Definite Noun Phrases in Statistical Machine Translation into Scandinavian Languages

Sara Stymne
Linköping University, Linköping, Sweden
Xerox Research Centre Europe, Grenoble, France
sara.stymne@liu.se

Abstract

The Scandinavian languages have an unusual structure of definite noun phrases (NPs), with a noun suffix as one possibility of expressing definiteness, which is problematic for statistical machine translation from languages with different NP structures. We show that translation can be improved by simple source side transformations of definite NPs, for translation from English and Italian, into Danish, Swedish, and Norwegian, with small adjustments of the preprocessing strategy, depending on the language pair. We also explored target side transformations, with mixed results.

1 Introduction

One problem for statistical machine translation is when the source language has a different structure in some respect than the target language. One such issue is the unusual realization of definite noun phrases in Scandinavian languages. Definiteness can be expressed in two ways in Scandinavian languages, either by a definite article or by a suffix on the head noun. This is problematic for translation from languages that only use definite articles, such as English or Italian, leading to problems such as wrong noun forms and spurious definite articles in the translation output.

It has previously been shown that definite noun phrases can successfully be handled by a preprocessing step for translation between English and Danish (Stymne, 2009b). In this study, source language noun phrases were transformed to be similar in structure to target language NPs. In this paper we show that preprocessing of definiteness also can be successful for translation from English into Swedish and Norwegian, and from Italian to Danish, using the same basic strategy as in Stymne (2009b). However, some small careful modifications to the original English-Danish preprocessing strategy were necessary.

2 Definiteness

In the Scandinavian languages there are two mechanisms for expressing definiteness, by using a definite article or by using a suffix on the head noun. These mechanisms can also be used in combination, so called double definiteness. The distribution rules for these two mechanisms are quite strict in all Scandinavian languages, but they differ somewhat between them. The noun phrase realization is different in NPs with a pre-modifier such as an adjective or numeral and in NPs without pre-modifiers. Table 1 shows the allowed and disallowed combinations in Swedish, Norwegian Bokmål[1] and Danish, which are the target languages we focus on in this paper, and compares it to English and Italian, the source languages we focus on. There are similar phenomena in the other Scandinavian languages – Norwegian Nynorsk, Icelandic, and Faroese – as well.

As can be seen in Table 1 there is a difference in pre-modified noun phrases, where the definite article is used, and in simple noun-phrases where the suffix is used. In Swedish and Norwegian there is double definiteness in pre-modified noun phrases, something that never occurs in Danish, where only the definite article is used in pre-modified noun phrases. The definite article, *den/det/de* (inflected for gender and number), coincides with

[1] There are two written varieties of Norwegian, Bokmål and Nynorsk. We will use the term Norwegian to refer to Norwegian Bokmål.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 289–296
Leuven, Belgium, May 2011

NP type	**Swedish**	**Norwegian**	**Danish**	**English**	**Italian**
sg, -mod	hund**en**	hund**en**	hund**en**	**the** dog	**il** cane
	***den** hund(**en**)	***den** hund(**en**)	***den** hund(**en**)		
sg, +mod	**den** svarta hund**en**	**den** svarte hund**en**	**den** sorte hund	**the** black dog	**il** cane nero
	***den** svarta hund	***den** svarte hund	***den** sorte hunden		
	*svarta hund**en**	*svarte hund**en**	*sorte hund**en**		
pl, -mod	hundar**na**	hunde**ne**	hunde**ne**	**the** dogs	**i** cani
pl, +mod	**de** svarta hundar**na**	**de** svarte hunde**ne**	**de** sorte hunde	**the** black dogs	**i** cani neri

Table 1: Definite noun phrases in Swedish, Norwegian, and Danish, contrasted to English and Italian. NP type shows number (singular or plural), and if the NP is modified by an adjective or not. The grammaticality judgments are for a definite reading of the definite articles *den/de*. In some cases these examples are acceptable with a demonstrative reading of *den/de*, see Table 2.

the demonstrative article, so some of the ungrammatical examples in Table 1 are grammatical in a demonstrative reading, see Table 2.

