Leveraging WordNet and Wikipedia for Multilingual Glosses

Isabelle Augenstein, Amin Kiem und Christoph Mayer | Institut für Computerlinguistik, Universität Heidelberg | SWP 2010

Motivation

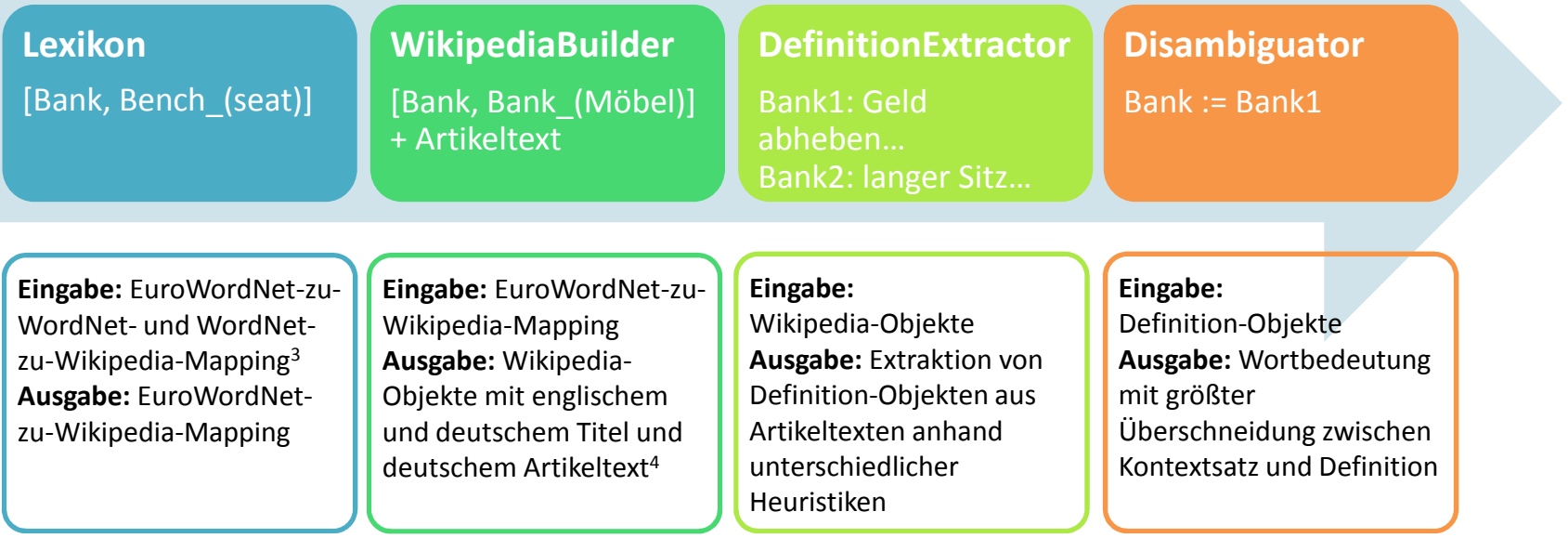
- **Problem:** Lexikalische Ambiguität
  - Bank (Möbel) vs. Bank (Geldinstitut)
- **Lösung:** WSD mit Simplified LESK[1]
  - Annahme: die Definition der besten Lesart hat die größte Überschneidung mit dem Kontextsatz
  - Kontextsatz: Franz will sich **Geld** von der Bank **leihen**.
  - Definition Bank A: *Geldinstitut, bei dem man **Geld** einzahlen oder **leihen** kann*
  - Definition Bank B: *langer Sitz für mehr als eine Person*

Evaluation

- **Goldstandard** von Broscheit et al. (2010) [2]
  - Von 40 Wörtern mit jeweils 20+ Testsätzen (richtige Lesart annotiert als GermaNet 5 Sense) sind 18 Nomen
- **Einschränkungen**
  1. Beschränkung des WSD-Systems auf Nomen
  2. Unvollständiges WordNet-zu-Wikipedia-Mapping
  3. Versionskonflikte: GermaNet 5 (Goldstandard) vs. Germanet 1.2 (EuroWordnet-zu-WordNet-Mapping)
  4. Ausschluss von Mono-Sense-Wörtern
  5. Veralteter Wikipedia-Dump: Wegfall der Wörter mit fehlenden deutschen Seiten

Ausblick

- Aktualisierung bzw. Vervollständigung der Ressourcen
- Heuristiken verbessern oder neu hinzufügen



Heuristik	Precision
Sentence Model	26,7 %
Paragraph Model	0 %
Text Model	30 %
Word TFIDF	0 %
Sentence TFIDF	0 %
Normalized Sentence TFIDF	26,7 %

[1] Kilgarriff, Adam (2000). *Framework and results for English SENSEVAL.* In: Computers and the Humanities 34 (1-2), Special Issue on SENSEVAL

[2] Broscheit, Samuel et al. (2010). *Rapid Bootstrapping of Word Sense Disambiguation Resources for German.* In: Proceedings of the 10. Konferenz zur Verarbeitung Natürlicher Sprache, Saarbrücken

[3] Navigli, Roberto und S. P. Ponzetto (2010). *BabelNet: Building a Very Large Multilingual Semantic Network.* In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden

[4] Zesch, Torsten et al. (2008). *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.* In: Proceedings of the Conference on Language Resources and Evaluation (LREC).