Bridge the GAP: Multi-lingual Models For Ambiguous Pronominal Coreference Resolution in South Asian Languages

Rahothvarman P*
IIIT Hyderabad
rahothvarman.p@research.iiit.ac.in

Adith John Rajeev*
IIIT Hyderabad
adith.r@research.iiit.ac.in

Kaveri Anuranjana
IIIT Hyderabad
kaveri.a@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Coreference resolution, the process of determining what a referring expression (a pronoun or a noun phrase) refers to in discourse, is a critical aspect of natural language understanding. However, the development of computational models for coreference resolution in low-resource languages, such as the Dravidian (and more broadly all South Asian) languages, still remains a significant challenge due to the scarcity of annotated corpora in these languages. To address this data scarcity, we adopt a pipeline that translates the English GAP dataset into various South Asian languages, creating a multi-lingual coreference dataset mGAP. Our research aims to leverage this dataset and develop two novel models, namely the joint embedding model and the cross attention model for coreference resolution with Dravidian languages in mind. We also demonstrate that cross-attention captures pronoun-candidate relations better leading to improved coreference resolution. We also harness the similarity across South Asian languages via transfer learning in order to use high resource languages to learn coreference for low resource languages.

1 Introduction

Coreference resolution involves identifying and linking referring expressions (pronouns or noun phrases) to their respective referents. Accurate resolution is essential for discourse-level tasks such as dialogue understanding (Tseng et al., 2021), machine translation (Stojanovski and Fraser, 2019), summarization (Steinberger et al., 2007), question answering (Castagnola, 2002) and sentiment analysis (De Clercq and Hoste, 2020). Prominent datasets such as OntoNotes 5.0 (Pradhan et al., 2013a) (English, Chinese & Arabic), ParCorFull (Lapshinova-Koltunski et al., 2018) (English & German) and TransMuCoRes (Mishra et al., 2024) (31 South Asian languages) have focused on multilingual coreference resolution.

Transformer-based methods have been proposed for multilingual coreference resolution (Martinelli et al., 2024; Liu et al., 2022). Additionally, ChatGPT (Chen, 2024), despite its advances on the WinoGrand Challenge, could not learn linguistic features of Chinese such as zero pronouns. South Asian languages (SALs) exhibit similar unique traits which we investigate in this paper.

Pronominal coreference resolution, the most common type of coreference (Lappin and Leass, 1994), identifies the referents of pronouns. In languages with complex grammatical structures such as pro-drop, gender-neutral pronouns or elaborate gender agreement — resolving pronominal coreference is quite challenging. Indo-European and Sino-Tibetan are the two largest language families with 4.7 billion speakers together (eth, 2024). Despite its prevalance, pronominal coreference resolution remains unexplored in SALs. Addressing this gap is crucial for the growing need for NLP solutions tailored to these populations. We aim to bridge this gap by making the following key contributions:

1. **Multilingual GAP (mGAP)**: mGAP is a multilingual ambiguous pronoun resolution corpus of 8,908 ambiguous pronoun-name pairs derived from the GAP Coreference Dataset for 27 SALs. This includes a manually translated **Gold** subset for few languages to address automatic translation errors and a pronoun lexicon, **PronounLex**.

2. **Coreference Resolution Multilingual Models**: We develop multilingual models for coreference resolution and train them on mGAP. We also demonstrate that cross-attention improves pronoun resolution.

3. **Transfer Learning for South Asian languages**: We explore transfer learning by train-

[*] These authors contributed equally to this work.

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 104–114
January 19, 2025. ©2025 International Committee on Computational Linguistics

ing our models on one language and testing on other SALs. This provides insights into the cross-lingual adaptability of coreference resolution systems and the similarity between various languages.

2 Related Work

Several datasets have been developed for coreference resolution. OntoNotes (Pradhan et al., 2013b) spans English, Chinese, and Arabic, providing coreference annotations along with syntactic, semantic, and discourse-level information. It adopts a span-detection approach, where models identify text spans referring to the same entity, offering a comprehensive framework for coreference resolution. LitBank (Bamman et al., 2019) contains longer documents annotated with ACE entity categories, including person, location, geopolitical entity, facility, organization, and vehicle. The Winograd Schema Challenge (WSC) (Levesque et al., 2012) serves as a key benchmark for evaluating models under ambiguous pronoun resolution scenarios that demand contextual reasoning beyond simple linguistic cues to handle complex pronoun disambiguation.

The GAP Coreference Dataset (Webster et al., 2018) contains 8,908 coreference-labeled pairs of ambiguous pronouns and candidate names, sampled from Wikipedia. It provides a gender-balanced dataset designed to evaluate gender bias in language models. The current state-of-the-art on the GAP dataset is achieved by the Coref-MTL(Liu et al., 2023) model, which attains an overall score of 92.72 and demonstrates a bias score of 99.76. This model jointly learns to identify mentions and establish coreferential links.

