Modeling Punctuation Prediction as Machine Translation

Stephan Peitz, Markus Freitag, Arne Mauser,
Hermann Ney

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

Abstract

Punctuation prediction is an important task in Spoken Language Translation. The output of speech recognition systems does not typically contain punctuation marks. In this paper we analyze different methods for punctuation prediction and show improvements in the quality of the final translation output. In our experiments we compare the different approaches and show improvements of up to 0.8 BLEU points on the IWSLT 2011 English French Speech Translation of Talks task using a translation system to translate from unpunctuated to punctuated text instead of a language model based punctuation prediction method. Furthermore, we do a system combination of the hypotheses of all our different approaches and get an additional improvement of 0.4 points in BLEU.

1. Introduction

Spoken language translation (SLT) is an important application of automatic speech recognition (ASR) and machine translation (MT). It takes one of the most natural forms of human communication – speech – as input and makes it accessible to speakers of another language. In recent years, large research projects have been focussed on speech translation such as the European TC-Star project where the focus was on speech-to-speech translation of speeches in the European parliament and the DARPA-funded GALE project where TV-shows and broadcast news were translated from Arabic and Chinese to English for intelligence purposes. Even applications for mobile phones are now available to a broader audiences. For example, the Google Translate application for Android mobile phones has been downloaded more than 10 million times already [1], a number clearly stating the rising acceptance of speech translation technologies in the general public.

The speech translation process is typically divided in two distinct parts. First, there is automatic speech recognition, that provides a transcription of the spoken words. The second part is the translation of the recognized words by the machine translation system.

Almost all state-of-the-art ASR systems recognize sequences of words, but do not provide punctuation marks. In regular spoken text, punctuation marks are not made explicit. Humans can often infer punctuation by prosodic, syntactic or semantic cues, but the tasks is difficult for more casual and conversational speech as grammatic rules are often only loosely observed. This makes the evaluation of speech recognition results with punctuation more ambiguous and therefore recognitions systems tend to be optimized for output without punctuation.

Most MT system however are trained on text data with proper punctuation marks. Therefore, for an acceptable level of quality in translation, the systems expect correctly punctuated texts. Furthermore, the output of MT is expected to have correct punctuation in the output, even when translating from speech. Therefore, punctuation prediction has to be done at some stage of the speech translation process.

In this work, we will model punctuation prediction as machine translation and compare the different stages at which the prediction is done. The method is compared to the most common existing methods and evaluated by the accuracy of the predicted punctuation as well as by the quality of the final translation output.

The paper is organized as follows. In Section 2, a short overview of the published research on punctuation prediction is given. In Section 3, we recapitulate different approaches for punctuation prediction. We present our approach using a statistical phrase-based machine translation system in Section 4, followed by Section 5 describing the system combination. Finally, Section 6 describes the experimental results, followed by a conclusion.

2. Related Work

This paper is based on the work of [1]. Amongst others they presented three different approaches to restore punctuation in already segmented ASR output. In addition to implicit punctuation generation in the translation process, punc-

[1] https://market.android.com/details?id=com.google.android.apps.translate&hl=en

tuation was predicted as pre- and postprocessing step. For punctuation prediction they used the HIDDEN-NGRAM tool from the SRI toolkit [2]. The implicit punctuation generation worked best on IWSLT 2006 corpus, but on TC-STAR 2006 corpus they achieved better results with punctuation prediction on source and target. They pointed out that on small corpora like IWSLT 2006 falsely inserted punctuation marks in the source side deteriorated the performance of the translation system. However, the IWSLT corpus became larger in the last years and therefore we verify the results within IWSLT 2011 SLT task. Furthermore, we use in addition for the punctuation prediction a phrase-based statistical machine translation system.

Using MT for punctuation prediction was first described in [3]. 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 [4] 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. Moreover, they achieved further improvements by combining these two methods using a majority voting procedure. In our work, we further investigate this approach and compare it with the HIDDEN-NGRAM tool at different stages at which the prediction is done. In our analysis we consider translation quality at the end of the translation pipeline as well as the accuracy of the punctuation prediction. In contrast to the majority vote, we do a system combination of the hypotheses of all different approaches.

The approach described in [5] is based on conditional random fields (CRF). 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 novel 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 F1-measure. Both in terms of BLEU and in terms of precision, recall and F1-measure 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 phrase-based machine translation system for punctuation prediction has the advantage to integrate more features beside the language model.

