

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Pfaffenwaldring 5B
D - 70569 Stuttgart

Bachelorarbeit

**Improving SMT-based Synonym
Extraction across Word Classes
by Distributional Reranking
of Synonyms and Hypernyms**

Maximilian Bräuninger

Studiengang: Informatik

Prüfer: Dr. Sabine Schulte im Walde

Betreuer: Dr. Sabine Schulte im Walde,
Marion Di Marco

begonnen am: 14.02.2017

beendet am: 14.08.2017

CR-Klassifikation: I.2.7

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben.

Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet.

Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens.

Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht.

Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

M. Bräuninger

Stuttgart, 10.08.2017

Declaration

I hereby declare that the work presented in this thesis is entirely my own.

I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations.

Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before.

The electronic copy is consistent with all submitted copies.

M. Bräuninger

Stuttgart, 10.08.2017

Contents

1	Introduction	5
2	Related Work	6
3	Data	7
3.1	Parallel Corpus	7
3.2	Target Sets	8
4	Statistical Machine Translation	10
4.1	Definition	10
4.2	Alignment	11
4.3	Fast Align	14
5	Synonym Extraction	16
5.1	Data Preprocessing	17
5.2	Dictionary Creation	18
5.3	Extraction	20
5.4	Probabilities and Ranking	20
6	Re-ranking	22
6.1	Weeds Precision	22
6.2	Cosine Similarity	23
7	Results	24
7.1	Gold Standard	25
7.2	Unfiltered	25

7.3	Re-Ranked	30
7.4	Particle Verbs	32
7.5	Manual Evaluation	33
8	Conclusion	35

1 Introduction

Finding synonyms is a very interesting and important task in natural language processing (NLP). Synonyms possess a number of different application areas. The creation of Thesauri (Lin et al. (2003)) springs to mind. Synonyms are also useful in automatic machine translation (Carbonell et al. (2006)). Furthermore they are not only useful in the creation but also in the evaluation process of automatic machine translations since they can help recognize different, equally correct, translations of a given word or sentence like for example "to drive" and "to ride" for the German word "fahren" (Lavie and Denkowski (2009)).

This thesis will focus on the extraction of synonyms for German nouns, adjectives and verbs. The presented method is based on Bannard and Callison-Burch (2005). The basic idea is to take a set of German words and first translate them into English pivots. In a second step the English pivots found in the earlier step are then re-translated into German words. The words retrieved in this second step are synonym candidates for the initial German word. During the steps a translation probability count is kept between the original word, the pivots and the synonym candidates. These probabilities are used to calculate a synonym probability in order to allow the candidates to be ranked. In a last step, the obtained synonym candidates are re-ranked using two different distributional semantic measures, one, using a new approach, by trying to filter out hypernyms, the other trying to measure the similarity of the synonym candidate and the target word. The results are then checked against a gold standard obtained by the German Dictionary's website *Duden*¹.

¹ <http://www.duden.de>

2 Related Work

The underlying idea in this thesis is mainly based on Bannard and Callison-Burch (2005). They were the first ones trying to extract synonym candidates using a multilingual parallel corpus and SMT-techniques. Their idea is to translate a target word into different pivots of another language and then re-translate those pivots into synonym candidates for the target word. Other papers dealing with this topic include Wittmann et al. (2014). They also use the approach of Bannard and Callison-Burch (2005) however they focus on particle verbs in particular. Furthermore they try to improve their results using different re-ranking strategies, which will also be done in this thesis. Other work regarding the automatic extraction of synonyms mainly focuses on using monolingual corpora. For example Barzilay and Lee (2003) use comparable monolingual corpora, in their case articles, written by different newswire agencies, about the same topic. Barzilay and McKeown (2001) use a monolingual parallel corpus, specifically various English translations of a foreign text.

Another field of research which is of interest for this thesis is distributional semantics, especially the recognition of hypernyms in contrast to synonyms. A lot of research has been done on the field of hypernym identification using distributional semantics. Lenci and Benotto (2012) have tested several directional similarity measures in order to identify hypernyms. These methods rely on the asymmetrical relationship between a hypernym and the hyponyms. The hypernym possesses an overall broader meaning semantically, which is not included in the hyponym. Among the methods investigated by them is the one used in this thesis which has been created by Weeds and Weir (2003), the so called "weedsPrecision value", which measures the inclusion of the features of one term t_1 within another term t_2 .

3 Data

In order to obtain satisfying synonym candidates it is important to use the right datasets, meaning a suitable multilingual parallel corpus and sets of target words. The goal of this thesis is the extraction of synonyms of a variety of different word classes (see 3.2). Since, in this case, the objective is to retrieve general language synonyms, the context, language and words used in the parallel corpus should preferably be very general and not use special terminology. If the context of a dataset is too specialized this could lead to a distortion of the extracted synonyms towards this topic. For example in a medical context the word "heart" will most likely be linked to an organ whereas in a lyrical context the same word might also be interpreted as a symbol of love. Clearly these two different interpretations will lead to different synonyms which is why the parallel corpus should be as general as possible in order to not alter the results.

3.1 Parallel Corpus

The bilingual parallel corpus used in this thesis consists of the Europarl corpus v7 (Koehn (2005)) extracted from the proceedings of the European Parliament. In its newest version the corpus consists of over 60 million words in each language. Besides the two languages, German and English, used in this case, the corpus also contains the translations of 19 other languages. (French, Italian, Spanish, Portuguese, Romanian, Dutch, Danish, Swedish, Bulgarian, Czech, Polish, Slovak, Slovene, Finnish, Hungarian, Estonian, Latvian, Lithuanian and Greek). Other parts of the corpus used in the thesis consist of the News Commentary v10 corpus and the Common Crawl Corpus, consisting of crawled webpages. The corpus meets the requirements as it offers bilingual sentence-aligned translations in German and English. Furthermore the nature of the European Parliament, news and webcrawling suggests that the texts used will not all be specialized toward one context. While overall

1 Jahr 179830	1 europäisch 290700	1 geben 167013
2 Kommission 161847	2 neu 155759	2 machen 99729
3 Herr 144770	3 groß 144774	3 finden 97587
4 Land 130437	4 gut 92548	4 sagen 89402
5 Parlament 130389	5 erst 79256	5 bieten 86139
6 Union 120576	6 wichtig 66910	6 gehen 83270
7 Hotel 119185	7 weit 64734	7 liegen 78866
8 Präsident 109126	8 hoch 63438	8 kommen 78818
9 Bericht 106145	9 politisch 62625	9 stehen 78717
10 Mitgliedstaat 90256	10 international 59974	10 stellen 76623
11 Frage 89323	11 verschieden 53141	11 sehen 69671
12 Rat 81073	12 eigen 52905	12 lassen 62835
13 Frau 81051	13 gemeinsam 49715	13 unterstützen 58236
14 Zeit 76726	14 letzt 48642	14 führen 56238
15 Problem 75327	15 klein 47823	15 bestehen 53661
16 Bereich 72860	16 nah 37663	16 tun 51852
17 Mensch 67432	17 national 36789	17 bringen 50660
18 Vorschlag 63138	18 sozial 36213	18 nehmen 50004
19 Entwicklung 62325	19 öffentlich 34849	19 erhalten 49753
20 Seite 58628	20 wirtschaftlich 33620	20 erreichen 49601

Figure 1: Top 20 Tokens nouns, adjectives, verbs from left to right

the political and European nature of part of the corpus' origin is quite obvious, for example "European", "political", "parliament" or "commission" are amongst the most common words, the overall bandwidth and comprehensiveness discussed in the European Parliament mixed with the other corpora leads to a rather mixed use of language. Therefore it can be concluded that overall the corpus represents a language general enough to achieve satisfying results. Furthermore the corpus offers the sentence alignment required to apply the SMT-methods required to create English pivots and German re-translations.

