UPC's Bilingual N-gram Translation System

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert
Patrik Lambert, José A. R. Fonollosa, Marta R. Costa-jussà, Maxim Khalilov

Department of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC), Barcelona 08034, Spain
{canton, rbanchs, jmcrego, agispert, lambert, adrian, mruiz, khalilov}@gps.tsc.upc.edu

Abstract

This paper describes the UPC's bilingual n-gram approach to statistical machine translation, which implements the log-linear combination of a bilingual n-gram translation model with six additional feature functions. A brief description of the complete system is presented and special attention is devoted to the novel features and reordering strategies that have been recently implemented. Translation results for the Spanish-to-English and English-to-Spanish tasks considered during the TC-STAR's second evaluation campaign are presented and discussed. Finally, improvements achieved in translation accuracy with respect to the previous year's system are also evaluated and discussed

1. Introduction

The UPC's statistical machine translation approach implements a log-linear combination of feature functions (Och and Ney, 2002) along with a translation model which is based on bilingual n-grams (de Gispert and Mariño, 2002). This translation model differs from the phrase-based translation approach (Koehn et al., 2003) in two basic issues: training data is monotonously segmented and the model considers n-gram probabilities instead of relative frequencies. The original version of the system (Mariño et al., 2005) implemented four feature functions along with the bilingual n-gram model. For this second evaluation campaign, some novel feature functions and reordering strategies that consider POS information have been implemented. A detailed description for each of these new features is presented in this work.

Translation results for three of the four tasks considered during TC-STAR's second evaluation campaign are also presented and discussed. More specifically, these tasks are: EPPS[1] Spanish-to-English, EPPS English-to-Spanish and CORTES[2] Spanish-to-English. For each of these tasks, three different translation conditions were considered: final text edition, verbatim transcriptions and automatic speech recognition.

The paper is structured as follows. Section 2. describes the bilingual n-gram translation model and section 3. presents a brief overview of the original system feature functions and implementation. Next, section 4. describes in detail the new feature functions and reordering strategies implemented. Section 5. presents and discusses the translation experiments and their results and, finally, section 6. presents some conclusions and further work.

2. UPC's Translation Model

The UPC's translation model has been derived from the finite-state approach; more specifically, from the work of Casacuberta (2001; 2004). However, different from it, where the translation model is implemented by using a finite-state transducer, the UPC's system implements a bilingual 5-gram model. It actually constitutes a language model of bilingual units, referred to as tuples, which approximates the joint probability between source and target languages by using 5-grams (de Gispert and Mariño, 2002), such as described by the following equation:

$$p(T,S) \approx \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, \ldots, (t,s)_{k-4}) \quad (1)$$

where t refers to target, s to source and $(t,s)_k$ to the k^{th} tuple of a given bilingual sentence pair.

Tuples are extracted from Viterbi alignments, which are automatically computed by using GIZA++ (Och, 2003), according to the following two constraints (Crego et al., 2004):

- tuple extraction should produce a monotonic segmentation of bilingual sentence pairs, and
- no smaller tuples can be extracted without violating the previous constraint.

According to this, tuples can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual corpus. Figure 1 presents a simple example illustrating the unique tuple segmentation for a given pair of sentences.

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 "translations" from figure 1. As seen from the figure, "translations" appears embedded into tuple 4. If a similar situation is encountered for all occurrences of "translations" in the training corpus, then no translation probability for an independent occurrence of such word will exist.

To overcome this problem, the tuple 5-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). While tu-

[1] European Parliament Plenary Sessions
[2] Spanish Parliament Speeches

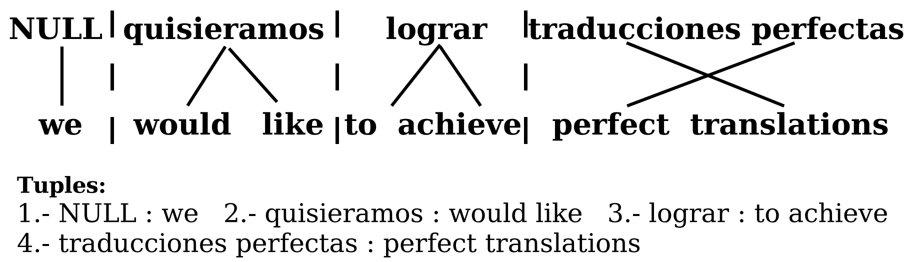

Figure 1: Example of tuple extraction.

ples are extracted from the union set of alignments computed in both directions, source-to-target and target-to-source, these embedded-word translation probabilities are computed from the intersection set of 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 1 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 according to a weight based on IBM model1 (Crego et al., 2005a). In this way, no target word remains linked to NULL, and tuples with NULL source side are not extracted.

