Normalizing German and English Inflectional Morphology to Improve Statistical Word Alignment

Simon Corston-Oliver and Michael Gamon

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{simonco, mgamon}@microsoft.com

Abstract. German has a richer system of inflectional morphology than English, which causes problems for current approaches to statistical word alignment. Using Giza++ as a reference implementation of the IBM Model 1, an HMM-based alignment and IBM Model 4, we measure the impact of normalizing inflectional morphology on German-English statistical word alignment. We demonstrate that normalizing inflectional morphology improves the perplexity of models and reduces alignment errors.

1 Introduction

The task of statistical word alignment is to identify the word correspondences that obtain between a sentence in a source language and the translation of that sentence into a target language. Of course, fluent translation performed by expert human translators involves reformulation that obscures word alignment. However, in many domains, automatically identified word alignments serve as an important source of knowledge for machine translation.

We describe a series of experiments in which we apply morphological normalizations to both the source and target language before computing statistical word alignments. We consider the case of aligning English and German, two closely related languages that differ typologically in ways that are problematic for current statistical approaches to word alignment.

We perform a series of experiments using the Giza++ toolkit (Och and Ney, 2001). The toolkit provides an implementation of IBM Model 1 and Model 4 (Brown et al., 1993) as well as an HMM-based alignment model (Vogel, Ney and Tillman, 1996), together with useful metrics of model perplexity. We perform five iterations of IBM Model 1, which attempts to find simple word translations without consideration of the position of the words within the sentence. The word-alignment hypotheses yielded by this first stage serve as input for five iterations of HMM alignments, which in turn serve as input for five iterations of IBM Model 4. Model 4, which models phenomena such as the relative order of a head and a modifier, is the most sophisticated model considered here. Clustering of words was performed using JCLUSTER (Goodman, 2001).

These word alignment models take a naïve view of linguistic encoding. Sentences are conceived of as little more than a sequence of words, mapped one-to-one or one-to-N from the source language to the target language. Recent research has attempted

to improve machine translation by considering the linguistic structure that exists between the level of the word and the level of the sentence (see, for example, Alshawi, Bangalore and Douglas, 2000; Marcu and Wong, 2002; Koehn et al., 2003). Relatively little research has been directed towards considerations of the role of word-internal structure.

Brown et al. (1993), considering the case of English-French machine translation, perform some orthographic regularizations such as restoring the elided *e* at the end of the relative pronoun *qu'*, or separating the portmanteau *des* "of.the.PLURAL" into its components *de les*. They also speculate that additional morphological analysis to identify the relations among inflected forms of verbs might improve the quality of their models. Not until very recently have results been reported on evaluating the improvements obtainable through morphological processing, the most recent being the work by Nießen and Ney (2004) and Dejean et al. (2003).

Before presenting the experimental results, we briefly outline the salient morphological differences between English and German.

2 Morphological Facts

English and German are historically related; both languages are in the Western branch of the Germanic family of Indo-European. Despite this close historical relation, the modern-day languages differ typologically in ways that are problematic for statistical approaches to word alignment.

German has pervasive productive noun-compounding. English displays its Germanic roots in the analogous phenomenon of the noun group—sequences of nouns with no indication of syntactic or semantic connection. As a general rule, English noun groups translate in German as noun compounds. The converse does not always obtain; German compounds occasionally translate as simple English nouns, other times as nouns with prepositional, adjectival, or participial modifiers. When using models such as those of Brown et al. (1993), which allow one-to-one or one-to-N alignments, we would expect this asymmetry to result in poor alignment when English is the source language and German is the target language.

The order of constituents within the clause is considerably more variable in German and long distance dependencies such as relative clause extraposition are more common than in English (Gamon et al., 2002). In German, so-called separable verb prefixes may occur bound to a verb or may detach and occur in long distance relationships to the verb. Adding to the confusion, many of these separable prefixes are homographic with prepositions.

The languages differ greatly in the richness of their inflectional morphologies. Both languages make a three way distinction in degree of adjectives and adverbs. In nominal inflections, however, English makes only a two way distinction in number (singular vs. plural) whereas German makes a two way distinction in number (singular and plural), a four way distinction in grammatical case (nominative, accusative, genitive and dative) and a three way distinction in lexical gender (masculine, feminine, neuter). Nominal case is realized in the German noun phrase on

the noun, the determiner and/or pre-nominal modifiers such as adjectives. Vestiges of this case marking remain in the English pronominal system, e.g. *I/me/my*.

