POS-Aware Neural Approaches for Word Alignment in Dravidian Languages

Antony Alexander James, Parameswari Krishnamurthy
International Institute of Information Technology, Hyderabad, India
antonyalexander8942@gmail.com, param.krishna@iiit.ac.in

Abstract

This research explores word alignment in low-resource languages, specifically focusing on Telugu and Tamil, two languages within the Dravidian language family. Traditional statistical models such as FastAlign, GIZA++, and Eflomal serve as baselines but are often limited in low-resource settings. Neural methods, including SimAlign and AWESOME-align, which leverage multilingual BERT, show promising results by achieving alignment without extensive parallel data. Applying these neural models to Telugu-Tamil and Tamil-Telugu alignments, we found that fine-tuning with POS-tagged data significantly improves alignment accuracy compared to untagged data, achieving an improvement of 6–7%. However, our combined embeddings approach, which merges word embeddings with POS tags, did not yield additional gains. Expanding the study, we included Tamil, Telugu, and English alignments to explore linguistic mappings between Dravidian and an Indo-European languages. Results demonstrate the comparative performance across models and language pairs, emphasizing both the benefits of POS-tag fine-tuning and the complexities of cross-linguistic alignment.

1 Introduction

Word alignment is an essential task in natural language processing (NLP), for machine translation (MT) and cross-lingual information transfer. In this research, we focus on the Dravidian languages i.e. Tamil and Telugu, alongside an Indo-European language i.e. English, to analyze the agglutinative nature of Dravidian languages in contrast with English. The Dravidian language family comprises 26 languages, linguistically classified into three groups: South, Central, and North (Krishnamurti, 2003). Dravidian languages exhibit an agglutinative structure, where words are formed by combining morphemes, with each morpheme retaining its meaning and function. Based on this structure, we hypothesize that intra-family word alignment (e.g., Telugu & Tamil) may be more accurate than alignment across language families (e.g., English & Dravidian). In this study, we conduct word alignment on four language pairs: Telugu-Tamil, Tamil-Telugu, English-Telugu, and English-Tamil.

Traditional word alignment models, particularly statistical ones like IBM Models (Brown et al., 1993), GIZA++ (Och and Ney, 2003), FastAlign (Dyer et al., 2013) and Efloma(Östling and Tiedemann, 2016), face limitations in low-resource languages like Telugu and Tamil, where parallel data is scarce. While these models perform well in high-resource settings due to their reliance on abundant parallel data, they are less effective for low-resource languages. Recently, neural network-based models utilizing multilingual embeddings from BERT (Devlin, 2018) and XLM-RoBERTa (Conneau, 2019) have shown promise in overcoming these data limitations, generating word alignments even with minimal parallel corpora.

SimAlign (Sabet et al., 2020) and AWESOME-align (Dou and Neubig, 2021) are two neural models that utilize multilingual contextual embeddings to align words across languages. SimAlign computes alignments based on embedding similarity, using alignment strategies like Argmax, Itermax, and Match. In contrast, AWESOME-align applies a softmax-based alignment extraction process that predicts word alignments by calculating alignment probabilities between source and target embeddings. To improve alignment accuracy, AWESOME-align fine-tunes BERT-based models on parallel corpora using techniques like Masked Language Modeling (MLM) and Translation Language Modeling (TLM). These techniques help the model learn cross-lingual representations by predicting masked tokens within and across sentences, further enhancing alignment quality. Together, these neural approaches have demonstrated substantial improvements in alignment accuracy

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 154–159
January 19, 2025.

for low-resource languages, outperforming traditional statistical models (Sabet et al., 2020).

In addition to leverage multilingual BERT, we fine-tuned a mBERT model on POS-tagged English, Telugu, and Tamil paired data and applied it within both AWESOME-align and SimAlign to assess alignment accuracy improvements. We conducted alignment tasks before and after fine-tuning, comparing POS-tagged and untagged data to evaluate the impact of POS information. We also explored a novel approach by combining word and POS tag embeddings into enriched vectors, using two strategies: addition (summing embeddings) and concatenation (merging into an extended vector). Word alignments were extracted from these combined embeddings via cosine similarity to measure source-target word similarity. Although this combination approach aimed to leverage both semantic and syntactic information, results showed it did not significantly outperform alignments based solely on word embeddings.

