Tapping BERT for Preposition Sense Disambiguation

Anonymous ACL submission

Abstract

Prepositions are frequently occurring polysemous words. Disambiguation of prepositions is crucial in tasks like semantic role labelling, question answering, text entailment, and noun compound paraphrasing. In this paper, we propose a novel methodology for preposition sense disambiguation (PSD), which does not use any linguistic tools. In a supervised setting, the machine learning model is presented with sentences wherein prepositions have been annotated with 'senses'. These 'senses' are IDs in what is called 'The Preposition Project (TPP)'. We use the hidden layer representations from pre-trained BERT and its variants. The latent representations are then classified into the correct sense ID using a Multi-Layer Perceptron. The datasets used for this task are from SemEval-2007 Task-6 and Oxford English Corpus (OEC). Our methodology gives an accuracy of 86.85% on the SemEval task, which is better than the state-of-the-art.

1 Introduction

Prepositions are among the foremost commonly used terms and they are the most ambiguous words in English (Baldwin et al., 2009). According to the British National Corpus (Clear, 1993, BNC), prepositions account for four of the top 10 most frequently used terms in English (*of*, *to*, *in*, and *for*). They can impart different context to other parts of the sentences i.e. noun, verbs etc. It can be tricky to identify the meaning (or 'sense') of the preposition in a given sentence.

This ambiguity renders the task of disambiguation of prepositions to be a significant one. It helps in semantic role labeling (Ye and Baldwin, 2006), where the task is to identify predicates, extract their arguments, and label the arguments with predefined semantic roles. In most cases, the predicate is a verb, and argument is a noun phrase (subject) or a preposition phrase (direct object or indirect object). Consider the following example:

John ate some rice ***with***$_1$ *lentil soup* ***with***$_2$ *a spoon* ***with***$_3$ *his friend.*

#	Sense ID	Relation	Complement
1	1(1)	Accompanier	anything that can accompany the attachment point
2	4(3)	Means	an instrument in the action described by the attachment point
3	9(7)	Concomitant	sb or sth linked with subject via the POA

Table 1: Senses of the three occurrences of *with* in the above example. (sb: somebody; sth: something)

Table 1 explains how a prepositions can take different meanings. TPP gives several insights for each preposition sense in detail. Select terms like Sense ID, semantic relation and complement properties have been noted for the above example. Sense IDs are essentially labels defined by TPP. The integer outside the bracket refers to the super sense, and the integer inside the bracket refers to the sense of the preposition. Mapping for supersenses and list of all possible meanings for any preposition can be found on the TPP website[1].

Understanding the sense of the preposition can help us understand fine semantic relations in the sentence. This can be instrumental in other NLP tasks, like question-answering. In the example from Table 1, if the sense of *with* (Accompanier, Means, Concomitant) is identified, we can answer questions like - *What does John eat with rice?* (lentil soup), *What does John eat rice with?* (spoon), *Who does John eat with?* (friend). Further, SemEval-2013 Task4 (Hendrickx et al., 2013) discusses that prepositions are a preferred choice when it comes to understanding and expressing a relation between the components of a noun compound. Following up on this, Ponkiya et al. (2018) states that prepositional paraphrasing is a crucial step in the paraphrasing of noun compounds. Sim-

[1] https://www.clres.com/

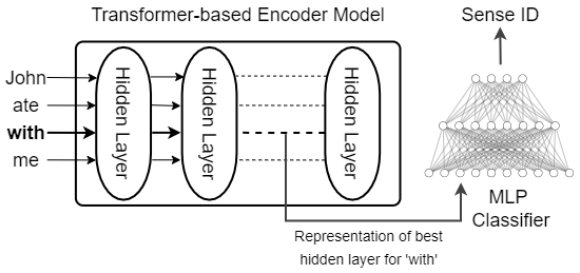

Figure 1: Model architecture for PSD. Latent representation of a transformer-based pre-trained encoder is passed through an MLP for each preposition.

ilarly, the application of PSD can also be seen in text entailment, phrasal verb paraphrasing, etc.

In this paper, we make use of more recent word representations from pretrained transformer-based language models like BERT (Devlin et al., 2019). Du et al. (2019) indicates that different hidden layers of language models learn different things. We identified the preposition-specific hidden layer, which improved the performance. Our approach achieved slightly better accuracy than the current state-of-the-art, without using any linguistic tool.