3. Punctuation Prediction Strategies

As mentioned in Section 1, there are three stages at which punctuation can be predicted: before, during, and after translation. Each of the stages requires a different translation system.

- When we predict the punctuation before translation, a regular text translation system can be used, that expects correctly punctuated input.
- To predict punctuation in the translation process, we need a translation system without punctuation marks in the source language, but with punctuation marks in target language.
- Punctuation prediction in the target language requires a translation system training without any punctuation.

The different approaches are visualized in Figure 1.

For all approaches, we assume that the segmentation of the speech recognition output is given and corresponds to at least sentence-like units. The level of annotation can vary from predicting only sentence-end punctuation marks such as full stops and questions marks, or a richer annotation that also contains commas and more challenging types of punctuation such as parentheses and quotation marks. In this work, we consider all kind of punctuation.

In the following subsections, we will describe each prediction strategy and the consequences on the translation pipeline in more detail.

3.1. Prediction in the Source Language

Predicting punctuation in the source language means, that the output from the speech recognition system, that does not contain any punctuation marks is augmented with punctuation using automatic methods. The main advantage of this methods is that no modification to the training data or the translation system are required and a standard text translation system can be used.

In order to provide a good input for the translation system and to get a good translation, the punctuation prediction has to be accurate as possible. Errors in the predicted punctuation can affect the translation system output quality. The main reason for this is that longer phrase or rule matches in the translation system are prevented by incorrectly inserted punctuation marks. Therefore, for this approach, the accuracy of the prediction is crucial for the final translation system performance.

The prediction performance itself is influenced by the error rate of the speech recognition system. Recognition errors, that already by themselves are impairing the translation quality may lead to additional punctuation prediction errors that further increase translation error rate.

Nevertheless the approach of predicting punctuation marks in the source language remains attractive because of

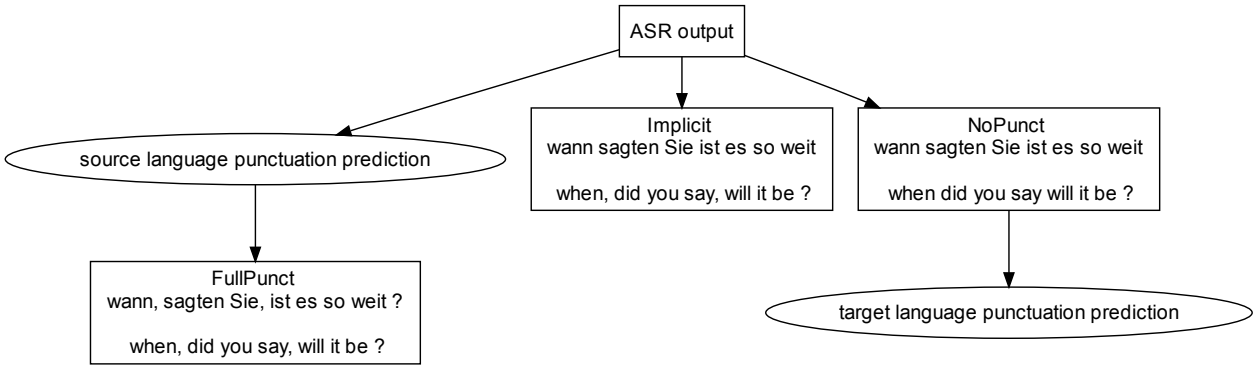

Figure 1: The three different stages where punctuation prediction can be done.

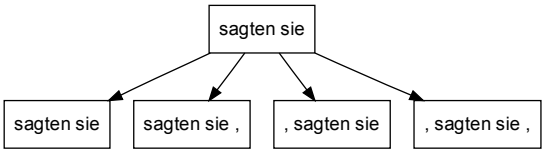

Figure 2: The German source phrase "sagten sie" has four possible translation options in the IMPLICIT system

the above stated simplicity in retaining the universal translation system. In the remainder of this work, we will refer to this translation system as FULLPUNCT.

3.2. Implicit Prediction

Natural languages have a variety of different punctuation systems and rules. Typically, the punctuation in one language cannot always directly be taken over into another language. As translation systems learn from real data, they also implicitly learn to handle the different punctuation styles in source and target language.

