Neural Network Language Models for Translation with Limited Data

Maxim Khalilov, José A. R. Fonollosa

Centre de Recerca TALP
Universitat Politècnica de Catalunya
Barcelona, 08034 Spain
{khalilov,adrian}@gps.tsc.upc.edu

F. Zamora-Martínez, M.J. Castro-Bleda,
S. España-Boquera

Dep. de Lenguajes y Sistemas Informáticos
Universidad Politécnica de Valencia
Valencia, 46022 Spain
{fzamora,mcastro,sespana}@dsic.upv.es

Abstract

In this paper we present how to estimate a continuous space Language Model with a Neural Network to be used in a Statistical Machine Translation system. We report results for an Italian-English translation task obtained on a small corpus (about 150 K tokens), that can be considered a task with a lack of training data. Different word history length included in the connectionist language model (n-gram order) and distinct continuous space representation (i.e. words appearing in the training corpus more than k times) are considered in the study. The experimental results are evaluated by means of automatic evaluation metrics correlated with fluency and adequacy of the generated translations.

1. Introduction

Language modeling is an essential step in many Natural Language Processing applications, and particularly in the Statistical Machine Translation (SMT) task. Techniques for language modeling can be classically decomposed into two main approaches. The first type of models comprises traditional grammars, for example, synchronous context-free grammars. The second approach includes purely statistical corpus-based probabilistic models, which is a powerful and simple method for language modeling. The so-called n-gram models, which assign high probability to frequent sequences of words by considering the history of only $n-1$ preceding words in the utterance, has become a "de facto" standard for language modeling in the state-of-the-art SMT systems.

The approach presented in this paper can be considered as a coherent and natural evolution of the probabilistic Language Models (LMs): we propose to use a continuous LM trained in the form of a Neural Network (NN).

The use of continuous space representation of language has successfully applied in recent NN approaches to language modeling [32, 3, 8]. However, the use of Neural Network Language Models (NN LMs) in state-of-the-art SMT systems is not so popular. The only comprehensive work refers to [28], where the target LM is presented in the form of a fully-connected Multilayer Perceptron.

The Basic Travel Expression corpus [30] from the tourist domain has been used in our experiments. This corpus is characterized by extremely limited amount of training data (about 150 K of tokens in the English part of the training corpus), as compared to other translation tasks (for example, Europarl corpus of parliament speeches contains 35 M words). We decided to consider the translation between two European languages with distinct inflection, but a similar word order, i.e. Italian to English translation.

The lack of vast bilingual resources requires special techniques for the integration into a machine translation system. With regard to language modeling, the recently presented specific algorithms include: (a) techniques dealing with class-based n-gram LMs [22, 31]: some promising works include factored LM representation allowing for different lexical and syntactical text clusterization [19]; (b) LM adaptation to the particular translation task [17, 18] and (c) other techniques which include synchronous context-free grammar LMs, as shown in [9], continuous space LMs, and other non-trivial language modeling algorithms, such as NN LMs.

The article is structured as follows: in Section 2 we describe the novel feature presented in the paper, i.e. NN LMs and its training algorithm. In Section 3 we give some background of the SMT and briefly outline the n-gram-based SMT system. Section 4 presents our experimental setup and Section 5 concludes the article with the results and the leading discussions.

2. Neural Network Language Models

A different approach to the widely-used statistical language models based on n-grams consists on using Neural Networks. A NN LM is a statistical LM which follows the same equation as n-grams:

$$p(w_1 \dots w_{|W|}) \approx \prod_{i=1}^{|W|} p(w_i | w_{i-n+1} \dots w_{i-1}) \qquad (1)$$

and where the probabilities that appear in that expression are estimated with a NN. The model naturally fits under the probabilistic interpretation of the outputs of the NNs: if a NN is trained as a classifier, the outputs associated to each class are estimations of the posterior probabilities of the defined classes. The demonstration of this assertion can be found in a number of places, for example in [5].

