# Cross-Lingual Bootstrapping of Semantic Lexicons: The Case of FrameNet

# Cross-lingual Bootstrapping of Semantic Lexicons: The Case of FrameNet

**Sebastian Padó**
Computational Linguistics, Saarland University
P.O. Box 15 11 50, 66041 Saarbrücken, Germany
pado@coli.uni-sb.de

**Mirella Lapata**
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
mlap@inf.ed.ac.uk

**Abstract**

This paper considers the problem of unsupervised semantic lexicon acquisition. We introduce a fully automatic approach which exploits parallel corpora, relies on shallow text properties, and is relatively inexpensive. Given the English FrameNet lexicon, our method exploits word alignments to generate frame candidate lists for new languages, which are subsequently pruned automatically using a small set of linguistically motivated filters. Evaluation shows that our approach can produce high-precision multilingual FrameNet lexicons without recourse to bilingual dictionaries or deep syntactic and semantic analysis.

## Introduction

Shallow semantic parsing, the task of automatically identifying the semantic roles conveyed by sentential constituents, is an important step towards text understanding and can ultimately benefit many natural language processing applications ranging from information extraction (Surdeanu *et al.* 2003) to question answering (Narayanan & Harabagiu 2004) and machine translation (Boas 2002).

Recent advances in semantic parsing have greatly benefited from the availability of resources that document the surface realisation of semantic roles, thus providing richly annotated data for training statistical semantic role labellers (Gildea & Jurafsky 2002). The FrameNet lexical resource (Baker, Fillmore, & Lowe 1998) has played a central role in this enterprise. In FrameNet, meaning is represented by frames, i.e., schematic representations of situations. Semantic roles are frame-specific, and are called frame elements. The database associates frames with lemmas (verbs, nouns, adjectives) that can evoke them (called frame-evoking elements or FEEs), lists the possible syntactic realisations of their semantic roles, and provides annotated examples from the British National Corpus (Burnard 1995).

Table 1 exemplifies the COMMITMENT frame. The frame has four roles, a SPEAKER who commits him/herself to do something, an ADDRESSEE to whom the commitment is made, a MESSAGE or a TOPIC expressing the commitment, and a MEDIUM used to transmit the MESSAGE. The frame elements are realised by different syntactic expressions. For instance, the SPEAKER is typically an NP, whereas the MESSAGE is often expressed as a clausal complement (see the expressions in boldface in Table 1). The COMMITMENT frame can be evoked by *promise*, *consent*, *pledge*, and several other verbs as well as nouns (see the list of FEEs in Table 1).

| Frame: COMMITMENT | | |
|---|---|---|
| Frame Elements | SPEAKER | **Kim** promised to be on time. |
| | ADDRESSEE | Kim promised **Pat** to be on time. |
| | MESSAGE | Kim promised Pat **to be on time.** |
| | TOPIC | The government broke its promise **about taxes.** |
| | MEDIUM | Kim promised **in writing** to sell Pat the house. |
| FEEs | consent.v, covenant.n, covenant.v, oath.n, vow.n, pledge.v, promise.n, promise.v, swear.v, threat.n, threaten.v, undertake.v, undertaking.n, volunteer.v | |

Table 1: Example frame from the FrameNet database

An important first step towards semantic parsing is the creation of a semantic lexicon listing appropriate frames for lemmas together with their corresponding semantic roles. The English FrameNet currently contains 513 frames covering 7125 lexical items and has been under development for approximately eight years; German, Spanish, and Japanese FrameNets, though in their infancy, are also under construction. Since manual lexicon construction is costly, time consuming, and rarely ever complete, automatic methods for acquiring FrameNets for new languages could significantly speed up the lexicon development process. In this paper, we propose an automatic method which employs parallel corpora for the acquisition task and holds promise for reducing the manual effort involved in creating semantic lexicons.

