# Annotating Data Selection for Improving Machine Translation

*Keiji Yasuda, Hideo Okuma, Masao Utiyama and Eiichiro Sumita*

National Institute of Information and Communications Technology
keiji.yasuda,hideo.okuma,mutiyama,eiichiro.sumita@nict.go.jp

## Abstract

In order to efficiently improve machine translation systems, we propose a method which selects data to be annotated (manually translated) from speech-to-speech translation field data. For the selection experiments, we used data from field experiments conducted during the 2009 fiscal year in five areas of Japan. For the selection experiments, we used data sets from two areas: one data set giving the lowest baseline speech translation performance for its test set, and another data set giving the highest. In the experiments, we compare two methods for selecting data to be manually translated from the field data. Both of them use source side language models for data selection, but in different manners. According to the experimental results, either or both of the methods show larger improvements compared to a random data selection.

## 1. Introduction

As a result of the drastic technical innovation advances in spoken language processing, speech-to-speech translation systems are now starting to be used in actual fields [1, 2]. In addition to the development of basic technologies, the efficient usage of field data is also an important challenge to be addressed for system performance improvement.

A speech-to-speech translation system consists mainly of three subsystems: an automatic speech recognition (ASR) subsystem, a machine translation (MT) subsystem and a speech synthesis subsystem. While the simplest and most effective usage of field data is to annotate (transcribe and translate) all of the field data and use this for ASR and MT training, annotation is expensive and time consuming.

In this paper, we propose a method for selecting useful field data to be manually translated in sentence units. In previous studies on ASR [3], positive results were obtained by selectively annotating (transcribing) field data. There have also been many studies on domain adaptation research handling data selection [1, 4, 5, 6] in ASR and MT researches, However, there has been little research done from a data annotation[1] point of view.

Typical MT domain adaptation research handles the selection of productive training sentences from out-of-domain monolingual or parallel corpus. In the task setting, we can use source and target language information to select training sentences from the parallel corpora. However, only source language information is available in our annotation data selection task. In this paper, we propose two methods that use source side language models. These methods use the source-side language model in the different manners to select productive field data for manual translation.

Section 2 introduces the field experiments that the data used for this paper originates from. Section 3 explains the proposed selection method. Section 4 details the selection experiments using the field experiment data. Section 5 concludes the paper.

[1] Different from ASR, annotation means manual translation here.

## 2. Field Experiments

The data sets used in this paper are extracted from the user log data of speech-to-speech translation field experiments done in fiscal year 2009 [2]. In this section, we explain the field experiments.

### 2.1. Outline

As shown in Fig. 1, the field experiments were carried out in five areas across Japan. The total budget for the five projects was 985 million yen, which was funded by the Ministry of Internal Affairs and Communications.

The two main purposes for the field experiments were:

1. To improve speech to speech translation technology by using the user log data.

2. To promote the early realization of actual speech to speech translation services to make things easier for visitors to Japan.

Each project was lead by either different private companies or by joint public-private ventures. A speech-to-speech translation engine was provided by the National Institute of Information and Communications Technology (NICT). This engine can translate in both directions for Japanese-English, Japanese-Chinese and Japanese-Korean. Using this engine, each project developed their own speech-to-speech translation interfaces operated on PCs or mobile devices such as smartphones. In addition to the speech-to-speech translation system, each project also developed travel-related applications, such as a GPS navigation system and a travel informa-

tion system. A schematic diagram of the system configuration is shown Fig. 2.

In the adaptation experiments, we used data sets from two areas. One is the data set from Hokkaido, which gave the lowest speech translation performance on its test set. The other data set is from Kyushu, which gave the highest.

### 2.2. System configuration of MT system

Fig. 3 gives a flowchart of the MT subsystem. As shown in the figure, the MT system consists of 2 main components: the statistical based machine translation (SMT) and a translation memory.

