Towards Semi-Automatic Generation of Proposition Banks for Low-Resource Languages

Alan Akbik
IBM Research
Almaden Research Center
San Jose, CA 95120

Vishwajeet Kumar
IIT Bombay
CS and Engineeering
Mumbai, India

Yunyao Li
IBM Research
Almaden Research Center
San Jose, CA 95120

{akbika,yunyaoli}@us.ibm.com vishwajeetkumar86@gmail.com

Abstract

Annotation projection based on parallel corpora has shown great promise in inexpensively creating Proposition Banks for languages for which high-quality parallel corpora and syntactic parsers are available. In this paper, we present an experimental study where we apply this approach to three languages that lack such resources: *Tamil*, *Bengali* and *Malayalam*. We find an average quality difference of 6 to 20 absolute F-measure points vis-a-vis high-resource languages, which indicates that annotation projection alone is insufficient in low-resource scenarios. Based on these results, we explore the possibility of using annotation projection as a starting point for inexpensive data curation involving both experts and non-experts. We give an outline of what such a process may look like and present an initial study to discuss its potential and challenges.

1 Introduction

Creating syntactically and semantically annotated NLP resources for low-resource languages is known to be immensely costly. For instance, the Proposition Bank (Palmer et al., 2005) was created by annotating predicate-argument structures in the Penn Treebank (Marcus et al., 1993) with shallow semantic labels: *frame* labels for verbal predicates and *role* labels for arguments. Similarly, the SALSA (Burchardt et al., 2006) resource added FrameNet-style annotations to the TIGER Treebank (Brants et al., 2002), the Chinese Propbank (Xue, 2008) is built on the Chinese Treebank (Xue et al., 2005), and so forth. Since each such layer of annotation typically requires years of manual work, the accumulated costs can be prohibitive for low-resource languages.

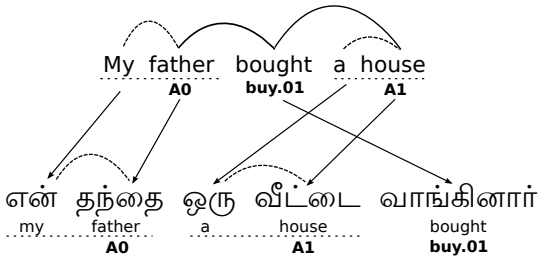

Figure 1: Annotation projection on a pair of very simple sentences. English Propbank frame (**buy.01**) and role (**A0**, **A1**) labels are projected onto aligned Tamil words. Furthermore, the typed dependencies between the words "my father" and "a house" (dotted lines) are projected onto their Tamil equivalents.

Recent work on **annotation projection** offers a way to inexpensively label a target language corpus with linguistic annotation (Padó and Lapata, 2009). This only requires a word-aligned parallel corpus of labeled English sentences and their translations in the target language. English labels are then automatically projected onto the aligned target language words. Refer to Figure 1 for an example.

Low-resource languages. However, previous work that investigated Propbank annotation projection has focused only on languages for which treebanks - and therefore syntactic parsers - already exist. Since syntactic information is typically used to increase projection accuracy (Padó and Lapata, 2009; Akbik et al., 2015), we must expect this approach to work less well for low-resource languages. In addition, low-resource languages have fewer sources of high-

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 993–998,
Austin, Texas, November 1-5, 2016.

