QCRI at WMT12: Experiments in Spanish-English and German-English Machine Translation of News Text

Francisco Guzmán, Preslav Nakov, Ahmed Thabet, Stephan Vogel
Qatar Computing Research Institute
Qatar Foundation
Tornado Tower, floor 10, PO box 5825
Doha, Qatar
{fguzman,pnakov,ahawad,svogel}@qf.org.qa

Abstract

We describe the systems developed by the team of the Qatar Computing Research Institute for the WMT12 Shared Translation Task. We used a phrase-based statistical machine translation model with several non-standard settings, most notably tuning data selection and phrase table combination. The evaluation results show that we rank second in BLEU and TER for Spanish-English, and in the top tier for German-English.

1 Introduction

The team of the Qatar Computing Research Institute (QCRI) participated in the Shared Translation Task of WMT12 for two language pairs:[1] Spanish-English and German-English. We used the state-of-the-art phrase-based model (Koehn et al., 2003) for statistical machine translation (SMT) with several non-standard settings, e.g., data selection and phrase table combination. The evaluation results show that we rank second in BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) for Spanish-English, and in the top tier for German-English.

In Section 2, we describe the parameters of our baseline system and the non-standard settings we experimented with. In Section 3, we discuss our primary and secondary submissions for the two language pairs. Finally, in Section 4, we provide a short summary.

[1]The WMT12 organizers invited systems translating between English and four other European languages, in both directions: French, Spanish, German, and Czech. However, we only participated in Spanish→English and German→English.

2 System Description

Below, in Section 2.1, we first describe our initial configuration; then, we discuss our incremental improvements. We explored several non-standard settings and extensions and we evaluated their impact with respect to different baselines. These baselines are denoted in the tables below by a #number that corresponds to systems in Figures 1 for Spanish-English and in Figure 2 for German-English.

We report case insensitive BLEU calculated on the news2011 testing data using the NIST scoring tool v.11b.

2.1 Initial Configuration

Our baseline system can be summarized as follows:

- Training: News Commentary + Europarl training bi-texts;
- Tuning: news2010;
- Testing: news2011;
- Tokenization: splitting words containing a dash, e.g., *first-order* becomes *first @-@ order*;
- Maximum sentence length: 100 tokens;
- Truecasing: convert sentence-initial words to their most frequent case in the training dataset;
- Word alignments: directed IBM model 4 (Brown et al., 1993) alignments in both directions, then *grow-diag-final-and* heuristics;
- Maximum phrase length: 7 tokens;
- Phrase table scores: forward & reverse phrase translation probabilities, forward & reverse lexical translation probabilities, phrase penalty;

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 298–303,
Montréal, Canada, June 7-8, 2012.

- Language model: 5-gram, trained on the target side of the two training bi-texts;
- Reordering: lexicalized, *msd-bidirectional-fe*;
- Detokenization: reconnecting words that were split around dashes;
- Model parameter optimization: minimum error rate training (MERT), optimizing BLEU.

2.2 Phrase Tables

We experimented with two non-standard settings:

Smoothing. The four standard scores associated with each phrase pair in the phrase table (forward & reverse phrase translation probabilities, forward & reverse lexical translation probabilities) are normally used unsmoothed. We also experimented with Good-Turing and Kneser-Ney smoothing (Chen and Goodman, 1999). As Table 1 shows, the latter works a bit better for both Spanish-English and German-English.

	es-en	**de-en**
Baseline (es:#3,de:#4)	29.98	22.03
Good Turing	29.98	22.07
Kneser-Ney	**30.16**	**22.30**

Table 1: **Phrase table smoothing.**

Phrase table combination. We built two phrase tables, one for News Commentary + Europarl and an additional one for the UN bi-text. We then merged them,[2] adding additional features to each entry in the merged phrase table: F_1, F_2, and F_3. The value of F_1/F_2 is 1 if the phrase pair came from the first/second phrase table, and 0.5 otherwise, while F_3 is 1 if the phrase pair was in both tables, and 0.5 otherwise. We optimized the weights for all features, including the additional ones, using MERT.[3] Table 2 shows that this improves by +0.42 BLEU points.