2 Related Work

Litkowski (2013) used lemmatizer, dependency parser as well as WordNet to extract features to get the senses of the prepositions to achieve state-of-the-art accuracy of 85.7% on the task of preposition sense disambiguation using the SemEval-2007 dataset. Average accuracy within 5% of the state-of-the-art, has been achieved in Srikumar and Roth (2013) and Hovy et al. (2010). Recently, Gong et al. (2018) used combined word vectors of left and right context as well as a context interplay vector as features to train the sense classifier. Hassani and Lee (2017) used deep convolutional neural networks along with lexical and syntactic features as well as word embeddings for sense disambiguation of spatial prepositions.

The contextualized BERT (Devlin et al., 2018) embeddings have been shown to be capable of clustering polysemic words into distinct sense regions in the embedding space (Wiedemann et al., 2019). Among the work done on the use of embeddings for word sense disambiguation (WSD) is Peters et al. (2018) that incorporated the pre-trained ELMo embeddings as WSD features. A study by Du et al. (2019) fine-tuned BERT and used various internal representations from the BERT encoder as features for WSD. Gessler and Schneider (2021) uses contextualized embeddings from BERT and its variants as representations of different senses of the prepositions. The sentences from the dataset are ranked according to the cosine similarity of these representations with those of the tokens from the dataset.

3 Our Approach

The crux of our approach is in using the embeddings learnt by the pre-trained transformer models. We hypothesize that the latent representations learnt by pre-trained transformer models have 'sense discriminative' capabilities, for each preposition. Using a preposition-specific Multi-Layer Perceptron (MLP) classifier, these representations are classified into corresponding senses.

We use BERT and its variants to get a contextual representation of a preposition. We provide a sentence as input to the model and get a representation of the preposition from different layers. We use a development set to decide which hidden-layer representation is best for each of the prepositions. If the i-th token in the sentence S is the preposition, then the representation from j-th layer, v_{ij} is

$$v_{ij} = BERT(S, i, j) \quad (1)$$

We treat j (hidden layer number) as a hyperparameter and use a development set to fine-tune. We feed the contextual representation v_{ij} to a single hidden layer MLP network, with softmax as final activation, to predict the sense of the preposition.

4 Experiments

In our approach, the sentence and the preposition are passed as inputs to the model. The pre-trained transformer models then generate the latent representation of the preposition in the sentence to be classified by the MLP classifier.

4.1 Datasets

We use the SemEval-2007 Task 6 dataset for testing our methodology for PSD. The corpus consists of 24,633 sentences in total. The dataset uses a repository of 334 senses for 34 prepositions. Surprisingly, 75 senses do not have a single example.

We also used the Oxford English Corpus (OEC) dataset[2] for our experiments as it contains data for much more prepositions than the SemEval-2007 dataset. This corpus contains 7,650 sentences covering 635 senses for 259 prepositions (including

[2] http://oxforddictionaries.com/us/words/the-oxford-english-corpus

Dataset	System	Techniques Used	Accuracy(%)
SemEval	Ye and Baldwin (2007)	chunker, dependency parser, named entity extractor, WordNet	69.3
	Litkowski (2013)	lemmatizer, dependency parser, WordNet	85.7
	Gonen and Goldberg (2016)	multilingual corpus, aligner, dependency parser	81.3
	Gong et al. (2018)	Word2Vec with fixed context size	80.0
	Our System	pretrained transformers, multi-layer perceptrons	**86.9**
OEC	Gong et al. (2018)	Word2Vec with fixed context size	40.0
	Our System	pretrained transformers, multi-layer perceptrons	**63.2**

Table 2: Performance of our system for PSD compared with reported results on various datasets.

phrases judged to be multi-word expressions displaying preposition behaviour).

For the SemEval dataset, we use the custom test and train sets provided by the task to train our model. For the OEC corpus, as no training and test data is given separately, we divide the corpus into training and test set using 80:20 split and then evaluate our models on the test set thus obtained. We carry the experiments for various train-test splits (every time we train the model from scratch) and report average accuracy.

