Influence of Accurate Compound Noun Splitting on Bilingual Vocabulary Extraction

Marcin Junczys-Dowmunt

Abstract. The influence of compound noun splitting on a German-Polish bilingual vocabulary extraction task is investigated. To accomplish this, several unsupervised methods for increasingly accurate compound noun splitting are introduced. Bilingual evidence from a parallel German-Polish corpus and co-occurrence counts from the web are used to disambiguate compound noun analyses directly. These collected splits serve as training data for a probabilistic model that abstracts away from the errors made by the direct methods and reaches an f-measure of 95.10%. Furthermore, these methods are evaluated in terms of word alignment quality and extraction accuracy where linguistically accurate methods are found to outperform the corpus-based methods proposed in the literature. A comparison of alignment quality achieved with the best splitting method and the baseline implies that the effort to build supervised splitting methods might result in minimal or no performance gains.

1 Introduction

During the work on the automatic extraction of German compound nouns and their Polish equivalents from a large parallel corpus, we noticed that the splitting of compound nouns has the most beneficial effect on extraction accuracy. A simple splitting method consistently resulted for all investigated corpus sizes in an improvement of more than 20% compared to attempts where no splitting was applied. Encouraged by this result, we investigate whether further improvements can be achieved when more sophisticated splitting methods are employed. We evaluate several unsupervised methods for compound noun splitting using empirical evidence from the corpus and from the web. Using the "one sense per corpus" assumption (Fung 1998) for compound noun constituents as a base, a probabilistic model of compound nouns is introduced that is trained on the data obtained from the direct methods. A second model that allows for exceptions from the previous assumption in the face of strong evidence is proposed. The probabilistic models are shown to outperform the methods they were trained on and reach results only slightly worse than models trained on manually annotated training data.

Our approach to the extraction of bilingual phrase pairs[1] relies on the application

1. For similar approaches to bilingual vocabulary and terminology extraction see for instance Dagan and Church (1998).

of the alignment models implemented in GIZA++ (Och and Ney 2003). Polish words that have been aligned with a German compound noun are extracted as equivalents of this noun. We compare the influence of the introduced splitting methods on the quality of alignments and equivalence pairs for a test set of compound nouns that have been manually annotated with their Polish counterparts.

2 Corpus statistics

The corpus we use for our extraction task and for the corpus-based splitting methods is the German-Polish part of the third release of the JRC-Acquis parallel corpus (Steinberger et al. 2006). The JRC-Aqcuis is basically a subset of the *Acquis Communautaire*, the total body of European Union law. The German-Polish language pair compromises 23,322 parallel texts with 1,231,766 alignment links between sentences. In order to reduce the vocabulary size, we deleted links that consisted mainly of foreign language material (i.e. other than German or Polish respectively) in either language. After this deletion the German half of the corpus contains 26,704,419 tokens which correspond to 287,754 types, whereas the Polish half consists of 25,405,924 tokens and 221,014 types. Numeric tokens are ignored. In the German half of the corpus 2,163,620 compound noun tokens were collected which correspond to 142,443 compound noun types. Comparing these numbers with the German corpus statistics, we see that only 8.1% of the tokens are compound nouns, but they account for 50.1% of the overall vocabulary. This is consistent with the findings of other researchers in the field. For instance Baroni et al. (2002) identified 7% of the tokens and 47% of the types of a comparably large newswire corpus to be nominal compounds, whilst Schiller (2005) reports lower percentages (5.5% and 43%) for a newspaper corpus.

3 Working definition of a German compound noun

For our needs we define a German compounded word c (not necessarily a noun) as a string consisting of alphabetical characters and optionally a hyphen, written without spaces that can be split up into a sequence $\mathbf{s}_1^n = s_1 s_2 \dots s_n$ of n segments. Segments are required to cover the whole string, but are not allowed to overlap. Segments s_1 to s_{n-1} are called *modifier segments*, the last segment s_n is distinguished and is denoted as the *head segment*. We call a sequence of segments a *segmentation*. The set of possible segmentations for a compound noun c is denoted by $\mathrm{Seg}(c)$.

