Further Experiments with Shallow Hybrid MT Systems

Christian Federmann[1], **Andreas Eisele**[1], **Hans Uszkoreit**[1,2],
Yu Chen[1], **Sabine Hunsicker**[1], **Jia Xu**[1]
1: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Saarbrücken, Germany
2: Universität des Saarlandes, Saarbrücken, Germany
{cfedermann,eisele,uszkoreit,yuchen,sabine.hunsicker,jia.xu}@dfki.de

Abstract

We describe our hybrid machine translation system which has been developed for and used in the WMT10 shared task. We compute translations from a rule-based MT system and combine the resulting translation "templates" with partial phrases from a state-of-the-art phrase-based, statistical MT engine. Phrase substitution is guided by several decision factors, a continuation of previous work within our group. For the shared task, we have computed translations for six language pairs including English, German, French and Spanish. Our experiments have shown that our shallow substitution approach can effectively improve the translation result from the RBMT system; however it has also become clear that a deeper integration is needed to further improve translation quality.

1 Introduction

In recent years the quality of machine translation (MT) output has improved greatly, although each paradigm suffers from its own particular kind of errors: statistical machine translation (SMT) often shows poor syntax, while rule-based engines (RBMT) experience a lack in vocabulary. Hybrid systems try to avoid these typical errors by combining techniques from both paradigms in a most useful manner.

In this paper we present the improved version of the hybrid system we developed last year's shared task (Federmann et al., 2009). We take the output from an RBMT engine as basis for our hybrid translations and substitute noun phrases by translations from an SMT engine. Even though a general increase in quality could be observed, our system introduced errors of its own during the substitution process. In an internal error analysis, these degradations were classified as follows:

- the translation by the SMT engine is incorrect
- the structure degrades through substitution (because of e.g. capitalization errors, double prepositions, etc.)
- the phrase substitution goes astray (caused by alignment problems, etc.)

Errors of the first class cannot be corrected, as we have no way of knowing when the translation by the SMT engine is incorrect. The other two classes could be eliminated, however, by introducing additional steps for pre- and post-processing as well as improving the hybrid algorithm itself. Our current error analysis based on the results of this year's shared task does not show these types of errors anymore.

Additionally, we extended our coverage to also include the language pairs English↔French and English↔Spanish in both directions as well as English→German, compared to last year's initial experiments for German→English only. We were able to achieve an increase in translation quality for this language set, which shows that the substitution method works for different language configurations.

2 Architecture

Our hybrid translation system takes translation output from a) the Lucy RBMT system (Alonso and Thurmair, 2003) and b) a Moses-based SMT system (Koehn et al., 2007). We then identify noun phrases inside the rule-based translation and compute the most likely correspondences in the statistical translation output. For these, we apply a factored substitution method that decides whether the original RBMT phrase should be kept or rather be replaced by the Moses phrase. As this shallow substitution process may introduce problems at

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 77–81,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

phrase boundaries, we afterwards perform several post-processing steps to cleanup and finalize the hybrid translation result. A schematic overview of our hybrid system and its main components is given in figure 1.

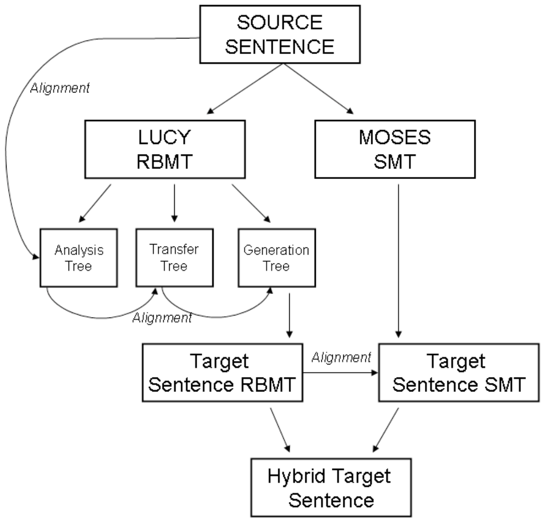

Figure 1: Schematic overview of the hybrid MT system architecture.

2.1 Input to the Hybrid System

Lucy RBMT System We obtain the translation as well as linguistic structures from the RBMT system. An internal evaluation has shown that these structures are usually of a high quality which supports our initial decision to consider the RBMT output as an appropriate "template" for our hybrid translation approach. The Lucy translation output can include additional markup that allows to identify unknown words or other, local phenomena.

The Lucy system is a transfer-based MT system that performs translation in three phases, namely *analysis*, *transfer*, and *generation*. Intermediate tree structures for each of the translation phases can be extracted from the Lucy system to guide the hybrid system. Sadly, only the 1-best path through these three phases is given, so no alternative translation possibilities can be extracted from the given data; a fact that clearly limits the potential for more deeply integrated hybrid translation approaches. Nevertheless, the availability of the 1-best trees already allows to improve the translation quality of the RBMT system as we will show in this paper.

