End to End Multilingual Coreference Resolution for Indian Languages

Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao
AU-KBC Research Centre,
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract

This paper describes an approach on an end to end model for Multilingual Coreference Resolution (CR) for low resource languages such as Tamil, Malayalam and Hindi. We have done fine tune the XLM-Roberta large model on multilingual training dataset using specific languages with linguistic features and without linguistic features.XLM-R with linguistic features achieves better results than the baseline system. This shows that giving the linguistic knowledge enriches the system performance.The performance of the system is comparable with the state of the art systems.

1 Introduction

The task of CoreferenceResolution (CR) has attracted considerable attention in natural language processing due to its importance in deep language understanding and its potential as a subtask in a variety of complex natural language processing problems. It is a key module in many NLP systems and is used in many higher-level NLP tasks, such as Information Extraction, Cause-effect identification, Cross lingual information retrieval, document summarization and question answering. CR is the task of finding all referring expressions for a referent (an anaphor) that refers to the same real-world entity (the antecedent) with in a document or text. A referring expression or a mention is a noun phrase (NP) , a named entity (NE), or a clause, which refer to an entity in the real world known as the referent (Sapena, E.et.al.,2013). The referents (anaphors) can be a pronoun, a reflexive, a reciprocals, a noun, in short anything that requires a reference in front. A grouping of referring expressions with the same referent or antecedent is called a coreference chain or cluster. The goal of a coreference resolution system is to identify such chains or clusters in a given document and give each chain as output.

The span of coreferential expressions can be from a single sentence or can be separated by one or more sentences. In many cases, it is seen that the entire document needs to be considered for identifying the referent accurately. The coreference resolution task can be divided into two subtasks, 1) Identify entity mentions and group them together according to the real-world entity they refer to and 2) To resolve the task of anaphora resolution which is closely related to coreference resolution (Sukthanker et al., 2020). Multilingual coreference resolution is the process of determining if expressions in multiple languages refer to the same entity.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 discusses the data and section 4 outlines our approach. Whereas the Section 5 discuss the evaluation and Section 6 the conclusion.

2 Related Work

In general, coreference resolution is solved by first predicting the coreference mentions and subsequently performing coreference linking (clustering) of the predicted mentions. In recent years this method got changed and an end-to-end approach proved to be better performing (Lee et al., 2017, 2018; Joshi et al., 2019, 2020).

Recently, different approaches have been presented for coreference resolution in English. Latent antecedents (Bohnet, B et.al., 2018; Fernandes ER.et.al.,2014) or neural models

(Wiseman S,et.al., 2015; Wiseman S,et.al.,2016) have gained popularity and achieved state-of-the art results. Hybridisation of techniques has been also proposed, for example in (Lee et.al.,2017), which presents a hybrid architecture of their Stanford system. This approach incorporates both rule-based and statistical machine learning methods.

Interest in coreference resolution in languages other than English has been increasing in the last few years (Ogrodniczuk M et.al.,2016; 2017), and because of this many papers about adaptations of coreference resolution systems to a language other than the one for which they were created have been published. The Beautiful Anaphora Resolution Toolkit (BART) (Versley Y et.al., 2008) has been adapted to many languages: originally created for English, its flexible modular architecture facilitates its portability to other languages. There has therefore been a lot of work on extending the BART coreference toolkit to languages other than English including Italian (Poesio M. et al., 2010), German (Broscheit S. et al,2010), Polish (Kopeć M. etal.,2012), Arabic and Chinese (Uryupina O etal.,2012), and Indian languages (Sikdar UK etal., 2016). There are other systems using CRFs such as (Lalitha Devi et al., 2019, Ram, Sundar et al., 2019).

The first end-to-end neural coreference resolution system was introduced by (Leeet al., 2017), and many subsequent neural coreference resolution systems are based on their model. The end-to-end approach has been improved by (Kirstain et al. 2021) not to explicitly construct the span representations, and by (Dobrovolskii 2021) to consider only the word level, ignoring the span level altogether during coreference linking. Simultaneously, (Wu et al. 2020) formulated coreference resolution in a question answering setting.

