

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

Pfaffenwaldring 5b – 70569 Stuttgart – Germany

{wittmamz|wellermn|schulte}@ims.uni-stuttgart.de

Abstract

We present a method for the extraction of synonyms for German particle verbs based on a word-aligned German-English parallel corpus: by translating the particle verb to a pivot, which is then translated back, a set of synonym candidates can be extracted and ranked according to the respective translation probabilities. In order to deal with separated particle verbs, we apply re-ordering rules to the German part of the data. In our evaluation against a gold standard, we compare different pre-processing strategies (lemmatized vs. inflected forms) and introduce language model scores of synonym candidates in the context of the input particle verb as well as distributional similarity as additional re-ranking criteria. Our evaluation shows that distributional similarity as a re-ranking feature is more robust than language model scores and leads to an improved ranking of the synonym candidates. In addition to evaluating against a gold standard, we also present a small-scale manual evaluation.

Keywords: Synonym extraction, distributional similarity, particle verbs.

1. Introduction

Synonyms are important in many NLP tasks and applications, such as thesaurus creation (Curran, 2003; Lin et al., 2003), machine translation (Carbonell et al., 2006; van der Plas and Tiedemann, 2006) and machine translation evaluation (Lavie and Denkowski, 2009). In this paper, we present a method to extract synonyms for German particle verbs from word-aligned bilingual data. German particle verbs are productive compositions of a base verb and a prefix particle (such as *anfangen*, *nachrennen*). They represent a challenging target group among multi-word expressions, as they may occur as one unit (i.e. particle and verb in one word), or in separated form (verb and particle are separated), as illustrated by example (1).

- (1) *Er nahm den Mantel wegen der starken Hitze ab.*
He took off the coat because of the intense heat.

This property can lead to problems, not only in applications such as parsing, word alignment and machine translation, but also with regard to low-level NLP tasks such as part-of-speech tagging and lemmatization, since it is difficult to treat the verb and its particle as one unit when they appear separated. As part of a larger project, we are interested in the meaning and the compositionality of German particle verbs. To this end, synonyms of particle verbs are an important means (i) to address the particle verb meaning through paraphrasing, and (ii) to address the meaning components of the constituents (i.e. the notoriously ambiguous particles, and the base verbs).

In the presented approach for synonym extraction, we use English translations of German particle verbs as pivots and the respective back-translations are considered as synonym candidates, which are then filtered and ranked. In order to further improve the ranking, we apply and compare two re-ranking strategies based on contextual, monolingual information: (i) rating the synonym candidate in the context of the original particle verb in a sentence by means of a language model and (ii) calculating the contextual similarity

of synonym candidates and particle verbs based on distributional window co-occurrence.

The paper is structured as follows: we explain the pre-processing applied to the German part of our bilingual data in order to deal with separated particle verbs, and then present the process of extracting particle verbs and their synonym candidates. We then compare re-ranking the obtained candidates according to language model scores and distributional similarity: while the language model scores only have a marginal influence, using distributional similarity considerably improves the ranking. We also study the effects of reducing morphological richness (i.e. lemmatization) on word alignment and subsequently on the extraction of synonym candidates. In the evaluation, we measure the precision of the top-ranked synonym candidates by means of comparison with a gold standard containing comprehensive lists of synonyms for a given particle verb. This automatic evaluation is supplemented with a small-scale manual evaluation.

2. Related work

We mainly follow the method described by Bannard and Callison-Burch (2005), who were the first to extract synonyms on the basis of pivots and back-translations in parallel corpora, for different phrase types. In contrast to their approach, we apply the method to German particle verbs, which requires a suitable pre-processing step. Other work on the automatic extraction of paraphrases has focused on using monolingual parallel corpora, such as multiple translations of novels (Barzilay and McKeown, 2001), or monolingual comparable corpora, such as collections of articles about the same event (Barzilay and Lee, 2003).