3.2 Target Sets

To extract synonym candidates, appropriate target word sets have to be created. To ensure precision comparability and evaluability several target sets are built. Each set contains a predetermined, fixed amount of words. The words in the target sets are the German target words which will later be used to extract synonym candidates. This thesis focuses on synonyms for German

nouns (e.g. "Tisch - table"), attributive adjectives (e.g. "das *große* Haus - the *big* house"), and full verbs (e.g. "komm! - come!", "gehen - to go", "wir *kommen* an - we are *arriving*") (Schiller et al. (1995)). Overall three target synonym sets are created with each set only containing either lemmatized nouns (NN), lemmatized attributive adjectives (ADJA) or lemmatized full verbs (VV). Each set contains 300 words evenly distributed between 100 high, 100 medium and 100 low frequency words regarding the number of their appearances in the original text. In this case words will be considered highly frequent if their token count is within the top 25% of the word class, medium appearing if their word count is in the top 50-25% and as low if the word count is below the top 50%. In order to avoid very uncommon or nonsensical words, the low frequency words are also required to appear at least twelve times in the text. Figure 1 shows the 20 tokens with highest frequency count found in the German Europarl corpus used to create the test sets.

As shown in Table 1 there are far more different nouns than full verbs or attributive adjectives with roughly 40000 eligible nouns and less than 25% of that amount for verbs and adjectives. Regarding the distribution though, there appear to be far more nouns with a lower token count thus lowering the 25% and 50% frequency threshold significantly in comparison to the verbs and adjectives. Another aspect regarded in the creation of the test sets is to ensure to have a reliable way to later compare the extracted synonym candidates. For the sake of achieving this comparability all words in the target sets are required to have at least a certain amount (one test set with at least two "Duden synonyms" and one with at least ten "Duden synonyms") of synonyms on *Duden*, the source used to find the gold standard synonyms the results will later be checked against in this thesis. The different test sets regarding the minimum amount of "Duden synonyms" are created in the hope of observing differences in the precision rate of the highest ranked synonym candidate when checked against the gold standard.

Overall this leads to six different test sets all containing 300 target words

	Token count	Token count >12	Max. Token count	>25% Threshold	>50% Threshold
NN	145,972	42,732	179,830	1,747	131
VV	10,026	5,708	167,013	6,266	775
ADJA	20,801	9,229	290,700	2,858	2,292

Table 1: The amount and distribution of the token counts of the German Europarl corpus

for synonym extraction ordered by word class to ensure comparability of the differences between word classes in synonym extraction.

4 Statistical Machine Translation

Since the basic idea of synonym extraction in this thesis requires a word to be translated from its source language (in this case German) into a target language (here English) and to then re-translate the pivots found in the first translation step, it is crucial to use a reliable and accurate method of translating words into different languages.

Already in 1949, Warren Weaver, an American mathematician, proposed to use cryptanalytic and statistical methods from communication theory in order to solve the problem of computer based text translation (Brown et al. (1993)). Warren’s idea proved to be too complex for the computational power of the 1950s and 60s but the significance of translation problems in NLP and the ongoing development of computer power lead to further research on the topic. One result of this research is Statistical Machine Translation (SMT), which treats translations as a machine learning problem using large parallel corpora (as the one described in chapter 3) to create translations (Lopez (2008)).

4.1 Definition

We are given a string f in the source language F with vocabulary V_F . We now transform f into another string e of the target language E with vocab-

Der Pirat hat den Schatz gefunden

The pirate has found the treasure

Figure 2: Translationally equivalent sentences in German and English

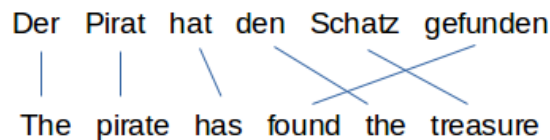


Figure 3: Translationally equivalent sentences in German and English with word alignment

ulary V_E . In order to be able to translate sentences from language F into language E our goal is to transform a string $f = f_1, f_2, f_3, \dots, f_m; f_x \in V_F$ into $e = e_1, e_2, e_3, \dots, e_l; e_x \in V_E$ with f and e being translationally equivalent (Lopez (2008)).

Figure 2 shows a German sentence and a translationally equal English sentence. As depicted in Figure 3 it is possible to form pairs of German and English words which are translations of each other. These words within the sentences are called *aligned* words. This allows us to break down translational equivalence into a number of smaller word equivalence problems (Lopez (2008); Koehn (2009)).

4.2 Alignment

Brown et al. (1993) introduced five different models (IBM Model 1-5), each of them assigning a probability to all of the possible word alignments between two translationally equivalent sentences. Equivalent to the description in section 4.1 the overall goal in this paper is to assign a probability to each

And the program has been implemented
Le programme a été mis en application

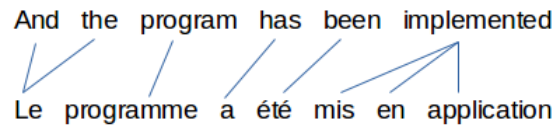
A word alignment diagram showing the mapping between English and French. The English sentence "And the program has been implemented" is aligned with the French sentence "Le programme a été mis en application". Blue arrows indicate the following alignments: "And" to "Le", "the" to "programme", "program" to "a", "has" to "été", "been" to "mis", and "implemented" to "en application".

Figure 4: Example of a word alignment with independent English words used by Brown et al. (1993)

The balance was the territory of the aboriginal people
Le reste appartenait aux autochtones

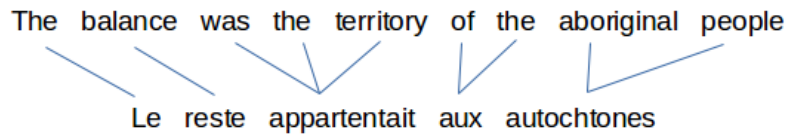
A word alignment diagram showing the mapping between English and French. The English sentence "The balance was the territory of the aboriginal people" is aligned with the French sentence "Le reste appartenait aux autochtones". Blue arrows indicate the following alignments: "The" to "Le", "balance" to "reste", "was" to "appartenait", "the" to "aux", "territory" to "autochtones", and "of the aboriginal people" to "autochtones".

Figure 5: Example of a word alignment with independent French words used by Brown et al. (1993)

The poor don't have any money
Les pauvres sont demunis

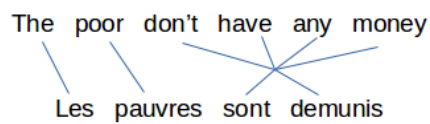
A word alignment diagram showing the mapping between English and French. The English sentence "The poor don't have any money" is aligned with the French sentence "Les pauvres sont demunis". Blue arrows indicate the following alignments: "The" to "Les", "poor" to "pauvres", "don't" to "sont", "have" to "demunis", and "any money" to "demunis".