The languages have similar systems of tense, mood and aspect. Verbal inflection distinguishes past versus non-past, with weak vestiges of an erstwhile distinction between subjunctive and indicative mood. Many complexes of tense, aspect and mood are formed periphrastically. The most notable difference between the two languages occurs in the morphological marking of person and number of the verb. Aside from the irregular verb *be*, English distinguishes only third-person singular versus non-third-person singular. German on the other hand distinguishes first, second and third person by means of inflectional suffixes on the verb. In the data considered here, drawn from technical manuals, first and second person inflections are extremely uncommon.

3 The Problem of Morphology

Let us now consider how these linguistic facts pose a problem for statistical word alignment. As previously noted, the correspondence between an English noun group and a German noun compound gives rise to an N-to-one mapping, which the IBM models do not allow. Differences in constituent order, however, are really only a problem when decoding, i.e. when applying a statistical machine translation system: it is difficult to model the movement of whole constituents by means of distortions of words.

The homography of separable prefixes and prepositions adds interference when attempting word alignment.

The most glaring deficiency of the IBM models in the face of the linguistic facts presented above concerns related word forms. The models do not recognize that some words are alternate forms of other words, as opposed to distinct lexical items. To put this another way, the models conflate two problems: the selection of the appropriate lexical item and the selection of the appropriate form, given the lexical item.

Since the models do not recognize related word forms, the effect of inflectional morphology is to fragment the data, resulting in probability mass being inadvertently smeared across related forms. Furthermore, as Och and Ney (2003) observe, in languages with rich morphology, a corpus is likely to contain many inflected forms that occur only once. We might expect that these problems could be resolved by using more training data. Even if this were true in principle, in practice aligned sentences are difficult to obtain, particularly for specific domains or for certain language pairs. We seek a method for extracting more information from limited data using modest amounts of linguistic processing.

With this brief formulation of the problem, we can now contrast the morphological operations of this paper with Nießen and Ney (2000), who also consider the case of German-English word alignment. Nießen and Ney perform a series of morphological operations on the German text. They reattach separated verbal prefixes to the verb, split compounds into their constituents, annotate a handful of high-frequency function words for part of speech, treat multiword phrases as units, and regularize words not

seen in training. The cumulative effect of these linguistic operations is to reduce the subjective sentence error rate by approximately 11-12% in two domains.

Nießen and Ney (2004) describe results from experiments where sentence-level restructuring transformations such as the ones in Nießen and Ney (2000) are combined with hierarchical lexicon models based on equivalence classes of words. These equivalence classes of (morphologically related) words have the same translation. The classes are obtained by applying morphological analysis and discounting morphological tags that do not change the translation into the target language. The statistical translation lexicon which results from clustering words in equivalence classes is considerably smaller (65.5% on the Verbmobil corpus).

The morphological operations that we perform are the complement of those performed by Nießen and Ney (2000). We do not reattach separated verbal prefixes, split compounds, annotate function words for part of speech, merge multiword phrases or regularize unseen words. Rather, we normalize inflectional morphology, reducing words to their citation form. Since it is not clear what the citation form for German determiners ought to be, we normalize all forms of the definite article to the nonce word *DefDet*, all forms of the indefinite article to *IndefDet*, and all demonstratives to *Proxldet* ("proximal determiner") and *DistlDet* ("distal determiner"). We perform one additional operation on all German text, i.e. even in the scenarios characterized below as involving no inflectional normalization, we separate contractions into their constituents, in a similar fashion to what Brown et al. (1993) do for French. For example, the portmanteau *zum* "to.the.DATIVE" is replaced with the two words *zu dem*. When morphological regularization is applied this is then rendered as *zu DefDet*.

The following examples illustrate the effects of morphological processing. Words that are stemmed are shown in italics.

English

Before. If your computer is connected to a network, network policy settings may also prevent you from completing this procedure.

After. if your computer *be connect* to *Indefdet* network, network policy *setting* may also prevent you from *complete Proxldet* procedure.

German

Before. Anwendungen installieren, die von Mitgliedern der Gruppe Benutzer erfolgreich ausgeführt werden können.

After. *Anwendung* installieren, die von *Mitglied DefDet* Gruppe Benutzer erfolgreich *ausführen* werden können.

Aligned English sentence. Install applications that Users can run successfully.

4 Data

We measured the effect of normalizing inflectional morphology on a collection of 98,971 aligned German-English sentence pairs from a corpus of technical manuals and help files for computer software. The content of these files is prosaic and the translations are, for the most part, fairly close.

