Bilingual N-gram Statistical Machine Translation

José B. Mariño **Rafael E. Banchs** **Josep María Crego**
Adrià de Gispert **Patrik Lambert** **José A. R. Fonollosa**
Marta Ruiz
Universitat Politècnica de Catalunya
Department of Signal Theory and Communications, Barcelona 08034, Spain
(canton,rbanchs,jmcrego,agispert,lambert,adrian,mruiz)@gps.tsc.upc.edu

Abstract

This paper describes a statistical machine translation system that uses a translation model which is based on bilingual n-grams. When this translation model is log-linearly combined with four specific feature functions, state of the art translations are achieved for Spanish-to-English and English-to-Spanish translation tasks. Some specific results obtained for the EPPS (European Parliament Plenary Sessions) data are presented and discussed. Finally, future research issues are depicted.

1 Introduction

Statistical machine translation (SMT) constitutes a research area that has gained much attention worldwide during the last years. The idea of machine translation can be traced back to the early fifties (Knight, 1997). However, it wasn't until the beginning of the nineties when this technology experienced an actual growth which was motivated by the development of computer resources needed to allow the implementation of translation algorithms based on statistical methods (Brown *et al.*, 1990) and (1993).

The first SMT systems were based on the noisy channel approach, which models the probability of a target language sentence T given a source language sentence S as a translation model probability $p(S|T)$ times a target language model probability $p(T)$. In recent systems such an approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och and Ney, 2002). Additionally, original word-based translation models (Brown *et al.*, 1993) have been replaced by phrase-based translation models (Zens *et al.*, 2002) and (Koehn *et al.*, 2003), which are estimated from aligned bilingual corpora by using relative frequencies.

On the other hand, the translation problem has also been approached from the finite-state perspective as the most natural way for integrating speech recognition and machine translation into a speech-to-speech translation system (Riccardi *et al.*, 1996), (Vidal, 1997), (Bangalore and Riccardi, 2001) and (Casacuberta, 2001). In this kind of systems the translation model constitutes a finite-state network which is learned from training data.

The translation system described in this paper implements a translation model based on the finite-state perspective, (de Gispert and Mariño, 2002) and (de Gispert *et al.*, 2004), which is used along with a log-linear combination of four additional feature functions (Crego *et al.*, 2005). The implemented translation model, which is referred to as *tuple n-gram model*, differs from the well known phrase-model approach (Koehn *et al.*, 2003) in two basic issues. First, training data is monotonously and uniquely segmented into bilingual units called tuples. And second, the model considers n-gram probabilities instead of relative frequencies.

This paper is structured as follows. Next section describes in detail the bilingual n-gram translation model. Then, section 3 presents a description of the overall system. Section 4 presents and discusses some translation results obtained for the EPPS (European Parliament Plenary Sessions) data, specifically for Spanish-to-English and English-to-Spanish translation tasks. Finally, some discussion and conclusions are presented along with some further intended work.

2 The Tuple N-gram Model

As already mentioned, the translation model used here is based on bilingual n-grams. This section describes in detail this translation model, which is called the tuple n-gram model. This model is actually a language model of a particular language composed by bilingual units

which are referred to as tuples (de Gispert and Mariño, 2002). This model approximates the joint probability between source and target languages by using 3-grams as described by the following equation:

$$p(T,S) \approx \prod_{n=1}^{N} p((t,s)_n | (t,s)_{n-2}, (t,s)_{n-1}) \quad (1)$$

where t refers to target, s to source and $(t,s)_n$ to the n^{th} tuple of a given bilingual sentence pair. It is important to notice that, since both languages are linked up in tuples, the context information provided by this translation model is bilingual.

Tuples are extracted from a word-to-word aligned corpus. More specifically, word-to-word alignments are performed in both directions, source-to-target and target-to-source, by using GIZA++ (Och and Ney, 2000). Then, tuples are extracted from the union set of alignments according to the following constraints (Crego *et al.*, 2004):

- a monotonous segmentation of each bilingual sentence pairs is produced,
- no word inside the tuple is aligned to words outside the tuple, and
- no smaller tuples can be extracted without violating the previous constraints.

As a consequence of these constraints, only one segmentation is possible for a given sentence pair. Figure 1 presents a simple example illustrating the tuple extraction process.