A large body of research has exploited distributional models to paraphrasing or synonym extraction by relying on the contextual similarity of two words or phrases, most prominently Lin (1998), Sahlgrén (2006), Padó and Lapata (2007); we use this method for re-ranking the obtained synonym candidates. Similar methods for extracting synonyms include Wang et al. (2010), which uses patterns found in

newspapers and probabilities of verbs co-occurring with a pattern, Blondel and Senellart (2002) in which synonyms are extracted from a dictionary by using a graph representation of words used in definitions (with vertices between words that are contained in each other’s definition) and a websearch algorithm, or Dang et al. (2009), where the focus lies on the context around target words, composed of vectors constructed from surrounding n-grams.

With respect to future work in the field, specifically the addition of word sense disambiguation (cf. discussion in section 7), the method used in Diab and Resnik (2002) may be of interest; it consists in using a sense inventory for English (target language) to determine a predominant word sense for a group of English words aligned with the same French word (source language), then projecting the predominant word sense over to the source word.

3. Data pre-processing

The fact that particle verbs often occur in separated form is problematic for many applications, including word alignment and statistical machine translation. In addition, English and German tend to have diverging sentence orderings, which adds further problems to the task of word alignment as it is difficult to align words that are positioned at a large distance from each other. This applies particularly to verbs, which can appear at the very end of German clauses, while the corresponding English verbs tends to be at the beginning of a clause. Collins et al. (2005) and Fraser (2009) showed that for SMT applications, it is helpful to restructure the source-side language in such a way that the new structure imitates that of the target language. We applied reordering steps following Fraser (2009) to the parsed German part of the bilingual data; the reordering includes moving verbs from the verb-final position to a sentence-initial position corresponding to the expected English structure, and moving separate particles in front of the respective verb, as illustrated in examples (2) and (3).

- (2 a) *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**.*
- (2 b) *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.*
- (3 a) *Die Einkommen **steigen steil an** ...*
*The incomes **rise strongly PART** ...*
- (3 b) *Die Einkommen **an steigen steil** ...*
*The incomes **PART rise strongly** ...*

This pre-processing aims at improving the alignment quality and also allows to conveniently extract separated particle verbs and treat them in the same way as non-separated occurrences; in the synonym extraction step, this helps to

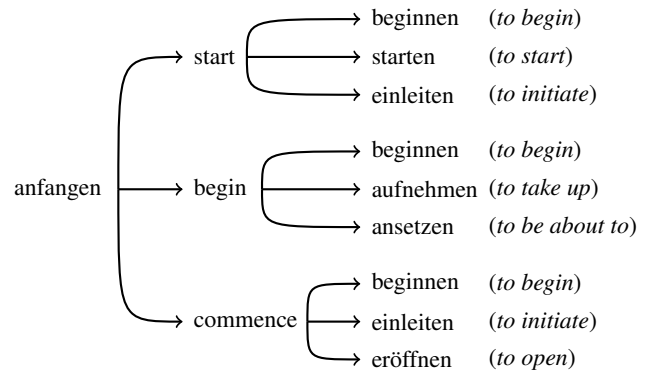


Figure 1: Synonym extraction based on pivots: the back translations of the pivots of the verb *anfangen* (*to begin*) form the set of synonym candidates.

avoid problems caused by incomplete particle verbs occurring in the synonym candidate sets.

The reordered German part of the data and the English part are then word aligned. As German is a morphologically rich language, the data was lemmatized in a further pre-processing step (see section 6.3 for more details).

4. Synonym extraction

The method for synonym extraction consists in first gathering all target-language translations (pivots) of the input verb, and then translating all pivots back, which results in a set of synonym candidates. Figure 1 illustrates this process: starting with one particle verb, several pivots are found via word alignment. Their back translations then constitute the set of synonym candidates of the starting verb.

