Edinburgh System Description for the 2006 TC-STAR Spoken Language Translation Evaluation

Abhishek Arun, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch Hieu Hoang, Philipp Koehn, Miles Osborne, David Talbot

School of Informatics
University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland, UK
{a.arun,a.axelrod-2,a.c.birch-mayne,c.callison-burch,h.hoang,d.r.talbot}@sms.ed.ac.uk
{miles,pkoehn}@inf.ed.ac.uk

Abstract

In this paper we describe the Edinburgh University statistical machine translation system, as used for the TC-STAR 2006 evaluation campaign. We participated in the primary Final Text Edition track for the Spanish to English and English to Spanish translation tasks, using only the provided datasets for training our translation and language models. We obtained the highest WNM/Recall score in both language pairs and had competitive results for all other evaluation metrics.

1. Introduction

This document describes the first TC-STAR Spoken Language Translation submission from the University of Edinburgh's Statistical Machine Translation group. We participated in the primary track with text data input provided by the European Parliament (Final Text Editions), using only the resources supplied on the evaluation campaign website.

Our MT system was originally developed for translation of European parliament texts from German to English (Koehn et al., 2003). We have previously extended our system to work on the DARPA challenges on Chinese and Arabic (Koehn, 2004a; Koehn et al., 2005b), as well as on speech data in Asian languages (Koehn et al., 2005a). Although we were limited to only participating in the Spanish to English and the English to Spanish translation tasks because of time constraints, we welcomed the chance to work with another European language pair.

The next section of this paper provides a brief overview of our phrase-based translation system in its out-of-the-box form. We then present two new additions to our standard system, namely a recaser and the ability to use higher-order ngram language models during decoding. In section 3., we describe the experiments run to optimise key components of our system, particularly the selection of reordering limits and language models. Lastly, we report our results on the TC-STAR translation evaluation and provide some analysis.

2. System Description

Our system employs a phrase-based statistical machine translation model (Koehn et al., 2003), implemented by the Pharaoh decoder (Koehn, 2004b).

In phrase-based SMT models, the input ("foreign") sentence is segmented into so-called phrases, which are sequences of adjacent words that are not necessarily linguistically motivated. Each foreign phrase is mapped into the target language ("English"). Phrases are allowed to be reordered during translation; see Figure 1 for an illustration.

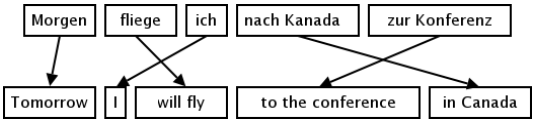

Figure 1: Phrase-Based SMT: Input sentence is segmented into phrases, which are then mapped onto output phrases.

2.1. Log-Linear Phrase-Based Model

Mathematically, our machine translation system employs a log linear approach to search for the most probable English output sentence $\mathbf{e}$ given some foreign input sentence $\mathbf{f}$.

The Pharaoh decoder selects the most likely translation by maximising the sum of probabilities over a set of feature functions $h_m(\mathbf{e}, \mathbf{f})$ that are scaled by weights λ_m:

$$\begin{aligned} \hat{\mathbf{e}} &= \arg\max_e p(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_e \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \end{aligned}$$

The log linear model provides a natural framework to integrate many components and to weight them relative to each other based on their performance. Our standard phrase-based MT system uses the following feature functions:

- phrase translation probability (in both directions)
- lexical translation probability (in both directions)
- word penalty
- phrase penalty
- language model score
- linear reordering penalty
- lexicalised reordering weight

The weights used to scale the feature functions are found via Minimum Error Rate Training, using a method suggested by Och (2003). This parameter training was performed using a small held-out development set, and using

the BLEU score of the system (Papineni et al., 2002) as the optimisation metric.

2.1.1. Phrase and Lexical Translation Probability Features

The most important component of the system is the phrase translation probability table. To create the phrase translation table, we extracted phrase pairs from the training corpus by first aligning the words in the corpus and extracting phrase pairs that are consistent with the word alignment. We then assign probabilities to the obtained phrase translations. By now, inducing phrase-based translation models from word-level alignments is common practice in SMT.

