Translation Combination using Factored Word Substitution

Christian Federmann[1], **Silke Theison**[2], **Andreas Eisele**[1,2], **Hans Uszkoreit**[1,2], **Yu Chen**[2], **Michael Jellinghaus**[2], **Sabine Hunsicker**[2]
1: Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Saarbrücken, Germany
2: Universität des Saarlandes, Saarbrücken, Germany
{cfedermann,eisele,uszkoreit}@dfki.de, {sith,yuchen,micha,sabineh}@coli.uni-sb.de

Abstract

We present a word substitution approach to combine the output of different machine translation systems. Using part of speech information, candidate words are determined among possible translation options, which in turn are estimated through a precomputed word alignment. Automatic substitution is guided by several decision factors, including part of speech, local context, and language model probabilities. The combination of these factors is defined after careful manual analysis of their respective impact. The approach is tested for the language pair German-English, however the general technique itself is language independent.

1 Introduction

Despite remarkable progress in machine translation (MT) in the last decade, automatic translation is still far away from satisfactory quality. Even the most advanced MT technology as summarized by (Lopez, 2008), including the best statistical, rule-based and example-based systems, produces output rife with errors. Those systems may employ different algorithms or vary in the linguistic resources they use which in turn leads to different characteristic errors.

Besides continued research on improving MT techniques, one line of research is dedicated to better exploitation of existing methods for the combination of their respective advantages (Macherey and Och, 2007; Rosti et al., 2007a).

Current approaches for system combination involve post-editing methods (Dugast et al., 2007; Theison, 2007), re-ranking strategies, or shallow phrase substitution. The combination procedure applied for this pape tries to optimize word-level translations within a "trusted" sentence frame selected due to the high quality of its syntactic structure. The underlying idea of the approach is the improvement of a given (original) translation through the exploitation of additional translations of the same text. This can be seen as a simplified version of (Rosti et al., 2007b).

Considering our submission from the shared translation task as the "trusted" frame, we add translations from four additional MT systems that have been chosen based on their performance in terms of automatic evaluation metrics. In total, the combination system performs 1,691 substitutions, i.e., an average of 0.67 substitutions per sentence.

2 Architecture

Our system combination approach computes a combined translation from a given set of machine translations. Below, we present a short overview by describing the different steps in the derivation of a combined translation.

Compute POS tags for translations. We apply part-of-speech (POS) tagging to prepare the selection of possible substitution candidates. For the determination of POS tags we use the Stuttgart TreeTagger (Schmid, 1994).

Create word alignment. The alignment between source text and translations is needed to identify translation options within the different systems' translations. Word alignment is computed using the GIZA++ toolkit (Och and Ney, 2003), only one-to-one word alignments are employed.

Select substitution candidates. For the shared task, we decide to substitute nouns, verbs and adjectives based on the available POS tags. Initially, any such source word is considered as a possible substitution candidate. As we do not want to require substitution can-

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 70–74,
Athens, Greece, 30 March – 31 March 2009.

didates to have exactly the same POS tag as the source, we use groups of "similar" tags.

Compute decision factors for candidates. We define several decision factors to enable an automatic ranking of translation options. Details on these can be found in section 4.

Evaluate the decision factors and substitute. Using the available decision factors we compute the best translation and substitute.

The general combination approach is language independent as it only requires a (statistical) POS tagger and GIZA++ to compute the word alignments. More advanced linguistic resources are not required. The addition of lexical resources to improve the extracted word alignments has been considered, however the idea was then dropped as we did not expect any short-term improvements.

3 System selection

Our system combination engine takes any given number of translations and enables us to compute a combined translation out of these. One of the given system translations is chosen to provide the "sentence skeleton", i.e. the global structure of the translation, thus representing the *reference system*. All other systems can only contribute single words for substitution to the combined translation, hence serve as *substitution sources*.

3.1 Reference system

Following our research on hybrid translation trying to combine the strengths of rule-based MT with the virtues of statistical MT, we choose our own (usaar) submission from the shared task to provide the sentence frame for our combination system. As this translation is based upon a rule-based MT system, we expect the overall sentence structure to be of a sufficiently high quality.

