Enriching Phrase-Based Statistical Machine Translation with POS Information

Miriam Kaeshammer and **Dominikus Wetzel**
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{miriamk,dwetzel}@coli.uni-sb.de

Abstract

This work presents an extension to phrase-based statistical machine translation models which incorporates linguistic knowledge, namely part-of-speech information. Scores are added to the standard phrase table which represent how the phrases correspond to their translations on the part-of-speech level. We suggest two different kinds of scores. They are learned from a POS-tagged version of the parallel training corpus. The decoding strategy does not have to be modified. Our experiments show that our extended models achieve similar BLEU and NIST scores compared to the standard model. Additional manual investigation reveals local improvements in the translation quality.

1 Introduction

Currently, the most prominent paradigm in statistical machine translation (SMT) are phrase-based models (Koehn et al., 2003), in which text chunks (*phrases*) of one language are mapped to corresponding text chunks in another language. This standard approach works only with the surface forms of words and no linguistic information is used for establishing the mapping between phrases or generating the final translation. It has been shown, however, that integrating linguistic knowledge, e.g. part-of-speech (POS) or morphological information, in pre- or post-processing or directly into the translation model improves the translation quality (cf. Section 2).

Factored translation models (Koehn and Hoang, 2007) are one extension of the standard phrase-based approach, which allow to include rich linguistic knowledge into the translation model. Additional models for the specified factors are used, which makes decoding computationally more complex as the mapping between the factors can result in an explosion of translation options.

With this work, we explore a different approach to integrate linguistic knowledge, in particular POS information, into the phrase-based model. The standard phrase (translation) table is enriched with new scores which encode the correspondence on the POS level between the two phrases of a phrase pair; for example the probability of "translating" the POS sequence of one phrase into the POS sequence of the other phrase. We propose two methods to obtain such *POS scores*. These extra scores are additional feature functions in the log-linear framework for computing the best translation (Och and Ney, 2002). They supply further information about the phrase pairs under consideration during decoding, but do not increase the number of translation options.

The presented extension neither makes use of hand-crafted rules nor manually identified patterns. It can therefore be performed fully automatically. Furthermore, our approach is language-independent and does not rely on a specific POS tagger or tag set. Adaptation to other language pairs is hence straightforward.

This paper first describes related work and then introduces our extended translation model. Evaluation results are reported for experiments with a German-English system. We finally discuss our work and suggest possible further extensions.

2 Related Work

There are several strategies for improving the quality of standard phrase-based SMT by incorporating linguistic knowledge, in particular POS information.

One such approach is to modify the data in a pre-processing step. For example, Collins et al. (2005) parse the sentences of the source language and restructure the word order, such that it matches the target language word order more

Proceedings of the Student Research Workshop associated with RANLP 2011, pages 33–40,
Hissar, Bulgaria, 13 September 2011.

closely. Language-specific, manually devised rules are employed. Popović and Ney (2006) follow the same idea, but make use of manually defined patterns based on POS information: e.g. local adjective-noun reordering for Spanish and long-range reorderings of German verbs. Essentially, this strategy aims at facilitating and improving the word alignment. Another example along those lines is (Carpuat, 2009). Surface words in the training data are replaced with their lemma and POS tag. Once the improved alignment is obtained, the phrase extraction is based on the original training data, thus a different decoding strategy is not necessary. Another data-driven approach is presented in (Rottmann and Vogel, 2007), where word reordering rules based on POS tags are learned. A word lattice with all reorderings (including probabilities for each) is constructed and used by the decoder to make more informed decisions.

Another strategy is concerned with enhancing the system's output in a post-processing step. Koehn and Knight (2003) propose a method for noun phrases where feature-rich reranking is applied to a list of n-best translations.

Instead of the above pre- or post-processing steps, Koehn and Hoang (2007) present factored models which allow for a direct integration of linguistic information into the phrase-based translation model. Each surface word is now represented by a vector of linguistic factors. It is a general framework, exemplified on POS and morphological enrichment. In order to tackle the increasing translation options introduced by additional factors, the decoding strategy needs to be adapted: translation options are precomputed and early pruning is applied. Factored models including POS information (amongst others) are employed for example by Holmqvist et al. (2007) for German-English translation and Singh and Bandyopadhyay (2010) for the resource-poor language pair Manipuri-English.

