How Lexical is Bilingual Lexicon Induction?

Anonymous ACL submission

Abstract

In contemporary machine learning approaches to bilingual lexicon induction (BLI), a model learns a mapping between the embedding spaces of a language pair. Recently, retrieve-and-rank approach to BLI has achieved state of the art results on the task. However, the problem remains challenging in low-resource settings, due to the paucity of data. The task is complicated by factors such as lexical variation across languages. We argue that the incorporation of additional lexical information into the recent retrieve-and-rank approach should improve lexicon induction. We demonstrate the efficacy of our proposed approach on XLING, improving over the previous state of the art by an average of 2% across all language pairs.

1 Introduction

Bilingual lexicon induction (BLI) is fundamental to many downstream NLP applications, such as machine translation (Qi et al., 2018; Duan et al., 2020), cross-lingual information retrieval (Vulić and Moens, 2015), document classification (Klementiev et al., 2012), dependency parsing (Guo et al., 2015; Ahmad et al., 2019), and language acquisition and learning (Yuan et al., 2020). In addition, it facilitates model sharing between high-resource and their aligned low-resource languages.

Contemporary approaches to BLI involve alignment of embeddings trained on monolingual corpora into a shared vector space. A challenge of this approach is *hubness* – the problem of high density regions in cross-lingual embedding space where, in the alignment space, the embedding of a term in a source language is surrounded by a dense cluster of terms in the target language. These hub terms are difficult to align and are worthy of investigation. The recent cross-domain similarity local scaling (CSLS) addresses this by normalizing distances by the average distance of each term's embedding to its nearest neighbors (Conneau et al., 2017). While it would be desirable to take advantage of CSLS in a state-of-the-art BLI model such as BLICEr (Li et al., 2022b), computing nearest neighbors is prohibitively expensive. While performance is better due to a pairwise cross-attention mechanism, this affects our ability perform an approximate nearest neighbour lookup.

We propose instead to address the hubness problem by including simple lexical features. We start with the observation that the lexical similarity of a pair of languages tends to be indicated by a relatively high rank correlation of term frequency, particularly for certain parts of speech. Figure 1 shows, by part of speech, the Spearman's rank correlation of corresponding terms in the 5k vocabularies in the XLING corpus (Glavaš et al., 2019). All language pairs have a positive rank correlation. This is especially so for proper nouns (PROPN) and nouns and the least so for verbs. This suggests that including term frequency and part of speech as features to the model can improve alignment of terms in high-density regions of the embedding space. Indeed, our approach improves the state of the art by 2.75% and 1.2% on the semi-supervised and supervised splits, respectively, of the XLING benchmark. An additional benefit of our approach is that it does not incur the computational overhead of the more complex CSLS for the pairwise approach.

2 Related Work

Methods for BLI can be classified into unsupervised, semi-supervised and supervised approaches. While purely unsupervised methods (Conneau et al., 2017; Grave et al., 2018) have yielded impressive results on many language pairs, minimal supervision through a small seed dictionary has helped improve performance considerably especially when relatively low-resource languages are considered. Supervised and semi-supervised approaches typically assume a dictionary of 5k and 1k word cor-

