Hybrid Machine Translation Architectures within and beyond the EuroMatrix project

Andreas Eisele[1,2], Christian Federmann[2], Hans Uszkoreit[1,2], Hervé Saint-Amand[1], Martin Kay[1,3], Michael Jellinghaus[1], Sabine Hunsicker[1], Teresa Herrmann[1], Yu Chen[1]

1: Saarland University, Saarbrücken, Germany
2: DFKI GmbH, Saarbrücken, Germany
3:Stanford University, USA

Abstract. This paper presents two hybrid architectures combining rule-based and statistical machine translation (RBMT and SMT) technology. In the first case, several existing MT engines are combined in a multi-engine setup, and a decoder for SMT is used to select and combine the best expressions proposed by different engines. The other architecture uses lexical entries found using a combination of SMT technology with shallow linguistic processing, which are then included in a rule-based system. The first architecture has been implemented in the framework of the EuroMatrix project, where results from the recent evaluation campaign give important insights into the strengths and weaknesses of the approach relative to other participating RBMT and SMT systems. The second architecture has been developed in collaboration with the European Patent Office who uses the enhanced RBMT system in the framework of their European Machine Translation Project.

1 Introduction

Recent work on statistical machine translation has led to significant progress in coverage and quality of translation technology [1, 2] but so far, most of this work has focused on translation into English, where relatively simple morphological structure and abundance of monolingual training data helped to compensate for the relative lack of linguistic sophistication of the underlying models. As SMT systems are trained on massive amounts of data, they are typically quite good at capturing implicit knowledge contained in co-occurrence statistics, which can serve as a shallow replacement for the world knowledge that would be required for the resolution of ambiguities and the insertion of information that happens to be missing in the source text but is required to generate well-formed text in the target language. In previous years, decades of work went into the implementation of MT systems (typically rule-based) for frequently used language pairs[1], and these systems quite often contain a wealth of linguistic knowledge about the languages involved, such as fairly complete mechanisms for morphological

[1] See [3] for an extensive, regularly updated list of commercial MT systems

and syntactic analysis and generation, as well as a large number of bilingual lexical entries spanning many application domains. It is an interesting challenge to combine the different types of knowledge into integrated systems that could then exploit both linguistic knowledge contained in the rules of one or several conventional MT system(s) and non-linguistic knowledge that can be extracted from large amounts of text. The EuroMatrix project (see www.euromatrix.net) has been exploring such combinations of rule-based and statistical knowledge sources, one of the approaches being an integration of existing rule-based MT systems into a multi-engine architecture. This paper describes several incarnations of such multi-engine architectures within the project. A careful analysis of the results will guide us in the choice of further steps towards the construction of hybrid MT systems for practical applications.

2 Merging multiple MT results via a SMT decoder

2.1 Architecture

Combinations of MT systems into multi-engine architectures have a long tradition, starting perhaps with [4]. Multi-engine systems can be roughly divided into simple architectures that try to select the best output from a number of systems but leave the individual hypotheses as is on the one hand [5–10], and more sophisticated setups on the other hand that try to recombine the best parts from multiple hypotheses into a new utterance that can be better than the best of the given candidates, as described in [11–16]. Recombining multiple MT results requires finding the correspondences between alternative renderings of a source-language expression proposed by different MT systems. This is generally not straightforward, as different word order and errors in the output can make it hard to identify the alignment. Still, we assume that a good way to combine the various MT outcomes will need to involve word alignment between the MT output and the given source text, and hence a specialized module for word alignment is a central component of our setup. Additionally, a recombination system needs a way to pick the best combination of alternative building blocks. When judging the quality of a particular configuration, both the plausibility of the building blocks as such and their relation to the context need to be taken into account. The required optimization process is very similar to the search in a SMT decoder that looks for naturally sounding combinations of highly probable partial translations. Instead of implementing a special-purpose search procedure from scratch, we transform the information contained in the MT output into a form that is suitable as input for an existing SMT decoder. This has the additional advantage that it is simple to combine resources used in standard phrase-based SMT with the material extracted from the rule-based MT results; the optimal combination can essentially be reduced to the task of finding good relative weights for the various phrase table entries. This architecture is described in more detail in [17], where also examples of results are given. It should be noted that this is certainly not the only way to combine systems. In particular, as this proposed