The implicit punctuation prediction approach proposed by [1], takes this idea and assume that the source language as produced by the speech recognition system does not contain any punctuation marks and the target language uses regular, full punctuation. We will refer to this system as IMPLICIT.

The training data for this machine translation system is preprocessed by removing all punctuation marks from the source language data, while the target language data is kept untouched. The removal is done after the word alignment step. The punctuation marks in the target sentence which are aligned with punctuation marks in the source sentences become non-aligned. In the phrase extraction two different phrase pairs for the same source side are extracted: One containing the punctuation mark and one without punctuation. In Figure 2 the German source phrase "sagten sie" has four possible translation options.

For IMPLICIT, we can still use the same language model as for the regular FULLPUNCT system. However, we need to change the translation model by re-extracting phrase and word lexicon models. Another disadvantage of this method is, that prediction and translation are not separate components in the speech translation pipeline. This makes systematic translation errors harder to track down, and separate optimization and execution of the components impossible.

3.3. Prediction in the Target Language

The prediction in the target language is done after translation. The translation system needed for this method does not have any punctuation in source and target language. All punctuation marks are removed from the training data as well as from the development and test sets. This results in the machine translation system NOPUNCT. After the translation process, the punctuation marks have to be inserted in the same way as when predicting punctuation marks in the source language.

This method has two major disadvantages. First, in addition to the translation model also the target language model used in translation has to be rebuilt. The second and more severe disadvantage for the final system performance is, that the translation produces errors, that make the punctuation prediction less accurate. This includes errors resulting from incorrect speech recognition that are propagated through the translation system as well as errors introduced in the translation process itself. As translation error tends to be higher than speech recognition error, accurate prediction of punctuation marks in the target language is conceptually more error-prone than the other methods presented above.

4. Punctuation Prediction with Statistical Machine Translation

Inspired by [1], we use a phrase-based machine translation system to predict punctuation marks. Instead of using a bilingual translation system, where source language and target language are different natural languages, we use a system where we translate into a natural language with proper punctuation from the same natural language without punctuation.

The motivation for this procedure is that additional models of the translation system and the possibility to automatically tune the system for good performance with respect to an

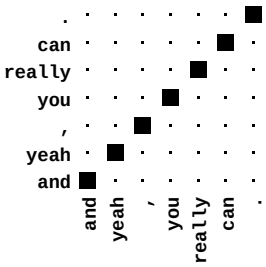

Figure 3: monoton alignment with punctuation marks

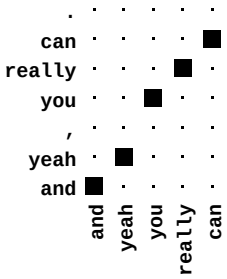

Figure 4: punctuation marks in the source sentence become non-aligned

evaluation measure help to predict punctuation correctly. In contrast, the HIDDEN-NGRAM is based on a language model only, but a phrase-based MT system is able to use further features e.g. the phrase translation probabilities.

We train our system monolingual using the source language training corpus. In order to train the system, we create two versions of the source language text: one without punctuation and one with punctuation. For phrase extraction, the alignment is assumed to be monotone. Similar to the implicit prediction approach, the punctuation marks in the target sentence which are aligned with punctuation marks in the source sentences become non-aligned. In Figure 3 and Figure 4 is one example for deleting the punctuation marks in the source sentence. Now, we are able to train a monolingual MT system for unpunctuated to punctuated text.

The tuning set for the parameter tuning is constructed by removing the punctuation marks from the regular development set source text. As reference we use the original source text with the punctuation left intact.

The phrase-based MT system used in this work for the punctuation prediction is an in-house implementation of the state-of-the-art MT decoder described in [6]. We use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, an 9-gram source language model and three binary count features. Due to the fact, that we use a monotone alignment, the reordering model is dropped. We also allow longer phrases to capture punctuation dependencies. The optimization is done with standard MERT [7] on 200-best lists with

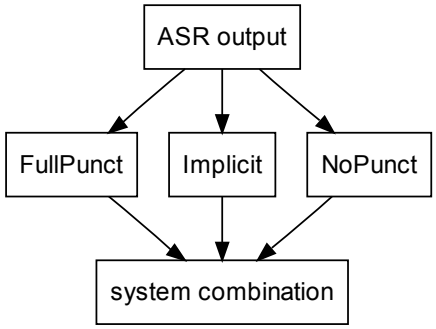

Figure 5: System combination of the translation result coming from different punctuation prediction methods.