For the SMT, we employed a log-linear model as a phrase-based SMT [7]. This model expresses the probability of a target-language word sequence ($e$) for a given source language word sequence ($f$) given by

$$P(e|f) = \frac{\exp\left(\sum_{i=1}^{M} \lambda_i h_i(e,f)\right)}{\sum_{e'} \exp\left(\sum_{i=1}^{M} \lambda_i h_i(e',f)\right)} \quad (1)$$

where $h_i(e,f)$ is the feature function, $\lambda_i$ is the feature function's weight, and $M$ is the number of features. We can approximate Eq. 1 by regarding its denominator as constant. The translation results ($\hat{e}$) are then obtained by

$$\hat{e}(f, \lambda_1^M) = \mathrm{argmax}_e \sum_{i=1}^{M} \lambda_i h_i(e,f) \quad (2)$$

We used the following eight features [7] for the translations.

1. Phrase translation probability from the source language to the target language
2. Phrase translation probability from the target language to the source language
3. Lexical weighting probability from the source language to the target language
4. Lexical weighting probability from the target language to the source language
5. Phrase penalty
6. Word penalty
7. Distortion weight
8. 5-gram language model probability

For the MT training, we use two kinds of corpora. One kind is the BTEC corpus [8]. The other types are regional expression corpora built by each project. These regional expression corpora contain dialect expressions and regional proper nouns. The corpus sizes for the Hokkaido and Kyushu projects are 3000 and 5095 sentence pairs, respectively.

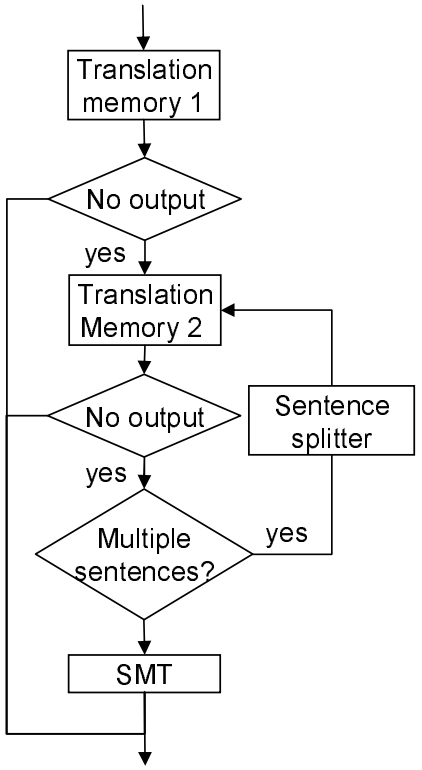


Figure 3: A flow of the Machine Translation subsystem.

First, we train two different models using these corpora. Then, by using the formula below, linearly interpolate two models [2].

$$\begin{aligned} h_{baseline}(e,f) & \\ = \quad & \mu h_{btec}(e,f) \\ & +(1-\mu) h_{regional}(e,f) \end{aligned} \quad (3)$$

Here, $h_{regional}(e,f)$ and $h_{btec}(e,f)$ are the regional model, and the BTEC model, respectively. $\mu$ is the interpolation weight. We empirically decided that the value of $\mu$ is 0.9.

These corpora are also used for translation memory. "Translation memory 1" and "Translation memory 2" in Fig. 3 use the regional expression corpus and the BTEC corpus, respectively.

## 3. Selection Method for Annotation

In this section, we go into detail about the selection method. Unlike for the field experiments, we henceforth use simple SMT models. The usage of the regional expression corpus is also different. We simply concatenate the BTEC corpus and the regional expression corpus on the corpus level, then train a single set of SMT models. These changes are to simplify the experimental system by eliminating system combination processes and linear interpolation weight tuning.

