Manual Analysis of Structurally Informed Reordering in German-English Machine Translation

Teresa Herrmann, Jan Niehues, Alex Waibel

Interactive Systems Laboratories
Karlsruhe Institute of Technology
{teresa.herrmann,jan.niehues,alexander.waibel}@kit.edu

Abstract

Word reordering is a difficult task for translation. Common automatic metrics such as BLEU have problems reflecting improvements in target language word order. However, it is a crucial aspect for humans when deciding on translation quality. This paper presents a detailed analysis of a structure-aware reordering approach applied in a German-to-English phrase-based machine translation system. We compare the translation outputs of two translation systems applying reordering rules based on parts-of-speech and syntax trees on a sentence-by-sentence basis. For each sentence-pair we examine the global translation performance and classify local changes in the translated sentences. This analysis is applied to three data sets representing different genres. While the improvement in BLEU differed substantially between the data sets, the manual evaluation showed that both global translation performance as well as individual types of improvements and degradations exhibit a similar behavior throughout the three data sets. We have observed that for 55-64% of the sentences with different translations, the translation produced using the tree-based reordering was considered to be the better translation. As intended by the investigated reordering model, most improvements are achieved by improving the position of the verb or being able to translate a verb that could not be translated before.

Keywords: Statistical Machine Translation, Word Reordering, Manual Evaluation

1. Introduction

Different word orders between languages are one of the main problems in statistical machine translation. Especially between German and English the position of the verb leads to problems during translation. While in English, the verb is typically at the second position in the sentence, the rules for the position of the verb in a German sentence differ depending on the sentence type. In a main clause, the German verb is at the second position of that clause, as in English. In a German subordinate clause, the verb is at the final position. In case the German verb consists of several parts, e.g. auxiliary and participle, the rules get more complicated: in a main clause the finite auxiliary is in the second position and the participle is at the final position with possible objects or prepositional phrases in between. In a subordinate clause, both the auxiliary and participle are located at the final position, however, with switched order.

Adhering to those rules during translation is not straight forward in a surface-based statistical translation system, since the system has no notion about the type of clause which is being translated. A typical phrase-based system treats one phrase[1] at a time without looking at context beyond phrase boundaries. In addition, wrong word positions are penalized more than missing words by the automatic metric commonly used to measure translation quality. Hence, the system tends to lose the German verbs at the final position producing an English sentence lacking the verb or important parts of the verb.

In order to address the reordering problem, many approaches have been proposed. A very popular one is to detach reordering from the translation process and perform reordering of the source sentence before translation. Such pre-reordering approaches differ in the linguistic knowledge used in the model. One may use word-level information such as the actual words or their parts-of-speech (POS), as well as information about the sentence structure, such as syntactic parse trees or dependency trees. Typically the reordering model consists of rules determining the procedure of rearranging the source language words according to the target language word order. Then monotone translation can be performed. Using this method, improvements of the translation quality could be shown in the automatic metrics measuring the quality of machine translation systems.

Even though evaluation campaigns recently put more emphasis on human judgements of machine translation quality, such manual evaluations have been reported only on a small scale due to their expensiveness.

We perform an exhaustive analysis of the impact of a structure-aware reordering approach in statistical machine translation. With this analysis we want to investigate whether using a reordering approach relying on structure also results in better structure in the translation output.

We examine translations of three different types of data: news texts, TED talks and university lectures. Since these data types exhibit different text characteristics, we also expect the effects of the reordering model to differ. For each of the data types we generated translations by a system using part-of-speech based reordering rules only, and a system using part-of-speech-based and syntactic tree-based reordering rules together. Then we perform a manual analysis of the two translations on a sentence-per-sentence basis. We analyze the impact of the additional tree-based reordering rules on the different data sets, i.e. how many sentences are affected. In addition, we address the translation quality by a pairwise comparison of the two translations generated by the two systems. Third, we investigate the improvements and degradations introduced by the tree-based model

[1] A phrase here consists of a sequence of words as opposed to phrases in the linguistic sense.

