Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation

Marion Weller[1,2], **Fabienne Cap**[2], **Stefan Müller**[1]
Sabine Schulte im Walde[1], **Alexander Fraser**[2]
[1] IMS, University of Stuttgart
{wellermn;muellesn;schulte}@ims.uni-stuttgart.de
[2] CIS, Ludwig-Maximilian University of Munich
{cap;fraser}@cis.uni-muenchen.de

Abstract

The paper presents an approach to morphological compound splitting that takes the degree of compositionality into account. We apply our approach to German noun compounds and particle verbs within a German–English SMT system, and study the effect of *only splitting compositional compounds* as opposed to an aggressive splitting. A qualitative study explores the translational behaviour of non-compositional compounds.

1 Introduction

In German, as in many other languages, two (or more) simplex words can be combined to form a compound. This is a productive process, leading to a potentially infinite number of sound German compounds. As a consequence, many NLP applications suffer from coverage issues for compounds which do not appear or appear only infrequently in language resources. However, while many compounds are not covered, their component words are often found in lexical resources or training data. Compound processing allows access to these component words and thus can overcome these sparsity issues.

We use Statistical Machine Translation (SMT) as an example application for compound processing. Our SMT system translates from German to English, where compounds are usually split in the German source language prior to training and decoding. The benefits are obvious: vocabulary size is reduced and the languages are adjusted in terms of granularity, as exemplified by the compound *Holzzaun*. This results in better alignment quality and model estimation. Compound splitting also enables the translation of compounds not occurring in the parallel data, if the parts have been seen and can thus be translated individually. However, these assumptions only hold for *compositional* compounds like *Holzzaun* ('wooden fence'), whose meanings can be derived from the meanings of their constituents, namely *Holz* ('wood') and *Zaun* ('fence'). In contrast, the splitting of *non-compositional* compounds may lead to translation errors: e.g. the meaning of *Jägerzaun* ('lattice fence') cannot be represented by the meanings of its constituents *Jäger* ('hunter') and *Zaun* ('fence'). Here, an erroneous splitting of the compound can lead to wrong generalizations or translation pairs, such as *Jäger* → *lattice*, in the absence of other evidence about how to translate *Jäger*. When splitting compounds for SMT, two important factors should thus be considered: (1) *whether* a compound is compositional and should be split, and if so (2) *how* the compound should be split. Most previous approaches mainly focused on the second task, *how* to split a compound, e.g. using frequency statistics (Koehn and Knight, 2003) or a rule-based morphology (Fritzinger and Fraser, 2010), and all of them showed improved SMT quality for compound splitting. The decision about *whether* the compound is compositional and should be split at all has not received much attention in the past.

Holzzaun < wooden, fence — *1:n alignment*; Holz — wooden, Zaun — fence — *1:1 alignment*

In this work, we examine the effect of *only splitting compositional compounds*, in contrast to splitting all compounds. To this end, we combine (A) an approach relying on the distributional similarity between compounds and their constituents, to predict the degree of compositionality and thus to determine *whether* to split the compound with (B) a combination of morphological and frequency-based features

to determine *how* to split a compound. We experiment with this novel semantically-informed compound splitting on the source-side data of a German-English SMT system. As far as we know, we are the first to study the impact of compositionality-aware compound splitting in SMT. We evaluate our systems on a standard and on a specifically created test set, both for noun compounds and particle verbs. Our results show that phrase-based SMT is generally robust with regard to over-splitting non-compositional compounds, with the exception of low-frequency words. This is in line with corresponding assumptions from previous work. Furthermore, we present a small-scale study about the translational behaviour of non-compositional compounds, which can surprisingly often be translated component-wise.

2 Related Work

We combine morphology-based compound splitting with distributional semantics to improve phrase-based SMT. Here, we discuss relevant work of compound splitting in SMT and distributional semantics.