3 Extended Translation Model

The general idea is to integrate POS information into the translation process by adding one or several *POS scores* to each phrase pair in the standard phrase table which represents the translation model and usually contains phrase translation probabilities, lexical weightings and a phrase penalty. The additional scores reflect how well the POS sequence which underlies one phrase of the pair corresponds to the POS sequence of the other phrase of the pair. Two concrete methods to calculate this correspondence will be described in Section 3.2. The new scores can be integrated into the log-linear framework as additional feature functions.

Figure 1 shows two phrase pairs from a German-English phrase table. In this particular case, the POS scores should encode the correspondence between `ART ADJA NN` from the German side and `DT JJ NNS VBN` (a) or `DT JJ NNS` (b) from the English side. Intuitively, `ART ADJA NN` corresponds better to `DT JJ NNS` than to `DT JJ NNS VBN`. Phrase pair (b) should therefore have higher POS scores.

The transition from the standard translation model to the extended one can be broken up into two major steps: (1) **POS-Mapping**, which is the task of mapping each phrase pair in the standard phrase table to its underlying pair of POS sequences (henceforth *POS phrase pair*), and (2) **POS-Scoring**, which refers to assigning POS scores to each phrase pair based on the previously determined POS phrase pair.

3.1 POS-Mapping

Obtaining the part-of-speech information for each phrase in the phrase table cannot be achieved by tagging the phrases with a regular POS tagger. They are usually written for and trained on full sentences. Phrases would therefore get assigned incorrect POS tags, since a phrase without its context and the same phrase occurring in an actual sentence are likely to be tagged with different POS sequences.

Since the phrase pairs in the phrase table originate from specific contexts in the parallel training corpus, we require a phrase to have the same POS sequence as it has in the context of its sentence. Consequently, our approach takes the following steps: First, both sides of the training corpus are POS-tagged. Secondly, the untagged phrases in the phrase table and their tagged counterparts in the corpus are associated with each other to establish a mapping from phrase pairs to POS phrase pairs. This procedure is consequently not called POS-Tagging, but rather POS-Mapping.

Our approach is to apply the same phrase extraction algorithm again that has been used to obtain the standard phrase table. Phrase pairs are ex-

(a) die möglichen risiken ||| the possible risks posed ||| 1.0 [...] 0.155567 0.000520715
(b) die möglichen risiken ||| the possible risks ||| 0.1 [...] 0.178425 0.0249141

Figure 1: Two phrase pairs, each with the first standard translation score and two new POS scores.

tracted from the POS-tagged parallel training corpus, thereby taking over the word alignments that have been established for the parallel sentences to extract standard phrase pairs before. In the resulting *word/POS phrase table*, a token is a combination of a word with a POS tag. For this to work, words and POS tags must be delimited by any special character other than a space. Thanks to the reused word alignments, the word/POS phrase table contains each phrase pair of the standard phrase table at least once. If a phrase pair occurs with several different POS sequences in the training data, the word/POS phrase table contains an entry for each of them.

By matching the standard phrase table against the word/POS phrase table, the POS phrase pair(s) for each standard phrase pair are obtained. The word/POS phrase table is hence used as the mapping element between phrase pairs and their corresponding POS phrase pairs. The result of this POS-Mapping step is a $1 : k$ (with $k \geq 1$) mapping from phrase pairs to POS phrase pairs. The POS phrase pairs are the basis for calculating the POS scores as explained in the following subsection.

An alternative approach to POS-Mapping would be a search for the phrases in the tagged sentences. This however requires elaborate techniques such as indexing.

3.2 POS-Scoring

We propose two different kinds of POS scores to encode the correspondence on the POS level between the two phrases of a phrase pair: *POS Phrase Translation (PPT)* and *POS Phrase Frequency (PPF)* scores.

PPT scores PPT scores encode how likely it is to "translate" one POS phrase into another POS phrase. The idea behind those scores and also the way how they are obtained is very similar to the scores in a standard phrase table, namely translation probabilities and lexical weightings. The difference is that the tokens that constitute the phrases are POS tags. Consequently, phrase pair extraction and phrase pair scoring (maximum likelihood estimation for translation probability and lexical weighting in both translation directions) is performed on a version of the parallel training corpus, in which each word is substituted by its POS tag. Again, as we did in Section 3.1 to obtain the word/POS phrase table, the word alignments that were established to extract the standard phrase pairs are reused.