The training set for a LM is a sequence $w_1 w_2 \dots w_{|W|}$ of words from a vocabulary Ω. In order to train a NN to predict the next word given a history of length $n-1$, each input word must be encoded. A natural representation is a local encoding following a "1-of-$|\Omega|$" scheme. The problem of this encoding for tasks with large vocabularies (as is the case) is the huge size of the resulting NN. We have solved this problem following the ideas of [3], learning a distributed representation for each word.

Figure 1 illustrates the architecture of the feed-forward NN used to estimate the NN LM. The input is composed of words $w_{i-n+1}, \dots, w_{i-1}$ of Equation (1). Each word is represented using a local encoding. P is the projection layer of the input words, formed by $P_{i-n+1}, \dots, P_{i-1}$ subsets of projection units. The subset of projection units P_j represents the distributed encoding of input word w_j. The weights of this projection layer were linked, that is, the weights from each local encoding of input word w_j to the corresponding subset of projection units P_j are the same for all input words j.

H denotes the hidden layer and the output layer O has $|\Omega|$ units, one for each word of the vocabulary. Trained as a classifier, this NN predicts the posterior probability of each word of the vocabulary given the history, i.e., $p(w_i | w_{i-n+1} \dots w_{i-1})$.

In order to achieve a good configuration (topology and parameters) for each NN LM in the translation task, exhaustive scanning using a tuning set was performed. The activation function for the hidden layers was the *hyperbolic tangent* function and the *softmax* function was chosen for the output units. Best configurations used a projection layer of 32 units for each word.

To illustrate the huge sizes of the NNs used, Table 1 shows the topology and number of weights of the selected NN LMs for a vocabulary of 2 148 words (words with less than k=5 occurrences were discarded from the Basic Travel Expression corpus) and a vocabulary of 3 093 (corresponding to k=3). The third columns shows the topology of the used NNs (number of input, projection, hidden and output units) and the last column shows the number of weights (first, the weights replicated $n-1$ times at the projection layer and, secondly, the weights at the hidden and output layers).

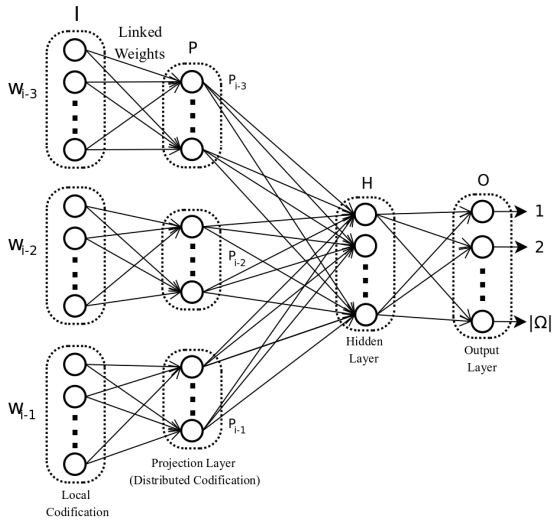

Figure 1. Architecture of the continuous space NN LM. The input words are $w_{i-n+1}, \dots, w_{i-1}$ (in this example, the input words are w_{i-3}, w_{i-2}, and w_{i-1} for a 4-gram). I, P, H and O are the input, projection, hidden and output layer, respectively, of the Multilayer Perceptron.

3. SMT system

SMT is based on the principle of translating a source sentence s into a sentence in the target language t. The problem is formulated in terms of source and target languages and is defined according to the following Equation (2) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (3):

$$\hat{t} = \arg\max_t \{ p(t \mid s) \} = \qquad (2)$$

$$= \arg\max_t \{ p(s \mid t) \cdot p(t) \} . \qquad (3)$$

This decomposition made according to the Bayes rule is called *noisy channel* approach, and the first systems following this approach performed translation on the word level [6]. However, modern state-of-the-art SMT systems

Table 1. Sizes of the selected NN LMs configurations.

