentions contains a gap, that is some non-aligned words are surrounded by aligned words, we include the non-aligned words in the annotation. Projected mentions consisting of only one word are kept only if the word is not a function word.

We use at least two coreference resolution tools in order to perform annotation merging. The underlying intuition is that if different tools detect the same mentions, and thus the same clusters, it is more likely the annotation is correct.

Several coreference and alignment tools have been integrated into the framework. The choice of a specific tool is treated as a parameter, allowing flexible configurations. Figure 1 illustrates the interaction between the three modules. Execution of the pipeline can be configured via a config.json file, which specifies all parameters and the combination of coreference, merging, and alignment procedures. A direct execution example is shown below:

```
python comealps.py --file-source
    source_document --file-target
    target_document --file-format 'IWSLT'
     --target 'de' --method align --
    coreference fast_coref --alignment-
    method awesome
```

3.1. Coreference Annotation

The first step of the framework perfoms coreference resolution on the source language using a selected tool among the following implemented: NeuralCoref, FastCoref and Maverick. COME-ALPs accepts input files in plain text or XML format, as commonly used in WMT and IWSLT datasets. The output of this step is a dictionary representing clusters and the mentions belonging to each cluster. Each cluster contains mention identifiers along with the text span. This data structure serves as the foundation for subsequent merging, alignment, and projection processes. A representative dictionary example is shown below:

```
{ { "1": {
        "144-156": {"text": "a local mall
            ", }
        "319-337": {"text": "the
            Westfield Mall", } ... }
  ... } }
```

In addition to the dictionary, the annotated text is exported while preserving the original document structure. Each mention is marked with a <mention> tag, which includes both its mention and cluster IDs. For evaluation purposes, the annotations are also generated in CoNLL format.[4]

This step is modular: any coreference resolution tool that outputs a dictionary of mentions and clusters in the same format can be used. As long as the dictionary conforms to this structure, the subsequent merging, alignment, and projection steps can be executed seamlessly.

3.2. Merging Coreference Annotations

To increase the reliability of the generated coreference clusters, we introduce a merging step before alignment and projection to the target language. This step performs a cross-validation between clusters produced by two coreference tools (called method A and B), retaining only those that are detected by both, at least partially.

The merging process proceeds as follows: (i) clusters generated by the two tools are compared, and two clusters are considered to correspond if

[4]https://universaldependencies.org/format.html

at least $\alpha\%$ of their mentions overlap. Each cluster from method A is aligned to only one cluster from method B, the one with the highest overlap; (ii) once clusters are aligned, each cluster from method A is extended by including mentions detected in the aligned cluster from method B. For overlapping mentions within a cluster, the longest mention span is retained, and (iii) as this step functions as a validation process, only clusters from method A that have a corresponding aligned cluster in method B are retained for subsequent processing.

Note that *merging coreferences* is a process applicable at $original$ or $projected$ mentions.

3.3. Alignment and Projection

In this step, coreference annotations are projected from the source language to the target language or system output. Projection to system-generated data is more challenging, as translation quality can significantly affect the accuracy of the annotation. The process begins with token-level alignment between the source and target sequences. This alignment allows the framework to project source-language mentions onto the corresponding target tokens. To maintain contiguous mention spans, all tokens that fall between aligned tokens of a projected mention are also included in the target mention. A critical aspect of this step is the preservation of mention and cluster identifiers. When projecting annotations, the framework transfers the IDs from the source language to the target, ensuring that corresponding mentions remain consistently linked across languages. This design enables reliable cross-lingual analysis and evaluation of coreference phenomena in parallel or system-generated corpora.

After projecting coreferences using different alignment tools, the resulting annotations can be merged following the same procedure described in Section 3.2.

4. COME-ALPs Dataset Annotation

In this section, we demonstrate the capabilities of our framework for annotating parallel datasets. First, we present a precision-based evaluation of coreference resolution in parallel corpora. Next, we introduce the new annotated resources generated with our framework. Finally, we provide a use case illustrating how these resources can be used to analyze the interaction between multilingual coreference resolution in context-aware neural machine translation (CA-NMT).

Model Metric	NeuralCoref + Alignment	FastCoref + Alignment	Merge + Alignment
MUC	39.2	33.4	**40.0**
B^3	**33.7**	27.4	**33.7**
$CEAF_e$	28.1	29.9	**36.1**
$CEAF_m$	36.2	35.4	**41.0**

Table 1: Precision results of four coreference resolution metrics, over discourse phenomena in ParCorFull

4.1. Precision valuation in parallel corpora

To evaluate our framework, we annotate parallel corpora and assess the quality of coreference resolution in the target language. Our evaluation focuses on the complete pipeline, including merging and alignment, to determine their impact on annotation quality across datasets.