4.2 Training

We experiment with BERT-base, BERT-large, DistilBERT (Sanh et al., 2019), Big Bird (Zaheer et al., 2020), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2019) using the Hugging-face library (Wolf et al., 2020) to load pertained models. The experiments consisted of choosing the best hidden layer of transformers for extracting the latent representation. Since we use the pre-trained transformer models for extracting representations for the prepositions, we froze the transformers' parameters during the MLP classifier training. So, the training only optimizes the parameters of the MLP classifier. We use Adam optimizer (default parameters) and the Cross-Entropy Loss for the training.

For evaluating the overall performance on the task, we find the accuracy of individual classifiers. We then compute the average overall prepositions for comparison with the baseline.

5 Results and Discussions

The results for our experiments using variants of the BERT model are shown in Table 3. The classifier for each representation was trained with the hidden layer that gave the best validation accuracy. It is observed that the classifier based on the representations from Big Bird performs the best amongst all the variants. The comparison of our system with other systems has been shown in Table 2, and as can be seen, our system outperforms all the existing systems. Additionally, our system is language-agnostic and does not require linguistic tools or static word embedding, unlike other work done in PSD.

Model	Accuracy(%)
BERT - base	85.4
BERT - large	86.1
DistilBERT	81.5
RoBERTa	83.8
Big Bird	86.9
ALBERT	83.4

Table 3: Performance of our system when using different pre-trained encoder models on the SemEval dataset.

Error Analysis

Some prepositions like *through*, *on* and *after* give accuracy less than 75%. We studied the misclassified samples to understand why the model is unable to learn their senses well. One big reason is the data imbalance. This can be validated while comparing the prepositions *on* and *of*. They both have a similar number of classes (20 and 17 respectively), but *of* has almost 4 times the data than *on*. Given this, it can be predicted that *of* should have significantly higher accuracy than *on*. Surprisingly, accuracy for *of* (0.84) is faintly better than accuracy for *on* (0.83). After deeper analysis, it appears that though *of* has the highest number of datapoints per sense in the dataset, the distribution is highly skewed. 9 out of 17 senses of *of* get only 8% of the data, while the 4 most frequent senses enjoy about 75% of the data. Many senses with few data points ($\sim$10) were mostly classified as a 3(1b), a sense with a high number ($\sim$700) of data. For preposition *on*, there are cases of improper distribution, but much fewer than *of*, which creates the disparity in the result.

We observed that if the difference of data between two senses was not large, the model failed to

distinguish between the senses properly. For example, among the senses of preposition *of*, the sense 11(6) represents *'noun representing the subject of action denoted in the POA'* and the sense 12(6a) represents *'noun representing the object of action denoted in the POA'*. The former sense was often predicted as the latter suggests that the model fails to identify the finer nuance in their meanings. On the contrary, if the meaning of the sense is starkly different from others, it can be easy for the model to correctly predict the sense. For example, for the preposition *above*, the sense 9(3) only had 8 sentences, but still got 100% accuracy during testing. The reason behind this becomes clear after looking at how all the 5 senses of *above* are defined. Sense 9(3) complements an *established norm* or a *specified amount* like '*above average*', '*above $ 19*', '*above 90%*'. This makes it easier for the model to identify the pattern and differentiate among senses.

Robustness of our approach

The sense of any preposition is highly influenced by the governor (attachment) and the complement of the preposition (Gessler and Schneider, 2021). Changing a word can change the preposition sense, which makes it crucial to discuss our model's robustness to minor changes in the input. We do this experiment for four randomly chosen prepositions. We report sentences for which the sense changes in Table 4 and sentences for which sense doesn't change in Table 5. The changes in input were made by altering the complement of the preposition (changing the prepositional phrase), the associated verb (governor) in the sentence, or a combination of these. Since actual labels for these sentences are not available, we verified them manually by going through their definition.

