Cross-Lingual Sense Alignment via Neural Machine Translation

Daniel Chen & Nikolai Lyssogor
Department of Linguistics, Department of Computer Science
University of Colorado
{daniel.chen-1, nikolai.lyssogor}@colorado.edu

1 Introduction

Proposition Bank (PropBank) (Palmer et al., 2005) is a corpus of text data containing annotations of sense entries that disambiguate the semantics of a verbal proposition. PropBank corpora currently include linguist-annotated, gold standard PropBank sense labels for English, Hindi, Chinese, Arabic, Finnish, Portuguese, Basque, Turkish, French, Spanish, German, Italian. A PropBank-annotated corpus contains valuable, linguist-verified contexts for disambiguation of polysemous or homonymous tokens, as well as the argument structure of the given proposition. In machine translation tasks where more literal, strict translations are preferred, such as probing for understanding of the source language's grammatical structure, having a MT-derived corpus of aligned PropBank senses can provide valuable insight into scaling a MT model for more coarse-grained or fine-grained translation.

To have humans create PropBank corpora for low-resourced languages at the same scale as English PropBank is a generally infeasible task, due to the overrepresentation of English annotators in linguistics research. The current state-of-the-art models for generating non-English PropBank corpora using the shallow semantics of English PropBank roles are part of the IBM Universal PropBank (Akbik et al., 2015). These models utilizes parallel corpora where both languages consist of real world text. The usage of machine translation as a tool for constructing the parallel corpus allows for eventual broader coverage of languages not covered by the IBM Universal PropBank, which currently houses corpora for Chinese, Finnish, French, German, Italian, Portuguese, and Spanish, all labeled with English PropBank senses.

We propose a novel mapping task that automatically projects existing PropBank senses in a source language (L1[1]) onto machine-translated text in a target language (L2[2]). PropBank is a language-specific corpus of text data containing annotations of sense entries that disambiguate the semantics of a verbal proposition. The output of the projection task is a semantic mapping from each sense *s* of a particular L1, PropBank-labeled verbal token to the corresponding lemmas in L2 that share the shallow semantics of each sense *s*. An automatic machine translation pipeline that produces accurate projections of existing L1 PropBank senses can quickly generate high-quality shallow semantic annotations for L2 target languages. The task also evaluates the efficacy of the DeepL translation model in generating semantically consistent translations. Creating a dataset of multilingual associations between verbs in L1 and their possible translations in L2 can assist with starting a new PropBank corpus for L2 that is built off of a gold standard L1 PropBank corpus. The produced dataset can also be used for multilingual semantic role labeling (SRL) tasks that utilize multilingual language models like mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020).

[1] source language and L1 are used interchangeably
[2] target language and L2 are used interchangeably

2 Related Work

2.1 Automatic Generation of PropBank Corpora

Our paper differs from previous work on automatic PropBank labeling because we focus on aligning L1 senses with L2 tokens for the purposes of building a linguistically-rich, MT-constructed corpus for a semantically under-resourced language, rather than evaluating a downstream task of building a model that aims to produce identical semantic roles in L1 and L2. Given that the success of the translation model in producing semantically transferable L2 translations already produces confounding factors, it is more appropriate to evaluate the translation model's linguistic merit alone, rather than producing and evaluating a secondary classifier (ostensibly for automatic SRL) that uses the translation model's output as a given.

Fei et al. (2020) create machine-translated corpora for the purpose of improving performance on downstream SRL tasks. They exploit a multilingual BiLSTM model to perform cross-lingual SRL, where the projection component aligns the target word with the highest alignment probability to a token equipped with an existing semantic role and POS tag. They use Google Translate as their translation model, while we will use DeepL, which covers less languages (29 to Google's 133) but is considered more accurate. AK and Yıldız (2019) uses WordNet synsets (synonymous sets) to perform semantic alignment, since Turkish data is already tagged with English WordNet. Akbik et al. (2015) predates usage of machine translation to create corpora, but presents a worthwhile experiment on a parallel French-English corpus, where

Language	Sentence	PropBank Sense
German	(1) Seine Arbeit wird von ehrenamtlichen Helfern und Regionalgruppen des Vereins **unterstützt**.	unterstützt.01
English (DeepL)	His work **is supported by** volunteers and regional groups of the association.	support.01
German	(2) Die Bombe **schlug** ein Loch in die Straße.	schlagen.01
English (DeepL)	The bomb **made** a hole in the road	hit.01
German	(3) Unsere Mannschaft **schlug** den Gegner (mit) 2:1.	schlagen.02
English (DeepL)	Our team **beat** the opponent (with) 2:1	defeat.01

Table 1: Projection of PropBank Sense via DeepL MT (German sentences from Universal PropBank 1.0 (Jindal et al., 2022) and Collins Online German Dictionary)

the French corpus is annotated with English PropBank labels as much as possible. We aim to quickly generate an L2 corpus using only L1 PropBank labels, a task that could be evaluated with this parallel corpus of real language data.

2.2 Multilingual Language Models

Devlin et al. (2019) released BERT, a landmark transformer-based pre-trained language model that can be fine-tuned on a variety of downstream tasks, including semantic role labeling. The authors also released mBERT, a language model trained in an unsupervised manner on the concatenation of the Wikipedia corpora from 104 languages. Conneau et al. (2020) improved on mBERT with their language model XLM-R by using a much larger corpus and model capacity. This was shown to outperform many monolingual BERT models.

The usage of multilingual language models allows for a shared semantic space to perform alignments between tokens of different languages. Our work seeks to establish a narrower iteration of that shared semantic space by creating explicit linguistic mappings between polysemous L1 tokens and the potentially disparate L2 tokens that correspond to each of the L1 token's senses. This language-to-language task is not as broad a coverage as a multilingual language model, but can be used to identify the gaps in those models' linguistic coverage.