Rule Type	Example Rule	
POS	*VVIMP VMFIN PPER*	→ 2 1 0
	*VAFIN * VVPP*	→ 0 2 1
Tree	*VP PTNEG NP VVPP*	→ 0 2 1

Figure 1: *Rule Types*

and classify them according to word categories and their role in the sentence. As a fourth aspect, we examine the correlation between changes introduced by the tree-based reordering model and the overall translation performance.

2. Reordering Models

We use two approaches based on sequences of part-of-speech (POS) sequences of the words in the sentence (Rottmann and Vogel, 2007; Niehues and Kolss, 2009), which operate on the word-level only. There are two types of POS-based reordering rules: Continuous rules consist of a sequence of POS tags on the left hand side and an indexed representation of their target order on the right hand side of the rule. A discontinuous rule consist of a sequence of POS tags with placeholders on the left hand side. The right hand side of the rule contains the reordered indices where the tags matched by the placeholder are assigned one index.

In addition we perform reordering based on constituents of syntactic parse trees (Herrmann et al., 2013). In contrast to the POS-based rules, the tree-based rules are able to take sentence structure into account. The tree-based rules address reordering within one constituent of a syntactic tree. The rule consists of the head category and the the child categories of the constituent on the left hand side of the rule. The right hand side represents the reordered sequence of the children where each child constituent is assigned one index. Examples for the rule types are presented in Figure 1.

2.1. Training and Application

For the training of the reordering rules a parallel corpus and a word alignment is required. In addition, we need the POS tags and parse trees for the source side of the corpus. Using this information we learn rules that rearrange the source words in the order of the aligned target words. Before translation a word lattice is created that includes the original source sentence as the monotone translation path. Then the reordering rules are applied to the source text. Finally, the word lattice including all reordering variants is used as input to the decoder. The decoder then searches for the best reordering variant while producing the most probable translation. For more details refer to the descriptions of POS-based rules and tree-based rules in the respective papers.

3. Related Work

Since the development cycles for machine translation systems are becoming shorter, automatic metrics are a popular method for measuring the quality of machine translation systems or their included models quickly and in a reproducible fashion (Papineni et al., 2002; Snover et al., 2006; Lavie and Denkowski, 2009). Since typical metrics for translation quality do not correlate well with reordering quality, explicitly measuring the reordering quality can provide insights on just this aspect (Birch, 2011). However, human judgement stays an important factor and is applied as an additional or even main decision criterion for translation quality in evaluation campaigns for machine translation systems (Bojar et al., 2013; Federico et al., 2012). A classification scheme for human error analysis of machine translation is presented in (Vilar et al., 2006). This scheme is also applied in a tool for performing manual error analysis for machine translation (Stymne, 2011), which allows choosing between error classification methods and adding customized error classes. An extensive error analysis of different machine translation systems translating from English and Spanish to Catalán distinguishes linguistic error classes such as orthographic, lexical, morphological, semantic and syntactic errors (Farrús et al., 2012).

A framework towards an automatic error analysis directed in particular at different types of linguistic errors in machine translation is proposed in (Popovic and Ney, 2011), also presenting a human error analysis as a reference for their automatic system. Another framework for semi-automatic error analysis makes use of manual and automatic annotations regarding several characteristics of input documents and connects them with system performance (Kirchhoff et al., 2007).

While inspired by these works on manually assessing the quality of translation system output and the presented error classification schemes, we focus on the comparison of two different translation system outputs. Therefore, we use a customized classification where improvements and degradations between different translations are analyzed instead of errors with regard to a reference translation.

4. Analysis

We perform an analysis of two translation outputs, one using a reordering model based on word-level information only and one using word-level and sentence structure information. By assessing the translation quality and determining the types of improvements and degradations introduced by the more structure-aware reordering model we investigate whether the structure-aware model indeed produces translations with better sentence structure compared to a reordering model using only word-level information. We analyze four different aspects in our comparison of two translation outputs for three different data sets.

