LIMSI’s Contribution to the WMT’16 Biomedical Translation Task
Julia Ive, François Yvon, Aurélien Max

To cite this version:
Julia Ive, François Yvon, Aurélien Max. LIMSI’s Contribution to the WMT’16 Biomedical Translation Task. First Conference on Machine Translation, Aug 2016, Berlin, Germany. pp.469-476, 10.18653/v1/W16-2337. hal-01388658

HAL Id: hal-01388658
https://hal.archives-ouvertes.fr/hal-01388658
Submitted on 3 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
LIMSI's Contribution to the WMT'16 Biomedical Translation Task

Julia Ive1,2, Aurélien Max1, François Yvon1
LIMSI, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France,1
Cochrane France, INSERM U1153, 75181 Paris, France2
{firstname.lastname}@limsi.fr

Abstract
The article describes LIMSI’s submission to the first WMT’16 shared biomedical translation task, focusing on the sole English-French translation direction. Our main submission is the output of a MOSES-based statistical machine translation (SMT) system, rescored with Structured OUtput Layer (SOUL) neural network models. We also present an attempt to circumvent syntactic complexity: our proposal combines the outputs of PBSMT systems trained either to translate entire source sentences or specific syntactic constructs extracted from those sentences. The approach is implemented using Confusion Network (CN) decoding. The quality of the combined output is comparable to the quality of our main system.

1 Introduction
The paper provides the details of LIMSI’s submission to the first shared biomedical translation task at WMT’16. For our main submission we built a phrase-based statistical machine translation (SMT) system using MOSES and attempted to improve the quality of its output by rescoring its $n$-best list with Structured OUtput Layer (SOUL) neural network models.

Our secondary submission was designed to mitigate the negative effects of syntactic complexity of sentences. This complexity creates a challenge for the phrase-based SMT (PBSMT) paradigm that only sees a sentence as a sequential structure. To overcome this problem, the output of PBSMT systems can be combined with the output of "syntax-aware" MT systems (rule-based, syntax-based, etc. (Freitag et al., 2014b; Avramidis et al., 2015; Li et al., 2015)).

As the building of the latter type of systems can be costly, we propose a light-weight alternative that combines the outputs of several PBSMT systems trained for the translation of (a) entire sentences, and (b) separate continuous and discontinuous syntactic constructions extracted from those sentences. The combination is performed using confusion network (CN) decoding. The quantitative difference with the baseline is rather small, but our comparative analysis of this system allows us to better understand its potential and limitations.

2 Systems Overview
In all our experiments we used the MOSES implementation of the phrase-based approach to SMT (Koehn et al., 2007).

2.1 Additional Parallel Data
The translation of scientific abstracts in the biomedical domain is a task that is characterized by the availability of high-quality in-domain corpora. In all our experiments, we used the English-French Cochrane corpus of medical review abstracts, which resembles the shared task data (Ive et al., 2016).1 This corpus was split in two parts: titles (COCHRANE-TITLES) and abstracts (COCHRANE-ABS). The same split was performed for the SCIELO corpus (SCIELO-TITLES and SCIELO-ABS, respectively). We will further refer to the union of all the provided task data and of the COCHRANE data as the IN-DOMAIN-DATA. Additionally, we used the data distributed for the WMT’14 medical task,2 even though its relatedness to the SCIELO test data is lesser.

---

1http://www.translatecochrane.fr/corpus
2http://statmt.org/wmt14/medical-task
2.2 Additional Monolingual Data

As additional monolingual data we used the full French dataset provided by the organizers of the WMT’15 translation task.\(^3\)

2.3 Preprocessing and Word Alignment

Tokenization and detokenization for both source (English) and target (French) texts were performed by our in-house text processing tools (Déchelotte et al., 2008). Additionally, the MEDLINE-TITLES corpus provided with the shared task was cleaned as follows: we excluded source sentences with generic comments instead of translations (e.g., ”In Process Citation”). This reduced the count of the original corpus sentences by 3%. Details on the WMT’14 and WMT’15 data preprocessing schemes can be found in (Pécheux et al., 2014; Marie et al., 2015). The statistics regarding the preprocessed data are in Table 1. Word alignments were computed using fast_align (Dyer et al., 2013).