Our method leverages the existing English FrameNet to overcome the resource shortage in other languages by exploiting the translational equivalences present in word-aligned data. We focus on inducing a frame lexicon, i.e., a list of lemmas for each frame, under the assumption that FrameNet descriptions for English port reliably to other languages (Erk *et al.* 2003). The lexicon induction task is a prerequisite to learning how a frame's semantic roles are syntactically realised. Our method can yield a FrameNet lexi-

con for any language for which English parallel corpora are available. To show its potential, we investigate the induction of FrameNet lexicons for German and French.

In the following section, we provide an overview of related work on automatic construction of multilingual resources. Then, we introduce our FrameNet-based lexicon induction algorithms. Next, we present our experimental framework and data. Evaluation results and their discussion conclude the paper.

## Related Work

The automatic construction of bilingual lexicons has a long history in machine translation (see Melamed (1996) and the references therein). Word-level translational equivalences can be extracted from parallel texts without recourse to corpus external resources, simply by using word-alignment methods (Och & Ney 2003) and filtering techniques aimed at discarding erroneous translations (Melamed 1996). Mann & Yarowsky (2001) induce bilingual lexicons using solely on-line dictionaries and cognate word pairs.

Translational correspondences are also exploited for inducing multilingual text analysis tools (Yarowsky, Ngai, & Wicentowski 2001; Hwa *et al.* 2002). English analyses (e.g., syntax trees, POS tags) are projected onto another language through word-aligned parallel text. The projected annotations can be then used to derive monolingual tools without additional annotation cost.

Fung & Chen (2004) automatically construct an English-Chinese bilingual FrameNet by mapping automatically FrameNet entries to concepts listed in HowNet[1], an on-line ontology for Chinese, without access to parallel texts.

The present work extends previous approaches on automatic lexicon construction and annotation projection by inducing a FrameNet compliant lexicon rather than a generic word-to-word translation mapping. Our work exploits word alignment as a proxy for translational equivalence while making use of the already available English FrameNet. Our lexicon construction process is unsupervised and portable across languages and domains. The proposed method does not make use of bilingual dictionaries. Such dictionaries may be unavailable for low density languages or indeed inappropriate for lexicon induction purposes. Despite their simplicity, dictionary lookup strategies often suffer from problems caused by translational ambiguity; there are often one-to-many correspondences in a bilingual dictionary. Furthermore, translational equivalence is rarely quantified in manually constructed resources, whereas this information is naturally expressed by word alignment probabilities.

## FrameNet Lexicon Induction

We induce FrameNet lexicons for new languages by adopting a generate-and-prune approach. In a first step, we obtain word alignments $s_i \rightsquigarrow_1 t_j$ between a source and a target language from a parallel corpus. The alignments constitute a repository of token-level *translation pairs* for frames, from which we generate a large type-level *FEE candidate list* for the target language. This initial list is somewhat noisy and contains many spurious translations. In a second step, we prune spurious translations using *incremental filters*, thus acquiring FrameNet-compliant entries. Filters can apply both at the token and type level, placing constraints either on admissible translation pairs or on FEE *candidate probability*.

[1]See `http://www.keenage.com/zhiwang/e_zhiwang.html`.

| | |
|---|---|
| $s, t$ | word types in source and target language |
| $s_i$, $t_j$ | word tokens of $s$ and $t$ |
| $f$ | frame in English FrameNet |
| $p_c(f\|t)$ | candidate probability of frame $f$ for type $t$ |
| $s_i \rightsquigarrow_1 t_j$ | true iff $s_i$ and $t_j$ are one-to-one aligned |
| $tp(s_i, t_j, f)$ | true iff $s_i$ and $t_j$ form a translation pair for $f$ |
| $fee(f, s)$ | true iff $s$ is a listed FEE for $f$ |
| $wsd(f, s_i)$ | true iff $f$ is the disambiguated frame for $s_i$ |
| $cont(s_i)$ | true iff $s_i$ is a content word (Adj, N, or V) |
| $cand(t, f)$ | true iff $t$ is a FEE candidate for $f$ |

Table 2: Notation overview

### FEE Candidate List Generation

For the reasons outlined in the previous paragraph, we use automatic methods for finding translation equivalences that allow for rapid lexicon acquisition across languages and domains. We obtain an initial FEE candidate list from a parallel corpus using the statistical IBM word alignment models (Och & Ney 2003). Other alignment techniques could also serve as a basis for generating FEE candidates; our method is not specific to the IBM models.