4.1. Data

The three data sets used in our analysis represent different genres. The first data set are news texts, which are written in formal style. They typically consist of grammatically correct, but longer and more complex sentences. The second data set consists of human transcripts of TED talks[2]. This type of presentations are perfectly practiced performances, so the speaker hardly make mistakes and spontaneous speech artifacts such as repetitions or stuttering are very rare. The transcripts are edited in subtitle style resulting in a more written form of sentences. The third data

[2] `http://www.ted.com`

set consists of human transcriptions of lectures and talks recorded at a university. Even though obvious spontaneous speech artifacts are removed from the data, no further editing is performed. Consequently, the style resembles more that of actual speech than it is the case with TED talks.

By examining those three types of data, which exhibit different text characteristics and vary in their degree of grammaticality, complexity and spontaneity, we want to assess the impact of the structure-aware reordering approach more thoroughly and find out how it performs in these different environments. For each of the data sets we analyzed between 100 and 166 sentence pairs.

4.2. Impact of Trees depending on genre

We first analyze how much the translations differ when using the word-level and the strucure-aware reordering model. The word-level reordering model only includes reordering rules based on parts-of-speech, whereas the structure-aware model additionally includes the tree-based reordering rules. Since the rest of the translation system is identical and only the reordering model is changed to produce the two translations, there might be sentences which remain unchanged. The first aspect of our analysis therefore considers the amount of sentences affected by the changed reordering model and how this impact varies across the data sets representing different genres.

4.3. Global Sentence Performance

Motivated by the sometimes inconclusive behavior of automatic metrics when measuring joint reordering and translation quality, the second part of the analysis is a manual evaluation of two translation outputs for each of the three data sets. The evaluation consists in a pairwise comparison of the translation quality of the two translations, one produced by the part-of-speech-based reordering model and the other one using the structure-aware reordering rules in addition. For each sentence the source sentence and the two translations are presented without revealing the system which generated the translation. The presentation of the translations takes place in random order to ensure anonymity. Then the overall better translation is chosen allowing ties. This assessment of the global translation performance on sentence basis is also performed on all three genres.

4.4. Local Phenomena

As the third part of the analysis, the changes introduced by the tree-based reordering rules are examined more thoroughly. Each change in the translated sentence is classified according to the three steps presented in Table 1. First, we determine whether it represents an improvement or a degradation of the translation quality. Then further classification is performed, defining the role of the changed word(s) in the sentence, either as its part-of-speech, its constituent role or globally affecting the subject-verb-object (SVO) structure[3]. Then a more fine-grained distiction according to the type of the change may be carried out. Since verbs are our special concern when translating between German and English, for verbs we distinguish between improved/degraded position, insertion, deletion, substitution by an improved/degraded verb form or a different word choice. For most other changes, we only discern between insertion/deletion and position changes.

change	role in sentence	type of change
improvement	verb	insertion
degradation	adv	deletion
	adj	position
	noun	substitution
	negation	word choice
	preposition	word form
	compound	
	PP	
	NP	
	SVO structure	
	. . .	

Table 1: Classes in the Classification Scheme

We provide statistics for total amounts of improvements and degradations introduced by the tree-based reordering model for the three genres and analyze which types of words or sentence parts are prominently affected by the model.

4.5. Local changes and global translation performance

In the last part of the analysis we have a deeper look at the correlation between local changes and the global translation performance on the sentence basis for the individual data sets. We examine how individual improvements and degradations affect overall sentence performance and whether we can draw conclusions about sentence quality when we observe certain changes.

5. Results

In this section we present the results of the analysis. We translated three data sets by applying two versions of the reordering model described in Section 2. within a phrase-based translation system. The first system uses only POS-based reordering and the other one uses POS-based and tree-based reordering together.

In Section 5.1. we give the statistics of the different data sets. Section 5.2. describes how much the data sets were affected by the tree-based reordering model compared to the POS-based reordering model. In addition, we draw the connection to the translation quality measured with an automatic metric. Afterwards, we present the results of the pairwise comparison of translation quality, which was performed manually. The fine-grained analysis is presented in Section 5.4., showing first the number of improvements and degradations introduced by the tree-based reordering and then a more detailed examination of the types of changes. The final section presents the analysis of the correlation between local changes and global sentence performance.

