# Multilingual Coreference Resolution via Cycle-Consistent Machine Translation

Anonymous ACL submission

## Abstract

Coreference resolution is a core NLP task, having a broad range of downstream applications, e.g. machine translation, question answering, document summarization, etc. While the task is well-studied in English, comparatively less attention is dedicated to coreference resolution in other languages, especially low-resource ones. To mitigate this gap, we propose a novel coreference resolution pipeline that harnesses machine translation (MT) from English to a target low-resource language, to generate or expand training data. To automatically validate the quality of the translated samples, we back-translate the samples and assess the similarity with the original English samples via cosine similarity in the latent space of a BERT model. The resulting similarity scores are integrated into the loss function to weight training samples according to their MT cycle consistency. Extensive experiments on four low-resource languages show that our pipeline brings significant performance gains in coreference resolution. Moreover, our pipeline enables accurate coreference resolution in languages where no previous corpora were available. We publicly release our code at https://anonymous.4open.science/r/NewCoref-D3B8/.

## 1 Introduction

Coreference resolution (CR) is a fundamental NLP task, which aims to identify all expressions in a text that refer to the same entity. The first attempts at solving the CR problem were heavily based on human-designed rules for the English language (Hobbs, 1978; Ng, 2005; Ponzetto and Strube, 2006; Raghunathan et al., 2010). These types of methods are limited by the difficulty of drawing a complete list of non-contradictory rules, and are exposed to problems associated with the statistical nature of language. The foundational work of Lee et al. (2017) was set to address CR by creating a fully trainable solution, without human-designed linguistic rules. The authors introduced the first end-to-end neural system, using a bidirectional LSTM to produce contextual span representations for joint mention detection in English. Deep models later benefited from the emergence of better neural encoders (Joshi et al., 2019), such as BERT (Devlin et al., 2019). While end-to-end models reach competitive results (Kirstain et al., 2021; Xu and Choi, 2020), they usually have many task-specific hyperparameters and are hard to tune, as stated by Zhang et al. (2023).

More recently, researchers introduced a new category of sequence-to-sequence solutions (Urbizu et al., 2020; Liu et al., 2022; Bohnet et al., 2023; Straka, 2023), aiming to generate text representations of entity clusters. Notably, CorPipe (Straka, 2023) won the CRAC 2023 shared task on multilingual coreference resolution, while CorPipeEnsemble ranked first in the CRAC 2025 (unconstrained) edition. Another direction of study is the use of zero-shot large language models (LLMs) via prompting. Le and Ritter (2024) found that, although the zero-shot performance of promoted LLMs is respectable, they still remain way below specialized state-of-the-art models, by 10-20% on benchmarks like CoNLL-2012/OntoNotes (Pradhan et al., 2012). The CRAC 2025 results (Novák et al., 2025) also indicate that zero-shot LLMs lag far behind specialized models, with a clear gap of about 13% in terms of F1.

These empirical observations underline the utility of task-specific datasets used to train and test specialized CR models. However, CR datasets in certain languages are small, outdated, or entirely missing. There are clear efforts to remedy this situation, e.g. the CRAC 2025 shared task (Novák et al., 2025) describes CorefUD as a harmonized multilingual collection of 22 datasets in 17 languages. By contrast, for low-resource languages, such as Romanian, we did not find any CR datasets