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4381–4386

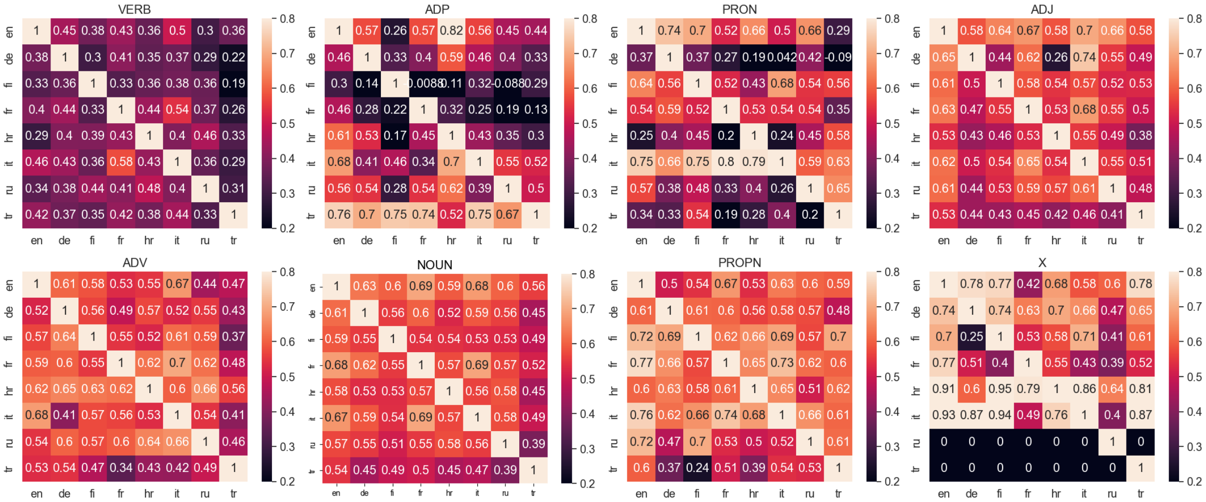

Figure 1: Spearman's Rank correlation of term frequencies derived from Common Crawl and Wikipedia. Cells containing a 0 have an insufficient (<10) number of terms in the source language for a particular part of speech.

respondences respectively for their training. In the semi-supervised setting, high-confidence alignments at each step are iteratively added as anchor points for subsequent training runs. Results on semi-supervised BLI have shown to improve by adopting a classification-based approach to iteratively refine and augment the seed translation dictionary (Karan et al., 2020). Such an approach allows for including arbitrary features such as term frequencies and sub-word information.

Recent semi-supervised and supervised approaches include ContrastiveBLI (Li et al., 2022a) and BLICEr which achieve state of the art results and serve as strong baselines for our work. ContrastiveBLI uses a familiar bi-encoder setup with hard negative sampling and contrastive learning. Two configurations for the bi-encoder are used:

C1: Fine-tuned bi-encoder on static fastText (Bojanowski et al., 2017) embeddings

C2: Fine-tuned bi-encoder on multi-lingual BERT (Devlin et al., 2018). C2 involves an additional step of a Procrustes mapping from C1 (300-dim) to the fine-tuned BERT (768-dim) embedding. The final embeddings are then a linear combination of the projected C1 and BERT representation.

BLICEr further improves performance through a reranking step using a fine-tuned cross-encoder based on xlm-roberta-large (Conneau et al., 2019). Instead of a simple binary classification over sampled hard negatives, a score polarization technique is described which increases or decreases CSLS scores on a base CLWE embedding (C1 or C2) based on the assigned label. The model is then trained to predict this score. Results in BLICEr include an additional step of linearly combining the cross-encoder score with CSLS of the base embedding for each candidate. We frequently allude to C1, C2, and BLICEr in subsequent sections.

3 Method

3.1 Retriever

We use the fastText-based C1 model described previously to retrieve top candidates for our reranker. C2, which leverages both fastText and multilingual-BERT, achieves better results both as a standalone BLI system as well as when used as a retriever in BLICEr. However, for simplicity, we only use the static fastText-based C1 model in our system and note that further improvement might be had from utilizing C2 as the retriever. For the supervised and semi-supervised systems, we utilize the C1 model trained on 5k and 1k data respectively. Consistent with recent work in BLI, we use the CSLS metric to score the nearest neighbors.

3.2 Base Reranker

Our ranking approach closely follows BLICEr in several respects. We score each source-target candidate pair using `xlm-roberta-large`[1]. The pairs are formatted – e.g., for English *apple* and French *pomme* – as `apple (english), pomme (français)!`. Also like BLICEr, we mine twenty hard negatives for each positive example to train the cross encoder for a binary classification objective.

