Better Punctuation Prediction with Hierarchical Phrase-Based Translation

Stephan Peitz, Markus Freitag, Hermann Ney

Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

Punctuation prediction is an important task in spoken language translation and can be performed by using a monolingual phrase-based translation system to translate from unpunctuated to text with punctuation. However, a punctuation prediction system based on phrase-based translation is not able to capture long-range dependencies between words and punctuation marks. In this paper, we propose to employ hierarchical translation in place of phrase-based translation and show that this approach is more robust for unseen word sequences. Furthermore, we analyze different optimization criteria for tuning the scaling factors of a monolingual statistical machine translation system. In our experiments, we compare the new approach with other punctuation prediction methods and show improvements in terms of F_1-Score and BLEU on the IWSLT 2014 German→English and English→French translation tasks.

1. Introduction

Spoken language translation (SLT) has become an important application of automatic speech recognition (ASR) and machine translation (MT). The challenge of SLT is to translate automatic transcribed speech rather than written text into another language. In recent years, several research projects such as QUAERO[1] and EU-Bridge[2] have been focussed on speech translation. Furthmore, the increasing number of available Android application for speech translation[3] indicates a growing interest in speech translation technologies in the general public.

The translation of speech is in general divided in two independent parts. First, ASR provides a automatic transcription of spoken words. Next, the recognized words are translated by an MT system.

As in speech punctuation is not made explicit, most ASR systems provide an output without punctuation marks only. Most MT systems however are trained on data with proper punctuation and expect written text with correct punctuation as input. Therefore, the output of ASR systems has to be enriched with punctuation marks. In MT an accurate punctuation of the input is crucial as the prediction errors affect the translation quality. In [1], a loss of up to 4 BLEU points was obtained if punctuation marks need to be predicted, compared to correct punctuation in the input.

In recent years several methods to predict punctuation were developed. These methods are based on n-gram language models, on conditional random fields (CRF) or on monolingual statistical machine translation (SMT) systems translating from unpunctuated text to text with punctuation. One of the advantages of an SMT system or CRF is that more features beside the language model can be integrated. Furthermore, punctuation prediction can be done before, after or during the actual translation. Following [1, 2], we use a phrase-based SMT system for punctuation prediction before the actual translation as starting point.

In this work, we propose to employ hierarchical translation in place of phrase-based translation. In phrase-based translation, the translation units are bilingual phrases which are pairs formed by a sequence of source language words and its translation. Since a sequence of words can be translated at once, local contextual information is preserved. In the context of punctuation prediction, such information is useful to predict punctuation marks depending of its surrounding words, e.g. commas. However, this approach has its limitation for unseen word sequences and dependencies beyond the local context, e.g. the dependency between a question word and a question mark. If a sequence of words was not seen in the training data, the phrase-based translation system will fall back on shorter phrases with less local contextual information. Thus, more prediction errors can occur. To generalize better and to model dependencies as described above, we need a more abstract form of phrases. In hierarchical translation, such phrases are defined since discontinuous phrases with "gaps" are allowed. Those phrases capture long-range dependencies between words. In terms of punctuation prediction, we want to model dependencies between words and punctuation marks. In addition, by using more abstract phrases, a punctuation prediction system based on hi-

[1] http://www.quaero.org/
[2] http://www.eu-bridge.eu/
[3] https://play.google.com/store/search?q=speech%20translation&c=apps

erarchical translation models is more robust for unseen word sequences and generalize better.

As already mentioned, another advantage of using an SMT system for punctuation prediction is that different features besides the language model can be applied. These features are combined in a log-linear model. In this work, we investigate the impact of different optimization criteria for tuning the scaling factors of the features with minimum error rate training.

In our experiments on the IWSLT 2014 German→English and English→French machine translation task, we show improvements in terms of F_1-Score and BLEU.

This paper is structured as follows. We start in Section 2 with a short overview of the published research on punctuation prediction. In Section 3, we recap the idea of modeling punctuation prediction as machine translation and discuss different optimization criteria for tuning the scaling factors of a monolingual MT system. We present our approach using a hierarchical phrase-based translation system for punctuation prediction in Section 4. Finally, Section 5 describes the experimental results, followed by a conclusion in Section 6.