setup gives the last word to the SMT decoder, there is the risk that linguistically well-formed constructs from one of the rule-based engines will deteriorate in the final decoding step. Alternative architectures are under exploration and one such approach will be described below. For experiments in the framework of the shared task of the 2008 ACL workshop on SMT [17] we used a set of six rule-based MT engines that are partly available via web interfaces and partly installed locally. In addition to these engines, we generated phrase tables from the training data following the baseline methodology given in the description of the shared task and using the scripts included in the Moses toolkit [18]. In order to improve alignment quality, the source text and the output text of the MT systems were aligned with the help of a modified version of GIZA++ that it is able to load given models and which is embedded into a client-server setup, as described in [19]. The original Moses phrase table and separate phrase tables for each of the RBMT systems were then combined into a unified phrase table. By combining domain-specific lexical knowledge learned from the training data with more general knowledge contained in the linguistic rules, the hybrid system can both handle a wider range of syntactic constructions and exploit knowledge that the RBMT systems possess about the particular vocabulary of the source text.

2.2 Results

We submitted the results of the hybrid system as well as the results from each of the rule-based systems (suitably anonymized) to the shared task of the WMT 2008 workshop. This gives us the opportunity to compare the results with many other systems under fair conditions, both using automatic evaluation metric and comparisons involving human inspection. Detailed results of this evaluation are

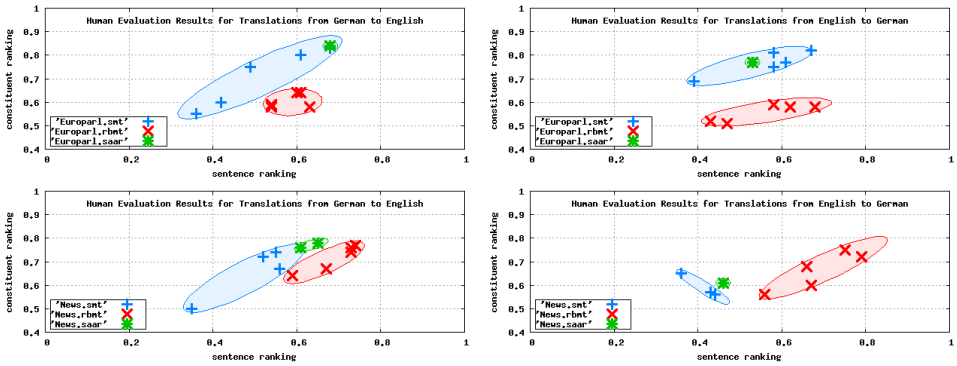

Fig. 1. Relative performance of system types for in-domain (EuroParl, upper row) and out-of-domain (News, lower row) data

documented in [20]. By condensing several of the tables into a joint plot, it

becomes easier to see some of the salient patterns contained in these datasets. Fig. 1 project the results of two different types of human evaluation into two-dimensional plots, and it is interesting to study the different behavior of the systems that depend strongly on whether the tests are done on data from the same or from a different domain as the training data. The plot displays the relative performance of the systems for the directions German ↔ English according to sentence ranking and constituent ranking. We do not reveal the identity of the systems but cluster them into SMT systems, RBMT engines, and our hybrid combination. As long as testing is done in domain, with English as the target language, the statistical approaches can adapt to the domain's typical expressions, and the best statistical systems are better than the best RBMT system in sentence ranking and much better in constituent ranking. For tests in a different domain, the rule-based systems are somewhat better than SMTs in sentence ranking but have only a very slight advantage in constituent ranking. Under both scenarios, the hybrid combination behaves similar to the SMT system but can obtain a slight improvement from the larger lexicon. For translations into German, RBMT systems generally perform better, but our hybrid architecture is currently not able not preserve this advantage over the SMT approach.