Cross-attention enables deeper interactions between pronouns, candidates and surrounding context, addressing limitations of dual-encoder models that rely on fixed vector representations (Agarwal and Bikel, 2020; Li and Zhang, 2024). It also captures dependencies across discourse in linguistically rich contexts (Liu et al., 2022). Inspired by recent advances in entity linking using cross-attention encoders, we propose applying cross-attention mechanisms to pronoun resolution.

2.1 SAL Coreference Resolution

In the context of SALs, coreference resolution has traditionally relied on rule-based approaches, which require extensive linguistic analysis and manual annotation. Initial work on Hindi (Dakwale et al., 2013) and Telugu (Jonnalagadda and Mamidi, 2015) made use of manually-crafted rules. Further strategies involved the integrating of Gender-Number-Person (GNP) features and using Conditional Random Fields (CRFs). These approaches were investigated in Hindi (Lalitha Devi et al., 2014) and Tamil (A and Lalitha Devi, 2012) to ascertain coreference relationships. Nevertheless, the application of GNP features in SALs presents a challenge due to their highly unique and intricate inflectional system, and thus would limit the scalability of such rule-based approaches.

Meanwhile, recent work on Chinese anaphora resolution demonstrated the ability of ChatGPT to accurately resolve anaphora on a Chinese Winograd Schema (Chen, 2024), thereby illustrating the significant potential of transformer-based models for non-English languages.

Despite these advancements, there is a significant gap in resources and models for South Asian languages especially those that are low resource. Previous efforts by Mishra et al. (2024) address this gap by introducing a dataset encompassing 31 South Asian languages, created using translation and word-alignment tools from OntoNotes and LitBank. The study demonstrates that 75% of the English references align with their predicted translations, showing promise in the accuracy of the dataset.

2.2 Transfer Learning

Major pre-trained multilingual models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) learn language agnostic representations and have set strong baselines across benchmarks like XGLUE (Liang et al., 2020), XCOPA (Edoardo M. Ponti and Korhonen, 2020) and XNLI (Conneau et al., 2018).

Radford (2018) highlight that complete fine-tuning of language models utilizing task-aware input transformations can enhance performance across diverse natural language understanding benchmarks, outperforming traditional discriminatively trained models. Additionally, Hu et al. (2020) demonstrate the effectiveness of zero-shot learning for cross-lingual transfer across diverse NLP tasks and languages, highlighting its potential in low-resource settings. Models jointly trained on multiple datasets with sampling for data augmentation outperform those trained on individual ones, achieving robust and state-of-the-art coreference resolution across varied domains (Toshniwal et al., 2021).

3 Dataset

In this research, we incorporate the human annotated corpus on gendered ambiguous pronoun, GAP (Webster et al., 2018) to create mGAP. The GAP dataset, sourced from Wikipedia consists of the following attributes, the **text**, **pronoun**, **candidateA** and **candidateB** along with their character offsets in the text. There are 4,454 contexts (a balanced set between masculine and feminine contexts) each of which contains two annotated names, this results in 8,908 pronoun–candidate pair labels.

3.1 Dataset creation

We follow a slightly modified version of the pipeline proposed by Mishra et al. (2024) to create mGAP. The pipeline consists of the following tasks: machine translation and alignment.

3.1.1 Machine Translation

nllb-200-1.3B (Costa jussà et al., 2022) an MT model based on Sparsely Gated Mixture of Experts based approach which has been trained on more than 200 languages. It's high coverage makes it suitable for translating even low resource languages. For the GAP dataset, we translate the text, pronoun and the candidates to the target language using nllb-200-1.3B model.

3.1.2 Text Alignment

The GAP dataset also annotated the character offsets of the the pronoun and candidates for each text. These offsets facilitated evaluation of additional span-based models. Section 4.2 also presents a span-based model which harnesses coreference cross-attention using these spans. To obtain the offsets in the target languages, we use **awesome-align** (Dou and Neubig, 2021), a multilingual BERT-based aligner. It leverages mBERT's rich multilingual representations, fine-tuning it for alignment resulting in broad language support and high-quality alignments.

However, the model was built to only return "positive" alignments based on a confidence threshold, which often excluded essential alignments, particularly for the pronouns and nouns. To resolve this issue, we modified the architecture to prioritize recall over precision, selecting the highest alignment for each word regardless of a threshold. While this approach could be more noisy for the general task of alignment, it effectively improved the identification of pronouns and nouns. Furthermore, analysis of the aligned data revealed that the model makes fewer errors with pronouns and nouns compared to other grammatical categories, underscoring the effectiveness of this approach.

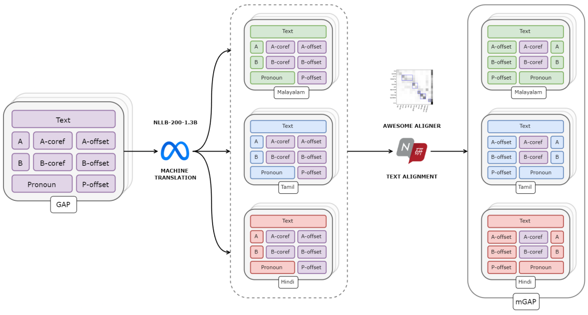

Figure 1: Translate and align pipeline used to generate mGAP. **nllb** (Costa jussà et al., 2022) translates the English sentences to respective SAL and **awesome-align** (Dou and Neubig, 2021) generates the candidate and pronoun span indices.