Sentence	Exp	Pred
I ate bread **with** *butter*.	1(1)	1(1)
I ate bread **with** *a fork*.	4(3)	4(3)
I ate bread **with** *disgust*.	7(5)	7(5)
It was his first visit **to** *Africa*.	1(1)	1(1)
It was his first *exposure* **to** *Americans*.	8(3)	8(3)
It *was similar* **to** *his first visit*.	10(4a)	10(4a)
I live **in** *India*.	1(1)	1(1)
I live **in** *the moment*.	3(2)	3(2)
I live **in** *the terror of dying*.	5(4)	5(4)
He read a book **by** *Roald Dahl*.	4(1c)	4(1c)
He read a book **by** *the end of the day*.	5(2)	5(2)
He *placed the* book **by** *the table*.	18(5)	18(5)

Table 4: Sentences with different preposition sense and predicted outputs (**Exp**: Expected, **Pred**: Predicted) [3]

Sentence	Exp	Pred
I ate bread **with** *butter*.	1(1)	1(1)
I ate bread **with** *butter and jam*.	1(1)	1(1)
I ate bread **with** *some butter*.	1(1)	5(3a)
It was his first visit **to** *Africa*.	1(1)	1(1)
It was his first *journey* **to** *America*.	1(1)	1(1)
It was his first visit **to** *that country*.	1(1)	1(1)
I live **in** *India*.	1(1)	1(1)
I live **in** *my home*.	1(1)	1(1)
I *sleep* **in** *my own house*.	1(1)	1(1)
He read a book **by** *Roald Dahl*.	4(1c)	4(1c)
He read a book **by** *a famous author*.	4(1c)	4(1c)
He *found* a book **by** *Roald Dahl*.	4(1c)	4(1c)

Table 5: Sentences with the same preposition sense and predicted outputs (**Exp**: Expected, **Pred**: Predicted) [3]

For most of the examples in Tables 4 and 5, the model outputs matched the expected senses, indicating that our model is robust to small changes in input. However, consider the third sentence for the preposition *with* in Table 5. Based on the manual evaluation, the preposition should have the sense 1(1), which means *accompanied by (another person or thing)*. However, we can see that the model outputs the sense 5(3), which describes *a material used for a purpose*. Both these sense descriptions could be similar for some cases, including our example. This ambiguity points to an important aspect that, even for humans, it could be difficult to decide one particular sense for a preposition in a sentence.

6 Conclusion and Future Work

This paper proposes a transformer-based method for PSD, which only relies on pre-trained language models. Among BERT variants, Big Bird performs the best, giving state-of-the-art results without relying on any linguistic machinery, thus drastically reducing the human effort required for the task. This methodology can also be extended to low resource languages, where linguistic resources are absent, as the BERT model is trained on the unannotated text corpus. In the future, we would like to investigate data augmentation techniques to expand the training dataset. The substitution of point of attachment or complement with 'similar' words seems promising.

[3]The description of exact senses, for each preposition, can be found at https://www.clres.com/db/TPPEditor.html

References

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. 2009. Prepositions in applications: A survey and introduction to the special issue. *Comput. Linguist.*, 35(2):119–149.

Jeremy H. Clear. 1993. *The British National Corpus*, page 163–187. MIT Press, Cambridge, MA, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*, pages 4171–4186. Association for Computational Linguistics.

Jiaju Du, Fanchao Qi, and Maosong Sun. 2019. Using bert for word sense disambiguation. *arXiv preprint arXiv:1909.08358*.

Luke Gessler and Nathan Schneider. 2021. Bert has uncommon sense: Similarity ranking for word sense bertology. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 539–547.

Hila Gonen and Yoav Goldberg. 2016. Semi supervised preposition-sense disambiguation using multilingual data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2718–2729.

Hongyu Gong, Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. Preposition sense disambiguation and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1510–1521. Association for Computational Linguistics.

Kaveh Hassani and Won-Sook Lee. 2017. Disambiguating spatial prepositions using deep convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition? dimensions of sense disambiguation for an interesting word class. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 454–462.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Ken Litkowski. 2013. Preposition disambiguation: Still a problem. *CL Research, Damascus, MD*, pages 1–8.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Girishkumar Ponkiya, Kevin Patel, Pushpak Bhattacharyya, and Girish Palshikar. 2018. Treat us like the sequences we are: Prepositional paraphrasing of noun compounds using LSTM. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1827–1836.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Vivek Srikumar and Dan Roth. 2013. Modeling semantic relations expressed by prepositions. *TACL*, 1:231–242.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Patrick Ye and Timothy Baldwin. 2006. Semantic role labeling of prepositional phrases. *ACM Trans. Asian Lang. Inf. Process.*, 5(3):228–244.

Patrick Ye and Timothy Baldwin. 2007. Melb-Yb. In *Proceedings of the 4th International Workshop on Semantic Evaluations - SemEval '07*, pages 241–244. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.