3. Basic System Implementation

As already mentioned, the UPC's translation system implements a log linear combination of the bilingual n-gram translation system described in the previous section along with six additional feature functions. Four of these six models correspond to the same feature functions implemented in the original system version, which was used during TC-STAR's first evaluation campaign. A very brief description of this original basic system is presented here. A more detailed description can be found in Mariño *et al.* (2005).

The search engine for the UPC's translation system was developed by Crego *et al.* (2005b). It implements a beam-search strategy based on dynamic programming and allows for threshold pruning and hypothesis recombination. All implemented feature functions are simultaneously taken into account by the search engine during the decoding stage.

Additionally, an optimization tool, based on a simplex method (Press et al., 2002), was developed and used for weighting each feature function contribution. This algorithm adjusts the log-linear weights so that a non-linear combination of translation *BLEU* (Papineni et al., 2002) and *NIST* is maximized over the provided development set for each task under consideration.

3.1. Target Language Model

This feature provides information about the target language structure and fluency. It favors those partial-translation hypotheses which are more likely to constitute correctly structured target sentences over those which are not. The model is implemented by using a word 4-gram model of the target language, which is computed according to the following expression:

$$h_{TL}(T) = log \prod_{k=1}^{K} p(w_k|w_{k-1}, \ldots, w_{k-3}) \qquad (2)$$

where w_k refers to k^{th} word in the considered partial-translation hypothesis.

3.2. Word Bonus Model

This feature introduces a bonus which depends on the partial-translation hypothesis length. This is done in order to compensate the system preference for short translations over large ones. The model is implemented through a bonus factor which directly depends on the total number of words contained in the partial-translation hypothesis, and it is computed as follows:

$$h_{WP}(T) = M \qquad (3)$$

where M is the number of words contained in the partial-translation hypothesis.

3.3. Source-to-Target Lexicon Model

This feature actually constitutes a complementary translation model. This model provides, for a given tuple, a translation probability estimate between the source and target sides of it. This feature is implemented by using the IBM-1 lexical parameters (Brown et al., 1993; Och et al., 2004). According to this, the source-to-target lexicon probability is computed for each tuple according to the following equation:

$$h_{LF}(T,S) = log \frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} q(t_i^n|s_j^n) \qquad (4)$$

where s_j^n and t_i^n 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 of words in each side of it. In the equation $q(.)$ refers to IBM-1 lexical parameters which are estimated from alignments computed in the source-to-target direction.

3.4. Target-to-Source Lexicon Model

Similar to the previous feature, this feature function constitutes a complementary translation model too. It is computed exactly in the same way the previous model is, with the only difference that IBM-1 lexical parameters are estimated from alignments computed in the target-to-source direction instead.

4. Novel System Features

In addition to the bilingual n-gram translation model and the feature functions described in the previous section, this year's version of the UPC's translation system implements some novel feature functions and reordering strategies, which exploits morpho-syntactic information of either the source or the target language. These features are described in detail within this section.

Additionally, the incidence of each presented feature on translation accuracy is illustrated by comparing translation *BLEU* and *mWER* for an experimental EPPS test set, when the corresponding feature is and is not included. The test set used for the experiments presented in this section corresponds to a subset of the previous year's evaluation campaign test set. On the other hand, trainings and optimizations[3] were carried on by using this year's training and development data sets, which are described in detail in section 5.

4.1. Target POS-tag Language Model

This feature implements a 5-gram language model of target POS-tags. This model is trained by considering POS-tags, instead of words, for the target side of the training corpus. Accordingly, the tuple translation unit is redefined in terms of a triplet which includes: a source string containing the source side of the tuple, a target string containing the target side of the tuple, and a POS string containing the POS-tags corresponding to the words in the target strings.

It is important to mention that the POS information contained in the triplet is not actually used for computing the bilingual translation model probabilities described in section 2. This information is used only during decoding by the 5-gram language model of target POS-tags in order to score the alternative POS-tag sequences associated to the competing partial-translation hypothesis. This procedure is illustrated in figure 2.