3.1.3 PronounLex

We used hand curated lexicon of pronouns for a specific SAL, PronounLex to perform a sanity check on the alignments to see how well the Aligner performs before and after the change in architecture of the Aligner. This step helps us gauge whether the system is correctly identifying and aligning pronouns in the target languages. This acts as a safeguard to ensure that the system has at least identified and aligned a pronoun, it is not foolproof since although the pronoun could be pointing to a word that belongs in the lexicon in the target language, it need not be the right pronoun. This approach still provides a simple check to monitor the aligner's accuracy and identify potential areas for improvement.

3.2 Gold Dataset

To scale the dataset across many SALs, we rely on nllb and awesome-align. However, errors can accumulate across translation and alignment stemming from each model's errors, affecting the final output. To address this, we include gold test sets by taking a subset of 200 random samples from the test split and manually cleaning the translations and alignments for a few key languages (Tamil, Malayalam, Kannada, Hindi and Bengali).

During the manual dataset creation, we observed certain linguistic phenomena that make mGAP challenging. One such feature is pro-drop, which omits the pronoun entirely. However, this was observed in a negligible fraction of samples across languages. Additionally, longer sentences often undergo phrase rearrangement. For Hindi, pronouns sometimes align with the object's gender instead of the subject's. Moreover, certain honorific pronouns lose gender distinction while verbs become gendered. These linguistic variations illustrate the unique challenges involved in resolving ambiguous pronouns in SALs.

4 Model Architecture

We introduce two coreference resolution models, each targeting distinct challenges. The Joint Embedding Model (JEM) is centered on harnessing the efficacy of the multilingual embeddings across multiple languages. The goal of this model is to investigate how pronoun resolution in multilingual contexts can be improved by leveraging shared representations across languages.

In contrast, the Cross Attention Model (CAM) relies on a cross-attention mechanism to capture the relationships between pronouns and the potential candidates. CAM investigates how architectural improvements can improve the coreference resolution procedure by specifically attending to candidate spans inside phrases, whereas JEM highlights the effectiveness of multilingual embeddings.

4.1 Joint Embedding Model (JEM)

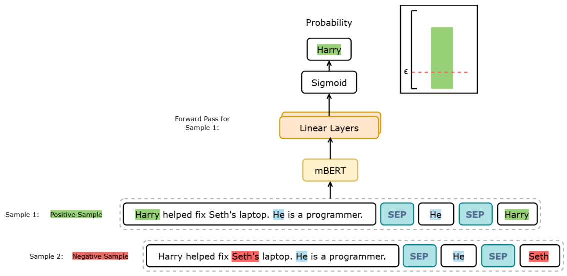
Probability
Harry
Sigmoid
ϵ
Forward Pass for Sample 1:
Linear Layers
mBERT
Sample 1:
Positive Sample
Harry helped fix Seth's laptop. He is a programmer.
SEP
He
SEP
Harry
Sample 2:
Negative Sample
Harry helped fix Seth's laptop. He is a programmer.
SEP
He
SEP
Seth

Figure 2: Architecture of Joint Embedding Model (JEM). Each sentence is passed twice with a positive and negative sample during training. The model outputs the probability of the pronoun referring to sampled candidate thresholded by ϵ.

Our Joint Embedding Model leverages multilingual BERT fine-tuned for coreference resolution (see Figure 2). We reformulate GAP's three-way classification task (candidate A, candidate B or neither (ϕ)) to a binary classification task that predicts whether the candidate present in the data point either corresponds to the pronoun or not. We selected this objective because our experiments revealed that the three-way classification objective led to suboptimal performance.

In this binary framework, each data point is sampled twice: once with the pronoun paired with its correct candidate (positive sample) and once with an incorrect candidate (negative sample). Cases where the pronoun refers to neither candidate are excluded during training and handled via a threshold-based mechanism at inference. We hypothesize that in the three-way setup, by focusing on a direct relationship between the pronoun and a single candidate, the model can learns relevant features without the distraction of multiple competing candidates (or even learning to predict neither). Additionally an increased number of samples further leads to better results.

We format the input sequence as follows:

$$x = [\text{[CLS]} \,;\, S \,;\, \text{[SEP]} \,;\, P ; \text{[SEP]} \,;\, C_i]$$

where x is the input to the model, S is the context sentence containing the ambiguous pronoun, P is the target pronoun requiring resolution, C_i are possible candidates ($C_i \in \{A, B\}$).

During training, we perform the binary classification by obtaining the probabilities for each candidate using the following formula:

$$\hat{y} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot h + b_1) + b_2)$$

where h is the $[CLS]$ token output from mBERT for the concatenated input, W_1 and W_2 are the weights of the linear layers, and b_1 and b_2 are the biases.

