Multilingual Lexicon Bootstrapping - Improving a Lexicon Induction System Using a Parallel Corpus

Patrick Ziering[1] **Lonneke van der Plas**[1] **Hinrich Schütze**[2]

[1]Institute for NLP, University of Stuttgart, Germany
[2]CIS, University of Munich, Germany
{Patrick.Ziering, Lonneke.vanderPlas}@ims.uni-stuttgart.de

Abstract

We address the task of improving the quality of lexicon bootstrapping, i.e., of expanding a semantic lexicon on a given corpus. A main problem of iterative bootstrapping techniques is the fact that lexicon quality degrades gradually as more and more false terms are added. We propose to exploit linguistic variation between languages to reduce this problem of semantic drift with a knowledge-lean and language-independent ensemble method. Our results on English and German show that lexicon bootstrapping benefits significantly from the multilingual symbiosis.

1 Introduction

High-quality semantic lexicons are an essential resource for many natural language processing (NLP) tasks like information extraction or anaphora resolution. Methods for automatically bootstrapping semantic lexicons given a seed list often struggle with lexicon accuracy decrease in higher iterations depending on corpus size (Igo and Riloff, 2009). One reason for this is *semantic drift*, which occurs when erroneous terms and/or contexts are introduced into and then dominate the iterative process (Curran et al., 2007). For instance, the ambiguity found in female names such as *Iris* and *Rose* may cause the induced terms to drift into flower names (McIntosh and Curran, 2009). Examples from the patent domain, that we are focusing on in this work, are PROCESSES that may drift into the semantic class of OBJECTS when terms such as *energy storage* and *spring coupling* are induced.

Previous work has used the cross-lingual correspondence between variations in linguistic structure and variations in ambiguity as a form of naturally occurring supervision in unsupervised learning for a number of tasks (Dagan et al., 1991; Snyder and Barzilay, 2010). On the lexical level, cross-lingual variations proved to remedy problems related to polysemy for synonym acquisition (Van der Plas and Tiedemann, 2010) and word sense disambiguation (Lefever and Hoste, 2010).

We hypothesize that cross-lingual divergences will be preeminently suitable to remedy problems related to semantic drift in iterative bootstrapping, where lexical and structural ambiguity give rise to erroneous terms and/or contexts. Languages are not isomorphic: ambiguous terms and contexts are frequently language-specific. In our example above, the English term *energy storage* is ambiguous, however, in German, each reading has its own translation. *Energy storage* is translated with *Energiespeicher* in the OBJECT reading and *Energiespeicherung* in the PROCESS reading.

Our multilingual ensemble lexicon bootstrapping system is inspired by Basilisk (Thelen and Riloff, 2002). Previous work has addressed semantic drift in Basilisk by conflict resolution between several classes (Thelen and Riloff, 2002), by using web queries (Igo and Riloff, 2009) and by combining Basilisk in an ensemble with an SVM tagger and a coreference resolution system (Qadir and Riloff, 2012). These approaches are monolingual. Instead we use a multilingual ensemble method where the induced lexicons of several languages constrain each other.

Apart from addressing semantic drift, the multilingual setting we propose has several other advantages. First, one language may leave implicit what another expresses directly in linguistic forms. In German, common nouns are capitalized and compound nouns are written as one word. We propagate German noun information via word alignment to English and thereby learn both single words as well as most multiword expressions (MWEs) without the need for a noun chunker or MWE recognizer.

Second, as a result of the multilingual ensem-

International Joint Conference on Natural Language Processing, pages 844–848,
Nagoya, Japan, 14-18 October 2013.

ble method we are able to induce lexicons for any language given a parallel corpus. We do not need seed lists for all languages, which are often sparse. Translating[1] the English seed list automatically results in high-quality lexicons for all other languages.

Finally, many pattern-based lexicon bootstrapping methods use pre-defined patterns which require language- and domain-specific syntactic analyses. Our multilingual approach makes use of a parallel corpus and tools from phrase-based machine translation that substitute for the necessity of pattern definition and once more guarantees a knowledge-lean and language-independent process.