As Table 1 shows, while the number of words is nearly identical in the German and English data sets, the vocabulary size in German is nearly twice that of the English, and the number of singletons in German is more than twice that of English.

Table 1. Corpus profile

	German	English
Words	1,541,002	1,527,134
Vocabulary	53,951	27,959
Singletons	26,690	12,417

5 Results

We perform stemming on the English and German text using the NLPWin analysis system (Heidorn, 2000). In the discussion below we consider the perplexity of the models, and word error rates measured against a gold standard set of one hundred manually aligned sentences that were sampled uniformly from the data.

The stemmers for English and German are knowledge-engineered components. To evaluate the accuracy of the stemming components, we examined the output of the stemmer for each language when applied to the gold standard set of one hundred sentences. We classified the stems produced as good or bad in the context of the sentence, focusing only on those stems that actually changed form or that ought to have changed form. Cases where the resulting stem was the same as the input, e.g. English prepositions or singular nouns or German nouns occurring in the nominative singular, were ignored. Cases that ought to have been stemmed but which were not in fact stemmed were counted as errors.

The English file contained 1,489 tokens; the German analogue contained 1,561 tokens. [1] As Table 2 shows, the effects of the morphological processing were overwhelmingly positive. In the English test set there were 262 morphological normalizations, i.e. 17.6% of the tokens were normalized. In German, there were 576 normalizations, i.e. 36.9% of the tokens were normalized. Table 3 presents a breakdown of the errors encountered. The miscellaneous category indicates places where unusual tokens such as non-breaking spaces were replaced with actual words, an artifact of tokenization in the NLPWin system. Compared to Table 1 morphological normalization reduces the number of singletons in German by 17.2% and in English by 7.8%.

[1] Punctuation other than white space is counted as a token. Throughout this paper, the term “word alignment” should be interpreted to also include alignments of punctuation symbols.

Table 2. Accuracy of morphological processing

	English	German
Good	248	545
Bad	14	31
Error %	4.5%	5.4%

Table 3. Analysis of morphological errors

	English	German
Failed to stem	1	20
Should not have stemmed	0	5
Wrong stem	7	5
Miscellaneous	6	1

As noted in the introduction, we used Giza++ to compute statistical word alignments for our data. We performed five iterations of IBM Model 1, followed by five iterations of HMM, and then five iterations of IBM Model 4. To evaluate the effect of stemming, we measured the perplexity of the final Model 4.

Raw perplexity numbers of the final Model 4 are not comparable across the different morphological processing scenarios we want to investigate, however. Perplexity is tied to vocabulary size, and if the vocabulary size changes (as it does as a result of morphological stemming), perplexity numbers change. In order to overcome vocabulary size dependence of the results, we use the differential perplexity between the Model 4 and a uniform model operating on the same vocabulary. Below we illustrate how this amounts to simply scaling the perplexity number by the target vocabulary size.

Perplexity PPL_{M4} regarding the model 4 probability P_i^{M4} for a target word w_i is:

$$PPL_{M4} = \exp\left[-\frac{1}{N}\sum_{i=1}^{N}\log P_i^{M4}\right] \quad \textbf{(1)}$$

where N is the size of the sample.

A uniform probability distribution for translation would always assign the probability $P_u = P(w_i) = 1/V$ to a target word, where V is the size of the target vocabulary. Perplexity based on the uniform model is defined as follows.

$$PPL_u = \exp\left[-\frac{1}{N}\sum_{i=1}^{N}\log P_u\right] = \exp\left[-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{V}\right)\right] \quad \textbf{(2)}$$

$$= \exp\left[-\frac{1}{N}N\log\left(\frac{1}{V}\right)\right] = V$$

We define the differential perplexity $DIFFPPL$ as the ratio of the perplexities PPL_{M4} and PPL_u which is equivalent to dividing the original perplexity PPL_{M4} by V:

$$DIFFPPL = PPL_{M4} / PPL_u = \frac{PPL_{M4}}{V} \quad (3)$$

In the remainder of this paper we will, for the sake of convenience, refer to this differential perplexity simply as "perplexity".