[2] This interpolation is only conducted on feature # 1 to 4 and 8 in section 2.2

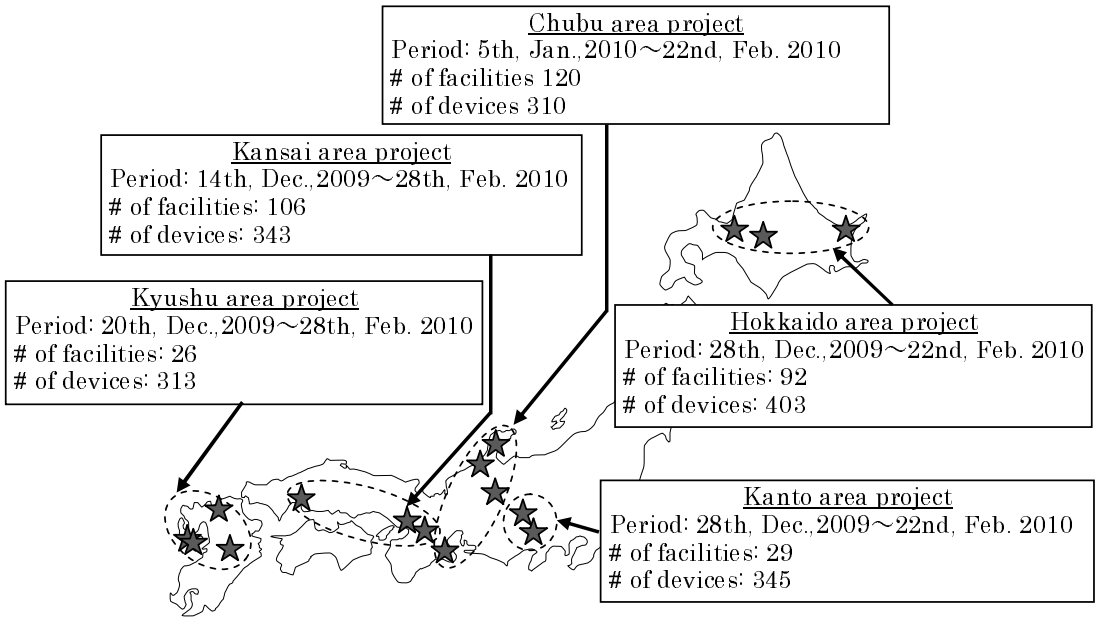


Figure 1: Overview of the five local projects.

In the proposed method, we use the source-side language model to score the sentences to be manually translated from the field data. As explained below, the proposed method selects $n$ sentences to be manually translated in incremental steps.

### 3.1. Scoring Function using Development Set Perplexity

The main idea of the first method is to select the sentences which minimize the development set perplexity calculated by the source-side language model. The actual process is as follows.

**Step 1** Train a set of SMT models using a corpus consisting of the BTEC ($C_{btec}$) and a regional expression corpus ($C_{regional}$).

**Step 2** Decode all of the transcribed field data ($C_{field}$) using the set of models trained in Step 1.

**Step 3** Take the sentences which contain out-of-vocabulary (OOV) words as selected sentences ($C_{selected}$) and update the $C_{field}$ as the remaining sentences.

**Step 4** For each sentence $s \in C_{field}$, train a source side-language model ($LM_{src1}$) using a corpus consisting of $C_{btec}$, $C_{regional}$, $C_{selected}$ and $s$. Using $LM_{src1}$, calculate the development set perplexity $PP(C_{dev}|s)$ by the following formula
$PP(C_{dev}|s) = p(C_{dev})^{-\frac{1}{n_{dev}}}$
where $C_{dev}$ is the development set, and $n_{dev}$ is the number of words in the development set.

**Step 6** Sort $C_{field}$ in increasing order using $PP(C_{dev}|s)$

**Step 7** Take the top $n$ sentences as selected sentences ($C_{selected}$) and update the $C_{field}$ as the remaining sentences.

**Step 8** If the size of $C_{selected}$ is sufficient, end. If not, go back to Step 4.

### 3.2. Scoring Function using Sentence Perplexity

The idea of the second method is selecting informative sentences by computing the informativeness using current corpus has. The second method also uses the source-side language model. This method calculates sentence unit perplexity for each sentence in the field data, then selects sentences which have a large sentence perplexity.

**Step 1 to 3** Same as section 3.2.

**Step 4** Train the source-side language models ($LM_{src2}$) using the corpus consisting of $C_{btec}$, $C_{regional}$ and $C_{selected}$.

