Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling

Alan Akbik*
Technische Universität Berlin, Germany
alan.akbik@tu-berlin.de

Laura Chiticariu **Marina Danilevsky** **Yunyao Li**
Shivakumar Vaithyanathan **Huaiyu Zhu**
IBM Research - Almaden
650 Harry Road, San Jose, CA 95120, USA
{chiti,mdanile,yunyaoli,vaithyan,huaiyu}@us.ibm.com

Abstract

Semantic role labeling (SRL) is crucial to natural language understanding as it identifies the predicate-argument structure in text with semantic labels. Unfortunately, resources required to construct SRL models are expensive to obtain and simply do not exist for most languages. In this paper, we present a two-stage method to enable the construction of SRL models for resource-poor languages by exploiting monolingual SRL and multilingual parallel data. Experimental results show that our method outperforms existing methods. We use our method to generate Proposition Banks with high to reasonable quality for 7 languages in three language families and release these resources to the research community.

1 Introduction

Semantic role labeling (SRL) is the task of automatically labeling predicates and arguments in a sentence with shallow semantic labels. This level of analysis provides a more stable semantic representation across syntactically different sentences, thereby enabling a range of NLP tasks such as information extraction and question answering (Shen and Lapata, 2007; Maqsud et al., 2014). Projects such as the Proposition Bank (PropBank) (Palmer et al., 2005) spent considerable effort to annotate corpora with semantic labels, in turn enabling supervised learning of statistical SRL parsers for English. Unfortunately, due to the high costs of manual annotation, comparable SRL resources do not exist for most other languages, with few exceptions (Hajič et al., 2009; Erk et al., 2003; Zaghouani et al., 2010; Vaidya et al., 2011).

As a cost-effective alternative to manual annotation, previous work has investigated the *direct projection* of semantic labels from a resource rich language (English) to a resource poor target language (TL) in parallel corpora (Pado, 2007; Van der Plas et al., 2011). The underlying assumption is that original and translated sentences in parallel corpora are semantically broadly equivalent. Hence, if English sentences of a parallel corpus are automatically labeled using an SRL system, these labels can be projected onto aligned words in the TL corpus, thereby automatically labeling the TL corpus with semantic labels. This way, PropBank-like resources can automatically be created that enable the training of statistical SRL systems for new TLs.

However, as noted in previous work (Pado, 2007; Van der Plas et al., 2011), aligned sentences in parallel corpora often exibit issues such as *translation*

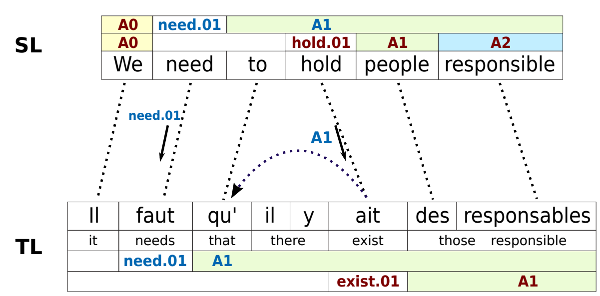


Figure 1: Pair of parallel sentences from $\texttt{French}_{\text{gold}}$ with word alignments (dotted lines), SRL labels for the English sentence, and gold SRL labels for the French sentence. Only two of the seven English SRL labels should be projected here.

*This work was conducted at IBM.

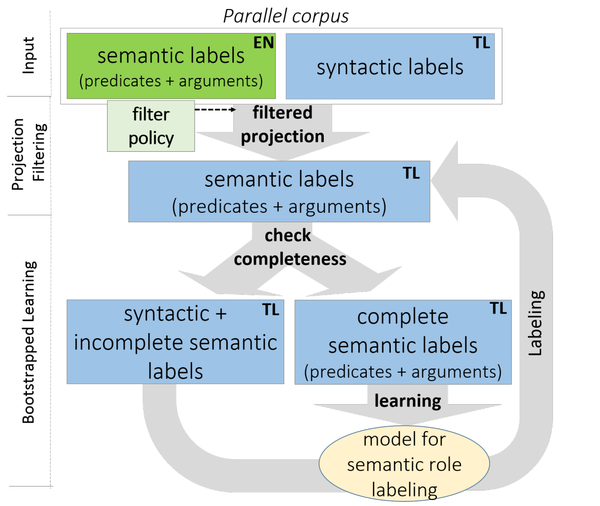


Figure 2: Overview of the proposed two-stage approach for projecting English (EN) semantic role labels onto a TL corpus.

shifts that go against this assumption. For example, in Fig. 1, the English sentence "*We need to **hold** people responsible*" is translated into a French sentence that literally reads as "*There need to **exist** those responsible*". Hence, the predicate label of the English word "*hold*" should not be projected onto the French verb, which has a different meaning. As the example in Fig. 1 shows, this means that only a subset of all SL labels can be directly projected.

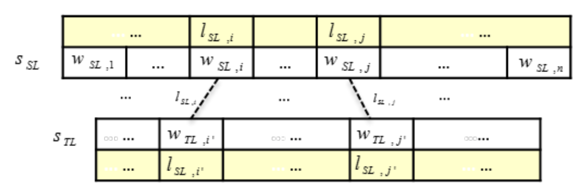
In this paper, we aim to create PropBank-like resources for a range of languages from different language groups. To this end, we propose a two-stage approach to cross-lingual semantic labeling that addresses such errors, shown in Fig. 2: Given a parallel corpus in which the source language (SL) side is automatically labeled with PropBank labels and the TL side is syntactically parsed, we use a *filtered projection* approach that allows the projection only of high-confidence SL labels. This results in a TL corpus with low recall but high precision. In the second stage, we repeatedly sample a subset of complete TL sentences and train a classifier to iteratively add new labels, significantly increasing the recall in the TL corpus while retaining the improvement in precision.