Moses SMT System We used a state-of-the-art Moses SMT system to create statistical phrase-based translations of our input text. Moses has been modified so that it returns the translation results together with the bidirectional word alignments between the source texts and the translations. Again, we make use of markup which helps to identify unknown words as these will later guide the factored substitution method. Both of the translation models and the language models within our SMT systems were only trained with lower-cased and tokenized Europarl training data. The system used sets of feature weights determined using data sets also from Europarl (test2008). In addition, we used LDC gigaword corpus to train large scale n-gram language models to be used in our hybrid system. We tokenized the source texts using the standard tokenizers available from the shared task website. The SMT translations are re-cased before being fed into the hybrid system together with the word alignment information.The hybrid system can easily be adapted to support other statistical translation engines. If the alignment information is not available, a suitable alignment tool would be necessary to compute it as the alignment is a key requirement for the hybrid system.

2.2 Aligning RBMT and SMT Output

We compute alignment in several components of the hybrid system, namely:

source-text-to-tree: we first find an alignment between the source text and the corresponding analysis tree(s). As Lucy tends to subdivide large sentences into several smaller units, it sometimes becomes necessary to align more than one tree structure to a given source sentence.

analysis-transfer-generation: for each of the analysis trees, we re-construct the path from its tree nodes, via the transfer tree, and their corresponding generation tree nodes.

tree-to-target-text: similarly to the first alignment process, we find a mapping between generation tree nodes and the actual translation output of the RBMT system.

source-text-to-tokenized: as the Lucy RBMT system works on non-tokenized input text and our Moses system takes tokenized input, we need to align the source text to its tokenized form.

Given the aforementioned alignments, we can then correlate phrases from the rule-based translation with their counterparts from the statistical translation, both on source or target side. As our hybrid approach relies on the identification of such phrase pairs, the computation of the different alignments is critical to obtain good combination performance.

Please note that all these tree-based alignments can be computed with a very high accuracy. However, due to the nature of statistical word alignment, the same does not hold for the alignment obtained from the Moses system. If the alignment process has produced erroneous phrase tables, it is very likely that Lucy phrases and their "aligned" SMT matches simply will not fit. Or put the other way round: the better the underlying SMT word alignment, the greater the potential of the hybrid substitution approach.

2.3 Factored Substitution

Given the results of the alignment process, we can then identify "interesting" phrases for substitution. Following our experimental setup from last year's shared task, we again decided to focus on *noun phrases* as these seem to be best-suited for in-place swapping of phrases. Our initial assumption is that SMT phrases are better on a lexical level, hence we aim to replace Lucy's noun phrases by their Moses counterparts.

Still, we want to perform the substitution in a controlled manner in order to avoid problems or non-matching insertions. For this, we have (manually) derived a set of *factors* that are checked for each of the phrase pairs that are processed. The factors are described briefly below:

identical? simply checks whether two candidate phrases are identical.

too complex? a Lucy phrase is "too complex" to substitute if it contains more than 2 embedded noun phrases.

many-to-one? this factor checks if a Lucy phrase containing more than one word is mapped to a Moses phrase with only one token.

contains pronoun? checks if the Lucy phrase contains a pronoun.

contains verb? checks if the Lucy phrase contains a verb.

unknown? checks whether one of the phrases is marked as "unknown".

length mismatch computes the number of words for both phrases and checks if the absolute difference is too large.

language model computes language model scores for both phrases and checks which is more likely according to the LM.

All of these factors have been designed and adjusted during an internal development phase using data from previous shared tasks.

2.4 Post-processing Steps

After the hybrid translation has been computed, we perform several post-processing steps to clean up and finalize the result:

cleanup first, we perform basic cleanup operations such as whitespace normalization, capitalizing the first word in each sentence, etc.

multi-words then, we take care of proper handling of multi-word expressions. Using the tree structures from the RBMT system we eliminate superfluous whitespace and join multi-words, even if they were separated in the SMT phrase.

prepositions finally, we give prepositions a special treatment. Experience from last year's shared task had shown that things like double prepositions contributed to a large extent to the amount of avoidable errors. We tried to circumvent this class of error by identifying the correct prepositions; erroneous prepositions are removed.

3 Hybrid Translation Analysis

We evaluated the intermediate outputs using BLEU (Papineni et al., 2001) against human references as in table 3. The BLEU score is calculated in lower case after the text tokenization. The translation systems compared are Moses, Lucy, Google and our hybrid system with different configurations:

Hybrid: we use the language model with case information and substitute some NPs in Lucy outputs by Moses outputs.