BLEU as optimization criterion. 200-best lists are chosen to get more different hypotheses.

5. System Combination

System combination is used to produce consensus translations from multiple translation hypotheses generated with different systems. We follow an approach similar to the one described in [8, 9]. The basic procedure is, that hypotheses from different translations systems are aligned on the word level to find corresponding parts. Based on these alignments, a weighted majority voting on aligned words and additional models are used to produce the consensus translation.

In the scope of this work, we will combine translation output from multiple punctuation prediction schemes. Figure 5 shows the basic idea how to use system combination in this task.

6. Experimental Evaluation

The methods presented in this paper were evaluated on the IWSLT 2011 English-to-French translation track [10]. IWSLT is an annual public evaluation campaign focused on speech translation. The domain of the 2011 translation task is lecture-type talks presented at TED conferences which are also available online[2]. Two different conditions were evaluated: Automatic and manual transcription of lectures. While the correct manual transcription also contained punctuation marks, the automatic transcription did not. The automatic transcription used in this work was the 1-best hypothesis from the speech recognition system. The in-domain training data (Table 1) also consisted of transcribed TED lectures as well as news commentaries and transcribed speeches from the European Parliament. For system tuning and testing, we use the provided 1-best ASR output as development set and test set (Table 2).

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

Table 2: Data statistics for the preprocessed English-French development and test sets used in the MT and SLT track. In the sets, numerical quantities have been replaced by a special category symbol.

		MT		SLT	
		English	French	English	French
dev	Sentences	934			
	Running words	20131	20280	17735	20280
	without Punct. Marks	17795			
	Vocabulary	3209	3717	3132	3717
test	Sentences	1664			
	Running words	31975	33814	27427	33814
	without Punct. Marks	27653			
	Vocabulary	3711	4678	3670	4678

Table 1: Data statistics for the preprocessed English-French parallel training corpus used in the MT track. In the corpus, numerical quantities have been replaced by a special category symbol.

	English	French
Sentences	2.0M	
Running words	54.3M	59.9M
without Punct. Marks	48.9M	
Vocabulary	136K	159K

Table 3: Data statistics for the preprocessed English without punctuation marks - English parallel training corpus used for punctuation prediction with the phrase-based MT system. In the corpus, numerical quantities have been replaced by a special category symbol.

	English without Punct.	English	French without Punct.	French
Sentences	107075			
Running words	1.8M	2.1M	1.9M	2.2M
Vocabulary	43554	43576	55640	55663

6.1. Punctuation Prediction Model

The 9-gram language model is trained on the given training corpus (Table 1). The phrase extraction is done on the English-without-punctuation to English corpus (Table 3) respectively on the French-without-punctuation to French corpus (Table 3). We use a modified development set as described in Section 4. We remove the punctuation of the development and test sets which are available in the MT task of IWSLT 2011 (Table 2).

6.2. Hierarchical phrase-based decoder for translation

The following MT system is given to compare all punctuation prediction strategies. We use the open source hierarchical phrase-based system Jane [11], which implements the hierarchical approach as introduced by [12]. The search is carried out using the cube pruning algorithm [13]. The models integrated into our Jane systems are: phrase translation probabilities in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four binary count features and an 4-gram language model. For a robust baseline we add a sparse discriminative word lexicon (DWL) model for lexical smoothing and triplets similar to [14]. The model weights are optimized with standard MERT [7] on 100-best lists.

6.3. Comparison of the punctuation prediction accuracy

To assess and compare the punctuation prediction performance of the approaches presented in Section 3, we remove all punctuation from test set of the correct manual transcription, and restore the punctuation marks with the HIDDEN-NGRAM as well as with our phrase-based decoder for punctuation prediction (PPMT). We use the original test set as reference. We verify the methods before the translation, because the translation process causes too many errors for measuring the accuracy of the prediction meaningful.

