The ISL Phrase-Based MT System for the 2007 ACL Workshop on Statistical Machine Translation

M. Paulik[1,2], K. Rottmann[2], J. Niehues[2], S. Hildebrand [1,2] and S. Vogel[1]
[1]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Institut für Theoretische Informatik, Universität Karlsruhe (TH), Karlsruhe, Germany
{paulik|silja|vogel}@cs.cmu.edu;{jniehues|rottmann}@ira.uka.de

Abstract

In this paper we describe the Interactive Systems Laboratories (ISL) phrase-based machine translation system used in the shared task "Machine Translation for European Languages" of the ACL 2007 Workshop on Statistical Machine Translation. We present results for a system combination of the ISL syntax-augmented MT system and the ISL phrase-based system by combining and rescoring the n-best lists of the two systems. We also investigate the combination of two of our phrase-based systems translating from different source languages, namely Spanish and German, into their common target language, English.

1 Introduction

The shared task of the ACL 2007 Workshop on Statistical Machine Translation focuses on the automatic translation of European language pairs. The workshop provides common training sets for translation model training and language model training to allow for easy comparison of results between the participants.

Interactive Systems Laboratories participated in the English ↔ Spanish Europarl and News Commentary task as well as in the English ↔ German Europarl task. This paper describes the phrase-based machine translation (MT) system that was applied to these tasks. We also investigate the feasibility of combining the ISL syntax-augmented MT system (Zollmann et al., 2007) with our phrase-based system by combining and rescoring the n-best lists produced by both systems for the Spanish → English Europarl task. Furthermore, we apply the same combination technique to combine two of our phrase-based systems that operate on different source languages (Spanish and German), but share the same target language (English).

The paper is organized as follows. In section 2 we give a general description of our phrase-based statistical machine translation system. Section 3 gives an overview of the data and of the final systems used for the English ↔ Spanish Europarl and News Commentary tasks, along with corresponding performance numbers. Section 4 shows the data, final systems and results for the English ↔ German Europarl task. In Section 5, we present our experiments involving a combination of the syntax-augmented MT system with the phrase-based MT system and a combination of the Spanish → English and German → English phrase-based systems.

2 The ISL Phrase-Based MT System

2.1 Word and Phrase Alignment

Phrase-to-phrase translation pairs are extracted by training IBM Model-4 word alignments in both directions, using the GIZA++ toolkit (Och and Ney, 2000), and then extracting phrase pair candidates which are consistent with these alignments, starting from the intersection of both alignments. This is done with the help of phrase model training code provided by University of Edinburgh during the NAACL 2006 Workshop on Statistical Machine Translation (Koehn and Monz, 2006). The raw rel-

Proceedings of the Second Workshop on Statistical Machine Translation, pages 197–202,
Prague, June 2007.

ative frequency estimates found in the phrase translation tables are then smoothed by applying modified Kneser-Ney discounting as explained in (Foster et al., 2006). The resulting phrase translation tables are pruned by using the combined translation model score as determined by Minimum Error Rate (MER) optimization on the development set.

2.2 Word Reordering

We apply a part-of-speech (POS) based reordering scheme (J. M. Crego et al., 2006) to the POS-tagged source sentences before decoding. For this, we use the GIZA++ alignments and the POS-tagged source side of the training corpus to learn reordering rules that achieve a (locally) monotone alignment. Figure 1 shows an example in which three reordering rules are extracted from the POS tags of an English source sentence and its corresponding Spanish GIZA++ alignment. Before translation, we construct lattices for every source sentence. The lattices include the original source sentence along with all the reorderings that are consistent with the learned rules. All incoming edges of the lattice are annotated with distortion model scores. Figure 2 gives an example of such a lattice. In the subsequent lattice decoding step, we apply either monotone decoding or decoding with a reduced local reordering window, typically of size 2.

