Simple Compound Splitting for German

Marion Weller-Di Marco
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Centrum für Informations- und Sprachverarbeitung, LMU München
dimarco@ims.uni-stuttgart.de

Abstract

This paper presents a simple method for German compound splitting that combines a basic frequency-based approach with a form-to-lemma mapping to approximate morphological operations. With the exception of a small set of hand-crafted rules for modeling transitional elements, our approach is resource-poor. In our evaluation, the simple splitter outperforms a splitter relying on rich morphological resources.

1 Introduction

In German, as in many other languages, two (or more) words can be combined to form a compound, leading to an infinite amount of new compounds. For many NLP applications, this productive word formation process presents a problem as compounds often do not appear at all or only infrequently in the training data. A typical NLP application that benefits from compound handling is statistical machine translation (SMT). For example, a compound that does not occur in the training data cannot be translated. However, the components of a compound often occur in the training data and can be used to translate a previously unseen compound. Thus, making the parts of a compound accessible through compound splitting when training an SMT system leads to a better lexical coverage and, consequently, to improved translation quality. Similarly, in an information retrieval scenario, information about the individual parts of a compound helps to generalize and can thus lead to improved performance.

The basis for successful compound handling in NLP applications is the decomposition of a complex compound into its components. This is not a trivial task, as the compound parts are not always just concatenated as in *Reis|feld* ('rice field'), but are often subject to morphological modifications. For example, the components can be connected with a transitional element, as the *-er* in *Bilder|buch* ('picture book'); or parts of the modifier can be deleted, for example *Kirch|turm* ('church tower'), where the final *-e* of the lemma *Kirche* is deleted. Furthermore, the modifier components can undergo non-concatenative morphological modifications such as changing a vowel in the word stem ("Umlautung"), for example *Buch* → *Büch-* in *Bücher|regal* ('book shelf').

To split compounds into meaningful parts, and in particular to obtain a lemmatized representation of the modifier, all these morphological operations need to be considered and modeled accordingly.

There are many approaches for compound splitting, ranging from simple substring operations (e.g. Koehn and Knight (2003)) to linguistically sound splitting approaches relying on high-quality morphological resources (e.g. Fritzinger and Fraser (2010)). This paper aims at the "middle ground" of this spectrum by combining a minimum amount of linguistic information with corpus-derived statistics. We present a simple method for compound splitting that makes linguistically informed splitting decisions, but requires only minimal resources. It relies on a small set of handcrafted rules to model transitional elements, but all other morphological operations (such as "Umlautung") are induced from a mapping of inflected word forms to the word lemma – this can be easily obtained from large part-of-speech tagged corpora. Our approach makes use of the fact that many of the forms that are taken on by compound modifiers are equal to inflected forms (typically plural or genitive forms) and thus can be observed in corpora. Thus, an explicit modeling of morphological operations for the modifier is often not necessary. Furthermore, we make use of part-of-speech information for a flat analysis of the compound, as illustrated below:

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 161–166,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

Häuserfassade → haus_NN fassade_NN
house front

Abfüllanlage → abfüllen_V anlage_NN
filling facility

In contrast to morphological resources, which typically involve a large amount of manual work, part-of-speech taggers are easily available and cheap to use even for very large corpora. We thus consider the presented splitting approach as essentially resource-poor.

In the remainder of the paper, we first outline the splitting method, and then evaluate the splitting quality on a set of more than 51,000 German nominal compounds. In a comparison with the splitting results obtained with a well-acclaimed splitter relying on a high-quality morphological resource, our simple splitter obtains competitive results.

2 Related Work

Koehn and Knight (2003) present a frequency-based approach to compound splitting for German. They use word frequencies derived from corpus data to identify compound parts. Different splitting analyses are then ranked based on the geometric mean of subword frequencies. They allow two linking elements (*-s* and *-es*), as well as the deletion of characters. Their basic approach is extended by part-of-speech tags and a bilingual lexicon to restrict the selection of splitting options. Despite the simplicity of the basic approach, they report imrovements in translation quality for German–English translation. Stymne (2008) extends the algorithm by Koehn and Knight (2003) with the 20 most frequent morphological transformations and explores the effect on factored machine translation.

