The pay-offs of preprocessing for German-English Statistical Machine Translation

İlknur Durgar El-Kahlout, Francois Yvon

Univ. Paris-Sud and LIMSI-CNRS, BP 133, 91430
Orsay cedex, France
Ilknur.Durgar@limsi.fr, Francois.Yvon@limsi.fr

Abstract

In this paper, we present the result of our work on improving the preprocessing for German-English statistical machine translation. We implemented and tested various improvements aimed at i) converting German texts to the new orthographic conventions; ii) performing a new tokenization for German; iii) normalizing lexical redundancy with the help of POS tagging and morphological analysis; iv) splitting German compound
words with frequency based algorithm and; v) reducing singletons and out-of-vocabulary words. All these steps are performed during preprocessing on the German side. Combining all these processes, we reduced 10% of the singletons, 2% OOV words, and obtained 1.5 absolute (7% relative) BLEU improvement on the WMT 2010 German to English News translation task.

1. Introduction

Most state-of-the-art statistical machine translation (SMT) systems rely on word forms to estimate their translation models and to perform translation. Using unanalyzed token poses a number of difficulties, especially when the source and target languages are morphologically different, as in the case for German and English, the pair under focus in this study.

First, morphological variation, especially on the German side, tends to blur the alignment regularities and to increase the level of noise in the phrase table. For instance, German word *Berg* has four forms with different case information *Berg, Berg, Berge, Berges* and all these forms tend to be aligned to English noun *mountain*. This fact is aggravated here by the compounding processes, which are again specially productive on the German side. Compounds are composed by concatenating word lemmas and there is no theoretical limit on the number of lemmas in a German compound word. This typically creates situations where one German word corresponds to several English words, a problematic configuration when aligning German with English. For instance, the German compound word;

- *Rinder*$_1$*kennzeichnungs*$_2$*- und*$_3$ *Rindfleisch*$_4$ *etikettierungs*$_5$*überwachungs*$_6$*aufgaben*$_7$ *übertragungs*$_8$ *gesetz*$_9$

which can be (literally) translated to the English phrase;

- *Cattle*$_1$ *marking*$_2$ *and*$_3$ *beef*$_4$ *labeling*$_5$*supervision*$_6$ *duties*$_7$ *delegation*$_8$ *law*$_9$

Morphological productivity also means more out-of-vocabulary (OOV) words, which cannot be translated, at test time. Finally, morphological mismatch between source and target also increases the level of lexical ambiguity when translating into the morphologically more complex language. For instance, when translating from English into German, the English phrases often lack the contextual information (eg. regarding agreement) that is needed to select the correct German phrase. As a result, for all practical reasons, German lexicon is large which causes serious data sparseness issues. In table 2, data sparseness problem is observed obviously. German has less total number of words but has many more distinct word forms (about 2.5 times).

These problems are well documented and can be addressed using morphological and/or syntactic information during a preprocessing and/or postprocessing step; such methods have indeed often proven effective in isolation for improving the overall machine translation quality.

In this paper, we combined a series of systematic experiments to investigate the effects of: i) converting old-written German words to new writing style according to the new orthography reform; ii) performing a new language-pair specific tokenization for German; iii) normalizing lexical redundancy by performing different level of lemmatization and pseudo-tags with the help of POS tagging and morphological analysis; iv) splitting German compounds with corpus-driven splitting algorithm and; v) reducing singletons and OOV words. All processes were performed as a preprocessing step on German. The main contribution of this paper is thus a systematic assessment of the effects of preprocessing in German, which shows that paying attention to these small, and often overlooked details, can results in gains that largely exceed those obtained with fancy translation models.

2. Related Work

Using morphology in statistical machine translation has been addressed by several researchers for translation from or into morphologically complex languages. Morphological preprocessing is all the more useful that the parallel resource is scarce. For the German-English language pair, Niessen and Ney [1] use morphological decomposition with base forms

and POS tags to introduce a hierarchical lexicon model, which improves translation results. Corston-Oliver and Gamon [2] and Koehn [3] normalize inflectional morphology by replacing word forms with stems both in German and English. Morphological analysis is also useful at test time: Yang and Kirchhoff [4] discuss the use of phrase-based backoff models to provide translations for unknown words, through morphological decomposition.