4.1. Methodology

In order to rank the candidates according to how likely they are to be valid synonyms, each candidate is assigned a probability. The *synonym probability* $p(e_2|e_1)_{e_2 \neq e_1}$ for a synonym candidate e_2 given a particle verb e_1 is calculated as the product of two translation probabilities: the *pivot probability* $p(f_i|e_1)$, i.e. the probability of the English phrase f_i being a translation of the particle verb e_1 , and the *return probability* $p(e_2|f_i)$, i.e. that the synonym candidate e_2 is a translation of the English phrase f_i . The final score is the sum over all pivots $f_{1..n}$:

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

The translation probabilities are estimated using relative frequencies based on counts in the parallel corpus.

4.2. Filtering

In order to decrease the amount of invalid synonym candidates, filtering heuristics were applied at the *pivot probability step* and the *return probability step*: obviously useless English translations containing only stop-words (e.g. articles) or punctuation were discarded as pivots. In the back-translation step, synonym candidates consisting only of stop-words or punctuation were removed, as well as candidates containing the input particle verb or no verb at all,

aufbauen	→	build	0.3820
	→	build up	0.0696
	→	develop	0.0693
	→	establish	0.0669
	→	create	0.0482
	→	base	0.0436
	→	rebuild	0.0374
	→	construct	0.0342
	→	set up	0.0315
	→	to build	0.0280

Figure 2: English pivots with probabilities for the particle verb *aufbauen* (*to build*).

gold	ranked synonyms	gloss	probability
+	bauen	<i>to build</i>	0.11184
+	schaffen	<i>to create/make</i>	0.08409
+	errichten	<i>to construct</i>	0.07393
(+)	entwickeln ¹	<i>to develop</i>	0.04699
-	ausbauen	<i>to extend</i>	0.02281
+	beruhen	<i>to be based</i>	0.02259
+	einrichten	<i>to set up</i>	0.01589
+	gestalten	<i>to design</i>	0.01414
+	bilden	<i>to form</i>	0.01212
+	basieren	<i>to base</i>	0.01210

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

assuming that a valid synonym of a verb has to contain at least one verb that is not the input verb itself. It is important to note that we do not restrict the set of synonyms to only particle verbs or verbs (as a one-word synonym), but allow any phrases as long as they contain at least one verb. Multi-word candidates containing the same words in a different order were gathered into one entry; this simplifies the comparison with the gold standard.

4.3. Examples

Figure 2 shows a subset of the pivots for the verb *aufbauen* (*to build up*), with its synonym candidates in table 1: depending on the context, they can be considered valid synonyms of the input verb. Note that *aufbauen* can have different meanings: *to build/set up sth.* or *to be based/founded on sth.*; the second meaning is represented by the entries *beruhen* and *basieren*.

Table 2 lists the top-ranked synonym candidates for the input verb *anfangen* (*to begin*): save for a few exceptions (*ausgehen*, *reichen*, *nehmen*), the obtained candidates can be considered valid synonyms of the input verb. Again, many of the synonym candidates are ambiguous; for example the verb *anlaufen* can have the meanings of *to begin*, *to start running*, *to tarnish*, *to accrue* or *to head for port*. For the evaluation, we consider a candidate to be correct if one

¹The entry in the gold standard is “*sich entwickeln*”, i.e. with a reflexive pronoun.

gold	ranked synonyms	gloss	probability
+	beginnen	<i>to begin</i>	0.36014
+	aufnehmen	<i>to take up</i>	0.04452
(-)	eingeleitet	<i>initiated</i>	0.02207
-	ausgehen	<i>to assume</i>	0.01767
+	einleiten	<i>to initiate</i>	0.01628
-	reichen	<i>to extend</i>	0.01568
+	starten	<i>to start</i>	0.01222
-	nehmen	<i>to take</i>	0.01158
+	anlaufen	<i>to begin</i>	0.00787
+	ansetzen	<i>to be about to</i>	0.00778

Table 2: The 10 top-ranked synonym candidates for the particle verb *anfangen* (*to begin*).

of its meanings is a valid synonym of the particle verb.

