Automatic Extraction of Synonyms for German Particle Verbs from Parallel Data with Distributional Similarity as a Re-Ranking Feature

Moritz Wittmann, Marion Weller, Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung – Universität Stuttgart
{wittmamz|wellermn|schulte}@ims.uni-stuttgart.de

Universität Stuttgart

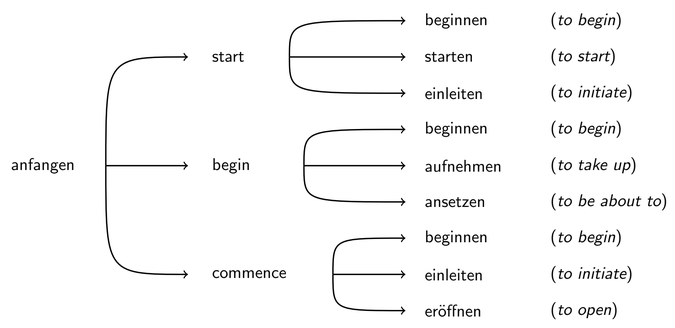
Introduction

- We present a method to extract synonyms for German particle verbs from word-aligned parallel data, based on [1].
- Synonyms are important in many NLP tasks and applications, such as thesaurus creation, machine translation and machine translation evaluation.
- German particle verbs are productive compositions of a base verb and a prefix particle
  - *anfangen* (*to begin*)
  - *nachrennen* (*to run after somebody*)
- Particle verbs may also occur as separate words:
  Er **nahm** *den Mantel wegen der starken Hitze* **ab**.
  He **took off** *the coat because of the intense heat.*
- We apply pre-processing in form of reordering the data.

Synonym Extraction

Synonym extraction consists in two steps:

- Gathering all English translations (pivots) of the German input verb
- Translating all pivots back to German, which results in a set of synonym candidates



- The *synonym probability* $p(e_2|e_1)_{e2 \neq e1}$ for a synonym candidate e_2 given a particle verb e_1 is calculated as the product of two translation probabilities:
  - The *pivot probability* $p(f_i|e_1)$ (the English phrase f_i is a translation of the particle verb e_1)
  - The *return probability* $p(e_2|f_i)$ (the synonym candidate e_2 is a translation of the English phrase f_i).

  The final score is the sum over all pivots $f_{1..n}$:

$$p(e_2|e_1)_{e2 \neq e1} = \sum_{i=1}^{n} p(f_i|e_1)p(e_2|f_i) \quad (1)$$

- In order to decrease the amount of invalid synonym candidates, various filtering heuristics were applied during the *pivot probability step* and the *return probability step*.
- Any phrases with at least one verb are allowed as synonyms. Candidates containing the same words in a different order were gathered into one entry.

gold	ranked synonyms	gloss	probability
+	bauen	*to build*	0.11184
+	schaffen	*to create/make*	0.08409
+	errichten	*to construct*	0.07393
(+)	entwickeln	*to develop*	0.04699
-	ausbauen	*to extend*	0.02281
+	beruhen	*to be based*	0.02259
+	einrichten	*to set up*	0.01589
+	gestalten	*to design*	0.01414
+	bilden	*to form*	0.01212
+	basieren	*to base*	0.01210

Table 1: The 10 top-ranked synonym candidates for the verb *aufbauen* (*to build up*).

Re-Ranking Strategies

To improve the ranking according to the synonym probability, we experimented with two re-ranking strategies.

Language Model Re-Ranking

- Synonym candidates are rated by a language model in the context of their respective particle verbs.
- We used 10 random sentences containing the particle verb as context for the synonym candidates; the perplexities obtained by the language model were averaged.
- This re-ranking strategy showed no improvements in the results.
- Language models depend too strongly on the sentences chosen for scoring (word-sense mismatches and incompatible subcategorization frames).