2.1 Compound Splitting in SMT

Compound splitting in SMT is a well-studied task. There is a wide range of previous work, including purely string- and frequency-based approaches, but also linguistically-informed approaches. All lines of research improved translation performance due to compound splitting. In Koehn and Knight (2003), compounds are split through the identification of substrings from a corpus. The splitting is performed without linguistic knowledge (except for the insertion of the filler letters "(e)s"), which necessarily leads to many erroneous splittings. Multiple possible splitting options are disambiguated using the frequencies of the substrings. Starting from Koehn and Knight (2003), Stymne (2008) covers more morphological transformations and imposes POS constraints on the subwords. Nießen and Ney (2000) and Fritzinger and Fraser (2010) perform compound splitting by relying on morphological analysers to identify suitable split points. This has the advantage of returning only linguistically motivated splitting options, but the analyses are often ambiguous and require disambiguation: Nießen and Ney (2000) use a parser for context-sensitive disambiguation, and Fritzinger and Fraser (2010) use corpus frequencies to find the best split for each compound. Other approaches use a two-step word alignment process: first, word alignment is performed on a split representation of the compounding language. Then, all former compound parts for which there is no aligned counterpart in the non-compounding language are merged back to the compound again. Finally, word alignment is re-run on this representation. See Koehn and Knight (2003) for experiments on German, DeNeefe et al. (2008) for Arabic and Bai et al. (2008) for Chinese. This blocks non-compositional compounds from being split if they are translated as one simplex English word in the training data (e.g. *Heckenschütze*, lit. 'hedge|shooter'; 'sniper') and aligned correctly. However, cases like *Jägerzaun*, 'lattice fence' are not covered.

In the present work, we identify compounds with a morphological analyser, disambiguated with corpus frequencies. Moreover, we restrict splitting to compositional compounds using distributional semantics. We are not aware of any previous work that takes semantics into account for compound splitting in SMT.

2.2 Distributional Semantics and Compounding

Distributional information has been a steadily increasing, integral part of lexical semantic research over the past 20 years. Based on the *distributional hypothesis* (Firth, 1957; Harris, 1968) that "you shall know a word by the company it keeps", distributional semantics exploits the co-occurrence of words in corpora to explore the meanings and the similarities of the words, phrases, sentences, etc. of interest.

Among many other tasks, distributional semantic information has been utilised to determine the degree of compositionality (or: semantic transparency) of various types of compounds, most notably regarding noun compounds (e.g., Zinsmeister and Heid (2004), Reddy et al. (2011), Schulte im Walde et al. (2013), Salehi et al. (2014)) and particle verbs (e.g., McCarthy et al. (2003), Bannard (2005), Cook and Stevenson (2006), Kühner and Schulte im Walde (2010), Bott and Schulte im Walde (2014), Salehi et al. (2014)). Typically, these approaches rely on co-occurrence information from a corpus (either referring to bags-of-words, or focusing on target-specific types of features), and compare the distributional features of the compounds with those of the constituents, in order to predict the degree of compositionality of the

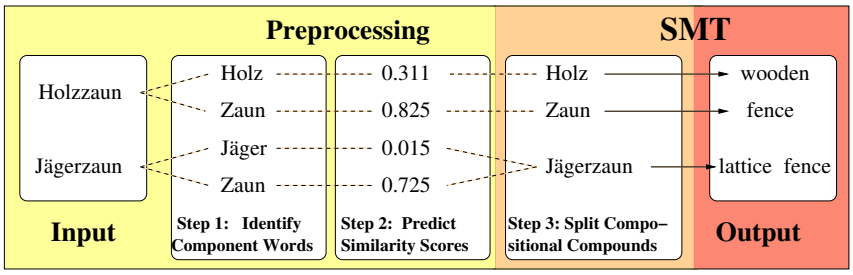

Figure 1: Semantically-informed compound processing in SMT.

compound. The underlying assumption is that a compound which is similar in meaning to a constituent (as in *Holzzaun–Zaun* ('wooden fence'–'fence') but not in *Löwenzahn–Zahn* ('lion|tooth (dandelion)'–'tooth')) is also similar to the constituent with regard to co-occurrence information.

Most related to this work on noun compounds, Reddy et al. (2011) relied on window-based distributional models to predict the compositionality of English noun compounds, and Schulte im Walde et al. (2013) compared window-based against syntax-based distributional models to predict the compositionality of German noun compounds. Zinsmeister and Heid (2004) used subcategorising verbs to predict compound–head similarities of German noun compounds. Most recently, Salehi et al. (2014) extended the previous approaches to take multi-lingual co-occurrence information into account, regarding English and German noun compounds, and English particle verbs.

3 Methodology

We integrate our semantically-informed compound splitting as a pre-processing step to the German source language of an SMT system. See Figure 1 for an illustration of our compound processing pipeline.

