Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing

Preslav Nakov*
EECS, CS division
University of California at Berkeley
Berkeley, CA 94720
nakov@cs.berkeley.edu

Abstract

We describe the experiments of the UC Berkeley team on improving English-Spanish machine translation of news text, as part of the WMT'08 Shared Translation Task. We experiment with domain adaptation, combining a small in-domain news bi-text and a large out-of-domain one from the Europarl corpus, building two separate phrase translation models and two separate language models. We further add a third phrase translation model trained on a version of the news bi-text augmented with monolingual sentence-level syntactic paraphrases on the source-language side, and we combine all models in a log-linear model using minimum error rate training. Finally, we experiment with different tokenization and recasing rules, achieving 35.09% Bleu score on the WMT'07 news test data when translating from English to Spanish, which is a sizable improvement over the highest Bleu score achieved on that dataset at WMT'07: 33.10% (in fact, by our system). On the WMT'08 English to Spanish news translation, we achieve 21.92%, which makes our team the second best on Bleu score.

1 Introduction

Modern Statistical Machine Translation (SMT) systems are trained on sentence-aligned bilingual corpora, typically from a single domain. When tested on text from that same domain, they demonstrate state-of-the art performance, but on out-of-domain test data the results can get significantly worse. For example, on the WMT'06 Shared Translation Task, the scores for French to English translation dropped from about 30 to about 20 Bleu points for nearly all systems when tested on *News Commentary* rather than *Europarl* text, which was used on training (Koehn and Monz, 2006).

Therefore, in 2007 the Shared Task organizers provided 1M words of bilingual *News Commentary* training data in addition to the 30M *Europarl* data, thus inviting interest in domain adaptation experiments. Given the success of the idea, the same task was offered this year with slightly larger training bi-texts: 1.3M and 32M words, respectively.

2 System Parameters

The team of the University of California at Berkeley (`ucb`) participated in the WMT'08 Shared Translation Task with two systems, English→Spanish and Spanish→English, applied to translating *News Commentary* text, for which a very limited amount of training data was provided. We experimented with domain adaptation, combining the provided small in-domain bi-text and the large out-of-domain one from the *Europarl* corpus, building two phrase translation models and two language models. We further added a third phrase translation model trained on a version of the news bi-text augmented with monolingual sentence-level syntactic paraphrases on the source-language side, and we combined all models in one big log-linear model using minimum error rate training. We also experimented with different tokenization and recasing ideas.

*After January 2008 at the Linguistic Modeling Department, Institute for Parallel Processing, Bulgarian Academy of Sciences, nakov@lml.bas.bg

Proceedings of the Third Workshop on Statistical Machine Translation, pages 147–150,
Columbus, June 2008.

2.1 Sentence-Level Syntactic Paraphrases

The idea of using paraphrases is motivated by the observation that, in many cases, the testing text contains pieces that are equivalent, but syntactically different from the phrases learned on training, which might result in missing the opportunity for a high-quality translation. For example, an English→Spanish SMT system could have an entry in its phrase table for *inequality of income*, but not for *income inequality*. Note that the latter phrase is hard to translate into Spanish where noun compounds are rare: the correct translation in this case requires a suitable Spanish preposition and a reordering, which are hard for the system to realize and do properly. We address this problem by generating nearly-equivalent syntactic paraphrases of the source-side training sentences, targeted at noun compounds. We then pair each paraphrased sentence with the foreign translation associated with the original sentence in the training data. The resulting augmented bi-text is used to train an SMT system, which learns many useful new phrases. The idea was introduced in (Nakov and Hearst, 2007), and is described in more detail in (Nakov, 2007).

Unfortunately, using multiple paraphrased versions of the same sentence changes the word frequencies in the training bi-text, thus causing worse maximum likelihood estimates, which results in bad system performance. However, real improvements can still be achieved by merging the phrase tables of the two systems, giving priority to the original.

2.2 Domain Adaptation

In our previous findings (Nakov and Hearst, 2007), we found that using in-domain and out-of-domain language models is the best way to perform domain adaptation. Following (Koehn and Schroeder, 2007), we further used two phrase tables.

2.3 Improving the Recaser

One problem we noticed with the default recasing is that unknown words are left in lowercase. However, many unknown words are in fact named entities (persons, organization, or locations), which should be spelled capitalized. Therefore, we prepared a new recasing script, which makes sure that all unknown words keep their original case.