2.3 Decoder and MER Training

The ISL beam search decoder (Vogel, 2003) combines all the different model scores to find the best translation. Here, the following models were used:

- The translation model, i.e. the phrase-to-phrase translations extracted from the bilingual corpus, annoted with four translation model scores. These four scores are the smoothed forward and backward phrase translation probabilities and the forward and backward lexical weights.
- A 4-gram language model. The SRI language model toolkit was used to train the language model and we applied modified Kneser-Ney smoothing.
- An internal word reordering model in addition to the already described POS-based reordering. This internal reordering model assigns higher costs to longer distance reordering.
- Simple word and phrase count models. The former is essentially used to compensate for the tendency of the language model to prefer shorter translations, while the latter can give preference to longer phrases, potentially improving fluency.

```
We all agree on that
PRP DT VB IN DT
En {4} esto {5} estamos {1} todos {2} de {} acuerdo {3}
⇒ PRP DT VB IN DT :  4 – 5 – 1 – 2 – 3
⇒ PRP DT VB:  2 – 3 – 1
⇒ PRP DT VB IN:  3 – 4 – 1 – 2
```

Figure 1: Rule extraction for the POS-based reordering scheme.

The ISL SMT decoder is capable of loading several language models (LMs) at the same time, namely n-gram SRI language models with n up to 4 and suffix array language models (Zhang and Vogel, 2006) of arbitrary length. While we typically see gains in performance for using suffix array LMs with longer histories, we restricted ourselves here to one 4-gram SRI LM only, due to a limited amount of available LM training data. The decoding process itself is organized in two stages. First, all available word and phrase translations are found and inserted into a so-called translation lattice. Then the best combination of these partial translations is found by doing a best path search through the translation lattice, where we also allow for word reorderings within a predefined local reordering window.

To optimize the system towards a maximal BLEU or NIST score, we use Minimum Error Rate (MER) Training as described in (Och, 2003). For each model weight, MER applies a multi-linear search on the development set n-best list produced by the system. Due to the limited numbers of translations in the n-best list, these new model weights are suboptimal. To compensate for this, a new full translation is done. The resulting new n-best list is then merged with the old n-best list and the optimization process is repeated. Typically, the translation quality converges after three iterations.

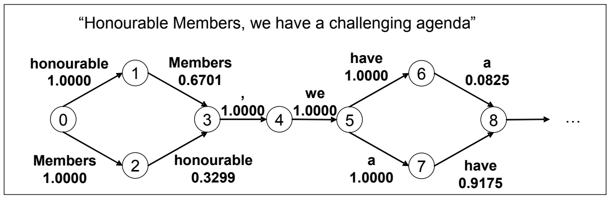

Figure 2: Example for a source sentence lattice from the POS-based reordering scheme.

	English	Spanish
sentence pairs	1259914	
unique sent. pairs	1240151	
sentence length	25.3	26.3
words	31.84 M	33.16 M
vocabulary	266.9 K	346.3 K

Table 1: Corpus statistics for the English/Spanish Europarl corpus.

3 Spanish ↔ English Europarl and News Commentary Task

3.1 Data and Translation Tasks

The systems for the English ↔ Spanish translation tasks were trained on the sentence-aligned Europarl corpus (Koehn, 2005). Detailed corpus statistics can be found in Table 1. The available parallel News Commentary training data of approximately 1 million running words for both languages was only used as additional language model training data, to adapt our in-domain (Europarl) system to the out-of-domain (News Commentary) task.
The development sets consist of 2000 Europarl sentences (dev-EU) and 1057 News Commentary sentences (dev-NC). The available development-test data consists of 2 x 2000 Europarl sentences (devtest-EU and test06-EU) and 1064 News Commentary sentences (test06-NC). All development and development-test sets have only one reference translation per sentence.