To produce an initial FEE candidate list, we employ a rather liberal definition of translation pair, covering all correspondences we gather from a parallel corpus; a source and target language token form a translation pair for frame $f$ if they are aligned and the source token is a FEE of $f$ (see Table 2 for an overview of the notation used in this paper).

$$tp(s_i, t_j, f) \equiv s_i \rightsquigarrow_1 t_j \wedge fee(f, s) \tag{1}$$

The probability that a target type $t$ is a candidate FEE of frame $f$ is defined as the number of translation pairs evoking $f$ over all translation pairs attested for $t$:

$$p_c(f|t) = \frac{\sum_s |\{tp(s_i, t_j, f)\}|}{\sum_f \sum_s |\{tp(s_i, t_j, f)\}|} \tag{2}$$

Without imposing any filters on the alignment output, we obtain an initial FEE candidate list for frame $f$ by accepting all candidates with non-zero probability (i.e., candidates with at least one translation pair for $f$):

$$cand(t, f) \equiv p_c(f|t) > 0 \tag{3}$$

### Candidate Filtering

We have developed three classes of filters to eliminate spurious candidate translations. These arise from shortcomings in the automatic alignment model which is inexact and noisy, the non-treatment of linguistic phenomena underlying translation mismatches (e.g., frame polysemy, multi-word expressions), and their interactions. We propose filters that specifically address different types of translation errors and make use of linguistic knowledge as well as knowledge-poor

| Example | Frame | German |
|---|---|---|
| He asks about her health. | QUESTIONING | (be)fragen |
| He asks for the jam. | REQUEST | bitten |
| He asks himself if it is true. | COGITATION | sich fragen |

Table 3: Example of frame polysemy for the verb *ask*

filters that generally measure alignment quality. The latter produce a ranking over translation candidates, thus allowing us to specify thresholds that directly influence the precision (and recall) of the induced lexicon.

**Alignment filters.** Alignment errors often arise from indirect correspondences (Melamed 1996), i.e., frequently co-occurring, but not semantically equivalent bilingual word pairs. Examples include light verb constructions (e.g., *take* $\rightsquigarrow_1$ *entscheiden* "decide" because of the frequent construction *take a decision*), but also collocations proper (e.g., *increase* $\rightsquigarrow_1$ *plötzlich* "suddenly" because of the frequent expression *plötzlicher Anstieg* "sudden increase").

German noun compounds (e.g., *Reisekosten* "travel expenses") and non-compositional multi-word expressions such as idioms (e.g., *kick the bucket*) are another source of erroneous translations. Multi-word expressions are primary examples of lexical translational divergence where words can correspond to phrases and vice versa. Such expressions are hard to treat adequately for most alignment models which typically produce word-to-word translations.

We have developed two filters that target alignment errors at the token level. The first filter imposes constraints on the grammatical categories of translation pairs. The filter simply discards translation pairs whose target token is not a content word, i.e., a verb, noun, or adjective:

$$tp(s_i, t_j, f) \equiv s_i \rightsquigarrow_1 t_j \wedge cont(t_j) \tag{4}$$

Since the English FrameNet covers almost exclusively these three parts of speech, we assume that this coverage is sufficient at least for related target languages. The constraint effectively discards a number of alignment errors such the case of *increase* $\rightsquigarrow_1$ *plötzlich* mentioned above.

