Joint WMT Submission of the QUAERO Project

*Markus Freitag, *Gregor Leusch, *Joern Wuebker, *Stephan Peitz, *Hermann Ney,
†Teresa Herrmann, †Jan Niehues, †Alex Waibel,
‡Alexandre Allauzen, ‡Gilles Adda,‡Josep Maria Crego,
§Bianka Buschbeck, §Tonio Wandmacher, §Jean Senellart
*RWTH Aachen University, Aachen, Germany
†Karlsruhe Institute of Technology, Karlsruhe, Germany
‡LIMSI-CNRS, Orsay, France
§SYSTRAN Software, Inc.
*surname@cs.rwth-aachen.de
†firstname.surname@kit.edu
‡firstname.lastname@limsi.fr §surname@systran.fr

Abstract

This paper describes the joint QUAERO submission to the WMT 2011 machine translation evaluation. Four groups (RWTH Aachen University, Karlsruhe Institute of Technology, LIMSI-CNRS, and SYSTRAN) of the QUAERO project submitted a joint translation for the WMT German→English task. Each group translated the data sets with their own systems. Then RWTH system combination combines these translations to a better one. In this paper, we describe the single systems of each group. Before we present the results of the system combination, we give a short description of the RWTH Aachen system combination approach.

1 Overview

QUAERO is a European research and development program with the goal of developing multimedia and multilingual indexing and management tools for professional and general public applications (http://www.quaero.org). Research in machine translation is mainly assigned to the four groups participating in this joint submission. The aim of this WMT submission was to show the quality of a joint translation by combining the knowledge of the four project partners. Each group develop and maintain their own different machine translation system. These single systems differ not only in their general approach, but also in the preprocessing of training and test data. To take the advantage of these differences of each translation system, we combined all hypotheses of the different systems, using the RWTH system combination approach.

1.1 Data Sets

For WMT 2011 each QUAERO partner trained their systems on the parallel Europarl and News Commentary corpora. All single systems were tuned on the newstest2009 dev set. The newstest2008 dev set was used to train the system combination parameters. Finally the newstest2010 dev set was used to compare the results of the different system combination approaches and settings.

2 Translation Systems

2.1 RWTH Aachen Single Systems

For the WMT 2011 evaluation the RWTH utilized RWTH's state-of-the-art phrase-based and hierarchical translation systems. GIZA++ (Och and Ney, 2003) was employed to train word alignments, language models have been created with the SRILM toolkit (Stolcke, 2002).

2.1.1 Phrase-Based System

The phrase-based translation (PBT) system is similar to the one described in Zens and Ney (2008). After phrase pair extraction from the word-aligned bilingual corpus, the translation probabilities are estimated by relative frequencies. The standard feature set also includes an n-gram language model, phrase-level IBM-1 and word-, phrase- and distortion-penalties, which are combined in log-linear fashion. Parameters are optimized with the Downhill-Simplex algorithm (Nelder and Mead, 1965) on the word graph.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 358–364,
Edinburgh, Scotland, UK, July 30–31, 2011.

2.1.2 Hierarchical System

For the hierarchical setups described in this paper, the open source Jane toolkit (Vilar et al., 2010) is employed. Jane has been developed at RWTH and implements the hierarchical approach as introduced by Chiang (2007) with some state-of-the-art extensions. In hierarchical phrase-based translation, a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. The search is typically carried out using the cube pruning algorithm (Huang and Chiang, 2007). The model weights are optimized with standard MERT (Och, 2003) on 100-best lists.

2.1.3 Phrase Model Training

For some PBT systems a forced alignment procedure was applied to train the phrase translation model as described in Wuebker et al. (2010). A modified version of the translation decoder is used to produce a phrase alignment on the bilingual training data. The phrase translation probabilities are estimated from their relative frequencies in the phrase-aligned training data. In addition to providing a statistically well-founded phrase model, this has the benefit of producing smaller phrase tables and thus allowing more rapid and less memory consuming experiments with a better translation quality.