In demonstrative NPs, shown in Table 2, there is always double definiteness in Norwegian, whereas only the demonstrative article is used in Danish. In Swedish, the use of the definite suffix depends on the choice of demonstrative article, with *den (här)* the suffix is used, but with *denna* no suffix is used. In possessive noun phrases, the indefinite noun form is always used, except in a Norwegian option with a final possessive pronoun, where the definite suffix is used.

NPs that are post-modified by a relative clause constitute an additional complication. In all three languages both types of definiteness marker are allowed in NPs with relative clauses, exemplified in Swedish in (1–2). The definite article tends to be used with restrictive relative clauses, and the suffix for non-restrictive relative clauses. But, just as in English where the choice of relative pronoun and commas can be used for this purpose, the distinction between the two cases is fuzzy, and there are many exceptions to the general tendency. Thus, we will not be further concerned with relative clause exceptions in this paper.

(1) Den hund som skällde är snäll
The dog that barked is nice

(2) Hunden som skällde är snäll
The dog, which barked, is nice

There are also other special cases with irregular behavior, such as name-like uses like *Vita huset* (*the White House*) where the definite article is not used, or in connection with what Dahl (2003) call selectors, inherently definite words, like *först* (*first*) or *höger* (*right*), where the realization varies. These cases will also be ignored.

In summary Danish is most regular with respect to the definite suffix, which is only used in NPs without pre-modifiers. In Swedish and Norwegian the definite suffix can be used in other constructions than pure definite NPs, such as demonstrative or possessive NPs.

In English and Italian the same base noun form is always used, see Tables 1 and 2, both with definite and demonstrative articles. In possessive noun phrases Italian uses both a definite article and a possessive adjective, contrary to the other languages that mostly use just a possessive pronoun. Italian adjectives can be pre- and post-modifiers to nouns, as in (3), whereas all other languages only have pre-modifying adjectives.

(3) il grande cane nero
the big dog black

3 Previous Work

Our work fits into a growing mass of work where either the source or target language is preprocessed before training a SMT system, in order to make the languages more similar. If the target language is modified, a postprocessing step is necessary. Such modifications have been targeted at many different phenomena, such as compound words and word order.

The current study is based on Stymne (2009b), who address the issue of definiteness in translation from English to Danish, by transforming English NPs to a structure similar to that of Danish NPs. Rule-based transformations based on part-of-speech were used. The results, using only one simple transformation, were very good with relative Bleu improvements of 7.7% and 22.1% on two different domains. Definiteness was also targeted by Samuelsson (2006), who transformed German text, based only on surface forms, for translation into Swedish. There were no improvements on translation from German to Swedish using this

NP type	**Swedish**	**Norwegian**	**Danish**	**English**	**Italian**
dem, -mod	**den (här)** hund**en** **denna** hund	**den** hund**en** **denne** hund**en**	**den** hund **denne** hund	**this** dog	**questo** cane
dem, +mod	**den (här)** svarta hund**en**	**den** svarte hund**en**	**den** sorte hund	**this** black dog	**questo** cane nero
poss, -mod	min hund	min hund hund**en** min	min hund	my dog	**il** mio cane
poss, +mod	min svarta hund	min svarte hund **den** svarte hund**en** min	min sorte hund	my black dog	**il** mio cane nero

Table 2: Demonstrative and possessive noun phrases in Swedish, Norwegian, and Danish, contrasted to English and Italian. NP type shows if the NP is dem(onstrative) or poss(essive), and if the NP is modified by an adjective or not.

approach, but for translation in the other direction, which included postprocessing of the modified German NPs, there was a relative Bleu improvement of 11.0%.