German is not the only problematic language and morphological analysis/decomposition has also proven useful for many other languages. For instance, Lee [5] uses a morphologically analyzed and tagged parallel corpus for Arabic-English SMT. Sadat and Habash [6] and Zollmann et *al.* [7] also exploit morphology in Arabic-English statistical machine translation. Popovic and Ney [8] investigate various ways of improving translation quality from inflected languages Spanish, Catalan and Serbian by using stems, suffixes and part-of-speech tags. Goldwater and McClosky [9] use morphological analysis on the Czech side and introduce lemmas and pseudo words in Czech to English SMT. Talbot and Osborne [10] reduce source and target vocabulary by clustering related words to translate from Czech, French and Welsh. Recently, Carpuat [11], working on a French to English SMT system, proposes to replace words from specific morphological classes with their lemmas.

Researches on exploiting morphology is generally focused on translating morphologically rich languages into English. The reverse translation direction is studied, for instance, in Minkov et *al.* [12] who use morphological postprocessing *on the target side* using structural information and information from the source side, to improve translation quality of translation into Russian and Arabic. Durgar El-Kahlout and Oflazer [13] use morphological analysis to separate some Turkish inflectional morphemes that have counterparts on the English side.

3. German Preprocessing

German is a West Germanic language belonging to the Indo-European language family. German is a member of highly inflected languages, in which compounding is also very productive: these two interesting linguistic properties of German contribute to make the translation from and into German a real challenge for SMT systems. In this section, we detail the various preprocessing steps that are applied to reduce the vocabulary and remove redundant inflections.

3.1. Spelling/Orthography Reform

In 1996, German-speaking countries agreed on an orthography reform (*Rechtschreibreform*) to unify German spelling and introduced systematic rules to reduce the ambiguities for the letter to sound correspondences, capitalization and the use of hyphen and punctuation.

Converting "old" Europarl data to the new spelling was first addressed by Fraser [14]. Fraser splits the training corpus into old and new portions based on the occurrences of the very common word *dass* and its old writing *daß*. He then mapped words that were identical except for specific characters such as *ß/ss, ue/ü, ae/ä and oe/ö* into the same class and selected the best representative for each class, based on the relative frequency of each variant. In this work, we mostly followed Fraser's approach, with some minor changes and additions. We processed all German monolingual data of the WMT 2010 campaign[1], amounting to $20M$ sentences and $359M$ words, and used *dass* and *muss* and their variants *daß* and *muß* to detect parts written with the old spelling. We mainly focused on three types of spelling rules[2];

3.1.1. Sounds and Letters

This part of the German spelling rules is aimed at removing inconsistencies between letter and sounds by regularizing the spelling of words that are formed from the same stem.

- *ß/ss*: According to the new spelling rules, *ß*, when preceded by a short vowel sound, should be written with a double *s*. By applying this rule, according the relative frequencies, German word *blaßen* is changed to *blassen*, but *maßlose* remains unchanged.
- **Umlauts;** *ue/ü*, *ae/ä* **and** *oe/ö*: It is common to write *ue, ae or oe* instead of ü, ä and ö on non-German keyboards as it may be hard to print Umlauts. To remove different variations of the same token, we selected the variant with the highest frequency. For instance, *Waesche* is replaced by *Wäche* and *gedrueckt* is replaced by *gedrückt*.
- **Triple Consonants**: triple consonants, when followed by a vowel, were reduced to two. In the new spelling, triple consonants are preserved no matter the following letter. Spelling variants differing only on the number of consonants are normalized accordingly. For instance, *Moselschiffahrt* is changed to *Moselschifffahrt* and *Schrotteile* is changed to *Schrottteile*.