While BLICEr demonstrated improvement in the supervised setting through score polarization,

[1]Available via https://huggingface.co/xlm-roberta-large.

		en-de	en-fi	en-fr	en-hr	en-it	en-ru	en-tr	de-*	fi-*	hr-*	it-*	ru-*	tr-*
	C1	50.4	42.15	61.65	35.65	59.60	42.50	38.15	41.89	35.81	40.26	65.63	48.61	32.06
	C2	50.85	45	62.5	42.35	61.05	46.05	41.05	44.75	39.39	44.68	66.77	50.26	35.57
	RCSLS+BLICEr	56.5	45.9	63.65	41.1	64.45	52.25	40.2	-	-	-	-	-	-
	C1+BLICEr	52.5	50.95	64.4	49.3	65.05	50.8	46.55	-	-	-	-	-	-
	C2(C1)+BLICEr	51.05	50.15	63	50.9	62.85	52.7	46.35	-	-	-	-	-	-
1k	XLM-R (Ours)	46.45	49.3	58.75	47.7	57.9	51.8	40.7	40.11	38.89	44.92	58.26	44.47	33.76
	LETOR(XLM+CSLS)	52.75	50.7	63.15	49	62.55	52.75	45.4	45.24	43.18	48.79	63.57	51.22	38.22
	LETOR+Freq	56.9	53	67.3	50.7	66.25	54.7	47.4	45.74	44.72	50.55	66.12	52.88	39.32
	LETOR+POS+Freq	58.2	53.15	67.3	50.75	66.3	54.75	47.74	46.51	44.98	50.44	67.22	53.04	39.25
	LETOR+POS+Freq+C1	**58.9**	**53.45**	**68.5**	**51.9**	**67.8**	**56.45**	**49.2**	**48.88**	**46.47**	**51.54**	**68.61**	**54.79**	**40.97**
	C1	54.9	44.6	65.05	40.7	63.45	49.15	41.35	44.21	39.21	43.18	66.51	50.1	35.38
	C2	57.75	47.17	67.2	47.2	65.6	50.5	44.74	47.17	42.71	48.22	67.86	52.33	38.66
	RCSLS + BLICEr	64	53.6	71.75	53.15	70.5	60.45	50.35	-	-	-	-	-	-
	C1+BLICEr	62.75	54.25	70.75	55.4	70.05	59.25	51.05	-	-	-	-	-	-
	C2(C1)+BLICEr	63.45	55.95	70.90	**57.55**	70.25	60.4	52.85	-	-	-	-	-	-
5k	XLM-R (Ours)	52.8	49.45	59	49.45	60.04	54.5	41.75	41.96	39.04	43.26	54.6	45	30.91
	LETOR(XLM+CSLS)	61.2	54.2	68.2	54.1	69.2	57.6	50.15	49.47	46.46	50.54	66.45	53.86	41.03
	LETOR+Freq	64.75	56.05	71.45	55.9	71.6	59.95	51.55	50.42	48.6	52.26	68.6	54.97	42.62
	LETOR+POS+Freq	64.75	57	72.4	56.65	72.6	61.05	52.35	51.57	48.6	53.1	69.96	56.08	42.57
	LETOR+POS+Freq+C1	**65.85**	**57.65**	**72.65**	57.05	**72.85**	**61.3**	**53.3**	**52.06**	**49.29**	**53.93**	**70.94**	**56.81**	**43.22**

Table 1: Results of our LETOR Method on XLING with 5k (supervised) and 1k (semi-supervised) data.

we maintain the simple binary objective in all our experiments. In the semi-supervised set, we use an additional 4k high-confidence pairs from C1 to augment the initial 1k seed dictionary. The model is fine tuned for one epoch on each language pair.

3.3 LETOR with XGBoost

We model our additional features through a Learning to Rank (Cao et al., 2007) objective using XGBoost (Chen and Guestrin, 2016). Each group consists of features belonging to the source word and all of its candidates. We use the following features as inputs to our LETOR model:

POS features: Source and candidate POS (categorical), and a binary label indicating a POS match.