Figure 6: Example of a general word alignment used by Brown et al. (1993)

pair of strings (f, e) consisting of a given string f in source language F (here French) and all possible strings e in target language E (here English). Brown et al. (1993) define $\Pr(e|f)$ as the likelihood of e being a correct translation of f . Now the goal is to find \hat{e} with $\Pr(\hat{e}|f) = \max$. Using Bayes' Theorem:

$$\Pr(e|f) = \frac{\Pr(e) \cdot \Pr(f|e)}{\Pr(f)}$$

Since $\Pr(f)$ is independent of e in order to maximize $\Pr(e|f)$ it is sufficient to maximize the numerator of the fraction to find \hat{e} :

$$\hat{e} = \operatorname{argmax}_e \Pr(e) \cdot \Pr(f|e)$$

Brown et al. (1993) state that computing the *language model probability* $\Pr(e)$ has been dealt with in other contexts (for example by Maltese and Mancini (1992)) so they focus on computing $\Pr(f|e)$ the so called *translation model probability*. To calculate this probability word alignments as described in section 4.1 are used. Brown et al. (1993) describe three different possible alignment types. Type one is shown in Figure 4. Here every word in f of the source language French is connected to exactly one word in e of the target language English. In Figure 5 at least one word in f is linked to more than one word in e . Figure 6 shows a more general case in which several words in f are connected to several words in e .

The amount of possible alignments sums up as follows. With $|f| = m, |e| = l$ there are lm possible connections to be drawn between words in f and in e . Overall this leads to 2^{lm} different possible alignments of $(f|e) = A(e, f)$. Brown et al. (1993) describe five different models, each capable of computing $\Pr(f|e)$ more or less precisely. The underlying idea of each of the five models is to calculate $\Pr(f|e)$ as the sum of conditional probabilities $\Pr(f, a|e)$

$$\Pr(f|e) = \sum_{a \in A(e, f)} \Pr(f, a|e)$$

Under the restriction of alignments only in the form depicted in Figure 4, where each French word is linked to either exactly no or exactly one English

Parallel Corpus	#Tokens	IBM Model 4	Log-linear
Chinese-English	17.6M	2.7	0.2
French-English	117M	17.2	1.7
Arabic-English	368M	63.2	6.0

Table 2: The time required (hours) to train alignment models in one direction according to Dyer et al. (2013)

word the alignment a between $f_1^m = f_1, f_2, f_3, \dots, f_m$ and $e_1^l = e_1, e_2, e_3, \dots, e_l$ can be described as $a_1^m = a_1, a_2, a_3, \dots, a_m$; $a_x \in (0, l), a_x \in \mathbb{N}$. In this case for a value a_i , i represents the position in f , and a_i represents the position the word is aligned to in e . If $a_i = 0$ the word is not aligned to any word in e at all. Now, without loss of generality a possible representation of $\Pr(f, a|e)$ is (Brown et al. (1993)):

$$\Pr(f, a|e) = \Pr(m|e) \prod_{j=1}^m \Pr(a_j | a_1^{j-1}, f_1^{j-1}, m, e) \Pr(f_j | a_1^{j-1}, f_1^{j-1}, m, e)$$

All five models have different calculation processes of $\Pr(f, a|e)$ each factoring in different factors within f and e . Models 1 and 2 both assume the length of the French string m to be equally distributed between all reasonable lengths regarding e and l . Contrary to Model 1, Model 2 assumes the connection probability between two words in f, e to depend on the positions of the words within the strings and l, m . This leads to $\Pr(f|e)$ being dependent on the word order in f and e (Brown et al. (1993)). Models 3, 4 and 5 on the other hand try to estimate m by choosing the number of French words each English word in e will be connected to. The alignment tool used in this thesis is based on Model 2.

4.3 Fast Align

In search of a simple, effective and well scaling word alignment model Dyer et al. (2013) have presented their version of Brown et al. (1993)'s IBM Model

Model	FR-EN	ZH-EN
IBM-Model 1 EM	29.0	56.2
IBM-Model 2 EM	21.4	53.3
IBM-Model 4 EM	10.4	46.5
log-linear EM	18.5	45.4
log-linear \sim Dir	16.6	44.1
IBM-Model 1 \sim Dir	26.6	53.6

Table 3: The alignment error rate (AER) as described by Dyer et al. (2013). Lowest is best. EM standing for expectation maximization, \sim Dir for variational Bayes

Parallel Corpus	IBM Model 4	Log-linear
Chinese-English	34.1	34.7
French-English	27.4	27.7
Arabic-English	54.5	55.7

Table 4: Translation quality (BLEU) according to Dyer et al. (2013). Highest is best.

2. Model 2 was chosen since both IBM Model 1 and 2 support exact inference in $\Theta(|f| \cdot |e|)$ (Dyer et al. (2013)). Thus both Models are still widely used in task such as rapid large scale experimentation or parallel data mining. According to Dyer et al. (2013) both IBM Models are suboptimal. As mentioned in section 4.2 Model 1 does not pay attention to the word order in f, e which proves to be a problematic assumption. Model 2 on the other hand factors in the alignment structures but is overparameterized, which leads to overfitting. By creating a simple log-linear reparameterization of IBM Model 2 (Dyer et al. (2013)) that outperforms the more sophisticated IBM Model 4 on three large-scale translation tasks, while training the model is consistently ten times faster, Dyer et al. (2013) have created a highly potent and useful tool for a lot of different word alignment tasks, like this one. As depicted

in Table 2 the time required to train the alignment models in one direction is roughly 10% of the time taken by IBM Model 4. Furthermore Dyer et al. (2013) use different standards to measure the quality of their model. Table 3 shows the alignment error rate, a combination of precision and recall, requiring a perfect alignment to possess all of the "required" alignments while perhaps containing some of the "possible" ones (Koehn (2009), Mihalcea and Pedersen (2003)). Here IBM Model 4 and the log-linear Model perform quite equally with one test-set and evaluation method (French-English, expectation maximization) favoring Model 4 and the other (Chinese-English, variational Bayes) favoring the log-linear Model. Another metric used in order to measure the translation quality is the bilingual evaluation understudy (BLEU). BLEU compares the machine's output to that of a human. The closer the mechanical output, the higher its quality (Papineni et al. (2002)). BLEU uses a modified precision value, limiting the maximum amount of appearances of a word in the translation candidate to the number of appearances in the reference translation. Table 4 shows how Dyer et al. (2013)'s log-linear Model outscores Brown et al. (1993)'s Model 4 in all three test cases.