2.4 Changing Tokenization/Detokenization

We found the default tokenizer problematic: it keeps complex adjectives as one word, e.g., *well-rehearsed*, *self-assured*, *Arab-Israeli*. While linguistically correct, this is problematic for machine translation due to data sparsity. For example, the SMT system might know how to translate into Spanish both *well* and *rehearsed*, but not *well-rehearsed*, and thus at translation time it would be forced to handle it as an unknown word. A similar problem is related to double dashes '--', as illustrated by the following training sentence: "*So the question now is what can China do to freeze--and, if possible, to reverse--North Korea's nuclear program.*"

Therefore, we changed the tokenizer, so that it puts a space around '-' and '--'. We also changed the detokenizer accordingly, adding some rules for fixing erroneous output, e.g., making sure that in Spanish text ¿ and ?, ¡ and ! match. We also added some rules for numbers, e.g., the English 1,185.32 should be spelled as 1.185,32 in Spanish.

3 The UCB System

As Table 1 shows, we performed many experiments varying different parameters of the system. Due to space limitations, here we will only describe our best system, $\text{news}_{10} \prec \text{euro}_{10} \prec \text{par}_{10}$.

To build the system, we trained three separate phrase-based SMT systems (max phrase lengths 10): on the original *News Commentary* corpus (*news*), on the paraphrased version of *News Commentary* (*par*), and on the *Europarl* dataset (*euro*). As a result, we obtained three phrase tables, T_{news}, T_{par}, and T_{euro}, and three lexicalized reordering models, R_{news}, R_{par}, and R_{euro}, which we had to merge.

First, we kept all phrase pairs from T_{news}. Then we added those phrase pairs from T_{euro} which were not present in T_{news}. Finally, we added to them those from T_{par} which were not in T_{news} nor in T_{euro}. For each phrase pair added, we retained its associated features: forward phrase translation probability, reverse phrase translation probability, forward lexical translation probability, reverse lexical translation probability, and phrase penalty. We further added three new features – P_{news}, P_{euro}, and P_{par} – each of them was 1 if the phrase pair came from that system, and 0.5 otherwise.

	BLEU		Toke-	News Comm.			Europarl			Tuning	
Model	DR	IR	nizer	slen	plen	LM	slen	plen	LM	#iter	score
1	2	3	4	5	6	7	8	9	10	11	12
I. Original Tokenizer											
news$_7$ (baseline)	32.04	32.30	def.	40	7	3	–	–	–	8	33.51
news$_7$	31.98	32.21	def.	100	7	3	–	–	–	19	33.95
news$_{10}$	32.43	32.67	def.	100	10	3	–	–	–	13	34.50
II. New Tokenizer											
- II.1. Europarl only											
euro$_7$	29.92	30.19	new	–	–	–	40	7	5	10	33.02
euro$_{10}$	30.14	30.36	new	–	–	–	40	10	5	10	32.86
- II.2. News Commentary only											
par$_{10}$	31.17	31.44	new	100	10	3	–	–	–	8	33.91
news$_{10}$	32.27	32.53	new	100	10	3	–	–	–	12	34.49
news$_{10}\prec$par$_{10}$	32.09	32.34	new	100	10	3	–	–	–	24	34.63
- II.3. News Commentary + Europarl											
-- II.3.1. using Europarl LM											
par$_{10}$	32.88	33.16	new	100	10	3	–	–	5	11	35.54
news$_{10}$	33.99	34.26	new	100	10	3	–	–	5	8	36.16
news$_{10}\prec$par$_{10}$	34.42	34.71	new	100	10	3	–	–	5	17	36.41
-- II.3.2. using Europarl LM & Phrase Table (max phrase length 7)											
$\star$news$_{10}$+euro$_7$+par$_{10}$	32.75	32.96	new	100	10	3	40	7	5	27	35.28
$\star$news$_{10}$+euro$_7$	34.06	34.32	new	100	10	3	40	7	5	28	36.82
news$_{10}\prec$euro$_7$	34.05	34.31	new	100	10	3	40	7	5	9	36.71
news$_{10}\prec$par$_{10}\prec$euro$_7$	34.25	34.52	new	100	10	3	40	7	5	14	36.88
news$_{10}\prec$euro$_7\prec$par$_{10}$	34.69	34.97	new	100	10	3	40	7	5	10	37.01
-- II.3.3. using Europarl LM & Phrase Table (max phrase length 10)											
$\star$news$_{10}$+euro$_{10}$+par$_{10}$	32.74	33.02	new	100	10	3	40	10	5	36	35.60
news$_{10}\prec$euro$_{10}\prec$par$_{10}$	**34.85**	**35.09**	**new**	**100**	**10**	**3**	**40**	**10**	**5**	**12**	**37.13**

