DFKI System Combination using Syntactic Information at ML4HMT-2011

Christian Federmann, Sabine Hunsicker, Yu Chen, Rui Wang
Language Technology Lab
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
{cfedermann,sabine.hunsicker,yuchen,wang.rui}@dfki.de

Abstract

We present a substitution approach for the combination of machine translation outputs. Using a translation template derived from the output obtained from a rule-based translation engine, we identify parts of the template that could possibly be improved by adding in segments from other MT output. Substitution candidates are determined based on their part-of-speech. Alternative translations from the additional engines are retrieved by using word alignment. Substitution is based on several decision factors, such as part-of-speech, local left-/right-context, and language model probabilities. Our approach differs from other methods as it puts its main focus on preserving the syntactic structure inherited from the rule-based translation template. For the language pair Spanish-English an improvement in BLEU score can be observed.

1 Introduction

Statistical machine translation (SMT) systems have seen a lot of research progress during the last decade. They have effectively outperformed many existing, rule-based machine translation approaches due to their data-driven nature: SMT systems can be trained on large, parallel data sets and they can be tuned according to automated scoring metrics. This is often impossible for rule-based MT (RBMT) engines, in particular if they rely on hand-crafted rules and if they do not involve an overall probability model. This clearly indicates that such systems can profit from further research to catch up with and perhaps even beat current state-of-the-art statistical systems.

Rule-based translation output can have certain advantages over statistically translated content: the syntactic structure of the output is usually correct and complete and the word forms are properly generated. While this is often not fully reflected by standard automatic evaluation metrics such as BLEU (Papineni et al., 2001), it sometimes shows in manual evaluation where human evaluators notice the syntactic quality (i.e., grammaticality) of the output and rank RBMT output better than the automatic scores.

It is interesting to note that recently rule-based systems were able to outperform their statistical opponents in several open evaluation events (Callison-Burch et al., 2009; Callison-Burch et al., 2011). Furthermore, different machine translation paradigms seem to produce output containing complementary errors (Thurmair, 2009). Hence, it makes sense to search for effective ways of combining different systems in order to benefit from the respective advantages of different paradigms while trying to avoid their individual shortcomings. Therefore, we are more focusing on integrating systems of different types instead of applying general system combination techniques because previous results showing correlations between systems suggest that combining them has a great impact on the performance of the combined results (Macherey and Och, 2007).

Previous approaches on system combination include, among others, direct selection from the candidate translations (Callison-Burch and Flournoy, 2001; Akiba et al., 2001), combining word lattices

or n-best lists (Frederking and Nirenburg, 1994), hypothesis regeneration with an SMT decoder (Chen et al., 2007; Eisele et al., 2008; Chen et al., 2009) and ROVER-like voting schemes on confusion networks (Jayaraman and Lavie, 2005; Matusov et al., 2006; Rosti et al., 2007; He et al., 2008; Leusch et al., 2009). The last approach constructs a confusion network based on pairwise word alignments of the translation hypothesis, which might be re-ordered. The voting module selects the best consensus translation from the confusion network based on several statistical models. The target language model plays an important role in the voting procedure. It is very likely that the final translation does not resemble any of the hypotheses from the individual systems.

In this shared task, we follow the constituent substitution approach for system combination proposed by (Federmann et al., 2009). The substitution method is similar to voting on a confusion network that has a fixed backbone, however taking more linguistic information into account. Similar work has been reported in (Habash et al., 2009; Espana-Bonet et al., 2011). We choose the translations from an RBMT system as our fixed backbones, or "translation templates" in the hope of retaining the better syntactic structures created by such a system. The consensus translation is then produced by replacing complete constituents in the translation template rather than isolated words. Corresponding phrases in the other candidate translations are identified through word alignments back to the original source sentences. Our substitution algorithm is guided by several decision factors, including part-of-speech, local context, and a language model.

The remainder of this paper is structured as follows. In Section 2, we describe our system combination approach for the ML4HMT shared task and explain our substitution algorithm. Our experiments and results with the resulting combination system are presented in Section 3. Finally, we conclude and provide an outlook on future work in Section 4.

2 System Combination Approach

Our system combination approach is based on previous work on constituent substitution for system combination. One system is chosen as providing the translation template while the remaining systems provide alternative translation variants (on a segment level) which maybe substituted into the template according to a set of decision factors that are derived from syntactic features.

2.1 Finding the right translation template

The organisers of the ML4HMT shared task provided us a data set containing a development set of 1,025 sentences and a test set including 1,026 sentences. For each of these sentences, the source text, the corresponding reference translation, and the translation output as well as various annotations from five machine translation systems were available as source data. Depending on the MT system, the level of annotation details varied greatly and the overall annotation was very heterogeneous which, in our view, made it difficult to make equal use of all annotations/systems. This might be something that could be improved for future work on this data.

We chose the translations by the Lucy RBMT system (Alonso and Thurmair, 2003) as our translation backbone. There are two reasons for this:

1. As a rule-based system, Lucy creates structurally sound sentences. The drawbacks of missing vocabulary coverage and incorrect lexical choice can be made up by mining other translations for better translation variants.