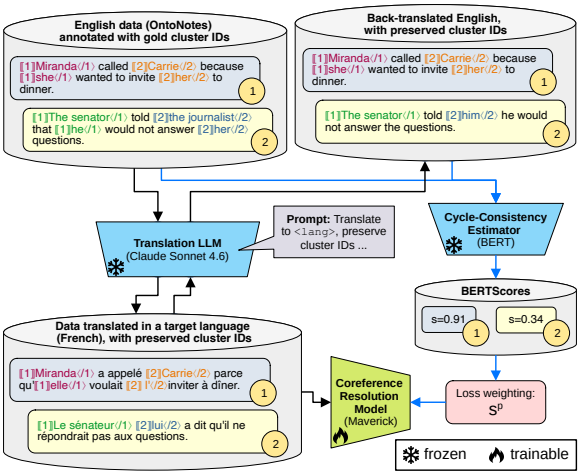


Figure 1: Overview of the proposed pipeline for coreference resolution. An LLM, namely Claude Sonnet 4.6 (Anthropic, 2026), is prompted to translate annotated samples from English to the target language and back. The cycle consistency of back-translations is estimated via BERTScore (Zhang et al., 2020). Finally, the CR model, namely Maverick (Martinelli et al., 2024), is trained on the target language, weighting the loss of each translated sample with $s^p$, where $s$ represents the BERTScore of the respective sample, and $p$ is a hyperparameter that controls the importance of cycle consistency. Best viewed in color.

that can be used in the evaluation and training of specialized models. To make things worse, it is reasonable to expect that the zero-shot performance in such languages is even lower.

To this end, we propose a novel CR framework that leverages existing English resources via machine translation (MT) to generate new training data in a target low-resource language. As illustrated in Figure 1, we employ back-translation and assess the overlap between original and back-translated English samples, where the overlap is given by the cosine similarity computed in the embedding space of a pre-trained BERT model (Devlin et al., 2019; Zhang et al., 2020). We conjecture that the utility of a translated data sample is proportional to its cycle consistency, i.e. the cosine similarity of its back-translation. Therefore, we integrate the cosine similarity between original and back-translated English samples into the loss function, to weight the importance of translated samples according to their cycle consistency.

To validate the proposed framework, we perform experiments across four low-resource languages: French, Hungarian, Romanian and Russian. While three of these languages have small-scale publicly available CR datasets, there are no CR resources for Romanian. The results confirm that our cycle-consistent MT augmentation framework can significantly boost performance in CR across all four languages, in both training dataset expansion and training dataset generation scenarios.

In summary, our contribution is threefold:

- We propose a novel CR framework based on MT to generate new training samples for low-resource languages, and modulate sample importance according to MT cycle consistency.
- We conduct comprehensive experiments across four low-resource languages, showing that the proposed framework can significantly boost CR performance.
- We manually curate a coreference resolution test set for Romanian, thus enabling the evaluation of CR systems for this low-resource language.

## 2 Method

Our model extends Maverick (Martinelli et al., 2024) with three modifications, to make it suitable for CR in low-resource languages. First, we replace the English-only encoder DeBERTa-v3-large (He et al., 2023) with mmBERT-base (Marone et al., 2025), a multilingual encoder pre-trained on 200+ languages. In this way, a single model can be employed across multiple languages. Second, we separate training into two phases: (i) train the mention detector with the frozen encoder, and (ii) fine-tune the encoder and the coreference heads using gold mentions as input, isolating the linking signal from mention-detection noise. Third, we augment the bilinear coreference scorer (Lee et al., 2017) with MT cycle consistency, providing a discriminative signal, independent of encoder representations.

**Generating data via MT.** The lack of large-scale CR resources in many languages motivates our MT-based augmentation strategy. We employ a highly capable LLM to perform MT, namely Claude Sonnet 4.6 (Anthropic, 2026). As illustrated in Figure 1, each source (English) document is translated using Claude Sonnet (Anthropic, 2026) via zero-shot prompting (the exact prompt is specified in Table 3). The prompt instructs the model to produce a fluent target-language translation, while preserving every $[\![k]\!] \ldots \langle/k\rangle$ span around the target-language equivalent of each English mention, thus maintaining all cluster identifiers $k \in \{1, 2, ..., K\}$, where $K$ is the number of entities.

**Back-translation quality scoring.** Translation errors introduce noise into the projected annotations.

To quantify this noise per document, we employ back-translation to complete the translation cycle: each target language translation is itself submitted to Claude Sonnet 4.6 with a symmetric prompt requesting translation back to English, while preserving all cluster markers. The back-translated English text is then compared with the original English source via BERTScore (Zhang et al., 2020), yielding a per-document quality score $s \in [0, 1]$. Our intuition is that a high-fidelity translation followed by a faithful back-translation recovers text semantically close to the source, whereas a translation that drops or misaligns mentions produces a divergent back-translation.