Distributional Similarity Re-ranking

- The distributional similarity between the particle verb and its synonym candidates is used to improve the ranking: we assume that similar words share similar contexts.
- Distributional similarity is computed as the cosine similarity of the respective context vectors (content words within a window of 10 words to each side), using local mutual information instead of co-occurrence frequencies extracted from a large corpus.
- In order to facilitate the computation and comparison of cosine similarity, the synonym candidates were restricted to single verbs.

top-5 candidates not reordered	top-5 candidates reordered: distr.-sim.
erfüllen (*to fulfil*)	zusammentreten (*to convene*)
entsprechen (*to comply with*)	zusammentreffen (*to meet*)
treffen (*to meet*)	tagen (*to meet*)
erreichen (*to reach*)	zusammenfinden (*to congregate/gather*)
einhalten (*to keep to*)	begegnen (*to meet/encounter*)

Table 2: The top-5 synonym-candidates for the verb *zusammenkommen* (*to come together*) before and after re-ranking using distributional similarity. Highlighted verbs occur in the gold standard.

Experiments and Evaluation

- German is a morphologically rich language: we compare variants of simplifying the surface forms by lemmatization.
- For evaluation, the top-ranked candidates are compared to a gold standard.
- Additionally, we present a small-scale manual evaluation.

Creation of a Gold Standard

- The synonym entries of the gold standard were looked up in the online dictionary *Duden*.
- Out of the 500 most frequent German particle verbs (freq ≥ 15) in our data, 138 have 30 or more synonyms listed (this ensures that a precision of 1 can be reached when evaluation the 30 top-ranked synonym candidates).

Data

- We used the DE-EN version of *Europarl* (1.5M parallel sentences)
- Word alignment was computed using *GIZA++*.
- The English side was tagged with *TreeTagger*; for the reordered German part, we used *SMOR* to obtain lemmatized forms.
- Distributional similarity was computed based on the *SdeWaC corpus* (880M words).
- We applied reordering steps to the parsed (*BitPar*) German text:
  - Move verbs to a sentence-initial position, corresponding to the expected English structure:
    - *dass* **sich** *die ersten Länder möglichst an den Wahlen zum Europäischen Parlament im Jahre 2004* **beteiligen können**.
      that **refl-pronoun** *the first countries if possible at the elections of the European Parliament in the year 2004* **participate can**.
    - *dass die ersten Länder* **können beteiligen sich** *möglichst an den Wahlen zum Europäischen Parlament im Jahre 2004 .*
      that the first countries **can participate refl-pronoun** *if possible at the elections of the European Parliament in the year 2004.*
  - Move separate particles in front of the respective verbs:
    - *Die Einkommen* **steigen** *steil* **an**.
      The incomes **rise** *strongly* **PART**.
    - *Die Einkommen* **an steigen** *steil.*
      The incomes **PART rise** *strongly.*

Results and Evaluation

- Various combinations of alignments and lemmatization were tested and compared in order to find the best one.

	EN
a	inflected
b	lemmatized

	DE
c	lemm. particle verbs
d	lemm. ADJ, V, N
e	lemm.

	Files A	Files In	top 1	top 5	top 30
1	a-e	a-d	58.6956	44.0579	22.2946
2	a-d	b-d	57.9710	43.9130	22.0048
3	a-e	b-d	57.2463	43.3333	21.9082
4	a-d	a-d	57.2463	43.9130	22.2463
5	a-c	b-d	56.5217	43.3333	22.1014
6	b-c	b-d	56.5217	40.0000	20.3623
7	a-c	a-d	55.7971	43.0434	22.1014
8	b-c	a-d	54.3478	40.4347	20.0000

Table 3: Precision for different combinations of pre-processing strategies. The 3 best systems are highlighted in each range. **A**: files used for alignment and **In**: input for synonym extraction.

- English inflection (*number* on nouns and *third-person marking* on verbs) provides useful information for the alignment,
- The morphologically more complex German (*number*, *gender*, *case*, *strong/weak inflection* on nominal phrases and richer *verbal inflection*) benefits from lemmatization.
- While the language model approach failed to improve the scores, distributional similarity re-ranking leads to considerable increases.

	top 1	top 5
no re-ranking	58.6956	44.0579
language model	58.6956	44.0579
distributional similarity	63.7681	49.7101

Table 4: Results for two re-ranking strategies for the best system (1) from table 3.