5 Synonym Extraction

As already mentioned the essential idea of this thesis is based on the method described by Bannard and Callison-Burch (2005). The first step is to take the translations, suggested by word alignment, of a German target word. The English translations of this word act as pivots. In the second step the pivots found are re-translated to German via word alignment. The synonym candidate set of the initial word now consists of all the re-translations gathered in the second step. This process is illustrated in Figure 7. In this example the German verb "essen" ("to eat") is translated into three English pivots, "to eat", "to dine" and "to consume". The re-translations of these pivots now lead to eight different German synonym candidates for the word "essen". In a later step, these synonym candidates will be ranked according to a synonym

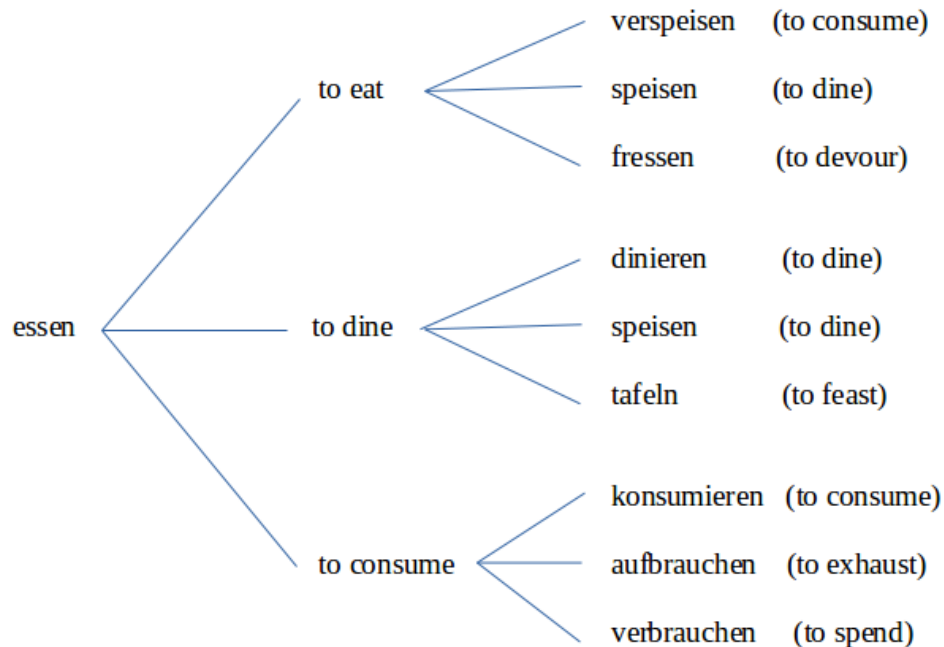


Figure 7: Illustration on how to obtain synonym candidates via translation and re-translation.

probability (see section 5.4).

5.1 Data Preprocessing

Using the word-alignment tools described in section 4 it is important to preprocess the input data in order to improve the alignment quality. Wittmann et al. (2014) propose to preprocess the input data in order to achieve the highest precision rates for synonym candidates. The best combinations achieved by Wittmann et al. (2014) consist of (partially) lemmatized German words and inflected English words. They conclude that in English the information, for example number on nouns and third-person marking on verbs, provided by inflection are useful for the overall quality of the alignment. This is not the

san francisco - es sein noch nie leicht , eine rational gespräch über die wert von gold zu führen .

san francisco - it has never been easy to have a rational conversation about the value of gold .

Figure 8: German and its corresponding English sentence taken from the Europarl data set.

case for the morphologically more complex German with information such as number, gender, case, strong/weak inflection on nominal phrases and richer verbal inflection (Wittmann et al. (2014)). Since they are only interested in the extraction of synonyms for particle verbs they have also tested the effect of only lemmatizing the German particle verbs, especially trying to combine multi-word particle verbs into single word particle verbs, with good results. However since this thesis focuses on a broader spectrum of words the whole German part of the parallel corpus is lemmatized.

The tool used for lemmatization and tagging is the tree tagger. The tree tagger is a tagger based on decision trees rather than on Markov models, leading to very exact results (Schmid (1994), Schmid (1999)).

5.2 Dictionary Creation

At first, two word dictionaries, German-English and English-German, are created using the alignment file output by the fast align tool. The alignment file consists of rows of data pairs. The i -th row contains the word alignment pairs for the i -th sentence in the source and target language. Figure 8 depicts the 13th sentences of the parallel corpus. Each word alignment is represented by two numbers $(n - m)$, with n indicating the n -th word in the sentence of the source language and m the m -th word in the sentence of the target language. An example is depicted in Figure 9. Figure 10 illustrates the word alignments shown in Figure 8 and Figure 9.

The creation of the German-English dictionary proceeds as follows. First,

0-0 1-1 2-2 3-3 6-4 6-5 6-6 7-7 8-8 6-9 9-10 10-11 11-12 12-13 13-14 14-15 15-16 16-17 19-18

Figure 9: The alignment created for Figure 8 by fast-align.

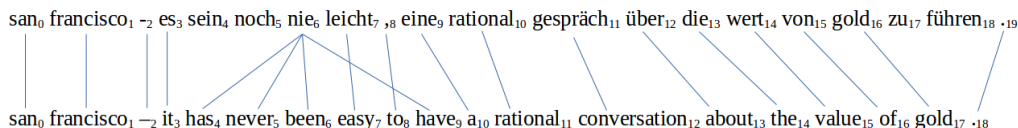


Figure 10: Illustration of the alignment shown in Figure 9.

all words within the German part of the text are linked to their appearances within each sentence. For example if the string s appears a total amount of four times in the parallel corpus in sentences i, j, k, l and within each of these four sentences s is the (zero based) s_i, s_j, s_k, s_l word then $s \rightarrow \{(i, s_i), (j, s_j), (k, s_k), (l, s_l)\}$.

Now, using the word appearance map $s \rightarrow \{(x_1, s_{x_1}), (x_2, s_{x_2}), \dots, (x_n, s_{x_n})\}$ the rows x_1, x_2, \dots, x_n of the alignment file are searched for pairs $(s_{x_k} - m)$. These pairs indicate that the m -th word in the English sentence x_k is a translation, or part of a translation of s . To prevent nonsense or overly long translations the amount of appearances of s_{x_k} within the alignment file in row x_k is limited to a maximum of three in the German-English dictionary. In the English-German dictionary, in order to only obtain single word synonym candidates, the maximum is limited to one. Furthermore a given set of stop words is filtered out of the dictionary in this step. The English-German dictionary is created in the same fashion. This now leaves us with two dictionaries each containing all German or English words within the Europarl dataset and their respective translations into the other language according to the word alignment. A depiction of the whole process is shown in Figure 11.

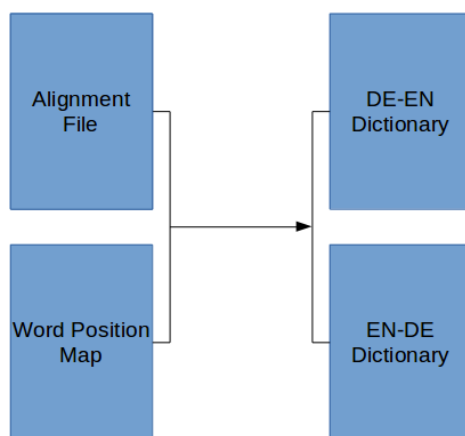


Figure 11: Illustration of the dictionary creation process.