NN LM		NN Topology	
Vocabulary	n-gram	Input–Projection–Hidden–Output	# Weights
k=5	3-gram	2×2 148–2×32–64–2 148	2×68 768 + 143 788
2 148	4-gram	3×2 148–3×32–64–2 148	3×68 768 + 145 828
k=3	3-gram	2×3 093–2×32–64–3 093	2×99 008 + 205 205
3 093	4-gram	3×3 093–3×32–64–3 093	3×99 008 + 207 253

operate with bilingual units extracted from the parallel corpus based on the word-to-word alignment. An enhancement of the SMT systems consists of calculating the posterior probability as a log-linear combination of a set of feature functions [4, 25]. Using this technique, it is possible to combine M feature models in the determination of the translation hypothesis, as shown below in Equation (4):

$$\hat{t} = \arg\max_{t} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t, s) \right\}, \quad (4)$$

where the feature functions h_m refer to the system models, namely bilingual translation model, target LM and additional feature models; and the set of λ_m refers to the weights corresponding to these models which are estimated according to a log-linear model, so that the recombined weights are optimized to maximize the translation scores on the development set (see Sections 4.1 and 4.2).

3.1. Translation model

Most of modern state-of-the-art SMT follow the *phrase-based* translation approach. The basic idea is to segment the given source word sequence into monolingual phrases, afterwards translate them and compose the target sentence [21, 25].

Another approach to SMT is the *n-gram-based* approach, which we follow in the framework of the study. It regards translation as a stochastic process maximizing the joint probability $p(s, t)$, leading to a decomposition based on bilingual n-grams, typically implemented by means of a Finite-State Transducer [7]. It operates with tuples that are extracted from a word-to-word alignment according to certain constraints, explained in details in [15]. The translation model is represented in the form of a 4-gram LM estimated using Kneser-Ney discounting, where the language is composed by tuples. The tuples induce a unique segmentation of the pairs of sentences, as shown in [12]. In this way the context used in the translation model is bilingual, it not only takes the target sentence into account, but both languages linked in tuples.

3.2. Other features

Besides the bilingual translation model, the baseline translation system implements a log-linear combination of several other features:

- *An statistical n-gram target LM.* 4-gram word-based model is used in the system that accounts for the target language statistical dependencies.

- *A connectionist target LM.* Different word history length included in the connectionist language model (n-gram order) and distinct continuous space representation (i.e. words appearing in the training corpus more than k times) are considered in this study.

- *A word penalty model.* A word penalty model is used to compensate the systems preference for short output sentences. Technically, the penalization depends on the total number of words in the partial translation hypothesis.

- *A source-to-target and a target-to-source lexicon models.* This model uses word-to-word IBM Model 1 probabilities [24] to estimate the lexical weights of each tuple. The target-to-source lexicon model is the same as the source-to-target lexicon model for the opposite translation direction. We used Giza++ [1] word-to-word direct and backward alignments respectively.

- *Extended word reordering.* An extended monotone distortion model based on the automatically learned reordering rules was used in the experiments for the Italian-English translation task. Reordering patterns are extracted in training from the crossed links found in the word alignment, on the next step, the monotone search graph is extended with reorderings following the patterns found in training. Once the search graph is built, the decoder traverses the graph looking for the best translation. The above mentioned distortion model is presented in [14].

3.3. Decoding and optimization

The MARIE decoder was used as a search engine for the translation system. The details can be found in [13]. The decoder implements a beam-search algorithm with pruning capabilities. The feature functions described above were taken into account in the decoding process. Given the development set and references, the log-linear combination of weights can be adjusted using the simplex optimization method [23] to maximize the score function according to a combination of automatic evaluation metrics (see Section 4.2) [26]. Detailed explanation of the standard automatic metrics to evaluate the translation quality, along with the optimization criteria that was used to tune the translation system, are presented in Sections 4.1 and 4.2.