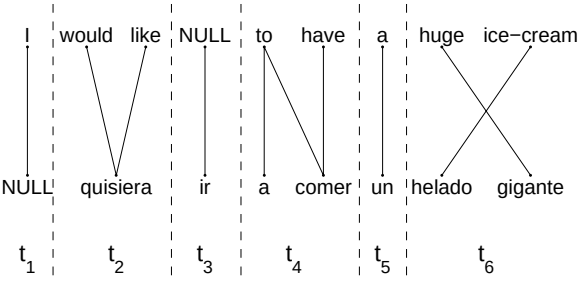

Figure 1: Example of tuple extraction from an aligned bilingual sentence pair.

Once tuples have been extracted, the tuple vocabulary is pruned by using histogram counts. This pruning is performed by keeping the N most frequent tuples with same source sides. In the case of the EPPS data, a value of $N = 20$ provided a good trade off between translation quality and computational expenses for Spanish-to-English translations; while a value of $N = 30$ provided the best trade off in the case of English-to-Spanish. After pruning, the tuple 3-gram model is trained by using the SRI Language Modeling toolkit (Stolcke, 2002) and the improved back-off smoothing method proposed by Kneser and Ney (1995).

Two important issues regarding this translation model must be considered. First, it often occurs that an important amount of single-word translation probabilities are left out of the model. This happens for all those words that appear always embedded into tuples containing two or more words. Consider for example the word "ice-cream" from figure 1. As seen from the figure, "ice-cream" appears embedded into tuple t_6. If a similar situation is encountered for all occurrences of "ice-cream" in the training corpus then no translation probability for an independent occurrence of such word will exist.

To overcome this problem, the tuple 3-gram model is enhanced by incorporating 1-gram translation probabilities for all the embedded words detected during the tuple extraction step (de Gispert *et al.*, 2004). These 1-gram translation probabilities are computed from the intersection of both, source-to-target and target-to-source, alignments.

The second important issue has to do with the fact that some words linked to NULL end up producing tuples with NULL source sides. Consider for example the tuple t_3 from figure 1. Since no NULL is actually expected to occur in translation inputs, such a kind of tuple cannot be allowed. This problem is solved by preprocessing the union set of alignments before the tuple extraction is performed. During this preprocessing, any target word that is linked to NULL is attached to either its precedent word or its following word[1]. In this way, no target word remains linked to NULL, and tuples with NULL source sides will not occur during tuple extraction.

3 The Translation System

As already mentioned in the introduction, the translation system presented here implements a log-linear combination of feature functions

[1]In the present version of the system, target words aligned to NULL are always attached to the following word. Further work in this area is proposed in the last section.

along with the tuple n-gram model. This section describes the log-linear model and each of the four specific feature functions that are used. Finally, a brief description of the customized decoding tool that is used is presented.

3.1 Log-Linear Model Framework

According to the maximum entropy framework (Berger *et al.*, 1996), the corresponding translation hypothesis T, for a given source sentence S, is defined by the target sentence that maximizes a log-linear combination of feature functions $h_i(S,T)$, as described in the following equation:

$$\underset{T}{argmax} \sum_i \lambda_i h_i(S,T) \quad (2)$$

where the λ_i's constitute the weighting coefficients of the log-linear combination and the feature function $h_i(S,T)$ corresponds to a logarithmic scaling of the i^{th}-model probabilities.

These weights are computed via an optimization procedure which maximizes the translation BLEU (Papineni *et al.*, 2002) over a given development set. This optimization is performed by using an in-house developed optimization algorithm, which is based on a simplex method (Press *et al.*, 2002).

3.2 Implemented Feature Functions

The proposed translation system implements a total of five feature functions:

- tuple 3-gram model,
- target language model,
- word penalty model,
- source-to-target lexicon model, and
- target-to-source lexicon model.

The first of these functions is the tuple 3-gram model, which was already described in the previous section.

The second feature function implemented is a target language model. This feature actually consists of an n-gram model, in which the probability of a translation hypothesis is approximated by the product of word 3-gram probabilities:

$$p_{LM}(T_k) \approx \prod_{n=1}^{k} p(w_n|w_{n-2},w_{n-1}) \quad (3)$$

where T_k refers to the partial translation hypothesis and w_n to the n^{th} word in it.

This model is trained from the target side of the bilingual corpus by using the SRI Language Modeling toolkit and, again, the Kneser-Ney smoothing method.

An extended target language model might also be obtained by considering additional information from other available monolingual sources. These extended target language models are actually computed by performing a log-linear combination of independently computed target language models. The weights of the log-linear combination are adjusted so perplexity, with respect to a given development set, is minimized. In the particular case of the EPPS data used, no significant improvement was achieved by using extended target language models.