During inference, we implement the 3-way classification as follows. For each data point, the model evaluates the label l, by computing the probabilities for both candidates A and B (using their respective BERT representations h_a and h_b) and comparing them with the threshold:

$$\hat{y_a} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot h_a + b_1) + b_2)$$

$$\hat{y_b} = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot h_b + b_1) + b_2)$$

$$l = \begin{cases} A, & \text{if } \hat{y_a} \leq \hat{y_b} \text{ and } \hat{y_a} > \epsilon \\ B, & \text{if } \hat{y_b} > \hat{y_a} \text{ and } \hat{y_b} > \epsilon \\ \phi, & \text{if } \hat{y_a} \leq \epsilon \text{ and } \hat{y_b} \leq \epsilon \end{cases}$$

If the probabilities for both candidates are below a pre-defined threshold ϵ (0.2 in our case), the model classifies the pronouns as referring to "neither". Otherwise, it classifies the pronoun to the candidate with the higher probability.

This architecture enables the model to effectively resolve ambiguous pronouns by leveraging contextual information. In the future, this approach could be also implemented for more than 3 classes, or can be implemented after identifying possible candidate spans within the sentence.

4.2 Cross Attention Model (CAM)

We present our other approach, a multi-headed cross attention network that computes the similarity between the candidates and pronouns from the underlying vector representations of the pronouns and candidates. This architecture consists of two main components, namely the candidate / pronoun encoders and the coreference resolver. Figure 3, illustrates the architecture.

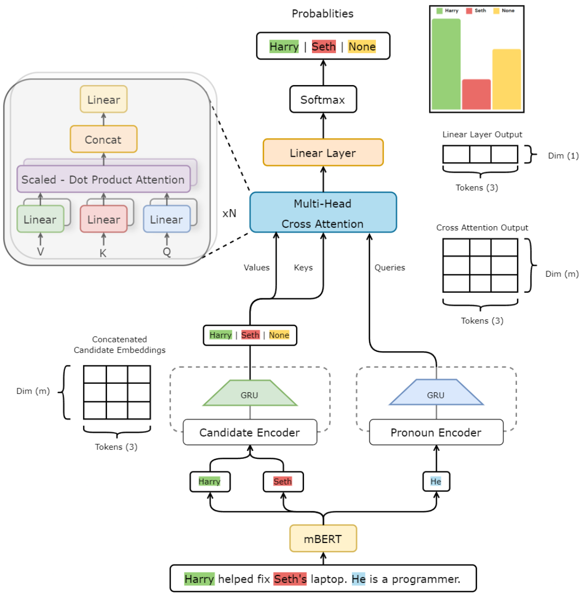

Figure 3: Architecture of Cross Attention Model (CAM). The model outputs a probability distribution over the possible candidates.

4.2.1 Candidate-Pronoun Encoder

Since the candidates and pronouns in SALs can span several tokens, we employed a Gated-Recurrent Unit (GRU) (Cho et al., 2014) based feature aggregation layer for the candidate-pronoun encoder to address this variation in token length for the candidates-pronouns between language. This application case is well suited to the GRU's capacity to maintain sequential information while aggregating the embeddings into a fixed-length representation. It efficiently condenses the multiple-token words $\in R^{N_t \text{ x } m}$ (where N_t represents the token length) into a single, cohesive vector so that the resolver can process it more easily.

While the pronoun encoder encodes the pronoun P as $E_p \in R^{1 \text{ x } m}$, the candidate A, B are independently encoded as $E_a \in R^{1 \text{ x } m}$, $E_b \in R^{1 \text{ x } m}$ respectively. The "neither" class $E_n \in R^{1 \text{ x } m}$ is represented with a zero vector. It is then concatenated with the candidate embeddings as shown below. Since all three potential outputs (A, B or neither) will now be represented consistently as E_t, this should simplify the categorization work. By relying on the model to learn this structure, we avoid arbitrary threshold-based cutoffs methods which were proposed in the previous sections.

$$E_t = [[E_a] : A; [E_b] : B; [E_n] : N]$$

4.2.2 Coreference Resolver

The network is given the concatenated candidate embeddings $E_t \in R^{3 \text{ x } m}$ as Key K and value V, and the pronoun embedding $E_p \in R^{1 \text{ x } m}$ as query Q. The cross attention is defined as follows :

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{m} e^{x_j}},\ \sigma(x_i) \in (0,1)^m \quad (1)$$

$$Sim(Q,K) = \sigma(\frac{QK^T}{\sqrt{d_k}}),\ d_k = \frac{m}{n_{heads}} \quad (2)$$

$$Attn = Sim(Q,K) * V \quad (3)$$

This attention mechanism allows the model to attend to the most relevant parts of the concatenated candidate representations in relation to the pronoun's representation. The model dynamically focuses on the specific features of candidate A, candidate B and even the absence of an appropriate candidate (indicated by the zero vector for "neither").

4.3 Implementation & Hyper-parameters

For JEM, the models were trained on a single NVIDIA GTX 1080Ti for 40 epochs, using a batch size of 8. For optimizer, we used the Adam optimizer with a learning rate of 10^{-5} and Binary Cross Entropy (BCE) loss. We used early stopping to prevent overfitting.