3 Feeding SMT phrases into a rule-based MT system

3.1 Motivation and Architecture

The architecture described in the previous section places a strong emphasis on the statistical models and can be seen as a variant of SMT where lexical information from rule-based engines is used to increase lexical coverage. However, it is also true that rule-based MT engines frequently suffer from insufficient lexical coverage while the ability of SMT to automatically induce lexical entries from existing translations can be seen as one of its key advantages. It is therefore interesting to investigate how automatically extracted lexical knowledge can be used to increase the coverage of a rule-based MT system. Such an arrangement leaves the control of the translation process with the rule-based engine, which has the advantage that well-formed syntactic structures generated by linguistic rules cannot be broken up by the SMT components. But as rule-based systems typically lack mechanisms for ruling out implausible results, they cannot easily cope with errors that enter the lexicon from misalignments, with examples that fail to generalize, and with expressions that strongly depend on the given context. Entries derived from statistical alignments need therefore be carefully filtered to keep the error rate at an acceptable level. Furthermore, the information that can be extracted from word alignments lacks the linguistic information required in rule-based systems. While corresponding expressions in a parallel corpus occur as fully inflected forms, the entries in a bilingual dictionary are normalized forms with morphological information defining all possible inflected forms. Even if a parallel corpus happens to include different inflected forms of a lemma, the collection of forms is a (typically very incomplete) random sample of the full paradigm, and it is therefore not always possible to derive all the

inflected forms of the lemma. Despite these additional difficulties, an infrastructure for the extraction of lexical entries was built up in the framework of a joint project between the DFKI and the European Patent Office (EPO), where the EPO wants to provide a translation facility for patent documents, which will be available to their examiners and eventually also to the general public. The translation itself is performed by an external service provider, using a rule-based MT engine, whereas the contribution of DFKI is the extraction and manual validation of additional lexical entries for the relevant technical fields. The architecture combines several modules from statistical and rule-based approaches to MT. On the one hand, parallel texts are sent through the statistical alignment machinery, based on GIZA++ that is also used to obtain word and phrase alignments in SMT. On the other hand, the texts are linguistically enriched by part of speech (PoS) tags and lemma information. The two representations are then combined and filters based on PoS sequences on both sides are used to obtain a set of candidates for the lexicon. A list of acceptable pairs of PoS sequences is generated by inspecting several hundred of the most frequently occurring PoS sequences and excluding those that either do not form a pair of linguistic phrases or where the interpretation on both sides is incompatible. Morphological classification is applied to these lexical entries to augment them with inflection classes, following the open lexicon interchange format (OLIF) standard [21]. Even if statistical alignment and linguistic preprocessing can lead us a long way towards the automatic creation of lexical entries, it is crucial to manually inspect and correct the results because rule-based MT systems have no other mechanism for preventing errors from incorrect lexical entries. In cases of technical terminology, the validation of the terminology requires both linguistic and technical competence and it may be necessary to distribute some steps over different groups of people. In order to facilitate this process, we have built up a web-based front end for lexical database maintenance such that the extracted lexical entries are stored in a centralized way and various parts of validation and quality control can be distributed over arbitrary workplaces that have access to the internet. The validation workflow consists of several steps where entries are first checked for monolingual wellformedness and properties like morphological head, gender, inflectional class, and the possibility of plural forms. This part of the interface is built such that the validator does not see internal codes for the inflectional class but rather a small set of distinctive full forms and has an option to correct these. In a second round, the corresponding forms from two languages are shown in combination and the validator can rule out those cases where the forms do not convey the same meaning. This round also deals with disambiguation; whereas generally for a term in one language the most frequently appearing translation in the same subject domain is used, the validator has the option of disapproving certain expressions. Disallowed expressions will then still be understood in the source text but avoided in the target text in favor of expressions that appeared less frequently. In a third round, the DB interface is used by representatives of the participating patent authorities for quality control by domain experts.