We obtained word alignments by using the GIZA++ toolkit (Och and Ney, 2003) on the training corpus in both translation directions. The two sets of alignments were then symmetrised using the **grow-diag-final** method previously described in Koehn et al. (2005a). This particular method of symmetrising — called the *refined method* (Och and Ney, 2003) — overcomes the inability of the IBM Models implemented in GIZA++ to map one target (English) word to multiple source (foreign) words.

Next, we collected phrase pairs that were consistent with the word-level alignments that were extracted. We define a *consistent* phrase pair as one where the words in the phrase pair are aligned to only with each other, and no words outside of the phrase pair are aligned to any words in the phrase pair. The extracted phrase pairs were assigned probabilities by unsmoothed relative frequency, and the translation probabilities were lexically weighted as in (Koehn et al., 2003).

2.1.2. The Word and Phrase Penalties

The word and phrase penalties simply add a constant factor for each word or phrase generated, to bias the model towards shorter output.

2.1.3. The Reordering Model Features

The basic reordering model only considers the linear distance that a phrase needs to be moved in order to align with its translation. This movement distance is measured on the foreign side. The linear reordering penalty simply adds a cost factor, δ^n, for all movements over n words.

Our system also includes a lexicalised reordering model (Koehn et al., 2005a) as a feature. For each phrase pair, we learn how likely it is to either directly follow the previous phrase (monotone), to swap positions with a previous phrase (swap), or to not connect to the previous phrase at all (discontinuous). These three types of reordering are illustrated in Figure 2.

Reordering is modeled in a bidirectional manner, taking into account both the previous and the next translated phrases. The phrase pairs are classified by reordering type during extraction, based on their alignments within the sentence grid:

- monotone: a word alignment point to the top left exists
- swap: an alignment point to the top right exists
- discontinuous: no alignment points to the top left nor top right

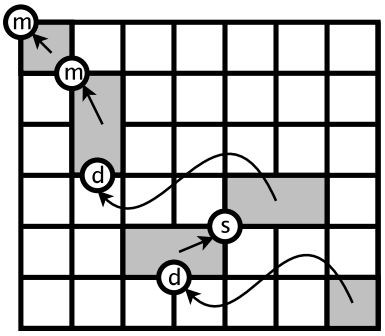

Figure 2: Possible orientations of phrases: monotone (m), swap (s), or discontinuous (d)

By smoothing the counts from classifying the phrase pairs by reordering type, we can estimate orientation probability distributions:

$$p_r(\text{orientation}|\bar{e}, \bar{f})$$

2.1.4. Language Model Score

During decoding, the candidate translation is created from left to right. As target translation hypotheses are created in our baseline system, their language model score is computed by conditioning on the two previous target words already generated. This is the standard trigram language model.

For example:

$$\begin{aligned} p(Mary\ did\ not\ slap) = \quad & p(Mary|START, START)\times \\ & p(did|Mary, START)\times \\ & p(not|Mary\ did)\times \\ & p(slap|did\ not) \end{aligned}$$

One of the recent additions to our system is the ability to use higher-order language models during decoding, and this TC-STAR translation task presented us with the opportunity to test it. For example, when our system is run using a 4-gram language model, the phrase in Equation 2.1.4. is scored as:

$$\begin{aligned} p(Mary\ did\ not\ slap) = \quad & p(Mary|START)\times \\ & p(did|Mary, START)\times \\ & p(not|START\ Mary\ did)\times \\ & p(slap|Mary\ did\ not) \end{aligned}$$

We used the SRI Language Modeling toolkit (Stolcke, 2002) to train smoothed 3-gram, 4-gram (without 4gram singletons) and 5-gram (without 4gram and 5gram singletons) language models for both Spanish and English on the respective monolingual training datasets provided.