In this way, a *POS phrase table* is trained which has four scores attached to each POS phrase pair. Those are the desired PPT scores. Due to the reused word-alignment, it contains all POS phrase pairs that also occur in the word/POS phrase table.

The standard phrase table is combined with the new PPT scores via the mapping from phrase pairs to POS phrase pairs introduced in Section 3.1. As this is a $1 : k$ mapping, it needs to be decided which of the k POS phrase pairs and corresponding scores to use. Currently, we decide for the POS phrase pair for which the sum of the scores is maximal and use the corresponding PPT scores $\hat{s}$:

$$\hat{s} = \operatorname*{argmax}_{s_k} \sum_{i=1}^{|s_k|} s_k(i) \qquad (1)$$

where k ranges over the POS phrase pairs which are mapped to the current phrase pair, s_k are the (four) PPT scores of the kth POS phrase pair and i is an index into these scores. This decision rule is a crucial point in the extended model where additional experiments using other techniques should be conducted.

From the four PPT scores in $\hat{s}$, several extended translation models have been derived which differ in the number of scores that are added to the standard phrase table: i. all 4 PPT scores, ii. only the phrase translation probabilities (PPT scores 1 and 3), iii. only the lexical weightings (PPT scores 2 and 4) and iv. only the inverse phrase translation probability (PPT score 1).

As an example, the last two scores on each line in Figure 1 are PPT scores (phrase translation probabilities) that have been obtained with the described method. Indeed both are higher for (b), which coincides with our expectation.

PPF score The PPF score encodes the raw frequency of POS phrase pairs; more specifically how often a POS phrase pair occurs in the word/POS phrase table (see Section 3.1). The intuition behind it is that POS phrase pairs which correspond to more than one distinct surface phrase pair are more reliable than POS phrase pairs that produce only one type of phrase pair. The latter could for example originate from a wrong alignment. This score abstracts away from directly counting the phrase pair occurrences in the parallel training corpus, which is information that is already incorporated in the standard phrase table scores.

To combine the obtained counts with the standard phrase table, we again use the $1 : k$ mapping from phrase pairs to POS phrase pairs and select the maximum out of the k PPF scores. As an example, phrase pair (a) in Figure 1 receives a PPF score of 289, while phrase pair (b) has PPF score 9735, according to the most frequent underlying POS phrase pair.

We anticipate the issue that shorter phrase pairs get higher counts, since their corresponding POS sequences are more likely to occur in the word/POS phrase table. This seems to result in a bias towards selecting shorter phrases during decoding, which stands in contrast to a phrase penalty which favors longer phrases that is commonly employed in phrase-based translation systems. We assume that the tuning procedure will find weights for the feature functions such that those two complement each other.

4 Experiments

For our experiments we used the Moses phrase-based SMT toolkit (Koehn et al., 2007; Koehn, 2010) to train translation systems from German to English.

4.1 Data

As training data we used the German and English documents from the Europarl Corpus Release v5 (Koehn, 2005), excluding the standard portion (Q4/2000). The data was sentence-aligned, tokenized and lowercased by the provided scripts. Sentences longer than 40 tokens on either language side were removed with their translations from the training corpus, resulting in about 1.1 million sentence pairs. From the held-out data 3000 sentences for development and 2000 sentences for testing were randomly chosen.

To generate the POS-tagged version of the tokenized training data, we applied the OpenNLP 1.4 POS tagger[1] using the provided German and English models. Afterwards, the POS-tagged training corpus was lowercased.

For the language model, we used the English side of the complete training corpus containing the lowercased data (about 1.5 million sentences). The model was generated with the SRILM toolkit 1.5.8[2] using 3-grams and Kneser-Ney discounting.

4.2 Setup

We used the Moses training script with the standard parameters except for the alignment heuristic (`grow-diag-final-and`) together with GIZA++ (Och and Ney, 2003) to train the standard translation model. The obtained word alignment was used for phrase extraction and scoring in order to construct the word/POS phrase table and the POS phrase table.

We tuned our extended systems as well as the standard system with minimum error rate training (MERT) (Och, 2003). For the extended models, the tuning script that comes with Moses needs to be adapted slightly. Additional triples specifying initialization and randomization ranges for the weights of our additional feature functions have to be inserted. Because of the possibility that the MERT algorithm gets trapped in a local maximum, several tuning runs for the same model with the same development data were performed.

We skipped recasing and detokenization, since we are only interested in the effect of our extended model with respect to the baseline.