2 Multilingual Lexicon Bootstrapping

Monolingual bootstrapping

```
 1: lexicon ← seed
 2: for int $i = 0$; $i < m$; $i{+}{+}$ do
 3:     patterns ←patternsOf(lexicon)
 4:     score(patterns)
 5:     patterns ← return-top-k(patterns,$20 + i$)
 6:     terms ← termsOf(patterns) $-$ lexicon
 7:     score(terms)
 8:     lexicon ← lexicon $\cup$ return-top-k(terms,$t$)
 9: end for
10: return lexicon
```

Figure 1: Basilisk (Thelen and Riloff, 2002)

Our basic algorithm is inspired by Basilisk (Thelen and Riloff, 2002), an algorithm developed for monolingual lexicon bootstrapping as shown in Figure 1. The starting point is a lexicon initialized with a given seed set. Then the lexicon is expanded iteratively. First Basilisk ranks all patterns containing terms from the lexicon (lexicon terms) (lines 3-4) based on the RlogF score:

$$\text{RlogF}(\text{pattern}_i) = F_i/N_i \log_2(F_i)$$

where F_i is the number of lexicon terms occuring in pattern$_i$ and N_i is the total number of terms occuring in pattern$_i$. Then Basilisk ranks all non-lexicon terms that occur in the 20+i highest-ranked patterns (where $i + 1$ is the number of performed iterations) (lines 5-7) based on the AvgLog score:

$$\text{AvgLog}(\text{term}_i) = 1/P_i \sum_{j=1}^{P_i} \log_2(F_j + 1)$$

where P_i is the number of patterns containing term$_i$. Finally Basilisk adds the t (originally 5) highest-ranked terms to the lexicon (line 8) and process repeats.

Multilingual ensemble framework

```
 1: for L$_i$ in $\{$L$_1, \ldots,$ L$_n\}$ do
 2:     B$_i$ ← initialize Basilisk for L$_i$
 3:     B$_i$.final ← $\{\}$
 4: end for
 5: while $\exists i$ (size(B$_i$.final)$< l$) do
 6:     for B$_i$ in $\{$B$_1, \ldots,$ B$_n\}$ do
 7:         B$_i$.iterate($m$, $t$)
 8:     end for
 9:     consensusCheck($\{$B$_1$,...,B$_n\}$)
10: end while
11: return (B$_1$.final,...,B$_n$.final)
```

Figure 2: Multilingual bootstrapping

We adapted Basilisk to a multilingual setting as shown in Figure 2. Key to the framework is the multilingual consensus check. In the **consensusCheck**, for each Basilisk process B$_i$ we intersect its lexicon with the translations of the lexicons of all other Basilisk processes B$_k$. We translated the lexicons from L$_k$ to L$_i$ using the bilingual dictionary DICT$_{k\leftrightarrow i}$ extracted from the corresponding phrase table. If the lexicon intersection is non-empty, the consensus terms are added to the final list of B$_i$ and the temporary lexicon is reset to the seed terms and the final list. If the intersection is empty, the lexicons are maintained completely leading to a higher chance of non-emptiness in the subsequent multilingual iteration.

We first initialize a Basilisk process B$_i$ for each language L$_i$ (Figure 1, line 1; Figure 2, lines 1-4). For each Basilisk process, we introduce a *final* list, that contains only lexicon terms that survived the consensus check. As long as at least one Basilisk process has a final list containing less than l terms, each Basilisk process performs m iterations of learning the top t terms[2] each (Figure 1, line 2-9; Figure 2, lines 6-8). Multilingual bootstrapping is finished, when all Basilisk processes have a final list of at least l terms (Figure 2, line 11).

[1] Potential ambiguity in translated seeds will be taken care of, because our ensemble learning prevents false terms resulting from erroneous seeds to be added to the lexicon.

[2] Thelen and Riloff (2002) originally set $t = 5$ - this would be inefficient with our ensemble lexicon bootstrapping on state-of-the-art machines because most of the time, there would be no consensus terms. We set $m = 2$ and $t = 25$, which seems to be a good trade-off between time and accuracy.

3 Experiments

Although our method is multilingual and language-independent we restrict our demonstration of its potential to two languages: German and English.

The parallel corpus. We use patent data distributed by the European Patent Office (EPO[3]) between 1998 and 2008. Most European patents provide their claims (the part of a patent defining the scope of protection) in German, English and French. We constructed a German-English parallel corpus out of 177,317 patent documents.

Creation of Moses phrase table. For each unordered language pair, we create a MOSES (Koehn et al., 2007) phrase table in several steps. We first apply sentence alignment (GARGANTUA Braune and Fraser (2010)), then word alignment (MGIZA++ Gao and Vogel (2008)) to the data. And finally, we apply the statistical machine translation tool MOSES to the parallel word-aligned data. The resulting data structure is a phrase table of word-aligned phrases in two languages as shown in Figure 3, where the third line indicates the word alignment.

