# Statistical Post Editing and Dictionary Extraction: Systran/Edinburgh submissions for ACL-WMT2009

**Loic Dugast**[*,**] and **Jean Senellart**[*]
[*]SYSTRAN S.A.
La Grande Arche
1, Parvis de la Défense
92044 Paris
La Défense Cedex
France

**Philipp Koehn**[**]
[**]School of Informatics
University of Edinburgh
10 Crichton Street,
Edinburgh
United Kingdom

## Abstract

We describe here the two Systran/University of Edinburgh submissions for WMT2009. They involve a statistical post-editing model with a particular handling of named entities (English to French and German to English) and the extraction of phrasal rules (English to French).

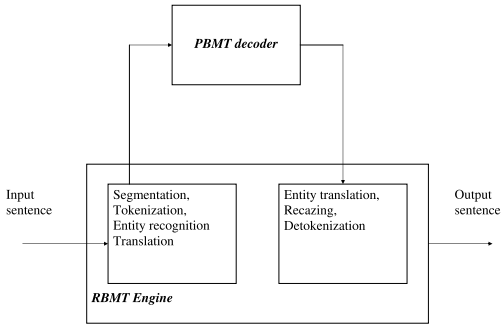


Figure 1: Translation with PBMT post-editing

## 1 Introduction

Previous results had shown a rather satisfying performance for hybrid systems such as the Statistical Phrase-based Post-Editing (SPE) (Simard et al., 2007) combination in comparison with purely phrase-based statistical models, reaching similar BLEU scores and often receiving better human judgement (German to English at WMT2007) against the BLEU metric. This last result was in accordance with the previous acknowledgment (Callison-Burch et al., 2006) that systems of too differing structure could not be compared reliably with BLEU. We participated in the recent Workshop on Machine Translation (WMT'09) in the language pairs English to French and German to English. On the one hand we trained a Post-Editing system with an additional special treatment to avoid the loss of entities such as dates and numbers. On the other hand we trained an additional English-to-French system (as a secondary submission) that made use of automatically extracted linguistic entries. In this paper, we will present both approaches. The latter is part of ongoing work motivated by the desire to both make use of corpus statistics and keep the advantage of the often (relative to automatic metrics's scores) higher rank in human judgement given to rule-based systems on out-of-domain data, as seen on the WMT 2008 results for both English to French and German to English (Callison-Burch et al., 2008).

## 2 Statistical Post Editing systems

### 2.1 Baseline

The basic setup is identical to the one described in (Dugast et al., 2007). A statistical translation model is trained between the rule-based translation of the source-side 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. Figure 1 shows the translation process.

Here are a few additional details which tend to improve training and limit unwanted statistical effects in translation:

- Named entities are replaced by special tokens on both sides. By reducing vocabulary and

*Proceedings of the 4th EACL Workshop on Statistical Machine Translation* , pages 110–114,
Athens, Greece, 30 March – 31 March 2009. 

combined with the next item mentioned, this should help word alignment. Moreover, 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 source). This was added to the parallel text to improve the word alignment.
- Rule-based output and reference translations are lowercased before performing alignment, leaving the recasing job up to the rule-based engine.
- Singleton phrase pairs are deleted from the phrase table to avoid overfitting.
- Phrase pairs non cohesive regarding entities are also discarded. We make the hypothesis that entities are always passed to the target language and all entities in the target language originate from the source language. This point is discussed in section 2.2.

We will discuss some of these details further in the upcoming sections.

Due to time constraints, we did not use the Giga French-English Parallel corpus provided for the workshop. We only made use of the News Commentary and the Europarl corpora. We used additional in-domain news corpora to train 5 grams language models, according to the baseline recommendations. Weights for these separate models were tuned through the Mert algorithm provided in the Moses toolkit (Koehn et al., 2007), using the provided news tuning set.