2. Additionally, of all five systems included in the workshop data, only the Lucy system provides analysis trees of the source sentence. Other systems only include trees for the target side of the translation, with many of the systems providing no syntactic information at all.

As our substitution approach is based on identifying *interesting*[1] phrases in the source sentence which are then linked to target language translations via word alignment, we decided to use the translations from the Lucy system as our translation template.

2.2 Reconstructing Lucy parse trees

The organisers of the workshop provided a flattened representation of the Lucy parse trees. Using some heuristics derived from the development set, we designed an algorithm to approximate the original deep

[1]Where *interesting* means *suitable for substitution within our system combination experiments*, e.g., noun phrases.

tree structures. For example, the XML fragment shown in Figure 1 describes the Spanish phrase *la inflación europea*. The noun phrase consists of:

- the determiner (*la*),
- the noun (*inflación*), and
- an adjective phrase (*europea*).

Our heuristics include a mapping which children a node is allowed to have: a node of the category **NO**, e.g., can either be a normal noun (**NST**) or a pronoun (**PRN**). Other part-of-speech categories are not legal wrt. the training data available from the ML4HMT development set.

With those heuristics, we built an XML parser which traverses the flattened XML tree representations and generates corresponding, *approximated tree structures* with a deeper structure. Figure 2 shows the syntactic tree we create from the XML fragment depicted in Figure 1. This deep tree is only an approximation of the original tree and does not contain all information that would be contained within parse trees generated from the original Lucy RBMY system, but it is nevertheless suitable to be used in our approach as we only consider substituting single words inside the candidate phrases we find in the source text parse trees.

2.3 Substitution algorithm

Previously, we have presented a language-independent substitution approach to system combination. Although this work also used rule-based machine translations as backbone[2], we exclusively relied on SMT systems to obtain alternative translation fragments. In this workshop we have access to translation output from systems that follow a variety of paradigms, however.

Lucy is an example for a rule-based MT system.

Apertium is also rule-based, whereas

Metis follows a hybrid approach and translates using a bilingual dictionary and a monolingual target language corpus.

MaTrEx includes several translation modules, but for this workshop a standard phrase-based

[2]We also used Lucy RBMT translation output in this previous work, but worked on original parse trees, not approximated tree structures.

```
<metanet:token id="s1_t2_r1_d1_k4">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NP"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k5">
<metanet:annotation type="alo" value="la"/>
<metanet:annotation type="can" value="el"/>
<metanet:annotation type="cat" value="DETP"/>
<metanet:string>la</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k6">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NO"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k7">
<metanet:annotation type="alo" value="inflación"/>
<metanet:annotation type="can" value="inflación"/>
<metanet:annotation type="cat" value="NST"/>
<metanet:string>inflación</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k8">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AP"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k9">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="A"/>
<metanet:string>europea</metanet:string>
</metanet:token>
<metanet:token id="s1_t2_r1_d1_k10">
<metanet:annotation type="alo" value="europea"/>
<metanet:annotation type="can" value="europeo"/>
<metanet:annotation type="cat" value="AST"/>
<metanet:string>europea</metanet:string>
</metanet:token>
```

Figure 1: Flattened representation of a Lucy parse tree.

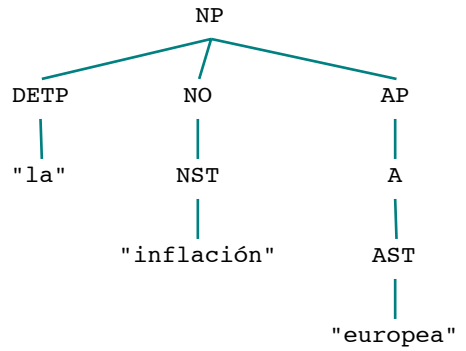

Figure 2: Approximated tree structure.

SMT model (Moses (Koehn et al., 2007)) was used.

Joshua provides output from a hierarchical phrase-based SMT model.

Using the approximated parse trees, we identify *interesting phrases* suitable for substitution: we consider noun, verb and adjective phrases. These are derived from the trees structures, while the potential substitution fragments from the other systems' output are linked using word alignment. Word alignment is computed using GIZA++ (Och and Ney, 2000). Each candidate translation by the four additional systems is evaluated according to the following features:

Matching POS? We only substitute if the part-of-speech of the candidate matches the reference, i.e., the translation template. This way we will not destroy the syntactic structure.

Majority vote Two or more systems may offer the same candidate translation. We prefer more frequent candidate fragments.

Context We take into account the part-of-speech of the surrounding tokens, left and right, to ensure that the fragment will fit into the context.

Language Model The candidate fragments as well as their -1 left and -1 right context are scored using a language model trained on EuroParl (Koehn, 2005).

3 Experiments

We tried out several combinations of features in our substitution system. In this section, we report on our experiments with the ML4HMT data set and provide results from comparing our system combination results to the baseline Lucy RBMT translation output. In our experiments, we translated from Spanish→English.