Our contributions are: (1) We propose *filtered projection* focused specifically on raising the precision of projected labels, based on a detailed analysis of direct projection errors. (2) We propose a *bootstrap learning approach* to retrain the SRL to iteratively improve recall without a significant reduction of precision, especially for arguments; (3) We demonstrate the effectiveness and generalizability of our approach via an extensive set of experiments over 7 different language pairs. (4) We generate PropBanks for each of these languages and release them to the research community.[1]

2 Stage 1: Filtered Annotation Projection

Stage 1 of our approach (Fig. 2) is designed to create a TL corpus with high precision semantic labels.

Direct Projection The idea of direct annotation projection (Van der Plas et al., 2011) is to transfer semantic labels from SL sentences to TL sentences according to word alignments. Formally, for each pair of sentences s_{SL} and s_{TL} in the parallel corpus, the word alignment produces alignment pairs ($w_{\mathrm{SL},i}$, $w_{\mathrm{TL},i'}$), where $w_{\mathrm{SL},i}$ and $w_{\mathrm{TL},i'}$ are words from s_{SL} and s_{TL} respectively. Under direct projection, if $l_{\mathrm{SL},i}$ is a predicate label for $w_{\mathrm{SL},i}$ and ($w_{\mathrm{SL},i}$, $w_{\mathrm{TL},i'}$) is an alignment pair, then $l_{\mathrm{SL},i}$ is transferred to $w_{\mathrm{TL},i'}$; If $l_{\mathrm{SL},j}$ is a predicate-argument label for ($w_{\mathrm{SL},i}$, $w_{\mathrm{SL},j}$), and ($w_{\mathrm{SL},i}$, $w_{\mathrm{TL},i'}$) and ($w_{\mathrm{SL},j}$, $w_{\mathrm{TL},j'}$) are alignment pairs, then $l_{\mathrm{SL},j}$ is transferred to ($w_{\mathrm{TL},i'}$, $w_{\mathrm{TL},j'}$), as illustrated below.



Filtered Projection As discussed earlier, direct projection is vulnerable to errors stemming from issues such as translation shifts. We propose *filtered projection* focused specifically on improving the precision of projected labels. Specifically, for a pair of sentences s_{SL} and s_{TL} in the parallel corpus, we retain the semantic label $l_{\mathrm{SL},i}$ projected from $w_{\mathrm{SL},i}$ onto $w_{\mathrm{TL},i'}$ if and only if it satisfies the filtering policies. This results in a target corpus containing fewer labels but of higher precision compared to that obtained via direct projection.

In the rest of the section, we analyze typical errors in direct projection (Sec. 2.2), present a set of filters to handle such errors (Sec. 2.3), and experimentally evaluate their effectiveness (Sec. 2.4).

[1]The resources are available on request.

Error Class	Number
Translation Shift: Predicate Mismatch	37
Translation Shift: Verb→Non-verb	36
No English Equivalent	8
Gold Data Errors	6
SRL Errors	5
Verb (near-)Synonyms	4
Light Verb Construction	3
Alignment Errors	1
Total	100

Table 1: Breakdown of error classes in **predicate** projection.

Error Class	Number
Non-Argument Head	33
SRL Errors	31
No English Equivalent	12
Gold Data Errors	11
Translation Shift: Argument Function	6
Parsing Errors	4
Alignment Errors	3
Total	100

Table 2: Breakdown of error classes in **argument** projection.

2.1 Experimental Setup

Data For experiments in this section and Sec. 3, we used the gold data set compiled by (Van der Plas et al., 2011), referred to as $\texttt{French}_{\text{gold}}$. It consists of 1,000 sentence-pairs from the English-French `Europarl` corpus (Koehn, 2005) with French sentences manually labeled with predicate and argument labels from the English Propbank.

Evaluation In line with previous work (Van der Plas et al., 2010), we count synonymous predicate labels sharing the same VERBNET (Schuler, 2005) class as true positives.[2] In addition, we exclude modal verbs from the evaluation due to inconsistent annotation.

Source Language SRL Throughout the rest of the paper, we use CLEARNLP (Choi and McCallum, 2013), a state-of-the-art SRL system, to produce semantic labels for English text.

2.2 Error Analysis

We observe that direct projection labels have both low precision and low recall (see Tab. 3 (*Direct*)).

Analysis of False Negatives The low recall of direct projection is not surprising; most semantic labels in the French sentences do not appear in the corresponding English sentences at all. Specifically, among 1,741 predicate labels in the French sentences, only 778 exist in the corresponding English sentences, imposing a 45% upper bound on the recall for projected predicates. Similarly, of the 5,061 argument labels in the French sentences, only 1,757 exist in the corresponding English sentences, resulting in a 35% upper bound on recall for arguments.[3]

[2]For instance, the French verb *sembler* may be correctly labeled as either of the synonyms: *seem.01* or *appear.02*.

[3]This upper bound is different from the one reported in (Van der Plas et al., 2011) which corresponds to the inter-annotator agreement over manual annotation of 100 sentences.