### 2.2 Trimming

In a statistical translation model, trimming of the phrase table had been shown to be beneficial (Johnson et al., 2007). For our post-editing model, we can afford to perform an even more aggressive trimming of the phrase table, since the rule-based system already provides us with a translation and we only aim at correcting the most frequent errors. Therefore, we suppress all unique phrase pairs before calculating the probabilities for the final phrase table.

| **Rule-Based French** | **Reference French** |
|---|---|
| __ent_date | et |
| __ent_date | __ent_numeric et |
| __ent_numeric de golfe . | du golfe __ent_date . |
| décennie | __ent_numeric ans |
| et __ent_numeric . | . |

Table 1: Examples of problematic phrase pairs

### 2.3 Avoiding the loss of entities

Deleted and spurious content is a well known problem for statistical models (Chiang et al., 2008). Though we do not know of any study proving it, it seems obvious that Named Entities that would be either deleted or added to the output out of nowhere is an especially problematic kind of error for the translation quality. The rule-based translation engine benefits from an entity recognition layer for numbers, dates and hours, addresses, company names and URIs. We therefore "trim" (delete) from the extracted phrase pairs any item that would not translate all entities from the source (i.e. the RBMT output) to the target or add spurious entities which were not present in the source side of the phrase pair. Table 1 illustrates the kind of phrase pairs that are excluded from the model. For example, the first phrase pair, when applied, would simply erase the date entity which was expressed in the source sentence, which we of course do not want.

## 3 Rule Extraction

The baseline Systran rule-based system is more or less a linguistic-oriented system that makes use of a dependency analysis, general transfer rules and dictionary entries, and finally a synthesis/reordering stage. The dictionary entries have long been the main entry point for customization of the system. Such lexical translation rules are fully linguistically coded dictionary entries, with the following features attached: part-of-speech, inflection category, headword and possibly some semantic tags. Table 2 displays a sample of manually-entered entries. These entries may both match any inflected form of the source and generate the appropriate (according to general agreement rules and depending on the source analysis) target inflection.

Motivations for adding phrasal dictionary entries

| POS | English | French | headword_English | headword_French |
|---|---|---|---|---|
| Noun | college level | niveau d'études universitaires | level | niveau |
| Adverb | on bail | sous caution | on | sous |
| Verb | badmouth | médire de | badmouth | médire |

Table 2: Example dictionary entries

(compound words) are twofold: first, just as for statistical translation models which went from word-based to phrase-based models, it helps solve disambiguation and non-literal translations. Second, as the rule-based engine makes use of a syntactic analysis of a source sentence, adding unambiguous phrasal chunks as entries will reduce the overall syntactic ambiguity and lead to a better source analysis.

### 3.1 Manual customization through dictionary entries

The Systran system provides a dictionary coding tool (Senellart et al., 2003). This tool allows the manual task of coding entries to be partially automated with the use of monolingual dictionaries and probabilistic context-free grammars, while allowing the user to fine-tune it by correcting the automatic coding and/or add more features. However, this remains first of all a time-consuming task. Moreover, it is not easy for humans to select the best translation among a set of alternatives, let alone assign them probabilities. Last but not least, the beneficial effect on translation is not guaranteed (especially, the effect on the rule-based dependency analysis).

### 3.2 Automatic extraction of dictionary entries

The problem consists of selecting relevant phrase pairs from a set, coding them linguistically and assign them probabilities. The extraction setup as depicted in figure 2) starts from a parallel corpus dataset. The baseline procedure is followed (word alignment using GIZA++ and use of common heuristics to extract phrase pairs (Koehn et al., 2007)) to extract phrase pairs. At this stage the "phrases" are plain word sequences, not necessarily linguistically motivated. Some statistical information is attached to each phrase pair: frequency of the pair and lexical weights in both directions. Each unique phrase pair is then processed by our dictionary coding tool which tries to map both word

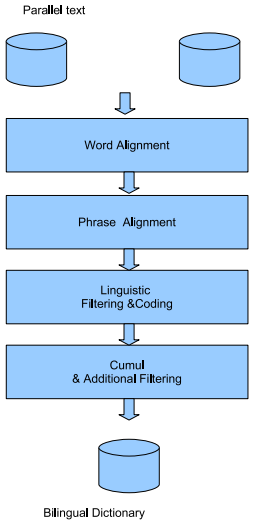


Figure 2: Extraction pipeline: from parallel texts to bilingual dictionary

sequences to a given category. If both sides are mapped to the same category, the phrase pair, now lemmatized, is retained as a bilingual entry. Otherwise, the candidate is excluded. Given that a bilingual entry with a same lemma may have various inflectional forms in corpus, we then sum the lemma counts. Finally, in the current setup, we only keep the most frequent translation for each source.

For our secondary submission for English-French, we extracted such entries from both the News Commentary and the Europarl corpus.