For every segment s there exists at least one corresponding lexeme l. A sequence $\mathbf{l}_1^n = l_1 l_2 \dots l_n$ of n lexemes where every lexeme l_i corresponds to the segment s_i in the segmentation $\mathbf{s}_1^n$ is called a *decomposition*. Similar to segments, we distinguish

between *modifier lexemes* and *head lexemes*. Additionally, we define $\mathrm{Dec}(\mathbf{s})$ as the set of decompositions corresponding to a segmentation $\mathbf{s}$. The set of all possible decompositions of a compound noun c is defined as $\mathrm{Dec}(c) = \bigcup_{\mathbf{s}\in\mathrm{Seg}(c)} \mathrm{Dec}(\mathbf{s})$.

With the help of these definitions the process of compound noun identification can be reduced to a search for strings for which a decomposition into at least two lexemes exists where the head lexeme is a noun. In order to reduce false indentifications (for instance *Verbraucher* or *folgende*), the word list produced from the German half of the corpus is filtered. Every word that does not begin with a capital letter or is included in a list of known non-compounded words is discarded.

4 Splitting of compound nouns

4.1 Corpus-based splitting method

Koehn and Knight (2003) propose to consider every capitalized word which can be split into two or more words occurring in the German part of the corpus as a compound word. The inventory of segments is limited to corpus words that are tagged as nouns, verbs, adjectives, adverbs, and negation particles. Compound nouns are allowed to be segments themselves. Originally, no distinctions between modifier and head segments are made. However, head segments are limited to words that have been tagged as nouns. By collecting the frequency $C(s)$ of every segment s in the corpus, the best-scored segmentation $\hat{\mathbf{s}}$ is found as follows:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}_1^n\in\mathrm{Seg}(c)}{\operatorname{argmax}} \sqrt[n]{\prod_{k=1}^{n} C(s_k)} \tag{1.1}$$

This method will not split a word if its frequency is higher than the geometrical mean of the frequencies of its segments. This makes sense if we have no knowledge of whether a given word is indeed a compound or just a simple word that could be incorrectly split. On the other hand, a number of compound nouns which could be split correctly may remain unsplit. We will refer to this method as CORPUS.

4.2 Lexicon-based splitting method

The source of the segments for the second method is the German-Polish translation lexicon of the POLENG MT system (Jassem 2006) which provides us with inflectional forms of approximately 90,000 non-composed lexemes. The set of head segments simply consists of all inflected forms of non-composed nouns.

Adding correct nominal modifier segments is a more challenging task. According to Fuhrhop (1998) most productive German linking elements are in fact paradigmatic, i.e. the form of a nominal modifier of a compound corresponds to one or more inflected forms of the noun. Usually this is the base form (in most cases), the plural nominative, or the singular genitive. All of these forms are treated as possible modifier segments. Other phenomena at the segmentation border include the addition of *-s* for each base form and the possible omission of a final *e* or combinations thereof. For verbs, adjectives, adverbs, and numerals the generation of segments is less complicated. Typically it suffices to add the stems and allow for an additional *-e* after verbs with final plosives.

As before we collect frequencies for all segments that can be observed in the corpus, with the remaining segments assigned a frequency of 1. Assigning zero would cause the geometrical mean of the segment frequencies to be zero as well and we would lose the scores implied by other segments in the same segmentation with possibly high frequencies.

$$\hat{\mathbf{s}} = \underset{\mathbf{s}_1^m \in \mathrm{Seg}(c)}{\operatorname{argmax}} \sqrt[m]{\prod_{k=1}^{m} C(s_k)} \qquad (1.2)$$

$$\text{where } m = \min\{n : \mathbf{s}_1^n \in \mathrm{Seg}(c) \wedge n > 1\}$$

The scoring function (1.1) is replaced by (1.2). Following Schiller (2005) we prefer the segmentation with the least elements, but not less than two. If there is more than one such segmentation the geometrical mean of word frequencies is used as a back-up. We denote this splitting method as LEX.

This method is used for the identification of compound nouns and the creation of undisambiguated splits and decompositions. About 36.4% of the found compound nouns have only one decomposition, the rest is ambiguous. More than 10 decompositions are possible for about 7%, a small number of compounds nouns (0.1%) have more than 100 decompositions. In most cases only one decomposition makes sense. The large number of mainly spurious decompositions is a negative effect of the simple approach to the generation of segments described above.

5 Disambiguation of decompositions

5.1 Disambiguation by bilingual evidence

Since we are conducting our experiments with splitting methods on compound nouns originating from a parallel corpus, taking advantage of a bilingual dictionary is straightforward for the disambiguation of splitting results. This has been proposed

by Koehn and Knight (2003) who employ an automatically extracted dictionary in combination with their corpus-based splitting method for the reduction of structural ambiguities.