4.3 Results

Table 1 shows the automatic evaluation of the outcome of the conducted experiments. Our extended model with all four PPT scores (t1) achieved the best results, followed by the baseline (t2) in terms of BLEU and our PPF model according to NIST. However, the reported scores are similar and the differences in performance between the extended models and the baselines are insignificant.

For two models, we report the performance of the systems that were obtained with two independent tuning instances (t1 and t2) in Table 1. The varying scores indicate the importance of the tuning step.

[1] `http://opennlp.sourceforge.net/`

[2] `www.speech.sri.com/projects/srilm/`

		BLEU	NIST
Standard Model	Baseline (t1)	25.59	6.7329
	Baseline (t2)	25.83	6.7817
Extended Model	all 4 PPT scores (t1)	**25.88**	**6.8091**
	all 4 PPT scores (t2)	25.60	6.7835
	PPT scores 1 and 3	25.58	6.7651
	PPT scores 2 and 4	25.66	6.7758
	PPT score 1	25.61	6.7590
	PPF score	25.73	6.7882

Table 1: Performance of the models on the test set.

To sum up, according to the automatic evaluation, none of our extended models clearly outperforms the baseline. This could suggest on the one hand that the additional POS scores do not lead to better translation models. One the other hand, BLEU and NIST might just not be able to reflect our improvements in the translation models and quality. They are automatic metrics and we only provide one reference translation for each sentence in the development and test data. Consequently, further inspection of the translated data is necessary.

4.4 Manual Investigation

Out of the 2000 test sentences, our extended model with all four PPT scores (t1) provides the same translation as the baseline (t2) in 613 cases (470 for t2 of the extended model). To find out about the variations that occur in the translations which differ, we manually inspected some sample sentences. As follows, we will present and describe the examples in Figure 2. In (2a) – (2f) the translation of our extended model is better than the one provided by the baseline system.

The baseline translation in (2a) is neither understandable nor grammatical. Our model accomplishes to translate the two genitive constructions and provides a suitable translation for the verb, which is missing completely in the baseline. In (2b) the relative clause construction in the scope of the negation is missing in the baseline. This leads to a severe change in meaning. The sentence provided by the extended system, in contrast, is fully meaningful and understandable. Obviously, it is not perfect; for example, *philosophical sense* lacks a determiner.

The baseline system provides ungrammatical translations that are hardly understandable for the test sentences in (2c), (2d) and (2e). In (2c) *wie* is not translated as the interrogative pronoun, and in (2d) the infinitive verb is missing. Our extended system produces good translations for both sentences. The test sentence in (2e) is difficult for machine translation because the verb in the subordinate clause is omitted from the first part of the conjunction. In fact, both systems cannot handle it. However, the extended system at least achieves to put the right content words into the two parts of the coordination; only the verb in the first part is missing.

Example (2f) shows that our extended model helps at conveying the semantics of the source sentence. The translation given by the baseline is not completely wrong, but it fails at expressing the *possibility* of the conflict and also the *process* of getting into a conflict. The translation of our extended system (*which could come into conflict*) conveys both.

There are also sentences within the test set, on which the baseline system performs better than the extended model. In (2g), the translation given by our model lacks a conjugated verb. (2h) shows an instance of a wrongly translated pronoun by our extended system. The sentence is furthermore ungrammatical whereas the translation by the baseline system is acceptable.

The given examples have revealed that the differences in the translations provided by the baseline system and our extended system are generally local. Often only a small number of words is affected. However, even local changes lead to better translations as shown in the examples (2a) – (2f). It is left to quantify these results to check whether the extended translation model overall introduces more improvements or deteriorations.

The examples in Figure 2 also illustrate why BLEU and NIST do not show a difference between the extended system and the baseline: Even if a translation is acceptable, it is usually very different from the provided reference translation. The small improvements are consequently not reflected in the automatic score.

5 Discussion & Future Work

With our extended model, we are able to incorporate linguistic information into the otherwise pure statistical MT approach. We have realized our approach within the framework provided by Moses and its tools, but other phrase-based SMT systems could be extended in the same way. Once the

scores encoding the additional information are calculated, almost no modification to existing code is necessary.

The presented method does not make use of any language-specific behavior or patterns, which leaves it open to any language combination, provided that there are POS taggers for the involved languages available. Since no hand-crafted rules need to be designed for the extension, our approach can be applied to new language pairs with only a minimum amount of time and effort. Moreover, any POS tagger with any POS tag set can be used in order to annotate the training data. It is also noteworthy that POS tagging is only needed during training and not during decoding.