2.2. Recaser

Our standard translation system is trained on and produces lowercased text. As the TC-STAR evaluation was on original-cased data, we implemented a recaser to capitalise our output translations.

The recaser is a log-linear translation system with only two component features, a translation model and a language model. The decoding task then is to find the most probable

original-cased sentence $\mathbf{E}$ that can be made from a lowercased sentence $\mathbf{e}$.

$$\begin{aligned}\hat{\mathbf{E}} &= \arg\max_{E} p(\mathbf{E}|\mathbf{e}) \\ &= \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{E}, \mathbf{e})\end{aligned}$$

We trained the recaser's translation model on a parallel original-cased and lowercased version of the provided target-language dataset. Word alignments were trivial to extract, as each word in the original-case sentence is mapped to the corresponding word in the lowercased sentence. The recaser's phrase table thus consists only of single-word phrases, and the phrase translation probabilities are learned by maximum likelihood estimation (MLE). For the recaser's language model, we simply trained a smoothed trigram language model on the original-case dataset using the SRILM toolkit.
As this was a very simple model, the scaling weights of the features in the recasing system were set by hand. Reordering was not allowed, so decoding was monotonic.

3. Experiments

We used the EPPS English/Spanish final text edition (FTE) files spanning May 2004-Jan 2005 and Apr 1996-Jan 2005, provided by ELDA, as training datasets. While the language models were trained using all of the 1,304,054 sentences in the training datasets, in order to reduce training time for the translation model, sentences more than 40 words long were excluded.
Training the translation model with lexicalised reordering table on the remaining 990,214 sentences took around 45 hours on an Intel Xeon 2.80GHz Linux machine with 5GB of RAM.
Once training was complete, we ran a series of experiments to tune our system. Our main goal was to find the most effective combination of ngram language model and reordering limit for the TC-STAR'06 evaluation task. We used the provided development set for minimum error rate training, and the 2005 test set as our testing/development set. Our post-processing consisted of dropping unknown words from the decoder output, and then passing the lowercased system output through the recaser described in section 2.2.. In this section, we report our system's BLEU scores for both case sensitive and case insensitive evaluation using the provided evaluation script.

3.1. Selecting The Language Model

A priori, we would expect translation quality to improve as we use higher order ngram language models. However, because of data sparseness, we might not reliably estimate the probabilities for the higher order ngrams.
Table 1 presents the number of distinct ngrams in the training data for each language, which might give an indication as to possible data sparseness issues. We note that there are more distinct ngrams in Spanish than in English, as Spanish is a morphologically richer language.

	1gram	2gram	3gram	4gram	5gram
English	113k	2,281k	9,540k	18,395k	24,181k
Spanish	156k	2,607k	9,899k	18,877k	24,732k

Table 1: Number of distinct ngrams in the training corpus.

	Reordering Limit						
	3	4	5	6	7	8	9
LM							
	Case insensitive						
3gram	55.3	56.5	56.4	57.0	56.4	55.6	56.1
4gram	56.1	55.9	55.8	56.5	56.5	56.7	56.0
5gram	56.7	56.6	57.1	**57.3**	56.1	56.5	57.1
	Case sensitive						
3gram	53.3	54.6	54.4	55.1	54.5	53.6	54.2
4gram	54.7	54.0	53.9	54.5	54.6	54.8	54.0
5gram	54.8	54.8	55.3	**55.4**	54.2	54.6	55.1

Table 2: Optimising the reordering limit (maximum word distance for phrase movement) and language model (LM) for Spanish to English translation. The table shows both case sensitive and case insensitive BLEU scores.

Nonetheless, conventional wisdom is that higher-order models are more useful in translation than trigrams, especially for a corpus this size. It is argued that perhaps rescoring with this larger n-gram model is sufficient, as it is significantly faster than the same language model integrated into the decoder. We cannot directly compare the use of a higher-order n-gram language model within the decoding step to using a trigram for decoding and a 5-gram for rescoring. This simply is because we have not implemented a rescoring step; all our work to date has centered on improving the decoder's capabilities.