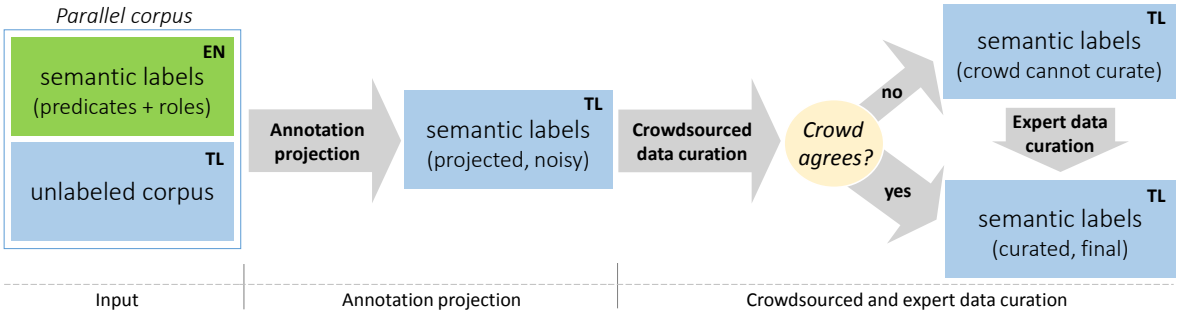

Figure 2: Proposed process of using annotation projection in a parallel corpus from English (EN) to a target language (TL) as basis for crowdsourced data curation. Experts are only involved in cases where the crowd cannot agree on a label.

quality parallel data available, further complicating annotation projection.

Contributions. In this paper, we present a study in which we apply annotation projection to three low-resource languages in order to quantify the difference in precision and recall vis-a-vis high-resource languages. Our study finds overall F1-measure of generated Proposition Banks to be significantly below state-of-the-art results, leading us to conclude that annotation projection may at best be a *starting point* for the generation of semantic resources for low-resource languages. To explore this idea, we outline a potential semi-automatic process in which we use crowdsourced data curation and limited expert involvement to confirm and correct automatically projected labels. Based on this initial study, we discuss the potential and challenges of the proposed approach.

2 Annotation Projection

Annotation projection takes as input a word-aligned parallel corpus of sentences in a source language (usually English) and their target language translations. A syntactic parser and a semantic role labeler produce labels for the English sentences, which are then projected onto aligned target language words. The underlying theory is that parallel sentences share a degree of syntactic and, in particular, semantic similarity, making such projection possible (Padó and Lapata, 2009).

State-of-the-art. Previous work analyzed errors in annotation projection and found that they are often caused by non-literal translations (Akbik et al., 2015). For this reason, previous work defined lexical and syntactic constraints to increase projection quality. These include verb filters to allow only verbs to be labeled as frames (Van der Plas et al., 2011), heuristics to ensure that only heads of syntactic constituents are labeled as arguments (Padó and Lapata, 2009) and the use of verb translation dictionaries (Akbik et al., 2015) to constrain frame mappings.

Adaptation to low-resource languages. Low-resource languages, however, lack syntactic parsers to identify target language predicate-argument structures. This requires us to make the following modifications to the approach:

Target language predicates We define lexical constraints using verb translation dictionaries. This ensures that only target language verbs that are aligned to literal source language translations are labeled as frames.

Target language arguments To identify arguments, we project not only the role label of source language arguments heads, but the entire argument dependency structure. This is illustrated in Figure 1: Two dependency arcs are projected from English onto Tamil, giving evidence that arguments **A0** and **A1** in the Tamil sentence each consist of two words.

This step produces a target language corpus with semantically annotated predicate-argument structure.

3 Outline of a Data Curation Process

As confirmed in the experiments section of this paper, the quality of the Proposition Banks generated using annotation projection is significantly lower for low-resource languages. We therefore propose to use this approach only as a starting point for an inexpensive curation process as illustrated in Figure 2:

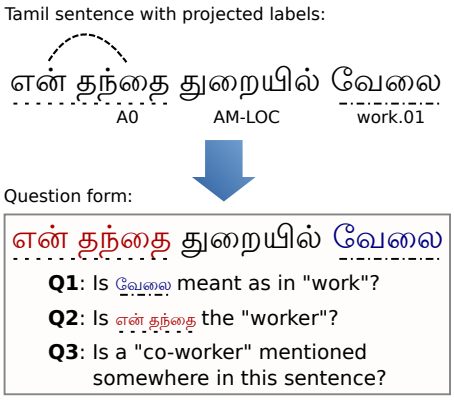

Figure 3: Example of how data curation questions may be formulated for the labels projected onto Tamil in Figure 1.