The problem of multiple word senses is quite prevalent when working with (particle) verbs, with regard to both the input verb and the obtained synonym candidates. While we do not control for different word senses within this work, we address this issue in section 7 where we discuss potential problems caused by ambiguous input verbs and outline strategies to deal with them.

5. Re-ranking strategies

For improving the ranking of the synonyms, we add two additional re-ranking features to the basic method of synonym extraction: (i) scores obtained from rating sentences with the synonym candidate in the context of the input particle verbs by means of a monolingual language model and (ii) distributional similarity of particle verbs and synonym candidates. As both features are based on monolingual context, they represent independent criteria in addition to the bilingual setup based on translation probabilities.

We discuss and compare the two methods with regard to their flexibility in terms of subcategorized elements and their ability to handle verbs with different word senses: while the language model approach seems to depend too strongly on the sentences chosen for rating, the method based on distributional similarity is more robust. This insight is also reflected by the results obtained by the two methods (cf. section 6.3).

5.1. Language model-based approach

Assuming that valid synonyms fit better into the context of the input verb than non-synonyms, the input verb is replaced by the synonym candidates and the altered context is rated in a language model. To this end, a set of 10 sentences for each particle verb was randomly selected from the corpus. This set was restricted to contain only infinitive forms of the particle verb, in order to avoid problems with verbal inflection. To minimize effects caused by different sentence lengths, only an 11-word window was scored by the language model (the target verb being in the middle). Based on the 10 test sentences, the average perplexity for each synonym candidate was calculated using the SRILM toolkit (Stolcke, 2002)². With a lower perplexity (ppl) cor-

²<http://www-speech.sri.com/projects/srilm>

original sentence	Ich frage mich , ob wir [beiden Parteien je klarmachen können , dass es hier] nur Verlierer geben kann. <i>I question me , whether we [both parties ever make clear can , that it here] only losers be can</i> <i>I wonder whether we can ever make it clear to both parties that there can be only losers.</i>
verb replaced with synonym candidate	Ich frage mich , ob wir [beiden Parteien je verdeutlichen können , dass es hier] nur Verlierer geben kann. <i>I wonder whether we can ever illustrate to both parties that there can be only losers.</i>

Figure 3: Example sentences for language model re-ranking (Sequence in brackets: window rated by the language model).

meaning 1: to stop	Damit die Getreidebauern ihre Produktion einstellen , werden sie selbstverständlich bezahlt : <i>that the grain growers their production stop , are they of course payed .</i> <i>the grain growers are of course payed for stopping their production.</i>
meaning 2: to adapt to	Die EU und die Mitgliedstaaten müssen sich um jeden Preis <u>auf</u> die Erfordernisse des Umweltschutzes einstellen ... <i>The EU and the member states must at all costs <u>to</u> the requirements of environmental protection adapt ...</i> <i>The EU and the member states must at all costs adapt to the requirements of environment protection ...</i>
meaning 3: to employ	<i>Es gibt kleine wettbewerbsfähige Fluggesellschaften , die Personal übernehmen bzw. einstellen könnten , ...</i> <i>There are small competitive airlines that personnel take over, or employ could , ...</i> <i>There are small competitive airlines that could take over or employ personnel , ...</i>

Figure 4: Example sentences for re-ranking candidates for the ambiguous verb *einstellen* (*stop*, *adapt to*, *employ*).

responding to a more predictable sentence, we use the following formula to re-rank the synonym candidates:

$$p_{new}(syn) = p(syn) + \alpha * \frac{1}{ppl_{average}(syn)} \quad (2)$$

For finding an optimal weight coefficient α , all values between 0.001 and 10.0 were tested. The top 30 reordered candidates for each verb were evaluated for each possible value for α .