Datasets and pipeline setup We annotate two datasets from ParCorFull2.0 (Lapshinova-Koltunski et al., 2022): DiscoMT composed of 10 documents with an average of 3,745 sentences; and, News, composed of 19 documents with an average of 28 sentences per document. All documents are correctly aligned in English-German. For coreference resolution, we use NeuralCoref and FastCoref, merging their outputs before projecting annotations to the target language using Awesome-Align. To assess the impact of merging, we also project annotations from each coreference tool independently. Evaluation focuses on precision of coreferent mentions, measured using standard metrics: MUC, B^3, $CEAF_e$, and $CEAF_m$. Scores are computed with the CorefUD scorer (Yu et al., 2023) using exact-match evaluation.

Table 1 presents the precision coreference resolution for four metrics. Each column corresponds to a different configuration: NeuralCoref + Alignment, FastCoref + Alignment, and the merged output of both tools + Alignment. The results show that the merged annotations consistently improve precision across all metrics, demonstrating the effectiveness of our merging strategy. Notably, the largest improvements are observed in $CEAF_e$ and $CEAF_m$, indicating that combining multiple tools helps capture more complete and accurate coreference chains. An unexpected observation is the relatively low precision of FastCoref + Alignment. While FastCoref typically achieves MUC-precision values above 80 on similar datasets (Otmazgin et al., 2022), in the projected setting its performance is lower than that of NeuralCoref.

Using our framework, we can detect coreference phenomena present in both languages, which is valuable for analyzing how discourse structures in a source language propagate to the target language

Language pair	en-fr	en-de	en-zh
#documents	909	2016	1523
#coref chains (src)	6775	12133	6424
#mentions (src)	28221	53879	28515
#coref chains (tgt)	6766	12115	6391
#mentions (tgt)	26728	50797	25463

Table 2: WMT24 statistics in source and target language.

Language pair	en-fr	en-de	en-zh
#documents	97	97	97
#coref chains (src)	2253	2746	2315
#mentions (src)	9362	9634	8756
#coref chains (tgt)	2246	2731	2256
#mentions (tgt)	8716	9098	7367

Table 3: IWSLT17 statistics in source and target language.

and potentially influence translation. Section 5 provides an illustrative example. Importantly, our goal is not to achieve state-of-the-art coreference resolution scores, but to establish cross-lingual links between mentions, enabling the study of discourse phenomena in parallel corpora.

4.2. Parallel annotated datasets

Following the pipeline described in the previous section (merging + alignment), we release two annotated parallel datasets produced using our proposed framework, COME-ALPs. In these datasets, the source-side annotations correspond to the merged output of NeuralCoref and FastCoref, while the target-side annotations result from the alignment and projection of the merged source annotations. This setup ensures that both sides of the parallel corpora share consistent coreference identifiers, enabling direct cross-lingual comparison and analysis.

The first dataset is based on the WMT24 test data (Kocmi et al., 2024). Using our procedure, we annotated both the source and target sides of the English–French and English–Chinese parallel datasets, generating more than 6,000 coreference chains per language pair. Table 2 presents detailed statistics for this dataset, including the number of documents, coreference chains, and mentions in both source and target languages.

The second dataset corresponds to the IWSLT17 test set (Cettolo et al., 2017), a well-known document-level translation benchmark. We applied the same annotation pipeline to the available English–French, English–German, and English–Chinese test data. In this case, COME-ALPs generated approximately 2,000 coreference chains per language pair. Table 3 summarizes the statistics for this dataset.

5. Evaluation of CA-NMT Models

In order to show a possible use case of multilanguage, parallel annoated data, we use the IWLST17 (English-French) data annotated with our framework to evaluate CA-NMT models. We inspire this evaluation from (Dinarelli et al., 2024), where two types of CA-NMT models are evaluated with three metrics based on attention weight patterns between a current sentence to be translated and one or more context sentences used by the model to disambiguate coreference phenomena.

5.1. Employed CA-NMT Models

CA-NMT models can generally be categorized into two main groups (Lupo et al., 2022a): concatenation-based and multi-encoder-based approaches. In this work, we analyze one representative model from each category. The first is a multi-encoder model, specifically a variant of the Hierarchical Attention Network (HAN) proposed by Lupo et al. (2022a), which leverages only the source-side context. The second is a concatenation-based model, derived from the Transformer architecture introduced by Lupo et al. (2022b), which uses both source and target context. Both models retain the standard Transformer configuration, consisting of six encoder and six decoder layers, with 512-dimensional hidden states, 2,048-dimensional feed-forward layers, and eight attention heads. Tokenization is performed using Byte Pair Encoding (BPE) with a shared input–output vocabulary of 32,000 tokens, consistent with previous work on the same datasets. All other hyperparameters, including those for model training, follow the setup described by Vaswani et al. (2017).