Additionally, we expanded our study to include English alongside Telugu and Tamil, adding a cross-linguistic perspective. By aligning English-Telugu and English-Tamil pairs, we aimed to uncover potential linguistic patterns between the Dravidian and European language families. Using our fine-tuned mBERT approach on both POS-tagged and untagged data, we evaluated alignment accuracy across these language pairs. This helped us explore how well our methods work in mapping relationships between languages from different families, offering initial insights into the linguistic connections between Dravidian and European languages.

2 Data Preparation

For this research, we used parallel datasets covering Telugu & Tamil, English-Telugu, and English-Tamil pairs to conduct word alignment experiments. The Telugu and Tamil dataset was sourced from in-house resources, while the English-Telugu and English-Tamil data were obtained from the publicly available Samanantar (Ramesh et al., 2022) corpus.

In-House Telugu and Tamil Dataset: The in-house Telugu and Tamil dataset contains 13,000 manually translated sentences.[1] Prepared over a year, it reflects careful effort by annotators to ensure accuracy and linguistic quality, making it a reliable source for studying alignment within the Dravidian language family.

2.1 Data Preprocessing

To prepare the data for word alignment tasks, each sentence in the parallel corpora was tokenized using NLTK's tokenizer, ensuring a consistent tokenization scheme across all languages.

2.2 Part-of-Speech (POS) Tagging

All sentences in English, Telugu, and Tamil were POS-tagged using the Trankit library (Van Nguyen et al., 2021). This step added syntactic information to each token, which was used in later stages of the experiment to assess the impact of POS-tagged data on word alignment accuracy.

2.3 Dataset Splitting and Annotation

The dataset is divided into two subsets: training and testing. For each language pair, we allocated 12,000 sentence pairs for training and 1,000 sentence pairs for testing. To ensure an accurate evaluation, the test dataset was manually annotated with gold-standard word alignments by expert annotators, establishing a reliable reference for alignment quality assessment.

2.4 Data Organization for Alignment Tools

The source and target corpora were organized based on the input requirements of the alignment tools used in this study. Each dataset was structured to match the specific formats expected by SimAlign and AWESOME-align, ensuring compatibility and streamlined processing for word alignment tasks.

3 Methodology

3.1 Word Alignment with SimAlign and AWESOME-align

We conducted word alignment tasks using SimAlign and AWESOME-align across four language pairs: Telugu-Tamil, Tamil-Telugu, English-Tamil, and English-Telugu. Both models used multilingual BERT (mBERT) embeddings to generate cross-lingual word alignments. Data preprocessing steps, such as tokenization and POS tagging, are detailed in the Data Preparation section.

Embedding Extraction: For AWESOME-align, embeddings were extracted from the 8th layer of mBERT, which captures a balance of syntax and semantics (Dou and Neubig, 2021). SimAlign, which operates at the subword level, averaged subword embeddings to obtain word-level representations.

[1] https://github.com/parameshkrishnaa/Alignment-Parallel-Data/

It considers all 12 layers of mBERT and can use a concatenation of these layers (mBERT[conc]), providing flexible options without additional fine-tuning (Sabet et al., 2020).

Alignment Computation: The alignments were computed based on similarity matrices generated from the contextualized embeddings of each word in the parallel sentences. For SimAlign, the alignments were calculated using three strategies:

Argmax: Aligning words based on their maximum similarity score. *Itermax:* Focusing on mutual consistency between source and target alignments. *Match:* Using a bipartite matching algorithm to optimize total similarity between words.

In AWESOME-align, alignments were generated by leveraging probability thresholding to produce the final alignment pairs. This stage provided a baseline comparison between pre-trained neural alignment models.