The second filter does not take any grammatical information into account; instead, it relies on bidirectional alignment to obtain high-precision alignment points. Following Koehn, Och, & Marcu (2003), we align a parallel corpus bidirectionally – foreign to English and English to foreign. This way, we obtain two word alignments which we consequently intersect. Bidirectional filtering exploits the fact that only reliable one-to-one correspondences will occur in both word alignments, resulting in cleaner (though sparser) alignments:

$$tp(s_i, t_j, f) \equiv s_i \rightsquigarrow_1 t_j \wedge t_j \rightsquigarrow_1 s_i \tag{5}$$

**Frame polysemy filters.** Source language words can usually evoke more than one frame. This is illustrated in Table 3 where the verb *ask* evokes three frames, QUESTIONING, REQUEST, and COGITATION, each realised by a different verb in the target language. Without any explicit handling of frame polysemy in the source language, all three German translations for *ask* will be listed for each frame.

We propose two filters to address this problem. The first one applies at the token level and uses frame disambiguation to eliminate erroneous frame assignments. More specifically, we employ a frame disambiguator in the source language which labels each token by the single most appropriate frame given its context; consequently, translation pairs are formed for disambiguated frames:

$$tp(s_i, t_j, f) \equiv s_i \rightsquigarrow_1 t_j \wedge wsd(f, s_i) \tag{6}$$

Similarly to the "most frequent sense" baseline commonly employed in word sense disambiguation (WSD), our second filter forces frame disambiguation at the type level by defaulting to the frame with the highest candidate probability for a given target word:

$$cand(t, f) \equiv p_c(f|t) = \max_{f'} p_c(f'|t) \tag{7}$$

This is clearly an approximation, since target language words can also be polysemous; however, our results indicate that here, as in WSD, considerable mileage can be gained by defaulting to the most frequent frame per target FEE.

**Information-theoretic filters.** Melamed (1997) proposes a measure of translational (in-)consistency based on information theory. Translational inconsistency is defined for every target word as the entropy of the probability distribution $p(s|t)$ of the aligned source words, given the target word. The measure captures the intuition that consistent translations will lead to a concentration of the probability mass on a few source words, resulting in low entropy. For our purposes, we assume that reliable translations are better FEE candidates, and we apply the filter by accepting all FEE candidates whose entropy is below a certain threshold $n$:

$$cand(t, f) \equiv -\sum_{s} [p(s|t) \log p(s|t)] < n \tag{8}$$

Translational inconsistency is a general-purpose filter and as such does not take the structure of FrameNet into account. Ideally, we would like to group the source words by frames so that within-frame variance (e.g., *versprechen* translated as either *vow* or *promise*) does not increase, but across-frame variance (e.g., *versprechen* translated as *vow* or *often*) does. In the absence of a semantic lexicon which covers all English words, we approximate frame-based entropy by collapsing all FEEs of the current frame into one "meta-type" (e.g., for COMMITMENT, all FEEs in Table 1 are treated as the same word). Again, we retain FEE candidates if the entropy lies below a threshold $n$:

$$\begin{aligned} cand(t, f) \equiv & -\sum_{fee(f,s)} p(s|t) \log \sum_{fee(f,s)} p(s|t) - \\ & \sum_{\neg fee(f,s)} [p(s|t) \log p(s|t)] < n \end{aligned} \tag{9}$$

## Evaluation Set-up

**Data.** We evaluated our bilingual lexicon induction framework on FrameNet release 1.1 for two language pairs, namely English-German and English-French. We used the Europarl corpus (Koehn 2002), a corpus of professionally translated proceedings of the European Parliament between 1997 and 2003. The corpus is available in 11 languages with approximately 20 million words per language and is aligned at the document and sentence level.

| DE Band | TP | FNr | AvgC |
|---|---|---|---|
| High | $>7836$ | 159 | 199.7 |
| Medium | $<7836$ | 159 | 64.7 |
| Low | $<959$ | 158 | 12.6 |
| **FR Band** | **TP** | **FNr** | **AvgC** |
| High | $>9528$ | 163 | 203.0 |
| Medium | $<9528$ | 163 | 56.6 |
| Low | $<1277$ | 162 | 11.3 |