**Per-document loss weighting.** Rather than applying a hard threshold to discard low-quality documents, which would lose potentially useful training data, we incorporate BERTScore directly into the training objective. Each document $D$ contributes with the following weight:

$$w_D = s_D^p, \tag{1}$$

where $s_D$ is the BERTScore between source and back-translated versions of document $D$, and $p \geq 0$ controls the strength of the penalty. Then, the weighted training objective becomes:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} w_D \cdot \mathcal{L}_D(\theta), \tag{2}$$

where $\mathcal{D}$ is the collection of translated documents (originally available in English), $\theta$ represents the parameters of the CR model, and $\mathcal{L}_D$ is the per-document loss for the current training phase. For training phase (i), $\mathcal{L}_D$ is the standard binary cross-entropy on mention start/end logits, while for phase (ii), it is the marginal log-likelihood over gold antecedents (Lee et al., 2017).

## 3 Experiments

**Datasets.** For French, we use the ANCOR corpus (Muzerelle et al., 2014), which contains 530 transcripts of spontaneous spoken French drawn from interviews, conversations, and oral surveys. For Hungarian, we use SzegedKoref (Vincze et al., 2018), a dataset comprising 320 short editorial and news documents annotated for nominal coreference. For Russian, we use RuCor (Toldova et al., 2014), a corpus formed of 180 texts covering news, scientific articles, blog posts and fiction. For French, Hungarian and Russian, where the native gold data is small or domain-restricted, we supplement the within-language training data with LLM-translated OntoNotes 5.0 documents to expand both volume and domain diversity. For data augmentation via MT, we choose OntoNotes 5.0 (Weischedel et al., 2013; Pradhan et al., 2012) as the source English corpus, as it spans a broad range of genres: newswire, broadcast news, broadcast conversation, magazine, web text, telephone speech, and biblical text. For Romanian, no publicly available coreference dataset exists. We therefore construct a Romanian dataset entirely from English documents drawn from OntoNotes 5.0 (Weischedel et al., 2013; Pradhan et al., 2012). The documents are translated with Claude Sonnet 4.6, which is instructed to preserve annotations. Further, the corresponding Romanian test set is manually verified and corrected by a native Romanian speaker to ensure that translation, mention boundaries, and coreference links are correct.

**Evaluation measures.** Following Martinelli et al. (2024), we employ three evaluation metrics, namely MUC (Vilain et al., 1995), $B^3$ (Bagga and Baldwin, 1998), and CEAF-E ($\phi_4$-CEAF) (Luo, 2005). For each of them, we report the precision (P), recall (R), and F1 scores. We also report the CoNLL F1 score, which is defined as the average F1 score of the MUC, $B^3$ and CEAF-E metrics.

**Hyperparameter setup.** We train Maverick (Martinelli et al., 2024) using AdamW, with a learning rate of $10^{-4}$ and a mini-batch size of 16. We use a gradient clipping of 1.0. We train using early stopping with a patience of 20 epochs, and select the best model via CoNLL F1 on validation data. All other hyperparameters are left to their default values. Our pipeline introduces a single extra hyperparameter into Maverick (Martinelli et al., 2024), namely the power $p$ in Eq. (1). We tune $p$ on validation data for one language (French), considering values for $p \in \{0.5, 1, 2, 3\}$. The optimal value $p = 3$ is kept across all languages to avoid overfitting in hyperparameter space.

**Results.** For French, Hungarian and Russian, we compare three alternatives: a base Maverick model (trained on original within-language data), a Maverick model that benefits from MT data (trained on both original and translated data), and a Maverick model that benefits from MT data, but penalizes MT samples according to their cycle consistency (back-translation quality). For Romanian, there is no within-language data available, so we replace the base Maverick model with a zero-shot LLM (we use the same LLM as for MT, namely Claude Sonnet 4.6). In Table 1, we present comparative results across four low-resource languages. The results