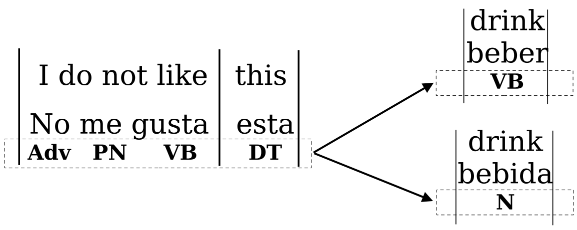

Figure 2: Augmented tuple and target POS-tag language model implementation.

[3]Notice that since independent optimizations were performed for each individual feature evaluation, all three experimental results presented in this section are not comparable among them.

Notice also that this feature does not require POS tagging of the output sentence, as the POS information is carried in the tuple. In this way, this feature helps, along with the target language model, to provide a better concatenation of tuples during the decoding process. Table 1 shows the incidence of the target POS-tag language model on translation accuracy for the experimental translation tasks.

Task	Target POS LM	BLEU	mWER
ES → EN	not included	0.541	34.98
	included	0.546	34.47
EN → ES	not included	0.471	40.22
	included	0.475	40.42

Table 1: *Incidence of the target POS-tag language model on translation accuracy.*

Notice from table 1 that a small improvement is clearly achieved for the Spanish-to-English direction when including the target POS-tag language model. However, for the English-to-Spanish translation direction, the effect is not clear at all.

4.2. Source POS-tag Language Model

This feature implements a word reordering strategy that is complemented by a 5-gram language model of reordered source POS-tags. As a first step, during training, reordering patterns for the source POS-tags are learned from the aligned corpus. These reordering patterns are identified by looking at the link crossings occurring in the aligned corpus and are classified according to the corresponding POS-tags of the source words involved.

During translation, the input sentence to be translated is replaced by a word-graph which includes alternative paths based on the POS reordering patterns learned during training. Then, the single sentence word-graph is augmented by adding as many paths as POS reordering patterns can be applied. This augmented word-graph is then used as the decoder's input. Notice that this procedure requires tagging the input sentence to be translated. This procedure is illustrated in figure 3, and further described in Crego (2006).

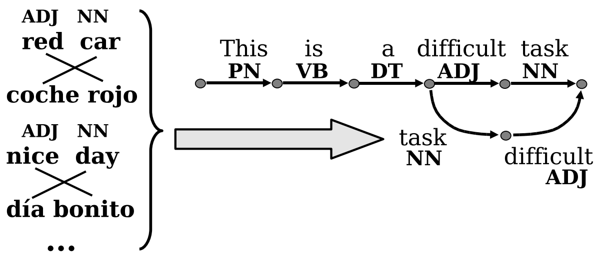

Figure 3: Example of word reordering based on source POS information.

As a second step, the reordering strategy illustrated in figure 3 is further supported by a 5-gram language model of reordered source POS-tags. In this sense, all link crossings identified during the reordering-pattern identification phase are unfolded by reordering the source words and their corresponding POS-tags while keeping the target side of the

corpus untouched. From this reordered sequence of POS-tags, a 5-gram language model is trained. During decoding, the 5-gram language model of reordered source POS-tags is used to help the decoder to select those more appropriate paths among all possible paths in the input word-graph.
Table 2 shows the incidence of the augmented word-graph input and its supporting source POS-tag language model on translation accuracy for the experimental translation tasks under consideration.

Task	Graph	POS LM	BLEU	mWER
ES → EN	no	no	0.552	34.40
	yes	no	0.556	34.23
	yes	yes	0.564	33.75
EN → ES	no	no	0.480	41.18
	yes	no	0.485	41.15
	yes	yes	0.489	40.29

Table 2: *Incidence of the augmented word-graph input and the supporting source POS-tag language model on translation accuracy.*

Notice from table 2 that some interesting improvements are clearly achieved for both translation directions. Additionally, it is obvious from the results that the inclusion of the source POS-tag language model during decoding further increases the contribution of the input word-graph.

4.3. Word Reordering Based on Alignment Block Classification

This constitutes an alternative word reordering strategy, which is implemented as a preprocessing stage instead of as while-decoding feature function. A detailed description for this method is provided in Costa-jussà and Fonollosa (2006). This strategy is intended to infer those most probable re-orderings for sequences of words, which are referred to as blocks, in order to achieve monotonic alignments from current data alignments, and generalize re-orderings for unseen pairs of blocks.
Given a word alignment, a list of alignment blocks is extracted. An alignment block is defined as a pair of consecutive source blocks for which their corresponding target translations appear in the opposite order. Then, by using a classification algorithm, the list is processed in order to decide whether two consecutive blocks have to be reordered or not. This classification algorithm should be able to cope with swapping examples seen during training, as well as infer properties that might allow reordering of pairs of blocks not seen together during training. So, the algorithm is implemented in two steps:

- In the first phase, the algorithm filters the constructed list of alignment blocks from possible erroneous alignments or ambiguities. Two basic criteria are used: pairs appearing less than N_{min} times are discarded, and pairs of blocks with swapping probability[4] less than a threshold are also discarded.
- In the second phase, the algorithm infers generalization groups from the filtered list of alignment blocks. This is done by grouping together all alignment blocks that have one side (source or target) in common with at least any other alignment block within the same group.