[3] Since we are analyzing English translation output, we expect an SVO sentence structure. When applying this classification to other languages, the correct sentence structure of that language should be applied instead.

5.1. Data statistics

As described in Section 4.1. we used three different data sets for our analysis. First, a news text, namely the news2011 data set from the WMT 2012 evaluation campaign (Callison-Burch et al., 2012) consisting of 3003 sentences. This data set contains various articles from the news domain, covering topics ranging from politics and economics to culture and archictecture. The second test set is the test2010 set from the IWSLT evaluation campaign (Federico et al., 2012). It contains 11 transcribed TED talks consisting of 1565 sentences in total. The topics range from medicine to video games. The lectures test set consists of seven lectures and talks at a university given by five individual speakers from the KIT lecture corpus (Stüker et al., 2012). The transcriptions sum up to 2300 sentences in total.

5.2. Impact of Trees depending on genre

The translation outputs are expectantly similar, since the only difference between the systems is the additional tree-based rules. However, there is a observable difference in the impact of the tree-based rules depending on the genre of the data sets.

Data set	size	POS	+Tree	
News	3003	21.98	22.45	+0.47
TED	1565	30.73	30.87	+0.14
Lectures	2300	25.64	25.65	+0.01

Table 2: *Translation accuracy (BLEU)*

The automatic assessment of translation quality using the BLEU score (Papineni et al., 2002) are presented in Table 2. It can be seen that only for the News data set a measurable difference between the translation quality can be achieved by adding the tree-based rules. For the translation of TED talks and Lectures, the automatic score stays practically the same. It is to be noted that this automatic measurement only represents the translation accuracy on the translated document as a whole.

In order to get a deeper insight into this genre-dependent behavior, we analyzed the impact of the tree-based model on the sentence level. Table 3 shows for each of the three data sets the amount of changed sentences due to the tree-based rules in relation to the total number of sentences in the translated document. For the News data, the translation of 75.5% of the sentences in the testset is changed due to the introduction of the tree-based rules. In contrast, the translation of speech data, i.e. the TED talks and university lectures, is a lot less affected by the tree-based rules. Only between 16.3 and 22.5% of the sentences exhibit a changed translation.

Data set	size	different	%
News	3003	2267	75.5
TED	1565	255	16.3
Lectures	2300	518	22.5

Table 3: *Impact of tree model*

A reason for this difference between written text and speech data may be due to their different textual characteristics. Written text tends to contain more complex sentences, which is the types of sentences where the tree-based reordering model can exert its strengths best. In spoken performances, overly complex sentences structures are typically avoided in order to facilitate comprehension on the part of the audience. Shorter and less complex sentences can be addressed well with the POS-based reordering rules, which explains why often word orders proposed by the tree-based model are not chosen for translation.

In order to confirm this assumption we examine different aspects of the data that could give an indication of the complexity of sentences. First of all, sentence length and the number of punctuation marks could be an indictor for complexity, since this increases parsing difficulty and could lead to erroneous parse trees.

Data set	sentence length (avg.)		# punct per sentence	
	all	subset	all	subset
News	20.83	23.29	4.8	5.1
TED	16.29	25.00	3.9	5.7
Lectures	19.30	27.01	4.3	4.8

Table 4: *Analysis of textual complexity*

Table 4 shows the two aspects mentioned above: average sentence length and number of punctuation marks per sentence both for the subsets of affected sentences and all sentences of the three data sets. As expected, the average amount of words per sentence as well as the number of punctuation marks is highest in the News data set. For the speech data sets, lectures contain longer sentences and more punctuation marks due to the specialized content in the university setting. TED talks are more general, popular talks directed at a broader audience where the appropriate presentation style consists of shorter, concise sentences. When considering only the subset of sentences affected by the tree-based rules, we can see the average sentence length as well as the number of punctuation marks increases for all data sets. This corresponds with our expectation that longer and complex sentences are explicitly targeted by the tree-based rules. For the subset, where the tree rules lead to different translations, the sentence length for the speech data is even longer than for text data. The reason might be that for the same sentence length, the structure of a written text is more complex than for a speech text. Therefore, the tree rules are already more important for shorter sentences.