3.1.2. Foreign Words

To adopt the words of foreign origin into the German language, some specific sounds are assimilated to their closest German sound. This is for instance the case for *ph* rewritten as *f*, *gh* as *g*, *rh* as *r* and *th* as *t*. With this new spelling rules, the word *sympthome* is spelled as *symptome* and *stephano* is written as *stefano*.

3.1.3. Use of Hyphens with Numbers

The spelling of compounds comprising numbers and regular words was also inconsistent in the old German spelling. For some combinations, a hyphen was used (*6-Kilogramm-Packung*) but omitted in some other cases (*12mal*). This am-

[1] http://www.statmt.org/wmt10/

[2] http://people.exeter.ac.uk/pjoyce/rechtschreibreform/indrules.html

biguity is solved by the new reform by assigning a hyphen to each number-word combination. We implemented the new spelling rule by inserting a hyphen between all number-word concatenations where a word is a sequence of three or more characters. For example, *400tonner* is rewritten as *400-Tonner*. We ignored the number-suffix combinations like *stel, sten, ern, etc.* as in for instance *100stel*.

3.2. Tokenization

Tokenization is an important, language specific process in machine translation. It is well-known that better tokenization often results in higher translation quality [15]. The tokenizer[3] used in our experiments does not chop off some German specific tokens that are redundant when translating into English. We therefore adapted the German tokenizer, focusing mainly on deleting unnecessary hyphens;

- **With numbers:** Number word combinations in English and German are different as mentioned above. Unlike English, words in German are compounded with numbers using a hyphen sign, as in *20-Tonner, 65-mal*. Accordingly, we replaced the hyphens in such combinations by a white space to separate the number and word. We also removed the hyphens in the number-suffix combinations, but this time with the effect of concatenating these tokens.

- **With hyphens** Some German compounds (notably those involving coordination) comprise an initial or trailing hyphen as in *Getrennt- und Zusammenschreibung*. These tokens are generally singletons and worsen the data sparseness problem. We therefore deleted all such hyphens that are not attached to any following or preceding token.

3.3. Removing the lexical redundancy

German has two numbers (singular/plural) and three morphological genders (masculine/feminine/neuter). German nouns, adjectives, determiners and pronouns are therefore inflected according to these categories, where the inflection is typically marked by a suffix change. Additionally, German verbs distribute case markers to their various dependents, using a system of four different cases (nominative/accusative/dative/genitive). Agreement takes place within the noun phrase, where all dependents of the noun should agree in gender, number and case with the head noun; agreement (in number and person) also takes place between the verb and its subject.

In this system, the inflection of adjectives is slightly more complicated than for the other parts-of-speech, as the inflection marks depends on the preceding token, where three different configurations yield different inflections: no preceding article, existence of definite or indefinite article.

As a result, German definite determiner could be marked in sixteen different ways according to the possible combinations of genders (3), case (4) and number (2)[4], which are fused in six different tokens *der, das, die, den, dem, des*. Except for the plural and genitive cases, all these forms are translated to the same English word *the*.

Having different word forms for a source side lemma that are systematically translated to the same target token is an instance of lexical redundancy in translation. This redundancy results in unnecessary large phrase translation tables that overload the decoder, as a separate phrase translation entry has to be kept for each word form. Our attempt to remove the lexical redundancy are similar to that of Corston-Oliver and Gamon [2]. These authors proposed to normalize all inflectional morphology by lemmatizing tokens on both the German and English side. With a very limited training data, they showed that reducing inflectional morphology decreases alignment error rate but did not report any experimental results of the effects on translation quality. This approach however causes the loss of critical information such as case and number. In our experiments, we investigated the effect of normalization on *translation quality* by various normalization strategies for the different word classes so as to reduce the German vocabulary size and to improve the robustness of the alignment probabilities while preserving all the necessary information. We used manually written patterns to remove the redundant information. A pattern typically defines those forms of a given morphological paradigm that should be considered equivalent when translating into English. These normalization patterns use the lemma information, as computed by the TreeTagger [16], together with the fine-grained POS information computed by the RFTagger [17], which uses a tagset containing approximately 800 tags. Table 1 displays the analysis of an example sentence.