The automatic evaluation represented in the BLEU/NIST scores showed only insignificant improvement for our extended system over the baseline. However, a manual investigation of the translated test data revealed qualitatively better translations. Some local phenomena seem to be handled better in the linguistically informed model. Certainly, in order to make reliable judgments, human evaluation of a representative set of translations is needed.

Tuning the weights of the feature functions is an essential step for obtaining a good translation system (cf. (Koehn, 2010)). The effect of different tuning instances on the translation output and thus BLEU/NIST can be seen in our experimental results in Table 1. Accordingly, it needs to be determined whether the MERT algorithm is still capable of finding good weights when more than the standard weights need to be tuned. A review of the literature did not clarify the impact of the number of weights on MERT tuning. Other tuning algorithms could be considered. Furthermore, MERT relies on automatic evaluation metrics. Because of their shortcomings (cf. (Callison-Burch et al., 2006)), the tuning approach might not exploit the full potential of the additionally encoded linguistic information. An improvement would be to include a human-based evaluation component in MERT (cf. (Zaidan and Callison-Burch, 2009)).

A very important further step would be to fully compare our approach to factored models (using POS information on the source and target side) (Koehn and Hoang, 2007) under the same experimental conditions as reported in this work. From a theoretical point of view, the main difference between our approach and the factored models is that the linguistic information is explicitly encoded in several phrase tables in the latter, while in the former it is implicit in the additional score(s) in just one phrase table. As mentioned before in Section 2, factored models have the shortcoming of a drastic rise of translation options during decoding. Our approach, in contrast, does not change the number of translation options. It rather provides more informed phrase pair selection criteria by means of the POS scores. The decoding strategy therefore does not need to be adapted.

Interestingly, Koehn and Hoang (2007) report only minor improvements in BLEU for their English-German system when using only the surface form and POS in the factored models. However, they report a greater improvement when also adding morphological information. This could suggest that POS information on its own is not informative enough to improve the BLEU score.

There are various ways to improve and extend the presented approach. One crucial point where we have made a rather ad hoc decision is the procedure in Equation 1. Ideally, one would want to use the POS scores that are optimal with respect to the translation result. Furthermore, this procedure should be improved such that it only considers the subset of POS scores that is actually used in the final phrase table.

Possible extensions of the models in our fashion are not only tied to POS information. One could for example incorporate more structured information such as dependency relations. This information would be assigned to a word just like the POS tag has been. More specifically, we suggest to consider the following two approaches: (1) Tokens get assigned the number of their dependants, e.g. Peter`/0` likes`/2` Mary`/0`. (2) Dependent tokens get assigned a tuple specifying their dependency type and their head word, e.g Peter`/(subj,likes)` likes`/(root,nil)` Mary`/(obj,likes)`. As this approach might run into data sparsity problems, as a variant, the dependency type could be omitted. Once one of the above syntax taggings is generated, the mapping and scores for the phrase table can then be obtained just as before with the POS-Mapping/Scoring approach.

6 Conclusion

We have described a language-independent approach to incorporate linguistic information such

as POS tags into phrase-based SMT. We achieved this by enriching the phrase pairs in the standard phrase table with additional POS scores which reflect the correspondence between the underlying POS sequences of the phrases of each pair. Two kinds of POS scores have been proposed: *POS Phrase Translation* scores from a learned phrase table based on POS sequences and *POS Phrase Frequency* scores which are raw counts of POS sequence pairs. To assign the scores to the standard phrase pairs, they have been mapped to their underlying POS sequences (via a word/POS phrase table). In order to extract the same phrases across all phrase tables, the word alignment of the standard phrase table has been reused. In experiments for German-English, automatic evaluation showed minor differences in performance between the extended systems and the baseline. Additional manual inspection of the results revealed promising local improvements. Compared to the factored models, our extension uses linguistic information implicitly, does not provide additional translation options and therefore does not introduce further complexity for decoding.

Acknowledgments

This work was part of a project seminar at Saarland University. We would like to thank our supervisor Andreas Eisele from DFKI for the ideas that got us started and his support. Part of this research was funded through the European Community's Seventh Framework Programme under grant agreement no. 231720, EuroMatrix Plus.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL*, pages 249–256.