Table 1: **English→Spanish translation experiments with the WMT'07 data: training on *News Commentary* and *Europarl*, and evaluating on *News Commentary*.** Column 1 provides a brief description of the model used. Here we use *euro*, *news* and *par* to refer to using phrase tables extracted from the *Europarl*, the *News Commentary*, or the *Paraphrased News Commentary* training bi-text; the index indicates the maximum phrase length allowed. The $\prec$ operation means the phrase tables are merged, giving priority to the left one and using additional features indicating where each phrase pair came from, while the $+$ operation indicates the phrase tables are used together without priorities. The models using the $+$ operation are marked with a $\star$ as a reminder that the involved phrase tables are used together, as opposed to being priority-merged. Note also that the models from II.3.1. only use the Spanish part of the *Europarl* training data to build an out-of-domain language model; this is not indicated in column 1, but can be seen in column 10. Columns 2 and 3 show the testing Bleu score after applying the Default Recaser (*DR*) and the Improved Recaser (*IR*), respectively. Column 4 shows whether the default or the new tokenizer was used. Columns 5, 6 and 7 contain the parameters of the *News Commentary* training data: maximum length of the training sentences used (*slen*), maximum length of the extracted phrases (*plen*), and order of the language model (*LM*), respectively. Columns 8, 9 and 10 contain the same parameters for the *Europarl* training data. Column 11 shows the number of iterations the MERT tuning took, and column 12 gives the corresponding tuning Bleu score achieved. Finally, for the WMT'08 competition, we used the system marked in bold.

We further merged R_{news}, R_{euro}, and R_{par} in a similar manner: we first kept all phrases from R_{news}, then we added those from R_{euro} which were not present in R_{news}, and finally those from R_{par} which were not in R_{news} nor in R_{euro}.

We used two language models with Kneser-Ney smoothing: a 3-gram model trained on *News Commentary*, and a 5-gram model trained on *Europarl*.

We then trained a log-linear model using the following feature functions: language model probabilities, word penalty, distortion cost, and the parameters from the phrase table. We set the feature weights by optimizing the *Bleu* score directly using minimum error rate training (Och, 2003) on the development set. We used these weights in a beam search decoder to produce translations for the test sentences, which we compared to the WMT'07 gold standard using *Bleu* (Papineni et al., 2002).

4 Results and Discussion

Table 1 shows the evaluation results using the WMT'07 *News Commentary* test data. Our best English→Spanish system $\text{news}_{10}\prec\text{euro}_{10}\prec\text{par}_{10}$ (see the table caption for explanation of the notation), which is also our submission, achieved 35.09 Bleu score with the improved recaser; with the default recaser, the score drops to 34.85.

Due to space limitations, our Spanish→English results are not in Table 1. This time, we did not use paraphrases, and our best system $\text{news}_{10}\prec\text{euro}_{10}$ achieved 35.78 and 35.17 Bleu score with the improved and the default recaser, respectively.

As the table shows, using the improved recaser yields consistent improvements by about 0.3 Bleu points. Using an out-of-domain language model adds about 2 additional Bleu points, e.g., news_{10} improves from 32.53 to 34.26, and $\text{news}_{10}\prec\text{par}_{10}$ improves from 32.34 to 34.71. The impact of also adding an out-of-domain phrase table is tiny: $\text{news}_{10}\prec\text{euro}_{7}$ improves on news_{10} by 0.05 only. Adding paraphrases however can yield an absolute improvement of about 0.6, e.g., 34.31 vs. 34.97 for $\text{news}_{10}\prec\text{euro}_{7}$ and $\text{news}_{10}\prec\text{euro}_{7}\prec\text{par}_{10}$. Interestingly, using an out-of-domain phrase table has a bigger impact when paraphrases are used, e.g., for $\text{news}_{10}\prec\text{par}_{10}$ and $\text{news}_{10}\prec\text{euro}_{7}\prec\text{par}_{10}$ we have 34.71 and 34.97, respectively. Finally, we were surprised to find out that using the new tokenizer does not help: for news_{10} the default tokenizer yields 32.67, while the new one only achieves 32.53. This is surprising for us, since the new tokenizer used to help consistent on the WMT'06 data.

5 Conclusions and Future Work

We described the UCB system for the WMT'08 Shared Translation Task. By combining in-domain and out-of-domain data, and by using sentence-level syntactic paraphrases and a better recaser, we achieved an improvement of almost 2 Bleu points[1] over the best result on the WMT'07 test data[2], and the second best Bleu score for this year's English→Spanish translation of news text.

In future work, we plan a deeper analysis of the obtained results. First, we would like to experiment with new ways to combine data from different domains. We also plan to further improve the recaser, and to investigate why the new tokenizer did not help for the WMT'07 data.

References

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227. Association for Computational Linguistics.

Preslav Nakov and Marti Hearst. 2007. UCB system description for the WMT 2007 shared task. In *Workshop on Statistical Machine Translation*, pages 212–215.

Preslav Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, UCB/EECS-2007-173.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

[1] Note however that this year we had more training data compared to last year: 1.3M vs. 1M words for *News Commentary*, and 32M vs. 30M words for *Europarl*.

[2] In fact, achieved by our system at WMT'07.