Silver Data for Coreference Resolution in Ukrainian: Translation, Alignment, and Projection

Pavlo Kuchmiichuk
University of Rochester
pavlo.kuchmiichuk@rochester.edu

Abstract

Low-resource languages continue to present challenges for current NLP methods, and multilingual NLP is gaining attention in the research community. One of the main issues is the lack of sufficient high-quality annotated data for low-resource languages. In this paper, we show how labeled data for high-resource languages such as English can be used in low-resource NLP. We present two silver datasets for coreference resolution in Ukrainian, adapted from existing English data by manual translation and machine translation in combination with automatic alignment and annotation projection. The code is made publicly available[1].

1 Introduction

Coreference resolution is a task that requires clustering the mentions in text that refer to the same underlying entity (Poesio and Artstein, 2008). For example, in a sentence "John asked Dan to drive him to work because his car was broken.", "John", "him", and "his" belong to the same cluster. In the past few years, significant progress has been achieved for English coreference resolution. To compare, in 2017 the average F_1 score of the model with the best performance in the CoNLL-2012 shared task (Pradhan et al., 2012) was 67.2 (Lee et al., 2017), and in 2021 the score of the best model increased to 81.0 (Dobrovolskii, 2021).

While the task of entity coreference resolution is relatively well-researched in English, it remains uncharted territory for many other languages. Models developed with English in mind often fail to perform on the same level if used for a different language (Joshi et al., 2020).

One of the prevalent issues when working with low-resource languages is the lack of annotated data. It is often complicated to find or compile high-quality datasets even for English, and labeled data for complex tasks in other languages is much rarer. Such data scarcity hinders NLP progress, as many state-of-the-art models require large amounts of labeled texts for training which are not available for low-resource languages (Ruder, 2020; Fincke et al., 2022). Building new high-quality datasets is essential for expanding NLP research to the low-resource setting.

One way of mitigating the data scarcity issue is adapting the data collected for high-resource languages. In this paper, we present two silver datasets annotated for coreference resolution in Ukrainian, both of which are built using existing English labeled data as a basis. First, we manually translated the Winograd Schema Challenge Dataset (Levesque et al., 2012) into the Ukrainian language. This dataset contains pairs of sentences with an ambiguous anaphor that can only be resolved using world knowledge and reasoning. Second, we used a machine translation model to translate texts from OntoNotes 5.0 (Weischedel et al., 2011) into Ukrainian, followed by automatically aligning and projecting annotations based on the cross-attention layer of an encoder-decoder model. Our approach allows efficiently creating silver datasets based on existing high-quality data, which can then be used to extend language model training to low-resource languages.

2 Related Work

In this section, we analyze different techniques commonly utilized to help with data scarcity issues in multilingual NLP.

2.1 Cross-lingual Transfer Learning

Methods such as cross-lingual transfer learning make it possible for the models to learn meanings of words across different languages simultaneously. Large pre-trained multilingual language models can transfer the knowledge learned from labeled data available in abundance for high-resource languages to low-resource ones.

[1] https://github.com/pkuchmiichuk/ua-coref

Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP), pages 62–72
May 5, 2023 ©2023 Association for Computational Linguistics

Cross-lingual transfer learning relies on finding a shared cross-lingual space for languages in the system. Aligning the source and target embedding spaces is one of the methods commonly used for this. Recently, pre-trained multilingual encoders have also been shown to yield good performance on various NLP tasks (Xu and Murray, 2022).

Pires et al. (2019) demonstrate how mBERT, a language model pre-trained on 104 languages, is able to generalize quite well for NLP tasks in different languages. The authors perform NER and POS tagging experiments to show that mBERT performs the cross-lingual transfer quite well, considering the model does not see any markers that denote the input language on the pre-training stage. Instead, mBERT is able to capture multilingual representations of words. The representations capture useful linguistic information in a language-agnostic way, which allows the model to handle knowledge transfer even across languages with different scripts.

Wu and Dredze (2019) reaffirm this conclusion after exploring the performance of mBERT on tasks such as document classification, natural language inference, named entity recognition, POS tagging, and dependency parsing. While learning multilingual representations, the model also retains language-specific information which contributes to its capabilities.