3.1.1 Fine-tuning the Multilingual BERT Model

To improve alignment accuracy, mBERT was fine-tuned on 12,000 parallel sentence pairs for each language pair, following the AWESOME-align approach (Dou and Neubig, 2021), with two main objectives:

Masked Language Modeling (MLM): Enhances understanding by training the model to predict masked tokens. Translation Language Modeling (TLM): Reinforces cross-lingual representations by processing source and target sentences together.

This fine-tuned model was then utilized in both SimAlign and AWESOME-align. This phase allowed us to directly compare the performance of the pre-trained mBERT model against its fine-tuned version, thereby assessing the improvement in alignment accuracy when fine-tuning is applied to low-resource parallel data.

3.2 Word Alignment on POS-Tagged Data

To examine the effect of Part-of-Speech (POS) information, we conducted additional alignment tasks using POS-tagged data across all language pairs.

Alignments were performed with both SimAlign and AWESOME-align, following the same procedures as in the initial experiments. This allowed us to compare alignment accuracy between untagged and POS-tagged data.

3.2.1 Fine-tuning on POS-Tagged Data

We further evaluated alignment accuracy by fine-tuning mBERT on POS-tagged parallel data. This fine-tuning process followed the same MLM and TLM objectives as previously described, using POS-tagged sentence pairs for each language pair.

After fine-tuning, alignments were computed with both SimAlign and AWESOME-align to assess the impact of POS-tagged data on alignment accuracy in low-resource settings.

3.3 Embedding Combination and Cosine Similarity for Word Alignment

The methodology for combining word and part-of-speech (POS) tag embeddings in our research is inspired from (Siekmeier et al., 2021). In their study, the authors demonstrated the effectiveness of integrating linguistic annotations, such as POS tags and named entity recognition (NER) tags, into neural machine translation models to improve translation accuracy. Specifically, they proposed a method for combining token and tag embeddings within the encoder of the neural translation system. Their approach yielded significant improvements in translation quality, particularly when working with named entity tags, indicating that embedding linguistic features at the token level can enhance performance in specific NLP tasks.

Building upon this concept, we applied a similar embedding combination technique to the word alignment task in our study. The goal was to leverage both semantic and syntactic information by combining word embeddings with their corresponding POS tag embeddings. This was accomplished in two distinct ways:

Addition: In this approach, the word embeddings and their respective POS tag embeddings were summed element-wise to create a single, combined vector. This method preserves the dimensionality of the original word embeddings while integrating syntactic features at each token level.

Concatenation: For this method, the word embeddings and POS tag embeddings were concatenated, resulting in a more comprehensive feature vector. This concatenation allows for the representation of both semantic and syntactic information simultaneously, capturing a richer linguistic context for each token.

After generating the combined embeddings, a **cosine similarity matrix** was applied to compute the alignment between words in the source and

target languages across multiple language pairs: Telugu-Tamil, Tamil-Telugu, English-Telugu, and English-Tamil. The cosine similarity matrix measures the angular similarity between vectors in the embedding space, allowing for the identification of corresponding word pairs based on their similarity in both the semantic and syntactic dimensions.

4 Baseline

We compare our results against three widely used statistical word alignment models that rely on parallel training data:

- **FastAlign** (Dyer et al., 2013) is based on IBM Model 2 (Brown et al.), valued for its speed and simplicity while maintaining reasonable alignment quality.
- **Eflomal** (Östling and Tiedemann, 2016) is a Bayesian alignment model that uses Markov Chain Monte Carlo inference and is known to outperform FastAlign in both speed and accuracy.
- **GIZA++** (Och and Ney, 2003) is a well-established tool implementing IBM Models 1 to 4 (Brown et al., 1993). It is widely used in machine translation, and we used standard settings, including five iterations of the Hidden Markov Model (HMM) (Eddy, 1996) phase.

These statistical models serve as the baseline for evaluating the performance of neural approaches, particularly in low-resource language pairs like Telugu and Tamil.