Verfahren zur selektiven Flüssigphasenhydrierung
the process for selective liquid phase hydrogenation
0-0 0-1 1-2 2-3 3-4 3-5 3-6

Figure 3: Main content in a phrase table entry

Extracting terms and patterns. For term extraction, we define one language as the term-specifying language L_{term} (i.e., the language that specifies the set of candidate terms for all languages) – in our case, we choose German since it expresses term boundaries very directly in its linguistic forms (capitalized nouns, single word compounding). German terms are defined as a capitalized token with at least 4 letters. For each unordered language pair $\{L_{term},L_i\}$, we define each term in L_i as a sequence of tokens that are aligned to a term in L_{term}. In Figure 3, "liquid phase hydrogenation" is defined as term since it is aligned to "Flüssigphasenhydrierung" .

For reducing errors due to poor word alignment[4], we apply MATE (Bohnet, 2010) part of speech (PoS) tagger on phrases in languages other than L_{term} and define a PoS filter that removes spurious tokens at the left and right boundaries. Figure 4 shows the PoS filter for English, that is adapted from Justeson and Katz (1995) to the task of filtering tokens. The aligned terms that pass the filtering are stored in a dictionary $\text{DICT}_{term\leftrightarrow i}$.

English **(JJ|VBG|NN)* NN (IN NN+)?**

Figure 4: PoS pattern for term filtering

Patterns are extracted from the phrase tables as well. For each phrase in L_{term} and L_i we use the remaining tokens surrounding each term as bootstrapping pattern associated with this term (e.g., "Verfahren zur selektiven <X>" is defined as pattern for "Flüssigphasenhydrierung").[5]

Our final data set contains roughly 19 million German and English term-pattern pairs. The dictionary $\text{DICT}_{DE\leftrightarrow EN}$ comprises 1.8 million entries.

Translating seed sets. We define one corpus language as the seed-definining language L_{seed} – in our case, we choose English since it provides the richest lexical resources. Then, for all other languages L_i we translate each seed term from L_{seed} to L_i using the most frequent translation in $\text{DICT}_{seed\leftrightarrow i}$.

Evaluation. We evaluated the multilingual bootstrapping system on two semantic classes motivated by the technical field of patents: PROCESS and TECHNICAL QUALITY.

PROCESS: A method or event that results in a change of state (e.g., *stretching*, *molding process*, *redundancy control*, ...).

TECHNICAL QUALITY: A basic or essential attribute which is measurable or shared by all members of a group (e.g., *power consumption*, *piston speed*, *light reflection index*, ...).

The sources for the English seed sets have been WordNet lexicographer classes (Ciaramita and Johnson, 2003) and Wikipedia[6] word lists.

For each semantic class and language we induced lexicons of 2000 terms. For each lexicon we evaluated a sample of 200 terms. Two annotators first rated 50 terms for each language and class as TRUE or FALSE. Then they discussed disagreements. Afterwards, they rated the remaining terms in each lexicon sample. We achieved a total inter-annotator agreement of $\kappa = .701$ (Cohen's

[3] www.epo.org

[4] Since our corpus is not large enough for perfect word alignment, it can be supported by a part of speech tagger. To keep the process completely language-independent, this step may also be skipped.

[5] We remove unique patterns because they do not contribute to lexicon bootstrapping.

[6] www.wikipedia.org

Kappa). For the results, we used the labeled lexicons of the annotator that finalized the task first.

4 Results

In our experiments we compare two methods. The first is the monolingual bootstrapping method[7] and the second is the bilingual ensemble bootstrapping method. For a proper comparison both methods make use of the same data as described in Section 3. Table 1 shows the accuracy of the induced lexicons for German and English when learned separately (lines 1-2) and when induced with the bilingual ensemble bootstrapping method (line 3)[8].

	Mode	Process	Technical Quality
1	DE	.730	.880
2	EN	.740	.895
3	DE / EN	.980† / .790	.960† / .955†

Table 1: Results of lexicon evaluation

Bilingual ensemble bootstrapping outperforms monolingual bootstrapping in both classes and languages. For the class TECHNICAL QUALITY there is a significant improvement in both languages (German: +.080; English: +.060). For the class PROCESS there is a significant improvement for German (+.250), whereas there is a nonsignificant improvement for English.