Pre- and postprocessing have also been used for compound words, both for translating from Germanic languages such as German (Nießen and Ney, 2000) and Swedish (Stymne and Holmqvist, 2008), and for translation into a Germanic language, which requires post-processing where split compounds are merged (Stymne, 2009a). Nießen and Ney (2000) explored several types of preprocessing for translation from German to English besides compound splitting, including merging of multi-word expressions, and separation of German verb prefixes, with good results. Preprocessing has also been used extensively for targeting word order differences between languages, either by using hand written rules targeting known differences between two languages (Collins et al., 2005), and automatically learnt rules (Xia and McCord, 2004) to reorder the source language.

Another type of preprocessing is morphological reduction, i.e. to remove some of the morphological information in one of the languages. Goldwater and McClosky (2005) used lemmatized Czech, with the addition of morphological tags both as separate words and as suffixes, for translation into English, and El-Kahlout and Yvon (2010) normalized German morphology by removing all distinctions that are not present in English, both with positive results. Fraser (2009) removed all German inflections for translation into German, and recreated it in a postprocessing step, however, with negative results. In these three studies, some information is removed before the translation process. It seems, however, that care has to be taken not to remove too much information.

All these approaches work on different levels of linguistic representations, and require different linguistic tools. The lowest possible level of representation is surface form, which does not require any linguistic processing, and is used in Samuelsson (2006). Methods based on part-of-speech (Stymne, 2009b), chunks (Zhang et al., 2007), or parse trees (Collins et al., 2005) are more commonly used. Some approaches also use morphological analyzers (Goldwater and McClosky, 2005). While there is more information on the higher level of linguistic representations, tools tend to make more errors, the more complex they are. There is thus a trade-off between the expressivity and generality of the representation used, and its correctness using automatic tools.

4 Preprocessing Strategies

Our main strategy used to improve the translation with respect to definiteness is to transform definite NPs in the source language, to make them similar in structure to NPs in the target language. We also explore the opposite, to transform the target to make it more similar to the source. These strategies are based on the assumption that definite noun phrases in the source language always are translated with definite noun phrases in the target language, which is not always the case. Further, we only focus on strict definite NPs, we do not take into account demonstrative clauses or possessive clauses, whose realization can differ in Swedish and Norwegian, and which always have non-definite nouns in Danish.

For the source side processing we need to identify definite NPs in English and Italian. The target side processing was only implemented for Swedish, and for that we identify Swedish definite NPs without pre-modifiers. We use part-of-speech tags and lemmas to identify definite noun phrases, obtained by an in-house Hidden-Markov-based part-of-speech tagger for Italian and English, and the Granska tagger for Swedish (Carl-

Language pair	Non-modified NPs	Modified NPs
English-Danish	remove-DEF, add-DEFSUFFIX	none
Italian-Danish	remove-DEF, add-DEFSUFFIX	move-ADJ
English-Swedish/Norwegian 1	remove-DEF, add-DEFSUFFIX	add-DEFSUFFIX
English-Swedish/Norwegian 2	remove-DEF	none

Table 3: Operators used to transform the source language for the different language pairs. Modified means pre-modified (or post-modified for Italian), by at least one adjective or numeral.

Orig En:	the central body is called ’ the european food authority ’ or ’ the authority ’ for short
for Da:	the central body is called ’ european food authority-DEF ’ or ’ authority-DEF ’ for short
for Sv/No 1:	the central body-DEF is called ’ the european food authority-DEF ’ or ’ authority-DEF ’ for short
for Sv/No 2:	the central body is called ’ the european food authority ’ or ’ authority ’ for short
Orig It:	nei fondi strutturali , notiamo problemi nell’ applicazione delle normative a tutti i livelli
for Da:	in il strutturali fondi , notiamo problemi in applicazione-DEF di normative-DEF a tutti livelli-DEF

Table 4: Example source side transformations

berger and Kann, 1999). Part-of-speech tags are used to identify nouns, adjectives and numerals, but definite articles are not distinguished from other articles in the tagsets used for Italian and English, so for them we use surface form in English, *the*, and lemma in Italian, where all definite articles are given the lemmas *lo* or *il*. In addition, prepositions and articles can be contracted in Italian, but this is also handled by the lemmas, where contractions are split and normalized, for instance *delle/dell’* to *de lo*. For Swedish, we have morphological tags, which identify nouns and articles as definite or indefinite.