For CAM, the models were trained on a single NVIDIA RTX 2070 Super Max-Q GPU for 100 epochs, using a batch size of 64. We employed the Adam optimizer with a learning rate of 10^{-5} and Categorical Cross Entropy (CCE) loss. Early stopping was used to prevent overfitting, and the best model was selected based on its performance on the validation set.

5 Results and Discussions

We trained our proposed approaches on the development sets of the mGAP dataset (28 languages) and evaluated it on the test sets of mGAP. The evaluation metrics used were F1 and Bias (Table 1). As reported by Webster et al. (2018), bias is the ratio of F1 scores of the female to male pronouns. A bias of less than one indicates the model predicts masculine pronouns better. Additionally, we assess the zero-shot transfer performance of both the models across various SALs.

5.1 Comparison of the Proposed Approaches

The Joint Embedding Model (JEM) and Cross-Attention Model (CAM) present two distinct approaches to pronominal anaphora resolution, each with different parameter footprints and computational requirements. JEM, which utilizes mBERT's CLS token through fully connected layers, requires training both the mBERT backbone and the additional FC layers. In contrast, CAM maintains a frozen mBERT backbone and only trains the cross-attention layer. Training only the Cross Attention layer significantly reduces the number of trainable parameters, leading to faster convergence, lower memory requirements and shorter training times.

Language	JEM		CAM	
	F1	Bias	F1	Bias
English	76.13	1.02	71.06	1.03
Assamese	49.27	0.99	57.08	0.98
Awadhi	67.7	0.98	61.81	0.97
Bengali	66.99	1.00	67.08	1.00
Bhojpuri	62.32	1.00	60.71	1.00
Burmese	56.24	0.95	55.06	1.02
Chhattisgarhi	42.95	0.98	63.46	0.95
Gujarati	65.92	0.95	66.48	0.97
Hindi	70.74	0.94	67.66	1.00
Kannada	68.38	0.98	65.30	1.03
Kashmiri	58.5	0.93	57.38	0.96
Magahi	65.84	0.97	62.21	0.96
Maithili	66.57	0.95	62.47	1.01
Malayalam	64.75	0.95	60.11	1.02
Marathi	64.65	0.95	65.32	1.02
Meitei	52.55	0.97	57.79	0.96
Nepali	62.89	1.04	65.16	0.98
Pashto	44.6	0.97	56.27	0.99
Persian	68.09	0.98	65.56	1.01
Punjabi	64.44	0.99	69.44	1.00
Santali	50.9	0.91	55.66	0.96
Sindhi	44.26	0.98	57.37	1.01
Tajik	56.42	0.92	59.23	1.03
Tamil	65.73	0.95	64.30	0.99
Telugu	70.02	1.04	66.84	1.00
Urdu	66.89	0.95	66.47	0.98
Uyghur	52.5	0.95	54.76	1.05
Uzbek	60.10	0.96	60.36	1.02
Average	60.94	0.97	62.23	0.99

Table 1: Results of our proposed approaches across English and 27 South Asian languages of mGAP

With an average F1 score of 62.23 across all languages, CAM performs marginally better than JEM, which averages 60.94. With JEM at 0.97 and CAM at 0.99, the bias levels of the two models are comparable. In terms of F1 scores, CAM often outperforms JEM, particularly for languages with less resources. With a large difference between its best score (76.13 for English) and lowest score (42.95 for Chhattisgarhi), JEM exhibits a greater variance in F1 scores across languages. CAM performs more consistently across languages and exhibits less variation in F1 score. CAM's greatest score (71.06 for English) and lowest score (55.06 for Burmese) fall within a more constrained range than JEM's.

5.2 Transfer Learning in SALs

In our research on zero-shot transfer learning, we looked at 27 different South Asian languages including English. In order to evaluate the model's cross-lingual generalization and pronominal anaphora resolution capabilities, we trained the models on one language and tested it on the remaining 27 languages for each experiment. With this setup, we were able to investigate the efficacy of

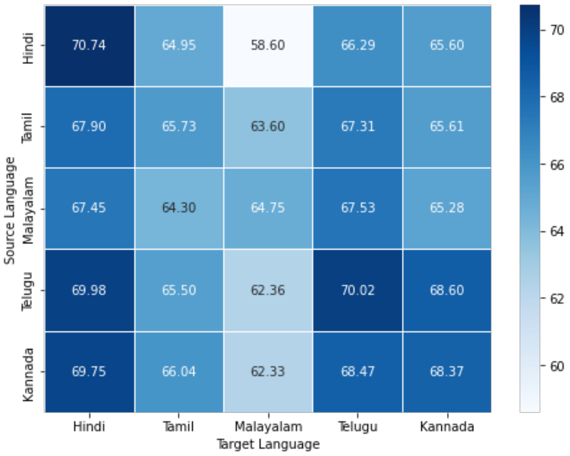

Figure 4: F1 scores of the Zero-Shot Transfer Experiments on JEM for a subset of languages in mGAP