Step 1: Crowdsourced data curation. Previous work has experimented with different approaches in crowdsourcing to generate frame-semantic annotations over text (Hong and Baker, 2011), including selection tasks (selecting one answer from a list of options) (Fossati et al., 2013) and marking tasks (marking text passages that evoke a certain semantic role) (Feizabadi and Padó, 2014). While these studies only report moderate results on annotator correctness and agreement, our goal is different from these works in that we only wish to curate projected labels, not generate SRL annotations from scratch. A related project in extending FrameNet with paraphrases (Pavlick et al., 2015) has shown that the crowd can effectively curate wrong paraphrases by answering a series of confirm-or-reject questions.

For our initial study, we generate human readable question-answer pairs (He et al., 2015) using the label descriptions of the English Propbank (see Figure 3). We generate two types of questions:

Label confirmation questions are confirm-or-reject questions on whether projected labels are correct (e.g. **Q1** and **Q2** in Figure 3). Workers further qualify their answers to indicate whether a sequence of words marked as an argument is incomplete.

Missing label questions are marking tasks which ask whether any core role labels of a frame are missing. For example, the BUY.01 frame has 5 core roles (labeled **A0** to **A4**), one of which is the ”price” (**A3**). Since no ”price” is labeled in the Tamil sentence in Figure 3, question **Q3** asks users to add this label if a ”price” is mentioned.

DATA SET	Bengali	Malayalam	Tamil
OPENSUBTITLES2016	75K	224K	21K
SPOKENTUTORIALS	31K	17K	32K
Total # sentences	106K	241K	53K

Table 1: Parallel data sets and number of parallel sentences used for each language.

Our goal is to effectively distribute a large part of the curation workload. In cases where the crowd unanimously agrees, we remove labels judged to be incorrect and add labels judged to be missing.

Step 2: Expert data curation. We also expect a percentage of questions for which non-experts will give conflicting answers[1]. As Figure 2 shows, such cases will be passed to experts for further curation. However, for the purpose of scalability, we aim to keep expert involvement to a minimum.

4 Experimental Study

We report our initial investigations over the following questions: (1) What are the differences in annotation projection quality between low- and high-resource languages?; and (2) Can non-experts be leveraged to at least partially curate projected labels?

4.1 Experimental Setup

Languages. We evaluate three low-resource languages, namely *Bengali*, an Indo-Aryan language, as well as *Tamil* and *Malayalam*, two South Dravidian languages. Between them, they are estimated to have more than 300 million first language speakers, yet there are few NLP resources available.

Data sets. We use two parallel corpora (see Table 1): OPENSUBTITLES2016 (Tiedemann, 2012), a corpus automatically generated from movie subtitles, and SPOKENTUTORIALS, a corpus of technical-domain tutorial translations.

Evaluation. For the purpose of comparison to previous work on high-resource languages, we replicate

[1]Common problems for non-experts that we observe in our initial experiments involve ambiguities caused by implicit or causal role-predicate relationships, as well as figurative usage and hypotheticals.

		PRED.	ARGUMENT			
LANG.	Match	P	P	R	F1	%Agree
Bengali	partial	1.0	0.84	0.68	0.75	0.67
PROJECTED	exact	1.0	0.83	0.68	**0.75**	
Bengali	partial	1.0	0.88	0.69	0.78	
CURATED	exact	1.0	0.87	0.69	**0.77**	
Malayalam	partial	0.99	0.87	0.65	0.75	**0.65**
PROJECTED	exact	0.99	0.79	0.63	**0.7**	
Malayalam	partial	0.99	0.92	0.69	0.78	
CURATED	exact	0.99	0.84	0.67	**0.74**	
Tamil	partial	0.77	0.49	0.59	0.53	0.75
PROJECTED	exact	0.77	0.45	0.58	**0.5**	
Tamil	partial	0.77	0.62	0.67	0.64	
CURATED	exact	0.77	0.58	0.65	**0.61**	
Chinese	partial	0.97	0.93	0.83	0.88	**0.92**
(Akbik et al., 2015)	exact	0.97	0.83	0.81	**0.82**	
German	partial	0.96	0.95	0.73	0.83	**0.92**
(Akbik et al., 2015)	exact	0.96	0.91	0.73	**0.81**	
Hindi	partial	0.91	0.93	0.66	0.77	0.81
(Akbik et al., 2015)	exact	0.91	0.58	0.54	**0.56**	