Contrary to Koehn and Knight (2003) we search for the single best decomposition common for all tokens of one compound noun. For this purpose the translational evidence that is available in all sentence pairs the compound noun appears in is taken into account simultaneously. The best decompositions are those for which evidence for the greatest number of lexemes is found most frequently. If no evidence has been found or if for several lexemes the same number of translations have been identified, the ambiguities are preserved. This disambiguation method is denoted as +DIC.

For this method the choice of an appropriate bilingual dictionary is crucial. Experiments with different dictionaries, hand-crafted and automatically extracted from GIZA++ translation tables, showed that the manually composed POLENG dictionary performs best. The noise in automatically produced dictionaries has a negative impact that persists even when thresholds are used.

5.2 Disambiguation by web-counts

The application of web statistics to the interpretation and bracketing of English compound nouns has been described by Lapata and Keller (2004). We test a similar approach to the disambiguation of splitting options of German compound nouns. Hit counts retrieved from GOOGLE for appropriately constructed queries serve as information about co-occurrences of compound noun segments and corresponding lexemes. The decomposition that receives the highest number of hits and its underlying segmentation are marked as accepted. The methods rely on two types of queries:

- Queries consist of selected forms of the lexemes belonging to a decomposition. For nouns the base form is used, for verbs and adjectives we use inflected forms to avoid confusion with homonymous nouns. This disambiguation method is marked as +WWW$_1$.
- Queries include all search keys from the previous method. Apart from that, the unsplit compound noun is added. This method is named +WWW$_2$.

Both types of search requests can be cascaded in cases where the queries extended with the unsplit compound noun return less than five hits for all decompositions. We then drop the compound noun from the query and repeat the search. This results in the method named +WWW$_3$.

The common weakness of all these approaches is their inability to disambiguate homonymous lexemes for which identical queries are generated. Also differences in the capitalization of the first letter, an important clue for the distinction of nouns from other words, cannot be captured this way.

5.3 Combined disambiguation

All of the disambiguation methods introduced preserve ambiguities if there was not enough evidence to choose a single best lexeme for a segment. We can assume that the described disambiguation methods fail for different compound nouns. For instance, homonymous lexemes can be resolved easily by dictionary look-up provided appropriate entries are available; data sparseness is hardly a problem for the web-based methods, but they cannot deal with homonyms. Therefore a combined approach should improve the general results. We cascade both methods in the following way:

- The first disambiguation method is applied to the analysis produced by a chosen splitting method.
- Only the best scored results are kept. If there are no ambiguities, the single best result has been found.
- If the remaining results are still ambiguous, the second method is applied.

According to our naming convention we label the lexicon-based splitting method as LEX+DIC+WWW$_3$ where the dictionary-based disambiguation method +DIC is applied before the web-based method +WWW$_3$.

6 A probabilistic splitting method

In this section a probabilistic model of compound nouns that uses the splitting knowledge acquired by the direct approaches as training data will be described. In a first step the compound nouns collected are analyzed using the method LEX+DIC+WWW$_3$ and only the best scored results are stored. Table 1 shows a sample of the collected data. $C_{\mathrm{M}}(s)$ denotes the number of times a modifier segment s occurred in all segmentations and $C_{\mathrm{M}}(s,l)$ counts how often a lexeme l was assigned to this segment. Fractional counts are added if lexical ambiguities could not be fully resolved. Obviously $C_{\mathrm{M}}(s) = \sum_l C_{\mathrm{M}}(s,l)$. The counts for heads, $C_{\mathrm{H}}(s)$ and $C_{\mathrm{H}}(s,l)$, are collected analogously.

For the modifier segment *rechts* the lexeme *Recht_N* (*law*) was assigned in 87.7% of the splits, but *Rechte_N* (*right hand* or *person with right-wing views*) was assigned in more than 7.2% of the splits which is still more than we would expect in a corpus of law-related texts. A manual check reveals that all assignments of *Rechte_N* to *rechts* are indeed incorrect, and similarly for *rechts_adv* (*on the right side*) where only *Rechtslenker* (*right-hand drive vehicle*) was correctly analysed. These errors are due to incorrectly generated segments, as for *Rechte_N*, or true ambiguities, as in the case of *rechts_adv*. In 813 out of 814 cases *Recht_N* would have been the correct choice for *rechts*.