The third feature function corresponds to a word penalty model. This function introduces a sentence length penalization in order to compensate the system preference for short target sentences caused by the presence of the previous target language model. This penalization depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

$$wp(T_k) = exp(\text{number of words in } T_k) \quad (4)$$

where, again, T_k refers to the partial translation hypothesis.

The fourth and fifth feature functions correspond to a forward and backward lexicon models. These models provide IBM 1 translation probabilities for each tuple based on the IBM 1 lexical parameters $p(t|s)$ (Och *et al.*, 2004). These lexicon models are computed according to the following equation:

$$p_{IBM1}((t,s)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i|s_n^j) \quad (5)$$

where s_n^j and t_n^i are the j^{th} and i^{th} words in the source and target sides of tuple $(t,s)_n$, being J and I the corresponding total number words in each side of it.

For computing the forward lexicon model, IBM model 1 lexical parameters from GIZA++ source-to-target alignments are used. In the case of the backward lexicon model, GIZA++ target-to-source alignments are used instead.

3.3 N-gram Based Decoder

The search engine used for the presented translation system was developed by Crego *et al.*

(2005). This decoder, which takes into account simultaneously all the five feature functions described above, implements a beam-search strategy and allows for three different pruning methods:

- *Threshold pruning*: In this case, hypotheses with scores below a predetermined threshold value are eliminated.[2]
- *Histogram pruning*: In this case, the maximum number of competing hypotheses in a given stack is limited to the K-best ranked ones. For the case of the EPPS data, a value of $K = 50$ provided a good trade off between translation quality and computational expenses for both directions, English-to-Spanish and Spanish-to-English.
- *Hypothesis recombination*: In this case, a risk free pruning method (Koehn, 2004) for recombining hypotheses is implemented. At any step of the search, two or more hypotheses are recombined if they agree in both the present tuple and the tuple 3-gram history.

All the results presented here were obtained by using the monotonous search modality of the decoder (i.e. without including reordering capabilities).

4 EPPS Translation Tasks

This section presents and discusses translation results obtained for the EPPS tasks in both directions, English-to-Spanish and Spanish-to-English. First the EPPS data as well as the training procedure are described. Then some experimental results are presented and discussed.

4.1 European Parliament Data

The EPPS data set corresponds to the parliamentary session transcriptions of the European Parliament and is currently available at the Parliament's website (`http://www.europarl.eu.int/`). In the case of the results presented here, we have used the version of the EPPS data that was made available by RWTH Aachen University through the TC-STAR consortium[3]. The training and test data used included session transcriptions from April 1996 until September 2004, and from November 15th until November 18th, 2004, respectively.

Table 1 presents some basic statistics of both, training and test, data sets for each considered language: English (eng) and Spanish (spa). More specifically, the statistics presented in Table 1 are the total number of sentences, the total number of words, the vocabulary size (or total number of distinct words) and the average number of words per sentence.

1.a.- Train data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.22 M	33.4 M	105 k	27.3
Spa	1.22 M	34.8 M	169 k	28.4

1.b.- Test data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1094	26.8 k	3.9 k	24.5
Spa	840	22.7 k	4.0 k	27.0

Table 1: Basic statistics for the considered training (a) and test (b) data sets (M and k stands for millions and thousands, respectively).

It can be observed from Table 1 that although the total number of words in the training corpus is very similar for both languages, this is not the case for the vocabulary sizes. Indeed the Spanish vocabulary is 60% larger than the English vocabulary. This can be explained by the more inflected nature of Spanish vocabulary, which is particularly evident in the case of nouns, adjectives and verbs which may have many different forms depending on gender, number, tense and mode. As will be seen in the discussion section, this difference between vocabulary sizes has important consequences in translation quality when translating from English into Spanish.

Also from Table 1, it can be observed that a different test set was used for each translation direction. In the English test set, 0.4% of the total number of words (112 words) did not occur in the training data. From these 112 words, only 81 corresponded to different words. Similarly, in the Spanish test set, 0.2% of the words (46 words) were not in the training data. In this case, 40 different words occurred.

[2]This type of pruning, although available in the decoder, was actually not used in experiments presented in this work.

[3]TC-STAR (Technology and Corpora for Speech to Speech Translation) is an European Community project funded by the *Sixth Framework Programme*. More information can be found at the consortium website: `http://www.tc-star.org/`

4.2 Preprocessing and Training

The training data was preprocessed by using standard tools for tokenizing and filtering. In the filtering stage, sentence pairs with a word ratio larger than 2.4 have been removed, as well as sentence pairs with at least one sentence of more than 100 words in length.