Analysis of False Positives While the recall produced by direct projection is close to the theoretical upper bound, the precision is far from the theoretical upper bound of 100%. To understand causes of false positives, we examine a random sample of 200 false positives, of which 100 are incorrect predicate labels, and 100 are incorrect argument labels belonging to correctly projected predicates. Tab. 1 and 2 show the detailed breakdown of errors for predicates and arguments, respectively. We first analyze the most common types of errors and discuss the residual errors later in Sec. 2.5.

• **Translation Shift: Predicate Mismatch** The most common predicate errors (37%) are translation shifts in which an English predicate is aligned to a French verb with a different meaning. Fig. 1 illustrates such a translation shift: label *hold.01* of English verb *hold* is wrongly projected onto the French verb *ait*, which is labeled as *exist.01* in $\texttt{French}_{\text{gold}}$.

• **Translation Shift: Verb→Non-Verb** is another common predicate error (36%). English verbs may be aligned with TL words other than verbs, which is often indicative of translation shifts. For instance, in the following sentence pair

s_{SL}	`We`	`know`	`what`	`happened`
s_{FR}	`On`	`connait`	`la`	`suite`
	We	know	the	result

the English verb *happen* is aligned to the French noun *suite (result)*, causing it to be wrongly projected with the English predicate label *happen.01*.

• **Non-Argument Head** The most common argument error (33%) is caused by the projection of argument labels onto words other than the syntactic head of a target verb's argument. For example, in Fig. 1 the label *A1* on the English *hold* is wrongly transferred to the French *ait*, which is not the syntactic head of the complement.

2.3 Filters

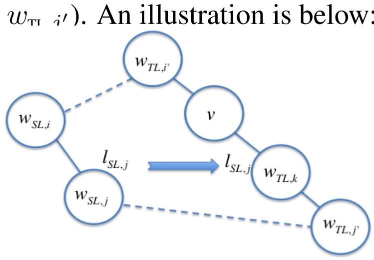
We consider the following filters to remove the most common types of false positives.

Verb Filter (VF) targets Verb→Non-Verb translation shift errors (Van der Plas et al., 2011). Formally, if direct projection transfers predicate label $l_{\text{SL},i}$ from $w_{\text{SL},i}$ onto $w_{\text{TL},i'}$, retain $l_{\text{SL},i}$ only if both $w_{\text{SL},i}$ and $w_{\text{TL},i'}$ are verbs.

Translation Filter (TF) handles both Predicate Mismatch and Verb→Non-Verb translation shift errors. It makes use of a translation dictionary and allows projection only if the TL verb is a valid translation of the SL verb. In addition, in order to ensure consistent predicate labels throughout the TL corpus, if a SL verb has several possible synonymous translations, it allows projection only for the most commonly observed translation.

Formally, for an aligned pair $(w_{\text{SL},i}, w_{\text{TL},i'})$ where $w_{\text{SL},i}$ has predicate label $l_{\text{SL},i}$, if $(w_{\text{SL},i}, w_{\text{TL},i'})$ is not a verb to verb translation from SL to TL, assign no label to $w_{\text{TL},i'}$. Otherwise, split the set of SL translations of $w_{\text{TL},i'}$ into synonym sets $S_1, S_2, \dots$; For each k, let W^k be the subset of S_k most commonly aligned with $w_{\text{TL},i'}$; If $w_{\text{SL},i}$ is in one of these W^k, assign label $l_{\text{SL},i}$ to $w_{\text{TL},i'}$; Otherwise assign no label to $w_{\text{TL},i'}$.

Reattachment Heuristic (RH) targets non-argument head errors that occur if a TL argument is not the direct child of a verb in the dependency parse tree of its sentence.[4] Assume direct projection transfers the predicate-argument label $l_{\text{SL},j}$ from ($w_{\text{SL},i}$, $w_{\text{SL},j}$) onto ($w_{\text{TL},i'}$, $w_{\text{TL},j'}$). Find the immediate ancestor verb of $w_{\text{TL},j'}$ in the dependency parse tree. Denote as $w_{\text{TL},k}$ its child that is an ancestor of $w_{\text{TL},j'}$. Assign the label $l_{\text{SL},j}$ to ($w_{\text{TL},i'}$, $w_{\text{TL},k}$) instead of ($w_{\text{TL},i'}$, $w_{\text{TL},j'}$). An illustration is below:



RH ensures that labels are always attached to the syntactic heads of their respective arguments, as determined by the dependency tree. An example of such reattachment is illustrated in Fig. 1 (curved arrow on TL sentence).

[4] In (Padó and Lapata, 2009), a similar filtering method is defined over constituent-based trees to reduce the set of viable nodes for argument labels to all nodes that are not a child of some ancestor of the predicate.

PROJECTION	PREDICATE P	PREDICATE R	PREDICATE F1	ARGUMENT P	ARGUMENT R	ARGUMENT F1
Direct	0.45	0.4	0.43	0.43	0.31	0.36
VF	0.59	0.4	0.48	0.53	0.31	0.39
TF	**0.88**	0.36	0.51	0.58	0.17	0.27
VF+RH	0.59	0.4	0.48	0.68	0.35	0.46
TF+RH	**0.88**	0.36	0.51	**0.75**	0.2	0.31
Upper Bound	1	0.45	0.62	1	0.35	0.51

Table 3: Quality of predicate and argument labels for different projection methods on $\texttt{French}_{\text{gold}}$, including upper bound.