[2]In theory, we should also re-normalize the conditional probabilities (forward/reverse phrase translation probability, and forward/reverse lexicalized phrase translation probability) since they may not sum to one anymore. In practice, this is not that important since the log-linear phrase-based SMT model does not require that the phrase table features be probabilities (e.g., F_1, F_2, F_3, and the phrase penalty are not probabilities); moreover, we have extra features whose impact is bigger.

[3]This is similar but different from (Nakov, 2008): when a phrase pair appeared in both tables, they only kept the entry from the first table, while we keep the entries from both tables.

	es-en
Baseline (es:#7)	30.94
Merging (1) News+EP with (2) UN	**31.36**

Table 2: **Phrase table merging.**

2.3 Language Models

We built the language models (LM) for our systems using a probabilistic 5-gram model with Kneser-Ney (KN) smoothing. We experimented with LMs trained on different training datasets. We used the SRILM toolkit (Stolcke, 2002) for training the language models, and the KenLM toolkit (Heafield and Lavie, 2010) for binarizing the resulting ARPA models for faster loading with the Moses decoder (Koehn et al., 2007).

2.3.1 Using WMT12 Corpora Only

We trained 5-gram LMs on datasets provided by the task organizers. The results are presented in Table 3. The first line reports the baseline BLEU scores using a language model trained on the target side of the News Commentary + Europarl training bi-texts. The second line shows the results when using an interpolation (minimizing the perplexity on the news2010 tuning dataset) of different language models, trained on the following corpora:

- the monolingual News Commentary corpus plus the English sides of all training News Commentary v.7 bi-texts (for French-English, Spanish-English, German-English, and Czech-English), with duplicate sentences removed (5.5M word tokens; one LM);
- the News Crawl 2007-2011 corpora, (1213M word tokens; separate LM for each of these five years);
- the Europarl v.7 monolingual corpus (60M word tokens; one LM);
- the English side of the Spanish-English UN bi-text (360M word tokens; one LM).

The last line in Table 3 shows the results when using an additional 5-gram LM in the interpolation, one trained on the English side of the 10^9 French-English bi-text (662M word tokens).

We can see that using these interpolations yields very sizable improvements of 1.7-2.5 BLEU points over the baseline. However, while the impact of adding the 10^9 bi-text to the interpolation is clearly visible for Spanish-English (+0.47 BLEU), it is almost negligible for German-English (+0.06 BLEU).

Corpora	es-en	de-en
Baseline (es:#1, de:#2)	27.34	20.01
News + EP + UN (interp.)	29.36	21.66
News + EP + UN + 10^9 (interp.)	**29.83**	**21.72**

Table 3: **LMs using the provided corpora only.**

2.3.2 Using Gigaword

In addition to the WMT12 data, we used the LDC Gigaword v.5 corpus. We divided the corpus into reasonably-sized chunks of text of about 2GB per chunk, and we built a separate intermediate language model for each chunk. Then, we interpolated these language models, minimizing the perplexity on the news2010 development set as with the previous LMs. We experimented with two different strategies for creating the chunks by segmenting the corpus according to (a) data source, e.g., AFP, Xinhua, etc., and (b) year of release. We thus compared the advantages of interpolating epoch-consistent LMs vs. source-coherent LMs. We trained individual LMs for each of the segments and we added them to a pool. Finally, we selected the ten most relevant ones from this pool based on their perplexity on the news2010 devset, and we interpolated them.

The results are shown in Table 4. The first line shows the baseline, which uses an interpolation of the nine LMs from the previous subsection. The following two lines show the results when using an LM trained on Gigaword only. We can see that for Spanish-English, interpolation by year performs better, while for German-English, it is better to use the by-source chunks. However, the following two lines show that when we translate with two LMs, one built from the WMT12 data only and one built using Gigaword data only, interpolation by year is preferable for Gigaword for both language pairs. For our submitted systems, we used the LMs shown in bold in Table 4: we used a single LM for Spanish-English and two LMs for German-English.