Table 4: Frame frequency bands

| High (828 DE / 799 FR) |
|---|
| PREVENTING, COMMUNICATION_RESPONSE, GIVING, DECIDING, CAUSE_CHANGE_OF_SCALAR_POSITION |
| Medium (366 DE / 347 FR) |
| EVALUATIVE_COMPARISON, TRAVEL, EMPLOYING, SENSATION, JUDGMENT_COMMUNICATION |
| Low (84 DE / 68 FR) |
| ADDING_UP, CONGREGATING, ESCAPING, SUSPICIOUSNESS, RECOVERY |

Table 5: Frames selected for evaluation

We used the default setting[2] of GIZA++ (Och & Ney 2003), a publicly available implementation of the IBM models and HMM word alignment models, to induce word alignments for both directions (source-target and target-source). We considered only one-to-one alignments, since we found one-to-many alignments to be unreliable. Moreover, we retrieved translation candidates from token (Viterbi) alignments rather than from the translation table; token alignments are a richer source of information regarding the model's decisions and do not involve setting arbitrary thresholds to separate genuine alignments from spurious ones. We removed all target words for which we observed less than five translation pairs in order to eliminate sparse and therefore unreliable alignments.

In further preprocessing steps, we applied the Stuttgart TreeTagger (Schmid 1994) to lemmatise and part-of-speech tag the English, German, and French data. We employed Erk's (2005) frame disambiguation system for labelling English tokens with their corresponding frames. This is the only available frame disambiguation system we are aware of; it employs a Naive Bayes classifier trained on the example sentences available in the English FrameNet database and yields an overall F-score of 74.7%.

**Gold standard.** In order to investigate how our approach performs across a range of frames with varying frequencies, frames were split into three bands based on an equal division of the number of translation pairs per frame in the corpus. The resulting bands for German (DE) and French (FR) are shown in Table 4 together with the number of translation pairs (TP), frames (FNr), and the average number of candidates per frame (AvgC) in each band. We randomly selected five frames from each band, ensuring that frames fell into the same bands in French and German.

The selected frames are shown in Table 5, together with the number of unfiltered FEE candidates for each band. All candidates (a total of 1278 for German and 1214 for French) were annotated by two annotators. The annotators had to classify FEEs as correct or spurious. Spurious FEEs were further classified as being the result of a frame polysemy or alignment error. The annotators were given detailed guidelines which operationalised the decisions and discussed borderline cases. The annotators had access to the list of FEE candidates and their English supports as well as to a concordance. Inter-annotator agreement was 85% ($\kappa = 0.79$) for German and 84% ($\kappa = 0.78$) for French. Disagreements were discussed and a gold standard was created by consensus. The annotators were linguistics graduate students and familiar with FrameNet; annotation speed was between 100 and 200 instances per hour.

[2]The training scheme involved five iterations of Model 1, five iterations of the HMM model, five iterations of Model 3, and five iterations of Model 4.

## Results

In this section, we assess the impact of the proposed filters on the bilingual FrameNet induction task. Since all filters are specified as conjunctions of constraints, they can be easily combined. We also evaluate our results against an approach that utilises a bilingual dictionary instead of automatic alignment. Finally, we compare our automatically acquired dictionary against the SALSA lexicon (Erk *et al.* 2003), a manually constructed FrameNet for German.

**Effects of Filtering.** Table 6 shows the quality of the induced lexicon prior to and following filtering for German and French, respectively. We first concentrate on the influence of alignment and frame polysemy filters, individually and in combination. Since these filters define constraints on translation pairs, no threshold optimisation takes place. In addition to precision and recall, Table 6 gives a breakdown of the amount of false positive FEEs in terms of polysemy (P) and alignment (A) errors.

The unfiltered lists (NoF) expectedly have a recall of 1.0 for both languages; precision is 0.35 for German and 0.30 for French. As can be seen from Table 6, NoF yields an F-score of 0.52 for German; one third of the induced FEEs are true positives, whereas two thirds are false positives, due to polysemy and alignment errors. NoF yields an F-score of 0.46 for French. Analysis of the translation pairs reveals a higher number of idiomatic divergences – translations where no sensible word alignment is possible – between English and French than between English and German.