2. Related Work

In recent years, several approaches for predicting punctuation have been presented.

The HIDDEN-NGRAM tool from the SRI toolkit [3] considers punctuation marks as hidden events occurring between words. The most likely hidden tag sequence is found using an n-gram language model trained on punctuated text. In this work, we will compare with this tool.

The approach described in [4] is based on conditional random fields. They extended the linear-chain CRF model to a factorial CRF model using two layers with different sets of tags for punctuation marks respectively sentence types. They compared their approach with linear-chain CRF model and the HIDDEN-NGRAM tool on the IWSLT 2009 corpus. Besides the comparison of the translation quality in terms of BLEU, they also compared the CRF models with the hidden event language model regarding precision, recall and F_1-Score. Both in terms of BLEU and in terms of precision, recall and F_1-Score the CRF models outperformed the hidden event language model. They claimed that using non-independent and overlapping features of the discriminative model as machine translation instead of a language model only helped. Similar to this approach, using a statistical machine translation system for punctuation prediction has the advantage to integrate more features beside the language model.

Using MT for punctuation prediction was first described in [5]. In this work, a phrase-based statistical machine translation system was trained on a pseudo-bilingual corpus. The case-sensitive target language text with punctuation was considered as the target language and the text without case information and punctuation was used as source language. They applied this approach as postprocessing step in evaluation campaign of IWSLT 2007 and achieved a significant improvement over the baseline. In [6] the same approach was employed as preprocessing step and compared with the HIDDEN-NGRAM tool within the evaluation campaign of IWSLT 2008. The HIDDEN-NGRAM tool outperformed the MT-based punctuation prediction. In addition to punctuation prediction using a monolingual MT system, performing segmentation of ASR output was described in [2]. In all mentioned papers using a monolingual MT system for punctuation prediction, the optimization criterion for tuning the scaling factors of such a system was not described. In this work, we will tune both the phrase-based and the hierarchical translation system against BLEU and F_α-Score and analyze the impact on the prediction accuracy and translation quality.

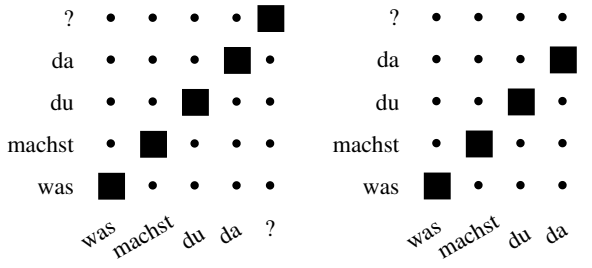
In [7], three different stages at which punctuation can be predicted are investigated: before, during and after the translation. Each of the stages requires a different translation system and has advantages and disadvantages. For predicting punctuation during the translation, additional punctuation prediction is not needed. The punctuation prediction before and after the translation was done with the HIDDEN-NGRAM tool. The implicit punctuation generation worked best on IWSLT 2006 corpus, but on TC-STAR 2006 corpus they achieved better results with punctuation prediction before and after the actual translation.

The impact of using a monolingual statistical machine translation system rather than the HIDDEN-NGRAM tool was analyzed in [1]. The authors report an improvement of 0.8 BLEU points by applying a monolingual statistical machine translation system before translation. An important advantage is that no modification of the actual translation system is needed. In our work, we follow this pipeline and replace the phrase-based translation model by a hierarchical translation model.

3. Modeling Punctuation Prediction as Machine Translation

Punctuation prediction using a statistical machine translation system is based on following pipeline. First, we extract the translation model for the SMT system from a pseudo-bilingual corpus. In order to create such a corpus, we need two versions of a monolingual corpus: one without punctuation (source text) and one with punctuation (target text). This is done by creating a monotone alignment (Figure 1(a)) and removing punctuation marks from the source sentences. The punctuation marks in the target sentences which are aligned with punctuation marks in the source sentences become unaligned (Figure 1(b)).