Language Models	es-en	de-en
Baseline (es:#5, de:#6)	30.31	22.48
GW by year	**30.68**	22.32
GW by source	30.52	22.56
News-etc + GW by year	30.60	**22.71**
News-etc + GW by source	30.55	22.54

Table 4: **LMs using Gigaword.**

2.4 Parameter Tuning and Data Selection

Parameter tuning is a very important step in SMT. The standard procedure consists of performing a series of iterations of MERT to choose the model parameters that maximize the translation quality on a development set, e.g., as measured by BLEU. While the procedure is widely adopted, it is also recognized that the selection of an appropriate development set is important since it biases the parameters towards specific types of translations. This is illustrated in Table 5, which shows BLEU on the news2011 testset when using different development sets for MERT.

Devset	es-en
news2008	29.47
news2009	29.14
news2010	**29.61**

Table 5: **Using different tuning sets for MERT.**

To address this problem, we performed a selection of development data using an n-gram-based similarity ranking. The selection was performed over a pool of candidate sentences drawn from the news2008, news2009, and news2010 tuning datasets. The similarity metric was defined as follows:

$$\texttt{sim}(f,g) = \texttt{2match}(f,g) * \texttt{lenpen}(f,g) \quad (1)$$

where `2match` represents the number of bi-gram matches between sentences f and g, and `lenpen` is a length penalty to discourage unbalanced matches.

We penalized the length difference using an inverted-squared sigmoid function:

$$\texttt{lenpen}(f,g) = 3 - 4 * \texttt{sig}\left(\left[\frac{|f|-|g|}{\alpha}\right]^2\right) \quad (2)$$

where $|.|$ denotes the length of a sentence in number of words, α controls the maximal tolerance to differences, and `sig` is the sigmoid function.

To generate a suitable development set, we averaged the similarity scores of candidate sentences w.r.t. to the target testset. For instance:

$$s_f = \frac{1}{|G|} \sum_{g \in G} \texttt{sim}(f, g) \quad (3)$$

where G is the set of the test sentences.

Finally, we selected a pool of candidates f from news2008, news2009 and news2011 to generate a 2000-best tuning set. The results when using each of the above penalty functions are presented on Table 6.

devset	es-en
baseline (es:#6)	30.68
selection ($\alpha = 5$)	**30.94**
selection ($\alpha = 10$)	30.90

Table 6: **Selecting sentences for MERT.**

The average length of the source-side sentences in our selected sentence pairs was smaller than in our baseline, the news2011 development dataset. This means that our selected source-side sentences tended to be shorter than in the baseline. Moreover, the standard deviation of the sentence lengths was smaller for our samples as well, which means that there were fewer long sentences; this is good since long sentences can take very long to translate. As a result, we observed sizable speedup in parameter tuning when running MERT on our selected tuning datasets.

2.5 Decoding and Hypothesis Reranking

We experimented with two decoding settings: (1) monotone at punctuation reordering (Tillmann and Ney, 2003), and (2) minimum Bayes risk decoding (Kumar and Byrne, 2004). The results are shown in Table 7. We can see that both yield improvements in BLEU, even if small.

2.6 System Combination

As the final step in our translation system, we performed hypothesis re-combination of the output of several of our systems using the Multi-Engine MT system (MEMT) (Heafield and Lavie, 2010).

	es-en	de-en
Baseline (es:#2,de:#3)	29.83	21.72
+MP	**29.98**	**22.03**
Baseline (es:#4,de:#5)	30.16	22.30
+MBR	**30.31**	**22.48**

Table 7: **Decoding parameters.** Experiments with monotone at punctuation (MP) reordering, and minimum Bayes risk (MBR) decoding.