For both languages, bidirectional alignment (BD) reduces the amount of spurious alignments, but results in a relative increase of polysemy errors due to the elimination of some true positives; part-of-speech filtering (POS) eliminates a smaller number of noisy alignments, but also incurs less polysemy errors. The frame disambiguation filter (FrD) disappointingly does not reduce the number of polysemy errors. The filter which selects the most frequent frame for a given target word (MaxFr) reduces both polysemy and alignment errors, however at the expense of recall. This tendency is more pronounced in the French data.

The most informative filter combinations are also shown

| Model (DE) | Rec. | Prec. | Fscore | %P | %A |
|---|---|---|---|---|---|
| NoF | 1.00 | 0.35 | 0.52 | 30 | 33 |
| BD | 0.70 | 0.47 | 0.56 | 37 | 15 |
| POS | 0.98 | 0.40 | 0.57 | 33 | 25 |
| FrD | 0.80 | 0.35 | 0.48 | 30 | 33 |
| MaxFr | 0.49 | 0.57 | 0.53 | 22 | 20 |
| POS FrD | 0.78 | 0.40 | 0.52 | 33 | 26 |
| **BD POS** | **0.68** | **0.50** | **0.58** | **38** | **11** |
| BD POS FrD | 0.54 | 0.50 | 0.52 | 38 | 10 |
| BD POS MaxFr | 0.36 | 0.68 | 0.47 | 24 | 6 |

| Model (FR) | Rec. | Prec. | Fscore | %P | %A |
|---|---|---|---|---|---|
| NoF | 1.00 | 0.30 | 0.46 | 31 | 38 |
| BD | 0.80 | 0.36 | 0.50 | 36 | 26 |
| POS | 0.99 | 0.33 | 0.49 | 34 | 34 |
| FrD | 0.79 | 0.29 | 0.43 | 32 | 38 |
| MaxFr | 0.31 | 0.60 | 0.41 | 25 | 14 |
| POS FrD | 0.78 | 0.32 | 0.46 | 34 | 32 |
| **BD POS** | **0.80** | **0.38** | **0.51** | **37** | **23** |
| BD POS FrD | 0.64 | 0.37 | 0.47 | 37 | 24 |
| BD POS MaxFr | 0.28 | 0.65 | 0.39 | 23 | 10 |

Table 6: Evaluation of alignment and polysemy filters of German (above) and French (below)

in Table 6. The highest F-score is obtained when bidirectional filtering is combined with part-of-speech tagging (BD POS). Again, when combining these two filters with MaxFr, precision increases (to 0.68 for German and 0.65 for French) while recall decreases significantly (to 0.36 for German and 0.28 for French).

Figure 1 shows the impact of the information-theoretic filters on the FrameNet induction task. Recall that these are threshold-dependent, i.e., their precision varies across different levels of recall. For both languages, entropy filtering (Entropy) can improve precision only marginally. Our FrameNet-specific entropy filter ($\text{Entropy}_{\text{mod}}$) systematically outperforms standard entropy at all recall levels; contrary to standard entropy, it consistently rewards lower recall with higher precision. Notice than when entropy is combined with bidirectional and part-of-speech filtering (BD POS Entropy/$\text{Entropy}_{\text{mod}}$) precision is improved, whereas unsurprisingly recall decreases. At a recall of 0.20, precision is approximately 0.70 for German and 0.50 for French.

Table 7 shows how the induced lexicon varies in size (average number of true positives per frame) across frequency bands (High, Medium, Low) before and after filtering. Expectedly, the largest number of FEEs is found in the High band. It is worth noticing that filters differ in their impact on different frequency bands: while MaxFr tends to remove high frequency FEEs, $\text{Entropy}_{\text{mod}}$ drastically reduces the number of low frequency FEEs.

**Comparison against a bilingual dictionary.** We next examined how our automatic methods compare against a simple dictionary lookup strategy. For this experiment we used LEO[3], a publicly available German-English bilingual

[3]See `http://dict.leo.org/`.