2.4 Filter Effectiveness

We now present an initial validation on the effectiveness of the aforementioned filters by evaluating their contribution to annotation projection quality for $\texttt{French}_{\text{gold}}$, as summarized in Tab. 3.

VF reduces the number of wrongly projected predicate labels, resulting in an increase of predicate precision to 59% (↑14 pp), without impact to recall. As a side effect, argument precision also increases to 53% (↑10 pp), since, if a predicate label cannot be projected, none of its arguments can be projected.

TF[5] reduces the number of wrongly projected predicate labels even more significantly, increasing predicate precision to 88% (↑43 pp), at a small cost to recall. Again, argument precision increases as a side effect. However, as expected, argument recall decreases significantly (↓14 pp, to 17%), as many arguments can no longer be projected.

RH targets argument labels directly (unlike VF and TF), significantly increasing argument precision and slightly increasing argument recall.

In summary, initial experiments confirm that our proposed filters are effective in improving precision of projected labels at a small cost in recall. In fact, TF+RH results in nearly 100% improvement in predicate and argument labels precision with a much smaller drop in recall.

2.5 Residual Errors

Filtered projection removes the most common errors discussed in Sec. 2.2. Most of the remaining errors

[5] In all experiments in this paper, we derived the translation dictionaries from the WIKTIONARY project and used VERBNET and WORDNET to find SL synonym groups.

come from the following sources.

SRL Errors The most common residual errors in the remaining projected labels, especially for argument labels, are caused by mistakes made by the English SRL system. Any wrong label it assigns to an English sentence may be projected onto the TL sentence, resulting in false positives.

No English Equivalent A small number of errors occur due to French particularities that do not exist in English. Such errors include certain French verbs for which no appropriate English PropBank labels exists, and French-specific syntactic particularities.[6]

Gold Data Errors Our evaluation so far relies on $\texttt{French}_{\text{gold}}$ as ground truth. Unfortunately, $\texttt{French}_{\text{gold}}$ does contain a small number of errors (e.g. missing argument labels). As a result, some correctly projected labels are being mistaken as false positives, causing a drop in both precision and recall. We therefore expect the true precision and recall of the approach to be somewhat higher than the estimate based on $\texttt{French}_{\text{gold}}$.

3 Stage 2: Bootstrapped Training of SRL

As discussed earlier, the TL corpus generated via filtered projection suffers from low recall. We address this issue with the second stage of our method.

Relabeling The idea of relabeling (Van der Plas et al., 2011) is to first train an SRL system over a TL corpus labeled using direct projection (with VF filter) and then use this SRL to relabel the corpus, effectively overwriting the projected labels with potentially less noisy predicted labels.

We first present an analysis on relabeling in concert with our proposed filters (Sec. 3.1), which motivates our bootstrap algorithm (Sec. 3.2).

3.1 Analysis of Relabeling Approach

We use the same experimental setup as in Sec. 2, and produce a labeled French corpus for each filtered annotation method. We then train an off-the-shelf SRL system (Björkelund et al., 2009) on each generated corpus and use it to relabel the corpus.

We measure precision and recall of each resulting TL corpus against $\texttt{French}_{\text{gold}}$ (see Tab. 4). Across all experiments, relabeling consistently improves recall over projection. The results also show how different factors affect the performance of relabeling.

PROJECTION	PREDICATE			ARGUMENT		
SRL training	P	R	F1	P	R	F1
DIRECT						
–	0.45	0.40	0.43	0.43	0.31	0.36
relabel (SP)	0.49	0.57	0.53	0.52	0.43	0.47
relabel (OW)	0.66	0.60	0.63	0.71	0.37	0.49
VERB FILTER (VF)						
–	0.59	0.40	0.48	0.53	0.31	0.39
relabel (SP)	0.57	0.55	0.56	0.61	0.42	0.50
relabel (OW) (Van der Plas et al., 2011)	0.56	0.55	0.56	0.69	0.31	0.43
PROPOSED (TF+RH)						
–	0.88	0.36	0.51	0.75	0.20	0.31
relabel$_{\text{full data}}$(SP)	**0.83**	0.58	0.68	**0.75**	0.41	0.53
relabel$_{\text{full data}}$(OW)	0.78	0.51	0.62	0.73	0.35	0.47
relabel$_{\text{comp. sent.}}$(SP)	0.80	0.64	0.71	0.68	0.48	0.56
relabel$_{\text{comp. sent.}}$(OW)	0.62	0.60	0.61	0.55	0.40	0.47
bootstrap (iter. 3)	0.78	0.68	**0.73**	0.71	0.55	**0.62**
bootstrap (terminate)	0.77	**0.70**	**0.73**	0.64	**0.60**	**0.62**

Table 4: Experiments on $\texttt{French}_{\text{gold}}$, with different projection and SRL training methods. SP=Supplement; OW=Overwrite.

Supplement vs. Overwrite Projected Labels The labels produced by the trained SRL can be used to either *overwrite* projected labels as in (Van der Plas et al., 2011), or to *supplement* them (supplying labels only for words w/o projected labels). Whether to overwrite or supplement depends on whether labels produced by the trained SRL are of higher quality than the projected labels. We find that while predicted labels are of higher precision than directly projected labels, they are of lower precision than labels post filtered projection. Therefore, for filtered projection, it makes more sense to allow predicted labels to only *supplement* projected labels.