Hybrid LLM: same as Hybrid but we use a larger language model.

Table 1: Intermediate results of BLEU[%] scores for WMT10 shared task.

System	de→en	en→de	fr→en	en→fr	es→en	en→es
Moses	18.32	12.66	22.26	20.06	24.28	24.72
Lucy	16.85	12.38	18.49	17.61	21.09	20.85
Google	25.64	18.51	28.53	28.70	32.77	32.20
Hybrid	17.29	13.05	18.92	19.58	22.53	23.55
Hybrid LLM	17.37	13.73	18.93	19.76	22.61	23.66
Hybrid SG	17.43	14.40	19.67	20.55	24.37	24.99
Hybrid NCLM	17.38	14.42	19.56	20.55	24.41	24.92

Hybrid SG: same as Hybrid but the NP substitutions are based on Google output instead of Moses translations.

Hybrid NCLM: same as Hybrid but we use the language model without case information.

We participated in the translation evaluation in six language pairs: German to English (de→en), English to German (en→de), French to English (fr→en), English to French (en→fr), Spanish to English (es→en) and English to Spanish (en→es). As shown in table 3, the Moses translation system achieves better results overall than the Lucy system does. Google's system outperforms other systems in all language pairs. The hybrid translation as described in section 2 improves the Lucy translation quality with a BLEU score up to 2.7% absolutely.

As we apply a larger language model or a language model without case information, the translation performance can be improved further. One major problem in the hybrid translation is that the Moses outputs are still not good enough to replace the Lucy outputs, therefore we experimented on a hybrid translation of Google and Lucy systems and substitute some unrelaible NP translations by the Google's translations. The results in the line of 'Hybrid SG' shows that the hybrid translation quality can be enhanced if the translation system where we select substitutions is better.

4 Internal Evaluation of Results

In the analysis of the remaining issues, the following main sources of problems can be distinguished:

- Lucy's output contains structural errors that cannot be fixed by the chosen approach.
- Lucy results contain errors that could have been corrected by alternative expressions from SMT, but the constraints in our system were too restrictive to let that happen.
- The SMT engine we use generates suboptimal results that find their way into the hybrid result.
- SMT results that are good are incorporated into the hybrid results in a wrong way.

We have inspected a part of the results and classified the problems according to these criteria. As this work is still ongoing, it is too early to report numerical results for the relative frequencies of the different causes of the error. However, we can already see that three of these four cases appear frequently enough to justify further attention. We observed several cases in which the parser in the Lucy system was confused by unknown expressions and delivered results that could have been significantly improved by a more robust parsing approach. We also encountered several cases in which an expression from SMT was used although the original Lucy output would have been better. Also we still observe problems finding to correct correspondences between Lucy output and SMT output, which leads to situations where material is inserted in the wrong place, which can lead to the loss of content words in the output.

5 Conclusion and Outlook

In our contribution to the shared task we have applied the hybrid architecture from (Federmann et al., 2009) to six language pairs. We have identified and fixed many of the problems we had observed last year, and we think that, in addition to the increased coverage in laguage pairs, the overall quality has been significantly increased.

However, in the last section we characterized three main sources of problems that will require further attention. We will address these problems in the near future in the following way:

1. We will investigate in more detail the alignment issue that leads to occasional loss of content words, and we expect that a careful inspection and correction of the code will in all likelihood give us a good remedy.

2. The problem of picking expressions from the SMT output that appear more probable to the language model although they are inferior to the original expression from the RBMT system is more difficult to fix. We will try to find better thresholds and biases that can at least reduce the number of cases in which this type of degradation happen.

3. Finally, we will also address the robustness issue that leads to suboptimal structures from the RBMT engine caused by parsing failures.

Our close collaboration with Lucy enables us to address these issues in a very effective way via the inspection and classification of intermediate structures and, if these structures indicate parsing problems, the generation of variants of the input sentence that facilitate correct parsing.

Acknowledgments

This work was supported by the EuroMatrixPlus project (IST-231720) which is funded by the European Commission under the Seventh Framework Programme. The authors want to thank Michael Jellinghaus and Bastian Simon for help with the inspection of intermediate results and classification of errors.

References

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proc. of the Ninth MT Summit*.

Christian Federmann, Silke Theison, Andreas Eisele, Hans Uszkoreit, Yu Chen, Michael Jellinghaus, and Sabine Hunsicker. 2009. Translation combination using factored word substitution. In *Proceedings of the Fourth Workshop on Statistical Machine Transla tion*, pages 70–74, Athens, Greece, March. Association for Computational Linguistics.

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 *Proc. of ACL Demo and Poster Sessions*, pages 177–180, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report RC22176(W0109-022), IBM.