5 Evaluation Measures

To evaluate alignment accuracy, we used the following measures:

- **Precision**: Measures the proportion of correct alignments out of all alignments made by the model.

$$Precision = \frac{|A \cap G|}{|A|}$$

- **Recall**: Measures the proportion of correct alignments out of all alignments in the gold-standard set.

$$Recall = \frac{|A \cap G|}{|G|}$$

- **Alignment Error Rate (AER)**: Provides an overall error rate by combining precision and recall. Lower AER indicates better alignment accuracy.

$$AER = 1 - \frac{2 \times |A \cap G|}{|A| + |G|}$$

- **F1 Score**: Balances precision and recall, providing a single accuracy score.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In these formulas, A represents the alignments predicted by the model, G represents the gold-standard alignments, and $|A \cap G|$ is the count of correct alignments.

6 Results

Neural vs. Statistical Models: Neural models (SimAlign and AWESOME-align) outperformed statistical models (GIZA++, FastAlign, and Eflomal) across all language pairs. The biggest improvements were seen in the Telugu-Tamil and Tamil-Telugu pair as shown in table [1] suggesting that neural models with multilingual embeddings work especially well for low-resource languages. SimAlign (Sabet et al., 2020), in particular, demonstrated the best performance among the neural models, especially for Telugu-Tamil and Tamil-Telugu pairs, due to shared linguistic features. In contrast, the improvements were smaller for English-Telugu and English-Tamil as shown in [1], likely because these language pairs lack similar structural features.

Effect of Fine-Tuning: Fine-tuning the multilingual BERT model improved alignment accuracy for all pairs, with the largest gains again in the Telugu-Tamil and Tamil-Telugu pair. Using Masked Language Modeling (MLM) and Translation Language Modeling (TLM) helped the model better understand cross-lingual connections, especially in Dravidian languages with shared grammatical structures.

Impact of POS-Tagged Data (Before and After Fine-Tuning): POS tagging was most beneficial for the Telugu-Tamil and Tamil-Telugu pairs after fine-tuning as shown in table [1], compared to untagged alignments. While the improvement was notable, the results were very low for the English-Telugu and English-Tamil pairs as shown in table[1]. This suggests that morphologically complex

Model	Type	Language pair							
		Telugu - Tamil		*Tamil - Telugu*		*English - Telugu*		*English - Tamil*	
		F1 ↑	**AER ↓**	**F1 ↑**	**AER ↓**	**F1 ↑**	**AER ↓**	**F1 ↑**	**AER ↓**
Fast-align	*untagged*	56.6	43.4	58.3	41.7	25.7	74.3	20.4	79.6
Giza++	*untagged*	54.7	45.3	56.8	43.7	23.3	76.7	17.8	82.2
Eflomal	*untagged*	65.5	34.5	67.7	32.3	27.8	72.2	12.4	87.6
SimAlign_inter	*untagged*	80.1	19.9	82.7	17.3	47	53	53.1	46.9
SimAlign_itermax	*untagged*	78.2	21.8	80.5	19.5	52.6	47.4	**54.8**	**45.2**
SimAlign_mwmf	*untagged*	73.2	26.8	75.5	24.5	52.9	47.1	52.4	47.6
Awesome_Align	*untagged*	64.2	35.8	66	34	23.2	76.8	31.8	68.2
SimAlign_inter_f	*untagged*	84.1	16	86.4	13.6	59.9	40.1	36.2	63.8
SimAlign_itermax_f	*untagged*	82.4	17.6	84.2	15.8	**65.6**	**34.4**	38.2	61.8
SimAlign_mwmf_f	*untagged*	74.3	25.7	76.4	23.6	65.3	34.7	37.6	62.4
Awesome_Align_f	*untagged*	65.9	34.1	67.8	32.2	37	63	27.1	72.9
SimAlign_inter	*tagged*	79.3	20.7	79.3	20.7	51.1	48.9	30.9	69.1
SimAlign_itermax	*tagged*	73.8	26.2	73.8	26.2	51.7	48.3	28.1	71.9
SimAlign_mwmf	*tagged*	70.8	29.2	70.8	29.2	49.1	50.9	26.6	73.4
Awesome_Align	*tagged*	71.7	28.3	68.1	31.9	30	70	19.2	80.8
Embed_add	*tagged*	41.2	58.8	34.4	65.6	20.5	79.5	12.6	87.4
Embed_concat	*tagged*	43.8	56.2	35.4	64.6	22.2	77.8	14.9	85.1
SimAlign_inter_f	*tagged*	**91.7**	**8.3**	**92.5**	**7.5**	64.6	35.4	37.3	62.7
SimAlign_itermax_f	*tagged*	84.4	15.6	84.6	15.4	60.1	39.9	31	69
SimAlign_mwmf_f	*tagged*	77.2	22.8	77.6	22.4	56.8	43.2	28.3	71.7
Awesome_Align_f	*tagged*	76.5	23.5	76.2	23.8	36.7	63.3	26.2	73.8
Embed_add_f	*tagged*	35.2	64.8	40.2	59.8	29.4	70.6	12.8	87.2
Embed_concat_f	*tagged*	37.9	62.1	58.9	41.1	34.4	65.6	15.3	84.7