Macherey et al. (2011) present an unsupervised method to compound splitting that does not rely on any handcrafted rules for transitional elements or morphological operations. Their method uses a bilingual corpus to learn morphological operations. Ziering and van der Plas (2016) take this idea a step further, but avoid relying on parallel corpora and instead learn “morphological operation patterns” based on inflectional information derived from lemmatized monolingual corpora. Phenomena such as “Umlautung” are learned as a replacement operation between lemma and inflected form. Riedl and Biemann (2016) present a method based on the assumption that a compound’s components are semantically similar, to identify valid splitting points. Their method is based on a distributional thesaurus and a set of “atomic word units” obtained from corpus data. It does not include normalization of the modifier, but only identifies the splitting points of a compound. Fritzinger and Fraser (2010) use the morphological resource SMOR (Schmid et al., 2004) to obtain splitting points. Multiple splitting options are ranked according to the geometric mean of the subword frequencies.

The approach presented in this paper is based on a splitting method outlined in Weller and Heid (2012) where it is used as a basis for term alignment of bilingual vocabulary in a scientific domain. With the main focus on alignment, the paper does not provide much information on the splitting technique itself. We re-implemented and extended the splitting approach, and present it in more detail with a comparison to a state-of-the-art splitter by Fritzinger and Fraser (2010).

3 Simple Compound Splitting via Form-to-Lemma Mapping

The splitting approach presented in this paper is similar to the frequency-based approach by Koehn and Knight (2003), but is extended with a mapping from inflected forms to lemmas to approximate compounding morphology. Assuming that the components of a compound also occur as inflected forms, a frequency list of lemmatized word forms serves as training data, in combination with a small set of possible transitional elements. In the splitting process, this allows to map a modifier such as *Häuser*, which is also a plural form, to the lemma *Haus* (‘house’)[1]. Additionally, we also use part-of-speech tags in order to restrict splitting possibilities to content words only (e.g. adjectives, nouns, verbs) and to avoid incorrect splits into short, but highly frequent inflected words, such as splitting the simple word *Gründer* (‘founder’) into *grün|der* (‘green|the’), where *der* is a definite article. At the same time, the part-of-speech tags allow to label the components and thus to provide a flat analysis. In the splitting process, the part-of-speech-tags of the modifier(s) can vary between all tags available in the training data, whereas the tag of the compound head is equal to the tag of the entire compound

[1] Lemmatizing the modifier is not possible with the splitting algorithm by Koehn and Knight (2003), which outputs the observed modifier form minus potential transitional elements, leading to different representations for different modifier realizations of the same lemma, e.g. *länderspiel* → *länder|spiel* (‘country match’: international match) vs. *landeswährung* → *land|währung* (‘country currency’: national currency).

(which is part of the input to the splitting process).

The splitting process begins with partitioning the compound into two substrings, which can then be split again in two substrings, respectively[2]. To be accepted as a valid substring, the substring must be found in the list of lemmas (via form-to-lemma mapping), after being modified for transitional elements, if necessary (cf. section 4). In this first splitting step, it is however possible to keep an "intermediate substring" that is to be split into valid substrings at the next splitting step, as illustrated by the word *Breitflügelfledermaus* ('wide wing bat: serotine bat')

comp.	Breitflügelfledermaus
input	breitflügelfledermaus_NN
split-1	breitflügel_XX fledermaus_NN
split-2	breit_ADJ flügel_NN fledermaus_NN

The part *breitflügel* does not exist as an individual word, and thus cannot be found in the lemma and part-of-speech lists; in the second step, it is split into the adjective *breit* and the noun *fügel*, resulting in a correct analysis.