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:

- Class 1 contains sentence-end punctuation marks `.`,`?` and `!`,
- Class 1.1 contains `.`,
- Class 1.2 contains `?`,
- Class 2 contains only commas, and

Table 4: Punctuation statistics for test set of the correct manual transcription and training corpus.

	training		test	
	rel. freq.	abs. freq.	rel. freq.	abs. freq.
class 1	40.3 %	2078101	41.8 %	1808
class 1.1	96 %	1995420	91.7 %	1658
class 1.2	3.4 %	70533	8.2 %	148
class 2	50.8 %	2621911	50.8 %	2194
class 3	8.9 %	461098	7.4 %	320
all	100 %	5161110	100 %	4322

- Class 3 holds all the remaining punctuation marks such as ", ', ;,), and).

Table 4 lists the absolute and relative frequency for each class regarding the test and training data.

The accuracy is measured in precision (Prec.), recall (Rec.), and F-measure (F_1). The reference is the original test set of the correct manual transcription. Table 5 shows the results of this comparison.

The results in class 1 are quite similar, but the prediction of commas is more accurate when using PPMT. However, the greatest difference in accuracy is obtained in class 3 indicating that PPMT predicts more challenging types of punctuation better e.g. shown in Table 6. When we consider all punctuation, the precision of the prediction with the HIDDEN-NGRAM is slightly higher then PPMT. However, the recall of the prediction with the PPMT is better and this results in a higher F-measure.

6.4. Comparison of the translation quality

While a comparison of the punctuation prediction performance might be a good indicator of the overall accuracy of the method, we ultimately want to improve the quality of the translation output. In order to compare the different strategies, we measure the translation quality of all systems in BLEU [15] and TER [16]. BLEU measures the accuracy of the translation, so higher values in BLEU are better. TER is an error measure, with lower values indicating better quality.

We built five different experimental setups with regards to the description in Subsection 6.2. To compare our new method, we use the HIDDEN-NGRAM tool with the same language model as applied in our phrase-based decoder for punctuation prediction. Thus, we get two systems for FULLPUNCT and two systems for NOPUNCT. The fifth system is IMPLICIT. Table 7 shows the comparison between the different translation system and both prediction tools.

FULLPUNCT with PPMT performs slightly better then IMPLICIT with 0.1 points in BLEU and 0.3 points in TER. However, the prediction with a phrase-based MT system outperforms both system using the HIDDEN-NGRAM tool as expected in Subsection 6.3. Using the FULLPUNCT system with PPMT, we get an improvement of 0.8 BLEU points and 0.7 TER points compared to FULLPUNCT using the HIDDEN-NGRAM tool. Similar improvement is obtained using the NOPUNCT systems. However, punctuation prediction in the source language leads to a better translation quality in terms of BLEU. We achieve an improvement of 0.7 BLEU points using FULLPUNCT instead of NOPUNCT.

An example of a prediction with the HIDDEN-NGRAM tool and our approach is given in Table 8. While the HIDDEN-NGRAM tool predicts the sentence end mark only, the phrase-based MT system has learned, that the phrase *right* within a text can be interpreted as question, which has to be separated by a comma from the sentence.

With the system combination of all five system, we get an additional improvement of 0.4 points in BLEU and 0.6 points in TER compared to the best single system FULLPUNCT with PPMT.

6.5. Punctuation Prediction with correct transcriptions

To analyze the impact of the errors of the ASR recognition on the translation quality, we take the data from the MT task and delete all punctuation to create a pseudo ASR output without any recognition errors. We do punctuation prediction with all described methods on our pseudo ASR output and compare it to the correct punctuation of the MT task. As you can see in Table 9, the correct punctuation has a higher score of 3.9 points in BLEU and 4.0 points in TER than our best punctuation prediction system. Furthermore, using the ASR output instead of the pseudo ASR output degrades our output 4.7 points in BLEU and 6.6 points in TER. Not only the ASR recognition but also the punctuation prediction needs more effort to improve the translation of spoken language.

7. Conclusion

In this paper, we compared different approaches for predicting punctuation in a speech translation setting. In contrast to [4] the translation-based punctuation prediction outperformed the language model based approach as well as implicit method in terms of BLEU and TER on the IWSLT 2011 SLT task.

The main advantage of modeling punctuation prediction as machine translation is that the translation system used in speech translations does not require special preprocessing for the task. Moreover, with the help of the RWTH system combination approach, we get an improvement in BLEU and TER compared to the best single system.