5.3 Extraction

In order to keep the file size within reasonable boundaries (about 1 GB) both dictionaries are not combined into one large German-German "synonym candidate dictionary" but are kept apart and then used to extract only the synonym candidates required. As already mentioned, overall this leads to six different extraction processes due to the six target lists created. The extraction process works in a linear fashion. First the target list is read. Then, looking at the German-English dictionary all English translations of the words are gathered (see pivots). In a second step, the pivots are inserted into the English-German dictionary. The respective translations are now mapped to the original target words, resulting in a (long) list of synonym candidates.

5.4 Probabilities and Ranking

Being presented a large amount of synonym candidates for each target word it is necessary to rank them according to their likelihood of actually being

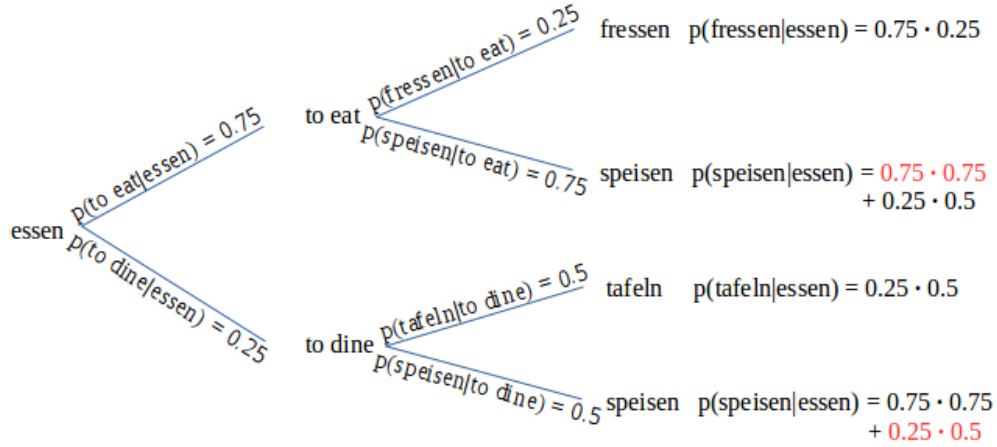


Figure 12: (Partial) extension of Figure 7 showing the calculation of the synonym probability. If a candidate word appears more than once, its part of the whole synonym probability sum is highlighted in red.

a valid synonym. Using the same method as Bannard and Callison-Burch (2005) the synonym probability $p(e_2|e_1)$, $e_1 \neq e_2$ of a synonym candidate e_2 given a target word e_1 is defined as follows.

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

With f_i representing an English pivot. Therefore the synonym probability is the sum over all pivots f_1, f_2, \dots, f_n with each summand consisting of two probabilities. The first one is the pivot probability $p(f_i|e_1)$ representing the likelihood of a pivot f_i being a translation of the German target word e_1 . The second probability is the so called return probability $p(e_2|f_i)$, describing the chance of the German phrase e_2 being a translation of the English pivot f_i . To calculate the translation probabilities relative frequencies obtained from the word alignment within the parallel corpus are used. An example illustrating the process of calculating the synonym probability is shown in Figure 12. Here the translation probabilities of the English pivots "to eat"

and "to dine" of the German word "essen (to eat)" are multiplied with the re-translation probabilities of their respective re-translations into German. The obtained values of same words, in this case of "speisen (to dine)" are added. This leads to three synonym probability values for the three synonym candidates.

6 Re-ranking

In the hope of improving the synonym extraction two re-ranking methods are implemented and investigated. Both re-ranking features rely on a vector space model in order to calculate the connections between the target words and their synonym candidates. In this case the data needed to create a reliable vector space model is taken from the DECOW 14ax web corpus (Schaefer and Bildhauer (2012), Schaefer (2015)). This corpus consists of over 11,660,000,000 tokens extracted mainly from websites with top-level domains de, at or ch. If a target word is not contained within the DECOW corpus, the word will be dropped and not be taken into consideration when calculating the precision values. If however a synonym candidate does not appear in the corpus, the corresponding target word will still remain in the calculation, but the synonym candidate will be dropped from the candidate list.

6.1 Weeds Precision

The first re-ranking feature relies on identifying hypernyms and ranking them lower in the candidates list, since hypernyms obviously are no synonyms and are regularly produced as synonym candidates using SMT-based synonym extraction methods. The hypernym identification method used here relies on using a directional (or asymmetrical) similarity measure, since hypernymy is an asymmetrical relation between words, with the hypernym being semantically broader than its hyponym (Lenci and Benotto (2012)). The directional similarity measure used in this thesis is called "weedsPrecision" (Weeds and

Weir (2003), Weeds et al. (2004) and Kotlerman et al. (2010)). In order to calculate "weedsPrecision" we are given two terms u, v . F_u, F_v are the sets of distributional features in their vector space model representation with $w_u(f_u), w_v(f_v)$; $f_u \in F_u, f_v \in F_v$ being the weights of the features. The precision is now defined as follows:

$$\text{weedsPrecision}(u, v) = \frac{\sum_{f \in (F_u \cap F_v)} w_u(f)}{\sum_{f \in F_u} w_u(f)}$$

Since the goal is to decrease the likelihood of a hypernym being regarded as a synonym candidate, the list of synonym candidates is re-ranked in the following way. The top 100 synonym candidates according to synonym probability are extracted as described in section 5. Each synonym probability value $p(e_2|e_1)$ of a pair of target word e_1 and synonym candidate e_2 is now multiplied by $1 - \text{weedsPrec}(e_1, e_2)$ which leads to a new synonym probability.

$$p_{\text{weeds}}(e_2, e_1) = p(e_2|e_1) \cdot (1 - \text{weedsPrec}(e_1, e_2))$$

For example, looking again at Figure 12, if $\text{weedsPrec}(\text{speisen}, \text{essen}) = 0.8$ and $\text{weedsPrec}(\text{fressen}, \text{essen}) = 0.1$ the synonym probability $p(\text{speisen}|\text{essen}) = 0.75 \cdot 0.75 + 0.5 \cdot 0.25 = 0.6875$ will now be changed to $p_{\text{weeds}}(\text{speisen}, \text{essen}) = 0.6875 \cdot (1 - 0.8) = 0.1375$ whereas $p_{\text{weeds}}(\text{fressen}, \text{essen}) = 0.75 \cdot 0.25 \cdot (1 - 0.1) = 0.16875$. This would then result in an actual re-ranking of the synonym candidate list, with "fressen" now being a higher ranked synonym candidate than "speisen".