| Language→ | | English | French | | | Hungarian | | | Romanian | | | Russian | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method→ | | Base | Base | +MT | $+s^p$ | Base | +MT | $+s^p$ | ZS | +MT | $+s^p$ | Base | +MT | $+s^p$ |
| **MUC** | **P** | 95.0 | 85.2 | 85.5 | **89.3** | 87.3 | 87.5 | **88.9** | 70.1 | 86.2 | **87.8** | 95.5 | 95.6 | **95.9** |
| | **R** | 95.7 | 82.3 | 83.4 | **88.0** | 88.9 | 89.5 | **90.8** | 66.0 | 84.8 | **86.5** | 95.8 | 96.2 | **96.5** |
| | **F1** | 95.4 | 83.7 | 84.4 | **88.6** | 88.1 | 88.5 | **89.8** | 68.0 | 85.5 | **87.1** | 91.7 | 92.0 | **92.4** |
| **$B^3$** | **P** | 88.1 | 80.8 | 80.2 | **85.6** | 84.2 | 84.2 | **85.4** | 65.8 | 81.5 | **83.2** | 92.1 | 92.0 | **92.3** |
| | **R** | 91.7 | 79.0 | 80.6 | **85.4** | 89.3 | 90.2 | **91.2** | 62.4 | 80.9 | **82.4** | 93.5 | 94.0 | **94.3** |
| | **F1** | 89.9 | 79.9 | 80.7 | **85.5** | 86.7 | 87.1 | **88.2** | 64.1 | 81.2 | **82.8** | 92.8 | 92.9 | **93.3** |
| **CEAF-E** | **P** | 92.7 | 78.1 | 78.6 | **83.7** | 87.8 | 88.2 | **89.1** | 62.8 | 80.1 | **81.9** | 92.6 | 92.8 | **93.1** |
| | **R** | 86.2 | 73.2 | 73.5 | **79.0** | 80.1 | 80.7 | **82.0** | 57.9 | 75.4 | **71.1** | 90.8 | 91.1 | **91.6** |
| | **F1** | 89.4 | 75.6 | 76.0 | **81.3** | 83.8 | 84.2 | **85.4** | 60.2 | 77.7 | **79.4** | 91.7 | 92.0 | **92.4** |
| **CoNLL** | **F1** | 91.6 | 79.7 | 80.4 | **85.1** | 86.2 | 86.5 | **87.8** | 64.1 | 81.5 | **83.1** | 93.4 | 93.6 | **94.0** |

Table 1: Coreference resolution results on four target languages (French, Hungarian, Romanian, Russian), measured with the official CoNLL-2012 scorer. Best score per language is highlighted in **bold**. Legend: **base** – Maverick trained on original target language data; **ZS** – zero-shot LLM (when no original training data is available); **+MT** – Maverick trained with translated examples; **+s$^p$** – Maverick trained with translated examples and cycle-consistent loss weighting. For reference, we report results on English with the **base** model.

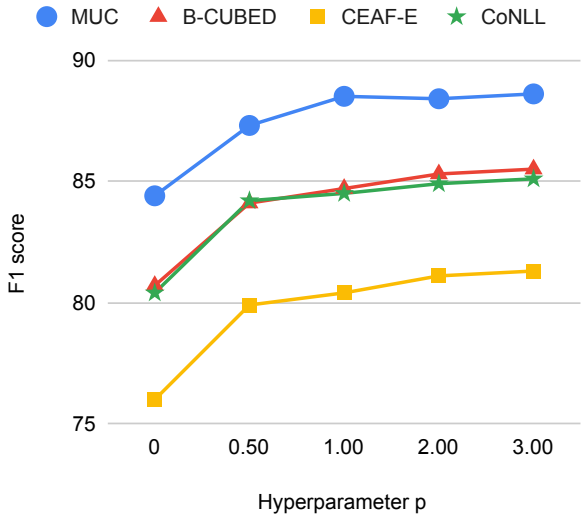


Figure 2: Ablation of hyperparameter $p$, which controls the impact of loss weighting in Eq. (2). Best viewed in color.

| Method | French | Hungarian | Romanian | Russian |
|---|---|---|---|---|
| Base | 79.7 | 86.2 | 64.1 | 93.4 |
| $+s^p$ (BLEU) | 83.7 | 86.8 | 81.5 | 93.7 |
| $+s^p$ (BERTScore) | **85.1** | **87.8** | **83.1** | **94.0** |

Table 2: BLEU vs. BERTScore comparison (in terms of CoNLL F1), as alternatives for the semantic similarity score $s$ used in our cycle-consistent loss weighting, across all four target languages.

indicate that MT-based data augmentation is beneficial, especially for Romanian, where there are no available coreference resolution corpora. Furthermore, we observe additional performance gains when introducing cycle-consistent loss weighting. Here, the improvements stem primarily from higher precision on MUC and $B^3$, suggesting that loss weighting based on $s^p$ helps suppress spurious mentions introduced by translation artifacts.