Table 2: Estimated precision and recall for Tamil, Bengali and Malayalam before and after non-expert curation. We list state-of-the-art results for German and Hindi for comparison.

earlier evaluation practice and English preprocessing steps (Akbik et al., 2015). After projection, we randomly select 100 sentences for each target language and pass them to a curation step by 2 non-experts. We then measure the inter-annotator agreement and the quality of the generated Proposition Banks in terms of predicate precision[2] and argument F1-score before and after crowdsourced curation[3].

4.2 Results

The evaluation results are listed in Table 2. For comparison, we include evaluation results reported for three high-resource languages: German and Chinese, representing average high-resource results, as well as Hindi, a below-average outlier. We make the following observations:

Lower annotation projection quality. We find that the F1-scores of Bengali, Malayalam and Tamil are 6, 11 and 31 pp below that of an average high-resource language (as exemplified by German in Table 2). Bengali and Malayalam, however, do surpass Hindi, for which only a relatively poor dependency parser was used. This suggests that syntactic annotation projection may be a better method for identifying predicate-argument structures in languages that lack fully developed dependency parsers.

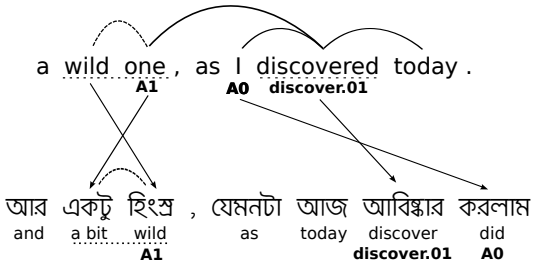

Figure 4: Example of a projection error. The verb *discover* in Bengali is a light verb construction. In addition, the pronoun *I* is not explicitly mentioned in the Bengali target sentence. This causes the pronoun *I* to be mistakenly aligned to the auxiliary of the light verb, causing it to be falsely labeled as **A0**.

Impact of parallel data. We note a significant impact of the size and quality of available parallel data on overall quality. For instance, the lowest-scoring language in our experiments, Tamil, use the smallest amount parallel data (see Table 1), most of which was from the SPOKENTUTORIALS corpus. This data is specific to the technical domain and seems less suited for annotation projection than the more general OPENSUBTITLES2016 corpus.

A qualitative inspection of projection errors points to a large portion of errors stemming from translation shifts. For instance, refer to Figure 4 for an English-Bengali example of the impact of even slight differences in translation: The English verb *discover* is expressed in Bengali as a light verb, while the pronoun *I* is dropped in the Bengali sentence (it is still implicitly evoked through the verb being in first person form). This causes the word alignment to align the English *I* to the Bengali auxiliary, onto which the role label **A0** is then incorrectly projected.

5 Discussion

In all three languages, we note improvements through curation. Argument F1-score improves to 77% (↑2 pp) for Bengali, to 74% (↑4 pp) for Malay-

[2]Since we do not ask missing label questions for predicates, we cannot estimate predicate recall.

[3]Following (Akbik et al., 2015), in the *exact* evaluation scheme, labels marked as correct and complete count as true positives. In *partial*, incomplete correct labels also count as true positives.

alam, and to 61% (↑11 pp) for Tamil on exact matches. Especially Tamil improves drastically, albeit from a much lower initial score than the other languages. This supports our general observation that crowd workers are good at spotting obvious errors, while they often disagree about more subtle differences in semantics. These results indicate that a curation process can at least be partially crowdsourced. An interesting question for further investigation is to what degree this is possible. As Table 2 shows, non-expert agreement in our initial study was far below reported expert agreement, with 25% to 35% of all questions problematic for non-experts.