The pattern used to identify English definite noun phrases is defined in (4), and consists of a definite article, possibly followed by an arbitary number of modifiers: adjectives or numerals, followed by at least one noun. The pattern used for Italian definite NPs is defined in (5), and it differs from English in that adjectives can be placed after the head noun, in addition to before. In practice though, allowing an arbitrary number of premodifiers are error prone, due to tagging errors, so we restrict transformations to noun phrases with a maximum of two pre-modifiers. For Swedish we are only interested in identifying definite NPs without pre-modifiers, and thus use the simplified pattern in (6), where we identify definite nouns which are not preceded by a pre-modifier or an article.

(4) `DEF-ART (ADJ|NUM)* NOUN+`

(5) `DEF-ART (ADJ|NUM)* NOUN+ ADJ*`

(6) `¬(ADJ|NUM|ART) NOUN-DEF`

We chose to use part-of-speech and lemmas since we believe that gives us enough information to extract the definite NPs we need. An alternative would have been to use a parser or chunker to identify noun phrases, but such tools generally have more errors than a POS-tagger. The patterns in (4–5) in practice constitute a chunker, though, but only for the definite NPs we need.

4.1 Source Side Processing

In order to perform the transformations we use two main operators, remove-DEFART and add-DEFSUFFIX, where the first one removes unnecessary definite articles in the source language, and the second adds a definite suffix to the head noun, which is often a single noun, but can be the last of many nouns for noun compounds. For Italian as a source language, we introduce a third operator, move-ADJ, which moves adjectives that are placed behind the noun, to before the noun. The choice of operators depends on if the identified noun phrase has adjectival and/or numeral modifiers or not. Swedish and Norwegian have the same structure of definite NPs, and can thus use the same strategies, whereas Danish has a different structure.

For English-to-Danish we follow the strategy described in Stymne (2009b). For noun phrases without pre-modifiers both remove-DEFART and add-DEFSUFFIX is used, since they are only marked with a definite suffix. For noun phrases with pre-modifiers, no operators are used, since they have the same structure as in English.

For Italian-to-Danish, the strategy is the same as for English-to-Danish, but we also take into account post-modifying adjectives, in addition to pre-modifiers, to distinguish the two classes of def-

Language pair	Decoder	Corpus	Sentences	Source words	Source, proc	Target words
English-Danish 1	Matrax	Automotive	168,046	1,526,759	1,479,186	1,395,661
English-Danish 2	Matrax+P+CS	Automotive	168,046	1,526,759	1,479.186	1,478,707
Italian-Danish	Matrax	Europarl	100,000	2,086,719	2,057,694	2,003,699
English-Norwegian	Matrax	Automotive	395,733	3,889,706	3,733,483	3,340,557
English-Swedish 1	Matrax	Automotive	327,596	3,454,887	3,295,481	2,870,623
English-Swedish 2	Moses+P	Europarl	701,157	15,043,321	14,385,253	13,603,062

Table 5: Experiment setup: languages, decoder, corpus and corpus statistics. +P on the decoder means that we used a sequence model based on part-of-speech, and +CS that compounds are processed.

inite NPs. The operator move-ADJ is used for definite noun phrases that contain an adjective that post-modifies the noun. In addition we also normalize the definite articles that are not removed, by replacing them with the lemmas *il* in singular and *lo* in plural.

For English-to-Swedish/Norwegian, we tried to mimic the strategy for English-to-Danish as closely as possible, while still taking into account the differences in realization between the languages. That means that for noun phrases without modifiers, the strategy is the same as for Danish, to use both remove-DEFART and add-DEFSUFFIX. For pre-modified phrases, though, we need to use add-DEFSUFFIX, since both types of definite markers are used there. We also decided to try a second strategy, where we do not use add-DEFSUFFIX, since the distribution of the definite suffixes are more complex in these languages in other types of phrases, and thus only used remove-DEFART in NPs without pre-modifiers. This transformation means that we lose information about definiteness in the source, and thus leave the choice of using a definite suffix or not mainly to the language model. The source side transformations for the different language pairs are summarized in Table 3, and exemplified in Table 4.