The rules we used take, for instance, the following form:

- **For articles, adjectives (only positive form) and pronouns (indefinite , possessive, demonstrative and relative pronouns);**
  - If a token has genitive case: replace with lemma+en (Ex. *des*, *der*, *des*, *der* $\rightarrow$ *d+en*)
  - If a token has plural number: replace with lemma+s (Ex. *die*, *den* $\rightarrow$ d+s)
  - All other gender, case and number: replace with lemma (Ex. *der*, *die*, *das*, *die* $\rightarrow$ d)

- **For nouns;**
  - Plural number: replace with lemma+s (Ex. *Bilder*, *Bildern*, *Bilder* $\rightarrow$ *Bild+s*))
  - All other gender and case: replace with lemma (Ex *Bild*, *Bilde*, *Bildes* $\rightarrow$ *Bild*;

[3] We used the tokenizer distributed by WMT 2010 organizers.

[4] For the plural forms, gender distinctions are neutralized and the same 4 forms are used for all genders .

Input	TT-POS	Lemma	RFT-POS
General	NN	General	N.Name.*2.*3.*4
Musharraf	NE	Musharraf	N.Name.Nom.Sg.*4
betrat*	VVFIN	betreten	VFIN.Full.3.Sg.Past.Ind
am	APPRART	am	APPRART.Dat.Sg.Masc
12.*	ADJA	12.	ADJA.Pos.Dat.Sg.Masc
Oktober*	NN	Oktober	N.Reg.Dat.Sg.Masc
1999	CARD	1999	CARD
die*	ART	d	ART.Def.Acc.Sg.Fem
nationale*	ADJA	national	ADJA.Pos.Acc.Sg.Fem
Bühne*	NN	Bühne	N.Reg.Acc.Sg.Fem
,	$,	,	SYM.Pun.Comma
als	KOKOM	als	CONJ.SubFin.-2
er	PPER	er	PRO.Pers.Subst.3.Nom.Sg.Masc
eine*	ART	ein	ART.Indef.Acc.Sg.Fem
gewählte*	ADJA	gewählt	ADJA.Pos.Acc.Sg.Fem
Regierung*	NN	Regierung	N.Reg.Acc.Sg.Fem
stürzte*	VVFIN	stürzen	VFIN.Full.3.Sg.Past.Ind
und	KON	und	CONJ.Coord.-2
ein*	ART	ein	ART.Indef.Acc.Sg.Neut
ehrgeiziges*	ADJA	ehrgeizig	ADJA.Pos.Acc.Sg.Neut
Nationbuilding-Projekt	NN	Nationbuilding-Project	N.Reg.Acc.Sg.Neut
ankündigte*	VVFIN	ankündigen	VFIN.Full.3.Sg.Past.Ind
.	$.	.	SYM.Pun.Sent

Table 1: *TreeTagger (TT) and RFTagger (RFT) outputs*

- **For main verbs (except auxiliary and modal verbs);**
  - The verbs with present tense and 3rd person singular is not changed.
  - All other verbs: lemma+past or lemma+pres, depending on the tense (Ex. stürzte → stürzen+past)

After complete normalization, the sentence given above becomes:

- **German Sentence:** *General Musharraf betrat am 12. Oktober 1999 die nationale Bühne, als er eine gewälte Regierung stürzte und ein ehrgeiziges Nationbuilding-Projekt ankündigte*
- **Normalization:** *General Musharraf betreten+past am 12. Oktober 1999 d national Bühne, als er ein gewält Regierung stürzen+past und ein ehrgeizig Nationbuilding-Projekt ankündigen+past*[5]
- **English Reference:** *General Musharraf appeared on the national scene on October 12, 1999, when he ousted an elected government and announced an ambitious "nation-building " project.*

Many experiments were carried out with different normalization schemes, involving the differential normalization of specific part-of-speech combinations (see Section 4).