2.1.4 Final Systems

For the German→English task, RWTH conducted experiments comparing the standard phrase extraction with the phrase training technique described in Section 2.1.3. Further experiments included the use of additional language model training data, reranking of n-best lists generated by the phrase-based system, and different optimization criteria.

A considerable increase in translation quality can be achieved by application of German compound splitting (Koehn and Knight, 2003). In comparison to standard heuristic phrase extraction techniques, performing force alignment phrase training (FA) gives an improvement in BLEU on newstest2008 and newstest2009, but a degradation in TER. The addition of LDC Gigaword corpora (+GW) to the language model training data shows improvements in both BLEU and TER. Reranking was done on 1000-best lists generated by the the best available system (PBT (FA)+GW). Following models were applied: n-gram posteriors (Zens and Ney, 2006), sentence length model, a 6-gram LM and IBM-1 lexicon models in both normal and inverse direction. These models are combined in a log-linear fashion and the scaling factors are tuned in the same manner as the baseline system (using TER$-$4BLEU on newstest2009).

The final table includes two identical Jane systems which are optimized on different criteria. The one optimized on TER$-$BLEU yields a much lower TER.

2.2 Karlsruhe Institute of Technology Single System

2.2.1 Preprocessing

We preprocess the training data prior to training the system, first by normalizing symbols such as quotes, dashes and apostrophes. Then smart-casing of the first words of each sentence is performed. For the German part of the training corpus we use the hunspell[1] lexicon to learn a mapping from old German spelling to new German spelling to obtain a corpus with homogeneous spelling. In addition, we perform compound splitting as described in (Koehn and Knight, 2003). Finally, we remove very long sentences, empty lines, and sentences that probably are not parallel due to length mismatch.

2.2.2 System Overview

The KIT system uses an in-house phrase-based decoder (Vogel, 2003) to perform translation. Optimization with regard to the BLEU score is done using Minimum Error Rate Training as described by Venugopal et al. (2005). The translation model is trained on the Europarl and News Commentary Corpus and the phrase table is based on a GIZA++ Word Alignment. We use two 4-gram SRI language models, one trained on the News Shuffle corpus and one trained on the Gigaword corpus. Reordering is performed based on continuous and non-continuous POS rules to cover short and long-range reorderings. The long-range reordering rules were also applied to the training corpus and phrase extraction was performed on the resulting reordering lattices. Part-of-speech tags are obtained using the TreeTag-

[1] http://hunspell.sourceforge.net/

ger (Schmid, 1994). In addition, the system applies a bilingual language model to extend the context of source language words available for translation. The individual models are described briefly in the following.

2.2.3 POS-based Reordering Model

We use a reordering model that is based on parts-of-speech (POS) and learn probabilistic rules from the POS tags of the words in the training corpus and the alignment information. In addition to continuous reordering rules that model short-range reordering (Rottmann and Vogel, 2007), we apply non-continuous rules to address long-range reorderings as typical for German-English translation (Niehues and Kolss, 2009). The reordering rules are applied to the source sentences and the reordered sentence variants as well as the original sequence are encoded in a word lattice which is used as input to the decoder.

2.2.4 Lattice Phrase Extraction

For the test sentences, the POS-based reordering allows us to change the word order in the source sentence so that the sentence can be translated more easily. If we apply this also to the training sentences, we would be able to extract also phrase pairs for originally discontinuous phrases and could apply them during translation of reordered test sentences.

Therefore, we build reordering lattices for all training sentences and then extract phrase pairs from the monotone source path as well as from the reordered paths. To limit the number of extracted phrase pairs, we extract a source phrase only once per sentence, even if it is found in different paths and we only use long-range reordering rules to generate the lattices for the training corpus.

2.2.5 Bilingual Language Model

In phrase-based systems the source sentence is segmented by the decoder during the search process. This segmentation into phrases leads to the loss of context information at the phrase boundaries. The language model can make use of more target side context. To make also source language context available we use a bilingual language model, an additional language model in the phrase-based system in which each token consist of a target word and all source words it is aligned to. The bilingual tokens enter the translation process as an additional target factor.