2.4 Language Models

We built an in-domain 6-gram language model (LM) (In-domain-LM1) combined with a 4-gram LM developed in the context of WMT’14 (In-domain-LM2); both are trained using the corresponding monolingual parts of the parallel data with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1996), using the SRILM (Stolcke, 2002) and KENLM (Heafield, 2011) toolkits. We also used an out-of-domain 4-gram LM (Out-of-domain-LM), described in (Marie et al., 2015).

2.5 SOUL

We also made use of Structured OUtput Layer (SOUL) neural network Language and Translation models (Le et al., 2011; Le et al., 2012a) as these have been shown to systematically improve our systems in recent evaluations (Le et al., 2012b; Allavuzen et al., 2013; Pécheux et al., 2014; Marie et al., 2015). The SOUL architecture can estimate LMs of higher \(n\)-gram order (e.g., \(n = 10\) instead of \(n = 4\)) for large output vocabulary; SOUL is used to rescore n-best lists of the MOSES system.

2.6 Development and Test Sets

In the absence of official development data, we chose our development (LIMSI_DEV) and internal test (LIMSI_TEST) data randomly out of the provided SCIELO-ABS and SCIELO-TITLES corpora. Each set contains 14% of the total count of SCIELO-TITLES sentences and 11% of the total count of SCIELO-ABS sentences.

Given the quantity of misspelled words in the data (e.g. “exernai” for “external”, “level” instead of “level”, etc.), we tried to select datasets with an OOV rate not higher than the rate of the rest of the SCIELO corpus, as compared to the vocabulary of the IN-DOMAIN-DATA (SCIELO data excluded) and WMT’14 medical data (e.g., for SCIELO-ABS the OOV rate ≈ 2%).

LIMSI_DEV and LIMSI_TEST were used to respectively tune and test our main PBSMT systems. LIMSI_TEST was further split into LIMSI_DEV2 and LIMSI_TEST2 for SOUL and system combination optimizations. Statistics for these datasets are in Table 1.

2.7 Evaluation Metrics

BLEU scores (Papineni et al., 2002) are computed using the cased multi-bleu.perl script and our own tokenizer for reference translations.

3 Baseline System

3.1 Details of System Building

For our baseline system, all the available IN-DOMAIN-DATA were used to train the translation models consisting of the phrase table (PT) and the lexicalized reordering models (msd-bidirectional-fe). We used the WMT’14 medical task parallel data to train additional models. More specifically, these models were used as back-off models to search for \(n\)-grams (up to \(n = 4\)) with no translation in the main models. The three LMs described in Section 2.4 were used. This system was tuned with kb-mira (Cherry and Foster, 2012) using 300-best lists.

3.2 Experiments and Results

The results of our baseline system are in Table 2. For our experiments with neural network models we took the 10-gram SOUL models trained for the LIMSI participation to WMT’12 (Le et al., 2012b). SOUL models define five additional features: a monolingual target LM score and four

---

\(^3\)http://www.statmt.org/wmt15/translation-task.html
translation model scores (Le et al., 2012a). The baseline 300-best list was reranked according to the combination of all baseline features and the SOUL features. Reranking allowed us to obtain an improvement of +1.17 BLEU over our baseline system. The system tends to perform better on the LIMSI-Test2-Titles part than on the LIMSI-Test2-Abs part. In the rest of this article, we focus our efforts on improving the translation quality of abstracts only.

4 Using Phrase-Based Statistical Machine Translation to Circumvent Syntactic Complexity

Scientific medical texts are characterized by a large quantity of compound terms and complex sentences. Their translation can be especially challenging for PBSMT due to its intrinsic limitations which include, among others, the generation of translations by mere concatenation and the inability to resolve long-distance relations between sentence components. These limitations can be overcome in PBSMT by combining with the outputs of “syntax-aware” MT systems (rule-base MT (RBMT), syntax-based MT (SBMT)) (Costa-Jussà et al., 2012; Avramidis et al., 2015; Li et al., 2015). The combination of system outputs is often performed with the help of Confusion Network (CN) decoding as an effective means to recombine translation alternatives at the word level (Deng et al., 2015; Freitag et al., 2014a; Freitag et al., 2014b; Zhu et al., 2013).

Less costly solutions seek to better explore the potential of phrase-based architectures. For instance, Hewavitharana et al. (2007) propose to improve the PBSMT outputs by separately translating noun phrases (NPs) extracted from source sentences.