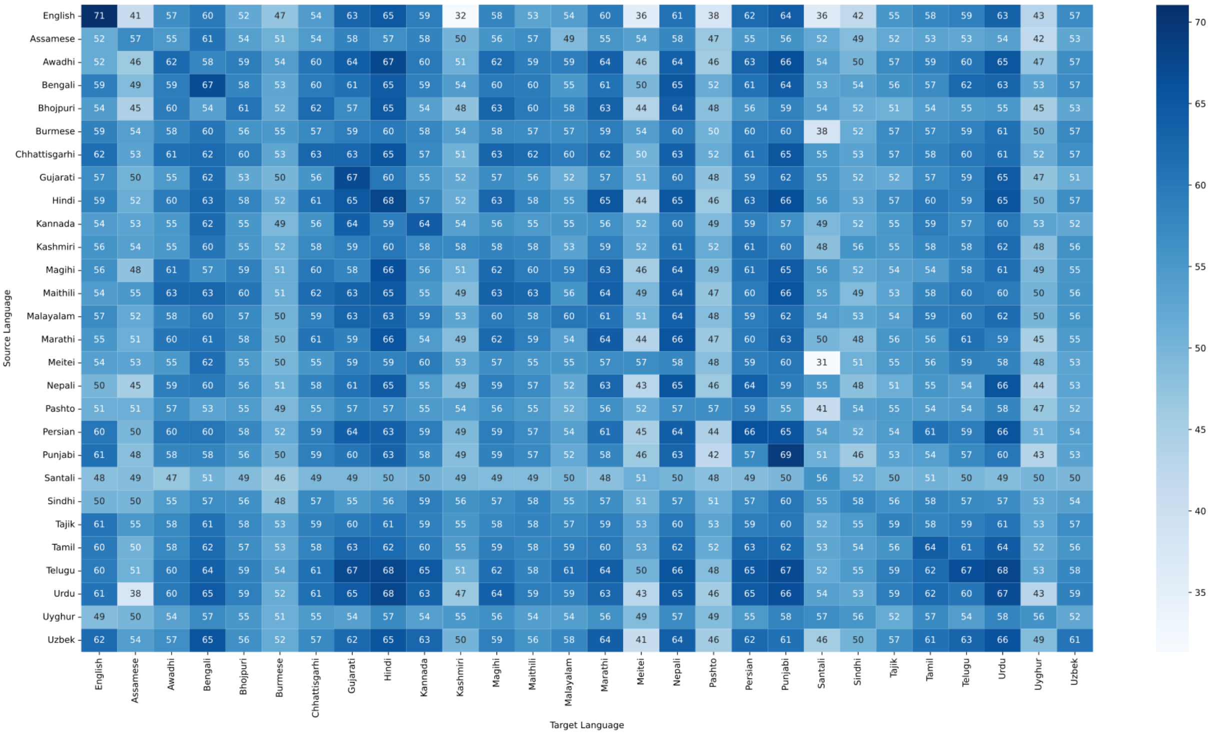

Figure 5: F1 scores of the Zero-Shot Transfer Experiments on CAM for all languages in mGAP

mBERT embeddings and the cross-lingual robustness of our architecture in a range of script-based and linguistic challenges.

5.2.1 JEM

Due to the large memory footprint of JEM, we limited our transfer learning trials to only five languages as it requires fine-tuning of mBERT (Figure 4). Consequently, we focused the assessment of the model's cross-lingual performance only on the Dravidian languages and Hindi.

In our experiments, we observed the models perform best when trained and evaluated on the same languages, with Telugu (70.02) and Hindi (70.74) exhibiting the highest self-performance. The high adaptability of Telugu and Kannada with one another (68.47 and 68.60) and with other languages is probably caused by the common Dravidian linguistic traits that mBERT has identified. Despite having a lower overall cross-lingual transferability, Malayalam outperforms Hindi (58.6) in other Dravidian languages like Tamil (63.6) and Telugu (62.36). These findings demonstrate the efficacy of mBERT in cross-lingual transfer learning, particularly among linguistically related populations.

5.2.2 CAM

With CAM, we were able to perform transfer learning for every language pair, enabling a comprehensive evaluation of cross-lingual performance across all 28 languages (Figure 5).

The analysis of CAM's zero shot transfer matrix reveals that the performance is relatively symmetric - if language A transfers well to language B, the reverse is often true as well. Similar to JEM, the model works better when trained and tested on the same language, as seen by the higher F1 scores along the diagonal for several languages. The languages (like Santali, Sindhi, Ughyur) that make poor sources and target are those that which mBERT isn't trained on. Indo-Aryan languages (Hindi, Bengali, Urdu, Punjabi, Marathi, and Gujarati) have high mutual scores (60–68%), particularly as pairings between Hindi and Urdu and Bengali and Hindi. Punjabi shows decent transfer with Hindi, Urdu, and Persian and vice-versa, likely because of its less rigid gender system, especially for loanwords. Tamil, Telugu and Kannada are all Dravidian languages that fare well together (58–65%), despite variations in script. Telugu has a high degree of cross-family transmission with Indo-Aryan languages. Persian and Tajik have higher scores because of script alignment, but Pashto, Persian, and Tajik all perform moderately (56–61%). It is evident that common linguistic traits have a greater impact than script similarity alone because Santali, which is linguistically and script-wise isolated, routinely ranks lower (48–50%). In general, cross-lingual performance is improved by shared scripts, such as Devanagari across Indo-Aryan languages. Nevertheless, it appears that language family predicts transfer success more accurately than script sharing.