6.2 Cosine Similarity

The second re-ranking feature implemented tries to re-rank the list according to the similarity of the two vectors representing a pair of target word and synonym candidate. This is possible due to the high semantical similarity between synonyms, resulting in high similarity of the vectors. To calculate the similarity of the two vectors the angle between both vectors using the cosine similarity value is calculated. Two terms, represented by their vectors,

are considered as equal if their angle is close to 0° , and thus their cosine value is close to one. On the other hand two vectors are considered semantically different if their angle is close to 90° , leading to a cosine value close to zero. The cosine value of two terms u, v with their sets of distributional features F_u, F_v and their weights $w_u(f_u), w_v(f_v)$, $f_u \in F_u, f_v \in F_v$ is calculated as follows:

$$\text{cosineSimilarity}(u, v) = \frac{\sum_{f \in (F_u \cap F_v)} w_u(f) \cdot w_v(f)}{\sqrt{\sum_{f \in F_u} w_u(f)^2} \cdot \sqrt{\sum_{f \in F_v} w_v(f)^2}}$$

In this case the goal is to rank synonym candidates higher, if their "cosineSimilarity" value with the target word is high. Therefore the synonym probability value is multiplied by the cosine similarity value, leaving us with the new synonym probability:

$$p_{\text{cosine}}(e_2, e_1) = p(e_2|e_1) \cdot \text{cosineSimilarity}(e_1, e_2)$$

Again, the top 100 candidates are re-ranked in the following manner. Given a target word e_1 and a synonym candidate e_2 with $\text{cosineSimilarity}(e_1, e_2)$ and a synonym probability $p(e_2|e_1)$ the new probability calculates as follows: $p_{\text{cosine}}(e_2, e_1) = p(e_2|e_1) \cdot \text{cosineSimilarity}(e_1, e_2)$. Exemplary (see Figure 12 again), for $\text{cosineValue}(\text{essen}, \text{speisen}) = 0.9$ and $\text{cosineValue}(\text{essen}, \text{fressen}) = 0.7$ both original synonym probabilities would be changed in the following way. $p_{\text{cosineSimilarity}}(\text{speisen}, \text{essen}) = 0.6875 \cdot 0.9 = 0.61875$ and $p_{\text{cosineSimilarity}}(\text{fressen}, \text{essen}) = 0.75 \cdot 0.25 \cdot 0.7 = 0.13125$. In this case "speisen" still would be the higher ranked synonym candidate compared to "fressen".

7 Results

The results, both unfiltered and re-ranked, are investigated regarding the precision of valid synonyms extracted (see section 7.1).

	2 Synonyms	10 Synonyms				
	Precision at 1	Precision at 1	Precision at 5	Precision at 10	highest Precision at 5	highest Precision at 10
ADJA	57%	62%	42%	33%	100%	90%
NN	37%	48%	33%	25%	100%	80%
VV	44%	44%	32%	25%	100%	90%

Table 5: Average precision rates of the different word categories when compared to the gold standard.

7.1 Gold Standard

In order to calculate the precision values a reliable gold standard containing reasonable synonyms for the selected target words is required. As already mentioned, in this case the online dictionary of *Duden* is used to determine the validity of a given synonym candidate. A candidate will be regarded as a valid synonym only if it also appears in the synonym section of the target word at *Duden*. Furthermore the gold standard is already used in the creation process of the target lists (see section 3.2) since valid target words within each test set are required to contain at least two respectively ten synonyms in the gold standard.

7.2 Unfiltered

Table 5 shows the average precision values when comparing the top unfiltered synonym candidates. To ensure a sufficient number of (gold standard) "Duden synonyms" is available for each target word, a distinction is made between the target sets requiring a minimum of two "Duden synonyms" and the ones containing target words with at least ten "Duden synonyms". Target words contained in the minimum two "Duden synonym" sets are only checked for the precision at one. This means only the highest rated synonym candidate (with regard to synonym probability) is taken into account. This leads to a precision value of either zero, if the top candidate is not contained in the gold standard or one, if the candidate is in the gold standard. Target words contained in the at least ten "Duden synonyms" sets are furthermore

	2 Synonyms	10 Synonyms		
	Precision at 1	Precision at 1	Precision at 5	Precision at 10
ADJA low	51%	56%	38%	27%
ADJA medium	62%	71%	45%	36%
ADJA high	57%	59%	45%	37%
NN low	24%	40%	22%	15%
NN medium	40%	52%	36%	27%
NN high	46%	53%	40%	32%
VV low	33%	37%	20%	16%
VV medium	54%	49%	34%	26%
VV high	44%	50%	41%	33%

Table 6: Average precision rates of the different word categories, subdivided by the different frequencies of the target words. Highest values in red.

checked for their precision at five and precision at ten. The overall average precision values for the two "Duden synonym" target sets range from 37% for nouns (NN) over 44% for full verbs (VV) to 57% for attributive adjectives (ADJA). When comparing these values to the precision at one values of the target sets with at least ten "Duden Synonyms" the values for ADJA (62%) and VV (44%) have a rather small difference within a range of 5 percentage points. However the difference between both NN values (49% for the "at least ten set") is larger with 11 percentage points. Overall the synonym extraction seems to perform slightly better with words that have a higher overall synonym count in the gold standard. This could be caused by the simple fact that words that contain more synonyms in the gold standard are words with an overall broader meaning and are semantically more likely to possess synonyms. Furthermore the sheer chance of hitting one of at least ten words is significantly higher than that of hitting one of at least two words.

When looking at the precision at five and precision at ten values the precision rates take a drop, however ADJAs still have the highest precision rate (42% at five, 33% at ten) whereas both NN and VV have lower rates (NN 33%, VV 32% at five, both 25% at 10), indicating that only one fourth of

the ten highest ranked synonym candidates are actual synonyms, according to *Duden*. Overall the results suggest that without any re-ranking strategies the extraction of attributive adjectives is more accurate than the extraction of nouns or full verbs.

A further precision rate shown in Table 5 is the highest precision rate achieved for one target word at five and ten synonyms respectively. While for five synonyms each word category has at least one target word with a precision rate of 100%, for ten synonyms no category reaches this value. Both ADJA and VV have a maximum precision rate of 90%, meaning one of the ten highest ranked synonym candidates according to synonyms probability is not a synonym regarding the gold standard. NN has a maximum rate of 80% therefore two synonym candidates are invalid synonyms according to *Duden*.

Table 6 shows the same precision values as described above, however each word category is divided into three different categories. Each category consists only of words that appear with a certain frequency within the text (see section 3.2). The highest precision rate for each word category is highlighted in the table. When having a closer look at the differences in precision rates with regard to word frequency it seems like words that have a medium (top 25-50%) or high frequency (top 25%) perform significantly better than words that appear with a low frequency (<50%) within the corpus. This difference could be caused by the quality of the alignment and therefore the translation and re-translation process which is improved by a higher amount of appearances. However the quality is not improved after a certain amount of appearances, which is why there seems to be no difference between high and medium frequency precision values. Especially the precision values of nominal nouns with a low frequency appear to be lower than expected. This could be caused by the fact that overall there is a high amount of nouns with low word counts thus lowering the <50% threshold significantly in comparison to verbs and adjectives (see Table 1).

After manually evaluating the negative results there appear to be five main

reasons that cause the extraction method to fail.

1. The ambiguous meaning of an English pivot.

The first and probably most common error produced by the synonym extraction appears to be connected to the ambiguous nature of some of the English pivots when being re-translated into German as depicted in Figure 13.

2. Partial translations of compounds.

The second common mistake is illustrated in Figure 14. This kind of error is caused by only partial translations of German compound words.

3. The context of the corpus leads to very specialized translations and re-translations.

The third mistake is based on the context of the parallel corpus. As already mentioned the goal is to extract general synonyms thus the gold standard also contains general synonyms for the target words. However the political domain of the corpus sometimes leads to synonym candidates within a political or law context. This is shown in Figure 15.

4. Words can be considered as synonyms, however they do not appear in the gold standard.

The fourth mistake observed regularly is the simple possibility of a synonym candidate being a valid synonym of the target word, however the word is not contained in the gold standard. Section 7.5 will cover this problem more closely.