These results may explain the difference in the proportion of affected sentences for the different data sets shown in Table 3. The differences in automatic translation scores between data sets will also be related to this finding. Since a lot fewer sentences are changed in the speech data sets, the tree rules' influence on the whole document is lower and is less noticeable in the BLEU score.

In order to evaluate the impact of the tree-based rules on the translation quality without this bias of amount of unchanged sentences, we calculate the translation accuracy on a subset of the original data set consisting of the changed sentences only. Table 5 shows the automatic translation scores for these subsets.

Data set	size	POS	+Tree	
News	2267	21.38	21.87	+0.49
TED	255	27.10	27.51	+0.41
Lectures	518	23.53	23.60	+0.07

Table 5: *Translation accuracy on subsets (BLEU)*

These new scores show that for the TED data it was indeed the case that the lower number of affected sentences led to a underestimation of the impact of the tree-based reordering on the automatically measured translation quality. For the News data, the impact was already obvious, since the bigger part of the sentences were already affected by the tree-based model. Excluding the remaining sentences from the automatic scoring did not change the score much. We can therefore argue that the tree-based reordering affects the translation of the TED talks positively in a similar way as the News data, whenever the application of the tree-based reordering rules results in a changed translation.

However, the automatic translation score for the translation of lectures shows not much of a difference compared to the previous results in Table 2.

After investigating the impact of the tree-based reordering model in various ways, we examined changed translation hypotheses manually to find out whether the change introduced by the tree-based reordering resulted in a better translation.

5.3. Global Sentence Performance

From all the sentences which were translated differently due to the tree-based reordering, we extracted sentences from each of the data sets for manual analysis. Table 6 shows the exact amount of sentences analyzed for each data set. For TED and News data, the first 100 and 165 of the changed sentences of the document were chosen. For the lecture data, 166 sentences were chosen for analysis by taking an even amount from each of the individual transcribed lectures.

Data set	size
News	165
TED	100
Lectures	166

Table 6: *Amounts of manually analyzed data*

We analyzed the global sentence performance by comparing the two translation hypotheses using POS-based rules only and POS and tree-based reordering rules. Table 7 shows the results. We can see that in 55-64% of the cases, the system using tree-based rules produced a better translation, while the translation using POS-based reordering only was considered the better translation in 24 to 28% of the sentences. There are more tree wins for the speech data sets than for the news data. However, the amount of POS wins is bigger for the speech data, while the amount of ties is lower. This might be both due to the abovementioned easier structure of speech sentences and the mismatch of training and test data for the parser.

Data set	Tree win	tie	POS win
News	55.8	19.4	24.9
TED	64.0	8.0	28.0
Lectures	60.8	12.7	26.5

Table 7: *Manual sentence-level analysis (%)*

In contrast to the automatic evaluation, which only indicates an improvement on the TED and News talks, the manual evaluation shows that the translation quality is improved on all different data set using the tree-based reordering approach.

5.4. Local Phenomena

While the previous section presented an analysis of the global sentence performance, considering each translation hypothesis as a whole, now we investigate the local phenomena more thoroughly, i.e. the individual changes of words and structure between the two hypotheses. We identify the changed regions in each sentence pair and determine for each of the changes introduced by the tree-based system, whether it improves the translation or degrades it.

Table 8 shows the amounts of improvements (++) and degradations (- -) among the total number of changes in all analyzed sentences of each data set. The News data set includes the lowest number of changes per sentence. More changes per sentence can be found in the two speech data sets. Consequently, even though much less sentences are affected by the tree-based model in the speech data sets (16% and 22% vs. 75% of the sentences, cf. Table 3), more changes are introduced per sentence in the affected sentences (1.3 in speech vs. 1.1 in text data).

Data set	++	%	- -	%	total	per sent.
News	119	65.0	64	35.0	183	1.11
TED	92	70.2	39	29.8	131	1.31
Lectures	159	70.4	67	29.6	226	1.36

Table 8: *Local phenomena*

Tables 9 and 10 show what types of changes can be discerned in the improvements and degradations, respectively. We differentiate substitutions, insertions and deletions of words as well as position changes. Substitutions include different word choice and changed tense or other morphological change of the word form.