The results for the actual news2012 testset are shown in Table 8: the system combination results are our primary submission. We can see that system combination yielded 0.4 BLEU points of improvement for Spanish-English and 0.2-0.3 BLEU points for German-English.

3 Our Submissions

Here we briefly describe the cumulative improvements when applying the above modifications to our baseline system, leading to our official submissions for the WMT12 Shared Translation Task.

3.1 Spanish-English

The development of our final Spanish-English system involved several incremental improvements, which have been described above and which are summarized in Figure 1. We started with a baseline system (see Section 2.1), which scored 27.34 BLEU points. From there, using a large interpolated language model trained on the provided data (see Section 2.3.1) yielded +2.49 BLEU points of improvement. Monotone-at-punctuation decoding contributed an additional improvement of +0.15, smoothing the phrase table using Kneser-Ney boosted the score by +0.18, and using minimum Bayes risk decoding added another +0.15 BLEU points. Changing the language model to one trained on Gigaword v.5 and interpolated by year yielded +0.37 additional points of improvement. Another +0.26 points came from tuning data selection. Finally, using the UN data in a merged phrase table (see Section 2.2) yielded another +0.42 BLEU points. Overall, we achieve a total improvement over our initial baseline of about 4 BLEU points.

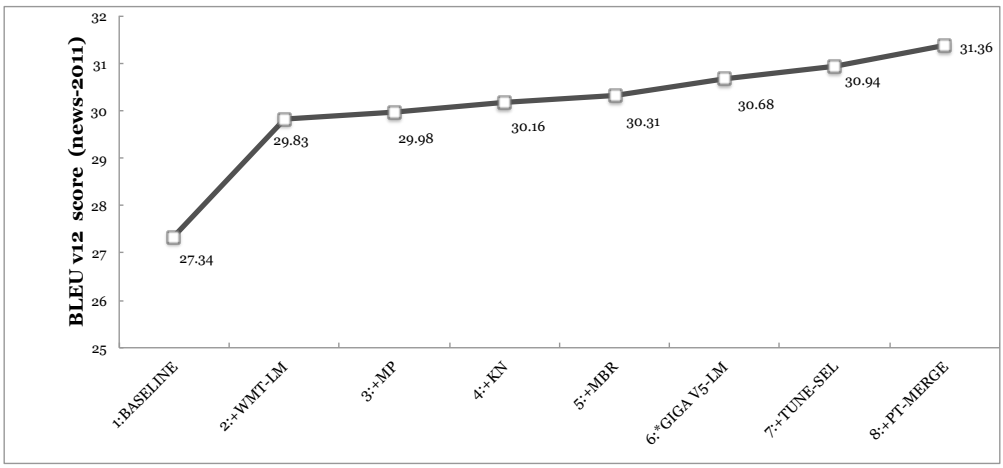

Figure 1: **Incremental improvements for the Spanish-English system.**

3.2 German-English

Figure 2 shows a similar sequence of improvements for our German-English system. We started with a baseline (see Section 2.1) that scored 19.79 BLEU points. Next, we performed compound splitting for the German side of the training, the development and the testing bi-texts, which yielded +0.22 BLEU points of improvement. Using a large interpolated language model trained on the provided corpora (see Section 2.3.1) added another +1.71. Monotone-at-punctuation decoding contributed +0.31, smoothing the phrase table using Kneser-Ney boosted the score by +0.27, and using minimum Bayes risk decoding added another +0.18 BLEU points. Finally, adding a second language model trained on the Gigaword v.5 corpus interpolated by year yielded +0.23 additional BLEU points. Overall, we achieved about 3 BLEU points of total improvement over our initial baseline.

3.3 Final Submissions

For both language pairs, our primary submission was a combination of the output of several of our best systems shown in Figures 1 and 2, which use different experimental settings; our secondary submission was our best individual system, i.e., the right-most one in Figures 1 and 2.