5. Words that share a relationship with the target words however they are no synonyms (e.g. hypernyms, antonyms).

The fifth and final mistake shown in Figure 17 is created by synonym candidates that share a semantical relation with the target word, however they are hypernyms, antonyms etc.

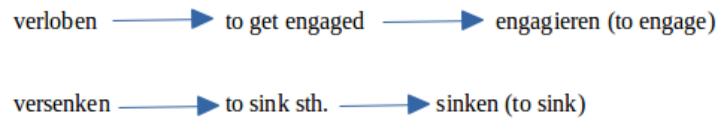


Figure 13: False synonyms created by the ambiguity of the English pivots.



Figure 14: False synonyms created by partial translations of compound words.



Figure 15: False synonyms created by a too specialized context of the corpus.

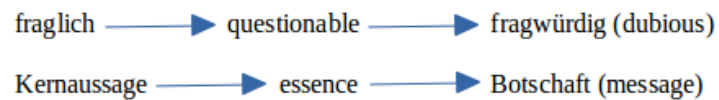


Figure 16: False synonyms created by an incomplete gold standard.

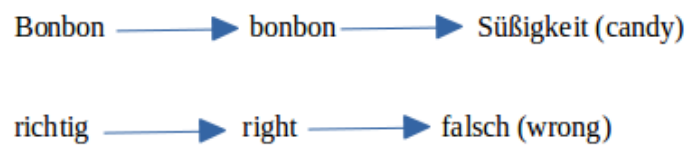


Figure 17: False synonyms created by a wrong semantic relation. The top example shows a hypernym, the bottom one an antonym.

		2 Synonyms	10 Synonyms				
		Precision at 1	Precision at 1	Precision at 5	Precision at 10	highest Precision at 5	highest Precision at 10
ADJA	weeds	45%	46%	37%	32%	100%	90%
	unranked	57%	62%	42%	33%	100%	90%
NN	weeds	22%	30%	24%	21%	100%	90%
	unranked	37%	48%	33%	25%	100%	80%
VV	weeds	33%	32%	26%	23%	80%	80%
	unranked	44%	44%	32%	25%	100%	90%

Table 7: Average precision rates of the different word categories after re-ranking using "weedsPrecision value" compared to the unranked values.

7.3 Re-Ranked

In this section, the re-ranked results, both using "weedsPrecision" and "cosineSimilarity", are investigated similar to the unfiltered results. No target words had to be filtered out since all of them appeared in the DECOW corpus enabling the calculation of the required re-ranking values.

Table 7 shows the re-ranked precision using the "weedsPrecision values". Obviously, re-ranking the synonym candidates did not have the desired effect of improving the precision values for the synonym candidates. Even worse, every single average precision value is significantly lower than its unranked counterpart. Especially looking at the precision at one and at five values, each is at least ten percentage points lower. When looking at the precision at ten values, they are quite similar to the unranked values. This could indicate that re-ranking using the "weedsPrecision" value mainly re-orders the highest ranked words however leaving them close within the top ten range, thus not changing this precision value as drastically.

When manually evaluating the "weedsPrecision values" there appear to be two problems that lead to the unsatisfying results.

1. Valid synonyms achieve very high "weedsPrecision values".
2. Synonym candidates do not appear in the DECOW corpus and thus are dropped from the candidate list.

Likely the first problem is the main issue with this method. When looking at

		2 Synonyms	10 Synonyms				
		Precision at 1	Precision at 1	Precision at 5	Precision at 10	highest Precision at 5	highest Precision at 10
ADJA	cosine	56%	62%	42%	34%	100%	90%
	unranked	57%	62%	42%	33%	100%	90%
NN	cosine	35%	49%	32%	25%	100%	90%
	unranked	37%	48%	33%	25%	100%	80%
VV	cosine	40%	41%	30%	24%	100%	80%
	unranked	44%	44%	32%	25%	100%	90%

Table 8: Average precision rates of the different word categories after re-ranking using "cosineSimilarity value" compared to the unranked values.

the data, rather often valid synonyms appear to score very high "weedsPrecision values". This fact leads to valid synonym candidates being down-ranked and dropping out of the top one, top five or even top ten spot in the synonym candidate list. To give an example, the target word "hinreißend (gorgeous)" has, according to *Duden*, the synonym "atemberaubend (breath taking)". This word happens to be in the top five of the unranked synonym candidate list. However it scores a "weedsPrecision value" of over 0.9. This results in the word "atemberaubend" dropping from the third highest ranked synonym candidate to the ninth highest, therefore dropping out of the top five. The second problem appears to happen mainly to composite words since most often these are the words that do not appear in the DECOW corpus. Since each of these words is filtered out some promising candidates drop out. For example the target word "Bonbon (bonbon)" contains the synonym candidate "Hartkaramelle (hard caramel)". However "Hartkaramelle" does not appear within the corpus and thus is erased from the list.

Table 8 shows the average precision values produced by re-ranking the synonym candidates using "cosineSimilarity values". When comparing these values to the unranked ones they appear to be quite similar. Most values have dropped by a few percentage points, while one has improved one point. This indicates that re-ranking the list with "cosineSimilarity values" does not seem to have a big impact on the overall precision. This could be caused by a lack of difference in the "cosineSimilarity values" compared to the corresponding synonym probability. While most high ranked (according to synonym probability) synonym candidates reach "cosineSimilarity values" of over 0.5 their

	Precision at 1	Precision at 5	Precision at 10	highest Precision at 5	highest Precision at 10
VV	44%	32%	25%	100%	90%
particle Verbs	57%	44%	35%	100%	80%

Table 9: Comparison of the full verb target set used in this paper and the target set of particle verbs used by Wittmann et al. (2014)

synonym probability values often differ by the factor ten or even more. This leads to less actual re-ranks than expected. Only some synonym candidates with a very low "cosineSimilarity value" drop out of the top one, top five or top ten. However the negative side effects of viable solutions dropping out due to their non appearance in the corpus described above also occurs here since the same corpus is used to create the vector space model. Overall this leads to results similar to the original one.

7.4 Particle Verbs

In this section the particle verbs used by Wittmann et al. (2014) are used to extract further synonym candidates. By doing so, possibly a statement can be made about their theory. These particle verbs consist of the top 500 particle verbs of the DE-EN Europarl corpus regarding their frequency (minimum frequency 15). Out of those particle verbs the ones with at least 30 synonyms in *Duden* are taken into account. Overall this leads to 138 particle verbs. In their paper, they conclude that extracting synonyms for particle verbs is especially difficult since particle verbs often possess different meanings. Table 9 shows the average precision of the full verb target list with at least ten synonym candidates previously used in this thesis in the first row. The second row shows the average precision values achieved by the 138 particle verbs of Wittmann et al. (2014). Comparing these values the particle verbs outscore the regular full verbs by over ten percentage points in precision at one, at five and at ten. The highest precision values are rather similar, with highest precision at ten being the only category where the full verbs have a higher precision rate than the particle verbs.