After having determined all possible splitting points and subwords, the resulting splitting possibilities are scored by the *geometric mean* of the lemma frequencies of the parts p_i of the respective splitting. If two splitting analyses have the same score, analyses with fewer explicit morphological operations to model transitional elements are preferred.

4 Modeling Transitional Elements

While many compounds can be formed seamlessly by concatenating two ore more words, some contain transitional elements linking the components. Many transitional elements are part of the inflectional inventory, and sometimes indicate a syntactic function such as *genitive* (e.g. *Tag**es**licht*; 'light of the day: daylight') or a plural (e.g. *Katz**en**futter*; 'food for cats: cat food '). This is, however, not always the case. The grammar *Duden* (Eisenberg et al. (1998), §879 ff.) lists the following transitional elements for noun compounds:

Noun+Noun This category has the most transitional rules, but many are part of the inflection inventory as either plural (pl) or genitive (gen) form and thus do not need to be modeled explicitly, but are covered by the form-to-lemma mapping:

add -en	Tat**en**drang	Tat|Drang	pl
add -n	Hase**n**braten	Hase|Braten	pl
add -ens	Herz**ens**güte	Herz|Güte	gen
add -ns	Glaube**ns**frage	Glaube|Frage	gen
add -es	Kind**es**wohl	Kind|Wohl	gen
add -er	Büch**er**regal	Buch|Regal	pl
add -e	Hund**e**hütte[3]	Hund|Hütte	pl
add -s	Museum**s**leiter	Museum|Leiter	gen
	Ansicht**s**karte	Ansicht|Karte	∅
rem. -e	Kirchturm	Kirch**e**|Turm	∅

From this set, only modifier forms resulting from the last two rules (*add -s*, *remove -e*) are not (entirely) covered by existing inflected forms: while *-s* often marks genitive forms, this transitional element can also occur in modifiers that do not have *-s* as inflection, including the group of nouns ending with frequent nominalization suffixes such as *-ung*, *-keit* or *-ion*. Similarly, the deletion of *-e* results in forms not covered by the inflectional inventory[4].

Verb+Noun There are only two modifications for compounds with a verbal modifier:

add -en	Schreibmaschine	schreib**en**|Maschine
add -n	Wanderweg	wander**n**|Weg

For verbal modifiers containing a nasal (*m, n*), an additional deletion of *-e-* might be required, for example *Rechengerät* → *rechn**en**\|Gerät*.

Other+Noun For all other modifiers (adjective, adverb, preposition), no modification is required.

Implemented Rules Based on the enumeration above, the morphological operations applied to the modifier are modeled as follows:

- Noun: *remove -s*
- Noun: *add -e*
- Noun: *remove -s*, *add -e*
- Verb: *add -en* (including deletion of *-e* in the context of *n,m*)
- Verb: *add -n*

All other morphological modifications are covered by mapping an inflected (plural or genitive) form to the lemma; this includes the phenomenon of

[2] This limits the number of splits to 4 components in total, which is sufficient for most applications, even though the number of components in a compound can be infinite.

[3] There can be some exceptions to this rule where the modifier form is not a plural form, e.g. *Mauseloch – Maus|Loch*.

[4] Both *add -s* and *remove -e* can actually only be applied to feminine nouns. However, as we only use basic POS-tags, this restriction is not used in the splitting process.

"Umlautung" which changes a vowel in the word stem when building the plural form, e.g. *Buch – Bücher*. Modeling more transitional elements is not necessary, and can even be harmful: for example, a *remove rule* for *-er* can result in incorrect analyses, as *-er* is not only a plural suffix, but can also represent a nominalization suffix that is part of the lemma, such as *Fisch**er**boot* → *Fischer|boot* ('fisherman boat') vs. **Fisch|boot* ('fish|boat').

5 Restricting Splitting Operations

In some cases, the selection of components or the application of particular transitional rules leads to incorrect splits. We employ two strategies to prevent some systematically occurring problems.