3.1 Target Compounds

German compounds are combinations of two (or more) simplex words. In some cases, a morphological transformation is required: for example, when combining the two nouns *Ausflug* ('excursion') and *Ziel* ('destination') → *Ausflug<u>s</u>ziel* ('excursion destination'), a filler letter (here: *"s"*) needs to be inserted. Other such transformations include more filler letters or the deletion/substitution of letters.

Noun compounds are formed of a head noun and a modifier, which can consist of nouns, verbs, adjectives or proper nouns.

Particle verbs are productive compositions of a base verb and a prefix particle, whose part-of-speech varies between open-class nouns, adjectives, and verbs, and closed-class prepositions and adverbs. In comparison to noun compounds, the constituents of German particle verbs exhibit a much higher degree of ambiguity: Verbs in general are more ambiguous than nouns, and the largest sub-class of particles (those with a preposition particle) is highly ambiguous by itself (e.g. Lechler and Roßdeutscher (2009) and Springorum (2011)). For example, in *<u>an</u>knabbern* ('to nibble <u>partially</u>'), the particle *an* expresses a partitive meaning , whereas in *<u>an</u>kleben* ('to glue <u>onto sth.</u>') *an* has a topological meaning (*to glue sth. onto an implicit background*). In addition, particle verb senses may be transparent or opaque with respect to their base verbs. For example, *abholen* 'fetch' is rather transparent with respect to its base verb *holen* 'fetch', whereas *anfangen* 'begin' is more opaque with respect to *fangen* 'catch'. In contrast, *einsetzen* has both transparent (e.g. 'insert') and opaque (e.g. 'begin') verb senses with respect to *setzen* 'put/sit (down)'. The high degree of ambiguity makes particle verbs a challenge for NLP. Moreover, particle and base verb can occur separately (*er fängt <u>an</u>*: 'he begins') or in one word (*dass er <u>anf</u>ängt*: 'that he begins'), depending on the clausal type. This makes consistent treatment of particle verbs difficult.

3.2 Identification of Component Parts

We use the rule-based morphological analyser SMOR (Schmid et al., 2004) to identify compounds and their constituents in our parallel training data (cf. Section 4). It relies on a large lexicon of word lemmas and feature rules for productive morphological processes in German, i.e., compounding, derivation and

inflection. In this paper, we will not consider splitting into derivational affixes (as needed for, e.g., Arabic and Turkish), but instead identify simplex words that may also occur independently. Moreover, we only keep noun compounds and particle verbs consisting of two constituents. The resulting set consists of 93,299 noun compound types and 3,689 particle verb types.

3.3 Predicting Compositionality based on Distributional Similarity

Starting from this set of compounds as derived from our parallel training data, we collected distributional co-occurrence information from two large German web corpora and the machine translation training data: (i) the German *COW* corpus (Schäfer and Bildhauer (2012), ∼9 billion words), (ii) the *SdeWaC* (Faaß and Eckart (2013), ∼880 million words), (iii) our MT parallel corpus (∼40 million words) and (iv) MT language model training data (∼146 million words). We relied on earlier work and used the 20,000 most frequent nouns from the SdeWaC as co-occurrence features, looking into a window of 20 words to the left and to the right of our target compounds and their constituents. We thus obtained a co-occurrence matrix of all compounds and their constituents with the 20,000 selected nouns. As co-occurrence strength (i.e., how strong is a co-occurrence between a target word and a co-occurring noun), we collected frequencies and transformed them into *local mutual information (LMI)* values, cf. Evert (2005). Finally, we calculated the distributional similarity between the compounds and their constituents, relying on the standard measure *cosine*. The cosine value is then used to predict the degree of compositionality between the respective compound–constituent pairs. For example, the cosine value of the pair *Baumschule–Baum*[1] is 0.38, while the cosine value of the pair *Baumschule–Schule* is only 0.01.

3.4 Semantically-Informed Compound Splitting

In the two preceding sections, we described how we identified component words and calculated distributional compositionality scores for all of the compounds found in our training data. Here, we give details on how we include the semantic information into the compound splitting process. Recall that we only want to split compositional compounds and keep non-compositional compounds together.