Analysis and discussion. To give the reader a better idea of how the bilingual ensemble method remedies semantic drift, we will comment on the asymmetric impact on performance, when high levels of ambiguity are present in one of the two languages.

We know from linguistic research (Ehrich and Rapp, 2000) that the German *ung*-ending is subject to sortal ambiguity. Words ending in *-ung* can be of various semantic types: processes, objects, events, and states. Many terms in the PROCESS class are described by nouns ending in *-ung*. Their sortal ambiguity gives rise to semantic drift from PROCESS to TECHNICAL QUALITY (e.g., *Belastung* can mean *charging* or *burden*), and to PROCESS-RELATED DEVICE (e.g., *Steuerung* can mean *steering* or *controller*). This sortal ambiguity of nouns in the PROCESS class does not have its counterpart in the English lexicon. It is therefore not surprising that we see a large improvement in the quality of the German lexicons, when English is used in the ensemble bootstrapping method. We achieve an improvement in German of +.250, the largest improvement overall.

In the present bilingual setting, we cannot prevent the ambiguity found in the German terms to influence the English terms. We believe that this is the reason for the asymmetric impact of bilingual bootstrapping on the class PROCESS, where we see only a small improvement in English (+.060). The positive effects from ensemble learning for the English PROCESS class is partly wiped out by the influence of high levels of ambiguity in German. In future work, we plan to add several languages to be able to prevent ambiguity in one language to overshadow the multilingual ensemble.

5 Conclusion

We address the problem of semantic drift in iterative bootstrapping. We propose a multilingual ensemble learning method for lexicon bootstrapping, in which lexicons for several languages are induced in parallel and constrain each other. This method exploits linguistic variation between languages to reduce the impact of lexical and structural ambiguity within one language. In a case study on German and English and the two semantic classes TECHNICAL QUALITY and PROCESS, we show that bilingual lexicon bootstrapping outperforms monolingual bootstrapping in all classes and languages.

In addition, our multilingual approach to lexicon bootstrapping is particularly knowledge-lean and language-independent. A parallel corpus, language-independent machine translation tools and seed lists of one single corpus language suffice to extract patterns, determine term boundaries and provide seed lists for an in principle unlimited number of languages.

Acknowledgments

This research was supported by the project TOPAS[9] (Tool platform for intelligent Patent Analysis and Summarization), its team members and the European Commission with its FP7-SME Program. The TOPAS Consortium is composed of five partners: Brügmann Software Inc., IALE Inc., Intellisemantic Inc., Pompeu Fabra University and University of Stuttgart. We thank all colleagues for their support. This research was further supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the SFB 732. Thanks also to Sebastian Ebert for conducting the creation of the phrase tables.

[7] Although the first method relies on a parallel corpus and multilingual preprocessing, we refer to it as the monolingual method because the learning is done monolingually.

[8] We mark each number with † if it significantly outperforms monolingual bootstrapping (z-test for proportions; $p < .05$).

[9] topasproject.eu

References

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *COLING 2010*.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *COLING '10*.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *EMNLP 2003*.

J. R. Curran, T. Murphy, and B. Scholz. 2007. Minimising Semantic Drift with Mutual Exclusion Bootstrapping. In *PACLING 2007*.

Ido Dagan, Alon Itai, and Ulrike Schwall. 1991. Two languages are more informative than one. In *ACL 1991*.

Veronika Ehrich and Irene Rapp. 2000. Sortale Bedeutung und Argumentstruktur: *ung*-Nominalisierungen im Deutschen. *Zeitschrift für Sprachwissenschaft*, 19(2):245–303.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*.

Sean P. Igo and Ellen Riloff. 2009. Corpus-based semantic lexicon induction with web-based corroboration. In *NAACL 2009*, pages 18–26.

J. Justeson and S. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, pages 9–27.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL '07*.

Els Lefever and Véronique Hoste. 2010. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In *SemEval 2010*.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *ACL-IJCNLP 2009*.

Ashequl Qadir and Ellen Riloff. 2012. Ensemble-based semantic lexicon induction for semantic tagging. In **SEM-2012*.

Benjamin Snyder and Regina Barzilay. 2010. Climbing the tower of babel: Unsupervised multilingual learning. In *ICML 2010*.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP 2002*.

Lonneke Van der Plas and Jörg Tiedemann. 2010. Finding medical term variations using parallel corpora and distributional similarity. In *OntoLex 2010*, Beijing, China.