[5]Selected tokens are marked with a star in Table 1.

3.4. Compound Splitting

As German language uses compounding extensively, compound words are one of the most challenging issues in German-English SMT. Combining nouns, verbs and adjectives to coin new words is a very common process. German compound words typically tend to align with more than one English word. But even when the compound parts are frequent enough, most of the compounds are rare. As words are freely conjoined, the vocabulary size increases vastly, yielding to sparse data problems that turn into unreliable word alignments, phrase extraction and poor parameter estimates.

Berton et *al.* [18] stated three facts with the Verbmobil project evaluation 95 corpus; i) one third of the vocabulary is compound words; ii) almost half of the OOV words in the test set are compound words; and iii) 90% of these compound words are composed of at least one "known" word.

Compound splitting has been addressed by some researchers. Niessen and Ney [1] used a morpho-syntactic analyzer to split the compounds. Koehn and Knight [19] introduced frequency based algorithm which compares the frequency of compound word and geometric mean of frequencies of different splitting options. Popovic et *al.* [20] compared linguistic and corpus-based compound splitting, and investigated the word alignments that are improved with splitting point information of compounds. Stymne [21] compared different corpus-based compound splitting combinations by changing the word length, scoring algorithm, number and POS of compound parts.

We focused on different parameters of compound split-

ting procedure than the previous works and carried out many experiments with different configuration settings. The focus of our experiments was the impact of length of candidate compound words and split parts, and different types of filler suffixes on the translation quality (see Section 4).

4. Experiments

4.1. Setup

We used all available data (*europarl-v5, news-commentary10*) distributed for WMT 2010 evaluation campaign. English texts were processed and normalized using in-house text processing tools in the tokenization and detokenization steps. German corpus is tokenized by our modified version of WMT 2010 tokenizer described in Section 3.2. All systems are built in "true-case". Table 2 displays some statistics regarding the corpus used in our experiments.

We generated word alignments using GIZA++ [22] with default settings. Systems were tuned using the Moses implementation of minimum error rate training (MERT) [23], using the news-test2008 as the development corpus. We used newstest2010 as the test corpus and obtained the translations with the Moses [24] decoder. All experiments use a similar setting and only differ on the preprocessing step on the German side. We used conventional 4-gram language models on the English side; these models were trained, using modified Kneser-Ney [25] smoothing, on the Gigaword English corpus. All results are reported in terms of BLEU [26] and NIST [27] scores.

4.2. Baseline Experiments

In the first three rows of Table 7, we present a first set of results that allows to evaluate the benefits of adapting German words to the spelling reform and of improving the tokenization process. We used the third configuration (system with new tokenization) as the baseline for all of the further experiments.

4.3. Normalization Experiments

Table 3[6] displays the normalization results for different normalization schemes. For each main category, a '+' in the corresponding column indicates that the category was subject to normalization.

Our experiments showed that normalization of adjectives, nouns and pronouns is important, and our best results are obtained when all three categories are normalized. During these experiments, we also tried to normalize the genitive case, resulting in lower scores for all conditions. These results indicate that genitive information is important for translation into English and should be kept for all the classes except the nouns. This is an expected result as genitive forms, when they express a possession, are generally translated with an extra preposition (typically *of*) on the English side. As a last note, normalizing determiners and verbs does not seem to improve the translation quality.

[6]As the verb normalization always decreases the system performance, we did not run any further experiments with it.