As can be seen, there is again a difference between the text and speech data sets. For the News data, nearly half of the improving changes (44%) are insertions of words, i.e. words appear in the translation that were not translated before. The rest of the changes are substitutions, i.e. different word choices (25%) and improved word positions (30%). For the two speech data sets, the biggest share of the improvements affect the position, (43 and 42%), while substitutions and insertions make up a smaller portion of the improvements. Deletions typically do not have a positive effect on the translation.

	++		
	News	TED	Lectures
substitution	25.2	23.9	30.2
word choice	20.2	19.6	23.3
word form	5.0	4.3	6.9
position	30.3	43.5	42.1
insertion	44.5	32.6	27.0
deletion	0.0	0.0	0.6
total	100.0	100.0	100.0

Table 9: *Local phenomena - types of improvements (%)*

Analyzing the types of negative changes showed that for News and TED data the main source of degradations is word substitutions, i.e. different word choices or word forms that change the translation quality for the worse. For the lectures it is the changed positions and deleted words that make up most of the negative changes, which is more than for the other two data sets. This might be a reason for the low BLEU improvement on the lecture test set observed in Table 5.

	- -		
	News	TED	Lectures
substitution	46.9	51.3	22.4
word choice	39.1	38.5	20.9
word form	7.8	12.8	1.5
position	34.4	25.6	43.3
insertion	0.0	7.7	0.0
deletion	18.8	12.8	34.3
total	100.0	100.0	100.0

Table 10: *Local phenomena - types of degradations (%)*

Tables 11 and 12 show the types of changes according to word classes and sentence constituents. Changes in word form, position, insertions and deletions related to a word class are analyzed. Different word choices leading to a better or worse translation are not taken into account. It is observable that the mainly affected word classes are verbs and adverbs throughout all data sets. Others are nouns and pronouns as well as prepositions. Regarding sentence structure, the position of whole prepositional phrases is one of the more prominently affected parts of the sentence.

The main word classes affected by degradations of translation quality are mainly the same as for improvements, as Table 12 shows. Although a lot less degradations are introduced by the tree-based reordering model, the changes still mainly affect the verbs, adverbs, nouns, pronouns, prepositions and prepositional phrases. As mentioned earlier, the main types of degradations are degraded position and erroneously removed words.

	++		
	News	TED	Lectures
verb	49	53	81
adverb	9	6	11
pronoun	0	7	5
noun	7	1	2
compound	2	0	3
determiner	3	0	1
adjective	1	0	0
preposition	8	1	2
conjunction	2	1	4
negation	1	0	1
interjection	0	0	1
PP	9	4	8
NP	1	1	2
SVO structure	3	0	0
clause	0	0	1
	95	74	122
word choice	24	18	37
total	119	92	159

Table 11: *Local phenomena - word classes (improvements)*

	- -		
	News	TED	Lectures
verb	14	9	18
adverb	2	0	8
pronoun	3	5	2
noun	3	0	4
compound	4	0	4
adjective	1	1	0
preposition	4	1	3
conjunction	0	2	3
interjection	0	0	1
PP	3	3	2
NP	3	0	5
SVO structure	1	1	1
clause	0	0	1
object	1	2	0
subject	0	0	1
	39	24	53
word choice	25	15	14
total	64	39	67

Table 12: *Local phenomena - word classes (degradations)*

5.5. Local changes and global translation performance

How are local changes correlated with the global translation performance? Table 13 shows how many of the positive changes in all word classes and in the verbs class only shown in Table 11 above were observed in a Tree win or POS win sentence. We can draw from these numbers that between 90.8 and 96.2 % of the improving changes in all classes result also in a globally improved translation quality. When we examine only the verbs, the tendency is similar. Between 83.7 and 95.1 % of the verbal improvements stem from a sentence produced by the tree-based reordering model and represent an improvement in translation quality over the sentence produced by the POS-based reordering model.