The splitting decision (to split/not split a compound) is based on the compositionality score of the compound that takes into account either one or both of the compound–constituent cosine values: if the predicted degree of compositionality is high, the compound is split. We consider and combine four different criteria: i) only the compound–modifier similarity (*mod*); (ii) only the compound–head similarity (*head*); a combination of the compound–modifier and the compound–head similarities, relying on (iii) the geometric mean (*geom*) or (iv) on the arithmetic mean (*arith*). We used different thresholds for each of these criteria throughout our experiments, with a specific focus on distinguishing the contributions of the modifiers vs. the heads in the splitting decision, following insights from recent work in psycholinguistic studies (Gagné and Spalding, 2009; Gagné and Spalding, 2011) as well as in computational approaches on noun compounding (Reddy et al., 2011; Schulte im Walde et al., 2013). Furthermore, we compare the effects of splitting with regard to two types of compounds, noun compounds and particle verbs: Both types are very productive and can generate a potentially infinite number of new forms.

4 Experimental Setting

This section gives an overview on the technical details of the SMT system and our data sets. Compound splitting is applied to all source-language data, i.e. the parallel data used to train the model, as well as the input for parameter tuning and testing.[2]

Translation Model Moses is a state-of-the-art toolkit for phrase-based SMT systems (Koehn et al., 2007). We use it with default settings to train a translation model and we do so separately for each of the different compound splittings. Word alignment is performed using GIZA++ (Och and Ney, 2003). Feature weights are tuned using Batch-Mira (Cherry and Foster, 2012) with *'-safe-hope'* until convergence.

Training Data Our parallel training data contains the Europarl corpus (version 4, cf. Koehn (2005)) and also newspaper texts, overall ca. 1.5 million sentences[3] (roughly 44 million words). In addition, we

[1] *Baum*|*schule*: 'tree|school' (tree nursery)

[2] Compounds not contained in the parallel data are always split, as they cannot be translated otherwise.

[3] Data from the shared task of the EACL 2009 workshop on statistical machine translation: `www.statmt.org/wmt09`

use an English corpus of roughly 227 million words (including the English part of the parallel data) to build a target-side 5-gram language model with SRILM (Stolcke, 2002) in combination with KENLM (Heafield, 2011). For parameter tuning, we use 1,025 sentences of news data.

Standard Test set 1,026 sentences of news data (test set from the 2009 WMT Shared Task): this set is to measure the translation quality on a standard SMT test and make it comparable to other work.

Noun/Verb Test set As our main focus lies on sentences containing compounds, we created a second test set which is rich in compounds. From the combined 2008-2013 Shared Task test sets, we extracted all sentences containing at least one noun compound for which we have compound-constituent similarity scores. Moreover, we excluded sentences containing nouns that are not in the parallel training data: such compounds can only be translated when split which allows to translate their components. The final test set consists of 2,574 sentences. Similarly, we also created a set rich in particle verbs (855 sentences).

Opaque Test set As the two first test sets mainly contain compositional compounds, we use a third test set consisting of sentences with only non-compositional compounds. The underlying compounds were chosen based on a list containing noun compounds and human ratings for compositionality (von der Heide and Borgwaldt (2009)). As before, the compounds must have occurred in the parallel data. The result is a list of 14 compounds, of which 11 have a low modifier-compound similarity and 3 have a low head-compound similarity. We then extracted sentences containing these compounds (5 per compound = 70 in total) from German newspaper data[4]. In contrast to the other sets, we use this test set in a qualitative study, to approximate the translation quality by counting the number of correctly translated compounds.

5 SMT Results

In this section, we present and discuss the results of our machine translation experiments. We first report results for two test sets in terms of a standard evaluation metric (BLEU) and then continue with a small-scale qualitative study on the translational behaviour of non-compositional compounds.

5.1 Compound Splitting within a Standard SMT Task

BLEU (Papineni et al., 2002) is a common metric to automatically measure the quality of SMT output by comparing n-gram matches of the SMT output with a human reference translation. Table 1 lists the results for our SMT-systems: we report on different compound-constituent scores and thresholds, for noun compounds and particle verbs respectively. Note that BLEU scores are not comparable across different test sets, but only illustrate system differences within one test set. We compare our systems to the scores of a *baseline system* (without compound processing) and an *aggressive split* system in which all noun compounds and particle verbs are split. The labels *DIST* and *FREQ* indicate how several possible splittings were disambiguated: *DIST* means we chose the splitting option having the higher geometric mean of the two compound-constituent scores, assuming that the variant expressing a higher compositionality score leads to the more probable splitting analysis. For *FREQ*, the decision is based on the geometric mean of corpus frequencies of the respective components of the compound, as is common practise for the disambiguation of multiple splitting options in SMT (Koehn and Knight, 2003; Fritzinger and Fraser, 2010). In terms of BLEU, there is little difference for these two variants. For further experiments, we thus decided to always use *FREQ* for disambiguation, assuming that components chosen by frequency are potentially better repre-