We could also choose to exploit the multilingual models to better refine the mappings in our task, by categorizing the associations of the multilingual embeddings for the polysemous L1 token with the various multilingual embeddings for the L2 tokens that ideally correspond to the same variations of the L1 token's embeddings. That is, the different senses of the L1 token ideally have a different embedding, and those same differences can ideally be matched with the L2 tokens we derive in our machine-translated mapping.

2.3 Human-in-the-Loop SRL

Wang et al. (2017) devised a novel method for annotating text with semantic role labels. The authors realized that some, but not all, text could be annotated by non-expert crowd source workers, while other text required an expert in order to annotate properly. The system the authors devised classifies text according to whether it requires an expert to annotate it or not, thus improving the efficiency and cost of semantic role labeling. Our work pursues a similar goal of quickly generating high-quality SRL-compatible data, given that specific senses possess varying semantic roles in their own PropBank frames.

If our MT corpus proves to be linguistically robust, Wang et al. (2017)'s metrics for classifying annotation difficulty via sentence complexity and frame / role features can be applied to mark which sense mappings require careful evaluation and refinement by linguists or which sense mappings could be crowdsourced. Some cases that might require special linguistic attention are verb phrases in the source language that map to a light verb with a noun object in the target language, such as *migliorare* in Italian (lit. improves, a transitive reading) to *gets better* (an intransitive reading) in English. This human-in-the-loop process would ultimately generate a human-verified gold standard parallel corpus where the fundamental structure has been quickly pre-supplied by the MT mapping.

3 Methods

PropBank corpora currently include linguist-annotated, gold standard PropBank sense labels for the following source languages: English, Hindi, Chinese, Arabic, Finnish, Portuguese, Basque, Turkish, French, Spanish, German, Italian. In order to leverage these linguistically rich gold standard PropBank annotations, we use a pre-existing machine translation model, DeepL, to translate an L1 sentence with PropBank annotations to an L2 that does not have existing PropBank annotations. After using an existing L2 syntactic parser (Stanford CoreNLP (Manning et al., 2014), Stanza Dependency Parser (Qi et al., 2020)) to extract relevant tokens (expected to be primarily verbal) from the machine-translated sentence, we align those tokens with the PropBank sense(s) from

the original L1 text.

Alignment occurs with either a partial or exact match of the potentially disparate span of L2 tokens to the L1 span, depending on the evaluation granularity (e.g. *chalk (it) up (to/as)* in English can have other noun phrases besides the prototypical *it* in between the head verb *chalk* and the necessary particle preposition *up*, and necessary but flexible alternatives for the secondary, "argumentative" preposition: *to* or *as*. Alignments from syntactic parsers will be contrasted to alignments from embedding similarity metrics like SimAlign (Sabet et al., 2020). Table 1 depicts an example input-output sequence of the projection of the L1 PropBank sense after alignment.

4 Experimental Design

The Universal PropBank 2.0 (Jindal et al., 2022) is a collection of automatically generated high-quality PropBank corpora, created by exploiting monolingual SRL and multilingual parallel data. Universal PropBank aims to encode target languages with the shallow semantics of the resource-rich English PropBank, that is, verbal propositions in target languages are annotated with English PropBank labels. The 2.0 iteration covers 23 languages and contains gold standard data for Polish, Portuguese, and French PropBanks, all manually annotated with English PropBank.

To evaluate our task of using the DeepL MT model to perform crosslingual PropBank alignment, we will first learn the mappings by translating gold English PropBank data into one of the Universal PropBank 2.0's gold standard L2's (Polish, Portuguese, French). We then evaluate the learned mappings on the gold standard parallel PropBanks, which contain real language data, to see how well the automatic mappings correspond to human mappings. The evaluation translates from an L2 into English as L1, directly porting over the English PropBank label that was learned for the L2 token via machine translation, where each L2 token corresponds to any number of L1 senses.

For example, sentence (3) in Table 1 can be evaluated as follows. Machine-translated gold standard English PropBank data mapped the verb *schlagen* in L2 German to the L1 English defeat.01 sense. Assuming sentence (3) is part of the evaluation dataset, we identify that the main verb in the German text is *schlagen* and assign that token the L1 English defeat.01 sense according to the learned mapping. If there are multiple L1 senses that can be chosen from the L2 token, we force DeepL to select a translation containing an exact match for one of the L1 lemmas. For example, we force *schlagen* to translate to *defeat* or *hit*, even though DeepL initially translates *schlagen* to *beat*. Whichever forced translation DeepL produces is the automatic PropBank label we select[3]. We then compare the automatic mapping to the gold standard human annotations that specify the L1 English PropBank label that should go with the L2 text, as well as the proper span that receives the PropBank label. Since this is not a traditional classification task that relies on training data — given that the goal is to cover the entire English PropBank lexicon as a source language — we calculate accuracy of the automatic English PropBank labels according to the gold standard corpora.

[3]either defeat.01 or hit.01 for sentence (3) in Table 1. DeepL selects "Our team defeated the opponent (with) 2:1" as the first alternate translation

5 Bibliographical References

References

Koray AK and Olcay Taner Yıldız. 2019. Automatic Propbank generation for Turkish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 33–41, Varna, Bulgaria. INCOMA Ltd.

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition banks for multilingual semantic role labeling. 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 397–407, Beijing, China. Association for Computational Linguistics.

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.

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.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1700–1711.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language

processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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*.

Chenguang Wang, Alan Akbik, Laura Chiticariu, Yunyao Li, Fei Xia, and Anbang Xu. 2017. CROWD-IN-THE-LOOP: A hybrid approach for annotating semantic roles. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1913–1922, Copenhagen, Denmark. Association for Computational Linguistics.