Impact of Sampling Method We are further interested in learning the impact of sampling the data on the quality of relabeling. For the best filter found earlier (TF+RH), we compare SRL trained on the entire data set ($_{\text{full data}}$) with SRL trained only on the subset of completely annotated sentences ($_{\text{comp. sent.}}$), where completeness is defined as:

Definition 1. *A direct component of a labeled sentence s_{TL} is either a verb in s_{TL} or a syntactic dependent of a verb. Then s_{TL} is **k-complete** if s_{TL} contains equal to or fewer than k unlabeled direct compo-*

[6]French negations, for instance, are split into a particle and a connegative. In the annotation scheme used in $\texttt{French}_{\text{gold}}$, particles and connegatives are labeled differently.

Algorithm 1 Bootstrap learning algorithm

Require: Corpus C_{TL} with initial set of labels L_{TL}, and resampling threshold function $k(i)$;
for $i = 1$ to ∞ **do**
  Let $k_i = k(i)$;
  Let $C_{\text{TL}}{}^{\text{comp}} = \{w \in C_{\text{TL}} : w \in s_{\text{TL}}, s_{\text{TL}} \text{is } k_i\text{-complete}\}$;
  Let $L_{\text{TL}}{}^{\text{comp}}$ be subset of L_{TL} appearing on $C_{\text{TL}}{}^{\text{comp}}$;
  Train an SRL on $(C_{\text{TL}}{}^{\text{comp}}, L_{\text{TL}}{}^{\text{comp}})$;
  Use the SRL to produce label set $L_{\text{TL}}{}^{\text{new}}$ on C_{TL};
  Let $C_{\text{TL}}{}^{\text{no.lab}} = \{w \in C_{\text{TL}} : w \text{ not labelled by } L_{\text{TL}}\}$;
  Let $L_{\text{TL}}{}^{\text{suppl}}$ be subset of $L_{\text{TL}}{}^{\text{new}}$ appearing on $C_{\text{TL}}{}^{\text{no.lab}}$;
  if $L_{\text{TL}}{}^{\text{suppl}} = \emptyset$ **then**
    Return the SRL;
  end if
  Let $L_{\text{TL}} = L_{\text{TL}} \cup L_{\text{TL}}{}^{\text{suppl}}$;
end for

*nents. 0-complete is abbreviated as **complete**.*

We observe that for TF+RH, when new labels supplement projected labels, relabeling over complete sentences results in better recall at slightly reduced precision, while including incomplete sentences into the training data reduces recall, but improves precision. While this finding may seem counterintuitive, it can be explained by how statistical SRL works. A densely labeled training data (such as comp. sent.) usually results in an SRL that generates densely labeled sentences, resulting in better recall but poorer precision. On the other hand, training data that is sparsely labeled results in an SRL that weighs the option of not assigning a label with higher probability, resulting in better precision and poorer recall. In short, one can control the tradeoff between precision and recall of SRL output by manipulating the completeness of the training data.

3.2 Bootstrap Learning

Building on the observation that we can sample data in such a way as to either favor precision or recall, we propose a bootstrapping algorithm to train an SRL iteratively over *k-complete* subsets of the data which are supplemented by high precision labels produced from previous iteration. The detailed algorithm is depicted in Algorithm 1.

Resampling Threshold Our goal is to use bootstrap learning to improve recall without sacrificing too much precision.

Proposition 1. *Under any resampling threshold, the set of labels L_{TL} increases monotonically in each iteration of Algorithm 1.*

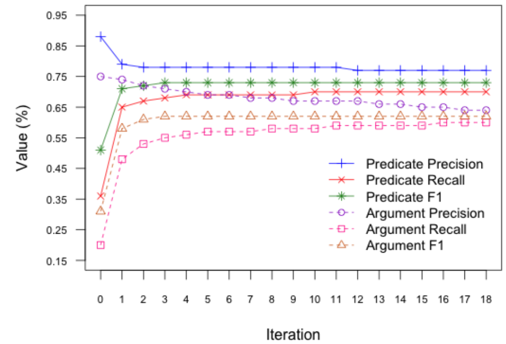


Figure 3: Values at each bootstrap iteration.

Since Prop. 1 guarantees the increase of the set of labels, we need to select a resampling function to favor precision while improving recall. Specifically, we use the formula $k(i) = \max(k_0 - i, 0)$, where k_0 is sufficiently large. Since the precision of labels generated by the SRL is lower than the precision of labels obtained from filtered projection, the precision of the training data is expected to decrease with the increase in recall. Therefore, starting with a high k seeks to ensure high precision labels are added to the training data in the first iterations. Decreasing k in each iteration seeks to ensure that resampling is done in an increasingly restrictive way to ensure that only high-quality annotated sentences are added to the training data, thus maintaining a high confidence in the learned SRL model.

3.3 Effectiveness of Bootstrapping

We experimentally evaluate the effectiveness of our model with $k_0 = 9$.[7] As shown in Tab 4, bootstrapping outperforms relabeling, producing labels with best overall quality in terms of F_1 measure and recall for both predicates and arguments, with a relatively small cost in precision.

While Algorithm 1 guarantees the increase of recall (Prop. 1), it provides no such guarantee on precision. Therefore, it is important to experimentally decide an early termination cutoff before the SRL gets overtrained. To do so, we evaluated the performance of the bootstrapping algorithm at each iteration (Fig. 3). We observe that for the first 3 iterations, F_1-measure for both predicates and arguments rises due to large increase in recall which offsets the smaller drop in precision. Then F_1-measure remains stable, with recall rising and pre-

[7] We found that setting k_0 to larger values had little impact on the final results .