Segment	$C_{\mathrm{M}}(s)$	Lexeme	$C_{\mathrm{M}}(s,l)$
rechts	814	Recht_N	731.7
		Rechte_N	58.8
		rechts_adv	23.5
steuer	730	Steuer_N	596.5
		Steuer_N2	103.0
		steuern_V	30.5

Segment	$C_{\mathrm{H}}(s)$	Lexeme	$C_{\mathrm{H}}(s,l)$
rechts	203	Recht_N	202.0
		Rechte_N	1.0
steuer	106	Steuer_N	99.5
		Steuer_N2	6.5

Table 1. Counts for *recht* and *steuer* as modifiers and heads.

6.1 One sense per corpus

This example suggests that the "One sense per corpus" hypothesis introduced by (Fung 1998) in the context of bilingual vocabulary extraction can also be applied to the disambiguation of compound noun constituents. Choosing the most probable lexeme for a segment complies with this hypothesis.

The probability of a modifier segment $Pr_{\mathrm{M}}(s)$ and the probability of a lexeme corresponding to a modifier segment $Pr_{\mathrm{M}}(l|s)$ are calculated as follows

$$Pr_{\mathrm{M}}(s) = \frac{C_{\mathrm{M}}(s)}{\sum_{s'} C_{\mathrm{M}}(s')}, \quad Pr_{\mathrm{M}}(l|s) = \frac{C_{\mathrm{M}}(s,l)}{\sum_{l'} C_{\mathrm{M}}(s,l')} \tag{1.3}$$

using simple Maximum Likelihood Estimates. Again, the probabilities for head segments and corresponding lexemes, $Pr_{\mathrm{H}}(s)$ and $Pr_{\mathrm{H}}(l|s)$, are calculated similarly. Smoothing methods are not applied. We can now express the probability of a single segmentation $\mathbf{s}_1^n$ by

$$Pr(\mathbf{s}_1^n) = \prod_{k=1}^{n-1} Pr_{\mathrm{M}}(s_k) Pr_{\mathrm{H}}(s_n). \tag{1.4}$$

The probabilistic equivalent of the scoring function (1.2) which allows us to choose the best segmentation can be stated as

$$\hat{\mathbf{s}} = \operatorname*{argmax}_{\mathbf{s} \in \mathrm{Seg}(c)} Pr(\mathbf{s}). \tag{1.5}$$

Analogously as for segments, we define the probability of a decomposition $\mathbf{l}_1^n$ given a segmentation $\mathbf{s}_1^n$ as

$$Pr(\mathbf{l}_1^n|\mathbf{s}_1^n) = \prod_{k=1}^{n-1} Pr_{\mathrm{M}}(l_k|s_k) Pr_{\mathrm{H}}(l_n|s_n). \tag{1.6}$$

Finally for the probability of a compound c corresponding to a sequence of lexemes $\mathbf{l}_1^n$ we have

$$Pr(\mathbf{l}_1^n|c) = \sum_{\mathbf{s}_1^n \in \mathrm{Seg}(c)} Pr(\mathbf{s}_1^n) Pr(\mathbf{l}_1^n|\mathbf{s}_1^n). \tag{1.7}$$

Since it is possible (although unlikely) that one decomposition is generated by more than one segmentation of the same compound noun, we sum over all of the found segmentations. Equation (1.7) offers us the means to search for the best decomposition $\hat{\mathbf{l}}$ of a given compound noun without the need of consulting external data. This is done in an analogy as for the best segmentation in equation (1.5) by

$$\hat{\mathbf{l}} = \underset{\mathbf{l} \in \mathrm{Dec}(c)}{\operatorname{argmax}} Pr(\mathbf{l}|c). \tag{1.8}$$

6.2 Weakening independence

As can be seen in equations (1.4) and (1.6) segments on different positions in a segmentation as well as the corresponding lexemes in a decomposition are assumed to be independent. This means that for a given modifier segment the same lexeme is chosen regardless of the choice made for the other segments and *vice versa*. Although this is consistent with the aforementioned "one sense per corpus" assumption for compound noun constituents, we would still prefer to be able to account for exceptions in cases when we have strong evidence for them. One example is *Steuergerät*, where the decomposition *steuern_V Gerät_N* (*steering device*) makes more sense than *Steuer_N Gerät_N* (*tax device*).