2.3 LIMSI-CNRS Single System

2.3.1 System overview

The LIMSI system is built with n-code[2], an open source statistical machine translation system based on bilingual n-grams.

2.3.2 n-code Overview

In a nutshell, the translation model is implemented as a stochastic finite-state transducer trained using a n-gram model of (source,target) pairs (Casacuberta and Vidal, 2004). Training this model requires to reorder source sentences so as to match the target word order. This is performed by a stochastic finite-state reordering model, which uses part-of-speech information[3] to generalize reordering patterns beyond lexical regularities.

In addition to the translation model, eleven feature functions are combined: a *target-language model;* four *lexicon models;* two *lexicalized reordering models* (Tillmann, 2004) aiming at predicting the orientation of the next translation unit; a weak distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system preference for short translations. The four *lexicon models* are similar to the ones use in a standard phrase based system: two scores correspond to the relative frequencies of the tuples and two lexical weights estimated from the automatically generated word alignments. The weights associated to feature functions are optimally combined using a discriminative training framework (Och, 2003), using the *newstest2009* data as development set.

The overall search is based on a beam-search strategy on top of a dynamic programming algorithm. Reordering hypotheses are computed in a preprocessing step, making use of reordering rules built from the word reorderings introduced in the tuple extraction process. The resulting reordering hypotheses are passed to the decoder in the form of word lattices (Crego and Mariño, 2007).

[2] `http://www.limsi.fr/Individu/jmcrego/n-code`

[3] Part-of-speech information for English and German is computed using the TreeTagger.

2.3.3 Data Preprocessing

Based on previous experiments which have demonstrated that better normalization tools provide better *BLEU* scores (K. Papineni and Zhu, 2002), all the English texts are tokenized and detokenized with in-house text processing tools (Déchelotte et al., 2008). For German, the standard tokenizer supplied by evaluation organizers is used.

2.3.4 Target n-gram Language Models

The English language model is trained assuming that the test set consists in a selection of news texts dating from the end of 2010 to the beginning of 2011. This assumption is based on what was done for the 2010 evaluation. Thus, a development corpus is built in order to create a vocabulary and to optimize the target language model.

Development Set and Vocabulary In order to cover different period, two development sets are used. The first one is *newstest2008*. However, this corpus is two years older than the targeted time period. Thus a second development corpus is gathered by randomly sampling bunches of 5 consecutive sentences from the provided news data of 2010 and 2011.

To estimate a LM, the English vocabulary is first defined by including all tokens observed in the Europarl and news-commentary corpora. This vocabulary is then expanded with all words that occur more that 5 times in the French-English giga-corpus, and with the most frequent proper names taken from the monolingual news data of 2010 and 2011. This procedure results in a vocabulary around 500k words.

Language Model Training All the training data allowed in the constrained task are divided into 9 sets based on dates on genres. On each set, a standard 4-gram LM is estimated from the 500k word vocabulary with in-house tools using absolute discounting interpolated with lower order models (Kneser and Ney, 1995; Chen and Goodman, 1998).

All LMs except the one trained on the news corpora from 2010-2011 are first linearly interpolated. The associated coefficients are estimated so as to minimize the perplexity evaluated on the *dev2010-2011*. The resulting LM and the 2010-2011 LM are finally interpolated with *newstest2008* as development data. This two steps interpolation aims to avoid an overestimate of the weight associated to the 2010-2011 LM.

2.4 SYSTRAN Software, Inc. Single System

The data submitted by SYSTRAN were obtained by the SYSTRAN baseline system in combination with a *statistical post editing* (SPE) component.

The SYSTRAN system is traditionally classified as a rule-based system. However, over the decades, its development has always been driven by pragmatic considerations, progressively integrating many of the most efficient MT approaches and techniques. Nowadays, the baseline engine can be considered as a linguistic-oriented system making use of dependency analysis, general transfer rules as well as of large manually encoded dictionaries (100k – 800k entries per language pair).