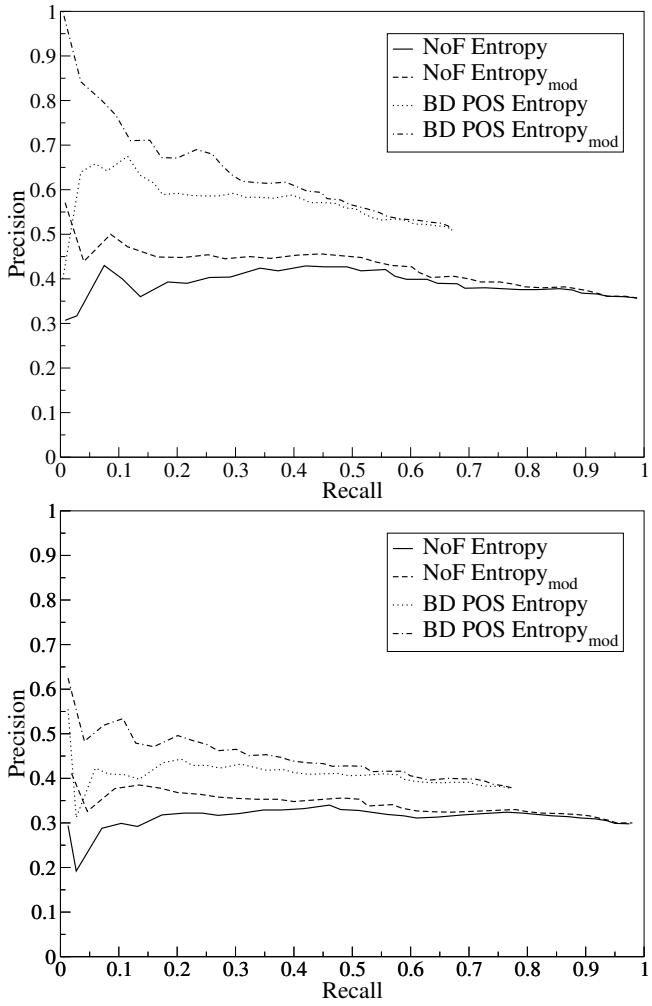


Figure 1: Precision of information-theoretic filters at varying recall levels for German (above) and French (below)

dictionary whose coverage exceeds 400,000 entries. We obtained a total of 984 FEE candidates for our 15 frames; these constitute 75% of the FEEs extracted from the parallel corpus. Of the FEEs obtained from the bilingual dictionary, 40% are true positives. Although there is hardly any noise (only 3% are alignment errors), a large number of spurious translations are due to polysemy (57%). These results indicate that even in cases where bilingual dictionaries are available, filtering is necessary for selecting appropriate FEEs.

**Comparison against SALSA.** Finally, we compared the output of our lexicon against SALSA (Erk *et al.* 2003), a manually constructed FrameNet for German. The lexicon's development is informed by the TIGER corpus, a 1.5M word corpus of newspaper text. We should expect a reasonable automatically created lexicon to exhibit some overlap with SALSA, even though the latter is still under construction. For the 15 frames used in our evaluation, SALSA contains 51 verbs; our automatically constructed lexicon (after applying BD POS filtering), contains 148 true positives. The two sets have 21 verbs in common, i.e., over 40% of the SALSA

| Band | High | | Medium | | Low | |
|---|---|---|---|---|---|---|
| | DE | FR | DE | FR | DE | FR |
| NoF | 51.2 | 37.0 | 32.0 | 29.6 | 6.8 | 5.8 |
| BD POS | 36.6 | 31.0 | 20.0 | 23.2 | 5.0 | 3.6 |
| MaxFr BD POS | 17.2 | 6.8 | 11.2 | 11.8 | 4.0 | 1.8 |
| BD POS Entropy$_{\mathrm{mod}}$ | 19.6 | 13.8 | 8.6 | 8.8 | 2.0 | 0.8 |

Table 7: Average number of true positives (FEEs) per band for different filters (Entropy$_{\mathrm{mod}}$: recall set to 0.33)

verb sample is covered by our lexicon.