First, the splitting allows to define stop-words that should not be used as compound components. This concerns, for example, high-frequent verb prefixes, such as *ge-, be-, ver-* or similar items, that cannot stand alone, but nonetheless occur in the training data. Alternatively, such entries can be excluded from the word/lemma lists used to estimate the splitting statistics, cf. section 6.

Furthermore, it is possible to forbid specific operations for particular nouns: this concerns words that are identical to other, unrelated words after removing or adding transitional elements. In contrast to the stop-word list, such words cannot be completely excluded; instead, the list specifies the word in combination with the forbidden operation.

For example, adding an *e* to the word *Reis* ('rice') changes the word to *Reis**e*** ('journey') – thus, the *add -e* operation should not be performed for this word. In the current implementation, there are 17 entries (of which 4 restrict the removal of *-s* and 13 restrict the addition of *-e*, corresponding to the two implemented modifier modifications for nouns). The list of restricted operations does not have a big impact on the overall performance: using the 17 entries results in 121 more correct splitting analyses in a test set of more than 51,000 nouns. However, it is useful to avoid systematic mis-splittings and can be easily extended.

6 Training Data and Categories

The training data consists of two lists: a mapping of inflected forms to lemmas with indication of the part-of-speech tag, and a lemma-POS-frequency list. Such lists can easily be derived from tagged corpora. Since the splitting routine relies on word frequencies, some simple cleaning steps help to improve splitting results: in particular high-frequent "non-words" can harm the splitting quality. Filtering the training data in order to remove such words is likely to be rewarded by better splitting outputs.

Since not all POS-tags make sense as modifier, the tags for this category are restricted to

- **adverbs** *wieder|Aufforstung* 're|forestation'
- **adjectives** *alt|Bestand* 'old|stock'
- **particles** *auf|Preis* 'sur|charge'
- **verbs** *wandern|Weg* 'hiking track'
- **nouns** *Apfel|Kuchen* 'apple cake')
- **proper nouns** *Adam|Apfel* 'adam's apple'

There is an additional "other" tag that can be used to add further categories if necessary, for example neoclassical items such as *hydro* to analyze terms of scientific domains.

As training data, we use a large German web-corpus (1.5 Mrd tokens, based on Baroni et al. (2009)), tagged with TreeTagger (Schmid, 1994). The corpus cleaning steps contain a mapping from old to new German orthography, as well as filtering out bad "short" words (up to length 3) using a dictionary [5]. All data is lowercased for splitting.

7 Evaluation

To evaluate our splitting method, we analyze the splitting analyses obtained for a gold standard and compare them with a state-of-the-art splitter (Fritzinger and Fraser, 2010) relying on the morphological resource SMOR (Schmid et al., 2004). SMOR is a comprehensive German finite-state morphology covering inflection, derivation and compounding. As gold standard, we use the binary split compound set developed for GermaNet (Henrich and Hinrichs, 2011), containing 51,230 noun compounds. For this task, all words in the test-set should be split into two parts.

To evaluate the splitting results, we use the measures *precision* and *recall* as defined in (Fritzinger and Fraser, 2010), adapted to the simpler setting of only rating correct vs. wrong splits, without deciding whether a word should be split or not:

- **precision**: $\frac{correct\,split}{correct\,split + wrong\,split}$
- **recall**: $\frac{correct\,split}{correct\,split + wrong\,split + not\,split}$

[5] Dictionary obtained from `dict.cc`

	correct split	wrong split	not split	P	R	F
SMOR Split	45,054	2,914	3,262	93.93	87.94	90.84
Simple Split	46,905	4,012	313	92.12	91.56	91.84

Table 1: Comparison of splitting results for "SMOR Split" and the presented method.

Without the need to decide whether a word should be split, the *accuracy* of splitting results corresponds to the *recall*. Furthermore, we compute the F-score as

$$F = 2\,\frac{precision * recall}{precision + recall} \qquad (1)$$

Table 1 shows the results of the two systems for the respective best split into two parts. A splitting analysis is counted as correct if both head and modifier are correct (i.e. exact string-match with the reference set). Part-of-speech tags are not part of the test-set and can thus not be evaluated.