Based on the information provided by the constructed generalization groups, the source sides of both bilingual training and development corpora, as well as the test data, are reordered. This modified training corpus is aligned once again and the translation model and feature function probabilities are computed from these new alignments. Finally, translations are computed from the modified input test data.
The proposed word reordering strategy happens to provide improvements in both alignment quality and translation accuracy. This is shown in table 3, were alignment quality is presented in terms of *AER* and translation accuracy in terms *BLEU* and *mWER*. This methodology has been only implemented for the Spanish-to-English direction.

Task	Reordering	AER	BLEU	mWER
ES → EN	no	20.64	0.552	34.46
	yes	19.36	0.562	33.68

Table 3: *Incidence of word reordering based on alignment block classification in both alignment quality and translation accuracy.*

5. TC-STAR Second Evaluation Results

The data sets used for experiments presented in these section correspond to those provided by ELDA for the 2006 TC-STAR[5] Second Evaluation Campaign, which are available through the ELDA's website at: `http://www.elda.org/tcstar-workshop/2006eval.htm`.
Results for three different tasks are presented here: English-to-Spanish (EPPS), Spanish-to-English (EPPS) and Spanish-to-English (CORTES). For each of these tasks, three different translation conditions were considered: final text edition, verbatim transcriptions, and automatic speech recognition (ASR).
The final text edition condition corresponds to the official transcripts of the respective parliament sessions, so it is actually a written language translation condition. On the other hand, the other two conditions are spoken language translation conditions. More specifically, the verbatim condition corresponds to literal transcriptions of parliamentary speeches, which include hesitations, repeated words and other spontaneous speech effects; and the ASR output condition corresponds to the output of an automatic speech recognition system, so it additionally includes speech recognition errors.
Table 4 presents basic statistics for the training data, which existed only for the case of the EPPS tasks. The CORTES Spanish-to-English task was performed by using the EPPS

[4]The swapping probability is defined as the ratio between the number of times that two blocks are swapped and the total number of times the same two blocks appear consecutive.

[5]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/`

training corpus. More specifically, the table presents the total number of sentences, the total number of running words and the vocabulary size for each language. This training corpus was aligned by using the GIZA++ alignment tool (Och, 2003).

Language	Sentences	Words	Vocabulary
English	1.28	34.9	0.106
Spanish	1.28	36.6	0.153

Table 4: *Basic statistics for the EPPS training data. The total number of sentences, words and the vocabulary size are provided in millions.*

Table 5, on the other hand, presents the total number of sentences contained in each condition's development and test data set, as well as the number of references available for computing the automatic error and accuracy measures.

Task	Data	Cond.	Sents.	Ref.
EN→ES EPPS	Dev.	FTE	735	2
		VBT	1194	2
		ASR	863	2
	Test	FTE	1117	2
		VBT	1155	2
		ASR	894	2
ES→EN EPPS	Dev.	FTE	430	2
		VBT	440	2
		ASR	440	2
	Test	FTE	894	2
		VBT	897	2
		ASR	1092	2
ES→EN CORTES	Dev.	FTE	380	2
		VBT	460	2
		ASR	460	2
	Test	FTE	888	2
		VBT	699	2
		ASR	1133	2

Table 5: *Total number of sentences, and available number of references, for each condition's development and test data set.*

Table 6 presents the *BLEU*, *NIST* and *mWER* scores obtained for the test data of each translation task and condition. For all results presented in table 6 for the ES→EN direction, the reordering strategy described in subsection 4.3. was implemented. In the EN→ES direction, no reordering strategy was used, except for the result identified as FTE⋆, for which the reordering strategy described in subsection 4.2. was used.
Notice from table 6, that a better performance is achieved for the EPPS task when translating from Spanish-to-English than when translating from English-to-Spanish. This is clearly due to the more inflected nature of Spanish vocabulary, which makes more difficult the translation in the English-to-Spanish direction. Notice also the significant difference in translation quality between the Spanish-to-English EPPS and CORTES tasks. This result makes a lot of sense since the training data was from the EPPS corpus.