3.2 Data Normalization

The ACL shared task is very close in form and content to the Final Text Editions (FTE) task of the TC-STAR (TC-STAR, 2004) evaluation. For this reason, we decided to apply a similar normalization scheme to the training data as was applied in our TC-STAR verbatim SMT system. Although trained on "verbatimized" data that did not contain any numbers, but rather had all numbers and dates spelled out, it yielded consistently better results than our TC-STAR FTE SMT system. When translating FTE content, the verbatim system treated all numbers as unknown words, i.e. they were left unchanged during translation. To compensate for this, we applied extended postprocessing to the translations that conducts the necessary conversions between Spanish and English numbers, e.g. the conversion of decimal comma in Spanish to decimal point in English. Other key points which we adopted from this normalization scheme were the tokenization of punctuation marks, the true-casing of the first word of each sentence, as well as extended cleaning of the training data. The latter mainly consisted of the removal of sections with a highly unbalanced source to target words ratio and the removal of unusual string combinations and document references, like for example "B5-0918/2000", "(COM(2000) 335 - C5-0386/2000 - 2000/0143(CNS))", etc.
Based on this normalization scheme, we trained and optimized a baseline in-domain system on accordingly normalized source and reference sentences. For optimization, we combined the available development sets for the Europarl task and the News Commentary task. In order to further improve the applied normalization scheme, we experimented with replacing all numbers with the string "NMBR", rather than spelling them out and by replacing all document identifiers with the string "DCMNT", rather than deleting them. This was first done for the language model training data only, and then for all data, i.e. for the bilingual training data and for the development set source and reference sentences. In the latter case, the respective tags were again replaced by the correct numbers and document identifiers during postprocessing. Table 2 shows the case sensitive BLEU scores for the three normalization approaches on the English ↔ Spanish Europarl and News Commentary development sets. These scores were computed with the official NIST scoring script against the original (not normalized) references.

3.3 In-domain System

As mentioned above, we combined the Europarl and News Commentary development sets when optimizing the in-domain system. This resulted in only one

Task	baseline	LM only	all data
Europarl	30.94	31.20	31.26
News Com.	31.28	31.39	31.73

Table 2: Case sensitive BLEU scores on the in-domain and out-of-domain development sets for the three different normalization schemes.

Task	Eng $\rightarrow$ Spa	Spa $\rightarrow$ Eng
dev-EU	31.29	31.77
dev-NC	31.81	31.12
devtest-EU	31.01	31.40
test06-EU	31.87	31.76
test06-NC	30.23	29.22

Table 3: Case sensitive BLEU scores for the final English $\leftrightarrow$ Spanish in-domain systems.

set of scaling factors, i.e. the in-domain system applies the same scaling factors for translating in-domain data as for translating out-of-domain data. Our baseline system applied only monotone lattice decoding. For our final in-domain system, we used a local reordering window of length 2, which accounts for the slightly higher scores when compared to the baseline system. The BLEU scores for both translation directions on the different development and development-test sets can be found in Table 3.

3.4 Out-of-domain System

In order to adapt our in-domain system towards the out-of-domain News Commentary task, we considered two approaches based on language model adaptation. First, we interpolated the in-domain LM with an out-of-domain LM computed on the available News Commentary training data. The interpolation weights were chosen such as to achieve a minimal LM perplexity on the out-of-domain development set. For both languages, the interpolation weights were approximately 0.5. Our second approach was to simply load the out-of-domain LM as an additional LM into our decoder. In both cases, we optimized the translation system on the out-of-domain development data only. For the second approach, MER optimization assigned three to four times higher scaling factors to the considerably smaller out-domain LM than to the original in-domain LM. Table 4 shows the results in BLEU on the out-of-domain development and development-test sets for both translation directions. While loading a second LM gives similar or slightly better results on the development set during MER optimization, we see consistently worse results on the unseen development-test set. This, in the context of the relatively small amount of development data, can be explained by stronger overfitting during optimization.