3.2 Substitution sources

For the implementation of our combination system, we need resources of potential substitution candidates. As sources for possible substitution, we thus include the translation results of the following four systems:

- Google (google)[1]
- University of Karlsruhe (uka)
- University of Maryland (umd)
- University of Stuttgart (stuttgart)

The decision to select the output of these particular MT systems is based on their performance in terms of different automatic evaluation metrics obtained with the IQMT Framework by (Giménez and Amigó, 2006). This includes BLEU, BLEU1, TER, NIST, METEOR, RG, MT06, and WMT08. The results, listing only the three best systems per metric, are given in table 1.

metric	best three systems		
BLEU1	google	uka	systran
	0.599	0.593	0.582
BLEU	google	uka	umd
	0.232	0.231	0.223
TER	umd	rwth.c3	uka
	0.350	0.335	0.332
NIST	google	umd	uka
	6.353	6.302	6.270
METEOR	google	uka	stuttgart
	0.558	0.555	0.548
RG	umd	uka	google
	0.527	0.525	0.520
MT06	umd	google	stuttgart
	0.415	0.413	0.410
WMT08	stuttgart	rbmt3	google
	0.344	0.341	0.336

Table 1: Automatic evaluation results.

On grounds of these results we anticipate the four above named translation engines to perform best when being combined with our hybrid machine translation system. We restrict the substitution sources to the four potentially best systems in order to omit bad substitutions and to reduce the computational complexity of the substitution problem. It is possible to choose any other number of substitution sources.

4 Substitution

As mentioned above, we consider nouns, verbs and adjectives as possible substitution candidates. In order to allow for automatic decision making amongst several translation options we define a set of factors, detailed in the following. Furthermore, we present some examples in order to illustrate the use of the factors within the decision process.

[1]The Google submission was translated by the Google MT production system offered within the Google Language Tools as opposed to the qualitatively superior Google MT research system.

4.1 Decision factors

The set of factors underlying the decision procedure consists of the following:

A: Matching POS. This Boolean factor checks whether the target word POS tag matches the source word's POS category. The factor compares the source text to the reference translation as we want to preserve the sentential structure of the latter.

B: Majority vote. For this factor, we compute an ordered list of the different translation options, sorted by decreasing frequency. A consensus between several systems may help to identify the best translation.

Both the reference system and the Google submission receive a $+1$ bonus, as they appeared to offer better candidates in more cases within the small data sample of our manual analysis.

C: POS context. Further filtering is applied determining the words' POS context. This is especially important as we do not want to degrade the sentence structure maintained by the translation output of the reference system.

In order to optimize this factor, we conduct trials with the single word, the -1 left, and the $+1$ right context. To reduce complexity, we shorten POS tags to a single character, e.g. $NN \rightarrow N$ or $NPS \rightarrow N$.

D: Language Model. We use an English language model to score the different translation options. As the combination system only replaces single words within a bi-gram context, we employ the bi-gram portion of the English Gigaword language model.

The language model had been estimated using the SRILM toolkit (Stolcke, 2002).

4.2 Factor configurations

To determine the best possible combination of our different factors, we define four potential factor configurations and evaluate them manually on a small set of sentences. The configurations differ in the consideration of the *POS context* for factor C (*strict* including -1 left context versus *relaxed* including no context) and in the usage of factor A *Matching POS* (+*A*). Table 2 shows the settings of factors A and C for the different configurations.

configuration	Matching POS	POS context
strict	disabled	-1 left
strict+A	enabled	-1 left
relaxed	disabled	single word
relaxed+A	enabled	single word

Table 2: Factor configurations for combination.

Our manual evaluation of the respective substitution decisions taken by different factor combination is suggestive of the "relaxed+A" configuration to produce the best combination result. Thus, this configuration is utilized to produce sound combined translations for the complete data set.

4.3 Factored substitution

Having determined the configuration of the different factors, we compute those for the complete data set, in order to apply the final substitution step which will create the combined translation.

The factored substitution algorithm chooses among the different translation options in the following way:

(a) Matching POS? If factor A is activated for the current factor configuration (+*A*), substitution of the given translation options can only be possible if the factor evaluates to True. Otherwise the substitution candidate is skipped.