Frequency features: Frequency ranks for the source and candidate described in section 1. In addition, we use the log-normalized raw frequency of source and candidate using wordfreq (Speer, 2022) which is derived from 8 different monolingual text corpora. We separately include the difference in frequency of the source-candidate pair.

Retriever & Reranker features: Raw logits returned from the base reranker (XLM-R) and CSLS score from the retriever (C1) for each pair.

Due to polysemy and synonymy, a group of candidates can consist of multiple positives as a result of synonymy in the target language. The listwise learning objective effectively shepherds our model into making better choices by taking into account relative candidate scores, their frequency alignment with the source and the part-of-speech information.

4 Results

We conduct our experiments on XLING which is a widely-used standard for BLI comprising 28 language pairs from 8 different languages. We choose XLING for its good mix of languages of differing typological similarities to compared to previous benchmarks (Conneau et al., 2017). The results from our modelling are presented in Table 1. We benchmark our results against BLICEr used in conjunction with different retrieval backbones - RCSLS (Joulin et al., 2018), C1, and C2. BLICEr only reports results on en-* XLING pairs, but we also report mean unidirectional accuracy of all other language pairs and compare results with C1 and C2 which are the best reported results on those pairs. LETOR-* rows use as input the raw logits from our own version of the fine-tuned XLM-R cross-encoder model. While this model is competitive with other baselines in the semi-supervised task, its standalone results are less impressive on the fully supervised set. This difference may be attributed to a more sophisticated sampling strategy and score polarization in BLICEr. We report results with a simple LETOR model using just the XLM-R logits and CSLS score, and also incremental changes from incorporating each of the features.

While we used XLM-R base, our final model still outperforms BLICEr in all en-* pairs on the 1k set, and 6 out of 7 pairs on the 5k set. Due to a more competitive cross-encoder baseline, the difference is more pronounced on the 1k set. We observe results from incorporating only the frequency-based features, as well as both POS and term frequency in our reranker. Part-of-speech information improves model accuracy in most cases, albeit marginally, however, best results are obtained when both features are used in conjunction. We further analyzed improvements on a per-POS basis and discovered the largest gains for nouns (7.3%) from amongst

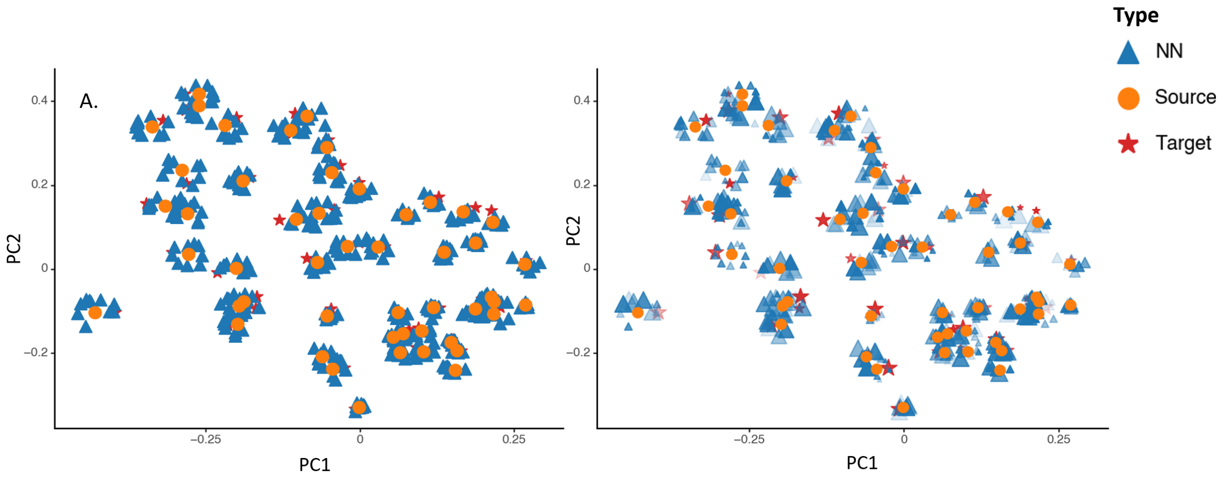

Figure 2: (LEFT) Nearest Neighbours, source and target. (RIGHT) Transparency signifies frequency difference of source-candidate pair, with point size indicating the likelihood of matching POS between source and target.