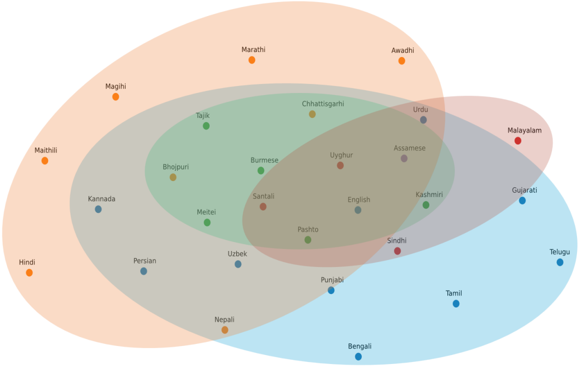

Figure 6: t-SNE Projections of Spectral Language Clusters obtained from Figure 5

We employed spectral clustering to find language groupings based on CAM's zero-shot pronominal coreference resolution performance (Figure 5). Groups were identified by t-SNE visualization (Figure 6): The clustering of the Dravidian languages (Kannada, Tamil, and Telugu) suggests that they have many structural traits and are highly transferable. Bengali, Marathi, and Hindi are Indo-Aryan languages that also clustered, indicating linguistic commonality. On the periphery, Malayalam implies less cross-lingual flexibility because of its distinct linguistic characteristics. This analysis highlights CAM's ability to capture nuanced cross-lingual relationships, enabling interpretation of model performance through a linguistic lens.

5.3 Gender Bias across SALs

As indicated in Table 1, JEM demonstrated gender bias that favours female pronouns in 5 out of the 28 languages. In contrast, CAM shows a preference for female pronouns in a broader range - 16 out of 28 languages. This could be attributed to CAM's ability to capture better representations from the dataset. Furthermore, CAM's average bias score is closer to 1 than JEM's, suggesting a more equitable performance across genders.

6 Conclusion and Future Work

In this study, we addressed data limitations and used multilingual transfer learning techniques to propose a comprehensive approach to resolving coreference in South Asian Languages that are severely under-resourced. We implement a pipeline that allowed us to produce a multilingual coreference dataset - mGAP, for 27 languages, with an addition manually-curated gold subset for a few key languages like Tamil, Malayalam, Hindi and Kannada. This dataset enabled us to test two novel model architectures, namely the Joint Embedding Model (JEM) and the Cross-Attention Model (CAM).

We evaluated the performance of multilingual embeddings and cross-attention architecture using JEM and CAM respectively. Strong zero shot transfer learning potential between a number of South Asian languages was validated by our results. The scores were also significantly impacted by the linguistic and cultural proximity of these languages. Additionally, we demonstrated the potential benefits of sequentially fine-tuning two languages, especially those with limited resources.

Ultimately, this work suggests practical methods for model adaption and offers insightful multilingual resources for coreference resolution in under-represented languages. Future research may expand on these findings by including additional under-resourced languages and exploring language-specific fine-tuning strategies for improved cross-lingual efficacy.

7 Limitations

Our approaches face a few constraints, primarily originating from limitations in the dataset and the challenges inherent in creating high-quality multilingual resources. We worked with only 27 out of the 31 languages used in Mishra et al. (2024) due to missing support for certain scripts in mBERT like Odia, Tibetan and Sinhalese.

Another significant limitation is the potential for error propagation throughout the dataset creation process, as each stage- translation and alignment- carries a risk of introducing inaccuracies. Although we modified the **awesome-align** model to enhance recall rather than precision, this adjustment may introduce further noise. Errors can accumulate through all these stages, affecting the overall quality of the dataset, and any model trained on this data.

Lastly, our gold standard dataset consists of merely 200 samples from the original dataset, which were manually annotated. Since creating gold data requires skilled annotators, the number of languages we currently cover, and the number of samples we annotate is greatly limited. The small size implies that it may not be able to cover all the linguistic and syntactical diversity found in larger, more varied datasets. While it provides a valuable benchmark for quality control, its limited scope may not fully capture the complexity of larger corpora.

References

2024. What are the largest language families? https://www.ethnologue.com/insights/largest-families/.

Akilandeswari A and Sobha Lalitha Devi. 2012. Resolution for pronouns in Tamil using CRF. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 103–112, Mumbai, India. The COLING 2012 Organizing Committee.

Oshin Agarwal and Daniel M. Bikel. 2020. Entity linking via dual and cross-attention encoders. *Preprint*, arXiv:2004.03555.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Luciano Castagnola. 2002. *Anaphora resolution for question answering*. Ph.D. thesis, Massachusetts Institute of Technology.

Shuangshuang Chen. 2024. Resolving chinese anaphora with chatgpt. In *2024 International Conference on Asian Language Processing (IALP)*, pages 31–36.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Preprint*, arXiv:1406.1078.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marta R. Costa jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Praveen Dakwale, Vandan Mujadia, and Dipti Misra Sharma. 2013. A hybrid approach for anaphora resolution in hindi. In *International Joint Conference on Natural Language Processing*.