4. Experiments

The experiment results were obtained on the Basic Travel Expression corpus, which includes data from a tourist domain. This corpus models a real situation when an Italian tourist appears in an English-speaking country and demands for simple explanations and other information useful for travellers. Along with regular senteces, like "Questo traghetto si sta dirigendo verso un'isola'.' (*"This ferry is heading for an island."*), it contains many colloquial or simple expressions, like "Hm! non mi sento bene." (*"Hm! I am not feeling well."*).

Automatic evaluation conditions were case-sensitive with tokenized punctuation marks. The development and test sets were provided with 7 reference translations. Basic Travel Expression corpus statistics can be found in Table 2. The number of words and the size of the vocabulary for the development and test reference English sets are calculated by average of the 7 references.

4.1. Translation scores

The BLEU score accounts for evaluation of the translation quality, by measuring the distance between a given translation and the set of reference translations using an n-gram LM (a 4-gram in the framework of this study) [26]. The NIST score is a sensitive metric of machine translation quality, based on the BLEU score, but weighting n-grams in order to provide less informative n-grams with higher weigths [16]. The METEOR score is an underestimated metric for the evaluation of machine translation output, which is calculated as an averaged mean of precision and benefited recall, considering stemms and synonyms matching (more details can be found in [2]).

Table 2. Statistics of the Basic Travel Expression corpus.

	Italian	English
Train		
Sentences	24.5 K	24.5
Words	166.3 K	155.4 K
Vocabulary	10.2 K	7.3 K
Development		
Sentences	489	489
Words	5.2 K	5.6 K
Vocabulary	1.2 K	1.7 K
Test		
Sentences	500	500
Words	6 K	7.3 K
Vocabulary	1.4 K	2.3 K

4.2. Baseline

The Italian part of the bilingual corpus was preprocessed. This step included tagging, lemmatization and separation of contractions as described in [11]. The optimization criteria to estimate the weights of the log-lineal model of Equation (4) was 100 BLEU + 4 NIST in the development set, following the point from [10].

A 4-gram target LM with unmodified Kneser-Ney backoff discounting and counts post-modification after discount estimation were generated using the SRI Language Modeling Toolkit [29]. The 4-gram was implicitly integrated into the SMT system and considered as the reference baseline, without taking into account the NN LM.

Tables 3 and 4 show BLEU, NIST and METEOR scores for the baseline system for the development and the test sets. Automatic evaluation was case insensitive and punctuation marks were not considered.

4.3. NN LMs experiments

Target NN LMs were trained on exactly the same training data as the 4-gram target LM. We considered two key parameters of the continuous NN LM: (*a*) *word frequency threshold* k: words with less than k occurrences were discarded; (*b*) *order of* n*-gram*: 3-gram and 4-gram were tested.

When reestimating the weights coefficients for the new log-lineal model with the NN LM, different start points were tried and the best set of weights due to the 100 BLEU + 4 NIST criteria was chosen. Table 3 and Table 4 show BLEU, NIST and METEOR scores when the NN LMs were

Table 3. Evaluation scores on the development dataset.

		BLEU	NIST	METEOR
Baseline		29.22	6.37	69.26
NN LM k=5	3-gram	30.02	6.31	69.44
	4-gram	30.07	6.17	69.19
NN LM k=3	3-gram	**30.54**	**6.44**	**69.61**
	4-gram	30.01	6.10	69.45

Table 4. Evaluation scores on the test dataset.

		BLEU	NIST	METEOR
Baseline		24.93	5.83	64.01
NN LM k=5	3-gram	25.17	5.86	63.70
	4-gram	25.07	5.79	63.99
NN LM k=3	3-gram	**25.23**	**6.02**	**64.10**
	4-gram	25.29	5.81	63.63

integrated as a part of the combined SMT system, for the development and the test sets.