The simple splitting system has a higher total number of correct splittings, and is thus better at recall/accuracy. However, the SMOR-based splitting system has a higher precision. In the combined measure F-score, the simple split system is slightly better.

Looking at the number of *unsplit* compounds, it becomes clear that the SMOR-based system employs a much more conservative splitting approach. This is due to several factors: First, some word forms are lexicalized in SMOR and thus remain unsplit, for example *Abend|Land* ('evening country: Occident'). This is often the case for non-compositional compounds, the splitting of which can turn out to be harmful in subsequent applications as their meaning cannot be derived from the parts as is the case with compositional compounds. Additionally, compounds containing a proper noun as modifier are likely not covered by SMOR's lexicon. Furtherore, the splitting approach itself is not designed to cover certain types of splittings, for example *auf_PART fahrt_NN* ('up|drive: driveway'), as particles cannot occur on their own, as opposed to nouns or verbs. The decision whether to split or not in such cases depends entirely on the application. In SMT applications, for example, it is generally assumed that over-splitting does not harm the translation quality, as the system can recover from this by translating split words as a phrase.

Summarizing, we can say that the presented simple splitting approach is competitive with a method relying on a high-quality morphological tool, despite being based only on tagged and lemmatized corpus data in combination with a small set of rules to cover transitional elements. The results show that the system is robust and nearly always produces a splitting analysis. This is due to the fact that it is independent of a hand-crafted lexicon, but rather relies on statistics derived from large corpora. As a result, even compounds containing proper names can be split, for example *Beaufort|skala* ('Beaufort scale') or *Bennett|känguru* ('Bennett kangaroo'). Furthermore, by choosing appropriate corpus data, the splitter can be easily adapted to a new domain.

8 Conclusion

We presented a simple compound splitter for German that relies on form–lemma mappings derived from POS-tagged data to approximate morphological operations. The use of part-of-speech tags restricts the splitting points, and furthermore provides a flat structure of the compounds. To model transitional elements, a small set of hand-crafted rules is defined, that can be extended with a list of words for which certain operations are forbidden.

In an evaluation of splitting performance using a gold standard of bipartite noun compounds, the presented approach performs better than a state-of-the art splitter relying on a high-quality morphological resource. While the SMOR-based approach might be at a slight disadvantage due its different splitting philosophy, the comparison shows that the relatively resource-poor simple approach is competitive, if not better, than a method using rich linguistic information.

9 Download

The compound splitter can be found at `www.ims.uni-stuttgart.de/data/SimpleCompoundSplitter`

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL), from the European Research Council (ERC) under grant agreement No 640550, from the DFG grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation (Phase Two).*

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Peter Eisenberg, Herrmann Gelhaus, Hans Wellmann, Helmut Henne, and Horst Sitta. 1998. *Duden – Grammatik der Deutschen Gegenwartssprache*, volume 4. Dudenverlag, Mannheim, Germany, 6th edition.

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 (WMT)*, pages 224–234, Uppsala, Sweden. Association for Computational Linguistics.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 420–426, Hissar, Bulgaria.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 187–193, Budapest, Hungary.

Klaus Macherey, Andrew M. Dai, David Talbot, Ashok C. Popat, and Franz Och. 2011. Language-independent Compound Splitting with Morphological Operations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404, Portland, Oregon.

Martin Riedl and Chris Biemann. 2016. Unsupervised Compound Splitting With Distributional Semantics Rivals Supervised Methods. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 617–622, San Diego, California.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1263–1266, Lisbon, Portugal.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

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, Gothenburg, Sweden.

Marion Weller and Ulrich Heid. 2012. Analyzing and aligning german compound nouns. In *Proceedings of the the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 2395–2400, Istanbul, Turkey.

Patrick Ziering and Lonneke van der Plas. 2016. Towards unsupervised and language-independent compound splitting using inflectional morphological transformations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–653, San Diego, California.