the most frequent POS types. This is consistent with our expectations in 1. Finally BLICEr reports results from using a linear combination of similarity scores using the cross-encoder as well as the CLWE backbone. For a more direct comparison, we do the same with our CLWE retriever (C1) which helps improve model performance across the board. Our approach yields improved results even in the absence of this additional step.

In Figure 2, we visualize a random sample of 50 baseline error cases in the en-de test set corrected by our LETOR model with (right) and without (left) the additional lexical features. Through the re-scaling of size of points with the probability of the POS matching, and transparency by frequency difference between source and candidate pairs, we observe how these features help the target stand out better in the right panel. This illustrates how our method tackles the hubness issue. While it is hard to disambiguate between close candidates in the embedding space, the LETOR model is able to turn to external cues in the form of these lexical features to help it make better predictions.

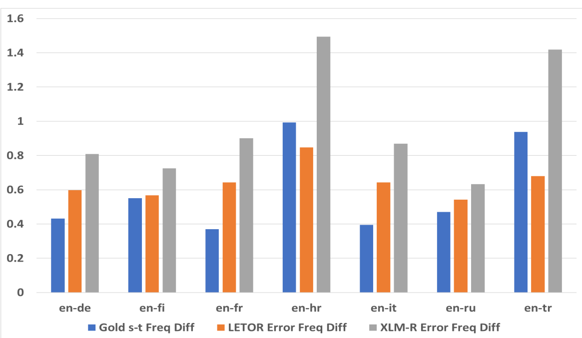

Figure 3: Mean absolute difference of term frequency.

To hone in on how the use of these lexical features affects a model, we do a post-hoc error analysis of our model on the test set using mean absolute difference of term frequency. Figure 3 shows the frequency difference in en-* pairs for the gold set and all error cases of XLM-R and LETOR. XLM-R consistently has higher frequency difference between source-predicted pairs. Conversely, predictions from the LETOR model have a frequency deviation that is more in-line with the gold distribution illustrating the models' higher proclivity to choose candidates with similar frequency.

5 Conclusion

Approaches to BLI have evolved to include full transformer based reranking methods. However, results on recent benchmarks indicate considerable scope of improvement still, particularly for low-resource or lexically dissimilar language pairs. While embeddings afford a rich semantic representation of individual words, we look towards supplementary features derived from individual monolingual corpora. Owing to the *hubness* issue we often retrieve many close candidates highlighting the need for better reranking and additional tools to deduce the correct correspondence. Our simple-yet-effective strategy of modeling lexical features using a ranking objective yields significant improvement over baselines. We are able to quantify their impact and demonstrate the efficacy of our approach across a wide array of language pairs. We hope this work inspires further research into both the acquisition and modelling of such features to further advance state of the art on bilingual lexicon induction.

6 Limitations

Our proposed approach uses a relatively simple learning-to-rank approach with XGBoost. This might be less effective at capturing complex, non-linear interactions between our features (POS types, term frequency, score from upstream models) than more sophisticated approaches such as Neural Networks. Also, as noted previously, we do not use the SOTA bi-encoder based model (C2) during our retrieval step due to compute and time constraints of training BERT-based bi-encoders for each individual language pair. Similarly we do not use scores from the SOTA cross-encoder, BLICEr, as input to the LETOR model. For these reasons, our approach might not fully exploit the extent of improvements made possible by incorporating such lexical features in the BLI task.