As can be observed, considerable improvements were obtained by using a NN LM. The best system configuration is highlighted in both Tables.

For the development dataset, the BLEU score for the NN LM experiments is always higher than for the baseline system. The METEOR score for the NN LM system is slightly higher than the reference one for most of the configurations.

Our previous experience shows that, for small translation tasks with a lack of training material, poor correlation of development and test results is frequent, although this has not been the case in these experiments. Considering development and test data results, the 3-gram k=3 NN LM system allows gaining up to 1.3 BLEU point for the development set and about 0.3 BLEU point for the test set. This difference is statistically significant for a 95% confidence interval and 1 000 resamples), using the bootstrap resampling method as described in [20].

Considering the NIST score, the baseline test results were exceeded for both 3-gram systems. Concerning METEOR score, only the 3-gram, k=3 system provides better LM generalization.

5. Discussion and error analysis

The architecure of a SMT system implies that the smaller the available training data, the worse the performance of a translation system. Obviously, new or specially adapted methods of limited information using in more efficient way are needed. The technique presented in this paper allows improving the performance of a SMT system having access to a small amount of training material by incorporating the NN LM.

The correlation of automatic and subjective human evaluation metrics (fluency and adequacy) is one of the main topics in the area of machine translation evaluation. As it was reported in [27] for small translation tasks fluency correlates best with BLEU and adequacy correlates best with METEOR, while the NIST metric has only moderate correlation to both subjective human evaluation metrics. Our work demonstrates the potential for NN LMs application in the SMT to improve translation fluency, while adequacy remains the same. The positive impact of higher n-gram is not clear, this is possibly due to the relatively short sentences provided within the Basic Travel Expression corpus; probably for a corpus with longer sentences this influence will be more considerable. Another possible issue is that higher n-gram order only slightly decreases translation quality, but, by other hand, it introduces more noisy translation hypotheses.

An example of a typical sentence from the Basic Travel Expression corpus is shown in Figure 2. The Italian expression "Oggi abbiamo a scelta" is translated by the baseline system as "Today we have selection at", whereas three of four NN LMs systems provide a more fluent translation "Today we have to choose from".

The contribution of this paper is to show the robustness of the NN LM even for highly limited training corpus. The in-domain NN LM provides a significantly better generalization of the target language, smoothed SMT output and improvement in the automatically evaluated translation scores.

6. Acknowledgments

This work was partially funded by the Spanish Government and FEDER under grants TEC2006-13964-C03 (AVIVAVOZ project) and TIN2005-08660-C04-02 (EDECAN project), by the Vicerrectorado de Innovación y Desarrollo of the Universidad Politécnica de Valencia under contract 4681 and under a FPU grant.

Source	Oggi abbiamo a scelta insalata ai frutti di mare insalata di patate e insalata mista.
References	Today we have a choice of seafood salad potato salad and wild vegetables salad.
	We are serving seafood salad potato salad and wild vegetables salad today.
	As for today's salad you can enjoy seafood potato and wild vegetables.
	For salad we have seafood potato and wild vegetables today.
	Today's selections are the seafood salad potato salad and wild vegetables salad.
	For today we have the seafood salad potato salad and wild vegetables salad.
	For today you can choose to have the seafood salad the potato salad or the wild vegetables salad.
Baseline	Today *we have selection at* the seafood salad potato salad and mixed salad.
3-gram k=5	Today *we have to choose from* the seafood salad potato salad and mixed salad.
4-gram k=5	Today we have selection at the seafood salad potato salad and mixed salad.
3-gram k=3	Today *we have to choose from* the seafood salad potato salad and mixed salad.
4-gram k=3	Today *we have to choose from* the seafood salad potato salad and mixed salad.

Figure 2. An example of translation.

References

[1] `http://www.fjoch.com/GIZA++.html`.