Looking at these results does not support Wittmann et al. (2014)'s theory. However the circumstances under which both target lists have been created are different. While the list created in this thesis only requires a target word to have at least ten synonyms, the particle verbs are required to have at least 30. This difference could cause the discrepancy in the precision values, since words with a high amount of synonyms are, mostly, common words, which as described in section 7.2 improves the quality of the alignment and therefore the overall quality of synonym extraction. Furthermore, as also described in section 7.2, the sheer amount of available synonyms increases the likelihood of a synonym candidate being valid.

7.5 Manual Evaluation

A possible problem mentioned in section 7.2 is connected to the nature of the gold standard. Only words listed in the synonym section of *Duden* are considered actual synonyms. However due to the multifaceted nature of semantics overall and synonyms in particular even a reliable source like *Duden* can not list every valid synonym for a certain word. Therefore in order to investigate the degree of this problem four native German speakers were asked to manually evaluate a part of the retrieved results of this thesis. Out of every test set, 25 synonym - synonym candidate pairs that are regarded as wrong were chosen at random, equaling 150 overall pairs. Then the four native speakers were asked to annotate whether, in their opinion, the pair is actually a non-synonym pair or whether it consists of two synonyms. The annotators were told to consider the different meanings a word can have regarding the context it is used in. Table 10 shows the percentage rate each annotator decided to rate a pair of synonym - synonym candidate as valid. The lowest rate is 19% by annotator 2, the highest is 38% by annotator 4. This leads to an overall average of roughly 30%, meaning one third of the pairs not considered valid by the gold standard were annotated to be a synonym by at least one of four native German speakers.

Annotator 1	Annotator 2	Annotator 3	Annotator 4
35%	27%	19%	38%

Table 10: Percentage of synonym - synonym candidate pairs each annotator considered to be valid synonyms

0 Annotators	1 Annotator	2 Annotators	3 Annotators	4 Annotators
43%	20%	18%	11%	8%

Table 11: Percentage distribution of the amount of annotators considering a synonym - synonym candidate pair as valid.

Table 11 shows the percentage agreement rate of the annotators. On 43% of the pairs, no annotator considered them to be actual synonyms, highly suggesting they are in fact no synonyms whatsoever. Among others, examples are "schallend (resounding)" and "durchschlagend (sweeping) or "blenden (to glare)" and "verstecken (to hide)". In 20% of the cases exactly one annotator chose a synonym - synonym candidate pair to be valid, while the other three annotators rated it as non-synonyms. An example of this category is "lernen (to learn)" and "erfahren (to experience)". 18% of the time two annotators decided the given pairs are synonyms while the other two decided they are not. A pair in this category includes "typisch (typical)" and "traditionell (traditional)". In 11% of the cases three of the four annotators decided to rate the given pair as synonyms. An example is "gleichlautend (conform)" and "identisch (identical)". 8% of the time all four annotators agreed in their decision to rate the synonym - synonym candidate pair as valid. Exemplary one could name "brauchbar (viable)" and "nützlich (useful)". Especially in the categories of three or four annotators agreeing in their decision to rate the pairs as valid synonyms one could argue these pairs are synonyms in a common word sense and thus belong into the gold standard. If only one or two annotators chose the synonym - synonym candidate pairs to be valid these words probably are synonyms within a special word context or only to some (native) speakers. They could be included into the gold standard but do not necessarily have to. However considering the results of this manual

evaluation, the gold standard in fact seems to miss synonyms thus lowering the precision rate of the extracted synonyms candidates.

8 Conclusion

Considering the results produced in the creation of this thesis one can conclude the extraction of synonyms using SMT and parallel corpora is a promising approach. The most important factor is to produce an accurate and high quality word alignment in order to find reasonable pivots and re-translations. This is especially important considering the failure of both re-ranking methods used in this thesis. Regarding the performance of both re-ranking methods, even though they did not perform as desired, they still can be considered valid options to improve the extraction. Looking at the differences observed in the quality of the extracted synonym candidates there appears to be a difference between adjectives and verbs, nouns. Adjectives perform significantly better. In my opinion this is caused by the overall nature of adjectives and the fact that most of them possess a lot of synonyms for example in *Duden*. Also the rarity of the target words plays an important role since more common words use to appear more often in a general parallel corpus thus improving the quality of the word alignment.

Possible and required improvements could include using different corpora to create the vector space model or using other, probably more sophisticated, distributional similarity measures in order to improve the distinction of the different similarity values of the vector pairs.

A further possible point of improvement could be found in the creation of the gold standard. Maybe combining multiple resources could improve its overall quality. Another possible change that could be tested is to not only extract the synonyms of the target word but also extract the synonyms of the synonym candidate and check them for the appearance of the target word. As a conclusion one could argue the method investigated in this thesis is useful, however still some improvements as suggested above can be tried in

order to improve the automatic synonym extraction using SMT even more.

References

- Colin Bannard and Chris Callison-Burch. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics, 2005.
- Regina Barzilay and Lillian Lee. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 16–23. Association for Computational Linguistics, 2003.
- Regina Barzilay and Kathleen R McKeown. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of the 39th annual meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics, 2001.
- Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- Jaime G Carbonell, Steve Klein, David Miller, Mike Steinbaum, Tomer Grassy, and Jochen Frey. Context-based Machine Translation. *Proceedings of the Association for Machine Translation of the Americas 2006*, pages 19–28, 2006.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2013*, 2013.
- Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation summit 2005*, volume 5, pages 79–86, 2005.

- Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009.
- Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389, 2010.
- Alon Lavie and Michael J Denkowski. The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine translation*, 23(2):105–115, 2009.
- Alessandro Lenci and Giulia Benotto. Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics, 2012.
- Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. Identifying Synonyms Among Distributionally Similar Words. In *Proceedings of International Joint Conference on Artificial Intelligence 2003*, volume 3, pages 1492–1493, 2003.
- Adam Lopez. Statistical Machine Translation. *ACM Computing Surveys (CSUR)*, 40(3):8, 2008.
- Giulio Maltese and Federico Mancini. An Automatic Technique to Include Grammatical and Morphological Information in a Trigram-based Statistical Language model. In *Proceedings of Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference*, volume 1, pages 157–160. IEEE, 1992.
- Rada Mihalcea and Ted Pedersen. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and*

- using parallel texts: data driven machine translation and beyond-Volume 3*, pages 1–10. Association for Computational Linguistics, 2003.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- Roland Schaefer. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, 2015.
- Roland Schaefer and Felix Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of Language Resource Evaluation Conference 2012*, pages 486–493, 2012.
- Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. *Universitäten Stuttgart und Tübingen*, 1995.
- Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.
- Helmut Schmid. Improvements in Part-of-Speech Tagging with an Application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Processing*, pages 13–26. Kluwer Academic Publishers, Dordrecht, 1999.
- Julie Weeds and David Weir. A General Framework for Distributional Similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88. Association for Computational Linguistics, 2003.

Julie Weeds, David Weir, and Diana McCarthy. Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1015. Association for Computational Linguistics, 2004.

Moritz Wittmann, Marion Weller, and Sabine Schulte im Walde. Automatic Extraction of Synonyms for German Particle Verbs from Parallel Data with Distributional Similarity as a Re-Ranking Feature. In *Proceedings of Language Resource and Evaluation Conference 2014*, pages 1430–1437, 2014.