NUS at WMT09: Domain Adaptation Experiments for English-Spanish Machine Translation of News Commentary Text

Preslav Nakov
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nakov@comp.nus.edu.sg

Hwee Tou Ng
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
nght@comp.nus.edu.sg

Abstract

We describe the system developed by the team of the National University of Singapore for English to Spanish machine translation of News Commentary text for the WMT09 Shared Translation Task. Our approach is based on domain adaptation, combining a small in-domain *News Commentary* bi-text and a large out-of-domain one from the *Europarl* corpus, from which we built and combined two separate phrase tables. We further combined two language models (in-domain and out-of-domain), and we experimented with cognates, improved tokenization and recasing, achieving the highest lowercased NIST score of 6.963 and the second best lowercased Bleu score of 24.91% for training without using additional external data for English-to-Spanish translation at the shared task.

1 Introduction

Modern Statistical Machine Translation (SMT) systems are typically trained on sentence-aligned parallel texts (bi-texts) from a particular domain. When tested on text from that domain, they demonstrate state-of-the art performance, but on out-of-domain test data the results can deteriorate significantly. For example, on the WMT06 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* instead of the *Europarl*[1] text, which was used for training (Koehn and Monz, 2006).

Subsequently, in 2007 and 2008, the WMT Shared Translation Task organizers provided a limited amount of bilingual *News Commentary* training data (1-1.3M words) in addition to the large amount of *Europarl* data (30-32M words), and set up separate evaluations on *News Commentary* and on *Europarl* data, thus inviting interest in domain adaptation experiments for the *News* domain (Callison-Burch et al., 2007; Callison-Burch et al., 2008). This year, the evaluation is on *News Commentary* only, which makes domain adaptation the central focus of the Shared Translation Task.

The team of the National University of Singapore (NUS) participated in the WMT09 Shared Translation Task with an English-to-Spanish system.[2] Our approach is based on domain adaptation, combining the small in-domain *News Commentary* bi-text (1.8M words) and the large out-of-domain one from the Europarl corpus (40M words), from which we built and combined two separate phrase tables. We further used two language models (in-domain and out-of-domain), cognates, improved tokenization, and additional smart recasing as a post-processing step.

2 The NUS System

Below we describe separately the standard and the nonstandard settings of our system.

2.1 Standard Settings

In our baseline experiments, we used the following general setup: First, we tokenized the par-

[1] See (Koehn, 2005) for details about the *Europarl* corpus.

[2] The task organizers invited submissions translating forward and/or backward between English and five other European languages (French, Spanish, German, Czech and Hungarian), but we only participated in English→Spanish, due to time limitations.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 75–79,
Athens, Greece, 30 March – 31 March 2009.

allel bi-text, converted it to lowercase, and filtered out the overly-long training sentences, which complicate word alignments (we tried maximum length limits of 40 and 100). We then built separate English-to-Spanish and Spanish-to-English directed word alignments using IBM model 4 (Brown et al., 1993), combined them using the *intersect+grow heuristic* (Och and Ney, 2003), and extracted phrase-level translation pairs of maximum length 7 using the *alignment template approach* (Och and Ney, 2004). We thus obtained a *phrase table* where each phrase translation pair is associated with the following five standard parameters: forward and reverse phrase translation probabilities, forward and reverse lexical translation probabilities, and phrase penalty.

We then trained a log-linear model using the standard feature functions: language model probability, word penalty, distortion costs (we tried distance based and lexicalized reordering models), and the parameters from the phrase table. We set all feature weights by optimizing Bleu (Papineni et al., 2002) directly using *minimum error rate training* (MERT) (Och, 2003) on the tuning part of the development set (`dev-test2009a`). We used these weights in a beam search decoder (Koehn et al., 2007) to translate the test sentences (the English part of `dev-test2009b`, tokenized and lowercased). We then recased the output using a monotone model that translates from lowercase to uppercase Spanish, we post-cased it using a simple heuristic, de-tokenized the result, and compared it to the gold standard (the Spanish part of `dev-test2009b`) using Bleu and NIST.