4.2 Target Side Processing

We also tried to preprocess the target side of the corpus for Swedish by adding articles that are present in English bare definite NPs, exemplified in (7). This transformation addresses the problem of the source side strategies for Swedish, where the first strategy creates markup only on definite nouns in pure definite NPs, and not in other contexts, such as in demonstrative NPs, and where the second strategy loses information present in English. The added articles will be present in the Swedish MT output, and we thus need a postprocessing step to remove them.

(7) grundvalen är det svåra beslutet
DEF grundvalen är det svåra beslutet
(the) basis-DEF is the hard decision-DEF

The added definite articles are separate tokens, *DEF*, which differ in surface form from the normal Swedish definite article, since we need to be able to identify them, in order to remove them in the postprocessing step. The tokens are added in NPs without pre-modifiers, where the head noun is in definite form. After translation, the *DEF* tokens are removed in the translation output.

5 Experiments

We used two standard phrase-based decoders, Matrax (Simard et al., 2005) and Moses (Koehn et al., 2007). Matrax allows noncontiguous bi-phrases, such as *jeopardize – bringe . . . i fare* (*bring . . . into danger*) for English-Danish, where words in the source, target, or both sides can be separated by gaps that have to be filled by other phrases at translation time. In the experiments we allowed up to four gaps per phrase pair. Moses, and most other phrase-based decoders can only use contiguous phrases. We use a 3-gram language model in Matrax, and a 5-gram model in Moses. In some experiments, we also used an additional sequence model based on part-of-speech. The sequence models were trained using the SRILM toolkit (Stolcke, 2002).

To train the decoder we used two different corpora, Europarl, proceedings of the European Parliament (Koehn, 2005), and an automotive corpus, collected from translation memory data. To reduce training times we did not use all data from Europarl. For the Italian-Danish experiment we randomly selected the sentences to use, and for English-Swedish we used version 2 of Europarl. The types and sizes of corpora used in the experiments are shown in Table 5. For English-Danish, the first experiment is repeated from Stymne (2009b), and in the second experiment a POS sequence model and compound processing as described in Stymne and Holmqvist (2008) were added. In all cases the number of words is higher in the source language than in the target language, but, as shown in the second last column of Table 5, the number of words is reduced and somewhat closer to the number of target words, after the source side definiteness processing. For test we used 1000 sentences, and for parameter optimization we used 1000 sentences for translation into Danish, 500 sentences with Moses, and 2000 sentences otherwise.

We trained systems with source side processing, which will be called DEF-proc for translation into Danish, and DEF-proc1, and DEF-proc2, for the two different strategies for translation into Swedish and Norwegian. For English–Swedish Europarl we also trained a system with target side processing, which is called Target-proc. We compare all results to baseline systems that do not use any transformations.

5.1 Results

Table 6 shows the results of the experiments, on the two standard metrics Bleu (Papineni et al., 2002) and NIST (Doddington, 2002) with one reference translation. Significance was tested using approximate randomization (Riezler and Maxwell, 2005), with $\alpha < 0.05$. Overall the results are much higher on the automotive corpus, than on Europarl, which is expected since that corpus is more homogenous, and has shorter sentences.

For English-Danish translation we see a large improvement of 5.44 Bleu points in the first experiment. In the second experiment, where we added a POS-sequence model and compound processing, the baseline is significantly better than the baseline of the first experiment. Again, definite processing gives an improvement, of 2.08 Bleu points, but it is smaller than in the first case, and the scores with definite processing are similar in the two experiments. This indicates a need to further explore the interactions of definite processing, and other types of preprocessing. For Italian-Danish translation, there is also a significant improvement, of 1.5 Bleu points.