3.2. Optimising Reordering Distance Limit

Reordering is measured by the movement of foreign phrases during translation. If while producing a monotone English sequence of words we translate the first foreign word, and then continue with the fifth foreign word, we measure this as a movement over three words (the intermediate foreign words numbered 2,3, and 4 are skipped). We mentioned in section 2.1.3. that our system can reorder phrases depending on their lexical context, but even then there is a limitation on the maximum allowable reordering movement.
Ideally, we would allow reordering of any distance, as movements over long distance do occur when translating. However, our previous experience has shown that even the lexicalised reordering model is not strong enough to correctly guide long distance movements. Nevertheless, we wanted to carry out experiments with differing reordering limits, especially as Spanish and English do not tend to reorder as heavily as other language pairs during translation.

3.3. Model Optimisation Results

The results of the optimisation experiments are shown in tables Table 2 and 3. For both translation directions, there does not seem to be a general trend as we optimise on re-

	Reordering Limit						
	3	4	5	6	7	8	9
LM							
	Case insensitive						
3gram	49.8	50.1	50.3	49.8	49.2	50.1	49.3
4gram	50.0	50.6	**50.8**	49.9	49.9	50.5	50.6
5gram	50.2	50.3	48.8	50.2	50.0	50.7	50.4
	Case sensitive						
3gram	48.4	48.8	49.0	48.5	47.9	48.8	48.0
4gram	48.7	49.2	**49.5**	48.6	48.6	49.2	49.4
5gram	48.9	49.0	47.8	48.9	48.8	49.4	49.1

Table 3: Optimising the reordering limit (maximum word distance for phrase movement) and language model (LM) for English to Spanish translation. The table shows both case sensitive and case insensitive BLEU scores.

Src	Asimismo, debatimos sobre el tema de **"Comunicar Europa"**
Ref	Likewise, we debated the subject **"Communicating Europe"**
Out	Furthermore, we are debating on the theme of *"communicating Europe"*
Src	De conformidad con el orden del da, se procede al debate sobre el **Informe del Consejo Europeo**
Ref	In accordance with the agenda, we will proceed with the debate about the **Report by the European Council**
Out	The next item is the debate on the *report of the European Council*

Table 4: Errors made by monolingual recaser that could be corrected using a bilingual recaser. *Src* refers to the source sentence, *Ref* is the reference target sentence and *Out* is our system's output

ordering limit and language model. 4gram and 5gram language models, in most cases, seem to give better results than the 3gram ones, but the improvements are not statistically significant and not consistent. For English to Spanish translation, we obtain the best result using a 4-gram language model with a reordering limit of 5, while from Spanish to English, the best result is obtained with a 5-gram language model and a reordering limit of 6. It is possible that the 5gram Spanish language model is affected by data sparseness as noted in Section 3.1..

While we are aware of the limited statistical significance of these results, we decided to use the aforementioned settings to translate the evaluation dataset for the TC-STAR'06 task. It is worth mentioning that larger ngram language models and higher reordering limits significantly slow down decoding. This is because higher order ngrams reduce the decoder's ability to recombine and thus reduce the number of hypotheses, and because higher reordering limits increase the decoder's search space exponentially.

3.4. Recaser Results

We find that the difference between case sensitive and case insensitive BLEU scores on the recased data is only around 1.5 to 2% across the board, which shows that our recaser is performing reasonably.

For this particular language pair, one of the ways in which the performance of the recaser could be further improved is by training our phrase table with a corpus of truecased data instead of original-cased data. In truecasing, the first letter of the first word of each sentence (unless it is a fully capitalized word) is lowercased. Using a truecased corpus, we would expect our phrase table to be less sparse and learn more accurate phrase translation probabilities. The recaser output could then be original-cased in a deterministic post-processing step.

Furthermore, we had set the weights of our recaser model features by hand, due to time constraints. In the future, we will experiment with setting the scaling weights using minimum error rate training.