Once the training data was preprocessed, word-to-word alignments were performed in both directions, source-to-target and target-to-source, by using GIZA++; and the intersection and union sets of both alignments were computed. For computing these alignments, five iterations for models IBM1 and HMM, and three iterations for models IBM3 and IBM4, were performed.

Then, a tuple set for each translation direction, Spanish-to-English and English-to-Spanish, was extracted from the union set of alignments and the resulting tuple vocabularies were pruned. Table 2 presents the total tuple vocabulary size, as well as tuple vocabulary sizes for two different pruning values $N = 30$ and $N = 20$. As already mentioned the best trade off between translation quality and computational expenses have been obtained for values of $N = 20$ and $N = 30$ for Spanish-to-English and English-to-Spanish, respectively. Notice from Table 2, that in both cases the resulting tuple vocabulary sizes are very similar.

Direction	Total	$N = 30$	$N = 20$
S → E	2.5 M	2.1 M	2.0 M
E → S	2.5 M	2.0 M	1.9 M

Table 2: Tuple vocabulary sizes for different pruning values (M stands for millions).

A tuple 3-gram translation model was trained for each translation direction by using the respective pruned tuple sets. Then, each model was enhanced by including the 1-gram probabilities for the embedded word tuples, which were extracted from the intersection set of alignments.

Table 3 shows the total number of n-grams contained in the models for each translation direction.

Similarly, the target language model and the forward and backward lexicon models were computed for each translation direction according to the procedures described in subsection 3.2.

Direction	1-grams	2-grams	3-grams
S → E	2.040 M	6.009 M	1.798 M
E → S	2.023 M	6.092 M	1.747 M

Table 3: Total number of n-grams for each translation direction (M stands for millions).

Once the models were computed, sets of optimal log-linear coefficients were estimated for each translation direction and system configuration according to the procedure described in subsection 3.1. As will be described in next section, four different system configurations were considered. For all these optimizations, a development data set of five hundred sentences was used. This data corresponded to parliamentary session transcriptions from October 21st until October 28th, 2004, so it was independent from both the training and the test data sets. The development data included three reference translations for both English and Spanish. Each optimization required to translate the development corpus between 20 and 100 times, depending on the system configuration.

Finally, the English and Spanish test data was translated into Spanish and English, respectively, by using the computed models and the estimated optimal coefficients for each system configuration and translation direction. The n-gram based decoding tool described in subsection 3.3 was used for generating the translations. The translation results are presented in the next section.

4.3 Task Results and Discussion

In order to evaluate the translation model performance and the feature contributions to the translation tasks, four different system configurations were considered in each translation direction:

- *Baseline System*: Only the tuple 3-gram translation model is used.
- *Target-reinforced System*: The translation model is used along with the target language and word penalty models.
- *Lexicon-reinforced System*: The translation model is used along with the forward and backward lexicon models.
- *Full System*: The translation model is used along with all the four additional feature functions.

The algorithms used for computing the evaluation measurements (mWER and BLEU) were the official TC-STAR evaluation tools distributed by ELDA (`http://www.elda.org/`). Two reference translations were available for each language test set.

Table 4 presents the mWER and BLEU values obtained for each system configuration and each translation direction.

Direction	System	mWER	BLEU
S → E	Baseline	39.55	0.476
	Target	39.61	0.485
	Lexicon	35.65	0.537
	Full	34.91	0.543
E → S	Baseline	44.45	0.428
	Target	44.67	0.436
	Lexicon	41.69	0.450
	Full	40.96	0.466

Table 4: Evaluation results for the Spanish-to-English and English-to-Spanish translation tasks.

As can be observed from Table 4, the inclusion of the four feature functions into the translation system definitively produces an important improvement in translation quality in both translation directions. Particularly, it becomes evident that features with the most impact on translation quality are the lexicon models. The target language model and the word penalty also contributed to improve translation quality, but in less degree.

Also, although it is more evident in the English-to-Spanish direction than in the opposite one, it can be noticed from the presented results that the contribution of target language and word penalty models is more relevant when the lexicon models are used (full system). In fact, as seen from λ_{LM} values in Table 5, when the lexicon models are not included, the target language model contribution to the overall translation system becomes significantly less important. This result suggests that including the lexicon models tends to favor short tuples over long ones, so the target language model becomes more important for providing target context information when the lexicon models are used. However more experimentation and research is required for fully understanding this interesting result.