The basic setup of the SPE component is identical to the one described in (L. Dugast and Koehn, 2007). A statistical translation model is trained on the rule-based translation of the source and the target side of the parallel corpus. This is done separately for each parallel corpus. Language models are trained on each target half of the parallel corpora and also on additional in-domain corpora. Moreover, the following measures – limiting unwanted statistical effects – were applied:

- Named entities are replaced by special tokens on both sides. This usually improves word alignment, since the vocabulary size is significantly reduced. In addition, entity translation is handled more reliably by the rule-based engine.
- The intersection of both vocabularies (i.e. vocabularies of the rule-based output and the reference translation) is used to produce an additional parallel corpus (whose target is identical to the source). This was added to the parallel text in order to improve word alignment.
- Singleton phrase pairs are deleted from the phrase table to avoid overfitting.
- Phrase pairs not containing the same number of entities on the source and the target side are also discarded.

- Phrase pairs appearing less than 2 times were pruned.

The SPE language model was trained 15M phrases from the news/europarl corpora, provided as training data for *WMT 2011*. Weights for these separate models were tuned by the MERT algorithm provided in the Moses toolkit (P. Koehn et al., 2007), using the provided news development set.

3 RWTH Aachen System Combination

System combination is used to produce consensus translations from multiple hypotheses produced with different translation engines that are better in terms of translation quality than any of the individual hypotheses. The basic concept of RWTH's approach to machine translation system combination has been described by Matusov et al. (2006; 2008). This approach includes an enhanced alignment and reordering framework. A lattice is built from the input hypotheses. The translation with the best score within the lattice according to a couple of statistical models is selected as consensus translation. A deeper description will be also given in the WMT11 system combination paper of RWTH Aachen University. For this task only the A2L framework has been used.

4 Experiments

We tried different system combinations with different sets of single systems and different optimization criteria. As RWTH has two different translation systems, we put the output of both systems into system combination. Although both systems have the same preprocessing, their hypotheses differ. Finally, we added for both RWTH systems two additional hypotheses to the system combination. The two hypotheses of Jane were optimized on different criteria. The first hypothesis was optimized on BLEU and the second one on TER−BLEU. The first RWTH phrase-based hypothesis was generated with force alignment, the second RWTH phrase-based hypothesis is a reranked version of the first one as described in 2.1.4. Compared to the other systems, the system by SYSTRAN has a completely different approach (see section 2.4). It is mainly based on a rule-based system. For the German→English pair, SYSTRAN achieves a lower BLEU score in each test set compared to the other groups. But since the SYSTRAN system is very different to the others, we still obtain an improvement when we add it also to system combination.

We obtain the best result from system combination of all seven systems, optimizing the parameters on BLEU. This system was the system we submitted to the WMT 2011 evaluation.

For each dev set we obtain an improvement compared to the best single systems. For newstest2008 and newstest2009 we get an improvement of 0.5 points in BLEU and 1.8 points in TER compared to the best single system of Karlsruhe Institute of Technology. For newstest2010 we get an improvement of 1.8 points in BLEU and 2.7 points in TER compared to the best single system of RWTH. The system combination weights optimized for the best run are listed in Table 2. We see that although the single system of SYSTRAN has the lowest BLEU scores, it gets the second highest system weight. This high value shows the influence of a completely different system. On the other hand, all RWTH systems are very similar, because of their same preprocessing and their small variations. Therefor the system combination parameter of all four systems by themselves are relatively small. The summarized ”RWTH approach” system weight, though, is again on par with the other systems.