The official BLEU scores, both cased and lowercased, for our primary and secondary submissions, as evaluated on the news2012 dataset, are shown in Table 8. For Spanish-English, we achieved the second highest BLEU and TER scores, while for German-English we were ranked in the top tier.

	news2012	
	lower	cased
Spanish-English		
Primary	34.0	32.9
Secondary	33.6	32.5
German-English		
Primary	23.9	22.6
Secondary	23.6	22.4

Table 8: **The official BLEU scores for our submissions to the WMT12 Shared Translation Task.**

4 Conclusion

We have described the primary and the secondary systems developed by the team of the Qatar Computing Research Institute for Spanish-English and German-English machine translation of news text for the WMT12 Shared Translation Task.

We experimented with phrase-based SMT, exploring a number of non-standard settings, most notably tuning data selection and phrase table combination, which we described and evaluated in a cumulative fashion. The automatic evaluation metrics,[4] have ranked our system second for Spanish-English and in the top tier for German-English.

We plan to continue our work on data selection for phrase table and the language model training, in addition to data selection for tuning.

[4]The evaluation scores for WMT12 are available online: `http://matrix.statmt.org/`

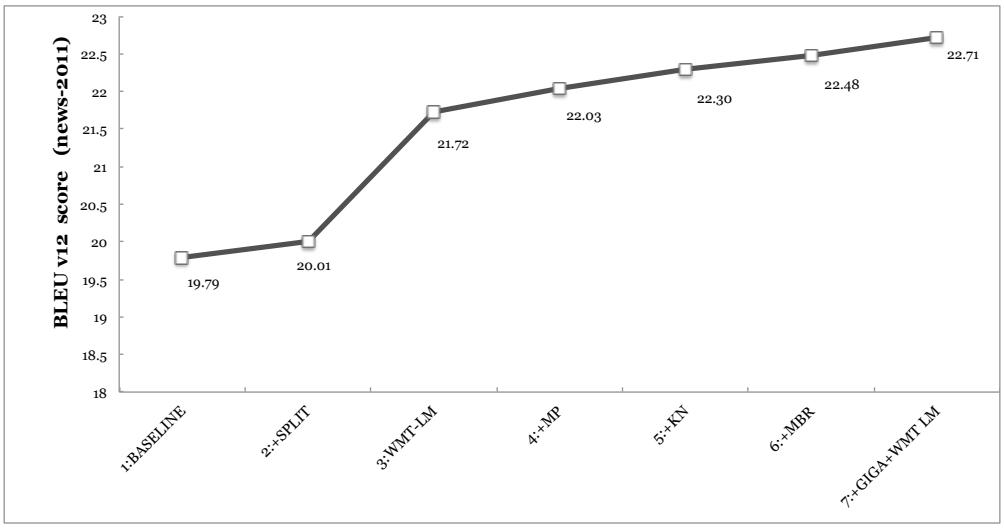

Figure 2: **Incremental improvements for the German-English system.**

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments, which have helped us improve the text of this paper.

References

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.

Stanley Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Kenneth Heafield and Alon Lavie. 2010. Combining machine translation output with open source: The Carnegie Mellon multi-engine machine translation scheme. *The Prague Bulletin of Mathematical Linguistics*, 93(1):27–36.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Edmonton, Canada.

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 ACL. Demonstration session*, ACL '07, pages 177–180, Prague, Czech Republic.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In Susan Dumais, Daniel Marcu, and Salim Roukos, editors, *Proceedings of the Annual Meeting of the North American chapter of the Association for Computational Linguistics*, HLT-NAACL '04, pages 169–176, Boston, MA.

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*, WMT '07, pages 147–150, Prague, Czech Republic.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Annual Meetig of the Association for Machine Translation in the Americas*, AMTA '06, pages 223–231.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of Intl. Conf. on Spoken Language Processing*, volume 2 of *ICSLP '02*, pages 901–904, Denver, CO.

Christoph Tillmann and Hermann Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.