Exp.	*ART*	*ADJ*	*PRO*	*NOUN*	*VERB*	*NIST*	*BLEU*
1	+					6.41	20.76
2		+				6.38	20.75
3			+			6.41	20.65
4				+		6.39	20.64
5					+	6.39	20.54
6		+		+		6.38	20.75
7	+			+		6.39	20.54
8	+	+		+		6.35	20.49
9		+	+	+		**6.42**	**20.85**
10		+		+	+	6.33	20.32
11		+	+	+	+	6.39	20.48
12	+	+	+	+		6.38	20.67
13	+	+	+	+	+	6.39	20.69

Table 3: *German-English normalization results*

4.4. Compound Splitting Experiments

Table 4 shows the results of various parameterization of the compound splitting algorithm. As mentioned in Section 4.2, the baseline is the system with the new tokenization. We used the corpus-based algorithm to handle compounding and experimented different splitting schemes similar to [21] by changing the candidate and subword lengths and the compound suffixes types. We used all suffixes and suffix types that are mentioned in this work. The complete list of suffixes are as follows :

- **Addition:** -s, -n, -en, -nen, -e, -es, -er, -ien
- **Truncation:** -e, -en, -n
- **Combination:** -us/-en, -um/-en, -um/-a, -a/-en, -on/-en, -on/-a -e/-i

As a minor change, we deleted the hyphens between compound word candidates and added this option in the split search space. For instance, for the compound word *Nationbuilding-Projekt*, the search space contains the following split options; *(Nationbuilding-Projekt), (Nationbuilding Projekt), (Nationbuilding -Projekt), (Nation building Projekt), (Nationbuilding -Projekt), (Nationbuilding- Projekt)*.

Experiments showed that compound word decomposition is crucial and helps vastly to improve translation results. We observed that choosing 4-8 and 5-10 as the minimum candidate-split lengths give better scores than the 3-6 configurations. The reason is very clear as many German prepositions, determiners and separable verb suffixes are of length 3 and the algorithm chooses the splits with higher frequencies, even when these tokens are not real split parts.

	Language	*Sentences*	*Total Words*	*Unique Words*	*Singletons*	*OOV(%)*
Train	English	$1.6M$	$44M$	$137K$	$57K$	–
	German		$42M$	$382K$	$191K$	–
Dev	English	2051	$49K$	$8K$	–	2.9
	German		$47K$	$10K$	–	5.0
Test	English	2489	$61K$	$9K$	–	3.0
	German		$61K$	$13K$	–	5.2

Table 2: *German-English corpora statistics*

Min. Split-Candidate Length	*Addition*	*Truncation*	*Combination*	*NIST*	*BLEU*
3-6	+			6.43	20.54
3-6	+	+		6.46	20.69
3-6	+	+	+	6.46	20.70
4-8	+			6.52	21.09
4-8	+	+		6.50	20.93
4-8	+	+	+	6.52	21.06
5-10	+			**6.53**	**21.09**
5-10	+	+		6.51	21.08
5-10	+	+	+	6.49	20.76

Table 4: *German-English compound splitting results*

4.5. Combined System

Both normalization and compound splitting help to increase the translation quality. To see the effect of the combination of these two methods, we split the compounds of the best normalization configuration which is adjective, noun and pronoun normalization. We used 4-8 as candidate-split minimum character lengths with only addition suffixes. Tables 5 and 6 show the number of preprocessed tokens after the normalization and the statistics of compound splitting on the normalized data. In table 5, the third column shows both the number of lemmatized words and pseudo words (lemma + pseudo tags). As seen in table 6, only a small portion of candidate compound words (about 20%) are split. We observed that frequency-based compound splitting generally tends to select splits with a small number (either 1, 2 or 3) of parts.