### 3.3 Validation of dictionary entries

The coding procedure, when applied to phrase pairs extracted from the corpus instead of manually entered entries, may generate rules that do not lead to an improved translation. Recall that we start from an existing system and only want to learn additional rules to adapt to the domain of the bilingual corpus we have at our disposal.

Now the problem consists of building the optimal

subset from the set of candidate entries, according to a translation evaluation metric (here, BLEU). Unlike the Mert procedure, we would like to do more than assign global weights for the whole set of translation rules, but instead make a decision for each individual phrasal rule.

As an approximate response to this problem, we test each extracted entry individually, starting from the lower n-grams to the longer (source) chunks, following algorithm 1. This results in dictionaries of 5k and 170k entries for the News Commentary and the Europarl parallel corpora, respectively.

**Algorithm 1** Dictionary Validation Algorithm

```
n=1
for n=1 to Nmax do
    map all n-gram entries to parallel sentences
    translate training corpus with current dictionary
    for each entry do
        translate all relevant sentences with current dictionary, plus this entry
        compute BLEU scores without and with the entry
    end for
    Select entries with better/worse sentences ratio above threshold
    add these entries to current dictionary
end for
```

## 4 Results

BLEU scores of the dictionary extraction experiments for the English-French language pair and three types of corpora are displayed in table 4. Table 3 shows results on the news test set. Post-editing setups were tuned on the news tuning set.

## 5 Conclusion and future work

We presented a few improvements to the Statistical Post Editing setup. They are part of an effort to better integrate a linguistic, rule-based system and the statistical correcting layer also illustrated in (Ueffing et al., 2008). Moreover, we presented a dictionary extraction setup which resulted in an improvement of 2 to 3 BLEU points over the baseline rule-based system when in-domain,as can be seen in table 4. This however improved translation very little on the "news" domain which was used for evaluation. We think that is a different issue, namely of domain adaptation. In order to push further this rule-extraction approach and according to our previous work (Dugast et al., 2007) (Dugast et al., 2008), the most promising would probably be the use of alternative meanings and a language model to decode the best translation in such a lattice. Another path for improvement would be to try and extract rules with more features, such as constraints of lexical subcategorization as they already exist in the manually entered entries. Finally, we would like to try combining the dictionary extraction setup with a Statistical Post-Editing layer to see if the latter supersedes the former.

## Acknowledgement

We would like to thank the anonymous reviewers for their comments and corrections.

## References

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *In proceedings of EACL 2006*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

David Chiang, Steve Deneefe, Yee S. Chan, and Hwee T. Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619, Honolulu, Hawaii, October. Association for Computational Linguistics.

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.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2008. Can we relearn an rbmt system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, U.S.A., June. Association for Computational Linguistics.

| System | BLEU |
|---|---|
| RBMT English-French | 20.48 |
| RBMT+SPE English-French | 21.90 |
| RBMT+Extracted dictionary English-French | 20.82 |
| RBMT German-English | 15.13 |
| RBMT+SPE German-English | 17.50 |

Table 3: Compared results of original RBMT system,post-editing and dictionary extraction: real-cased, untokenized NIST Bleu scores on the full newstest2009 set(%)

| System | **nc-test2007** (news commentary) | **test2007** (europarl) | **newstest2009** (news) |
|---|---|---|---|
| RBMT | 24.88 | 22.75 | 20.48 |
| RBMT +Dictionary extracted from News Commentary | 26.54 | - | 20.57 |
| RBMT +Dictionary extracted from Europarl | - | 25.55 | - |
| RBMT +Dictionary extracted from NC and Europarl, priority on NC | 26.65 | - | 20.82 |

Table 4: Results of dictionary extraction for English-French: real-cased, untokenized NIST Bleu scores (%)

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. 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 *ACL 2007, demonstration session*.

Jean Senellart, Jin Yang, and Anabel Rebollo. 2003. Technologie systran intuitive coding. In *in Proceedings of MT Summit IX*.

M. Simard, C. Goutte, and P. Isabelle. 2007. Statistical phrase-based post-editing. In *proceedings of the NAACL-HLT. 2007. NRC 49288.*

Nicola Ueffing, Jens Stephan, Evgeny Matusov, Loïc Dugast, George Foster, Roland Kuhn, Jean Senellart, and Jin Yang. 2008. Tighter integration of rule-based and statistical MT in serial system combination. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 913–920, Manchester, UK, August. Coling 2008 Organizing Committee.