We compute word alignments from English to German and from German to English, comparing four scenarios: None, Full, NP and Verb. The "None" scenario establishes the baseline if stemming is not performed. The "Verb" scenario performs stemming only on verbs and auxiliaries. The "NP" scenario performs stemming only on elements of the noun phrase such as nouns, pronouns, adjectives and determiners. The "Full" scenario reduces all words to their citation forms, applying to verbs, auxiliaries, and elements of the noun phrase as well as to any additional inflected forms such as adverbs inflected for degree. We remind the reader that even in the scenario labeled "None" we break contractions into their component parts. The results of stemming are presented in Figure 1 and Figure 2. For ease of exposition, the axes in the two figures are oriented so that improvements (i.e. reductions) in perplexity correspond to bars projected above the baseline. Bars projected below the baseline have a black top at the point where they meet the base plane. The base pane indicates the model perplexity when no stemming is performed in either language.

As Figure 1 illustrates, E-G perplexity is improved across the board if stemming is performed on the target language (German). If no stemming is done on the German side, stemming on the source language (English) worsens perplexity. Interestingly, the stemming of German verbs causes the largest improvements across all English stemming scenarios.

Figure 2 shows a remarkably different picture. If English is the target language, any stemming on either the German source or the English target yields worse perplexity results than not stemming at all, with the exception of tiny improvements when full stemming or verb stemming is performed on English.

The difference between the two graphs can be interpreted quite easily: when the target language makes fewer distinctions than the source language, it is easier to model the target probability than when the target language makes more distinctions than the source language. This is because a normalized term in the source language will have to align to multiple un-normalized words in the target across the corpus, smearing the probability mass.

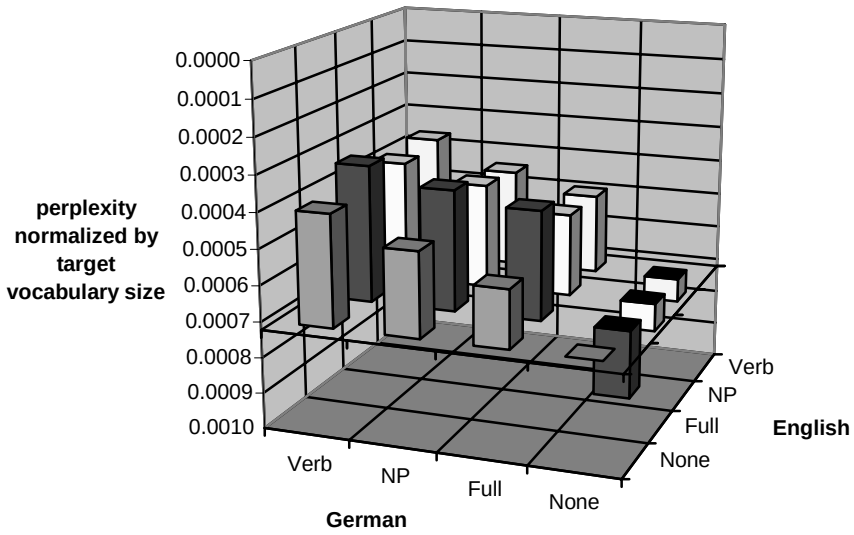

Fig. 1. English-to-German alignment

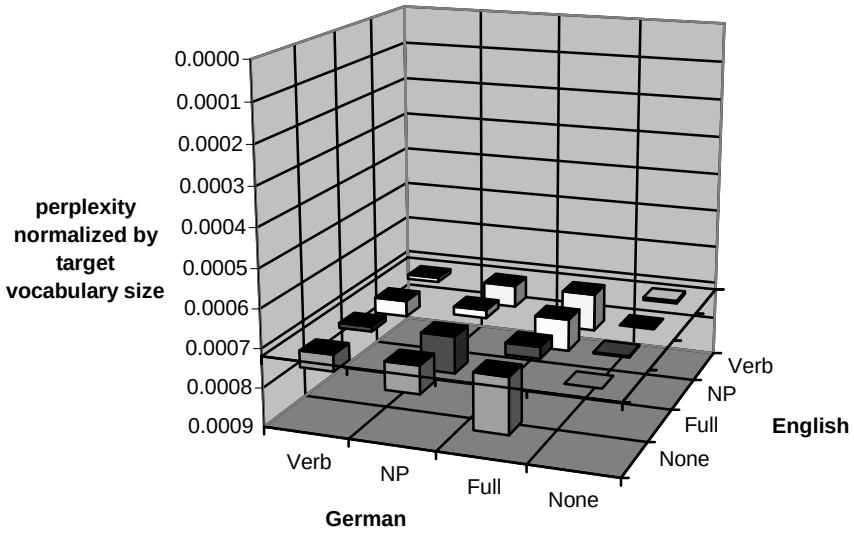

Fig. 2. German-to-English alignment

In order to assess the impact of these morphological operations on word alignment we manually annotated two sets of reference data. In one set of reference data, no