3.2 Results and Application

The proposed architecture was used to create translation dictionaries with technical vocabulary for all four language directions (EN paired with ES or DE, in both directions). In a first round of extraction work, about 40 million English-German sentence pairs and about 10 million English-Spanish sentence pairs have been processed and 2.3 million candidates for English-German term pairs as well as 0.8 million candidates for English-Spanish term pairs have been identified. About 90% of the extracted entries are pairs of noun phrases, which typically consist of multi-word expressions (MWEs) involving one or more adjectives or noun compounds. Often, English MWE correspond to one long German word, e.g. Empfängnisverhütungsmittelzusammensetzung (= contraceptive composition). An evaluation by the EPO showed that a significant subset of the identified term pairs are either correct or close enough to correct lexical entries that manual validation or correction was worthwhile. For each direction, 60000 lexical entries were selected by the EPO and manually validated by linguists at DFKI. Similar efforts for French and Italian are currently ongoing. As the entries are derived from documents for which the technical domain is known, it is possible to annotate the entries with the frequencies with which this translation is encountered in documents from this particular domain. Using this simple mechanism, it is possible to use knowledge of the IPC class of the source document to select the most appropriate translation of a given term in the source language. Comparisons of the translation quality with and without the automatically derived translation entries revealed a significant increase in lexical coverage using our model. The translation service has been made publicly available by the EPO and has processed more than 180000 documents by September 2007 [22].

4 Conclusion and Outlook

So far, we have presented two almost complementary ways to combine rule-based and statistical approaches to MT by integrating existing implementations into a larger architecture. In one case, rule-based MT engines are used to enrich the lexical resources available to the SMT decoder. In the other case, parts of the SMT infrastructure are used, together with linguistic processing and manual validation, to extend the lexicon of a rule-based MT engine. Both approaches have been implemented and show promising improvements to MT quality but as they are currently still in a somewhat prototypical state, it is still too early to give meaningful comparative evaluations. A further popular approach to the construction of hybrid MT architectures not discussed so far addresses the problem that the output of RBMT engines often sounds less natural and fluent in comparison with typical SMT results because standard RBMT approaches do not have access to statistical language models which are the main source of fluency (at least on a local, n-gram level) in the typical SMT setup. A fluency model can be integrated into a RBMT-based architecture via post-editing. This allows the replacement of output expressions by alternatives that fit the context better in the target language. A series of papers has explored this approach both within and beyond

the EuroMatrix project [23, 24], and results of such systems have been submitted to the shared task of the WMT08 workshop [20]. [19] investigates the effect of post-editing on the frequency of typical error types along an error classification inspired by [25] and compares BLEU scores with the results of the architecture proposed in Section 2. Similar types of evaluations are currently going on for more language pairs. Automatic post-editing of MT results can be applied to both architectures presented above and could be used to reduce the impact of disfluencies of the raw MT results. However, it should be clear that even if one or both of these approaches can be made to deliver significant improvements under fairly general conditions, the improvements will essentially only alleviate the problem of lexical coverage but will not touch some other well-known issues with the respective frameworks. One of the key problems of rule-based MT systems is their difficulty to deal with soft rules and preferences that are required for disambiguation and for picking the most natural expressions in the target language. Conversely, today's versions of SMT have obvious difficulties delivering syntactically well-formed utterances, especially when relevant constraints reach beyond the window size of the target language models. It is conceivable that a deeper integration of rule-based linguistic knowledge with corpus-based evidence will eventually be able to alleviate both problems in one integrated system. However, this will require an architecture that has simultaneous access to all relevant types of knowledge, which is beyond the relatively simple hybrid architectures presented here.

Acknowledgments

This work was supported by the EuroMatrix project funded by the European Commission (6th Framework Programme). Parts were supported by a contract between the European Patent Office and DFKI. We thank the colleagues in both projects for their invaluable help that made this work possible. We thank John Hutchins for suggesting significant improvements of the presentation.