Task	Cond.	BLEU	NIST	mWER
EN→ES EPPS	FTE	0.482	9.999	40.89
	FTE⋆	0.488	10.06	40.21
	VBT	0.440	9.500	44.66
	ASR	0.347	8.557	51.78
ES→EN EPPS	FTE	0.552	10.60	36.94
	VBT	0.520	10.61	38.84
	ASR	0.383	9.142	48.66
ES→EN CORTES	FTE	0.415	9.207	47.05
	VBT	0.446	9.640	45.18
	ASR	0.298	7.995	55.87

Table 6: BLEU, NIST *and* mWER *scores obtained for the test data of each translation task and condition. All* BLEU *scores are case sensitive and use as many translation references as described in table 5.*

Another important observation from table 6 is that translation accuracy deteriorates when moving from text translations (FTE) to speech translations (VBT and ASR). Nevertheless, in the particular case of the CORTES task, the best result was obtained for the verbatim condition. A more detailed evaluation of translation outputs is required.
Regarding the EPPS English-to-Spanish, for which the reordering strategy presented in subsection 4.2. was tested, it can be seen from table 6 that the FTE⋆ experiment performed slightly better than the other according to all evaluation metrics.
Finally, we also evaluate how much improvement, with respect to the previous year's system, has been achieved. In this sense, table 7 presents translation results for the test data used in the TC-STAR's First Evaluation Campaign by using the previous year's system (Mariño et al., 2005) and the current one. Only the final text edition condition for the EPPS task is presented.

Task	System	BLEU	NIST	mWER
EN→ES	2005	0.456	9.657	41.20
	2006	0.486	9.880	40.66
ES→EN	2005	0.524	10.55	35.11
	2006	0.555	10.79	33.68

Table 7: *Translation results for the 2005 test data by using the previous year's and current systems. Evaluation measures are case sensitive and were computed by using 2006's evaluation scripts.*

As seen from the table, performance has been improved in about 0.03 *BLEU* marks (in the 0 to 1 scale) for both translation directions.

6. Conclusions and Further Work

As can be concluded from the presented results, the current system proved to provide better results than the previous year one. However, Spanish-to-English translations continue to be significantly better than English-to-Spanish

translations, text translations better than speech translations, and morphology and reordering continue to be important problems. Although all the new features and reordering strategies presented happened to provide small, but clearly observable, improvements in translation accuracy, there is still a lot of work to be done in order to better exploit morpho-syntactic information for statistical machine translation purposes.
In this sense, further research should focus on:

- Reordering strategies, as well as non-monotonous decoding schemes.
- Using more complete and diverse linguistic information sources.
- Specific preprocessing strategies for verbatim and ASR output data.

7. Acknowledgments

This work was partly 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`), the European Social Fund and the Spanish Ministry of Education and Science.

8. References

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

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

F. Casacuberta. 2001. Finite-state transducers for speech input translation. In *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. In *Proc. of the 8th Int. Conf. on Spoken Language Processing*, pages 37–40, October.

J.M. Crego, A. de Gispert, and J.B. Mariño. 2005a. The talp ngram-based smt system for iwslt'05. In *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT 2005*, Pittsburgh (PA), October.

J.M. Crego, J.B. Mariño, and A. de Gispert. 2005b. A ngram-based statistical machine translation decoder. In *INTERSPEECH 2005*, Lisbon, September.

J.M. Crego. 2006. A reordering framework for statitstical machine translation. Internal report, Universitat Politecnica de Catalunya.

A. de Gispert and J.B. Mariño. 2002. Using x-grams for speech-to-speech translation. In *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. In *Proc. of the Int. Workshop on Spoken Language Translation*, pages 85–90, Kyoto, Japan, October.

M.R. Costa jussà and J.A.R. Fonollosa. 2006. Using reordering in statistical machine translation based on alignment block classification. Internal report, Universitat Politecnica de Catalunya.

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

J.B. Mariño, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M. Ruiz. 2005. Bilingual n-gram statistical machine translation. In *Proc. of the tenth Machine Translation Summit*, pages 275–282, Phuket, Thailand, September.

F.J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Ann. Meeting of the ACL*, pages 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. In *Proc. of the Human Language Technology Conf. NAACL*, pages 161–168, Boston, MA, May.

F.J. Och. 2003. Giza++ software. Internal report, RWTH Aachen University, `http://www.i6.informatik.rwth-aachen.de/`.

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

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