stemming had been performed for either language. In the other set, full stemming had been applied to both languages. The manual annotation consisted of indicating word alignments that were required or permissible (Och and Ney 2000, 2003). We then evaluated the alignments produced by Giza++ for these sentence pairs against the manually annotated gold standard measuring precision, recall and alignment error rate (AER) (Och and Ney 2003). Let A be the set of alignments produced by Giza++, S be the set of sure (i.e. required) alignments and P the set of possible alignments. The definition of precision, recall and AER is then:

$$\text{precision} = \frac{|A \cap P|}{|A|};\ \text{recall} = \frac{|A \cap S|}{|S|} \qquad (4)$$

$$\text{AER} = \frac{|A \cap P + A \cap S|}{|A + S|}$$

The results are presented in Table 4. Full stemming improves precision by 3.5% and recall by 7.6%. The alignment error rate is reduced from 20.63% to 16.16%, a relative reduction of 21.67%.

Table 4. Statistical word alignment accuracy

	No stemming	Full stemming
Precision	87.10%	90.24%
Recall	72.63%	78.15%
Alignment error rate	20.63%	16.16%

Note that the alignment error rates in Table 4 are much larger than the ones reported in Och and Ney (2003) for the English-German Verbmobil corpus. For the closest analogue of the Giza++ settings that we use, Och and Ney report an AER of 6.5%. This discrepancy is not surprising, however: Our corpus has approximately three times as many words as the Verbmobil corpus, more than ten times as many singletons and a vocabulary that is nine times larger.

6 Discussion

As noted above, the morphological operations that we perform are the complement of those that Nießen and Ney (2000) perform. In future research we intend to combine stemming, which we have demonstrated improves statistical word alignment, with the operations that Nießen and Ney perform. We expect that the effect of combining these morphological operations will be additive.

Additional work remains before the improved word alignments can be applied in an end-to-end statistical machine translation system. It would be most unsatisfactory to present German readers, for example, with only the citation form of words. Now that we have improved the issue of word choice, we must find a way to select the contextually appropriate word form. In many instances in German, the selection of

word form follows from other observable properties of the sentence. For example, prepositions govern certain cases and verbs agree with their subjects. One avenue might be to apply a transformation-based learning approach (Brill, 1995) to selecting the correct contextual variant of a word in the target language given cues from the surrounding context or from the source language.

Acknowledgements

Our thanks go to Chris Quirk and Chris Brockett for technical assistance with Giza++, and to Ciprian Chelba and Eric Ringger for discussions regarding the normalization of perplexity.

References

Alshawi, Hiyan, Shona Douglas, Srinivas Bangalore. 2000. Learning dependency translation models as collections of finite-state head transducers. *Computational Linguistics* 26(1):45-60.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case-study in part of speech tagging. *Computational Linguistics* 21(4):543-565.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263-311.

Dejean, Herve, Eric Gaussier, Cyril Goutte and Kenji Yamada. 2003. Reducing Parameter Space for Word Alignment. In *Proceedings from the HLT-NAACL 2003 workshop on Building Parallel Texts*. 23-26.

Gamon M., Ringger E., Zhang Z., Moore R., Corston-Oliver S. 2002. Extraposition: A case study in German sentence realization. In *Proceedings of COLING 2002*. 301-307.

Goodman, J. 2001. A Bit of Progress in Language Modeling, Extended Version. Microsoft Research Technical Report MSR-TR-2001-72.

Heidorn, George. 2000. Intelligent Writing Assistance. In R. Dale, H. Moisl and H. Somers, (eds.), *Handbook of Natural Language Processing*. Marcel Dekker.

Marcu, Daniel and William Wong. 2002. A Phrase-Based, Joint Probability Model for Statistical Machine Translation. EMNLP-02.

Nießen, Sonja and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. COLING '00: The 18th International Conference on Computational Linguistics. 1081-1085.

Nießen, Sonja and Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics* 30(2): 181-204.

Och, F. and H. Ney. 2000. Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. 440-447.

Och, F. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19-52.

Vogel, Stephan, Hermann Ney, Christoph Tillman. HMM-based word alignment in statistical translation. Proceedings of COLING '96: The 16th International Conference on Computational Linguistics. 836-841.