**Step 5** For each sentence $s \in C_{field}$ calculate sentence perplexity($PP(s)$) with the following formula
$PP(s) = p(s)^{-\frac{1}{n_s}}$
where $n_s$ is the number of words in $s$.

**Step 6** Sort $C_{field}$ in decreasing order using $PP(s)$

**Step 7** Take the top $n$ sentences as the selected sentences ($C_{selected}$) and update the $C_{field}$ as the remaining sentences.

**Step 8** Update $LM_{src2}$ (train a single language model using $C_{btec}$, $C_{btec}$ and $C_{selected}$).

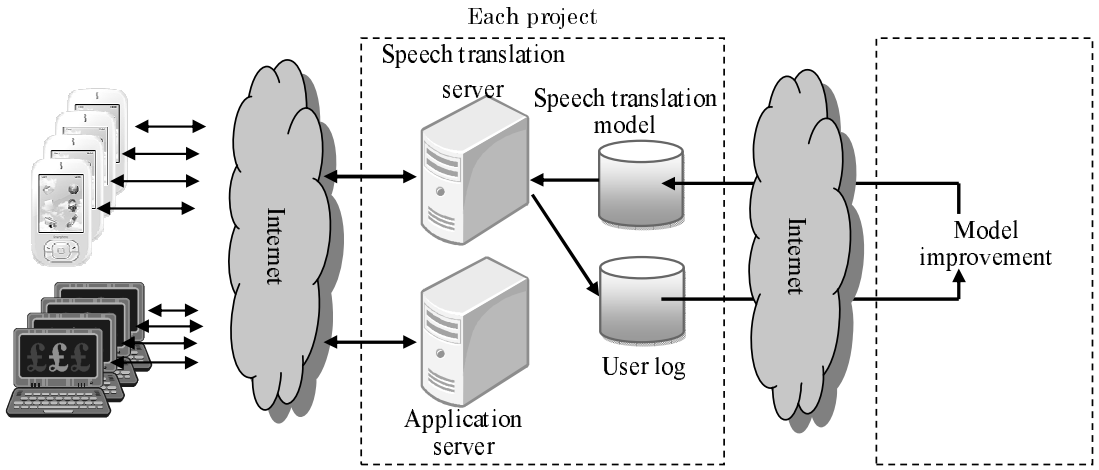


Figure 2: A schematic diagram of system configuration for the speech-to-speech translation experiment.

**Step 9** If the size of $C_{selected}$ is sufficient, end. If not, go back to Step 4.

## 4. Experiments

### 4.1. Experimental Settings

Table 1 shows data usage for the SMT models training. As shown in the table, we evaluate two baseline systems. "Baseline 1" is trained on the BTEC corpus and a regional expression corpus. Using the "Baseline 1", firstly, we translate all of the field data. We divide the field data into two parts, an OOV containing part (OOV sentences) and a non-OOV part (non-OOV sentences). In addition to the training data used for the "Baseline 1", "Baseline 2" uses OOV sentences and their manual translation. The baseline selection is a simple random selection. And, upper bound uses all of the field data and its manual translation.

Table 2 shows the details of data sets used for the annotation data selection experiments. For each of the data sets (Hokkaido and Kyushu), we randomly sampled 1000 sentences from the field data to be used for the development sets and test sets. These development sets are used for minimum error rate training[9] and annotation data selection. The test sets are for the MT performance evaluation.

In the experiments, we evaluate selection performance by computing the BLEU score [10] of the SMT system trained on the selected sentences including manual translation of the selected sentences. For the translation model and language model training, we use MOSES [11] and the SRI language model toolkit [12]. For the language model setting, we used a modified Kneser-Ney [13] 5-gram language model.

For the data selection, we use 3-gram language model and set $n$ (the number of selected sentence in each iteration) to be 1,000.