Language	Dep. Parser	Data Set	#Sentence
Arabic	STANFORD	`UN`	481K
Chinese	MATE-G	`UN`	2,986K
French	MATE-T	`UN`	2,542K
German	MATE-T	`Europarl`	560K
Hindi	MALT	`Hindencorp`	54K
Russian	MALT	`UN`	2,638K
Spanish	MATE-G	`UN`	2,304K

Table 5: Experimental setup .
Dependency parsers: STANFORD: (Green and Manning, 2010), MATE-G: (Bohnet, 2010), MATE-T: (Bohnet and Nivre, 2012), MALT: (Nivre et al., 2006). **Parallel corpora**: `UN`: (Rafalovitch et al., 2009), `Europarl`: (Koehn, 2005), `Hindencorp`: (Bojar et al., 2014). **Word alignment**: The `UN` corpus is already word-aligned. For others, we use the Berkeley Aligner (DeNero and Liang, 2007).

cision falling slightly at each iteration until convergence. To optimize precision and avoid overtraining, we set an iteration cutoff of 3. This combination of TF+RH filters, bootstrapping with $k_0 = 9$ and an iteration cutoff of 3 is used in the rest of our evaluation (Sec. 4), denoted as $\texttt{FB}_{best}$.

4 Multilingual Experiments

We use our method to generate Proposition Banks for 7 languages and evaluate the generated resources. We seek to answer the following questions: (1) What is the estimated quality for the generated PropBanks? How well does the approach work without language-specific adaptation? (2) Are there notable differences in quality from language to language; if so, why? We also present initial investigations on how different factors affect the performance of our method.

4.1 Experimental Setup

Data Tab. 5 lists the 7 different TLs and resources used in our experiments.[8] We chose these TLs because (1) they are among top 10 most influential languages in the world (Weber, 1997); and (2) we could find language experts to evaluate the results. English is used as SL in all our experiments.

Approach Tested For each TL, we used $\texttt{FB}_{best}$ (Sec. 3.3) to generate a corpus with semantic labels. From each TL corpus, we extracted all complete sentences to form the generated PropBanks.

[8]From each parallel corpus, we only keep sentences that are considered well-formed based on a set of standard heuristics. For example, we require a well-formed sentence to end in punctuation and not to contain certain special characters. For Arabic, as the dependency parser we use has relatively poor parsing accuracy, we additionally require sentences to be shorter than 100 characters.

		Predicate			Argument				
Lang.	Match	P	R	F1	P	R	F1	Agr	κ
Arabic	part.	0.97	0.89	0.93	0.86	0.69	0.77	0.92	0.87
	exact	0.97	0.89	0.93	0.67	0.63	0.65	0.85	0.77
Chinese	part.	0.97	0.88	0.92	0.93	0.83	0.88	0.95	0.91
	exact	0.97	0.88	0.92	0.83	0.81	0.82	0.92	0.86
French	part.	0.95	0.92	0.94	0.92	0.76	0.83	0.97	0.95
	exact	0.95	0.92	0.94	0.86	0.74	0.8	0.95	0.91
German	part.	0.96	0.92	0.94	0.95	0.73	0.83	0.95	0.91
	exact	0.96	0.92	0.94	0.91	0.73	0.81	0.92	0.86
Hindi	part.	0.91	0.68	0.78	0.93	0.66	0.77	0.94	0.88
	exact	0.91	0.68	0.78	0.58	0.54	0.56	0.81	0.69
Russian	part.	0.96	0.94	0.95	0.91	0.68	0.78	0.97	0.94
	exact	0.96	0.94	0.95	0.79	0.65	0.72	0.93	0.89
Spanish	part.	0.96	0.93	0.95	0.85	0.74	0.79	0.91	0.85
	exact	0.96	0.93	0.95	0.75	0.72	0.74	0.85	0.77

Table 6: Estimated precision and recall over seven languages.

Manual Evaluation While a gold annotated corpus for French ($\texttt{French}_{\text{gold}}$) was available for our experiments in the previous Sections, no such resources existed for the other TLs we wished to evaluate. We therefore chose to conduct a manual evaluation for each TL, each executed identically: For each TL we randomly selected 100 complete sentences with their generated semantic labels and assigned them to two language experts who were instructed to evaluate the semantic labels (based on their English descriptions) for the predicates and their core arguments. For each label, they were asked to determine (1) whether the label is correct; (2) if yes, then whether the boundary of the labeled constituent is correct: If also yes, mark the label as *fully correct*, otherwise as *partially correct*.

Metrics We used the standard measures of precision, recall, and F1 to measure the performance of the SRLs, with the following two schemes: (1) *Exact*: Only fully correct labels are considered as true positives; (2) *Partial*: Both fully and partially correct matches are considered as true positives.[9]

4.2 Experimental Results

Tab. 6 summarizes the estimated quality of semantic labels generated by our method for all seven TL. As can be seen, our method performed well for all

[9]Note that since the manually evaluated semantic labels are only a small fraction of the labels generated, the performance numbers obtained from manual evaluation is only an estimate of the actual quality for the generated resources.Thus the numbers obtained based on manual evaluation cannot be directly compared against the numbers computed over $\texttt{French}_{\text{gold}}$.