	News	++ TED	Lectures
all classes			
Tree wins	90.8	94.6	96.2
POS wins	5.0	5.4	4.4
verbs			
Tree wins	83.7	92.5	95.1
POS wins	12.2	7.5	6.2

Table 13: *Local vs. global (improvements) (%)*

Table 14 shows the correlation between degradations and glocal sentence quality. We have already established that a lot fewer negative changes than positive changes are introduced by the tree-based system. The previous table might indicate that a negative change should also correspond more likely with a worse translation quality of the output of the translation system using the tree-based reordering output, i.e. a POS win. When analyzing all word and constituent classes, the correlation between negative changes and POS wins is between 70.3 and 80.6 %. For the verbs, the correspondence is a little higher, between 71.4 and 88.9 %. However, the correlation is not as high as for positive changes with improved translation quality.

	News	- - TED	Lectures
all classes			
Tree wins	17.2	20.5	19.4
POS wins	70.3	79.5	80.6
verbs			
Tree wins	14.3	11.1	16.7
POS wins	71.4	88.9	83.3

Table 14: *Local vs. global (degradations) (%)*

Hence, we can conclude that local improvements introduced by the tree-based model will most likely coincide with an overall better translation quality of that given sentence. Local degradations are not necessarily to correspond with a degraded translation quality, although degradations in verbs have a more severe influence on the translation quality.

6. Conclusion

We have presented an in-depth analysis of a structure-aware word reordering approach for German to English phrase-based machine translation. We examined the changes in the translation output introduced by automatically learned tree-based reordering rules compared to part-of-speech-based reordering rules. We compared the results on three data sets which differ in genre and topic.

Our findings have shown that according to manual evaluation the structure-aware reordering approach helps produce an improved translation quality on all three data sets. The impact of the reordering model is higher on data that is well structured, grammatically correct texts, while fewer sentences were affected for the two speech data sets. When taking into account the affected sentences only, the translation quality as measured with the automatic metric BLEU behaved similarly on the News and the TED data.

The manual evaluation of sentence-level translation quality confirmed consistent improvements by the tree-based reordering model throughout all three data sets.

A similar behavior on the three data sets can also be reported for the local improvements in the sentence which include translations of words which were not translated before and improved word and consituent positions in the translated sentence. As intended in the design of the tree-based reordering model, verbs are the main cause for local improvements.

We observed a high correlation between local improvements in the sentence and an overall better sentence quality, while a local degradation not necessarily leads to a worse translation on the sentence level.

The presented work shows that a manual analysis, even though costly and time-consuming, can help understand the influence of a new model in a given translation framework. Especially in cases, where automatic metrics show inconsistent behavior, e.g. for different data or languages, it can prove important to measure the impact in form of the actual sentences that are affected and to inspect the introduced changes more thoroughly.

Although the presented work focused on the German-English language pair, the reordering model is language independent an can be applied to any language pair where there is a parser for the source language available. The analysis framework is also applicable to other languages. Since we expect the impact and affected word classes to be highly dependent on the language pair and even the translation direction, it would be interesting future work to perform such an analysis on another language pair or translation direction.

7. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

8. References

Alexandra Birch. 2011. *Reordering Metrics for Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Mireia Farrús, Marta R. Costa-Jussà, and Maja Popovic. 2012. Study and correlation analysis of linguistic, perceptual, and automatic machine translation evaluations. *JASIST*, 63(1):174–184.

Marcello Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Teresa Herrmann, Jan Niehues, and Alex Waibel. 2013. Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, Georgia, USA.

Katrin Kirchhoff, Owen Rambow, Nizar Habash, and Mona Diab. 2007. Semi-automatic error analysis for large-scale statistical machine translation systems. In *Proceedings of the MT Summit XI*, Copenhagen , Denmark.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, September.

Jan Niehues and Muntsin Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109-022), IBM Research Division, T. J. Watson Research Center.

Maja Popovic and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.

Sebastian Stüker, Florian Kraft, Christian Mohr, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2012. The kit lecture corpus for speech translation. In *LREC*, pages 3409–3414, Istanbul, Turkey.

Sara Stymne. 2011. Blast: A tool for error analysis of machine translation output. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, Portland, Oregon, USA.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, Genoa, Italy.