3 Data

In this work we have used aligned multilingual corpus for annotating for coreference. CR is a discourse phenomenon and it requires both syntactic and sematic information for resolution. In this study we have developed a corpus of aligned Hindi, Malayalam, Tamil texts to identify coreferring expressions. Our task focused on the same kind of coreference as considered in the past MUC competitions, namely the identity coreference. Identity coreference links nouns, pronouns and noun phrases (including proper names) to their corresponding antecedents.

We created our multilingual collection of data by translating the English data collected for the discourse translation project to Tamil, Hindi and Malayalam. The translation was done manually by bilingual experts. The data is from three domains, Agriculture, Health and Governance. The data is parallel aligned and annotated for anaphors, its antecedents and coreference chains. The details of the data is given in Table 1.

Table 1. Data Statistics

Sl No	Language	Number of sentences in the data	Total Coreference Chains
1	Tamil	50000	769
2	Malayalam	50000	769
3	Hindi	50000	769
Only similar chains are taken from all the 3 languages			

We annotated the corpus for anaphor-antecedentspairs and coreference manually and evaluated the inter annotator's agreement. The kappa score we received is 0.93 across all the three languages. Though all the languages under consideration here belongs to pro-drop languages we have not annotated the corpus for zero anaphors. Also we have not annotated singleton mentions. Both zero anaphors and the singleton mentions are not considered for this study. The data set annotated has the same coreference identifiers across three languages.

4 Our Approach

This Our model builds on the transformer-based end-to-end coreference resolution model that was originally proposed by Lee et al. (2017). This predicts the antecedents for all possible mention spans without detecting the mentions in the previous stage. We based our model on XLM Roberta large which is trained on 100 languages which includes Hindi, Tamil and Malayalam. Our coreference annotated dataset consists of Hindi, Tamil, Malayalam. XLM-Roberta, is an advanced multilingual model based on the transformer architecture and has shown promising results in various NLP tasks, including coreference resolution. XLM-Roberta is an extension of the Roberta model, designed for multilingual tasks. It is pre-trained on a large corpus of multilingual text using a masked language modelling objective. Its architecture allows it to capture semantic relationships and context effectively across multiple languages, making it suitable for tasks that require understanding deeper references within texts.

Training involves fine-tune the pre-trained XLM-Roberta model on the coreference resolution dataset for Hindi, Tamil and Malayalam. The dataset is parallel across languages as discussed in section 3. We fine-tune the model for each specific language.

Data Preparation: The data is formatted such that it can be given as input to the training system. A sequence of tokens representing the entire text and an attention mask to distinguish between real tokens and padding are given as the input.

Model Training (Fine-tuning): Incorporating linguistic information can significantly enhance the model. We include the following linguistic features:

i) Entity Type – Such as person, organization
ii) Word Category – Noun, Pronoun
iii) IsAntecedent – Yes/No and
iv) Number and Gender – Plural/Singular, Male/Female

To encode linguistic information, we add to each token representation the four feature information. In more detail, we convert the linguistic information into numerical information. For example for the Entity type, we use one-hot encode. And for binary features such as gender we use binary representation 0/1 (Eg: 1 for Female, 0 for male). This we name it as ling_emb. This representation is then concatenated with Roberta embedding of each token. This results in richer representation of each token. Along with the above step of embedding the following 3 steps are also involved in the process of fine tuning.

- **Selecting Loss Function**: Standard 'cross-entropy' loss function is used to train the model, ensuring that it can learn to minimize errors based on the combined feature representation
- **Batching**: We create batches that maintain the context around mentions to allow the model to learn the relationships effectively.
- **Optimizer and Learning Rate**: AdamW optimizer with a learning rate scheduler to adjust learning rates dynamically is used here.

We trained all the models on NVIDIA A80 GPUs using online learning (batch size 1 document). We limit the maximum sequence length to 6 non-overlapping segments of 512 tokens. During training, if the document is longer than 6 × 512 tokens, a random segment offset is sampled to take a random continuous block of 6 segments, and the rest of them are discarded. During prediction, longer documents are split into independent sub-documents (for simplicity, non-overlapping again). We use 50K steps for model fine-tuning on each language dataset. The training of each language took from 15 to 20 hours.