In order to weaken independence, we introduce another probability distribution assuming that head lexemes have preferences concerning the semantics of their modifiers, a tendency that has been described by Langer (1998). Since no bracketing methods are used, the linear order of lexemes in a decomposition is the only structural information available and we adopt the oversimplifying assumption that lexemes restrict the semantics of the lexeme that precedes them directly. This is in fact equivalent to expecting a left-branching binary structure for all compound nouns which should be correct for approximately 90%.[2]

The semantic information is obtained from the POLENG lexicon which is organized in an ontology that features more than 130 concepts. In order to avoid data sparseness, the concepts are mapped to a set of 17 concepts from the first three levels of the ontology tree. The concept assigned to a lexeme l is denoted by l^{sem}. Proceeding on the mentioned assumptions, we express the semantic probability of a decomposition $\mathbf{l}_1^n$ by

$$Pr_{\mathrm{SEM}}(\mathbf{l}_1^n) = \prod_{k=2}^{n} Pr(l_{k-1}^{\mathrm{sem}}|l_k) \tag{1.9}$$

2. 72% of the examined compounds are of length 2 where the branching structure is irrelevant. For the rest we can assume that about two thirds are left-branching. This has been shown to be true for complex English compound nouns by Lauer (1995) and we assume a similar distribution for German.

l_1	l_2	$Pr(l_1 l_2 \mid c)$ (*)	l_1^{sem}	$Pr(l_1^{\text{sem}} \mid l_2)$	$Pr(l_1 l_2 \mid c)$ (**)
Steuer_N	*Gerät_N*	**1.9093e-06**	POSSESSION	0.0012	2.3699e-09
Steuer_N2	*Gerät_N*	3.2736e-07	ARTIFACT	0.1207	3.9523e-08
steuern_V	*Gerät_N*	9.6950e-08	ACT	0.5135	**4.9782e-08**

Table 2. Example results for $c =$ *Steuergerät* without (*) and with (**) semantic context

and replace equation (1.7) with

$$Pr(\mathbf{l}_1^n|c) = Pr_{\text{SEM}}(\mathbf{l}_1^n) \sum_{\mathbf{s}_1^n \in \text{Seg}(c)} Pr(\mathbf{s}_1^n) Pr(\mathbf{l}_1^n|\mathbf{s}_1^n). \tag{1.10}$$

As before the empirically disambiguated decompositions are used as training data. Even a frequent lexeme like *Gerät_N* does not co-occur with all semantic concepts. Although we do not use the splitting method on compound nouns from outside the corpus, using MLE for the estimation of $Pr(l_{i-1}^{\text{sem}}|l_i)$ might introduce a number of zero probabilities in equation (1.9), due to data sparseness, and worse, due to errors produced by the empirical disambiguation methods. In order to avoid this influence of Pr_{SEM} on the scoring function (1.10), we apply smoothed counts computed with the help of the Simple Good-Turing (SGT) method (Gale 1994), calculating

$$Pr(l_{i-1}^{\text{sem}}|l_i) = \frac{C_{\text{SGT}}(l_{i-1}^{\text{sem}}, l_i)}{\sum_{l'_{i-1}} C(l'_{i-1} l_i)}. \tag{1.11}$$

Table 2 illustrates the differences between the two probabilistic models defined by equations (1.7) and (1.10) for our previous example *Steuergerät*. As said before, the decomposition that maximizes equation (1.7) is *Steuer_N Gerät_N*. Introducing contextual knowledge by equation (1.10) allows us to identify *steuern_V Gerät_N* as a better decomposition. Since the semantic concept POSSESSION was unseen for any lexeme preceding *Gerät_N* and the probability for the concept ACT appearing before *Gerät_N* is very high, we have strong evidence for an exception from the "one sense per corpus" assumption. In most cases however, the influence of the additional semantic context is only marginal.

7 Evaluation

In this section we present the evaluation of the proposed splitting methods under two different aspects: the linguistic correctness of the analyses and the influence of the splitting methods on the quality of bilingual alignment. The alignments are computed with GIZA++ (Och and Ney 2003). We use refined alignments produced from two alignment model trained in both directions as has been proposed by Och and Ney (2003). The input to GIZA++ has been lemmatized in order to reduce the size of the vocabulary.