	POS	*Total*	*Normalized*
Train	Adjective	2591461	2321191
	Noun	9417709	7776519
	Pronoun	4079385	3522816
Dev	Adjective	2562	2213
	Noun	12013	8585
	Pronoun	3394	415
Test	Adjective	3127	2748
	Noun	15635	11124
	Pronoun	4645	640

Table 5: *Normalization statistics*

	1	*2*	*3*	*4*	*5*	*6*
Train	9122206	1853046	158753	5913	168	3
Dev	8742	2589	230	14	2	–
Test	11098	3355	305	18	–	–

Table 6: *Number of parts after compound splitting*

4.6. Singleton and OOV Normalization

On top of the previous experiments, we also investigated the effect of singletons and OOV words on the translation quality. As shown in Table 2, half of the words in the German vocabulary are singletons. Moreover, during the normalization and compound splitting processes, we introduced some new words with pseudo tags and by marking compound parts. Singletons are problematic and harm the word alignment as they are seen in the training data just for once. To reduce the number of singletons we performed two actions : for every *new* singletons introduced by normalization and compound splitting, we removed the marking and pseudo tags, and for all other singletons we replaced the word by the corresponding lemma. Similarly, we also performed a lemmatization for all OOV words in the test data as OOV words cannot be translated because they do not occur in the training data. Table 7 shows all preprocessing steps after normalization.

The main objective of this work was to decrease the structural differences between German and English as much as possible. We showed that each preprocessing operation increase the translation quality but it is revealing to also investigate their net effects on the German corpus. Table 8 presents the statistics of the various German corpus we de-

System	*NIST*	*BLEU*
Baseline	6.24	20.03
Spelling Reform	6.35	20.45
+New Tokenization	6.39	20.55
+Normalization	6.42	20.85
+Compound Splitting	6.43	21.27
+Singleton Normalization	6.43	21.35
+OOV Normalization	6.43	21.46

Table 7: *Results on WMT 2010 newstest2010 test data*

rived. The first three columns present the changes in training data and the last one reports the OOV ratio in the test data. We observed that the improved preprocessing results a significant decrease in the German vocabulary, singletons and OOV ratio. Although the German vocabulary is still larger than the English one, both vocabularies tend to get closer in size. The impact is much less on the phrase table size. The reason behind this is the new tokens that are introduced by pseudo-words and compound split marking. For each new token, the phrase extraction process generated an extra phrase entry even a very similar entry (only with minor differences) exists.

5. Conclusions

This paper presented the results of improving preprocessing for German to English phrase-based statistical machine translation systems. German is a highly inflected language, where inflections marks vary depending on number, gender and case, while English has very poor inflection processes. Moreover, productive compounding processes in German yield a high number of one-to-many word alignments, which impact the phrase extraction and so translation quality. The main findings of our work can be summarized as follows: i) we have converted old-written parts of the corpora to new writing style according to the German spelling reform; ii) we have removed redundant German tokens by introducing a new language-pair specific tokenization; iii) we have considered various normalization schemes on different POS groups by taking German morphological features into account; iv) we have explored the effect of different word length and filler configurations on compound splitting and; v) we have experimented lemmatization of singletons and OOV words. We have showed that employing a better preprocessing of German provides a promising increase in translation quality. As a result of all these processes, we have decreased the German vocabulary by approximately 50%, the number of singletons by 10%, the OOV rate by 2%. We have also reported about 1.5 BLEU improvement on the translation quality.

Translation into German is a much more challenging task as it includes both restructuring normalized tokens and compound merging. Some research is reported for compound merging but there is not much effort to regenerate correct word forms for our best knowledge.

Acknowledgments This work has been partly financed by OSEO, the French State Agency for Innovation, under the Quaero program.

6. References

[1] S. Niessen and H. Ney, "Statistical machine translation with scarce resources using morpho-syntatic information," *Computational Linguistics*, vol. 30, no. 2, pp. 181–204, 2004.

[2] S. Corston-oliver and M. Gamon, "Normalizing german and english inflectional morphology to improve statistical word alignment," in *Proceedings of the Conference of the Association for Machine Translation in the Americas*, DC, USA, 2004, pp. 48–57.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit X*, Phuket, Thailand, 2005, pp. 79–86.

[4] M. Yang and K. Kirchhoff, "Phrase-based backoff models for machine translation of highly inflected languages," in *Proceedings of 11th Conference of the European Association of Computational Linguistics*, Trento, Italy, 2006, pp. 41–48.