Models that capture multilingual representations of words can be especially useful for word alignment, showing robust performance on different language pairs (Dou and Neubig, 2021). Researchers have also explored cross-lingual learning for coreference resolution in particular. Cruz et al. (2018) use a large Spanish corpora to create a model for Portuguese, leveraging FastText multilingual embeddings. Urbizu et al. (2019) work on coreference resolution for Basque, relying on English data from OntoNotes to train a cross-lingual model.

2.2 Domain Adaptation

Domain adaptation involves training a language model for a task in a specific domain without having enough data to train the model directly on in-domain data. As domain adaptation specifically aims to overcome lack of in-domain data issues, it is especially useful when working with low-resource languages. For example, one can consider data in a high-resource language such as English out-of-domain, and train language models for a target low-resource language using domain adaptation techniques.

Xu et al. (2021) introduce a gradual fine-tuning approach for domain adaptation. This contrasts with the common approaches to the task, using which the model is pre-trained on out-of-domain data and fine-tuned on in-domain data in one stage. Instead, the authors propose an iterative multi-stage fine-tuning method: the model is gradually fine-tuned on datasets composed both of out-of-domain and in-domain data. On each subsequent iteration, the percentage of in-domain data in the dataset increases. In training, this steers the model into target domain direction gently instead of using it as-is in zero-shot setting or directly performing one-stage fine-tuning on the whole target dataset. The gradual fine-tuning approach shows promising results for NLP tasks like dialogue state tracking and event extraction, outperforming both the pre-trained models and models fine-tuned using one-stage method. The authors conduct event extraction experiments on English and Arabic datasets, which highlights how the method can be utilized for working with low-resource languages. It is shown that gradual fine-tuning significantly improves results in comparison with baseline models.

Maurya et al. (2021) suggest using an additional pre-training step before fine-tuning the language model for solving natural language generation tasks in the target low-resource language. This allows the model to overcome the problem of mismatch between pre-training and fine-tuning objectives. Introducing an auxiliary task as an additional pre-training step improves the multilingual word representation and helps warm-start the model for performing a specific task in target language.

Xu and Murray (2022) use mixed fine-tuning to overcome the deficiencies of the common approach to target domain adaptation. Instead of focusing at one language at a time, mixed fine-tuning allows to use one multilingual model to handle many target languages at once and avoid overly language-specific models. A stochastic gradient surgery technique is used to mitigate the issue of conflicting gradients among different languages. The significant performance increase is specifically important for languages linguistically distant from English. This affirms that abruptly shifting to the target domain by one-stage fine-tuning can hinder the model, while mixed fine-tuning helps it to learn the representations more smoothly.

Knowledge transfer is also important for coreference resolution specifically. Xia and Van Durme (2021) demonstrate the effectiveness of continued training for multilingual coreference resolution, which involves first training a model on a source dataset until convergence, and then using it to train a second model on a target dataset. Yuan et al. (2022) use active learning for situations where no substantial in-domain, labeled data is available; this approach explores different sampling strategies for further labeling and continued training.

Domain adaptation methods such as gradual or mixed fine-tuning can be especially useful when using silver target datasets for training, helping to overcome the inherent noise problem.

2.3 Annotation Projection

Manual annotation for coreference resolution is a particularly challenging task because of the variety of coreference phenomena and the lack of standardized annotation guidelines. Automatic projection approach allows using the annotated data in the source language to transfer the linguistic annotation to unlabeled target data.

As outlined by Nateras (2022), common approaches to annotation projection usually utilize sentence-aligned parallel corpora or neural machine translation systems. Correct word alignment is crucial for the quality of the projected annotations; both automated and manual alignment methods have been proposed. As opposed to training the models exclusively on labeled data in the source language, target data with projected labels, while noisy, allows directly leveraging linguistic features of the target language.

Grishina (2019) provides a comprehensive overview of annotation projection methods applied to coreference resolution specifically. The discussed studies range from experimenting with manual projection of coreference chains (Harabagiu and Maiorano, 2000) to fully relying on translation-based approaches (Ogrodniczuk, 2013). In most of the works, English has been used as a source language; however, projecting from multiple other languages at the same time has been shown to improve the quality of the projected annotations. Grishina (2019) also conducts three annotation projection experiments using statistical word alignment with GIZA++ (Och and Ney, 2003) as well as mention extractors for the source and target languages.

Yarmohammadi et al. (2021) explore data projection and the use of silver data in zero-shot cross-lingual information extraction. The authors conduct experiments on English-Arabic annotation projection. Specifically, they translate the source text to the target language using a machine translation system, obtain word alignments using publicly available automatic tools, and directly project the annotations along the word alignments. The created silver data is then used to augment the training set.