Splitting method	Segmentations	Decompositions		
	Accuracy (%)	P (%)	R (%)	F (%)
LEX (baseline)	98.70	67.23	97.74	79.66
LEX+DIC	98.70	78.59	96.93	86.80
LEX+WWW$_1$	98.70	89.10	90.70	89.89
LEX+WWW$_2$	99.00	86.58	93.92	90.10
LEX+WWW$_3$	99.00	90.67	93.21	91.92
LEX+DIC+WWW$_3$	99.00	92.38	94.31	93.34
LEX+PROB	99.30	94.35	94.73	94.54
LEX+PROB+SEM	**99.30**	**95.05**	**95.15**	**95.10**
CORPUS	72.15	40.70	62.39	49.26
CORPUS+DIC	72.15	44.93	59.62	51.24

Table 3. Percentages for splitting accuracy, precision, recall and f-measure

7.1 Compound noun splitting

The performance of the described methods is evaluated on a test set of $N = 1000$ compound nouns that have been manually annotated with correct segmentations and decompositions. For segmentations accuracy is the only used measure. Correct segmentations (cr) are those for which all splitting points are identical with the splitting points in the manually created split. Accuracy is then calculated as $A = cr/N$.

In order to compare our results with a supervised splitting method, we use the same evaluation scheme for decompositions as has been proposed by Schiller (2005). Only "best scored" decompositions are taken into account. If there is no scoring method for decompositions (see CORPUS or LEX), all decompositions of the best segmentation are considered as results.

Among the set of results returned for one compound noun, the analyses which are identical with the manually disambiguated decomposition are true positives (tp) — in our case there is at most one for each compound noun. All other analyses that do not match the manual choice count as false positives (fp). Additionally a false negative (fn) is counted if the manual analysis is not among the results. Given the above values for all compound nouns in the test set, we calculate the overall precision (P), recall (R), and f-measure (F) in the standard way, where $P = tp/(tp + fp)$, $R = tp/(tp + fn)$, and $F = (2 \cdot P \cdot R)/(P + R)$.

Table 3 summarizes the results for the proposed splitting and disambiguation methods. As for segmentations, the task of correctly choosing all splitting points in a compound noun, all approaches based on the LEX method perform reasonably well.

For the web-based methods it seems reasonable to cascade disambiguation methods in such a way that the method with higher precision which is simultaneously more prone to data sparseness precedes the less precise but more robust method.

This is also true for the combination of dictionary look-up and web-counts which results in the best empirical splitting method LEX+DIC+WWW$_3$ with an f-measure of 93.34%.

Our claims from section 6.1 concerning the validity of the "one sense per corpus" hypothesis for compound noun elements seem to be confirmed by the results achieved by the probabilistic models. Abstracting from the data obtained from the unsupervised methods leads to an improved precision for LEX+PROB compared to the best empirical method. Method LEX+PROB+SEM additionally incorporates knowledge about exceptions from the above assumption and performs best among all investigated unsupervised approaches to compound noun splitting and analysis.

As for supervised methods, Schiller (2005) uses weighted finite state automata trained on two large sets of manually split and disambiguated compound nouns from newspaper texts. For an in-domain test set f-measures between 98.32% and 98.38% are reported. Our best result for decompositions, an f-measure of 95.10%, is not as high but still acceptable.

The results for the CORPUS and CORPUS+DIC methods are given for reasons of completeness. Since the aim of these methods is to establish one-to-one correspondences between bilingual equivalents rather than to provide linguistically correct analyses, they cannot be compared directly to our splitting methods. These results are nevertheless significant when we compare the alignment quality for different splitting methods in the next section.

7.2 Word alignment and extraction quality

We now turn to the original question how compound noun splitting quality affects the alignment quality for compound nouns. For the alignment and extraction task we annotated 1000 German compound nouns with their Polish equivalents. The compound nouns were randomly chosen from the identified compound nouns types, after which one token was selected for each type. There are no repetitions in the test set and it is distinct from the test set used in the previous section. In compliance with Och and Ney (2003) we distinguish between sure alignments (S) and possible alignments (P, where $S \subseteq P$) that additionally describe ambiguities, such as function words that are missing in the other language. We do not annotate the whole sentence that contains the compound noun, but only the alignment points associated to the given compound noun itself.