Another limitation of our work stems from ambiguity in the evaluation set of our benchmark dataset - XLING. Samples in XLING are constructed using word tuples derived from Google Translate. This approach does not account for issues arising due to polysemy and synonymy. The test set consists of a single target correspondence for each source word when, in practice, multiple correspondences might exist. Thus, a performance measure of any model evaluated on this test set, while indicative, does not fully reflect its efficacy on this task.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. 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 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online. Association for Computational Linguistics.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes. In *International Conference on Artificial Intelligence and Statistics*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Mladen Karan, Ivan Vulić, Anna Korhonen, and Goran Glavaš. 2020. Classification-based self-learning for weakly supervised bilingual lexicon induction. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6915–6922, Online. Association for Computational Linguistics.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2022a. Improving word translation via two-stage contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4353–4374, Dublin, Ireland. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Ivan Vulić, and Anna Korhonen. 2022b. Improving bilingual lexicon induction with cross-encoder reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4100–4116, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 363–372, New York, NY, USA. Association for Computing Machinery.

Michelle Yuan, Mozhi Zhang, Jordan Boyd-Graber, Leah Findlater, and Jordan Boyd-Graber. 2020. Interactive refinement of cross-lingual word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5984–5996, Online. Association for Computational Linguistics.

A Qualitative Examples

In Table 2, we show select examples from the en-de test set where the LETOR model is able to successfully map source words to the correct target correspondences. The words predicted with the baseline C1 model are very close alternatives from the target languages which translate to rotations, monochromatic, and sword for the source words motions, coloured, and spear respectively.

src	motions	coloured	spear
$rank_{src}$	15490	8450	13647
pos_{src}	NOUN	VERB	NOUN
$pred_{letor}$	bewegungen	farbig	speer
$rank_{letor}$	5855	19410	15249
pos_{letor}	NOUN	ADV	PROPN
$pred_{c1}$	rotationen	einfarbigen	schwert
$rank_{c1}$	122792	111085	7149
pos_{c1}	NOUN	ADJ	VERB

Table 2: Sample LETOR and C1 predictions (en-de)

The target words are much closer to the source words in relative frequency as shown by their ranks. The extra features help steer the LETOR model towards better predictions from amongst retrieved candidates that are very close in embedding space. We also plot the "motion" example in Figure 4. The correct translation "bewegungen" is better highlighted after applying transparency and size rescaling to indicate frequency difference and probability of part-of-speech match.

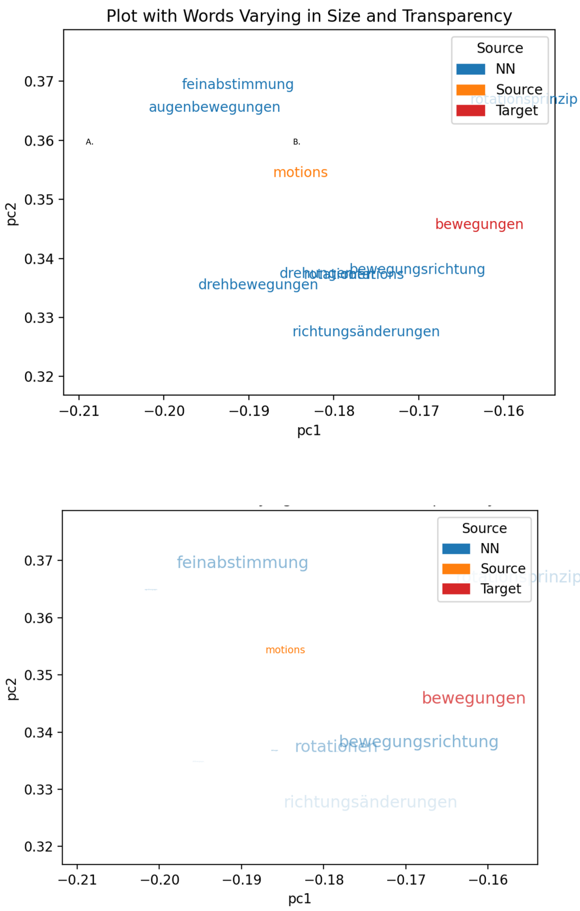

Figure 4: Example of BLI for "motion" without (top) and with (bottom) term frequency and POS information