Marine Carpuat. 2009. Toward Using Morphology in French-English Phrase-Based SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 150–154, Athens, Greece. ACL.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, Michigan. ACL.

Maria Holmqvist, Sara Stymne, and Lars Ahrenberg. 2007. Getting to Know Moses: Initial Experiments on German-English Factored Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic. ACL.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on EMNLP and CoNLL*, pages 868–876.

Philipp Koehn and Kevin Knight. 2003. Feature-Rich Statistical Translation of Noun Phrases. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 311–318, Sapporo, Japan.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL 2003*, pages 48–45.

Philipp Koehn, 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 *Annual Meeting of the ACL*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit*.

Philipp Koehn, 2010. *Moses. Statistical Machine Translation System*. User Manual and Code Guide.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of ACL*, pages 295–302.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of ACL*, pages 160–167.

Maja Popović and Hermann Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Skövde, Sweden.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-English Bidirectional Statistical Machine Translation Systems using Morphology and Dependency Relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91, Beijing, China. Coling 2010 Organizing Committee.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of Human-in-the-loop Minimum Error Rate Training. In *Proceedings of the 2009 Conference on EMNLP*, pages 52–61, Singapore. ACL.

German	sechsundsechzig prozent der gesamtbeschäftigung der gemeinschaft entfallen auf kleine und mittlere unternehmen [. . .]
Baseline	the community sechsundsechzig per cent of total employment in small and medium-sized enterprises [. . .]
4 PPT	sechsundsechzig % of total employment of the community is generated by small and medium-sized enterprises [. . .]
Reference	smes account for 66 % of total employment in the community [. . .]

(a) Handling genitive constructions and translating the main verb correctly

German	es gibt kein volk in europa , das im philosophischen sinne neutral ist .
Baseline	there are no people in europe , in the philosophical sense is neutral .
4 PPT	there is no people in europe , which is neutral in philosophical sense .
Reference	there is no nation in europe that is philosophically neutral .

(b) Handling a relative clause correctly

German	wie ist nun der konkrete stand der verhandlungen ?
Baseline	as is now the real state of negotiations ?
4 PPT	so what exactly is the real state of negotiations ?
Reference	what stage has actually been reached in these negotiations ?

(c) Translating interrogative pronoun properly

German	sie haben natürlich recht , immer wieder auf diese frage zu verweisen .
Baseline	you are right , of course , to this question again and again .
4 PPT	you are right , of course , always to refer to this question .
Reference	you are , of course , quite right to keep reverting to this question .

(d) Missing verb in baseline translated properly in our model

German	abschließend möchte ich noch sagen , dass die postdienstleistungen in schweden nicht schlechter und in gewisser weise sogar besser geworden sind .
Baseline	finally , i would like to say that the postal services in sweden and in some way not worse even improved .
4 PPT	finally , i would like to say that the postal services not worse in sweden and in some way have become even better .
Reference	in conclusion , i would like to say that the postal service in sweden has not deteriorated , in some respects it has even improved .

(e) Tricky coordinate construction with omitted verb

German	auf diese weise [. . .] könnten machtzentren geschaffen werden , die untereinander in konflikt geraten könnten .
Baseline	in this way [. . .] machtzentren could be created , in conflict with each other .
4 PPT	in this way [. . .] machtzentren could be created , which could come into conflict with each other .
Reference	[. . .] there is a real danger that this will result in conflicting centres of power .

(f) Conveying correct semantics

German	schließlich beruht jedes demokratische system auf dem vertrauen und dem zutrauen der menschen .
Baseline	finally , any democratic system is based on the confidence and the trust of the people .
4 PPT	finally , any democratic system based on the confidence and the trust of the people .
Reference	after all , any democratic system is built upon the trust and confidence of the people .

(g) Wrong translation due to missing verb

German	der rat möchte daran erinnern , dass seine politik stets darauf abzielt , ein möglichst hohes niveau des verbraucherschutzes zu gewährleisten .
Baseline	the council would like to remind you that its policy has always been at the highest possible level of consumer protection .
4 PPT	the council would like to remind you that his policy always aims , as a high level of consumer protection .
Reference	the council wishes to point out that its policy is always to afford consumers the highest possible level of protection .

(h) Wrong pronoun chosen for translation

Figure 2: Example sentences for comparing our PPT extended model with the baseline. (2a) – (2f) reveal improvements, (2g) – (2h) show weaknesses.