Given the pseudo-bilingual corpus and the modified alignment, we extract the translation model. In our work, we substitute the phrase-based translation model by a hierarchical translation model. Details about hierarchical translation are given in Section 4. In a next step, the scaling factors of



(a) Monotone alignment

(b) Punctuation marks in the target sentence become unaligned

Figure 1: Modification of the alignment

the monolingual translation system are tuned. We get a tuning set by removing the punctuation marks from a development set and use the original development set as reference. In this paper, we analyze different criteria used in the optimization of the scaling factors. We give further details in the following subsection.

3.1. Optimization Criteria

In most state-of-the-art SMT systems, MERT [8] is applied to optimize scaling factors of features using BLEU [9] as optimization criterion. However, the performances of systems predicting punctuation are measured and compared with the F_1-Score which is the harmonic mean of precision and recall. Thus, there is an inconsistency between optimization criterion and metric. Furthermore, the F_1-Score considers both precision and recall while BLEU is a metric which is based on n-gram precision and does not take recall into account. A criterion including recall is important because it ensures that the punctuation prediction system generates an appropriate amount of punctuation marks. In this work, we use F_α-Score as a more suitable optimization criterion. F_α-Score is a more general form of the F_1-Score, where α is a positive real number:

$$F_\alpha = (1+\alpha) \cdot \frac{(precision \cdot recall)}{\alpha \cdot precision + recall}$$

By varying the parameter α, more emphasis can be put on recall or precision. In this work, we will put more weight on recall and tune the systems with $\alpha = \{1,2,3,4\}$. We might lose precision and overgenerate punctuation marks, but this could be compensable for the actual translation system.

However, tuning a system on F_α-Score directly would not be practical as the positions of the punctuation marks would be ignored. For the optimization, we have to modify the F_α-Score and take the predecessors of the punctuation marks into account. In this work, we tune our monolingual translation systems using the modified F_α-Score as criterion with $\alpha = \{1,2,3,4\}$ and compare against systems tuned on BLEU.

4. Punctuation Prediction based on Hierarchical Translation

In hierarchical phrase-based translation [10], discontinuous phrases with "gaps" are allowed. The translation model is formalized as a synchronous context-free grammar (SCFG) and consists of bilingual rules, which are based on bilingual standard phrases and discontinuous phrases. Each bilingual rule rewrites a generic non-terminal X into a pair of strings $\tilde{f}$ and $\tilde{e}$ with both terminals and non-terminals in both languages

$$X \rightarrow \langle \tilde{f}, \tilde{e} \rangle.$$

Obtaining these rules is based on a heuristic extraction from automatically word-aligned bilingual training data. Just like in the phrase-based approach, all bilingual rules of a sentence pair are extracted given an alignment. The standard phrases are stored as *lexical rules* in the rule set. In addition, whenever a phrase contains a sub-phrase, this sub-phrase is replaced by a generic non-terminal X. With these hierarchical phrases we can define the *hierarchical rules* in the SCFG. The rule probabilities which are in general defined as relative frequencies are computed based on the joint counts $C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)$ of a bilingual rule $X \rightarrow \langle \tilde{f}, \tilde{e} \rangle$

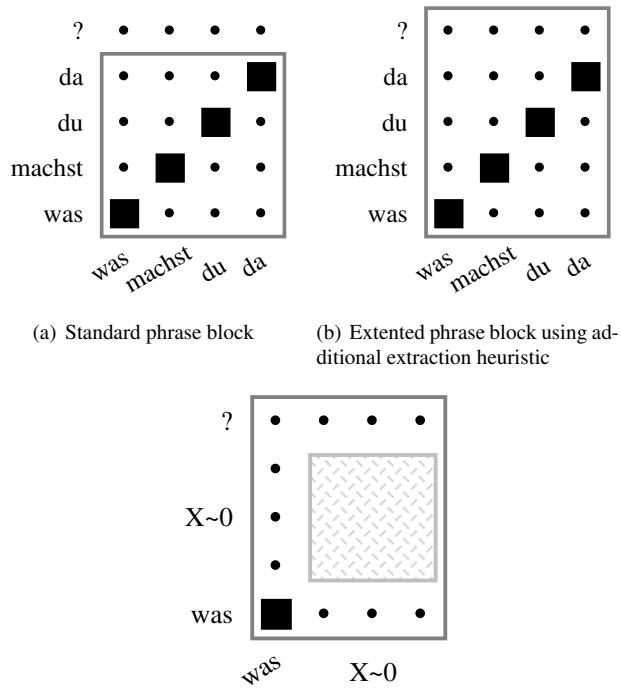
$$p(\tilde{f}|\tilde{e}) = \frac{C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)}.$$