Our approach to annotation projection presented in this work is similar to that of Yarmohammadi et al. (2021), with some important differences. First, we aligned the words based on cross-attention of the machine translation model rather than relying on statistical or embedding-based alignment. Second, when projecting a multi-token span, our approach allows multiple projected spans in the target text, while Yarmohammadi et al. (2021) decided to form a contiguous span containing all aligned tokens from the same source span, potentially including tokens not aligned to the source span in the middle.

3 Ukrainian Coreference Resolution

In this section, we present a survey of existing research for coreference resolution in Ukrainian.

Hlybovets (2018) focuses on building complex information processing systems using a concept of agent-based modeling. A coreference resolution module is presented as part of the bigger system. For detecting mentions, the author uses a rule-based NER system based on a generalized left-to-right (GLR) parser; the system can detect PER, ORG and LOC entities. A manually annotated news corpus is used for NER testing; the system achieves 0.48 F_1 score. To determine if the mentions are coreferent, methods from Soon et al. (2001) and Raghunathan et al. (2010) are adapted: the resulting system uses a multi-pass filtering sieve together with a decision tree classifier. The author does not report accuracy scores for this part of the system.

Pogorilyy and Kramov (2019) attempt to create a coreference resolution system for Ukrainian using a convolutional neural network. Following Clark and Manning (2016), coreference resolution is presented as a clustering task. Every entity in the text is considered a separate cluster at the initialization step. The task of the model is then to go over pairs of clusters and merge the ones that refer to the same entity.

For creating the clusters, the system proposed in Pogorilyy and Kramov (2019) uses a rule-based filtering sieves module and a multichannel CNN module. The rules are mostly based on direct string comparison with regular expressions, although some of them incorporate dictionaries of entity names scraped from Wikipedia. Then, pairs of clusters are given as an input to a convolutional neural network. Clusters are represented by averaging the word2vec embeddings of the corresponding entity words. The CNN module works as a binary classifier; to train it, the authors use the SEARN method adapted from Clark and Manning (2016). A dataset of Ukrainian news articles is used for training, testing, and evaluation. The model achieves 92.11 F_1 score for the B^3 coreference evaluation metric.

In Telenyk et al. (2021) the authors continue the work presented previously, now making some important changes. First, BiLSTM is trained instead of a CNN. Second, they perform feature analysis and conclude that word embeddings used for mention representation contribute a lot to the result, so they turn to ELMo embeddings instead of word2vec used previously. As for the results, the proposed model achieves 92.21 F_1 score for the B^3 coreference evaluation metric. Another important contribution is that both the pre-trained model and the dataset are made available to the public, which allows using them as a baseline for continuing research in this direction (Kramov, 2021).

The model can be used for different tasks; three endpoints are available to extract mentions, estimate coherence of the text and extract coreferent pairs. It also attempts to perform POS tagging and extract other grammatical features described in the previous checkpoint of the evaluated texts.

The Coreferent Clusters dataset presented by Kramov (2021) is, to our knowledge, the only publicly available dataset for coreference resolution in Ukrainian. It is distributed via Mendeley as a MYSQL database. The database contains a single table *word* with the texts and relevant information about the tokens: their parts of speech, case, animacy, gender, number, aspect, and mood. All mentions are labeled, including singletons, which are potentially coreferent but appear only once in a document. Each non-singleton mention belongs to a coreferent cluster. It is unclear whether the dataset was annotated manually or automatically, as the authors provide no specifics about the corpus creation process. Table 1 demonstrates the detailed statistics of Coreferent Clusters as well as the silver Ukrainian OntoNotes dataset presented in Section 5.

Overall, the task of coreference resolution in Ukrainian remains underresearched. High-quality annotated datasets are needed to appropriately evaluate the performance of existing models as well as train new ones using state-of-the-art NLP methods.

4 Data

In this section, we describe the base English datasets used to form the silver Ukrainian datasets. We explore a total of two source datasets: a smaller test dataset that allows for testing coreference resolution systems ability to deal with complicated anaphora ambiguity cases, as well as a large dataset commonly used for training coreference resolution models.