Both concatenation and multi-encoder architectures can, in principle, process an arbitrary number of contextual sentences from the past or future. However, most studies limit the context to a few preceding sentences, where relevant discourse information is concentrated, while also mitigating the computational cost associated with the attention mechanism.

In the *self-attention* mechanism of *Transformers* (Vaswani et al., 2017), used in the concatenation NMT model for attending to the context, queries, keys and values are the same vectors. In the HAN module (Miculicich et al., 2018), used in the multi-encoder model, queries are hidden states of the encoder for the current sentence, keys and values are previously encoded hidden states of the encoder for the context sentences. Thus, both models apply attention to contextual information, but in distinct ways: the multi-encoder model attends to each

context sentence individually, with a second-level attention mechanism distinguishing their relative importance, while the concatenation model attends to all context sentences at the same time.

5.2. Experimental Design

In order to analyze how context-aware NMT systems distribute attention over coreferent mentions across sentences in both source and target languages, we applied the following pre-processing steps to the IWSLT data annotated with our framework, using the two CA-NMT models considered.

First of all we translated the data with the two CA-NMT models. The models generated also attention weights from the current sentence to the context sentences, for the source-side context only in the multi-encoder model, for both source and target side context in the concatenation model. We used attention weights obtained as the average of all attention heads. In the multi-encoder model we used the attention heads of the first level of the HAN module. In the concatenation model we used attention heads from the last layer of the encoder, for the source-side context, or decoder, for the target-side context.

The second step was to align the system's input and output sequences to sentences in the IWSLT data. While alignment of input sequences should not be necessary, since NMT models are trained with their own tokenized data, input sequences are not exactly the same, thus they need to be re-aligned. Alignment was performed once again with awesome-align from the target annotated corpus and the system's output.

Using alignments, we retrieved tokens in the system's input and output sequences belonging to coreferent mentions, with the corresponding attention weights. At this point, we were able to compute attention scores over coreference links between mentions in the system's current sentence and mentions in the system's context sentences. These scores were used to perform quantitative automatic analysis based on three metrics introduced in (Dinarelli et al., 2024) and which we describe here for completeness.

In particular, we performed the following post-processing on the attention matrices: i) We filter out attention weights smaller or equal to the value w_u for a given context sentence, $w_u = \frac{1.0}{N}$, where N is the context sentence length. This post-processing filter out small attention weights and keeps only weights potentially meaningful for analyzing the model's behavior; ii) we re-normalize attention weights with respect to the maximum weight in a given context sentence. This post-processing converts into 1.0 the maximum weight, and it allows to compare the multi-encoder to the concatenation model. Since the latter processes concatenated sentences and attention weights sum up to 1, its attention weights have in general smaller values. Renormalizing attention weights over context sentences separately allows us to bring back values to the same scale as the multi-encoder model.

Attention-based metrics exploit coreferences annotated on the IWSLT corpus with our framework, by aligning the corpus data to the system input and output sequences. Once the alignment is performed, tokens in the system's input and output sequences belonging to coreferent mentions can be spotted, and scores for these coreference links can be computed with attention weights from the mentions in the current sentence to their corresponding antecedents in the context sentences.

We compute three evaluation metrics based on attention weights as follows:

1. *Max-weight* metric: is the percentage of coreference links for which the model gave the maximum attention weight with respect to the tokens in the context sentence. This metric reflects how the model has learned to exploit the context to disambiguate coreferences. Ideally, the model should assign a high weight (e.g., 1.0) to the correct coreference link and low weights (e.g., close to 0) to all other tokens in the context.

2. *Non-zero weight* metric: is the percentage of coreference links for which the model assigns an attention weight greater than zero. We expect a concentration of the attention weight distribution in the coreference links, and a minimum amount in the other tokens of the context sentence. This metric shows that the higher the attention from tokens requiring contextual disambiguation to the relevant context tokens, the more the model is correct in using the context.

3. *Average weight* metric: is the average attention weight the model gives to coreference links. This metric is computed by simply summing up the attention weights on all coreference links and dividing the sum by the number of coreference links.

We note that coreferent mentions may be composed of multiple tokens, and the attention mechanism of the model assigns a weight from each token in the current sentence to each token in the context sentence. In order to have only one attention weight for each coreference link, we chose to select the maximum weight. While this may give higher evaluation scores, we believe this choice does not change the overall picture.