Figure 3 shows an example sentence containing the particle verb *klarmachen* (*to make clear*) and a variant where the particle verb is replaced with the valid synonym candidate *verdeutlichen* (*to illustrate*). In this case, re-ranking based on language model scores is likely to succeed as both verbs can occur with the same subcategorized elements (subject – indirect object – subordinated *dass*-clause). Similarly, most of the synonyms of *klarmachen* can be expected to function with such a subcategorization frame, which allows to substitute the verbs without introducing structural problems. In contrast, the examples in figure 4 illustrate two problematic aspects of this approach:

- Synonymous verbs can have different subcategorization frames, which can lead to low ratings even for valid synonym candidates.
- For particle verbs with different meanings, it cannot be guaranteed that the meaning of the verb in the sentence corresponds to the meaning of the synonym candidates.

The particle verb *einstellen* has several meanings; the sentences in figure 4 represent the 3 meanings occurring in the set of the 10 randomly chosen sentences: *to stop* (5 sentences), *to adapt to* (4 sentences) and *to employ* (1 sentence). Further possible meanings, e.g. [*Temperatur*] *einstellen* (*to set [the temperature]*), seem to be less predominant in our data set and do not occur in the set of 10 random sentences. Thus, when substituting the original particle verb with synonym candidates, it might happen that a

valid synonym ends up in a sentence with a different meaning of the verb, leading to a bad rating for that synonym. Furthermore, synonymous verbs can require different subcategorized components: while *einstellen* in the sense of *to stop* or *to hire* subcategorizes a subject and a direct object (*to hire personnel* vs. *to stop production*), the meaning of *to adapt to* requires a prepositional phrase with the head *auf*, as well as a reflexive pronoun (*sich*). A possible synonym for this meaning would be *anpassen* which also requires a reflexive pronoun and a prepositional phrase, but with a different head (*an*).

In contrast to Bannard and Callison-Burch (2005), who achieve improvements by ranking synonym candidates according to language model scores, we concentrate on verbs, rather than nominal or adjectival phrases. The previous analysis makes it clear that substituting verbs with synonyms is problematic, even if there are no word sense problems, due to different possible subcategorization frames.

5.2. Distributional similarity

As a second re-ranking feature, we used the distributional similarity between the particle verb and its synonym candidates. Here, we take the context within a given window as an indicator for the similarity of the particle verb and its synonym candidates, assuming that similar words share similar contexts. Distributional similarity is computed as the cosine similarity of the respective context vectors; the context is defined as content words (nouns, adjectives, verbs and adverbs) within a window of 10 words to each side, using local mutual information instead of co-occurrence frequencies extracted from a large Web corpus (cf. section 6.2). For re-ranking, the translation probabilities and cosine similarities were multiplied. In order to facilitate the computation and comparison of cosine similarity, the synonym candidates were restricted to single verbs for this re-ranking approach.

Table 3 shows the effect of re-ranking based on distributional similarity for the particle verb *zusammenkommen* (*to come together*). Except for *treffen* (*to meet*), which is acceptable (even though not in the gold standard), none of the

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

Table 3: 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.

top-ranked candidates found by the basic method relying on translational probabilities is synonymous with *zusammenkommen*. Instead, the found synonym candidates represent the meaning of *to correspond to* or *to fulfill*, which is possibly caused by confusing alignments of *zusammenkommen* \leftrightarrow *meet* and *meet* [*requirement*] \leftrightarrow *erfüllen*. As *zusammentreffen* and [*Bedingung*] *erfüllen* are not similar in terms of cosine, the previously top-ranked synonym candidates are moved down in the list, allowing valid synonyms to move towards the top of the list. Evaluating against the gold standard, there are now 4 matches after re-ranking, in contrast to no match at all when ranked only according to translation probabilities.

In contrast to language model scores, where the choice of the sentences largely affects the outcome (e.g. in the case of word sense mismatches or different subcategorization frames), contextual similarity provides a general assessment that is independent from specific contexts. By relying on lemmatized content words co-occurring with each instance of the particle verbs for computing contextual similarity, this approach yields a more robust and general estimation of similarity than the perplexity scores obtained by the language model rating, which largely depend on the set of rated sentences.