Och and Ney propose the following measures to calculate precision, recall and alignment error rate (AER) for the sets S and P from the test set and an alignment A obtained from the word alignment process:

$$\mathrm{P} = \frac{|A \cap P|}{|A|}, \quad \mathrm{R} = \frac{|A \cap S|}{|S|}, \quad \mathrm{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}. \tag{1.12}$$

Splitting method	100K sentences		500K sentences		1.2M sentences	
	AER	EA	AER	EA	AER	EA
No splitting	56.38	21.5	48.19	27.9	—	—
CORPUS	31.87	37.9	25.51	45.6	—	—
CORPUS+DIC	31.61	38.2	25.49	45.7	—	—
LEX	26.64	43.8	20.91	51.2	17.58	56.2
LEX+PROB+SEM	25.99	44.8	20.56	51.4	17.57	56.1

Table 4. Percentages for Alignment Error Rate and Extraction Accuracy for chosen splitting methods

We define an additional measure for the quality of the extraction of the German-Polish translation pairs. Extraction accuracy (EA) is the percentage of compound nouns from the test set that were correctly and completely aligned with their sure Polish equivalent. These are raw numbers calculated from the original alignment data before the application of additional filtering and reconstruction techniques.

Table 4 presents the results for chosen splitting methods and various corpus sizes.[3] The general positive impact of compound noun splitting is obvious for all corpus sizes and was to be expected. What is more surprising is the significant superiority of the lexicon-based methods compared to the corpus-based splitting approaches. It can be clearly seen that one-to-one correspondence is less beneficial than linguistically motivated splitting. The improvement introduced by the linguistically more correct method LEX+PROB+SEM over the baseline method LEX is more significant for the smaller sets of training sentences and decreases with increasing corpus size.

8 Conclusions

We have shown two things. Firstly, unsupervised methods for compound noun splitting can reach results close to the performance of methods trained on manually disambiguated data. However, it would be interesting to see how the probabilistic models behave when the size of the training data varies. Approximately 140,000 compound nouns were disambiguated using the described empirical methods. Since no human interaction is required, it is no problem at all to increase the amount of training data, for instance with the release of a future, larger version of the JRC-Acquis.

Secondly, splitting methods that aim for full linguistic analyses of compound nouns achieve better results in alignment quality than methods that try to establish

3. For technical reasons for the complete corpus results are evaluated for the baseline and the best splitting method only. The computation for five different splitting methods with two directional models each would have been too time consuming.

one-to-one correspondences. The reasons for this may be due to better coverage of segments, the consistency of splits for all tokens of the same compound noun, and a reduced vocabulary since no compound noun remains unsplit. For smaller corpora the linguistically most accurate splitting method achieves better results than our baseline method, but the effect diminish with increasing corpus size. For large corpora the perfomance jump of 15% for splitting quality is not reflected in alignment quality or extraction accuracy. The statistical alignment models seem to be able to cope with minor inaccuracies concerning compound noun splitting. This implies that unsupervised methods are sufficiently accurate and no improvement could be achieved by employing models trained on better and therefore more costly data.

References

Baroni, Marco, Johannes Matiasek, and Harald Trost (2002). Predicting the components of German nominal compounds. In *Proceedings of the 15th European Conference on Artificial Intelligence, ECAI'2002*, 470–474.

Dagan, Ido and Ken Church (1998). Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation* 12(1-2):89–107.

Fuhrhop, Nanna (1998). *Grenzfälle morphologischer Einheiten*. Tübingen, Germany: Stauffenburg.

Fung, Pascale (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, 1–17.

Gale, William (1994). Good-Turing smoothing without tears. Statistics Research Reports from AT&T Laboratories 94.5, AT&T Bell Laboratories.

Jassem, Krzysztof (2006). *Przetwarzanie tekstów polskich w systemie tłumaczenia automatycznego POLENG*. Poznań, Poland: Wydawnictwo Naukowe UAM.

Koehn, Philipp 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).*, 347–354.

Langer, Stefan (1998). Zur Morphologie und Semantik von Nominalkomposita. In *Tagungsband der 4. Konferenz zur Verarbeitung natürlicher Sprache*, 83–97.

Lapata, Mirella and Frank Keller (2004). The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 121–128.

Lauer, Mark (1995). Corpus statistics meet the noun compound: some empirical results. In *Proceedings of the 33rd annual meeting of the Association for Computational Linguistics*, 47–54.

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

Schiller, Anne (2005). German compound analysis with *wfsc*. In *Finite-State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005*, 239–246.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR* abs/cs/0609058, informal publication.