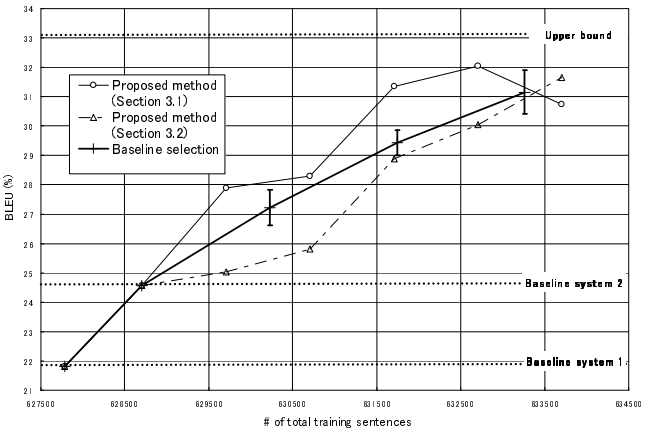


Figure 4: Evaluation results of annotation data selection (Hokkaido)

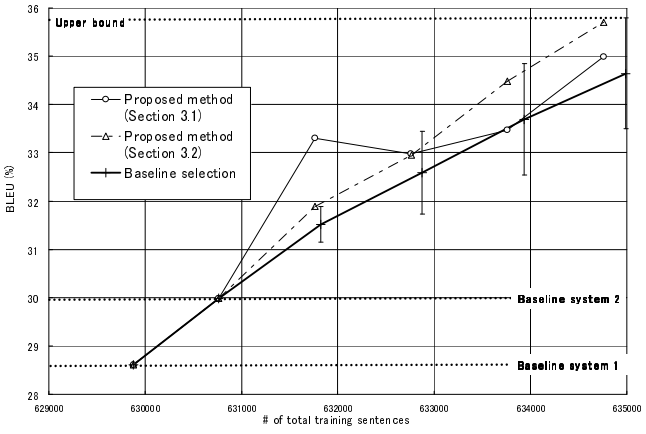


Figure 5: Evaluation results of annotation data selection (Kyushu)

Table 1: Data usage of SMT models' training

| System Type | BTEC Corpus | Regional Expression Corpus | Field Data Containing OOV | Field Data not Containing OOV |
|---|---|---|---|---|
| Baseline System 1 | Used | Used | Not used | Not used |
| Baseline System 2 | Used | Used | Used | Not used |
| Baseline Selection (Random Selection) | Used | Used | Used | Partly used (10% to 60%) |
| Proposed Selection | Used | Used | Used | Partly used (10% to 60%) |
| Upper Bound | Used | Used | Used | Used |

Table 2: Data set of annotation data selection experiments

| Project Area | Method | BTEC Corpus | Regional Expression Corpus | Field Data Containing OOV | Field Data not Containing OOV | Test set |
|---|---|---|---|---|---|---|
| Hokkaido | 624,782 | 3,000 | 923 | 7,596 | 500 | 500 |
| Kyushu | 624,782 | 5,905 | 885 | 5,287 | 500 | 500 |

### 4.2. Experimental Results

Figure 4 and 5 show the evaluation results of the Hokkaido data set and Kyushu data set, respectively. The vertical axis represents the BLEU score, and the horizontal axis represents the total number of the training sentences. In these figures, ◯ and △ show the results of the proposed methods. Figure 4 shows the results of the 1st to 5th iteration. Here, up to 5,000 sentences are selected out of 7,596 by the proposed method. Figure 5 shows the results of the 1st to 4th iteration. Up to 4,000 sentences are selected out of 5,287. To calculate the BLEU scores for proposed methods, we add selected sentences and their manual translations to the training corpus. Then, the SMT models are retrained.

Thick lines with error bars show the results of the baseline random selection. We carried out 5 random selections for each data size. For each selection, the same model's retraining procedures are taken as the proposed methods. The error bars represent standard deviations of BLEU scores.

Comparing "Baseline system 1" and "Baseline system 2", large improvements are obtained. These improvements are effects of adding OOV sentences and their manual translations into the training corpora.