(b) Majority vote winner? If the majority vote yields a unique winner, this translation option is taken as the final translation.

Using the $+1$ bonuses for both the reference system and the Google submission we introduce a slight bias that was motivated by manual evaluation of the different systems' translation results.

(c) Language model. If several majority vote winners can be determined, the one with the best language model score is chosen.

Due to the nature of real numbers this step always chooses a winning translation option and thus the termination of the substitution algorithm is well-defined.

Please note that, while factors A, B, and D are explicitly used within the substitution algorithm, factor C *POS context* is implicitly used only when computing the possible translation options for a given substitution candidate.

configuration	substitutions	ratio
strict	1,690	5.714%
strict+A	1,347	4.554%
relaxed	2,228	7.532%
relaxed+A	1,691	5.717%

Table 3: Substitutions for 29,579 candidates.

Interestingly we are able to obtain best results without considering the -1 left POS context, i.e. only checking the POS tag of the single word translation option for factor C.

4.4 Combination results

We compute system combinations for each of the four factor configurations defined above. Table 3 displays how many substitutions are conducted within each of these configurations.

The following examples illustrate the performance of the substitution algorithm used to produce the combined translations.

"Einbruch": the reference translation for "Einbruch" is "collapse", the substitution sources propose "slump" and "drop", but also "collapse", all three, considering the context, forming good translations. The majority vote rules out the suggestions different to the reference translation due to the fact that 2 more systems recommend "collapse" as the correct translation.

"Rückgang": the reference system translates this word as "drop" while all of the substitution sources choose "decline" as the correct translation. Since factor A evaluates to True, i.e. the POS tags are of the same nature, "decline" is clearly selected as the best translation by factor B *Majority vote* and thus replaces "drop" in the final combined translation result.

"Tagesgeschäfte": our reference system translates "Tagesgeschäfte" with "requirements", while two of the substitution systems indicate "business" to be a better translation. Due to the $+1$ bonus for our reference translation a tie between the two possible translations emerges, leaving the decision to the language model score, which is higher for "business".

4.5 Evaluation results

Table 4 shows the results of the manual evaluation campaign carried out as part of the WMT09 shared task. Randomly chosen sentences are presented to the annotator, who then has to put them into relative order. Note that each annotator is shown a random subset of the sentences to be evaluated.

system	relative rank	data points
google	-2.74	174
uka	-3.00	217
umd	-3.03	170
stuttgart	-2.89	163
usaar	-2.78	186
usaar-combo	-2.91	164

Table 4: Relative ranking results from the WMT09 manual evalution campaign.

Interestingly, our combined system is not able to outperform the baseline, i.e., additional data did not improve translation results. However the evaluation is rather intransparent since it does not allow for a strict comparison between sentences.

5 Conclusion

Within the system described in this paper, we approach a hybrid translation technique combining the output of different MT systems. Substituting particular words within a well-structured translation frame equips us with considerably enhanced translation output. We obtain promising results providing substantiated proof that our approach is going in the right direction.

Further steps in the future will include machine learning methods to optimize the factor selection. This was, due to limited amount of time and data, not feasible thus far. We will also investigate the potential of phrase-based substitution taking into account multi-word alignments instead of just single word mappings. Additionally, we would like to continue work on the integration of lexical resources to post-correct the word alignments obtained by GIZA++ as this will directly improve the overall system performance.

Acknowledgments

This work was supported by the EuroMatrix project (IST-034291) which is funded by the European Community under the Sixth Framework Programme for Research and Technological Development.

References

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Jesús Giménez and Enrique Amigó. 2006. IQMT: A framework for automatic machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

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.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007a. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York, April. Association for Computational Linguistics.

Antti-Veikko Rosti, Spyros Matsoukas, and Richard Schwartz. 2007b. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319, Prague, Czech Republic, June. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, September.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, Denver, Colorado.

Silke Theison. 2007. Optimizing rule-based machine translation output with the help of statistical methods. Master's thesis, Saarland University, Computational Linguistics department.