[2] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(2):1137–1155, 2003.

[4] A. Berger, S. D. Pietra, and V. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, 1996.

[5] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[6] P. Brown, J. Cocke, S. D. Pietra, V. D. Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[7] F. Casacuberta, E. Vidal, and J. M. Vilar. Architectures for speech-to-speech translation using finite-state models. In *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, pages 39–44, 2002.

[8] M. J. Castro and F. Prat. New Directions in Connectionist Language Modeling. In *Computational Methods in Neural Modeling*, volume 2686 of *LNCS*, pages 598–605. Springer-Verlag, 2003.

[9] E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation. In *Proceedings of the MT Summit IX. Intl. Assoc. for Machine Translation.*, 2003.

[10] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. The ITC-irst SMT system for IWSLT-2005. In *Proceedings of IWSLT 2005*, page 98104, 2005.

[11] J. Crego, A. de Gispert, P. Lambert, M. Khalilov, M. Costajussà, J. Mariño, R. Banchs, and J. Fonollosa. The TALP Ngram-based SMT System for IWSLT 2006. In *Proceedings of IWSLT 2006*, pages 116–122, 2006.

[12] J. M. Crego, J. Mariño, and A. de Gispert. Finite-state-based and Phrase-based Statistical Machine Translation. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 37–40, 2004.

[13] J. M. Crego, J. Mariño, and A. de Gispert. An Ngram-based Statistical Machine Translation Decoder. In *Proceedings of INTERSPEECH05*, pages 3185–3188, 2005.

[14] J. M. Crego and J. B. Mario. Improving statistical mt by coupling reordering and decoding. *Machine Translation*, 20(3):199–215, 2006.

[15] A. de Gispert and J. Mariño. Using X-grams for Speech-to-Speech Translation. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 1885–1888, 2002.

[16] G. Doddington. Automatic evaluation of machine translation quality using n-grams co-occurrence statistics. In *In HLT 2002 (Second Conference on Human Language Technology)*, pages 128–132, 2002.

[17] S. Hewavitharana, B. Zhao, A. S. Hildebrand, M. Eck, C. Hori, S. Vogel, and A. Waibel. The CMU statistical machine translation system for IWSLT2005. In *Proceedings of IWSLT 2005*, pages 63–70, 2005.

[18] M. Khalilov and J. A. R. Fonollosa. Language modeling for verbatim translation task. In *Proceedings of the IV Jornadas en Tecnología del Habla - the IV Biennial Workshop on Speech Technology*, pages 83–87, 2006.

[19] K. Kirchhoff. and M. Yang. Improved Language Modeling for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, 2005.

[20] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2004*, pages 388–395, 2004.

[21] D. Marcu and W. Wong. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of EMNLP02*, pages 133–139, 2002.

[22] A. Menezes and C. Quirk. Microsoft research treelet translation system: IWSLT evaluation. In *Proceedings of IWSLT 2006*, pages 105–108, 2005.

[23] J. Nelder and R. Mead. A simplex method for function minimization. *The Computer organization*, 7:308–313, 1965.

[24] F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of HLTNAACL04*, pages 161–168, 2004.

[25] F. J. Och and H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of ACL02*, pages 295–302, 2002.

[26] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL 2002*, pages 311–318, 2002.

[27] M. Paul. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of IWSLT06*, pages 1–15, 2006.

[28] H. Schwenk, D. Dchelotte, and J. L. Gauvain. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, 2006.

[29] A. Stolcke. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904, 2002.

[30] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152, 2002.

[31] H. Tsukada, T. Watanabe, J. Suzuki, H. Kazawa, and H. Isozaki. The NTT statistical machine translation system for IWSLT 2005. In *Proceedings of IWSLT 2006*, pages 128–133, 2005.

[32] P. Xu and F. Jelinek. Random forest in language modeling. In *Proceedings of EMNLP 2004*, pages 325–332, 2004.