For translation into Swedish and Norwegian, the first strategy, where nouns are marked with a suffix, led to significantly worse results than the baseline in both cases. The second strategy, which only uses remove-DEF, however, led to improvements in both cases, where the improvement for English-Swedish were statistically significant. These improvements were smaller than for Danish, however. For the second English–Swedish experiment, we also investigated target side preprocessing. This was not successful, with a significantly worse result on Bleu, and a somewhat worse NIST score, as the baseline.

Languages	System	Bleu	NIST
En-Da 1	Baseline	70.91	8.8816
	DEF-proc	76.35+	9.3629+
En-Da 2	Baseline	74.09	9.2328
	DEF-proc	76.17+	9.4342+
It-Da	Baseline	10.54	4.3924
	DEF-proc	12.04+	4.5754+
En-No	Baseline	58.57	8.8846
	DEF-proc1	56.59-	8.6943-
	DEF-proc2	59.08	8.9092
En-Sv 1	Baseline	61.20	9.7934
	DEF-proc1	58.84-	9.4898-
	DEF-proc2	62.05+	9.9129+
En-Sv 2	Baseline	21.63	6.1085
	DEF-proc2	22.03+	6.1778+
	Target-proc	21.31-	6.1018

Table 6: Translation results, a plus sign marks results that are significantly better than the baseline, and a minus sign marks significantly worse results.

We performed an initial error analysis of 50 short sample sentences from the second Swedish experiment, where the differences on the automatic metrics were quite small. The results of this analysis were somewhat different than what we expected based on the metric scores, with the lowest total number of errors for the target-proc system, which had 61 errors, compared to 71 for the abseline and 74 for the source side processing. The slightly higher number of errors in the system with source side processing were mainly due to wrong translations or insertions of function words, such as prepositions. Both systems with definiteness processing had a lower number of word order and punctuation errors than the baseline. The number of definiteness errors were approximately the same between the three systems, but they were all the wrong form of nouns in the system with source side processing, which is not surpsising since we removed the definite distinction in English bare NPs, whereas other types of definiteness errors also occured in the other two systems, such as spurious definite articles. This limited analysis did unfortunately not shed much light on the types of changes that were the result of adding definite processing, and further analysis is needed.

To illustrate the effects of the definiteness pro-

	Italian–Danish
Src:	Non pensa che dovremmo ormai esplorare nuovi modi per affrontare il problema delle nostre relazioni con la Birmania?
Ref:	Finder De ikke, at vi bör se på andre måder, hvorpå vi kan tackle problemet med vores relationer i Burma?
Baseline:	Tänker ikke at vi bör efterhånden resterende tid nye måder fat af vores forbindelser med den Burma?
DEF-proc:	Tänker ikke at vi bör overveje nye måder nu af vores forbindelser med Burma tackle problemet?
	English–Swedish
Src:	Men who commit murders rarely receive long prison sentences . . .
Ref:	Männen som utför morden får sällan långa fängelsestraff . . .
Baseline:	De män som begår morden sällan få långa fängelsestraff . . .
DEF-proc2:	Män som begår mord sällan få långa fängelsestraff . . .
Target-proc:	De män som begår morden sällan erhåller långa fängelsestraff . . .

Table 7: Sample translations

cessing, we will discuss two translation examples, shown in Table 7. In the Italian–Danish example, there is an unnecessary definite article in front of the proper name *Burma* in the baseline, which correctly is not there in the DEF-proc version. Overall the DEF-proc translation is a better translation than the baseline, mainly since it manages to translate the verbs *esplorare* (*explore*) and *affrontare* (*handle*), even though it is slightly problematic with a meaning shift of the first into *overveje* (*consider*) and a correct meaning, but wrong word order of the second, *tackle* (*handle*). Both these verb are, however, completely missing in the baseline translation. Both translations miss the main pronoun *De* (*you, polite*), which is not present in Italian, which is a pro-drop language. In the English–Swedish example, all three renderings of *Men who commit murders* are grammatically possible, but the baseline and Target-proc readings have lost the general reading of the source, and refers to specific *murders*. The rendering of the DEF-proc is actually more true to the generality of *murders* in the source than the reference, which might, however, have taken the context of surrounding sentences into account. In all MT sentences, there are problems with the placement of the adverb *sällan* (*rarely*), and the main verb *få* is non-finite in the baseline and DEF-proc systems, but has the correctly finite form *erhåller* in the Target-proc system, even though that is a worse lexical choice.