[5] Y.-S. Lee, "Morphological analysis for statistical machine translation," in *Proceedings of Human Language Technology conference / North American chapter of the ACL annual meeting - Companion Volume*, Boston, USA, 2004, pp. 57–60.

[6] F. Sadat and N. Habash, "Combination of arabic preprocessing schemes for statistical machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, Sydney, Australia, 2006, pp. 1–8.

[7] A. Zollmann, A. Venugopal, and S. Vogel, "Bridging the inflection morphology gap for Arabic statistical machine translation," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City, USA, 2006, pp. 201–204.

[8] M. Popovic and H. Ney, "Towards the use of word stems and suffixes for statistical machine translatio," in *4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004, pp. 1585–1588.

[9] S. Goldwater and D. McClosky, "Improving statistical MT through morphological analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005, pp. 676–683.

System	*Unique Words*	*Singletons(%)*	*Phrase Table Size*	*OOV(%)*
Spelling Reform	382292	50.4	9062239	5.2
+ New Tokenization	374725	49.9	9049745	5.1
+ Normalization	368492	51.9	8831898	5.0
+ Compound Splitting	220390	45.4	8944460	3.4
+ Singleton Normalization	205449	40.2	8916420	3.6
+ OOV Normalization	–	–	–	3.2

Table 8: *German Statistics for various experiments*

[10] D. Talbot and M. Osborne, "Modelling lexical redundancy for machine translation," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Sydney, Australia, 2006, pp. 969–976.

[11] M. Carpuat, "Toward using morphology in french-english phrase-based smt," in *Proceedings of the fourth ACL Wokshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 150–154.

[12] E. Minkov, K. Toutanova, and H. Suzuki, "Generating complex morphology for machine translation," in *Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic, June 2007, pp. 128–135.

[13] İlknur Durgar El-Kahlout and K. Oflazer, "Exploiting morphology and local word reordering in english to turkish phrase-based statistical machine translation," *To be Appear, Journal of IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[14] A. Fraser, "Experiments in morphosyntactic processing for translating to and from German," in *Proceedings of the Fourth ACL Workshop on Statistical Machine Translation*, Athens, Greece, 2009, pp. 115–119.

[15] D. Déchelotte, H. Schwenk, G. Adda, and J.-L. Gauvain, "Improved machine translation of speech-to-text outputs," in *Proceedings of INTERSPEECH'07*, Antwerp, Belgium, 2007.

[16] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994, pp. 44–49.

[17] H. Schmid and F. Laws, "Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK, 2008, pp. 777–784.

[18] A. Berton, P. Fetter, and P. Regel-brietzmann, "Compound words in large-vocabulary german speech recognition systems," in *Proceedings of the Ninth International Conference on Spoken Language Processing - ICSLP*, PA, USA, 1996, pp. 1165–1168.

[19] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proc. of the Conference of the European Chapter of the ACL*, Budapest, Hungary, 2003, pp. 187–193.

[20] M. Popovi, D. Stein, and H. Ney, "Statistical machine translation of german compound words," in *5th International Conference on Natural Language Processing - FINTAL*, Turku, Finland, 2006, pp. 616–624.

[21] S. Stymne, "German compounds in factored statistical machine translation," in *Proceedings of the 6th international conference on Advances in Natural Language Processing*, Berlin, Heidelberg, 2008, pp. 464–475.

[22] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the annual Meeting of the ACL*, Hongkong, China, 2000, pp. 440–447.

[23] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the ACL*, Sapporo, Japan, 2003, pp. 160–167.

[24] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *ACL, demonstration session*, Prague, Czech Republic, 2007.

[25] S. F. Chen and J. T. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting of the ACL*, Santa Cruz, NM, 1996, pp. 310–318.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation." in *Proceedings of the Annual Meeting of the ACL*, Philadelphia, PA, 2002, pp. 311–318.

[27] NIST, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," 2002. [Online]. Available: http://www.nist.gov/speech/tests/mt/doc/ngramstudy.pdf