## Discussion and Conclusions

In this paper we have presented a method for automatically inducing multilingual FrameNets from automatically aligned parallel corpora. We have assessed the approach on two languages, German and French, and have experimented with a variety of filtering techniques aimed at eliminating spurious translations. Combination of bidirectional alignment with part-of-speech filtering delivers the highest F-score for both languages. We have also shown that precision can be further increased using either entropy-based filtering or forcing frame disambiguation of FEE candidates.

The FEE candidate lists we acquire can be viewed as a high-precision *seed lexicon* which can be extended manually or automatically using monolingual resources or bootstrapping techniques (Riloff & Jones 1999). Note that preferring precision over recall does not necessarily produce a small lexicon. While the English FrameNet lists on average 13 FEEs per frame, our unfiltered candidate lists contain on average 30 (German) and 24 (French) true positives. A lexicon with a recall level of 30% would roughly correspond to a FrameNet-size lexicon. The filter combination BD POS MaxFr (see Table 6) yields approximately this recall at significant precision levels of 0.68 (German) and 0.65 (French).

Our filtering techniques successfully eliminated spurious alignments but had less impact on polysemy errors. These are due to subtle semantic distinctions across FrameNet frames, which require richer knowledge sources than our methods currently exploit. Although frame disambiguation seems promising, the current incompleteness of FrameNet makes frame disambiguation difficult, since the "correct" frame can be missing from the database or provide too little training data for the effective application of machine learning techniques. We have shown that a simple approach which defaults to the most frequent frame per FEE (analogously to the first-sense heuristic often employed in WSD) eliminates to some extent polysemy-related noise.

The approach taken in this paper relies on the availability of parallel texts and is relatively inexpensive. The proposed methods can be easily extended to other languages or domains. An important future direction lies in the development of a model that links the automatically induced FEEs with sentential constituents bearing frame-specific semantic roles (see the frame elements in Table 1). Further investigations on additional languages, text types, and genres will test the generality and portability of our approach. We will also examine whether the combination of automatic alignment with a bilingual dictionary will result in improved performance.


## Acknowledgements

The authors acknowledge the support of DFG (Padó; project SALSA-II) and EPSRC (Lapata; grant GR/T04540/01). Thanks to Ulrike Padó and Garance Paris for their annotation efforts and Katrin Erk and Katja Markert for useful discussions and comments on earlier versions of this paper.


## References


Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL 1998*, 86–90.

Boas, H. C. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of LREC 2002*, 1364–1371.

Erk, K.; Kowalski, A.; Pado, S.; and Pinkal, M. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of ACL 2003*, 537–544.

Erk, K. 2005. Frame assignment as word sense disambiguation. In *Proceedings of IWCS 2005*, 362–364.

Fung, P., and Chen, B. 2004. BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. In *Proceedings of COLING 2004*, 931–935.

Gildea, D., and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3):245–288.

Hwa, R.; Resnik, P.; Weinberg, A.; and Kolak, O. 2002. Evaluation of translational correspondance using annotation projection. In *Proceedings of ACL 2002*, 392–399.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL 2003*, 127–133.

Koehn, P. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished manuscript.

Mann, G., and Yarowsky, D. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL 2001*, 151–158.

Melamed, I. D. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of AMTA 1996*, 125–134.

Melamed, I. D. 1997. Measuring semantic entropy. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics at ANLP 1997*, 41–46.

Narayanan, S., and Harabagiu, S. 2004. Question answering based on semantic structures. In *Proceedings of COLING 2004*, 693–701.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–52.

Riloff, E., and Jones, R. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI 1999*, 1044–1049.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP 1994*, 44–49.

Surdeanu, M.; Harabagiu, S.; Williams, J.; and Aarseth, P. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, 8–15.

Yarowsky, D.; Ngai, G.; and Wicentowski, R. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*, 161–168.