6. Experiments and evaluation

In this section, we give an overview of the experiments and the underlying data sets and pre-processing steps. As German is a morphologically rich language, we compare variants of simplifying the surface forms (both English and German) by lemmatization. To assess the quality of the obtained synonym candidates, we measure the precision of the top-ranked candidates against a gold standard. In addition, we also present a small-scale manual evaluation for a set of 14 particle verbs.

6.1. Creation of a gold standard

The synonym entries of the gold standard were looked up in the online synonym dictionary by Duden³. For the 500

³www.duden.de

EN		DE	
a	inflected	c	lemmatized part. verbs
b	lemmatized	d	lemmatized ADJ, V, N
		e	lemmatized

Table 4: Pre-processing variants for word alignments.

	Files		top 1	top 5	top 30
	A	In			
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 5: Precision for different pre-processing strategies. The 3 best systems are highlighted in each range; with *A* specifying the files used for alignment, and *In* specifying the input for synonym extraction.

most frequent particle verbs (freq \geq 15) in our corpus, we chose verbs with at least 30 synonym entries in Duden⁴. This restriction guarantees that a precision of 1 can be reached when evaluating the 30 top-ranked synonym candidates. In total, 138 particle verbs meet this condition. The listed synonyms are not only one-word entries, but also contain multi-word entries, such as *klar werden* (*to become apparent*) for the particle verb *herausstellen* (*to emerge*).

6.2. Data

We used the DE-EN version of Europarl⁵ (1.5M parallel sentences). Applying the reordering rules to the German part required parsing; we used BitPar (Schmid, 2004). Word alignment was computed using GIZA++. The English side was tagged with TreeTagger (Schmid, 1994); for the reordered German part, we used SMOR (Schmid et al., 2004) to obtain lemmatized forms. For the language model re-ranking, we used the (non-reordered) German part of the parallel data. Distributional similarity was computed based on the corpus SdeWaC (880M words, Faaß and Eckart (2013)).

An important factor for working with parallel data is the quality of the word alignment. As German is a morphologically rich language, we studied the effect of lemmatization as a pre-processing step on word alignment (see table 4 for possible combinations). The input file for the synonym extraction is always lemmatized: this ensures that inflected variants are represented as one synonym candidate.

6.3. Results and evaluation

Table 5 lists the results of the different alignment combinations: while there is some variation, the setting with

⁴Duden provides a grouping of the synonyms according to word senses. As we do not control for word senses in this work, we do not make use of this information.

⁵www.statmt.org/europarl

	top 1	top 5	top 30
no re-ranking	58.6956	44.0579	22.2946
language model	58.6956	44.0579	22.1980
distributional similarity	63.7681	49.7101	23.7198

Table 6: Results for the two re-ranking strategies for the best system (1) from table 5. (For systems 3 and 4, the distributional similarity-based reordering even resulted in a precision of 65.22.)

inflected English data and (partially) lemmatized German data for alignment leads to the overall best result. These results indicate that English inflection (essentially *number* on nouns and *third-person marking* on verbs) provides useful information for the alignment, in contrast to the morphologically more complex German (*number*, *gender*, *case*, *strong/weak inflection* on nominal phrases and richer *verbal inflection*), where (partially) lemmatized input to the alignment leads to better results.

Table 6 shows the results of the two re-ranking methods: While the language model re-ranking has only a marginal effect, the distributional similarity-based approach leads to a considerable increase in precision. This confirms the assumption from section 5 where we assumed that distributional similarity provides a more reliable general assessment of the quality of synonym candidates, whereas in the case of language model scores, the choice of sentences has a large impact on the outcome: replacing verbs might generally lead to problems (e.g. in terms of different subcategorization requirements), particularly when considering that we do not control for different word senses.

6.4. Manual evaluation

In addition to precision values obtained by comparison with a gold standard, we also carried out a manual evaluation. While the gold standard comes from a trusted lexical resource, we were interested to see how humans perceive the obtained synonym candidates.