In both of the figures, the selection method using the development set perplexity (◯ in the Figures) gives better performance than the baseline random selection. This method performs very well, especially in selecting amounts of data. In the best case in Figure 5, a 1.78 point BLEU score improvement is obtained by the proposed method. To obtain the same performance by using baseline random selection, 2,000 more sentences would need to be manually translated.

The selection method using the sentence perplexity (△ in the Figures) also gives a better performance than the baseline random selection on the Kyushu data set (figure 5). However, the method gives a worse performance than the baseline random selection for the Hokkaido data set.

### 4.3. Future Works

In the experiments of this paper, we compare two language model based selection methods. Roughly speaking, the development set perplexity based method tends to select sentences which are close to the development set. Meanwhile, the test set perplexity-based method tends from select sentences which have new information to a currently available corpus. These two methods have totally different perspectives for selecting sentences from field data to be annotated.

Both of perspectives are very important for improving corpus based MT. Consequently, we will improve the selection method by using multiple scores given by the two methods in the near future.

The other challenge to be addressed is the selection of speech data to be transcribed. In the experiments in this paper, we only used manual transcriptions as the selection targets. For the efficient improvement of total performance of speech-to-speech translation systems, we will tackle data selection the speech data transcription.

## 5. Conclusion

In order to efficiently improve machine translation systems, we proposed a method which selects data to be annotated (manually translated) from speech-to-speech translation field data.

We carried out annotation data selection experiments using data from speech-to-speech translation field experiments conducted during fiscal year 2009 in Japan. The field experiments were conducted in 5 areas: the Hokkaido area, the Kanto area, the Chubu area, the Kansai area and the Kyushu area. In the selection experiments, we used data sets from two areas. One was the data set from Hokkaido, which gave the lowest speech translation performance on its test set. The other data set is from Kyushu, which give the highest speech

translation performance.

In the experiments, we evaluated selection performance by computing the BLEU score for an SMT system trained on the selected sentences including manual translation of the selected sentences. According to the experimental results, the selection method using the development set perplexity gives better performance than the baseline random selection. This method performs very well, especially when the method selects small amounts of data.

The other selection method which use the sentence perplexity also gives better performance than the baseline random selection for the Kyushu data set, but a worse performance than the baseline random selection on the Hokkaido data set. Considering this point, we conclude that the proposed methods can enable efficient annotation of the field data to improve MT performance.

In the future work, we will combine these two methods for better selection performance.

## 6. References


[1] N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A. W. Black, "Incremental adaptation of speech-to-speech translation," in *Proceedings of NAACL HLT 2009*, 2009, pp. 149–152.

[2] H. Kawai, R. Isotani, K. Yasuda, E. Sumita, U. Masao, S. Matsuda, Y. Ashikari, and S. Nakamura, "An overview of a nation-wide field experiment of speech-to-speech translation in fiscal year 2009 (in Japanese)," in *Proceedings of 2010 autumn meeting of Acoustical Society of Japan*, 2010, pp. 99–102.

[3] S. F. Hiroko Murakami, Koichi Shinoda, "A relative entropy based data selection approach for acoustic model training (in Japanese)," in *Proceedings of the ASJ spring meeting 2011*, no. 1-5-7, March 2011, pp. 17–20.

[4] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 23–31, 2005.

[5] N. Ueffing, G. Haffari, and A. Sarkar, "Semi-supervised model adaptation for statistical machine translation," *Machine Translation*, vol. 21, no. 2, pp. 77–94, 2007.

[6] K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, "Method of selecting training data to build a compact and efficient translation model," in *Proceedings of the Third International Joint Conference on Natural Language Processing*, 2008, pp. 655–660.

[7] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," *Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pp. 127–133, 2003.

[8] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), 2006, pp. 1674–1682.

[9] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds., 2003, pp. 160–167. [Online]. Available: http://www.aclweb.org/anthology/P03-1021.pdf

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, June 2007, pp. 177–180.

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[13] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Technical report TR-10-98, Center for Research in Computing Technology (Harvard University)*, 1998.