The translation probabilities are computed in source-to-target as well as in target-to-source direction. In the translation processes, these probabilities are integrated in the log-linear combination among other models such as a language model, word lexicon models, word and phrase penalty and binary features marking hierarchical phrases, glue rule and rules with non-terminals at the boundaries.

The translation process of hierarchical phrase-based approach can be considered as a parsing problem. Given an input sentence in the source language, this sentence is parsed using the source language part of the SCFG. Using the associated target part of the applied rule, a translation can be constructed. The language model score is incorporated by employing the cube pruning algorithm presented in [11].

In a standard translation task, hierarchical rules with up to two non-terminals are extracted. Using rules with one non-terminal, the translation system is able to model long-range dependency between terminals. Furthermore, rules with two non-terminals make it possible to perform reordering without an additional model. In other words, the reordering is modelled in the hierarchical translation model implicitly. In case of punctuation prediction, we perform monotone translation and reordering is not necessary. Thus, we extract rules with one non-terminal maximum.

For punctuation prediction, our goal is to capture long-range dependencies between words and punctuation using hierarchical rules. To be able to extract such rules, we add an heuristic to the rule extraction process as described in the next section.



(a) Standard phrase block

(b) Extented phrase block using additional extraction heuristic

(c) Hierarchical rule extracted from extended phrase block

Figure 2: Extraction heuristic applied for initial phrase blocks

4.1. Additional Phrase Extraction Heuristic

As mentioned in the Section 3, punctuation marks in the target sentences which are aligned with punctuation marks in the source sentences become unaligned. Applying the standard phrase extraction procedure [12], phrases with punctuation are not extracted (Figure 2(a)). In order to add phrases such as

$$\langle \text{was machst du da}, \text{was machst du da ?} \rangle$$

to the translation model, we apply a heuristic which allows for phrase blocks including non-aligned words which are adjacent to phrase boundaries (Figure 2(b)). By using such additional phrases as initial phrases in the hierarchical extraction process (Figure 2(c)), we are able to extract hierarchical rules which model long-range dependencies between words and punctuation marks, e.g.

$$X \rightarrow \langle \text{was } X\tilde{}0, \text{was } X\tilde{}0 \text{ ?} \rangle,$$
$$X \rightarrow \langle \text{machst du } X\tilde{}0, \text{machst du } X\tilde{}0 \text{ ?} \rangle.$$

In the first rule, the question mark on the target side is related to the German question word "was". In the second rule, the typical German word order for questions (verb "machst" before subject "du") triggers a question mark on the target side. Both rules are more abstract since the gap could be filled with any other phrases during decoding. Even for unseen word sequences, e.g. "was machst du heute", these rules match. Thus, punctuation prediction based hierarchical translation can generalize better and improve the prediction accuracy.

In the experimental evaluation, we will analyze if such rules influence the decoding process and affect the punctuation prediction accuracy. Note, for the phrase-based translation system, we apply the non-aligned word heuristic as well.

Table 1: Data statistics for the preprocessed German-without-punctuation→German parallel in-domain training corpus used for punctuation prediction with the monolingual MT systems.

	German without Punct.	German
Sentences	171721	
Running words	2.7M	3.3M
Vocabulary	119242	119266

5. Experimental Evaluation

Our approach to use hierarchical phrase-based translation for punctuation prediction was evaluated on the IWSLT 2014 German→English and English→French machine translation tasks. IWSLT is an annual public evaluation campaign focusing on spoken language translation. The domain is lecture-type talks presented at TED conferences. The translation part of the evaluation campaign is divided into two different tracks: translation of automatic and translation of manual transcriptions. While the correct manual transcription contains punctuation marks, the automatic transcription did not. As we focus on punctuation prediction in this work, we used for the experiments *pseudo ASR output* rather than real ASR output as input. Pseudo ASR output is created by removing punctuation marks from the manual transcriptions. Thus, recognition errors do not occur and case information is preserved.