Orphee De Clercq and Veronique Hoste. 2020. It's absolutely divine! can fine-grained sentiment analysis benefit from coreference resolution? In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–21, Barcelona, Spain (online). Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Olga Majewska Qianchu Liu Ivan Vulić Edoardo M. Ponti, Goran Glavaš and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Hemanth Reddy Jonnalagadda and Radhika Mamidi. 2015. Resolution of pronominal anaphora for Telugu dialogues. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 183–188, Trivandrum, India. NLP Association of India.

Sobha Lalitha Devi, Vijay Sundar Ram, and Pattabhi RK Rao. 2014. A generic anaphora resolution engine for Indian languages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1824–1833, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *KR*. AAAI Press.

Shi Li and Yongkang Zhang. 2024. Improving entity linking by combining semantic entity embeddings and cross-attention encoder. *J. Intell. Fuzzy Syst.*, 46(1):2899–2910.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Ruicheng Liu, Guanyi Chen, Rui Mao, and E. Cambria. 2023. A multi-task learning model for gold-two-mention co-reference resolution. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Ritwik Mishra, Pooja Desur, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2024. Multilingual coreference resolution in low-resource south asian languages. *Preprint*, arXiv:2402.13571.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013a. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013b. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing Management*, 43(6):1663–1680. Text Summarization.

Dario Stojanovski and Alexander Fraser. 2019. Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 140–150, Dublin, Ireland. European Association for Machine Translation.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. CREAD: combined resolution of ellipses and anaphora in dialogues. *CoRR*, abs/2105.09914.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Preprint*, arXiv:1810.05201.

A Appendix: Transfer Learning

We experiment with a sequential transfer learning methodology where we initially fine-tune a coreference resolution model using a source language, followed by additional finetuning on the target language. We aim to determine if this two-step process would improve cross-lingual performance.

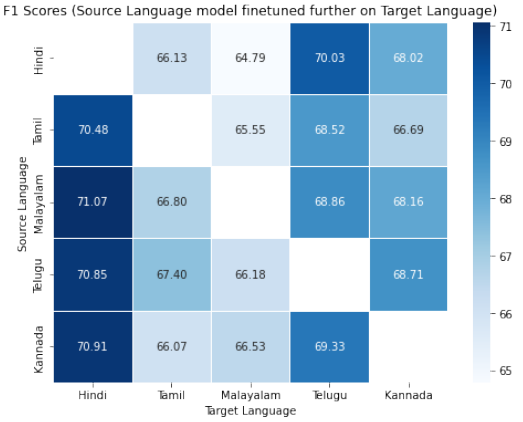

Figure 7: F1 Scores of JEM on a subset of the languages first finetuned on the source language, and then further on the target language.

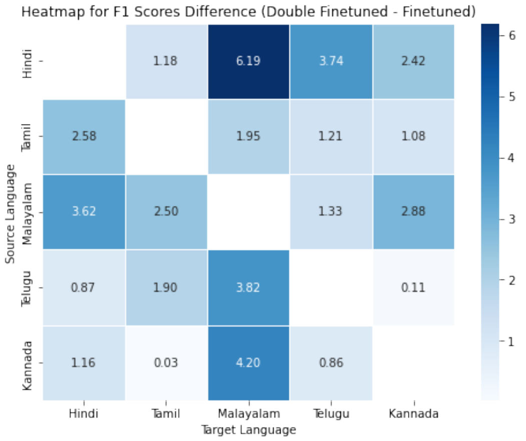

Figure 8: Difference in F1 scores from zero shot setting. (ref Fig. 4)

These experiments show us patterns in the cross-lingual training that highlight the impact of language families and linguistic distance between languages. The improvement from Hindi to the Dravidian languages show substantial growth with the double finetuning approach, with gains up to 6.19 F1 points (Hindi → Malayalam). This improvement suggests that zero shot transfer across these language families could be challenging due to the linguistic distance, and this gap could be effectively bridged by finetuning on the target language as well.

The case for Malayalam is especially interesting as a target language, as it consistently demonstrates the highest average improvements across various language sources. The significant enhancements noted when transitioning to Malayalam (6.19 from Hindi, 4.20 from Kannada) suggests that Malayalam could be distinct in its structural characteristics that render zero-shot transfer particularly difficult; however, these obstacles can be effectively mitigated through further finetuning. This observation has important implications for the allocation of resources in multilingual NLP initiatives that involve Malayalam.

In the context of the Dravidian language family, we observe more modest improvements resulting from double finetuning. These lesser gains likely indicate the stronger initial zero-shot transfer capabilities among these languages, which can be attributed to their shared linguistic traits and close linguistic distance. These results align with linguistic reasoning, indicating that models are more capable of transferring knowledge between closely related languages, even in zero-shot scenarios, thereby leaving limited opportunities for enhancement through additional fine-tuning. They also highlight the necessity of considering language family relationships when formulating transfer learning strategies for low-resource languages.