5 Conclusion

The four statistical machine translation systems of Karlsruhe Institute of Technology, RWTH Aachen and LIMSI and the very structural approach of SYSTRAN produce hypotheses with a huge variability compared to the others. Finally the RWTH Aachen system combination combined all single system hypotheses to one hypothesis with a higher BLEU compared to each single system. If the system combination implementation can handle enough single systems we would recommend to add all single systems to the system combination. Although the single system of SYSTRAN has the lowest BLEU scores and the RWTH single systems are similar we achieved the best result in using all single systems.

newstest2008		newstest2009		newstest2010		description
BLEU	TER	BLEU	TER	BLEU	TER	
22.73	60.73	22.50	59.82	25.26	57.37	sc (all systems) BLEU opt
22.61	60.60	22.28	59.39	25.07	56.95	sc (all systems - (1)) TER−BLEU opt
22.50	60.41	22.52	59.61	25.23	57.40	sc (all systems) TER−BLEU opt
22.19	60.09	22.05	59.31	24.74	56.89	sc (all systems - (4)) TER−BLEU opt
22.21	60.71	21.89	59.95	24.72	57.58	sc (all systems - (4,7)) TER−BLEU opt
22.22	60.45	21.79	59.72	24.32	57.59	sc (all systems - (3,4)) TER−BLEU opt
22.27	60.60	21.75	59.92	24.35	57.64	sc (all systems - (3,4)) BLEU opt
22.10	62.59	22.01	61.64	23.34	60.35	(1) Karlsruhe Institute of Technology
21.41	62.77	21.12	61.91	23.44	60.06	(2) RWTH PBT (FA) rerank +GW
21.11	62.96	21.06	62.16	23.29	60.26	(3) RWTH PBT (FA)
21.47	63.89	21.00	63.33	22.93	61.71	(4) RWTH jane + GW BLEU opt
20.89	61.05	20.36	60.47	23.42	58.31	(5) RWTH jane + GW TER−BLEU opt
20.33	64.50	19.79	64.91	21.97	61.44	(6) Limsi-CNRS
17.06	69.48	17.52	67.34	18.68	66.37	(7) SYSTRAN Software

Table 1: All systems for the WMT 2011 German→English translation task (truecase). BLEU and TER results are in percentage. FA denotes systems with phrase training, +GW the use of LDC data for the language model. sc denotes system combination.

system	weight
Karlsruhe Institute of Technology	0.350
RWTH PBT (FA) rerank +GW	0.001
RWTH PBT (FA)	0.046
RWTH jane + GW BLEU opt	0.023
RWTH jane + GW TER−BLEU opt	0.034
Limsi-CNRS	0.219
SYSTRAN Software	0.328

Table 2: Optimized systems weights for each system of the best system combination result.

Acknowledgments

This work was achieved as part of the QUAERO Programme, funded by OSEO, French State agency for innovation.

References

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

S.F. Chen and J.T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

D. Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

J.M. Crego and J.B. Mariño. 2007. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

D. Déchelotte, O. Galibert G. Adda, A. Allauzen, J. Gauvain, H. Meynard, and F. Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

L. Huang and D. Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

T. Ward K. Papineni, S. Roukos and W. Zhu. 2002. Bleu:

a method for automatic evaluation of machine translation. In *ACL '02: Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'95*, pages 181–184, Detroit, MI.

P. Koehn and K. Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of European Chapter of the ACL (EACL 2009)*, pages 187–194.

J. Senellart L. Dugast and P. Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 220–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40.

E. Matusov, G. Leusch, R.E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J.B. Mari no, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237.

J.A. Nelder and R. Mead. 1965. The Downhill Simplex Method. *Computer Journal*, 7:308.

J. Niehues and M. Kolss. 2009. A POS-Based Model for Long-Range Reorderings in SMT. In *Fourth Workshop on Statistical Machine Translation (WMT 2009)*, Athens, Greece.

F.J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

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

A. Birch P. Koehn, H. Hoang, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Rottmann and S. Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *TMI*, Skövde, Sweden.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, USA, September.

C. Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004*, pages 101–104. Association for Computational Linguistics.

A. Venugopal, A. Zollman, and A. Waibel. 2005. Training and Evaluation Error Minimization Rules for Statistical Machine Translation. In *Workshop on Data-drive Machine Translation and Beyond (WPT-05)*, Ann Arbor, MI.

D. Vilar, S. Stein, M. Huck, and H. Ney. 2010. Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

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

J. Wuebker, A. Mauser, and H. Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden, July.

R. Zens and H. Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 72–77, New York City, June.

R. Zens and H. Ney. 2008. Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Honolulu, Hawaii, October.