5.1. Punctuation Prediction

Both phrase-based and hierarchical translation models are trained on the provided in-domain training data (Table 1).

For tuning and testing our monolingual translation system, we used the provided manual transcribed development set and test sets (Table 2).

Training data as well as development and test set were modified as described in Section 3. A 5-gram language model, which was applied by both monolingual translation systems and the HIDDEN-NGRAM toolkit, was trained on the concatenation of the in-domain, europarl, news-commentary and commoncrawl corpora with the KenLM language model

Table 2: Data statistics for the preprocessed German-without-punctuation→German development and test sets for tuning and testing the punctuation prediction systems.

		German without Punct.	German
dev	Sentences	887	
	Running words	16521	19152
	Vocabulary	4029	4039
test	Sentences	1565	
	Running words	25483	30332
	Vocabulary	4976	4987

toolkit [13] using modified Kneser-Ney smoothing [14]. For creating the monolingual translation systems, we used an open-source translation tookit, which implements both phrase-based and hierarchical translation.

5.2. Bilingual Translation Systems

We set up translation systems for German→English and English→French to investigate the impact of better punctuation prediction on the translation quality in terms of BLEU and TER [15]. In order to analyze the effect of prediction errors on the translation quality, we compare with a setup with correct punctuation in the input. We employed phrase-based translation for both language pairs. Both systems were trained on all available bilingual and monolingual data provided by the IWSLT evaluation campaign.

5.3. Comparison of the Prediction Accuracy

The punctuation performance of our new approach using a hierarchical translation system (HPBT) is compared with a phrase-based translation system (PBT) and the HIDDEN-NGRAM tool. The accuracy is measured in precision (Prec.), recall (Rec.) and F_1-Score (F_1). Furthermore, we analyze the impact of different optimization criteria. Both translation systems were tuned on BLEU and F_α, where $\alpha = \{1,2,3,4\}$. Table 3 shows the result of this comparison for the German language.

The HPBT translation system tuned on F_2 performs best in terms of F_1-Score. For the PBT translation systems, tuning on F_2 leads to slightly better results. The performance of the HIDDEN-NGRAM toolkit is slightly better than the best PBT system. However, HIDDEN-NGRAM performs worse than the HPBT system tuned on F_2. In general, it seems tuning on F_α works better than tuning on BLEU. Although systems tuned on F_α tend to be less precise, the F_1-Score is higher compare to system tuned on BLEU. Best performance is achieved with $\alpha = 2$.

In the following, we define the PBT system tuned on BLEU as *baseline* and compare it against PBT tuned on F_2, HPBT tuned on F_2 and HIDDEN-NGRAM.

Table 3: Accuracy of the predicted punctuation on the test set of correct manual transcription without punctuation (German).

system	tuned on	Prec.	Rec.	F_1
PBT w/o heuristic	BLEU	86.7	24.4	43.9
PBT	BLEU	*82.7*	*67.5*	*74.3*
	F_1	82.6	67.5	74.3
	F_2	**78.3**	**71.4**	**74.7**
	F_3	76.6	72.2	74.4
	F_4	72.5	73.6	73.0
HPBT	BLEU	86.4	65.5	74.7
	F_1	81.8	71.0	76.0
	F_2	**77.0**	**75.4**	**76.2**
	F_3	75.9	75.2	75.6
	F_4	71.8	73.7	74.2
HIDDEN-NGRAM	-	**82.7**	**69.5**	**75.5**

Table 4: Three different classes of punctuation marks and their relative frequencies in the test set of the correct manual transcription.

Class	Punctuation marks	rel. freq.
1	. ? !	40,2%
2	,	53,3%
3	" ' ;) (	6,5%

To get further insight, on which level type of annotation the prediction methods are more or less accurate, we measure the accuracy regarding three different classes of punctuation marks (Table 4).