4.1 Winograd Schema Challenge Dataset

A Winograd schema is a pair of sentences that have only a slight difference in words, but contain an anaphora ambiguity that can only be resolved with world knowledge and reasoning. Such sentences can be quite complicated for coreference resolution models to solve, while human readers usually easily deal with them. An example of a Winograd schema are sentences such as:

1. The man couldn't lift his son because he was so weak.
2. The man couldn't lift his son because he was so heavy.

The pronoun "he" corefers with "the man" in the first sentence, and with "son" in the second.

The English Winograd Schema Challenge dataset contains 285 Winograd schema sentences that cover a wide range of linguistic features and world knowledge (Levesque et al., 2012). The size of the collection is understandably limited, as creating a large and diverse set of high-quality Winograd schemas is quite difficult. The goal of the Winograd Schema Challenge dataset is then not to provide a training dataset, but rather test language models that claim to have solved the problem of coreference resolution and pronoun disambiguation.

Translations of the WSC dataset are available in Chinese, Japanese, French, Portuguese, and Hebrew. Authors of French and Portuguese transla-

Dataset	Documents	Sentences	Tokens	Mentions	Clusters
Coreferent Clusters (Kramov, 2021)	2,528	17,122	361,534	24,257	8,538
Ukrainian OntoNotes	3,493	94,269	1,456,717	201,700	44,071

Table 1: The statistics of Ukrainian coreference resolution corpora. The counts for mentions do not include singletons.

tions made a few changes to the schemas in order to avoid unintended cues such as grammatical gender.

In a quite extensive survey of WSC, Kocijan et al. (2022) highlight three main methods commonly used to solve the Winograd Schema Challenge. Feature-based approaches rely on extracting relevant information in the form of sentence or word features and are usually rule-based. Neural approaches take advantage of semantic similarities between word embeddings or use RNNs for encoding the local context. Finally, the third group includes approaches that use large language models pre-trained on huge text corpora.

While the Winograd Schema Challenge has been largely overcome as originally formulated, the problem of commonsense reasoning still stands as one of the major NLP challenges. The low-resource setting makes solving the task even harder: having annotated collections such as the WSC dataset in other languages is vital to exploring different aspects of commonsense reasoning. We decided to manually translate the WSC dataset to Ukrainian, attempting to preserve the ambiguity of the schemas.

4.2 OntoNotes 5.0

OntoNotes 5.0 (Weischedel et al., 2011) is a large dataset containing annotations of syntactic parse trees, named entities, semantic roles, and coreference. The dataset contains texts of multiple genres such as telephone conversations, newswire, broadcast news, broadcast conversation, web text, and religious text. OntoNotes 5.0 is also multilingual, as it contains English, Chinese, and Arabic subsets.

The coreference annotation in OntoNotes connects coreferring instances of specific referring expressions, primarily noun phrases that introduce or access a discourse entity. Notably, the annotation does not include singletons–clusters containing only one entity. OntoNotes 5.0 is the primary dataset for experiments on coreference resolution, as it is used as a standard in CoNLL-2012 shared task (Pradhan et al., 2012).

We use OntoNotes 5.0 as a basis for automatically translating, aligning, and projecting annotations to create a silver Ukrainian version of it.

5 Silver Data Creation

In this section, we describe the methods and the process of creating silver Ukrainian coreference resolution datasets on the basis of high-quality English data.

5.1 Manual Translation

For the Ukrainian version of the Winograd Schema Challenge dataset, we manually translated the English schemas. In the process of translation, English proper names were replaced with Ukrainian ones. The resulting corpus contains 263 Winograd schema sentences. 22 sentences were excluded from the dataset, as no equivalent translation was found that would preserve the ambiguity. This is mostly due to specific ambiguous phrases used in English that would not retain their features when translated into Ukrainian. The resulting dataset can be used as a complex challenge for a coreference resolution system.

5.2 Machine Translation

While the WSC dataset translation in Ukrainian is tailored for complex cases of coreference resolution, it contains few sentences and can't be reliably used for training. Hence, we decided to build on OntoNotes 5.0 to compile a sufficient amount of data for further model training and evaluation. In particular, our approach is to take an annotated English dataset, translate it with a machine translation model, align the corresponding mentions in original and translated parallel texts, and project the annotations using the obtained alignment.

As the translation and alignment needs to be done automatically, we relied on using a high-quality machine translation model to bear this task. Specifically, we chose one of OpusMT models, as they are easily accessible through the *hugging-face* library, work with Ukrainian and English, and are generally of high quality (Tiedemann and Thottingal, 2020). The *Helsinki-NLP/opus-mt-en-*

uk model was used to translate from English to Ukrainian; this model achieves 50.2 BLEU score on Tatoeba.en.uk test set.