In our evaluation of the approach, we focus on the comparison to the Lucy baseline as our approach cannot be tuned with automated scoring metrics. Hence, it cannot be meaningfully compared to other systems in terms of BLEU scores.

3.1 Data sets

The WMT 2008 news test set of 2,051 sentences had been split into a development set of 1,025 sentences and a test set of 1,026 sentences. We used the development set data for the creation of the XML parser that approximates Lucy tree structures. We examined different combinations of features used in our substitution algorithm on the development data set.

3.2 Experimental results

In Table 1, we show the different feature configurations we tried. It is worth noting that each configuration performed better than the baseline, which was the Lucy RBMT system; this means that fragments from other systems actually did improve it. Table 2 presents results obtained from automated scoring metrics for the different system configurations applied on the development set data. Finally, in Table 3 we give the total number of substitutions that have been performed for each of the system configurations during our work on the development set.

Configuration	Matching POS?	Context
strict	yes	yes
pos	yes	no
context	no	yes
relaxed	no	no

Table 1: Feature configurations for experiments

Configuration	NIST	BLEU
baseline	5.0568	0.1516
strict	5.0937	0.1532
pos	5.0962	0.1534
context	**5.0984**	**0.1535**
relaxed	5.0932	0.1535

Table 2: Automated scoring results for development set.

The results shown in Table 2 indicate a possible improvement over the Lucy baseline. However, as the differences in BLEU between the configurations are not conclusive, we performed a manual evaluation of development set results. For example, the *context* feature disallows the substitution of *it is **saved*** by *it is **saves***. Removing this feature leads to

Configuration	# of substitutions
strict	412
pos	1,121
context	458
relaxed	1,317

Table 3: Substitution statistics for development set.

many more substitutions, which largely do not impact translation quality.

Based on our findings from the manual evaluation of development set results, we decided to use the *context* configuration in our final submission to the workshop. The context restriction includes part-of-speech matching implicitly, so adding this feature to the context restriction does not lead to any further improvements.

4 Conclusion

Whereas in previous work we only used translations generated by purely statistical MT systems as additional input, our system for the ML4HMT shared task could exploit output from systems of different paradigms. It remains to be investigated how this change affected resulting translation quality. The substitution approach showed improvements, although it was restricted to only single-word substitutions. In this hybrid setup we could retain the good syntactic structure of the RBMT output (which we used as translation template), while improving the lexical semantics by integrating translation fragments from other systems within the ML4HMT data set.

Future work includes expanding the substitution range to entire phrases and multi-word expressions. Restricting ourselves to single words has shown to help in retaining the good syntactic structure, but it also limits the impact of the additional systems on the baseline. By relaxing this restriction, we will open up our system to more extensive changes in the syntactic structure, which we will have to monitor closely to make sure we will not introduce translation candidates that will break the structure. Also, our features used for controlling the substitution algorithm are handcrafted at the moment; here we can see benefits from applying machine learning tools to actually learn helpful features from the given data. This will be an interesting extension of the system and would hopefully improve the substitution.

Acknowledgments

This work has been funded under the Seventh Framework Programme for Research and Technological Development of the European Commission through the T4ME contract (grant agreement no.: 249119). Part of this work was supported by the EuroMatrixPlus project (IST-231720) which is also funded by the European Community. We are grateful to two anonymous reviewers from LIHMT 2011 for their comments.

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Juan A. Alonso and Gregor Thurmair. 2003. The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA.

Chris Callison-Burch and Raymond Flournoy. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July. Association for Computational Linguistics.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of WMT07*, pages 193–196, Prague, Czech Republic, June. Association for Computational Linguistics.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining multi-engine translations with moses. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation, pages 42–46. Association for Computational Linguistics, 3.

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, June. Association for Computational Linguistics.

Cristina Espana-Bonet, Gorka Labaka, Arantza Diaz de Ilarraza, Llus Màrquez, and Kepa Sarasola. 2011. Hybrid machine translation guided by a rule-based system. In *Proceedings of the 13th Machine Translation Summit*, pages 554–561, September.

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 Translation*, pages 70–74. Association for Computational Linguistics, 3.

Robert E. Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *ANLP*, pages 95–100.

Nizar Habash, Bonnie J. Dorr, and Christof Monz. 2009. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation*, 23(1):23–63.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107, Honolulu, Hawaii, October. Association for Computational Linguistics.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, Budapest, Hungary.

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2009. The RWTH system combination system for WMT 2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 51–55, Athens, Greece, March. Association for Computational Linguistics.

Wolfgang Macherey and Franz J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 986–995, Prague, Czech Republic, June. Association for Computational Linguistics.

Evgeny Matusov, Nicola Ueffing, and Hermann 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*, pages 33–40, Trento, Italy, April.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

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.

Antti-Veikko I. Rosti, Spyridon Matsoukas, and Richard M. Schwartz. 2007. Improved word-level system combination for machine translation. In *ACL*.

Gregor Thurmair. 2009. Comparing different architectures of hybrid Machine Translation systems. In *MT Summit XII 2009*, August.