The result of this comparison is given in Table 5. In all three classes, HPBT outperforms both HIDDEN-NGRAM and PBT in terms of F_1-Score. However, the largest difference in accuracy is obtained in class 3.

Next, we investigate the usage of hierarchical rules in the decoding process and analyze whether such rules help to increase the prediction accuracy. This is done by counting the lexical and hierarchical rules applied during decoding. In particular, we count rules which introduce punctuation marks and compute the average target length of the applied rules. In this analysis, we compare PBT and HPBT (Table 6). While the PBT system uses short phrases with a limited local context (average target phrase length of 2.1 or 2.0), the HPBT system employs both lexical and hierarchical rules to insert punctuation marks. Even if only 20% of the rules which introduce punctuation marks are hierarchical, it seems that those rules help to improve the prediction accuracy.

We further examine these results using a prediction example (Table 7). In this example, both HIDDEN-NGRAM and PBT did not predict the question mark. However, HPBT is

Table 5: Accuracy of the predicted punctuation on the test of correct manual transcription without punctuation regarding three different classes of punctuation marks (German).

system	tuned on	class 1			class 2			class 3		
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
PBT	BLEU	88.4	73.9	80.5	84.8	75.9	80.1	56.7	21.3	30.9
PBT	F_2	88.7	73.8	80.5	78.5	83.5	80.9	58.8	27.1	37.1
HPBT	F_2	88.4	81.6	84.9	77.7	85.2	81.3	49.5	30.8	38.0
HIDDEN-NGRAM	-	87.8	81.6	84.6	84.8	75.1	79.7	39.9	12.6	19.1

Table 6: Comparison of the numbers of applied phrases introducing punctuation marks. Average target phrase length is given in parentheses.

system	tuned on	lexical rules	hierarchical rules
PBT	BLEU	2313 (2.1)	-
PBT	F_2	2549 (2.0)	-
HPBT	F_2	2234 (2.6)	442 (3.9)

Table 7: Examples for punctuation prediction on German pseudo ASR output using different prediction approaches.

system	tuned on	
pseudo ASR output		was machst du nur
PBT	BLEU	was machst du nur .
PBT	F_2	was machst du nur .
HPBT	F_2	was machst du nur ?
HIDDEN-NGRAM	-	was machst du nur .
correct punctuation		was machst du nur ?

able to produce the correct sentence end punctuation.

In this example, the word order "machst du" indicates that this sentence is a question. Using the HPBT system, the correct sentence end mark is introduced by applying following hierarchical rule:

$$X \rightarrow \langle \text{machst du } X^{\sim}0, \text{machst du } X^{\sim}0 \text{ ?} \rangle.$$

In this rule, a long-range dependency between the words "machst du" and the punctuation mark "?" exists. However, such a dependency is not modelled in the phrase-based translation system. The question mark can only be inserted by the phrase

$$\langle \text{nur}, \text{nur ?} \rangle.$$

A phrase with local contextual information about the word order, e.g.

$$\langle \text{machst du nur}, \text{machst du nur ?} \rangle,$$

was not seen in the extraction process and is not part of this translation model. Thus, the PBT system uses shorter phrases with less contextual information and it is more likely that a phrase producing an erroneous punctuation is used. Here, the following phrase was applied:

$$\langle \text{nur}, \text{nur .} \rangle.$$

In this example, it seems that the hierarchical system generalize better for unseen word sequences. Furthermore, the analysis shows that hierarchical rules influence the decoding process and help to improve the prediction accuracy in our experiments.

5.4. Comparison of the Translation Quality

In the introduction, we have mentioned the effect of punctuation errors on the translation quality. In the next experiments, we check whether a higher prediction accuracy results in an improvement of the translation quality in terms of BLEU and TER. We performed punctuation prediction with different setups and then translated the enriched pseudo ASR output. The translation was performed with the bilingual SMT systems described above. The result of this comparison is shown in Table 8. We lose up to 2.2 points in BLEU if punctuation marks need to be predicted. It seems that a higher prediction accuracy leads to a higher translation quality. The performance of HIDDEN-NGRAM and PBT tuned on BLEU is on the same level. By replacing the optimization criterion with F_2, we gain 0.2 points in BLEU. Using a hierarchical system improves the translation quality by additional 0.2 points in BLEU. TER is on the same level for all setups.