Manual Evaluation

- 4 German native-speakers were given a selection of 14 particle verbs and the respective 30 top-ranked synonym candidates.

verb	synonym candidate	P1	P2	P3	P4	gold
einstellen (*to cease*)	aussetzen (*to adjourn*)	yes	no	no	no	yes
einsetzen (*to intercede*)	verteidigen (*to defend*)	no	yes	no	no	yes
aufbauen (*to build up*)	entwickeln (*to develop*)	no	no	yes	no	no
festlegen (*lay down*)	niederlegen (*to put down*)	no	no	no	yes	no
zusteuern (*to head for*)	sich bewegen (*to move*)	no	no	no	no	yes
festhalten (*to record*)	hervorheben (*to emphasize*)	no	no	no	no	yes

Table 5: Individual annotation decisions in contrast to the gold standard for a subset of verb and synonym candidate pairs.

- Those candidates which were considered to be valid synonyms by at least two evaluators were counted when calculating the overall precision for the manual evaluation.
- The average agreement over the 14x30 synonym candidates between the four evaluators was 82.9%
  - all evaluators decide equally: 100%;
  - three evaluators decide equally: 75%;
  - otherwise: 50%

Verbs	P1	P2	P3	P4	Gold
aufbauen (*build up*)	46.67	36.67	53.33	46.67	50.00
einstellen (*set*)	50.00	36.67	33.33	46.67	43.33
festlegen (*determine*)	50.00	26.67	23.33	46.67	36.67
einsetzen (*use*)	40.00	26.67	6.67	40.00	33.33
umbringen (*kill*)	36.67	40.00	26.67	30.00	30.00
mitteilen (*inform*)	26.67	36.67	63.33	36.67	26.67
zusehen (*watch*)	46.67	20.00	43.33	36.67	26.67
darstellen (*represent*)	20.00	16.67	20.00	33.33	23.33
festhalten (*hold on to*)	33.33	16.67	10.00	26.67	23.33
aussetzen (*suspend*)	36.67	3.33	10.00	10.00	16.67
aufnehmen (*record*)	43.33	30.00	23.33	30.00	10.00
zusteuern (*head towards*)	6.67	26.67	23.33	40.00	10.00
aufgehen (*rise*)	13.33	30.00	6.67	16.67	0.00
vornehmen (*carry out*)	0.00	6.67	16.67	33.33	0.00
average	32.14	25.24	25.71	33.81	23.57

Table 6: The scores attributed to each verb by each of the four evaluators, as well as the gold standard evaluation score on the right.

Conclusion and Future Work

- We presented a method for the extraction of synonyms for German particle verbs using parallel data.
- In our evaluation we compared different pre-processing variants.
- The best system had a precision of 58.7% for the top-1-ranked synonym candidates; using distributional similarity for re-ranking leads to a further improvement (63.8%). A manual evaluation was carried out as well, with generally higher scores compared to the gold standard evaluation.
- One problem with this approach for synonym extraction is the lack of a method for dealing with word-sense ambiguity:
  - Controlling for word-senses may improve results and prove useful for applications which benefit from a word-sense distinction.
  - Word-sense ambiguity was also one of the reasons why the language-model re-ranking performed poorly.
- Improving the word alignments and recognizing multi-word expressions may significantly improve the results as well.
- Another possible strand of future work is the inclusion of more language pairs: as the respective translation and return probabilities are independent from each other for different language pairs, a combination of scores obtained from pivots of different languages should provide a better basis for ranking synonym candidates.

References

[1] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 597–604, Ann Arbour, 2005.

Acknowledgements

Funded by the DFG Research Project "Distributional Approaches to Semantic Relatedness" (Moritz Wittmann, Marion Weller) and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).