Inspired by this study, we propose to combine the baseline hypotheses with partial, local hypotheses by means of CN decoding. To obtain those partial hypotheses, we trained separate PBSMT systems to translate on the one hand the NPs (NP-SMT), often representing complex terms, and on the other hand, simplified variants of the source sentences where NPs are replaced by their syntactic head (NP-Reduced-SMT) (see Figure 1).

4.1 Methodology

A CN is a weighted directed acyclic graph where all the paths go through all the nodes (Mangu et al., 2000). There may be one or more arcs between two consecutive nodes. Arcs can here be considered as alternative translation choices for target words (including the empty NULL word).

Building a confusion network implies several decisions:

1. Choice of the main hypothesis (backbone) to guide the word order: This choice is crucial for the final translation quality (see e.g. (Hildebrand and Vogel, 2008)). In our case, we chose the 1-best baseline hypothesis as the backbone.

2. Choice of the word alignment strategy between the hypotheses: Alternative hypotheses are usually aligned to the backbone without taking their alignments with source tokens into account (Rosti et al., 2012; Rosti et al., 2008; Matusov et al., 2006). Following Du et al. (2009), we instead aligned hypotheses according to the source-target alignments produced by the decoder.

Figure 2 illustrates the hypothesis alignment

<table>
<thead>
<tr>
<th>Corpus</th>
<th># Lines</th>
<th># Tok. en</th>
<th># Tok. fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCIELO-ABS</td>
<td>≈8 K</td>
<td>≈200 K</td>
<td>≈280 K</td>
</tr>
<tr>
<td>SCIELO-TITLES</td>
<td>≈700</td>
<td>≈10 K</td>
<td>≈14 K</td>
</tr>
<tr>
<td>MEDLINE-TITLES</td>
<td>≈600 K</td>
<td>≈7 M</td>
<td>≈8 M</td>
</tr>
<tr>
<td>COCHRANE-ABS</td>
<td>≈140 K</td>
<td>≈3 M</td>
<td>≈5 M</td>
</tr>
<tr>
<td>COCHRANE-TITLES</td>
<td>≈8 K</td>
<td>≈90 K</td>
<td>≈130 K</td>
</tr>
<tr>
<td>IN-DOMAIN-DATA</td>
<td>≈800 K</td>
<td>≈10 M</td>
<td>≈13 M</td>
</tr>
<tr>
<td>COPPA</td>
<td>≈454 K</td>
<td>≈10 M</td>
<td>≈12 M</td>
</tr>
<tr>
<td>EMEA</td>
<td>≈324 K</td>
<td>≈6 M</td>
<td>≈7 M</td>
</tr>
<tr>
<td>PATTR-ABS</td>
<td>≈635 K</td>
<td>≈20 M</td>
<td>≈24 M</td>
</tr>
<tr>
<td>PATTR-CLAIMS</td>
<td>≈889 K</td>
<td>≈32 M</td>
<td>≈36 M</td>
</tr>
<tr>
<td>PATTR-TITLES</td>
<td>≈386 K</td>
<td>≈3 M</td>
<td>≈4 M</td>
</tr>
<tr>
<td>UMLS</td>
<td>≈2 M</td>
<td>≈8 M</td>
<td>≈8 M</td>
</tr>
<tr>
<td>WIKIPEDIA</td>
<td>≈8 K</td>
<td>≈17 K</td>
<td>≈19 K</td>
</tr>
<tr>
<td>WMT’14 medical task</td>
<td>≈6 M</td>
<td>≈160 M</td>
<td>≈190 M</td>
</tr>
<tr>
<td>WMT’15 translation task</td>
<td>≈6 M</td>
<td>≈160 M</td>
<td>≈190 M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set</th>
<th># Lines</th>
<th># Tok. en</th>
<th># Tok. fr</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMSI-Dev-Titles</td>
<td>100</td>
<td>1360</td>
<td>1834</td>
</tr>
<tr>
<td>LIMSI-Dev-Abs</td>
<td>900</td>
<td>24367</td>
<td>30560</td>
</tr>
<tr>
<td>LIMSI-Dev</td>
<td>1000</td>
<td>25727</td>
<td>32394</td>
</tr>
</tbody>
</table>

Table 1: Corpora used for training (left); development and test (right)
3. Arc scores. Each arc labelled \( u \) receives a score equal to the posterior unigram probability \( P(u|\varepsilon) \) of the system generating \( u \) at this position. \( P(u|\varepsilon) \) is computed as in (de Gispert et al., 2013):

\[
P(u|\varepsilon) = \frac{\sum_{E \in \varepsilon} \exp(\alpha H(E, F))}{\sum_{E' \in \varepsilon} \exp(\alpha H(E', F))},
\]

where \( \varepsilon \) is the space of translation hypotheses of a PBSMT system (a 10\( K \)-best list was chosen), and \( H(E, F) \) is the score assigned by the model to the sentence pair \( (E, F) \).