Task	Eng $\rightarrow$ Spa		Spa $\rightarrow$ Eng	
	interp	2 LMs	interp	2 LMs
dev-NC	33.31	33.28	32.61	32.70
test06-NC	32.55	32.15	30.73	30.55

Table 4: Case sensitive BLEU scores for the final English $\leftrightarrow$ Spanish out-of-domain systems.

4 English $\leftrightarrow$ German Europarl Task

The systems for the English $\leftrightarrow$ German translation tasks were trained on the sentence-aligned Europarl corpus only. The complete corpus consists of approximately 32 million English and 30 million German words.

We applied a similar normalization scheme to the training data as for the English $\leftrightarrow$ Spanish system. The main difference was that we did not replace numbers and that we removed all document references. In the translation process, the document references were treated as unknown words and therefore left unchanged. As above, we trained and optimized a first baseline system on the normalized source and reference sentences. However, we used only the Europarl task development set during optimization. To achieve further improvements on the German $\rightarrow$ English task, we applied a compound splitting technique. The compound splitting was based on (Koehn and Knight, 2003) and was applied on the lowercased source sentences. The words generated by the compound splitting were afterwards true-cased. Instead of replacing a compound by its separate parts, we added a parallel path into the source sentence lattices used for translation. The source sentence lattices were augmented with scores on their edges indicating whether each edge represents a word of the original text or if it was generated during compound splitting.

Table 5 shows the case-sensitive BLEU scores for the final German $\leftrightarrow$ English systems. In contrast to the English $\leftrightarrow$ Spanish systems, we used only monotonous decoding on the lattices containing the

task	Eng → Ger	Ger → Eng
dev-EU	18.58	23.85
devtest-EU	18.50	23.87
test06-EU	18.39	23.88

Table 5: Case sensitive BLEU scores for the final English ↔ German in-domain systems.

task	PHRA	SYNT	COMB
dev-EU	31.77	32.48	32.77
test06-EU	31.76	32.15	32.27

Table 6: Results for combining the syntax-augmented system (SYNT) with the phrase-based system (PHRA).

syntactical reorderings.

5 System Combination via n-best List Combination and Rescoring

5.1 N-best List Rescoring

For n-best list rescoring we used unique 500-best lists, which may have less than 500 entries for some sentences. In this evaluation, we used several features computed from different information sources such as features from the translation system, additional language models, IBM-1 word lexica and the n-best list itself. We calculated 4 features from the IBM-1 word lexica: the word probability sum as well as the maximum word probability in both language directions. From the n-best list itself, we calculated three different sets of scores. A position-dependent word agreement score as described in (Ueffing and Ney, 2005) with a position window instead of the Levenshtein alignment, the n-best list n-gram probability as described in (Zens and Ney, 2006) and a position-independent n-gram agreement, which is a variation on the first two. To tune the feature combination weights, we used MER optimization.

Rescoring the n-best lists from our individual systems did not give significant improvements on the available unseen development-test data. For this reason, we did not apply n-best list rescoring to the individual systems. However, we investigated the feasibility of combining two different systems by rescoring the joint n-best lists of both systems. The corresponding results are described in the following sections.

5.2 Combining Syntax-Augmented MT and Phrase-Based MT

On the Spanish → English in-domain task, we participated not only with the ISL phrase-based SMT system as described in this paper, but also with the ISL syntax-augmented system. The syntax-augmented system was trained on the same normalized data as the phrase-based system. However, it was optimized on the in-domain development set only. More details on the syntax-augmented system can be found in (Zollmann et al., 2007). Table 6 lists the respective BLEU scores of both systems as well as the BLEU score achieved by combining and rescoring the individual 500-best lists.