		nouns		**particle verbs**	
		stand.	**noun**	**stand.**	**verb**
baseline		21.00	21.08	21.00	20.29
aggr	DIST	22.00	22.02	21.02	20.11
	FREQ	22.04	21.88	21.11	20.21
0.05	head	21.77	21.58	–	–
	mod.	22.01	21.74	–	–
	geom.	21.99	21.71	–	–
	arith.	21.95	21.95	–	–
0.1	head	21.91	21.69	21.11	20.24
	mod.	22.01	21.63	20.98	20.43
	geom.	22.06	21.90	21.12	20.55
	arith.	22.05	21.73	21.08	20.34
0.15	head	21.80	21.67	21.10	20.09
	mod.	21.71	21.77	21.00	20.25
	geom.	21.78	21.64	20.84	20.30
	arith.	22.00	21.77	21.24	20.40
0.2	head	21.78	21.51	–	–
	mod.	21.78	21.45	–	–
	geom.	21.76	21.54	–	–
	arith.	22.02	21.79	–	–

Table 1: BLEU scores for all compound-constituent variations.

[4] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.html

rating	compound	gloss	mod.	head	translation
HIGHLY	Staats|bankrott	nation|bankruptcy	0.4779	0.6527	*national bankruptcy*
COMP.	Staats|gebilde	nation|structure	0.6955	0.3431	*national structure*
MEDIUM	Industrie|staat	industry|nation	0.0258	0.1488	*industrial nation*
COMP.	Staats|kasse	nation|cash box	0.0718	0.2757	*public purse, treasury*
LOW	Staats|spitze	nation|top	0.0024	0.0040	*top/head of state*
COMP.	Staats|monotheismus	nation|monotheism	0.0071	0.0071	*national monotheism*

Table 2: Examples for different compound-constitutent score ranges: HIGH: highly compositional, MEDIUM: cases of doubt, LOW: highly non-compositional, according to their scores.

sented in the training data. Thus, we first use frequencies to determine the best split option in the case of several possibilities, and then we apply distributional semantics to determine whether to split at all. The remainder of Table 1 reports on different variants of the semantically-informed splitting criteria we used. The notation *head/mod/geom/arith* indicates which (combination of) compound-constituent scores were applied as criterion, with the threshold indicated by the vertical number. We performed the first set of experiments with different thresholds for noun compounds, and then applied the medium-range thresholds to the particle verbs. Generally, there are no considerable differences between the systems with semantically restricted splitting and the *aggressive split* systems, even though there seems to be a slightly positive effect for particle verbs. Having a closer look, we find that for noun compounds on the standard test set, the best results (threshold: 0.1) are at the same level as the *aggressive split* systems; with some small losses in BLEU on some of the other settings.

5.2 Discussion

All settings clearly outperform the baseline system (without compound processing). This indicates that phrase-based SMT is rather robust with regard to non-semantic splitting as it is can often recover from over-splitting by translating the word sequence as a phrase. This is in line with previous observations of Koehn and Knight (2003). The results for the noun test set, which is biased towards containing more nominal compounds, even suggests that less splitting might harm the system, as the BLEU scores tend to drop when increasing the threshold. For particle verbs,[5] the picture is slightly different: first, splitting only particle verbs does not lead to a considerable improvement over the baseline, as in the case of noun compounds. For the verb test set, it even leads to a drop in BLEU. However, a more restricted splitting leads to improved BLEU scores, even though not significantly better than the un-split baseline system. Even though the handling of particle verbs needs to be refined in terms of dealing with their structural behaviour (split vs. unsplit depending on the sentence structure) or ambiguities of the particle verb, we consider this an encouraging result indicating that particle verbs can benefit from a semantically-informed splitting process.
There are several possible reasons why a more restricted splitting might not lead to an improvement, even though the idea of splitting only compositional compounds is intuitive and straightforward.