5.3 Alignment

After translating the sentences with the chosen model, different methods can be used to align the words in source and target sentences. This allows matching the spans corresponding to entity mentions in source and target sentences in order to project the annotations later.

Attention-based Alignment The attention-based approach used in Transformer models can help interpret how the model functions (Bahdanau et al., 2015; Belinkov and Glass, 2019). Attention shows how the model assigns weight to different input elements; in case of sequence-to-sequence machine translation models, it is possible to use this advantage to see which source tokens the model attends to when producing a target translation. The interpretability of attention weights has been the subject of various experiments, and while the saliency methods have been proven to work better, cross-attention can still provide important information about the functionality of the models (Vashishth et al., 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020).

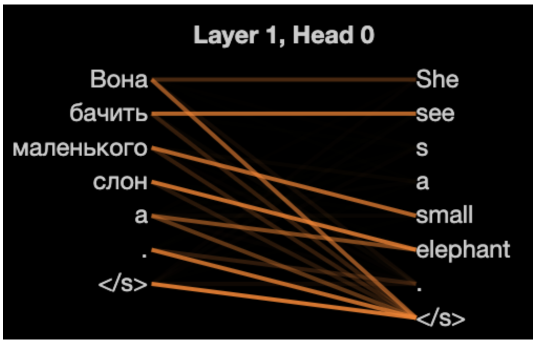

Figure 1: Cross-attention weight graph for one of the heads of the model.

For this approach, we used a visualization tool to determine if cross-attention of the machine translation model is enough for aligning the words. Using the *bertviz* package, we created attention weight graphs of all the layers and heads of the chosen model (Vig, 2019). The resulting visualization showed that for many layers, most of the attention is concentrated on the end token rather than other source words. Therefore, we decided to use only the cross-attention from the 0-th head of the 1-st layer, as it correlated the most with the intuitive judgment of how the words should be aligned.

Embedding-based Alignment Another method to align source and target sentences after machine translation relies on large pre-trained multilingual models that find a shared cross-lingual space for all the languages in the system.

The AWESOME aligner presented in Dou and Neubig (2021) is an example of this approach. Such method allows extracting the embeddings from multilingual models such as mBERT and using it to predict the alignment. Contrary to the attention-based alignment, aligning with AWESOME does not ensure every target word gets aligned with a corresponding source word.

Alignment based on Data Modification One more approach that can be utilized for this task is based in modifying the existing data to mark the specific entities in the source text. The marks then need to be preserved after translation, so that the entities can be recognised in the target text, For example, one could enclose the specific mentions in the source sentence in brackets or XML tags, then translate them, and look for marked words in translation, as the machine learning model often correctly reproduces the marks used. This makes it possible to locate the translated mention in the target text and align it with the corresponding one in the source.

This can be done using an iterative method. However, the machine translation model output can change quite drastically depending on which part of the input is marked. This also means that at each iteration, the translated sentence could be different, so it is unclear which of them to use as a "golden" translation.

For the final dataset, we decided to use the attention-based approach. We extracted the cross-attention weights for a specific translated sentence from the model, produced correct tokens from subtokens returned by the model, and aligned the relevant mentions. If most of the attention from the target token was concentrated on the *</s>* tag, we used the second highest-weighted source token for the alignment.

5.4 Projection

Projection is the next stage after aligning the source and target texts. This part depends significantly on the format the base corpus is presented in. OntoNotes 5.0 follows the standard format of CoNLL-2012 shared task, in which the coreference clusters contain the entities for each text. For projecting, only some information is necessary: namely, *cased_words*, which contains the tokenized document, *sent_id*, which contains ids of sentences the tokens appear in, and *clusters*, which contains the clusters of coreferring entities.

The attention-based alignment approach we have chosen relies on automatically translating texts from English into Ukrainian with a machine translation model. However, the model can only translate sentences rather than full texts. This makes the alignment and projection process non-trivial, as the whole text cannot be translated at once. Instead, every single sentence needs to undergo the process separately.

The general algorithm follows these stages:

1. Split the document into sentences.
2. For each span mentioned in a coreference cluster, find the specific sentence it appears in.
3. Modify the span indices so that they correspond to the sentence rather than the whole document.
4. Modify the new span representation so that it includes all the tokens in the span rather than only start and end word.