Some of the casing errors are caused by ambiguity in the data. For example, the training data has 40 occurences of the phrase "Berlin wall" and 150 occurences of "Berlin Wall", such that our recaser chooses the latter phrase as the correct cased translation of "berlin wall". However, in the TC-STAR'05 test set that we used to run our experiments and on which we report results, the phrase is always written as "Berlin wall".

Our casing model is trained using only monolingual data, but future work will incorporate the recent findings of Wang et al. (2006). They show that recasing accuracy can be significantly improved using a bilingual model that exploits case information from both source and target sentences, because MT output usually strongly preserves case from the input.

Such a model is estimated using features defined over both source and target texts. For example, one of the features Wang et al. (2006) introduce is an **upper-case translation feature** which is true when both the source and target phrases are in upper case. This feature aims to capture the idea that if a source word is cased, then the target word should be cased too, even if the word pair has not been seen in training data.

Table 4 shows some examples where a bilingual model with knowledge of the source sentence case information could have prevented case errors.

4. TC-STAR 2006 Evaluation Results

Our results on the final text edition task of the primary data track for English-Spanish and Spanish-English translations are shown in Table 5 and 6. We are very satisfied with our system's performance, given the corpus limitations of our selected track, and that we only had two weeks to train our translation and language models.

Our English-Spanish submission ranked significantly higher than our Spanish-English submission on most metrics. Our lower placement in Spanish-English may be due to only using a language model trained on the provided corpus, instead of taking advantage of the available larger monolingual English corpora. However, both of our submissions were ranked in first place on the WNM/Recall metric, which has been shown to strongly correlate with human adequacy(Babych and Hartley, 2004).

Language Pair	NIST cs*	BLEU cs*	NIST	BLEU	IBM	mWER cs*	mWER	mPER cs*	mPER	WNM/ Recall	WNM/ F-Measure
En-Sp	10.1132	0.4950	10.2003	0.5051	0.4942	39.6918	38.9415	30.5065	29.4770	0.4913	0.5099
Sp-En	10.1137	0.4559	10.3341	0.4724	0.4560	43.7428	42.7211	31.6688	30.1911	0.7470	0.7103

Table 5: Official results: The scores for our final text edition primary track submission to the TC-STAR'06 Evaluation Campaign (cs* = case sensitive).

Language Pair	NIST cs*	BLEU cs*	NIST	BLEU	IBM	mWER cs*	mWER	mPER cs*	mPER	WNM/ Recall	WNM/ F-Measure
En-Sp	6	4	5	4	3	4	4	5	4	1	5
Sp-En	14	13	13	11	12	13	13	13	11	1	7

Table 6: Official results: The rank among participants of our final text edition primary track submission to the TC-STAR'06 Evaluation Campaign (cs* = case sensitive).

5. Conclusions

In this paper, we outlined Edinburgh University's phrase-based statistical machine translation system. We presented two new additions to the system: a recaser, and support for higher-order ngram language models. Both of these extensions were tested on the Spanish-English and English-Spanish translation tasks, and each of these showed improvements over the baseline.

We obtained the best results on the TC-STAR'05 test data using a 4gram language model with a reordering limit of 5 for English-Spanish and using a 5gram language model with reordering limit of 6 for Spanish-English.

6. Acknowledgement

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

7. References

Bogdan Babych and Tony Hartley. 2004. Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 621–628, Barcelona, Spain, July.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005a. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *International Workshop on Spoken Language Translation 2005*.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. 2005b. Edinburgh system description for the 2005 NIST MT evaluation. In *Proceedings of Machine Translation Evaluation Workshop 2005*.

Philipp Koehn. 2004a. The foundation for statistical machine translation at MIT. In *Proceedings of Machine Translation Evaluation Workshop 2004*.

Philipp Koehn. 2004b. Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas, AMTA*, Lecture Notes in Computer Science. Springer.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*.

Wei Wang, Kevin Knight, and Daniel Marcu. 2006. Capitalizing machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.