Inconsistent Splitting Compositionality is a continuum rather than a binary decision, with the scores of many (compositional) compounds being in the medium range. Thus, it happens that some compounds containing a certain constituent are split, whereas others are not: such inconsistent splittings do not contribute to the generalization compound splitting aims for. Table 2 gives examples for compounds with different degrees of compositionality, which illustrate this issue: for *Industriestaat* ('industrial nation') and *Staatskasse* ('public purse') in the middle part of the table, a splitting decision based on the *head* scores for thresholds of 0.15 or 0.2 leads to inconsistent splitting. Only compounds with high scores, as the examples at the top of Table 2 are always split. The bottom part gives examples with comparatively low compound-constituent scores that would benefit from splitting, but which will not be split in any of our systems.

[5]Note that there are considerably less particle verbs than noun compounds in the standard test set and the parallel data.

compound	gloss	translation	unsplit	f	split	f
Seehunde	*sea|dogs*	*seals*	seals	5	seals	5
Flohmarkt	*flea|market*	*flea market*	flea market	5	flea market	5
Kopfsalat	*head|salad*	*lettuce*	lettuce	5	lettuce	5
Handtuch	*hand|cloth*	*towel*	towel	5	towel	5
Kronleuchter	*crown|candelabra*	*chandelier*	chandelier	5	crown leuchter	5
Gürteltiere	*belt|animal*	*armadillo*	armadillos	5	belt animals	5
Wasserhahn	*water|rooster*	*tap*	tap	5	water tap	2
					water supply	3
Meerschweinchen	*sea|piglet*	*guinea pig*	guinea pig	5	guinea pig	4
					sea pig	1
Taschenbuch	*pocket|book*	*paperback*	paperback	5	paper back	3
					pocket book	2
Kronkorken	*crown|cork*	*crown cap*	*kronkorken	5	crown corks	5
Taschenlampe	*pocket|lamp*	*flashlight*	*taschenlampe	5	pocket lamp	4
					bag lamp	1
Fleischwolf	*meat|wolf*	*meat grinder*	*fleischwolf	5	meat wolf	5
Marienkäfer	*Mary|bug*	*ladybug*	*marienkäfer	5	*marie käfer	5
Blockflöten	*block|flute*	*recorder*	*blockflöten	5	block might	4
					bloc might	1

Table 3: correct vs. wrong – Translation of non-compositional compounds (opaque test set) without being split (*unsplit*) vs. being *split* prior to translation. '*' highlights untranslated compounds.

Coverage of Opaque Compounds Another relevant factor concerns the frequency ranges of compounds that are most interesting for this approach. High/mid frequency compounds are usually well-covered by the training data of an SMT system, and in most cases they are translated correctly even if they have been split erroneously. This is due to the fact that split compounds can be learned and translated as a phrase if there were enough instances for the system to learn a valid translation. In the case of low-frequency compounds, the system is less likely to learn a correct translation from the parallel data. However, low-frequency compounds are not well covered by the system and splitting should thus be highly beneficial. Newly created, i.e. highly compositional compounds, tend to be of low frequency, as is illustrated by the example of *Staatsmonotheismus* (freq=1 in the parallel data) in Table 2. However, a wrong splitting decision for a non-compositional compound of low frequency is likely to lead to an incorrect translation as the SMT system has better statistics for the individual parts than for the sequence of the compounds constituents. We assume that for low-frequency compounds the distributional similarity scores are generally less reliable, even though using LMI helps to minimize this. To a certain extent, we expect non-compositional compounds –which are typically considered as lexicalized– to occur with higher frequencies than novel compositional compounds.[6] Furthermore, there are considerably more compositional than non-compositional compounds in standard text. Thus, being in favor of splitting in the case of low-frequency words should be reasonable in most contexts.

6 A Closer Look at Translating Opaque Compounds

In this section, we compare the translations of non-compositional compounds when they are unsplit and when they are split. We use a small test set containing 70 sentences, 5 for each of the 14 non-compositional compounds (see Section 4). Then we conduct a small-scale qualitative analysis focusing on the correct translation of opaque compounds.