Four German native-speakers (students of computational linguistics) were given a selection of 14 particle verbs and the respective lists of the 30 top-ranked synonym candidates. The particle verbs were chosen to cover the entire range of precision values (0% – 50%) in the previous gold standard evaluation. For each synonym candidate, the participants had to decide whether a synonym candidate was a valid synonym (assuming an appropriate context) or not. Those candidates which were considered valid synonyms by at least two evaluators were counted when calculating the overall precision for the manual evaluation. The results of this evaluation are presented in table 8. It can be seen that even if there is some variation between the annotators, they tend to consider more synonym candidates to be valid than the gold standard.

In an attempt to estimate the agreement of the annotators in the task of annotating the 14×30 synonym candidates, we used the following definition: if all or none of the evaluators considered a candidate a valid synonym, the agreement for the candidate is 100%. If three of the evaluators considered

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

Table 7: Individual annotation decisions in contrast to the gold standard for a subset of verb and synonym candidate pairs.

Verbs	P1	P2	P3	P4	Gold
aufbauen	46.67	36.67	53.33	46.67	50.00
einstellen	50.00	36.67	33.33	46.67	43.33
festlegen	50.00	26.67	23.33	46.67	36.67
einsetzen	40.00	26.67	6.67	40.00	33.33
umbringen	36.67	40.00	26.67	30.00	30.00
mitteilen	26.67	36.67	63.33	36.67	26.67
zusehen	46.67	20.00	43.33	36.67	26.67
darstellen	20.00	16.67	20.00	33.33	23.33
festhalten	33.33	16.67	10.00	26.67	23.33
aussetzen	36.67	3.33	10.00	10.00	16.67
aufnehmen	43.33	30.00	23.33	30.00	10.00
zusteuern	6.67	26.67	23.33	40.00	10.00
aufgehen	13.33	30.00	6.67	16.67	0.00
vornehmen	0.00	6.67	16.67	33.33	0.00
average	32.14	25.24	25.71	33.81	23.57

Table 8: This table shows the scores attributed to each verb by each of the four evaluators (considering the top 30 candidates), as well as the gold standard evaluation score on the right.

it a valid/invalid synonym, the agreement for the candidate is 75%. Otherwise the agreement is 50%. According to this method, the average agreement between the annotators was 82.9%.

While the annotators generally agree well, there are also pairs of verbs and synonym candidates for which it is more difficult to make a decision. Furthermore, there can also be a certain discrepancy with the gold standard, as illustrated in table 7. The fact that many of the verbs are highly ambiguous (e.g. for *einstellen* or *einsetzen*, 9 possible senses are listed in *Duden*) makes an evaluation even more difficult. In the end, the annotators have to rely on their own judgement and capability of finding a plausible context in which the synonym candidate could take the place of the original particle verb. However, this is often up to debate and also depends on personal preference.

7. Discussion

Our evaluations showed that the method of extracting synonyms for particle verbs by using back-translations of pivot translations leads to a respectable amount of valid synonyms. Re-ranking based on distributional similarity results in further improvement. However, during the evaluation, it also became clear that not controlling for word senses, and instead collecting all synonym candidates generated by all pivots into one set, is problematic.

In comparison to e.g. nominal phrases, (particle) verbs tend to have multiple senses more often than not. On average, the 138 particle verbs for which we extracted synonyms have 5.3 senses (according to Duden), ranging from 1 sense (4 verbs) to 14 senses (1 verb). While we obtained generally good results, both in terms of an automatic and a manual evaluation, taking into account the different word senses of particle verbs is necessary for two reasons. First, in possible application scenarios such as e.g. evaluation of MT, sets of synonyms grouped according to word senses are more useful than ungrouped sets of synonyms. Second, the word senses of a particle verb are often not evenly distributed; there is often a dominant sense accompanied by less frequent senses (cf. figure 2, which contains one very predominant pivot element compared with the rest of the pivots). In the process of (basic) synonym extraction, this can lead to a bias towards the most dominant meaning(s), as the less dominant senses are ranked lower due to smaller pivot and return probabilities. As a result, it might happen that less dominant word senses do not occur in the set of synonym candidates. Since the synonyms of the dominant senses are valid, we also cannot expect re-ranking based on distributional similarity to bring less dominant senses towards the top-n ranked candidates, as this re-ranking strategy will rather remove invalid candidates from the top of the list.