5.3 Combining MT Systems with Different Source Languages

(Och and Ney, 2001) describes methods for translating text given in multiple source languages into a single target language. The ultimate goal is to improve the translation quality when translating from one source language, for example English into multiple target languages, such as Spanish and German. This can be done by first translating the English document into German and then using the translation as an additional source, when translating to Spanish. Another scenario where a multi-source translation becomes desirable was described in (Paulik et al., 2005). The goal was to improve the quality of automatic speech recognition (ASR) systems by employing human-provided simultaneous translations. By using automatic speech translation systems to translate the speech of the human interpreters back into the source language, it is possible to bias the source language ASR system with the additional knowledge. Having these two frameworks in mind, we investigated the possibility of combining our in-domain German → English and Spanish → English translation systems using n-best list rescoring. Table 7 shows the corresponding results. Even though the German → English translation performance was approximately 8 BLEU below the translation performance of the Spanish → English system, we were able to improve the final translation performance by up to 1 BLEU.

task	Spa → Eng	Ger → Eng	Comb.
dev-EU	31.77	23.85	32.76
devtest-EU	31.40	23.87	32.41
test06-EU	31.76	23.88	32.51

Table 7: Results for combining the Spanish → English and German → English phrase-based systems on the in-domain tasks.

6 Conclusion

We described the ISL phrase-based statistical machine translation systems that were used for the 2007 ACL Workshop on Statistical Machine Translation. Using the available out-of-domain News Commentary task training data for language model adaptation, we were able to significantly increase the performance on the out-of-domain task by 2.3 BLEU for English → Spanish and by 1.3 BLEU for Spanish → English. We also showed the feasibility of combining different MT systems by combining and rescoring their resprective n-best lists. In particular, we focused on the combination of our phrase-based and syntax-augmented systems and the combination of two phrase-based systems operating on different source languages. While we saw only a minimal improvement of 0.1 BLEU for the phrase-based and syntax-augmented combination, we gained up to 1 BLEU, in case of the multi-source translation.

References

G. Foster, R. Kuhn, and H. Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proc. of Empirical Methods in Natural Language Processing*, Sydney, Australia.

J. M. Crego et al. 2006. N-gram-based SMT System Enhanced with Reordering Patterns. In *Proc. of the Workshop on Statistical Machine Translation*, pages 162–165, New York, USA.

P. Koehn and K. Knight. 2003. Empirical Methods for Compound Splitting. In *Proc. of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 187–193, Budapest, Hungary.

P. Koehn and C. Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Langauges. In *Proc. of the Workshop on Statistical Machine Translation*, pages 102–121, New York, USA.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of Machine Translation Summit*.

F.J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, Hongkong, China.

F. J. Och and H. Ney. 2001. Statistical Multi-Source Translation. In *Proc. of Machine Translation Summit*, pages 253–258, Santiago de Compostela, Spain.

F. J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160 – 167, Sapporo, Japan.

M. Paulik, S. Stueker, C. Fuegen, T. Schultz, T. Schaaf, and A. Waibel. 2005. Speech Translation Enhanced Automatic Speech Recognition. In *Proc. of the Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico.

TC-STAR. 2004. Technology and Corpora for Speech to Speech Translation. *http://www.tc-star.org.*

N. Ueffing and H. Ney. 2005. Word-Level Confidence Estimation for Machine Translation using Phrase-Based Translation Models. In *Proc. of HLT and EMNLP*, pages 763–770, Vancouver, British Columbia, Canada.

S. Vogel. 2003. SMT Decoder Dissected: Word Reordering. In *Proc. of Int. Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China.

R. Zens and H. Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation*, pages 72–77, New York, USA.

Y. Zhang and S. Vogel. 2006. Suffix Array and its Applications in Empirical Natural Language Processing. In *the Technical Report CMU-LTI-06-010*, Pittsburgh, USA.

A. Zollmann, A. Venugopal, M. Paulik, and S. Vogel. 2007. The Syntax Augmented MT (SAMT) system at the Shared Task for the 2007 ACL Workshop on Statistical Machine Translation. In *Proc. of ACL 2007 Workshop on Statistical MachineTranslation*, Prague, Czech Republic.