After that, the source sentences can be translated via a machine translation model. Since we now know the mentions present in each source sentence, the overall task amounts to aligning the sentences, extracting the specific target tokens aligned with source mentions tokens, producing the mapping between source and target mentions, and reconstructing the coreference clusters based on this mapping.

The main complications lie in the fact spans are often not equivalent in two languages. Alignment is done on the token level, and it is quite possible that words that form a continuous span in the source sentence will not form one in the target sentence. A few different alignment situations can happen: one source word may correspond to multiple target words, multiple source words may correspond to one target word, or source/target words may not align with a separate word at all.

Keeping the possible issues in mind, we wrote the alignment and projection scripts, which can then be used to form a mapping from source to target mentions. According to this mapping, we compile target coreference clusters for the silver Ukrainian version of OntoNotes 5.0. While somewhat noisy, the attention-based alignment usually correctly matches the entities and allows to properly project the annotations and form coreference clusters.

Detailed statistics of the created Ukrainian OntoNotes corpus are shown in Table 1. The OntoNotes dataset is significantly larger than Coreferent Clusters (Kramov, 2021), as the documents in OntoNotes generally contain longer sentences with more mentions.

6 Error analysis

In this section, we provide an analysis of the formed silver Ukrainian datasets.

6.1 Ukrainian WSC Dataset

Manually translating the schemas into Ukrainian does not only allow using the dataset as a complex challenge for a coreference resolution system, but also clearly illustrates the coreference differences in Ukrainian and English.

For some sentences, no equivalent existed that would preserve the ambiguity while keeping the content intact. Particular words or phrases that appear ambiguous in English may not retain those properties when translated into Ukrainian. To use such schemas properly, new pairs of sentences should be written with differing content. When translating, we attempted to keep the original content of the sentence unchanged when possible.

Manual translations can also be subpar sometimes. This approach to creating a dataset requires finding the middle ground between preserving as much original source content as possible and maintaining the overall schema form. While some of the Winograd Challenge Schema dataset pairs are grammatically correct in Ukrainian, native speakers may regard such sentences as less fluent than possible. The reason for that is the ambiguity itself: in fluent Ukrainian sentences, different devices would be used by speakers to remove the ambiguity and make the coreference resolution task easier.

For example, for the Winograd schema presented before, it would be more fluent to say "The man couldn’t lift his son because $\emptyset$ was so weak" in Ukrainian. This requires omitting the subject in the subordinate clause– a grammatically correct way of expressing the same content that will clearly resolve the ambiguity. The subject of "was" would be understood to correspond to the subject of the main clause.

The use of pronouns could be another example of the differences between two languages. It is grammatically correct and fluent to use demonstrative pronouns in Ukrainian where English can only use personal ones. Using the same example, "The man couldn't lift his son because *that was so heavy." can be used in Ukrainian, and "that" in this case will clearly point to the "son". Hence, in order to preserve both the ambiguity of the schema and its fluency in the target language, the content of the sentence must often be altered.

Similarly, other issues arise when resorting to literal translations of the schemas. Ukrainian nouns and pronouns all have grammatical gender, and personal pronouns are as such used for both proper and common nouns; in the sentences where English can use "it" for something denoting an object, eliminating the option for a coreference link from a pronoun like "he", Ukrainian may use "he" in place of both words, resulting in more ambiguities. Overall, literally translating the schemas does not work in many cases; other schemas should be presented to deal with this issue.

In sum, the error analysis shows that main complications in manual translation of complex datasets such as WSC arise when trying to relay the content perfectly in a different language. The distinct features of the source and target language may not allow for a trivial conversion; in this case, new original examples must be created for the resulting dataset.

6.2 Ukrainian OntoNotes Dataset

In the Ukrainian OntoNotes Dataset, the most prevalent errors are connected with the reliance on the machine translation model.

The quality of the English-Ukrainian translations produced by the OpusMT model used are sometimes below average. In some cases, the model produces the target text in a different language, which supposedly comes from incorrectly labeled data it was trained on. In addition, the task is made more complicated because of the nature of the English OntoNotes 5.0 dataset: its genre diversity presents a lot of problems for the machine translation model. As the model has not been trained on the text of some specific genres such as telephone conversations, the translation for such documents is of poor quality. For future work in this direction, we suggest choosing a different machine translation model to get translations that would be less noisy.