Table 1: *Comparison of Word Alignments Across Language Pairs Using POS-Tagged and Untagged Datasets. The 'Type' column indicates whether the dataset used was POS-tagged or untagged. Models with 'f' denote fine-tuned versions, and the best results for each metric are highlighted in bold ('F1 ↑': highest value value is the better & 'AER ↓': lowest value is the better).*

languages like Telugu and Tamil gain more alignment accuracy from added POS information, while POS tagging is less useful for English-inclusive pairs where structural differences are more pronounced.

Combined Embeddings: Combining word and POS embeddings (through addition or concatenation) didn't significantly improve alignment accuracy over using word embeddings alone, even for Telugu and Tamil pairs, results shown in the tables[1] by the model names 'Embed_add & Embed_concat'. Although it could capture both meaning and structure, it didn't provide practical gains for these language pairs.

7 Limitations

While neural models showed strengths in low-resource alignment, this study faced several limitations that affected the quality of results. Dataset Quality, The Samanantar dataset for English-Telugu and English-Tamil contained translation inconsistencies, with many sentences poorly matched. This made it harder for alignment models to learn accurate mappings. High-quality, carefully curated parallel data is needed for better alignment and cross-linguistic analysis. Computational Constraints, Limited computational resources restricted the level of fine-tuning and testing of larger models. This limitation reduced the ability to optimize hyperparameters and experiment with deeper models that might improve accuracy. More computational resources would allow for broader testing and potentially better alignment results.

8 Conclusion

This study shows that neural models work better than traditional statistical models for word alignment, especially among low-resource Dravid-

ian language pairs like Telugu and Tamil. Neural models consistently achieved higher accuracy, with SimAlign (Sabet et al., 2020) performing particularly well in Telugu-Tamil and Tamil-Telugu alignments, likely due to shared structural features within the Dravidian language family. However, this advantage was smaller when aligning Dravidian languages with English, which has a different structure.

Fine-tuning with POS-tagged data improved alignment accuracy the most in Telugu-Tamil and Tamil-Telugu pairs, as the POS information helped the model understand sentence structure better. In English-inclusive pairs (English-Telugu, English-Tamil), POS tagging had less impact, likely due to structural differences and some limitations in dataset quality.

Combining word and POS embeddings did not lead to additional accuracy gains. Although it aimed to capture both meaning and structure, this approach did not perform better than using word embeddings alone.

In summary, our findings shows the adaptability of neural models to the linguistic structures of Dravidian languages, showing promise for improving alignment in low-resource Dravidian language pairs. Future research could build on these results by experimenting with enhanced fine-tuning techniques, exploring additional syntactic or morphological features, and addressing the dataset quality issues in English-Dravidian pairs to improve alignment accuracy further.

References

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Sean R Eddy. 1996. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

Bhadriraju Krishnamurti. 2003. The dravidian languages. *The Cambridge University*.

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

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106(1):125.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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

Aren Siekmeier, WonKee Lee, Hongseok Kwon, and Jong-Hyeok Lee. 2021. Tag assisted neural machine translation of film subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 255–262.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. *arXiv preprint arXiv:2101.03289*.