To verify our improvements, we carried out additional experiments on the English→French translation task (Table 9). In this setup, we performed punctuation prediction both on pseudo ASR output and real ASR output. In terms of F_1-Score, HPBT outperforms both HIDDEN-NGRAM and PBT. However, on the real ASR output test set the performance of HPBT and PBT is on a same level in terms of translation quality.

Table 8: Impact of accuracy of punctuation prediction on the translation quality (German→English). Comparison with correct punctuation in the input.

system	tuned on	Prec.	Rec.	F_1	BLEU	TER
PBT	BLEU	82.7	67.5	74.3	27.3	53.3
PBT	F_2	78.3	71.4	74.7	27.5	53.4
HPBT	F_2	**77.0**	**75.4**	**76.2**	**27.7**	**53.2**
HIDDEN-NGRAM	-	82.7	69.5	75.5	27.2	53.2
correct punctuation					29.4	51.3

Table 9: Accuracy of the predicted punctuation on the test set of automatic (ASR) and correct manual transcription without punctuation (pseudo ASR) (English→French).

					pseudo ASR		ASR	
system	tuned on	Prec.	Rec.	F_1	BLEU	TER	BLEU	TER
PBT	BLEU	81.2	67.6	73.7	28.4	54.5	22.6	62.8
PBT	F_2	72.2	75.0	73.6	28.6	55.2	22.8	63.2
HPBT	F_2	**74.8**	**77.1**	**75.9**	**28.9**	**54.7**	22.7	62.7
HIDDEN-NGRAM	-	82.0	60.2	69.4	27.0	55.4	21.7	62.6
correct punctuation					31.9	50.1	-	-

6. Conclusion

In this paper, we introduced a new approach to predict punctuation with a monolingual hierarchical translation system. While phrase-based translation is limited to local context information, we are able to model long-range dependencies between words and punctuation marks by using hierarchical translation. In our experimental evaluation, we showed that our method improves the prediction accuracy and translation quality in terms of BLEU on the IWSLT German→English and English→French translation tasks. Furthermore, tuning a monolingual translation system for predicting punctuation on F_2 rather than BLEU improves the accuracy and translation quality.

In future work, we would like go to beyond the phrase level and investigate features which are operating on sentence level. In this way, quotes or parentheses could be modelled more accurate.

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

[1] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling Punctuation Prediction as Machine Translation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[2] E. Cho, J. Niehues, and A. Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system." in *IWSLT*, 2012, pp. 252–259.

[3] A. Stolcke, "Srilman extensible language modeling toolkit," in *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 901–904.

[4] W. Lu and H. T. Ng, "Better punctuation prediction with dynamic conditional random fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 177–186.

[5] H. Hassan, Y. Ma, and A. Way, "Matrex: the dcu machine translation system for iwslt 2007," in *Proceedings of the International Workshop on Spoken Language Translation 2007*, Trento, Italy, 2007.

[6] Y. Ma, J. Tinsley, H. Hassan, J. Du, and A. Way, "Exploiting Alignment Techniques in MaTrEx: the DCU Machine Translation System for IWSLT08," in *Proc. of the International Workshop on Spoken Language Translation*, Hawaii, USA, 2008, pp. 26–33.

[7] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *International Workshop on Spoken Language Translation*, Kyoto, Japan, Nov. 2006, pp. 158–165.

[8] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," Sapporo, Japan, July 2003, pp. 160–167.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[10] D. Chiang, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," Ann Arbor, Michigan, June 2005, pp. 263–270.

[11] Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[12] F. J. Och, C. Tillmann, H. Ney, *et al.*, "Improved alignment models for statistical machine translation," *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, 1999.

[13] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696. [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf

[14] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Computer Science Group, Harvard University, Cambridge, Massachusetts, USA, Tech. Rep. TR-10-98, Aug. 1998.

[15] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.