One possibility for word sense disambiguation is to use the pivot elements (cf. Bannard and Callison-Burch (2005)): considering the set of back translations obtained from one pivot as synonyms sharing one sense provides a good basis for handling word senses. In addition to grouping the obtained synonyms, this also allows to consider non-prominent pivots and their respective back translations, and thus ensures that all senses of the original particle verb occurring in the parallel data can be taken into account in the ranking step. As there are likely to be more pivots than word senses, a further step to cluster similar pivots is required.

Another problem is that of incorrect alignments: even though our pre-processing steps (lemmatization and re-ordering of German verbs) lead to a generally good alignment quality, low-frequency verbs in combination with bad word alignment can have a negative effect on the ranking of synonym candidates. Also, it can happen that parts of multi-word expressions are considered as pivots, which can lead to a meaning shift. For example, *meet* is a prominent pivot of *zusammentreffen* (to meet/gather). However, *meet* also frequently occurs in the context of *meet requirement*, which will generate invalid synonym candidates. Thus, heuristics for recognizing “appropriate” pivots for a given German verb could be helpful; this goes in line with the

idea of taking into account the individual pivots and the respective sets of generated back translations for the purpose of word sense handling.

8. Conclusion

We presented a method for the extraction of synonyms for German particle verbs using parallel data. In order to deal with separated particles, we applied reordering rules to the German part of the data. In our evaluation, we compared different pre-processing variants (with and without lemmatization); the best system has 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%). An additional manual evaluation of 14 particle verbs suggested that the precision obtained by comparison with the gold standard was slightly under-estimated. Furthermore, we discussed issues related to multiple word senses of the verb for which to extract compounds and outlined a method to control for word senses by grouping the synonym candidates according to the pivots.

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.

9. Acknowledgements

This work was 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).

10. References

- Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 597–604, Ann Arbour.
- Barzilay, R. and Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL*.
- Barzilay, R. and McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *In Proceedings of ACL*.
- Blondel, V. and Senellart, P. (2002). Automatic extraction of synonyms in a dictionary. In *Proceedings of TMW 2002*, Arlington, USA.
- Carbonell, J. G., Klein, S., Miller, D., Steinbaum, M., Grasziany, T., and Frey, J. (2006). Context-based Machine Translation. In *Proceedings of the Association for Machine Translation of the Americas*, pages 19–28, Boston, MA.
- Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*.
- Curran, J. (2003). *From Distributional to Semantic Similarity*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

- Dang, V., Xue, X., and Croft, W. B. (2009). Context-based quasi synonym extraction. In *CIIR Technical report*.
- Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceeding of the 40th Annual Meetings of the ACL*, Philadelphia, USA.
- Faaß, G. and Eckart, K. (2013). SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany.
- Fraser, A. (2009). Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of EACL WMT*.
- Lavie, A. and Denkowski, M. (2009). The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, 23:105–115.
- Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying Synonyms among Distributionally Similar Words. In *Proceedings of the International Conferences on Artificial Intelligence*, pages 1492–1493, Acapulco, Mexico.
- Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774, Montreal, Canada.
- Padó, S. and Lapata, M. (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.
- Sahlgren, M. (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University.
- Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of LREC 2004*.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of COLING*, Geneva, Switzerland.
- Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*.
- van der Plas, L. and Tiedemann, J. (2006). Finding Synonyms using Automatic Word Alignment and Measures of Distributional Similarity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 866–873, Sydney, Australia.
- Wang, W., Thomas, C., and Sheth, A. (2010). Pattern-based synonym and antonym extraction. In *Proceedings of ACMSE 2010*, Oxford (MS), USA.