Table 3 reports on correct translations for the non-compositional compounds for an experiment where they have been *split* or not split (*unsplit*) prior to translation. Even though all compounds occurred in the parallel data, five (which are marked with '*') cannot be translated by the unsplit system due to not being aligned correctly. The other compounds are translated correctly (marked with "+" in Table 3). In the course of our study, we found that many of the correct translations remain the same (*seals, flea market, lettuce, towel*). In the case of *guinea pig, paperback* and *tap* there are mixed results of correct and incorrect translations. Only in the cases of *chandelier* ("*crown leuchter*") and *armadillo* ("*belt animal*"),

[6] It has to be noted, though, that the model is influenced by the somewhat different domain of the parallel data (European Parliament proceedings, a standard data set for SMT).

compound	gloss	translation
Bärlauch	*bear|leek*	*bear leek*
Baumschule	*tree|school*	*tree nursery*
Löwenanteil	*lion|share*	*lion's share*
Fliegenpilz	*fly|mushroom*	*fly agaric*
Flohmarkt	*flea|market*	*flea market*

compound	gloss	translation
Handtasche	*hand|bag*	*handbag*
Hirschkäfer	*stag|beetle*	*stag beetle*
Hüttenkäse	*cottage|cheese*	*cottage cheese*
Kronkorken	*crown|cork*	*crown cap*
Teelicht	*tea|light*	*tea candle*

Table 4: Examples for (near) literal translation of non-compositional compounds.

which were translated correctly with the *unsplit* system, all translations obtained with the *split* system are wrong. Somewhat surprisingly, in some cases there even is a benefit from splitting the non-compositional compounds: *Kronkorken*, previously not translated at all, is correctly generated as *crown cork*. For other previously untranslated words, *Fleischwolf* and *Taschenlampe*, literal translations of the constituents are given: while *meat wolf* (instead of *meat grinder*) is probably not understandable, the translation of *Taschenlampe* as *pocket lamp* is certainly preferable to the untranslated compound.

Due to the observed unexpected translational behaviour of 2 of the 14 non-compositional compounds (*Flohmarkt* and *Kronkorken*), which can be translated literally and thus –in theory– benefit from splitting, we present a small study illustrating that this phenomenon is not as rare as one would intuitively expect. This study is not meant to be comprehensive, but rather to point out that the translational behaviour of non-compositional compounds can correspond to that of compositional compounds; Table 4 lists a few such examples. We assume that this behaviour is due to the fact that English and German are similar languages with a similar background. Thus, the "images" used in non-compositional words often tend to be similar. For some of the compounds (e.g. *Flohmarkt*) this is even true for some Romance languages, too (IT: *mercato delle pulci*, FR: *marché aux puces*) .

Generally, the SMT system should even be able to handle cases where the translation of one part is not strictly literal (e.g. *cap–cork* or *agaric–mushroom*). In comparison to a dictionary, which only lists few translations, the translation model offers a large choice of translation options that are not always strictly synonymous, but can cover a large range of related meanings. In combination with the target-side language model, this could allow to "guess" good translations of such compounds. However, the component-wise translation of non-compositional compounds only works if the source- and target language compounds contain the same number of constituents. For example, consider translating the word *Faultier* (*lazy|animal*: "*sloth*"): even if the SMT system offers the translation *faul–sloth*, it would also need to produce a translation for the constituent *tier*, probably resulting in something like *sloth animal*.

In conclusion, while phrase-based SMT is often able to recover from over-splitting by translating a word sequence as a phrase, this is not always necessary for opaque compounds as they can have a literal or near-literal translation. Thus, for explicitly handling non-compositional compounds in SMT, a monolingual estimation of compositionality is not the only relevant factor. The translational behaviour of compounds should also be taken into account.

7 Conclusion and Future Work

We studied the impact of compositionality in German-English SMT by restricting compound splitting to compositional compounds. The decision about compositionality is based on the distributional similarity between a compound and its constituents. We experimented with different threshold/score combinations on a standard and a specifically created test set. Our results indicate that phrase-based SMT is very robust with regard to over-splitting non-compositional noun compounds, with the exception of low-frequency compounds. Furthermore, we studied the translational behaviour of non-compositional compounds with a special focus on the fact that non-compositional compounds can in some cases be translated component-wise, leading to the conclusion that a monolingual estimation of compositionality is not sufficient for an optimal explicit handling of compounds in SMT applications.