**Ablation of loss weighting hyperparameter.** In Figure 2, we vary the hyperparameter $p$ in Eq. (1), considering values in the set $\{0, 0.5, 1, 2, 3\}$. Note that $p = 0$ turns off the BERTScore weighting in Eq. (2). The ablation of $p$ is performed on the French dataset. The results show that higher values of $p$ lead to better results, confirming that our cycle-consistent loss weighting is very useful. However, going from $p = 2$ to $p = 3$, we observe that the performance gains begin to saturate.

**BERTScore vs. BLEU.** In Table 2, we compare two alternatives to measure MT cycle consistency, namely BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). While both BLEU and BERTScore bring visible performance boosts across all four languages, BERTScore consistently outperforms BLEU. This is likely due to the fact that BLEU does not always capture semantic relations, such as synonymity.

## 4 Conclusion

We proposed a novel pipeline for coreference resolution in low-resource languages, which harnesses MT to augment existing datasets or generate new training data (for languages where CR resources were not previously available). We assessed MT cycle consistency and introduced it in the loss function of the CR model to modulate the importance of translated data samples accordingly. To validate our approach, we conducted CR experiments across five low-resource languages. Our results demonstrated that our pipeline leads to significant performance gains, and even enables CR in languages without existing resources. In future work, we aim to expand the list of low-resource languages.

## 5 Limitations

Our framework leverages the use of a highly capable LLM in the translation phase. While translated data is central to our framework, as it brings significant performance gains, LLM usage can also represent a downside of our framework, introducing some limitations, as detailed below.

First, LLMs are typically power-hungry models, having potentially negative effects on the environment due to their high energy consumption. As humanity will gradually move towards green energy production alternatives, the importance of the energy consumption problem of LLMs will diminish in the future. Moreover, we highlight that the translated data is meant to be reused multiple times to train and validate lighter models for coreference resolution. Hence, we limit LLM usage to the MT step, and refrain from fine-tuning LLMs for coreference resolution.

Second, potential biases of the LLM may eventually be transferred into the translated data, and later be inherited by the smaller coreference resolution model. We have manually inspected the translated examples and did not observe any age, gender, racial, or other kinds of biases. Nevertheless, our careful inspection does not completely exclude this possibility, especially for document genres and languages that are not included in our study.

## References


Anthropic. 2026. Claude sonnet 4.6 model card. Technical report, Anthropic.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING)*, pages 79–85.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style pre-training with gradient-disentangled embedding sharing. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5802–5807.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14–19.

Nghia T. Le and Alan Ritter. 2024. Are large language models robust coreference resolvers? In *Proceedings of Conference on Language Modeling (COLM)*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 993–1005.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 25–32.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13380–13394.

Judith Muzerelle, Anaïs Lefeuvre, Emmanuel Schang, Jean-Yves Antoine, Aurélie Pelletier, Denis Maurel, Iris Eshkol, and Jeanne Villaneau. 2014. ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 843–847.

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1081–1086.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2025. Findings of the fourth shared task on multilingual coreference resolution: Can LLMs dethrone traditional approaches? In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 95–118.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL (NAACL-HLT)*, pages 192–199.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 492–501.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution (CRAC)*, pages 41–51.

Svetlana Toldova, A. Roytberg, A.A. Ladygina, M.D. Vasilyeva, I.L. Azerkovich, M. Kurzukov, G. Sim, D.V. Gorshkov, A. Ivanova, A. Nedoluzhko, and Y. Grishina. 2014. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. In *Proceedings of the Annual International Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, volume 13, pages 681–694.