The posterior probabilities for the word (arcs) in NP-SMT and NP-Reduced-SMT hypotheses were rescaled to give more weight to local translation variants.

4. Choice of the combined hypotheses. The CN-DECODING diversity was increased by combining 30-best hypotheses from each system (baseline, NP-Reduced-SMT and NP-SMT).

Each path is the CN is finally scored as follows:

\[
S(E|F) = \alpha S_{post}(E) + \beta S_{LM}(E) + \gamma N_w(E),
\]

where \( S_{post} \) is the path posterior probability, \( S_{LM} \) is the interpolated LM score (In-domain-LM1 and Out-of-domain-LM), and \( N_w \) is the path length (excluding NULL arcs).

All CN-DECODING experiments, including the feature weight optimization (BLEU maximization), were performed using the SRI-LM (Stolcke, 2002) toolkit.

4.2 Details of System Building

We used the SCIELO-ABS and COCHRANE-ABS corpora, as well as LIMSIDEV and LIMSI TEST to create the NP-SMT and NP-Reduced-SMT training, development and test data. The NP-SMT source data contained the NPs extracted from the source side of all bitexts. The NP-Reduced-SMT target data contained the original source sentences with the NPs replaced by their heads (also preserving the associated article or possessive determiner) (Klein and Manning, 2003; Toutanova et al., 2003). The NP-SMT and NP-Reduced-SMT target data were created using the translations of the corresponding syntactic structures obtained from the fast_align...
source-target word alignments, where the non-aligned words were not considered. The NP-SMT training corpus was enriched with the titles and glossary corpora data (SCIELO-TITLES, COCHRANE-TITLES, MEDLINE-TITLES, PATTR-TITLES, UMLS, WIKIPEDIA).

These systems were built and tuned in a way similar to the baseline (see Figure 3). For each system, we prioritized NP-SMT or NP-Reduced-SMT model correspondingly, the other models being used as back-off models. LMs were built as explained in Section 2.4. We also used again the Out-of-domain-LM.

To evaluate these specialized systems, we compared the BLEU scores of the NP-SMT and NP-Reduced-SMT translations with artificial hypotheses derived from baseline hypotheses.

The results in Table 3 show small quality gains with our NP-SMT variant (+0.22 BLEU). Conversely, a slight decrease in quality (-0.35 BLEU) is observed for the NP-Reduced-SMT system. This is somewhat paradoxical, as we expected the simplified sentences to be easier to translate than the original sentences. This might be explained by the poor quality and frequent ungrammaticality of the NP-Reduced-SMT target side development and test sentences, the computation of which critically relies on word alignments.

<table>
<thead>
<tr>
<th>system</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP-SMT</td>
<td>27.47</td>
</tr>
<tr>
<td>NP-SMT + SOUL</td>
<td>28.46</td>
</tr>
<tr>
<td>base MOSES (NPs)</td>
<td>27.25</td>
</tr>
<tr>
<td>base MOSES + SOUL (NPs)</td>
<td>28.33</td>
</tr>
<tr>
<td>NP-Reduced-SMT</td>
<td>22.81</td>
</tr>
<tr>
<td>NP-Reduced-SMT + SOUL</td>
<td>23.53</td>
</tr>
<tr>
<td>base MOSES (NP-reduced)</td>
<td>23.16</td>
</tr>
<tr>
<td>base MOSES + SOUL (NP-reduced)</td>
<td>24.04</td>
</tr>
</tbody>
</table>

Table 3: NP-SMT and NP-Reduced-SMT performance for LIMSI/Test-2-Abs.

### 4.3 Experiments and Results

The resulting CN-DECODING 300-best lists were compared to the 300-best lists of the baseline system. On average, 11% of unique 1-grams from each CN-DECODING hypothesis search space are new (see Table 4), a significant proportion of novelty relative to our baseline system.