2.2 Nonstandard Settings

The nonstandard features of our system can be summarized as follows:

Two Language Models. Following Nakov and Hearst (2007), we used two language models (LM) – an in-domain one (trained on a concatenation of the provided monolingual Spanish *News Commentary* data and the Spanish side of the training *News Commentary* bi-text) and an out-of-domain one (trained on the provided monolingual Spanish *Europarl* data). For both LMs, we used 5-gram models with Kneser-Ney smoothing.

Merging Two Phrase Tables. Following Nakov (2008), we trained and merged two phrase-based SMT systems: a small in-domain one using the *News Commentary* bi-text, and a large out-of-domain one using the *Europarl* bi-text. As a result, we obtained two phrase tables, T_{news} and T_{euro}, and two lexicalized reordering models, R_{news} and R_{euro}. We merged the phrase table as follows. 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}. For each phrase pair added, we retained its associated features: forward and reverse phrase translation probabilities, forward and reverse lexical translation probabilities, and phrase penalty. We further added two new features, F_{news} and F_{euro}, which show the source of each phrase. Their values are 1 and 0.5 when the phrase was extracted from the *News Commentary* bi-text, 0.5 and 1 when it was extracted from the *Europarl* bi-text, and 1 and 1 when it was extracted from both. As a result, we ended up with seven parameters for each entry in the merged phrase table.

Merging Two Lexicalized Reordering Tables. When building the two phrase tables, we also built two lexicalized reordering tables (Koehn et al., 2005) for them, R_{news} and R_{euro}, which we merged as follows: We first kept all phrases from R_{news}, then we added those from R_{euro} which were not present in R_{news}. This resulting lexicalized reordering table was used together with the above-described merged phrase table.

Cognates. Previous research has shown that using cognates can yield better word alignments (Al-Onaizan et al., 1999; Kondrak et al., 2003), which in turn often means higher-quality phrase pairs and better SMT systems. Linguists define cognates as words derived from a common root (Bickford and Tuggy, 2002). Following previous researchers in *computational* linguistics (Bergsma and Kondrak, 2007; Mann and Yarowsky, 2001; Melamed, 1999), however, we adopted a simplified definition which ignores origin, defining cognates as words in different languages that are mutual translations and have a similar orthography. We extracted and used such potential cognates in order to bias the training of the IBM word alignment models. Following Melamed (1995), we measured the orthographic similarity using *longest common subsequence ratio* (LCSR), which is defined as follows:

$$\mathrm{LCSR}(s_1, s_2) = \frac{|\mathrm{LCS}(s_1, s_2)|}{\max(|s_1|, |s_2|)}$$

where $\mathrm{LCS}(s_1, s_2)$ is the *longest common subsequence* of s_1 and s_2, and $|s|$ is the length of s.

Following Nakov et al. (2007), we combined the LCSR similarity measure with *competitive linking* (Melamed, 2000) in order to extract potential cog-

nates from the training bi-text. Competitive linking assumes that, given a source English sentence and its Spanish translation, a source word is either translated with a single target word or is not translated at all. Given an English-Spanish sentence pair, we calculated LCSR for all cross-lingual word pairs (excluding stopwords and words of length 3 or less), which induced a fully-connected weighted bipartite graph. Then, we performed a greedy approximation to the maximum weighted bipartite matching in that graph (competitive linking) as follows: First, we aligned the most similar pair of unaligned words and we discarded these words from further consideration. Then, we aligned the next most similar pair of unaligned words, and so forth. The process was repeated until there were no words left or the maximal word pair similarity fell below a pre-specified threshold θ ($0 \leq \theta \leq 1$), which typically left some words unaligned.[3] As a result we ended up with a list C of potential cognate pairs. Following (Al-Onaizan et al., 1999; Kondrak et al., 2003; Nakov et al., 2007) we filtered out the duplicates in C, and we added the remaining cognate pairs as additional "sentence" pairs to the bi-text in order to bias the subsequent training of the IBM word alignment models.