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2020. Sequence to sequence coreference resolution. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 39–46.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia*, pages 45–52.

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy, and Richárd Farkas. 2018. SzegedKoref: A Hungarian coreference corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 401–405.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, and 1 others. 2013. OntoNotes Release 5.0. Technical report, Linguistic Data Consortium.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533.

Tianyao Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11493–11504.

# A Appendix

## A.1 Translation Prompt

To translate English documents annotated with entity clusters for coreference resolution, we employ Claude Sonnet 4.6 (Anthropic, 2026). The generic prompt used during translation from English to a target language <lang> is given in Table 3. In the prompt template, <lang> is replaced with one of target languages, namely French, Hungarian, Romanian and Russian. To translate documents back to English, we use a symmetric prompt. The employed prompt comprises precise rules, especially regarding the preservation of annotations, which are particularly important for the underlying CR task. We also exemplify the rules via an example, to further explain to the LLM how the translation should be performed.

## A.2 Compute Environment

We perform all our experiments on an academic compute environment, namely a workstation with a single Nvidia GeForce GTX 3090 GPU with 24 GB of VRAM. The reported results represent averages over three runs.

## A.3 Romanian Data Annotation

The annotator employed to verify and correct the English→Romanian translations is an adult who

```
You are translating English text to <lang> while PRESERVING coreference cluster annotations.

The input text contains inline coreference markers:
  -- [[N]]word[/N] marks a mention belonging to cluster N (an integer)
  -- All mentions of the SAME entity share the SAME cluster ID
  -- Markers can be nested: [[1]]the CEO of [[2]]Acme[/2][/1]

YOUR TASK:
1. Translate the entire text to fluent, natural <lang>.
2. CRITICAL: every English mention [[N]]...[/N] MUST appear in the <lang> translation with the
SAME cluster ID N, wrapping the <lang> equivalent of that mention.
3. Pronouns count as mentions. If "he" appears with ID 5 in English, the <lang> equivalent
pronoun (or whichever inflected form fits) MUST also be marked [[5]]...[/5].
4. If a mention is dropped because <lang> doesn't express it overtly (e.g. pro-drop subject),
still emit empty markers [[N]][/N] at the dropped position to preserve the cluster.
5. Keep brackets balanced and properly nested. Every [[N]] MUST have a matching [/N].
6. Do NOT introduce new cluster IDs. Use only the IDs present in the English text.
7. Output ONLY the <lang> text with annotations. No explanations, no preamble, no markdown
fences, just the annotated translation.

EXAMPLE INPUT:
[[1]]John[/1] went to [[2]]the store[/2]. [[1]]He[/1] bought [[3]]milk[/3], and
[[1]]John[/1] told [[4]]his wife[/4] about [[3]]it[/3].

EXAMPLE OUTPUT:
[[1]]Ion[/1] s-a dus [[2]]la magazin[/2]. [[1]]El[/1] a cumpărat [[3]]lapte[/3],
iar [[1]]Ion[/1] i-a spus [[4]]soţiei sale[/4] despre [[3]]asta[/3].

NOW TRANSLATE THIS TEXT (output the <lang> translation with annotations only):
{english_text}
```

Table 3: Prompt used for Claude Sonnet 4.6 (Anthropic, 2026) to translate documents annotated with coreference resolution clusters from English to a low-resource language <lang>, where <lang> is one of the following four languages: French, Hungarian, Romanian, Russian. The output example is shown for Romanian.

holds a master degree at a university located in Romania. The recruited annotator willingly agreed to engage in the annotation process, after agreeing to our terms and conditions. The authors provided accurate and complete instructions regarding the annotation task. The annotator was also given the LLM prompt. A fair compensation (25 EUR per hour) was paid to the annotator, upon completing the annotations. This is almost double the average wage in Romania (13.6 EUR per hour)[1]. The authors verified the manual annotations to confirm that the annotation task was carefully completed by the recruited annotator, according to the provided instructions.

### A.4 Romanian Data License Agreement

The Romanian version of OntoNotes 5.0 will be released under the LDC User Agreement for Non-Members.

[1] https://www.romania-insider.com/eurostat-romanians-working-hours-salaries-april-2026