Overall, we see some improvements with regard to definiteness in the systems with source side preprocessing, as in the examples discussed above, but there are also problems still left. We also see many other changes though, such as different lexical choices and word orders. One possible explanation for this can be that the word alignment changes when the two languages are more similar, and of more equal sentence length, which was the result of both types of definiteness processing.

6 Conclusion

We have shown that source side preprocessing targeting definite NPs is useful for translation into three Scandinavian languages on two different corpora using two different phrase-based decoders, as measured by automatic metrics. The attempt at target side preprocessing was not successful measured by automatic metrics, but had good results on an error analysis. There is a need for further analysis of the results, to try and pinpoint the reasons for the improvements on the automatic metrics, and to further investigate the effects of the preprocessing.

Care has to be taken when adjusting the source side processing strategy to a new language pair. When we performed the same type of transformation for translation into Swedish and Norwegian, as those that worked for Danish, both in this and previous work, the results were worse than the baseline. For translation into these languages, a more limited transformation were more useful. We believe that some treatment of definiteness is useful for translation into all Scandinavian languages, and that similar strategies as those described in this paper could also be useful for other source languages, and/or for translation into the other Scandinavian languages.

We see a much smaller effect of definite processing for translation into Swedish and Norwegian than into Danish. The definite suffix is used in more types of clauses in Swedish and Norwegian, than in Danish, which could partly explain this. Thus, it might be useful to design a more elaborate preprocessing strategy for these languages, taking other types of phrases than only simple definite NPs into account, possibly by using a machine learning method to decide where to apply transformations. There are also other possibilities of target side preprocessing, such as splitting off the definite suffix.

References

Carlberger, Johan and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software Practice and Experience*, 29:815–832.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540, Ann Arbor, Michigan, USA.

Dahl, Östen. 2003. Definite articles in Scandinavian: Competing grammaticalization processes in standard and non-standard varieties. In Kortmann, Bernd, editor, *Dialect Grammar from a Cross-Linguistic Perspective*, pages 147–180. Mouton de Gruyter, Berlin.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram cooccurence statistics. In *Proceedings of the Second International Conference on Human Language Technology*, pages 228–231, San Diego, California, USA.

El-Kahlout, İlknur Durgar and François Yvon. 2010. The pay-offs of preprocessing for German-English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 251–258.

Fraser, Alexander. 2009. Experiments in morphosyntactic processing for translating to and from German. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 115–119, Athens, Greece.

Goldwater, Sharon and David McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of the Human Language Technology Conference and the conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada.

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

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Riezler, Stefan and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*, pages 57–64, Ann Arbor, Michigan, USA.

Samuelsson, Yvonne. 2006. Nouns in statistical machine translation. Unpublished manuscript: Term paper, Statistical Machine Translation.

Simard, Michel, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of the Human Language Technology Conference and the conference on Empirical Methods in Natural Language Processing*, pages 755–762, Vancouver, British Columbia, Canada.

Stolcke, Andreas. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA.

Stymne, Sara and Maria Holmqvist. 2008. Processing of Swedish compounds for phrase-based statistical machine translation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 180–189, Hamburg, Germany.

Stymne, Sara. 2009a. A comparison of merging strategies for translation of German compounds. In *Proceedings of the EACL 2009 Student Research Workshop*, pages 61–69, Athens, Greece.

Stymne, Sara. 2009b. Definite noun phrases in statistical machine translation into Danish. In *Proceedings of the Workshop on Extracting and Using Constructions in NLP*, pages 4–9, Odense, Denmark.

Xia, Fei and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, Geneva, Switzerland.

Zhang, Yuqi, Richard Zens, and Hermann Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 21–28, Trento, Italy.