Improved (De-)tokenization. The default tokenizer does not split on hyphenated compound words like *nation-building*, *well-rehearsed*, *self-assured*, *Arab-Israeli*, *domestically-oriented*, *etc.* While linguistically correct, this can be problematic for machine translation since it can cause data sparsity issues. For example, the 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, *i.e.*, copy it to the output untranslated. 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.*" We changed the tokenizer so that it splits on '-' and '--'; we altered the detokenizer accordingly.

Improved Recaser. The default recaser suggested by the WMT09 organizers was based on a monotone translation model. We trained such a recaser on the Spanish side of the *News Commentary* bi-text that translates from lowercase to uppercase Spanish. While being good overall, it had a problem with unknown words, leaving them in lowercase. In a *News Commentary* text, however, most unknown words are named entities – persons, organization, locations – which are spelled with a capitalized initial in Spanish. Therefore, we used an additional recasing script, which runs over the output of the default recaser and sets the casing of the unknown words to the original casing they had in the English input. It also makes sure all sentences start with a capitalized initial.

Rule-based Post-editing. We did a quick study of the system errors on the development set, and we designed some heuristic post-editing rules, *e.g.*,

- **? or ! without ¿ or ¡ to the left:** if so, we insert ¿/¡ at the sentence beginning;
- **numbers:** we change English numbers like 1,185.32 to Spanish-style 1.185,32;
- **duplicate punctuation:** we remove duplicate sentence end markers, quotes, commas, parentheses, *etc.*

3 Experiments and Evaluation

Table 1 shows the performance of a simple baseline system and the impact of different cumulative modifications to that system when tuning on `dev-test2009a` and testing on `dev-test2009b`. The table report the Bleu and NIST scores measured on the detokenized output under three conditions: (1) without recasing ('*Lowercased*'), 2) using the default recaser ('*Recased (default)*'), and (3) using an improved recaser and post-editing rules *Post-cased & Post-edited*'). In the following discussion, we will discuss the Bleu results under condition (3).

System 1 uses sentences of length up to 40 tokens from the *News Commentary* bi-text, the default (de-)tokenizer, distance reordering, and a 3-gram language model trained on the Spanish side of the bi-text. Its performance is quite modest: 15.32% of Bleu with the default recaser, and 16.92% when the improved recaser and the post-editing rules are used.

System 2 increases to 100 the maximum length of the sentences in the bi-text, which yields 0.55% absolute improvement in Bleu.

System 3 uses the new (de-)tokenizer, but this turns out to make almost no difference.

[3] For *News Commentary*, we used $\theta = 0.4$, which was found by optimizing on the development set; for *Europarl*, we set $\theta = 0.58$ as suggested by Kondrak et al. (2003).

#	Bitext	System	Lowercased		Recased (default)		Post-cased & Post-edited	
			Bleu	NIST	Bleu	NIST	Bleu	NIST
1	news	*News Commentary* baseline	18.38	5.7837	15.32	5.2266	16.92	5.5091
2	news	+ max sentence length 100	18.91	5.8540	15.93	5.3119	17.47	5.5874
3	news	+ improved (de-)tokenizer	18.96	5.8706	15.97	5.3254	17.48	5.6020
4	news	+ lexicalized reordering	19.81	5.9422	16.64	5.3793	18.28	5.6696
5	news	+ LM: old+monol. *News*, 5-gram	22.29	6.2791	18.91	5.6901	20.55	5.9924
6	news	+ LM_2: *Europarl*, 5-gram	22.46	6.2438	19.10	5.6606	20.75	5.9570
7	news	+ cognates	23.14	6.3504	19.64	5.7478	21.32	6.0478
8	euro	*Europarl* ($\sim$ system 6)	23.73	6.4673	20.23	5.8707	21.89	6.1577
9	euro	+ cognates ($\sim$ system 7)	23.95	6.4709	20.44	5.8742	22.10	6.1607
10	both	Combining 7 & 9	**24.40**	**6.5723**	**20.74**	**5.9575**	**22.37**	**6.2506**