PROPBANK	#COMPLETE	%COMPLETE	#VERBS
Arabic	68.512	14%	330
Chinese	419,140	14%	1,102
French	248.256	10%	1145
German	44.007	8%	537
Hindi	1.623	3%	59
Russian	496.033	19%	1.349
Spanish	165.582	7%	909

Table 7: Characteristics of the generated PropBanks.

seven languages and generated high quality semantics labels across the board. For predicate labels, the precision is over 95% and the recall is over 85% for all languages except for Hindi. For argument labels, when considering partially correct matches, the precision is at least 85% (above 90% for most languages) and the recall is between 66% to 83% for all the languages. These encouraging results obtained from a diverse set of languages implies the generalizability of our method. In addition, the inter-annotator agreement is very high for all the languages, indicating that the results obtained based on manual evaluation are very reliable.

In addition, we make a number of interesting observations:

Dependency Parsing Accuracy The precision for exact argument labels is significantly below partial matches, particularly for Hindi (↓35 pp) and Arabic (↓19 pp). Since argument boundaries are determined syntactically, such errors are caused by dependency parsing. The fact that Hindi and Arbic suffer the most from this issue is consistent with the poorer performance of their dependency parsers compared to other languages (Nivre et al., 2006; Green and Manning, 2010).

Hindi as the Main Outlier The results for Hindi are much worse than the results for other languages. Besides the poorer dependency parser performance, the size of the parallel corpus used could be a factor: `Hindencorp` is one to two orders of magnitude smaller than the other corpora. The quality of the parallel corpus could be a reason as well: `Hindencorp` was collected from various sources, while both `UN` and `Europarl` were extracted from governmental proceedings.

Language-specific Errors Certain errors occur more frequently in some languages than others. An example are deverbal nouns in Chinese (Xue, 2006) in formal passive constructions with support verb 受. Since we currently only consider verbs for predicate labels, predicate labels are projected onto the support verbs instead of the deverbal nouns. Such errors appear for light verb constructions in all languages, but particularly affect Chinese due to the high frequency of this passive construction in the `UN` corpus.

	PREDICATE			ARGUMENT		
SAMPLE SIZE	P	R	F1	P	R	F1
100%	0.87	0.81	0.84	0.86	0.74	0.8
10%	0.88	0.8	0.84	0.87	0.72	0.79
1%	0.9	**0.76**	0.83	0.89	**0.67**	0.76

Table 8: Estimated impact of downsampling parallel corpus.

	PREDICATE			ARGUMENT		
HEURISTIC	P	R	F1	P	R	F1
none*	0.87	0.81	0.84	0.86	0.74	0.8
none**	0.88	0.8	0.84	**0.76**	**0.65**	0.7
customization*	0.87	0.81	0.84	**0.9**	0.74	0.81

Table 9: Impact of English SRLs (*=CLEARNLP, **=MATE-SRL) and language-spec. customization (*filter synt. expletive*).

Low Fraction of Complete Sentences As Tab. 7 shows, the fraction of complete sentences in the generated PropBanks is rather low, indicating the impact of moderate recall on the size of generated PropBanks. Especially for languages for which only small parallel corpora are available, such as Hindi, this points to the need to address recall issues in future work.

4.3 Additional Experiments

The observations made in Sec. 4.2 suggests a few factors that may potentially affect the performance of our method. To better understand their impact, we conducted the following initial investigation. SRL models produced in this set of experiments were evaluated using $\texttt{French}_{\text{gold}}$, sampled and evaluated in the same way as other experiments in this section for comparability.

Data Size We varied the data size for French by downsampling the `UN` corpus. As one can see from Tab. 8, downsampling the dataset by one order of magnitude (to 250k sentences) only slightly affects precision, while downsampling to 25k sentences has a more pronounced but still small impact on recall. It appears that data size does not have significant impact on the performance of our method.

Language-specific Customizations While our method is language-agnostic, intuitively language-specific customization can be helpful in addressing language-specific errors. As an initial experiment, we added a simple heuristic to filter out French verbs that are commonly used for "existential there" constructions, as one type of common errors for French involves the syntactic expletive *il* (Danlos, 2005) in "existential there" constructions such as *il faut* (see Fig. 1 (TL sentence) for an example) wrongly labeled with with role information. As shown in Tab. 9, this simple customization results in a small increase in precision, suggesting that language-specific customization can be helpful.

Quality of English SRL As noted in Sec. 2.5, errors made by English SRL are often prorogated to the TL via projection. To assess the impact of English SRL quality, we used two different English SRL systems: CLEARNLP and MATE-SRL. As can be seen from Tab. 9, the impact of English SRL quality is substantial on argument labeling.

4.4 Multilingual PropBanks

To facilitate future research on multilingual SRL, we release the created PropBanks for all 7 languages to the research community to encourage further research. Tab. 7 gives an overview over the resources.

5 Related Work

Annotation Projection in Parallel Corpora to train monolingual tools for new languages was introduced in the context of learning a PoS tagger (Yarowsky et al., 2001). Similar in spirit to our approach of using filters to increase the precision of projected labels, recent work (Täckström et al., 2013) uses token and type constraints to guide learning in cross-lingual PoS tagging.

Projection of Semantic Labels was considered for FrameNet (Baker et al., 1998) in (Padó and Lapata, 2009; Basili et al., 2009). Recently, however, most work in the area focuses on PropBank, which has been identified as a more suitable annotation scheme for joint syntactic-semantics settings due to broader coverage (Merlo and van der Plas, 2009), and was shown to be usable for languages other than English (Monachesi et al., 2007).

