University of Stuttgart
Germany

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
IJCNLP 2013

1. Monolingual Lexicon Bootstrapping

Initial system: Basilisk (Thelen and Riloff, 2002)

- Lexicon expansion by iteratively adding the highest-scored terms (e.g, GOLD) occurring in the highest-scored patterns (e.g., MADE OF <X>), based on a statistical scoring function.

```
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 ∪ return-top-k(terms,t)
9: end for
10: return lexicon
```

Problem: Semantic Drift

- False terms infect the lexicon and lead to a gradual degradation of lexicon quality.
- Semantic drift is frequently caused by polysemy.
  The polysemous PROCESS term *energy storage* might lead to a semantic drift to OBJECT.

Solution: Multilingual Ensemble Bootstrapping

Polysemy is frequently language-specific.

- We perform Basilisk on several languages in parallel.
- We retain only those concepts that have been induced in all languages after some iterations.

2. Building multilingual input data

Given a parallel corpus with any set of languages, we create a multilingual Basilisk input.

For each unordered language pair, we create a phrase table:

1. **Sentence alignment:** GARGANTUA (Braune and Fraser, 2010)
2. **Word alignment:** MGIZA++ (Gao and Vogel, 2008)
3. **Phrase table creation:** MOSES (Koehn et al., 2007)

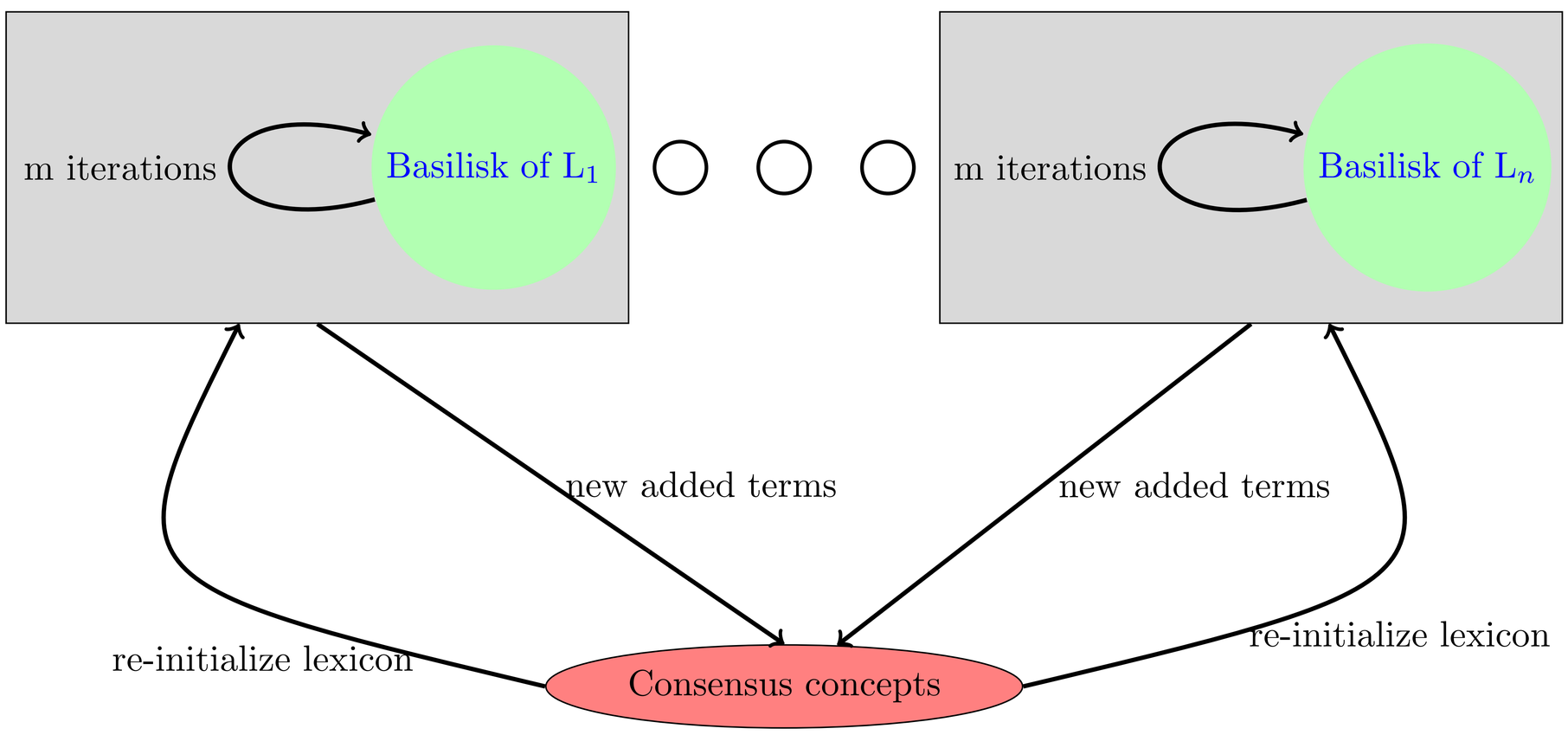
Verfahren zur selektiven Flüssigphasenhydrierung
the process for selective liquid phase hydrogenation
0-0 0-1 1-2 2-3 3-4 3-5 3-6

From each phrase table, we extract terms and patterns:

4. **Terms:**
   - Definition of a term-specifying language L_{term} (e.g., German, where a term is a capitalized token with at least 4 letters).
   - In a phrase table of $\{L_{term}, L_i\}$, we define each term in L_i as a token sequence aligned to a term in L_{term}.
     → Spurious tokens may be removed by using a PoS filter.
   - All aligned terms are stored in a dictionary $\text{DICT}_{term \leftrightarrow i}$.
5. **Patterns:**
   Pattern definition: tokens surrounding a term in a phrase

Our approach needs seed terms for only one language:

6. **Translating seeds:**
   - Definition of a seed-specifying language L_{seed} (e.g., English, since it provides the richest lexical resources).
   - For each language L_i, we translate each seed term from L_{seed} to L_i using the most frequent translation in $\text{DICT}_{seed \leftrightarrow i}$.



3. Advantages of Multilingual Lexicon Bootstrapping

+ Remedies problems related to semantic drift
+ No need for a (shallow) parser for term/pattern extraction
+ Language-independence
+ Learns both single words and MWEs
+ No need for a MWE recognizer given languages like German
+ Needs seed terms for only one language

4. Experiments

The parallel corpus

- 177K patents distributed by the European Patent Office (EPO).
- Patent claims are frequently multilingual.
- We constructed a German-English parallel corpus of patent claims.

The evaluation

- Baseline: Monolingual Basilisk with multilingual input part
- Semantic classes: PROCESS and TECHNICAL QUALITY
- Seeds: WordNet Supersense and Wikipedia
- Induced lexicons: 2000 terms
- Annotators labelled a 200-sample of each lexicon ($\kappa = .701$)

The results

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

- German PROCESS words ending on *-ung* are highly polysemous and thus give rise to semantic drift
- Using bilingual bootstrapping leads to an improvement (+.250) in German lexicon quality
- Similar drifts in later iterations in English can only partly be remedied due to this *ung*-polysemy
- Future work: More languages for addressing polysemy

Supported and funded by the patent research project TOPAS and the Deutsche Forschungsgemeinschaft (DFG SFB 732)