Table 1: **Impact of the combined modifications for English-to-Spanish machine translation on `dev-test2009b`.** We report the Bleu and NIST scores measured on the detokenized output under three conditions: (1) without recasing ('*Lowercased*'), (2) using the default recaser ('*Recased (default)*'), and (3) using an improved recaser and post-editing rules ('*Post-cased & Post-edited*'). The *News Commentary* baseline system uses sentences of length up to 40 tokens from the *News Commentary* bi-text, the default tokenizer and de-tokenizer, a distance-based reordering model, and a trigram language model trained on the Spanish side of the bi-text. The *Europarl* system is the same as system 6, except that it uses the *Europarl* bi-text instead of the *News Commentary* bi-text.

System 4 adds a lexicalized re-ordering model, which yields 0.8% absolute improvement.

System 5 improves the language model. It adds the additional monolingual Spanish *News Commentary* data provided by the organizers to the Spanish side of the bi-text, and uses a 5-gram language model instead of the 3-gram LM used by Systems 1-4. This yields a sizable absolute gain in Bleu: 2.27%.

System 6 adds a second 5-gram LM trained on the monolingual *Europarl* data, gaining 0.2%.

System 7 augments the training bi-text with cognate pairs, gaining another 0.57%.

System 8 is the same as *System 6*, except that it is trained on the out-of-domain *Europarl* bi-text instead of the in-domain *News Commentary* bi-text. Surprisingly, this turns out to work better than the in-domain *System 6* by 1.14% of Bleu. This is a quite surprising result since in both WMT07 and WMT08, for which comparable kinds and size of training data was provided, training on the out-of-domain *Europarl* was always worse than training on the in-domain *News Commentary*. We are not sure why it is different this year, but it could be due to the way the dev-train and dev-test was created for the 2009 data – by extracting alternating sentences from the original development set.

System 9 augments the *Europarl* bi-text with cognate pairs, gaining another 0.21%.

System 10 merges the phrase tables of systems 7 and 9, and is otherwise the same as them. This adds another 0.27%.

Our official submission to WMT09 is the post-edited *System 10*, re-tuned on the full development set: `dev-test2009a` + `dev-test2009b` (in order to produce more stable results with MERT).

4 Conclusion and Future Work

As we can see in Table 1, we have achieved not only a huge 'vertical' absolute improvement of 5.5-6% in Bleu from System 1 to System 10, but also a significant 'horizontal' one: our recased and post-edited result for *System 10* is better than that of the default recaser by 1.63% in Bleu (22.37% vs. 20.74%). Still, the lowercased Bleu of 24.40% suggests that there may be a lot of room for further improvement in recasing – we are still about 2% below it. While this is probably due primarily to the system choosing a different sentence-initial word, it certainly deserves further investigation in future work.

Acknowledgments

This research was supported by research grant POD0713875.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz Joseph Och, David Purdy, Noah Smith, and David Yarowsky. 1999. Statistical machine translation. Technical report, CLSP, Johns Hopkins University, Baltimore, MD.

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 656–663, Prague, Czech Republic.

Albert Bickford and David Tuggy. 2002. Electronic glossary of linguistic terms. http://www.sil.org/mexico/ling/glosario/E005ai-Glossary.htm.

Peter Brown, Vincent Della Pietra, Stephen Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH, USA.

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

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'05)*, Pittsburgh, PA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'07). Demonstration session*, pages 177–180, Prague, Czech Republic.

P. Koehn. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of the X MT Summit*, pages 79–86, Phuket, Thailand.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL'03)*, pages 46–48, Sapporo, Japan.

Gideon Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Annual Meeting of the North American Association for Computational Linguistics (NAACL'01)*, pages 1–8, Pittsburgh, PA, USA.

Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, Cambridge, MA, USA.

Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

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

Preslav Nakov, Svetlin Nakov, and Elena Paskaleva. 2007. Improved word alignments using the Web as a corpus. In *Proceedigs of Recent Advances in Natural Language Processing (RANLP'07)*, pages 400–405, Borovets, Bulgaria.

Preslav Nakov. 2008. Improving English-Spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, Columbus, OH, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'03)*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, PA, USA.