We also compared our approach to the MOSES xml-mode that enables to propose to the decoder alternative partial translations with their probability. Using 30-best lists of NP-SMT translations reranked by SOUL, we marked the source sentences with possible NP translations which competed with PT choices (inclusive option). Each NP translation variant was assigned a probability proportional to the $\prod_{0<n\leq n_p} P(u_n|c)$ of the 1-grams $u_n$ composing it. CN-DECODING decoding was performed according to the configuration described in Section 4.1, with the 30-best NP-SMT list reranked by SOUL.

Results in Table 5 confirm that CN-DECODING is superior here to MOSES xml-mode (+2.06 BLEU for LIMSI/Test-2-Abs).

<table>
<thead>
<tr>
<th>test set</th>
<th>MOSES base</th>
<th>MOSES + xml</th>
<th>CN-DECODING</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIMSI/Test2-Abs</td>
<td>32.38</td>
<td>29.59</td>
<td>32.84</td>
</tr>
<tr>
<td>LIMSI/Test2-Abs</td>
<td>29.20</td>
<td>28.79</td>
<td>28.85</td>
</tr>
</tbody>
</table>

Table 5: Results (BLEU) for different strategies of NP injection.

For the remaining CN-DECODING experiments, the 30-best lists of each system are reranked by SOUL prior to system combination.

We noticed that the NP-SMT and NP-Reduced-SMT hypotheses tend to be shorter than the corresponding local translations in the baseline output. We tried to reduce the negative impact on quality and avoided aligning baseline words to NULL in the CN-DECODING alignment procedure. We assigned the rest of the NULL arcs a very low probability of $p(\text{NULL}) = 0.001$ (compared to the previously assigned average score of all the other arcs between two consecutive nodes).

In this condition, the quality of CN-DECODING output reranked by SOUL shows an insignificant gain over the baseline MOSES + SOUL (+0.18 BLEU for LIMSI/Test2-Abs, see Tables 6, 2). It seems that the CN-DECODING procedure allowed our system to locally choose "good" translation variants, in spite of the quality decrease that we observed for NP-Reduced-SMT hypotheses (see
Table 3).

<table>
<thead>
<tr>
<th>test set</th>
<th>LIMSI Test2-ABS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN-DECODING</td>
<td>30.01</td>
</tr>
<tr>
<td>CN-DECODING + SOUL</td>
<td><strong>30.45</strong></td>
</tr>
</tbody>
</table>

Table 6: Results (BLEU) for CN-DECODING experiments.

4.4 Observations and Further Improvements

Manual inspection of the CN-DECODING output showed that the majority of the changes with respect to the baseline hypotheses concern introduction of synonyms, and only a few cases include the right choice of an article or of a grammatical form.

Our observations of the quality of the NP-SMT and NP-Reduced-SMT hypotheses suggest that the target development sets automatically created from source-target word alignments for those systems do not provide the right guidance for tuning, and also yield biased BLEU scores for these systems. More effort should be invested notably to compute better simplified versions of the original target sentences. Additionally, a more fine-grained procedure is required to estimate the quality of partial hypotheses before introducing them to CN-DECODING.

5 Conclusions

This paper described LIMSI’s submission to the shared WMT’16 biomedical translation task. We reported the results for the English-French translation direction. Our submitted system used MOSES and neural network SOUL models in a post-processing step.

In our experiments, we developed an approach aimed at mitigating the syntactic complexity which is a characteristic of a medical scientific publications. Our solution exploits the potential of phrase-based Statistical Machine Transla-

Acknowledgments

The work of the first author is supported by a CIFRE grant from the French ANRT.
References


Hai-Son Le, Thomas Lavergne, Alexandre Allauzen, Marianna Apidianaki, Li Gong, Aurélien Max, Artem Sokolov, Guillaume Wisniewski, and François Yvon. 2012b. LIMSI @ WMT12. In Proceedings of WMT, Montréal, Canada.


Benjamin Marie, Alexandre Allauzen, Franck Burilot, Quoc-Khanh Do, Julia Ives, Elena Knyazeva, Matthieu Labeau, Thomas Lavergne, Kevin Löser, Nicolas Pécheux, and François Yvon. 2015. LIMSI@WMT’15 : Translation task. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 145–151, Lisbon, Portugal, September.


Nicolas Pécheux, Li Gong, Quoc Khanh Do, Benjamin Marie, Yulia Ivanishcheva, Alexandre Allauzen, Thomas Lavergne, Jan Niehues, Aurélien Max, and François Yvon. 2014. LIMSI @ WMT14 Medical Translation Task. In Proceedings of WMT, Baltimore, Maryland.