Table 5 presents the optimized λ_i values[4] for each system configuration and each translation direction.

Direct.	System	λ_{LM}	λ_{wp}	λ_{fl}	λ_{bl}
S → E	Baseline	–	–	–	–
	Target	0.30	0.32	–	–
	Lexicon	–	–	0.48	0.06
	Full	0.48	0.28	0.48	0.13
E → S	Baseline	–	–	–	–
	Target	0.32	0.26	–	–
	Lexicon	–	–	0.17	0.07
	Full	0.80	0.75	0.23	0.18

Table 5: Optimized λ_i values for each system configuration and each translation direction.

Another important observation, which follows from comparing results between both translation directions, is that in all the cases Spanish-to-English translations are consistently and significantly better than English-to-Spanish translations. As already mentioned, this is clearly due to the more inflected nature of Spanish vocabulary. For example the single English word "the" can generate any of the four Spanish words "el", "la", "los" and "las". Similar situations occur with nouns, adjectives and verbs which may have many different forms in Spanish.

Additionally, a detailed review of about 130 translated sentence pairs, in each direction, was performed. This exercise resulted to be very useful since it allowed to identify the most common errors and problems related to the proposed SMT system in each translation direction. Three Spanish-to-English translation outputs are presented below for illustrative purposes. For each presented example, errors have been boldfaced and correct translations are provided in brackets:

- The policy of the European Union on Cuba **NULL must** [must not] change .
- To achieve these purposes , it is necessary NULL for the governments **to be allocated** [to allocate] , at least , 60 000 million NULL dollars a year ...
- In the UK we have **NULL** [already] **laws enough** [enough laws] , but we want to encourage NULL other States ...

[4]Since the log-linear weight of the tuple 3-gram model was fixed to unity, it has been omitted from the table.

A detailed analysis of all the reviewed translations revealed that most of translation problems encountered were basically related to the following four different types of errors:

- *Verbal Forms*: A great amount of wrong verbal tenses and auxiliary forms were detected. This problem turned out to be the most common one, reflecting the difficulty of the current statistical approach to capture the linguistic phenomena that shape head verbs, auxiliary verbs and pronouns into full verbal forms in each language.
- *Omitted Translations*: A large amount of translations involving tuples with NULL target sides were detected. Although in some cases these situations correspond to a correct translation, most of the time they result in omitted-word errors.
- *Reordering Problems*: The two specific situations that most commonly occurred were problems related to adjective-noun and subject-verb structures.
- *Concordance Problems*: Inconsistencies related to gender and number were the most commonly found.

5 Conclusions and Further Work

As can be noticed from the presented results, the tuple n-gram translation model, when used along with additional feature functions, provides state of the art translations for the two considered translation directions. Comparison between this translation system and a system that does not use the proposed bilingual n-gram based translation model has not been included here because of paper-length restrictions. However, a detailed comparison of the proposed system with other phrase-based translation systems is available through the TC-STAR consortium as a progress report (Ney *et al.*, 2005).

An important fact, that was also seen from the results, is that the quality obtained in the Spanish-to-English translation task is much better than the one obtained in the English-to-Spanish direction. According to this, significant efforts should be dedicated for properly exploiting morphological analysis and synthesis methods for improving English-to-Spanish translation quality.

Another interesting result was the evident important contribution of the lexicon models to the overall system performance. This result suggests further research is needed for studying the impact of these lexicon models at different stages of the translation process. In this way, a tuple-vocabulary pruning strategy, as well as a post-processing re-ranking procedure, based on these lexicon models seem to be very promising.

Additionally, four commonly-occurring types of translation errors were identified by reviewing a considerable amount of translated sentence pairs. This analysis has provided us with useful hints for future research and improvement of the proposed SMT system. However, more evaluation and discussion is required in this area for fully understanding such common translation failures and, then, implementing appropriate solutions.

In this way, our further research will focus on the following issues:

- The definition of the tuple, as a bilingual pair, will be revised in order to better handle the NULL occurrences in both the source and the target sides of it. As was mentioned in section 2, a better strategy for dealing with target words aligned to NULL is required. Similarly, a better handling of NULLs in the target side will result in less omitted-translation type of errors.
- Linguistic information must be used to cope with the observed morphology problems in the English-to-Spanish translation direction, as well as the more general problem of incorrect verbal form translations. In this sense, ongoing research on linguistic tuples classification is being done, in order to improve translation results. Preliminary results when detecting and classifying verb forms have been presented by de Gispert (2005).
- Reordering strategies, as well as non-monotonous decoding schemes, for the proposed SMT system must be developed and tested. As mentioned before, reordering problems specifically related to adjective-noun and subject-verb structures occur very often in Spanish-to-English and English-to-Spanish translations.