Metric NMT model	Max weight	Non-zero weight	Average weight
Multi-encoder (src)	59.2%	66.6%	0.2
Concat (src)	19.9%	40.8%	0.2
Concat (tgt)	28.9%	54.5%	0.3

Table 4: Quantitative results with three different evaluation metrics, over discourse phenomena in the IWSLT corpus, based on attention weights of CA-NMT models.

5.3. Quantitative analysis

Table 4 reports the results of our quantitative evaluation of CA-NMT models using three attention-based metrics: Max-weight, Non-zero weight, and Average weight. These metrics measure how the models attend to coreferent mentions in the context sentences, based on parallel coreference annotations from the IWSLT corpus. For the multi-encoder model, we report results only on the source side, as this system only has access to source-side contextual information. The results indicate that a significant portion of attention is focused on coreferent mentions, highlighting the model's ability to exploit context for disambiguation. For the concatenation model, which has access to both source and target context, attention is distributed across both parts. The Non-zero weight metric shows that nearly half of the coreferent mentions are attended, with average weights around 0.3, indicating moderate attention concentration on relevant mentions.

With respect to the use of our annotated corpora, we observe that target-side data is more frequently attended than source-side data. These findings are consistent with the original work of (Dinarelli et al., 2024), where a similarly annotated corpus (ParCorFull) was used to evaluate the models. This suggests that our framework produces results comparable to existing parallel corpora, while providing many more examples. Overall, these observations demonstrate the potential of our annotation framework to enable detailed analyses of complex neural systems and their handling of cross-sentence discourse phenomena.

6. Conclusion

In this work, we presented COME-ALPs, a framework for creating parallel, coreference-annotated datasets across multiple languages. Our framework allows the projection of coreference annotations from a source language to one or more target languages, maintaining consistent identifiers for mentions and clusters. This feature is crucial for enabling cross-lingual analyses, such as examining how attention is distributed over coreferent mentions in context-aware NMT models. We demonstrated the utility of our framework through two complementary evaluations. First, we evaluated the pipeline on existing multilingual datasets, including ParCorFull2.0, showing that merging multiple coreference tools improves annotation precision to capture cross-lingual coreference links. Second, we showcased a use case on IWSLT data, where the parallel annotations allowed us to analyze attention patterns in two CA-NMT models, revealing how these models distribute attention over coreferent mentions in both source and target contexts. As future work, we plan to integrate coreference resolution tools for target languages to further refine projections and enable more detailed studies of cross-lingual interactions. Additionally, we aim to explore standardization of tools and formats to facilitate broader adoption and reproducibility of parallel coreference resources.

7. Acknowledgements

This work was also supported by the CREMA project (**c**oreference **re**solution in **ma**chine translation), funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01.

8. Bibliographical References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages 1–31.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Chi Chen, Maosong Sun, and Yang Liu. 2020a. Mask-align: Self-supervised neural word alignment. *arXiv preprint arXiv:2012.07162*.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020b. Accurate word alignment induction from neural machine translation. *arXiv preprint arXiv:2004.14837*.

Marco Dinarelli, Dimitra Niaouri, Fabien Lopez, Gabriela Gonzalez-Saez, Mariam Nacklé, Emmanuelle Esperança-Rodier, Caroline Rossi, Di-

dier Schwab, and Nicolas Ballier. 2024. Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention. *Revue TAL : traitement automatique des langues*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse centric evaluation of machine translation with a densely annotated parallel corpus. *arXiv preprint arXiv:2305.11142*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Pavlo Kuchmiichuk. 2023. Silver data for coreference resolution in ukrainian: Translation, alignment, and projection. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 62–72.

Ekaterina Lapshinova-Koltunski, Pedro Augusto Ferreira, Elina Lartaud, and Christian Hardmeier. 2022. *ParCorFull2. 0: A parallel corpus annotated with full coreference*. Saarländische Universitäts-und Landesbibliothek.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022a. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022b. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. *arXiv preprint arXiv:2407.21489*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. *arXiv preprint arXiv:1810.02268*.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. *arXiv preprint arXiv:2209.04280*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. *arXiv preprint arXiv:1909.01383*.

Hao Wang and Yves Lepage. 2016. Combining fast_align with hierarchical sub-sentential alignment for better word alignments. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 1–7.

Juntao Yu, Michal Novák, Abdulrahman Aloraini, Nafise Sadat Moosavi, Silviu Paun, Sameer Pradhan, and Massimo Poesio. 2023. The universal anaphora scorer 2.0. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 183–194.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms giza++. *arXiv preprint arXiv:2004.14675*.

Boyuan Zheng, Patrick Xia, Mahsa Yarmohammadi, and Benjamin Van Durme. 2023. Multilingual coreference resolution in multiparty dialogue. *Transactions of the Association for Computational Linguistics*, 11:922–940.