5 Evaluation

We evaluate the trained models for all the 3 language. We consider the baseline system as the one which does not have linguistic features as explained in the previous section. The results are tested on our own dataset described in section 3. Table 2, shows the results obtained for the evaluation metric MUC.

As it can be observed from the results in Table 2, XLM-R with linguistic features achieves better results than the baseline. There is approximately 3% improvement in the results. This shows that giving the linguistic knowledge enriches the system performance.

Table 2. Results

Language	Baseline (MUC Score) P/R/F	XLM-R with Linguistic features (ling_emb) (MUC Score) P/R/F
Hindi	59.56/57.67/58.59	62.56/59.67/61.08
Tamil	59.96/58.27/59.10	63.26/59.97/61.57
Malayalam	58.96/57.97/58.46	63.16/59.88/61.47
English	60.16/59.78/59.96	65.22/63.77/64.45

Conclusion

We provide an end to end multilingual coreference resolution system for Indian languages which are resource poor using XLM Roberta with linguistic feature training. We have used the pre-trained model on multilingual training dataset and then fine tune for specific language, in this case Tamil, Malayalam and Hindi. We have developed a baseline system and a system using linguistic information. XLM-R with linguistic features achieves better results than the baseline. There is approximately 3% improvement in the results. This shows that giving the linguistic knowledge enriches the system performance. The performance of the system is comparable with the state of the art systems.

References

Broscheit S, Ponzetto SP, Versley Y, Poesio M. 2010. Extending BART to Provide a Coreference Resolution System for German. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010. Valletta, Malta; 2010. p. 164–167.

Joshi, M.; Levy, O.; Zettlemoyer, L.; Weld, D. 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); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 5802–5807.

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, September 2017. Association for Computational Linguistics.

Kopeć M, Ogrodniczuk M. 2012. Creating a Coreference Resolution System for Polish. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA); 2012. p. 192–195.

LiyanXu 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, Online, November 2020. Association for Computational Linguistics.

Poesio M, Uryupina O, Versley Y. 2010. Creating a Coreference Resolution System for Italian. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA); 2010. p. 713–716.

Recasens, M.; Màrquez, L.; Sapena, E.; Martí, M.A.; Taulé, M.; Hoste, V.; Poesio, M.; Versley, Y. Semeval-2010 task 1: Coreference resolution in multiple languages. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 1–8.

Rhea Sukthanker, SoujanyaPoria, Erik Cambria, and RamkumarThirunavukarasu. 2020. Anaphora and coreference resolution: A review. Information Fusion, 59:139–162, 2020.

Ritwik Mishra, PoojaDesur, Rajiv Ratn Shah, PonnurangamKumaraguru. Multilingual, Coreference Resolution in Low-resource South Asian Languages,arXiv:2402.13571 [cs.CL], https://doi.org/10.48550/arXiv.2402.1357

Sapena, E., Padró, L.,Turmo, J. 2013. A Constraint-Based Hypergraph Partitioning Approach to CoreferenceResolution. In Journal of *Computational Linguistics.* 2013, *39*, 847–884.

Sukthanker, R., Poria, S., Cambria, E.,Thirunavukarasu, R. 2018. Anaphora and Coreference Resolution: A Review. *arXiv*, arXiv:1805.11824.

Uryupina O, Moschitti A, Poesio M. 2012. BART Goes Multilingual: The UniTN/Essex Submission to the CoNLL-2012 Shared Task. In: Joint Conference on EMNLP and CoNLL—Shared Task. CoNLL'12. Jeju Island, Korea: Association for Computational Linguistics; 2012. p. 122–128.

Sikdar UK, Ekbal A, Saha S. 2016. A generalized framework for anaphora resolution in Indian languages. Knowledge-Based Systems. 2016;109:147–159.

Soraluze A, Arregi O, Arregi X, Díaz de Ilarraza A 2019. EUSKOR: End-to-end coreferenceresolution system for Basque. PLoS ONE 14(9): e0221801.https://doi.org/10.1371/journal.pone.0221801