6 Acknowledgments

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, `http://www.tc-star.org`).

References

S. Bangalore and G. Riccardi. 2001. "A finite state approach to machine translation". *Proc. IEEE ASRU*, Madonna di Campiglio, Italy.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. "A maximum entropy approach to natural language processing". *Computational Linguistics*, 22(1):39–71.

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

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. "The mathematics of statistical machine translation: parameter estimation". *Computational Linguistics*, 19(2):263–311.

F. Casacuberta. 2001. "Finite-state transducers for speech input translation". *Proc. IEEE ASRU*, Madonna di Campiglio, Italy.

J.M. Crego, J.B. Mariño, and A. de Gispert. 2004. "Finite-state-based and phrase-based statistical machine translation". *Proc. of the 8th Int. Conf. on Spoken Language Processing*, :37–40, October.

J.M. Crego, J.B. Mariño, and A. de Gispert. 2005. "A Ngram-based Statistical Machine Translation Decoder". Submitted to INTERSPEECH 2005.

A. de Gispert and J.B. Mariño. 2002. "Using X-grams for speech-to-speech translation". *Proc. of the 7th Int. Conf. on Spoken Language Processing.*

A. de Gispert, J.B. Mariño, and J.M. Crego. 2004. "TALP: Xgram-based spoken language translation system". *Proc. of the Int. Workshop on Spoken Language Translation*, :85–90. Kyoto, Japan, October.

A. de Gispert. 2005. "Phrase linguistic classification and generalization for improving statistical machine translation". To be published at the *ACL'05 Student Workshop.*

R. Kneser and H. Ney. 1995. "Improved backing-off for m-gram language modeling". *IEEE Inte. Conf. on Acoustics, Speech and Signal Processing*, :49–52, Detroit, MI, May.

K. Knight. 1997. "Automating knowledge acquisition for machine translation". *AI Magazine*, 18(4).

P. Koehn, F.J. Och, and D. Marcu. 2003. "Statistical phrase-based translation". *Proc. of the 2003 Meeting of the North American chapter of the ACL*, Edmonton, Alberta.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based SMT. *Proc. of the Conf. of the Association for Machine Translation in the Americas (AMTA)*,

H. Ney, V. Steinbiss, R. Zens, E. Matusov, J. González, Y. Lee, S. Roukos, M. Federico, M. Kolss, and R. Banchs. 2005. "SLT progress report". *TC-STAR Deliverable D5,* European Community project no. FP6-506738, Available on line at: `http://www.tc-star.org/pages/f_documents.htm`

F.J. Och and H. Ney. 2000. "Improved statistical alignment models". *Proc. of the 38th Ann. Meeting of the ACL*, Hong Kong, China, October.

F.J. Och and H. Ney. 2002. "Discriminative training and maximum entropy models for statistical machine translation". *Proc. of the 40th Ann. Meeting of the ACL*, :295–302, Philadelphia, PA, July.

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. 2004. "A smorgasbord of features for statistical machine translation". *Proc. of the Human Language Technology Conf. NAACL*, :161–168, Boston, MA, May.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. "Bleu: a method for automatic evaluation of machine translation". *Proc. of the 40th Ann. Conf. of the ACL*, Philadelphia, PA, July.

W.H. Press, S. Teukolsky, W. Vetterling, and B. Flannery. 2002. *Numerical Recipes in C++: the Art of Scientific Computing*, Cambridge University Press.

G. Riccardi, R. Pieraccini, and E. Bocchieri. 1996. "Stochastic automata for language modeling". *Computer Speech and Language*, 10(4):265–293.

A. Stolcke. 2002. "SRLIM: an extensible language modeling toolkit". *Proc. of the Int. Conf. on Spoken Language Processing* :901–904, Denver, CO.

E. Vidal. 1997. "Finite-state speech-to-speech translation". *Proc. of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, :111–114, Munich, Germany.

R. Zens, F.J. Och, and H. Ney. 2002. "Phrase-based statistical machine translation". *25th German Conf. on Artificial Intelligence*, :18–32, Aachen, September. Springer Verlag.