The alignment approach we chose also relies on the machine translation model, so naturally, its quality may be lower than expected for the same reasons. The alignment fully depends on the cross-attention layer of the encoder-decoder model. This may lead to mistakenly aligning unrelated words and then including them in the coreference clusters. Pruning the resulting clusters to get rid of such noise seems to be a promising future direction. In addition, other alignment methods such as statistical alignment or embedding-based alignment should be explored.

7 Conclusion

Creating new datasets is crucial in order to extend NLP research to the low-resource setting. Labeled data for languages such as English can be effectively utilized for this task. We present an approach to efficiently create silver data corpora for low-resource languages based on existing annotated data for high-resource languages such a English with machine translation, alignment, and annotation projection. We demonstrate how the suggested methods can be used to create two corpora for training and testing coreference resolution models for Ukrainian. The scripts for automatic translation, alignment, and projection, as well as the Ukrainian WSC dataset are made publicly available[2].

Future work will involve training and evaluating a baseline model using the created silver Ukrainian datasets. The suggested approaches may be improved by using different machine translation models or trying out better alignment methods.

Limitations

Our approach to creating silver data for Ukrainian on the basis of English annotated corpora is based on manual and machine translation. As the quality of the resulting translations is of utmost importance, the method has a few important limitations.

For manual translation of small sophisticated test datasets, the approach requires enrolling professional annotators with relevant experience. For an intricate corpus such as the WSC dataset, the annotation process may involve creating new content and altering the source documents significantly to

[2] https://github.com/pkuchmiichuk/ua-coref

preserve the specific features of the text required for the task.

For automatic translation, alignment, and annotation projection, a machine translation model from high-resource source language into the low-resource target language should be present. Such models may not exist whatsoever for many low-resource languages or exhibit poor quality, which limits the potential use of our approach.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2018. Exploring spanish corpora for portuguese coreference resolution. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 290–295.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. 2022. Language model priming for cross-lingual event extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10627–10635.

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

Sanda M. Harabagiu and Steven J. Maiorano. 2000. Multilingual coreference resolution. In *Sixth Applied Natural Language Processing Conference*, pages 142–149, Seattle, Washington, USA. Association for Computational Linguistics.

Andrii Hlybovets. 2018. Agent-based software systems for the search and analysis of information. *Doctor of Technical Sciences, Taras Shevchenko National University of Kyiv*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2022. The defeat of the winograd schema challenge.

Artem Kramov. 2021. Coreferent clusters (dataset and a pre-trained model).

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*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, Proceedings of the International Conference on Knowledge Representation and Reasoning, pages 552–561. Institute of Electrical and Electronics Engineers Inc.

Kaushal Kumar Maurya, Maunendra Sankar Desarkar, Yoshinobu Kano, and Kumari Deepshikha. 2021. ZmBART: An unsupervised cross-lingual transfer framework for language generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2804–2818, Online. Association for Computational Linguistics.

Luis Fernando Guzman Nateras. 2022. Modern cross-lingual information extraction. Area exam, University of Oregon, Computer and Information Sciences Department.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Maciej Ogrodniczuk. 2013. Translation- and projection-based unsupervised coreference resolution for polish. In *Language Processing and Intelligent Information Systems*, pages 125–130, Berlin, Heidelberg. Springer Berlin Heidelberg.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sergiy Pogorilyy and Artem Kramov. 2019. Coreferent pairs detection in Ukrainian texts using a convolutional neural network. *Visnyk Universytetu "Ukraina"*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA. Association for Computational Linguistics.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Sergii Telenyk, Sergiy Pogorilyy, and Artem Kramov. 2021. The complex method of coreferent pairs detection in a Ukrainian-language text based on a BiLSTM neural network. In *2021 IEEE 3rd International Conference on Advanced Trends in Information Theory (ATIT)*, pages 205–210.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. Deep cross-lingual coreference resolution for less-resourced languages: The case of Basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 35–41, Minneapolis, USA. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. *OntoNotes: A Large Training Corpus for Enhanced Processing*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. Gradual fine-tuning for low-resource domain adaptation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 214–221, Kyiv, Ukraine. Association for Computational Linguistics.

Haoran Xu and Kenton Murray. 2022. Por qué não utiliser alla språk? mixed training with gradient optimization in few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2043–2059, Seattle, United States. Association for Computational Linguistics.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.