A particular focus of our future work is therefore to quantify to which extent crowd-feedback can be valuable and how far the involvement of experts can be minimized for cost-effective resource generation. However, a Proposition Bank generated through this process would be peculiar in several ways:

Crowd semantics. First, generated Proposition Banks would be created in a drastically different way than current approaches that rely on experts to create and annotate frames. Effectively, the non-expert crowd would, to a large degree, shape the selection and annotation of English frame and role annotation for new target languages. An important question therefore is to what degree an auto-generated Propbank would differ from an expertly created one. In a related line of work (Akbik et al., 2016), we have conducted a preliminary comparison of an auto-generated Proposition Bank for Chinese and the manually created Chinese Proposition Bank (Xue and Palmer, 2005). Encouragingly, we find a significant overlap between both versions. Future work will further explore the usefulness of auto-generated Propbanks to train a semantic role labeler (Akbik and Li, 2016) and their usefulness for downstream applications in low-resource languages.

Partial syntactic annotation. Second, while curation of semantically labeled predicate-argument structure can be formulated as human intelligence tasks, this will not in all likelihood be possible for full parse trees. These Propbanks would therefore lack a treebank-style syntactic layer of annotation. Would an existing Propbank facilitate the future task of creating treebanks for low-resource languages? In other words, could the traditional order of first creating treebanks and then Propbanks be reversed?

PROPBANK	#SENTENCES	#LABELS	#FRAMES
Bengali	5,757	17,899	88
Malayalam	10,579	26,831	95
Tamil	3,486	11,765	68

Table 3: Number of labeled sentences, semantic labels and distinct frames of each auto-generated Propbank (*before* non-expert curation).

6 Conclusion and Outlook

We applied annotation projection to low-resource languages and found a significant drop in quality vis-a-vis high-resource languages. We then proposed and outlined a curation process for semi-automatically generating Proposition Banks and noted encouraging results in an initial study. To encourage discussion within the research community, we make our generated Proposition Banks for Bengali, Malayalam and Tamil (see Table 3 for an overview) publicly available[4].

References

[Akbik and Li2016] Alan Akbik and Yunyao Li. 2016. Polyglot: Multilingual semantic role labeling with unified labels. In *ACL 2016, 54th Annual Meeting of the Association for Computational Linguistics: Demostration Session*, page to appear.

[Akbik et al.2015] 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 *ACL 2015, 53rd Annual Meeting of the Association for Computational Linguistics Beijing, China*, pages 397–407.

[Akbik et al.2016] Alan Akbik, Xinyu Guan, and Yunyao Li. 2016. Multilingual aliasing for auto-generating proposition banks. In *COLING 2016, the 26th International Conference on Computational Linguistics (to appear)*.

[Brants et al.2002] Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, volume 168.

[Burchardt et al.2006] Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The salsa corpus: a german

[4]Datasets will be made available at this page: `http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=7454`

corpus resource for lexical semantics. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, volume 6.

[Feizabadi and Padó2014] Parvin Sadat Feizabadi and Sebastian Padó. 2014. Crowdsourcing annotation of non-local semantic roles. In *EACL*, pages 226–230.

[Fossati et al.2013] Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing framenet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 742–747, Sofia, Bulgaria, August. Association for Computational Linguistics.

[He et al.2015] Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal, September. Association for Computational Linguistics.

[Hong and Baker2011] Jisup Hong and Collin F Baker. 2011. How good is the crowd at real wsd? In *Proceedings of the 5th linguistic annotation workshop*, pages 30–37. Association for Computational Linguistics.

[Marcus et al.1993] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

[Padó and Lapata2009] Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, 36(1):307–340.

[Palmer et al.2005] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

[Pavlick et al.2015] Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015. Framenet+: Fast paraphrastic tripling of framenet. In *ACL (2)*, pages 408–413. The Association for Computer Linguistics.

[Tiedemann2012] Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of LREC 2012, Eighth International Conference on Language Resources and Evaluation*, pages 2214–2218.

[Van der Plas et al.2011] Lonneke Van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 299–304. Association for Computational Linguistics.

[Xue and Palmer2005] Nianwen Xue and Martha Palmer. 2005. Automatic semantic role labeling for chinese verbs. In *IJCAI*, volume 5, pages 1160–1165. Citeseer.

[Xue et al.2005] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238.

[Xue2008] Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational linguistics*, 34(2):225–255.