References

1. Callison-Burch, C., Koehn, P.: Introduction to statistical machine translation,. In: European Summer School for Language and Logic (ESSLLI). (2005)
2. Koehn, P., Monz, C.: Shared task: Exploiting parallel texts for statistical machine translation. In: Proceedings of the NAACL 2006 workshop on statistical machine translation, New York City (June 2006)
3. Hutchins, J.: IAMT compendium of translation software. 14th Edition (January 2008)
4. Frederking, R.E., Nirenburg, S.: Three heads are better than one. In: ANLP. (1994) 95–100
5. Tidhar, D., Küssner, U.: Learning to select a good translation. In: COLING. (2000) 843–849
6. Akiba, Y., Imamura, K., Sumita, E.: Using multiple edit distances to automatically rank machine translation output. In: Proceedings of MT Summit VIII, Santiago de Compostela, Spain (2001)

7. Callison-Burch, C., Flournoy, R.S.: A program for automatically selecting the best output from multiple machine translation engines. In: Proc. of MT Summit VIII, Santiago de Compostela, Spain (2001)
8. Akiba, Y., Watanabe, T., Sumita, E.: Using language and translation models to select the best among outputs from multiple mt systems. In: COLING. (2002)
9. Nomoto, T.: Multi-engine machine translation with voted language model. In: Proc. of ACL. (2004)
10. Eisele, A.: First steps towards multi-engine machine translation. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts. (June 2005)
11. Rayner, M., Carter, D.M.: Hybrid language processing in the spoken language translator. In: Proc. ICASSP '97, Munich, Germany (1997) 107–110
12. Hogan, C., Frederking, R.E.: An evaluation of the multi-engine MT architecture. In: Proceedings of AMTA. (1998) 113–123
13. Bangalore, S., Bordel, G., Riccardi, G.: Computing consensus translation from multiple machine translation systems. In: ASRU, Italy (2001)
14. Jayaraman, S., Lavie, A.: Multi-engine machine translation guided by explicit word matching. In: Proc. of EAMT, Budapest, Hungary (2005)
15. Matusov, E., Ueffing, N., Ney, H.: Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In: In Proc. EACL. (2006) 33–40
16. Rosti, A.V., Ayan, N.F., Xiang, B., Matsoukas, S., Schwartz, R., Dorr, B.J.: Combining translations from multiple machine translation systems. In: Proceedings of HLT-NAACL, Rochester, NY (April 22-27 2007) 228–235
17. Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., Chen, Y.: Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, ACL (June 2008) 179–182
18. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proc. of ACL Demo and Poster Sessions. (Jun 2007) 177–180
19. Theison, S.: Optimizing rule-based machine translation output with the help of statistical methods. Diploma thesis, Saarland University (2007)
20. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, ACL (June 2008) 70–106
21. Lieske, C., McCormick, S., Thurmair, G.: The open lexicon interchange format (olif) comes of age. In: Proceedings of MT Summit VIII: Machine Translation in the Information Age, Santiago de Compostela, Spain (September 2001) 211–216
22. Täger, W.: The European Machine Translation Programme. In: MT Summit XI Workshop on Patent Translation, Copenhagen (September 2007)
23. Dugast, L., Senellart, J., Koehn, P.: Statistical post-editing on SYSTRAN's rule-based translation system. In: Proceedings of WMT07, Prague, Czech Republic, Association for Computational Linguistics (June 2007) 220–223
24. Simard, M., Ueffing, N., Isabelle, P., Kuhn, R.: Rule-based translation with statistical phrase-based post-editing. In: Proceedings of WMT07, Prague, Czech Republic, Association for Computational Linguistics (June 2007) 203–206
25. Vilar, D., Xu, J., D'Haro, L.F., Ney, H.: Error analysis of statistical machine translation output. In: Proceedings of LREC 2006, Genoa (Mai 2006)