In future work, we would like to investigate special features for modelling parentheses or quotes. One advantage of using a MT system to predict punctuation is that additional model components and features can be added easily. Furthermore, the different optimization criteria e.g. F-measure or WER should also be analyzed.

Table 5: Accuracy of the predicted punctuation on the test of correct manual transcription without punctuation.

tool	**class 1**			**class 1.1**			**class 1.2**		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
HIDDEN-NGRAM	87.9	85.0	86.4	88.9	90.7	89.8	59.7	23.0	33.2
PPMT	88.2	81.7	84.8	89.0	87.5	88.2	63.4	17.6	27.5

tool	**class 2**			**class 3**			**all punct.**		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
HIDDEN-NGRAM	83.5	44.8	58.3	18.3	6.7	9.8	81.5	57.3	67.3
PPMT	80.6	59.3	68.3	47.2	22.7	30.7	80.7	64.2	71.5

Table 6: Example for prediction on pseudo ASR output.

system	**tool**	
pseudo ASR output	-	they say The plants talk to us
reference	-	they say , “ The plants talk to us . ”
FULLPUNCT	HIDDEN-NGRAM	they say The plants , talk to us .
FULLPUNCT	PPMT	they say , “ The plants talk to us .

Table 7: Results for the SLT tasks English-French, including used tool for punctuation prediction.

system	**tool**	**dev**		**test**	
		BLEU [%]	TER [%]	BLEU [%]	TER [%]
IMPLICIT	-	18.0	69.5	21.8	62.5
FULLPUNCT	HIDDEN-NGRAM	18.2	69.3	21.1	62.9
FULLPUNCT	PPMT	18.3	69.2	21.9	62.2
NOPUNCT	HIDDEN-NGRAM	17.3	67.9	20.4	62.8
NOPUNCT	PPMT	17.8	69.0	21.2	62.2
system combination		18.5	68.3	22.3	61.6

Table 8: Prediction example.

system	**tool**	
ASR output	-	but that ’s not really what this is about right and I open my hand
reference	-	but that ’s not really what this is about . right ? and then I open my hand up .
FULLPUNCT	HIDDEN-NGRAM	but that ’s not really what this is about right and I open my hand .
FULLPUNCT	PPMT	but that ’s not really what this is about , right ? and I open my hand .

8. Acknowledgements

This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

9. References

[1] 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.

[2] 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.

[3] H. Hassan, Y. Ma, and A. Way, “Matrex: the dcu ma-

Table 9: Results for the MT task English-French.

system	tool	dev		test	
		BLEU[%]	TER[%]	BLEU[%]	TER[%]
correct punctuation	-	27.5	57.0	30.8	50.9
IMPLICIT	-	24.0	61.7	26.6	55.9
FULLPUNCT	HIDDEN-NGRAM	24.4	60.6	26.6	55.1
FULLPUNCT	PPMT	24.4	60.5	27.0	55.0
NOPUNCT	HIDDEN-NGRAM	22.2	61.2	25.0	55.9
NOPUNCT	PPMT	22.9	62.3	26.0	56.2

chine translation system for iwslt 2007," in *Proceedings of the International Workshop on Spoken Language Translation 2007*, Trento, Italy, 2007.

[4] 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.

[5] 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.

[6] R. Zens and H. Ney, "Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 195–205.

[7] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[8] E. Matusov, N. Ueffing, and H. Ney, "Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006, pp. 33–40.

[9] E. Matusov, G. Leusch, R. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System Combination for Machine Translation of Spoken and Written Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1222–1237, 2008.

[10] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the iwslt 2011 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, 2011.

[11] D. Stein, D. Vilar, S. Peitz, M. Freitag, M. Huck, and H. Ney, "A guide to jane, an open source hierarchical translation toolkit," *The Prague Bulletin of Mathematical Linguistics*, no. 95, pp. 5–18, Apr. 2011.

[12] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[13] L. Huang and D. Chiang, "Forest Rescoring: Faster Decoding with Integrated Language Models," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 144–151.

[14] M. Huck, M. Ratajczak, P. Lehnen, and H. Ney, "A comparison of various types of extended lexicon models for statistical machine translation," in *Conference of the Association for Machine Translation in the Americas 2010*, Denver, Colorado, USA, Oct. 2010.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, IBM Research Report RC22176 (W0109-022), Sept. 2001.

[16] 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.