Direct projection of PropBank annotations was considered in (Van der Plas et al., 2011). Our approach significantly outperforms theirs in terms of recall and F_1 for both predicates and arguments (Section 3). A approach was proposed in (Van der Plas et al., 2014) in which information is aggregated at the corpus level, resulting in a significantly better SRL corpus for French. However, this approach has several practical limitations: (1) it does not consider the problem of argument identification of SRL systems, treating arguments as already given; (2) it generates rules for the argument classification step preferably from manually annotated data; (3) it has been demonstrated for a single language (French), and was not applied to any other language. In contrast, our approach trains an SRL system for both predicate and argument labels, in a completely automatic fashion. Furthermore, we have applied our approach to generate PropBanks for 7 languages and conducted experiments that indicate a high F_1 measure for all languages (Section 4).

Other Related Work A number of approaches such as model transfer (Kozhevnikov and Titov, 2013) and role induction (Titov and Klementiev, 2012) exist for the argument classification step in the SRL pipeline. In contrast, our work addresses the full SRL pipeline and seeks to generate SRL resources for TLs with English PropBank labels.

6 Conclusion

We proposed a two-staged method to construct multilingual SRL resources using monolingual SRL and parallel data and showed that our method outperforms previous approaches in both precision and recall. More importantly, through comprehensive experiments over seven languages from three language families, we show that our proposed method works well across different languages without any language specific customization. Preliminary results from additional experiments indicate that better English SRL and language-specific customization can further improve the results, which we aim to investigate in future work. A qualitative comparison against existing or under-construction PropBanks for Chinese (Xue, 2008), Hindi (Vaidya et al., 2011) or Arabic (Zaghouani et al., 2010) may be interesting, both for comparison of resources and for defining language-specific customizations. In addition, we plan to expand our experiments both to more languages as well as NomBank (Meyers et al., 2004)-style noun labels.

References

[Baker et al.1998] Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

[Basili et al.2009] Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *Computational Linguistics and Intelligent Text Processing*, pages 332–345. Springer.

[Björkelund et al.2009] Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

[Bohnet and Nivre2012] Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465. Association for Computational Linguistics.

[Bohnet2010] Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

[Bojar et al.2014] Ondřej Bojar, Vojtěch Diatka, Pavel Rychlỳ, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, Daniel Zeman, et al. 2014. Hindencorp–hindi-english and hindi-only corpus for machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

[Choi and McCallum2013] Jinho D. Choi and Andrew McCallum. 2013. Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

[Danlos2005] Laurence Danlos. 2005. Automatic recognition of french expletive pronoun occurrences. In *Natural language processing. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 73–78. Citeseer.

[DeNero and Liang2007] John DeNero and Percy Liang. 2007. The Berkeley Aligner. `http://code.google.com/p/berkeleyaligner/`.

[Erk et al.2003] K. Erk, A. Kowalski, S. Pado, and S. Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *ACL*.

[Green and Manning2010] Spence Green and Christopher D Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 394–402. Association for Computational Linguistics.

[Hajič et al.2009] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

[Koehn2005] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

[Kozhevnikov and Titov2013] Mikhail Kozhevnikov and Ivan Titov. 2013. Cross-lingual transfer of semantic role labeling models. In *ACL (1)*, pages 1190–1200.

[Maqsud et al.2014] Umar Maqsud, Sebastian Arnold, Michael Hülfenhaus, and Alan Akbik. 2014. Nerdle: Topic-specific question answering using wikia seeds. In Lamia Tounsi and Rafal Rak, editors, *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 81–85. ACL.

[Merlo and van der Plas2009] Paola Merlo and Lonneke van der Plas. 2009. Abstraction and generalisation in semantic role labels: Propbank, verbnet or both? In *ACL 2009*, pages 288–296.

[Meyers et al.2004] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for nombank. In *LREC*, volume 4, pages 803–806.

[Monachesi et al.2007] Paola Monachesi, Gerwert Stevens, and Jantine Trapman. 2007. Adding semantic role annotation to a corpus of written dutch. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 77–84.

[Nivre et al.2006] Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

[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.

[Pado2007] Sebastian Pado. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University. MP.

[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.

[Rafalovitch et al.2009] Alexandre Rafalovitch, Robert Dale, et al. 2009. United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit*, volume 12, pages 292–299.

[Schuler2005] Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

[Shen and Lapata2007] Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21. Citeseer.

[Täckström et al.2013] Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

[Titov and Klementiev2012] Ivan Titov and Alexandre Klementiev. 2012. Crosslingual induction of semantic roles. In *ACL*, pages 647–656.

[Vaidya et al.2011] Ashwini Vaidya, Jinho D Choi, Martha Palmer, and Bhuvana Narasimhan. 2011. Analysis of the hindi proposition bank using dependency structure. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 21–29. Association for Computational Linguistics.

[Van der Plas et al.2010] Lonneke Van der Plas, Tanja Samardžić, and Paola Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.

[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.

[Van der Plas et al.2014] Lonneke Van der Plas, Marianna Apidianaki, Rue John von Neumann, and Chenhua Chen. 2014. Global methods for cross-lingual semantic role and predicate labelling. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1279–1290. Association for Computational Linguistics.

[Weber1997] George Weber. 1997. Top languages: The world's 10 most influential languages. *Language Today*, December.

[Xue2006] Nianwen Xue. 2006. Semantic role labeling of nominalized predicates in chinese. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 431–438. Association for Computational Linguistics.

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

[Yarowsky et al.2001] David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

[Zaghouani et al.2010] Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised arabic propbank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226. Association for Computational Linguistics.