The relatively low impact of distinguishing the degree of compositionality might also be due to the fact that the task of translating noun compounds can be considered "easy", as the split components always occur adjacently. In contrast, handling other types of non-compositional structures (e.g. noun-verb or preposition-noun-verb combinations which are non-compositional) is a challenging task for future work.

Acknowledgements

This work was funded by the DFG Research Projects "Distributional Approaches to Semantic Relatedness" (Marion Weller, Stefan Müller) and "Models of Morphosyntax for Statistical Machine Translation – Phase 2" (Fabienne Cap, Alexander Fraser, Marion Weller) and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

References

Ming-Hong Bai, Keh-Jiann Chen, and Jason S Chang. 2008. Improving word alignment by adjusting chinese word segmentation. In *IJCNLP'08: Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 249–256.

Collin Bannard. 2005. Learning about the Meaning of Verb–Particle Constructions from Corpora. *Computer Speech and Language*, 19:467–478.

Stefan Bott and Sabine Schulte im Walde. 2014. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. In *Proceedings of the 9th Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *HLT-NAACL'12: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, volume 12, pages 34–35.

Paul Cook and Suzanne Stevenson. 2006. Classifying Particle Semantics in English Verb-Particle Constructions. In *Proceedings of the ACL/COLING Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 45–53, Sydney, Australia.

Steve DeNeefe, Ulf Hermjakob, and Kevin Knight. 2008. Overcoming vocabulary sparsity in mt using lattices. In *AMTA'08: Proceedings of the 8th Biennial Conference of the Association for Machine Translation in the Americas*.

Stefan Evert. 2005. *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Gertrud Faaß and Kerstin Eckart. 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*, pages 61–68, Darmstadt, Germany.

John R. Firth. 1957. *Papers in Linguistics 1934-51*. Longmans, London, UK.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the Fifth Workshop on Statistical Machine Translation*, pages 224–234. Association for Computational Linguistics.

Christina L. Gagné and Thomas L. Spalding. 2009. Constituent Integration during the Processing of Compound Words: Does it involve the Use of Relational Structures? *Journal of Memory and Language*, 60:20–35.

Christina L. Gagné and Thomas L. Spalding. 2011. Inferential Processing and Meta-Knowledge as the Bases for Property Inclusion in Combined Concepts. *Journal of Memory and Language*, 65:176–192.

Zellig Harris. 1968. Distributional Structure. In Jerold J. Katz, editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press.

Kenneth Heafield. 2011. Kenlm: faster and smaller language model queries. In *EMNLP'11: Proceedings of the 6th workshop on statistical machine translation within the 8th Conference on Empirical Methods in Natural Language Processing*, pages 187–197.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *EACL '03: Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–193, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL'07: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Demonstration Session*, pages 177–180.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *MT Summit'05: Proceedings of the 10th machine translation summit*, pages 79–86.

Natalie Kühner and Sabine Schulte im Walde. 2010. Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches. In *Proceedings of the 10th Conference on Natural Language Processing*, pages 47–56, Saarbrücken, Germany.

Andrea Lechler and Antje Roßdeutscher. 2009. German Particle Verbs with *auf*. Reconstructing their Composition in a DRT-based Framework. *Linguistische Berichte*, 220.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *COLING'00: Proceedings of the 18th International Conference on Computational Linguistics*, pages 1081–1085. Morgan Kaufmann.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51,.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL'02: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using Distributional Similarity of Multi-Way Translations to Predict Multiword Expression Compositionality. In *Proceedings of EACL 2014*.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. Smor: A German computational morphology covering derivation, composition and inflection. In *LREC '04: Proceedings of the 4th Conference on Language Resources and Evaluation*, pages 1263–1266.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA.

Sylvia Springorum. 2011. DRT-based Analysis of the German Verb Particle *"an"*. *Leuvense Bijdragen*, 97:80–105.

Andreas Stolcke. 2002. SRILM – an extensible language modelling toolkit. In *ICSLN'02: Proceedings of the international conference on spoken language processing*, pages 901–904.

Sara Stymne. 2008. German compounds in factored statistical machine translation. In *GoTAL '08: Proceedings of the 6th International Conference on Natural Language Processing*, pages 464–475. Springer Verlag.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter, Basis und Oberbegriffen. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Heike Zinsmeister and Ulrich Heid. 2004. Collocations of Complex Nouns: Evidence for Lexicalisation. In *Proceedings of Konvens*, Vienna, Austria.