LIMSI @ WMT'15 : Translation Task

Benjamin Marie[1,2,3], **Alexandre Allauzen**[1,2], **Franck Burlot**[1], **Quoc-Khanh Do**[1,2],
Julia Ive[1,2,4], **Elena Knyazeva**[1,2], **Matthieu Labeau**[1,2], **Thomas Lavergne**[1,2],
Kevin Löser[1,2], **Nicolas Pécheux**[1,2], **François Yvon**[1]
[1]LIMSI-CNRS, 91 403 Orsay, France
[2]Université Paris-Sud, 91 403 Orsay, France
[3]Lingua et Machina
[4]Centre Cochrane français
firstname.lastname@limsi.fr

Abstract

This paper describes LIMSI's submissions to the shared WMT'15 translation task. We report results for French-English, Russian-English in both directions, as well as for Finnish-into-English. Our submissions use NCODE and MOSES along with continuous space translation models in a post-processing step. The main novelties of this year's participation are the following: for Russian-English, we investigate a tailored normalization of Russian to translate into English, and a two-step process to translate first into simplified Russian, followed by a conversion into inflected Russian. For French-English, the challenge is domain adaptation, for which only monolingual corpora are available. Finally, for the Finnish-to-English task, we explore unsupervised morphological segmentation to reduce the sparsity of data induced by the rich morphology on the Finnish side.

1 Introduction

This paper documents LIMSI's participation to the machine translation shared task for three language pairs: French-English and Russian-English in both directions, as well as Finnish-into-English. Each of these tasks poses its own challenges.

For French-English, the task differs slightly from previous years as it considers user-generated news discusssions. While the domain remains the same, the texts that need to be translated are of a less formal type. To cope with the style shift, new monolingual corpora have been made available; they represent the only available in-domain resources to adapt statistical machine translation (SMT) systems.

For Russian-English, the main source of difficulty is the processing of Russian, a morphologically rich language with a much more complex inflectional system than English. To mitigate the effects of having too many Russian word forms, we explore ways to normalize Russian prior to translation into English, so as to reduce the number of forms by removing some "redundant" morphological information. When translating into Russian, we consider a two-step scenario. A conventional SMT system is first built to translate from English into a simplified version of Russian; a post-processing step then restores the correct inflection wherever needed.

Finally, for Finnish-into-English, we report preliminary experiments that explore unsupervised morphological segmentation techniques to reduce the sparsity issue induced by the rich morphology of Finnish.

2 Systems Overview

Our experiments use NCODE[1], an open source implementation of the n-gram approach, as well as MOSES, which implements a vanilla phrase-based approach.[2] For more details about these toolkits, the reader can refer to (Koehn et al., 2007) for MOSES and to (Crego et al., 2011) for NCODE.

2.1 Tokenization and word alignments

Tokenization for French and English text relies on in-house text processing tools (Déchelotte et al., 2008). All bilingual corpora provided by the organizers were used, except for the French-English tasks where the UN corpus was not considered.[3] We also used a heavily filtered version of the Common Crawl corpus, where we discard all sentences pairs that do not look like proper French/English parallel sentences. For all cor-

[1] http://ncode.limsi.fr

[2] http://www.statmt.org/moses/

[3] In fact, when used in combination with the Giga Fr-En corpus, no improvement could be observed (Koehn and Haddow, 2012).

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 145–151,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

pora, we finally removed all sentence pairs that did not match the default criteria of the MOSES script `clean-corpus-n.pl` or that contained more than 70 tokens.

Statistics regarding the parallel corpora used to train SMT systems are reported in Table 1 for the three language pairs under study. Word-level alignments are computed using `fast_align` (Dyer et al., 2013) with options ”-d -o -v”.

2.2 Language Models

The English language model (LM) was trained on all the available English monolingual data, plus the English side of the bilingual data for the Fr-En, Ru-En and Fi-En language pairs. For the French language model, we also used all the provided monolingual data and the French side of the bilingual En-Fr data. We removed all duplicate lines[4] and trained a 4-gram language model, pruning all singletons, with `lmplz` (Heafield et al., 2013).

2.3 SOUL

Neural networks, working on top of conventional n-gram back-off language models, have been introduced in (Bengio et al., 2003; Schwenk et al., 2006) as a potential means to improve conventional language models. As in our previous participations (Le et al., 2012b; Allauzen et al., 2013; Pécheux et al., 2014), we take advantage of the proposal of (Le et al., 2011). Using a specific neural network architecture, the *Structured OUtput Layer* (SOUL), it becomes possible to estimate n-gram models that use large output vocabulary, thereby making the training of large neural network language models feasible both for target language models and translation models (Le et al., 2012a). Moreover, the peculiar parameterization of continuous models allows us to consider longer dependencies than the one used by conventional n-gram models (e.g. $n = 10$ instead of $n = 4$).

3 Experiments for French-English

This year, the French-English translation task focuses on user-generated News discusssions, a less formal type of texts than the usual News articles of the previous WMT editions. Therefore, the main challenge for this task is domain adaptation, for which only monolingual data are distributed.

3.1 Development and test sets

Since this is the first time this translation task is considered, only a small development set of newsdiscusssions is available. In order to properly tune and test our systems, we performed a 3-fold cross-validation, splitting the 1,500 in-domain sentences in two parts. Each random split respects document boundary, and yields roughly 1,000 sentences for tuning and 500 sentences for testing. The source of the documents, the newspapers *Le Monde* and *The Guardian* are also known. This allows us to balance the proportion of documents from each source in the development and test sets. The BLEU scores for the French-English experiments are computed on the concatenation of each test set decoded using weights tuned on the corresponding 1,000 sentence tuning set.

3.2 Domain adaptation

The vast majority of bilingual data distributed for the translation task are News articles, meaning that they correspond to a more formal register than the News discussions. The only in-domain texts provided for this task are monolingual corpora. Nevertheless, these monolingual data have been used to adapt both the translation and language models. To adapt the bilingual data, we subsampled the concatenation of the noisy Common Crawl and Giga Fr-En corpus, which represent around 90% of all our bilingual data, using the so-called Modified Moore-Lewis (Axelrod et al., 2011) filtering method (`MML`). We kept all the Europarl and News-Commentary data. `MML` expects 4 LMs to score sentence pairs in the corpus we wish to filter: for the source and target languages, it requires a LM trained with in-domain data, along with an out-of-domain LM estimated on the data to filter.[5] The `MML` score of a sentence pair is the sum of the source and target's perplexity differences for both in-domain and out-of-domain LMs. Sentences pairs are ranked according to the `MML` score and the top N parallel sentences are used to learn the translation table used during decoding.

For LM adaptation, we used a log-linear combination of our large LM with a smaller one trained only on the monolingual in-domain corpus.[6]

[4] Experiments not reported in this paper showed no changes in BLEU score between keeping or removing duplicate lines, but removing duplicate lines conveniently reduced the size of the models due to singleton pruning.

[5] All language models for the `MML` scoring are 4-grams trained with `lmplz`.

[6] Corresponding respectively to 3.5 and 50 millions sen-

Corpus	Fr-En		Ru-En		Fi→En	
	Sentences	Tokens (Fr-En)	Sentences	Tokens (Ru-En)	Sentences	Tokens (Fi-En)
parallel data	24.3M	712.8M-597.7M	2.3M	45.7M-47.3M	2M	37.3M-51.7M
monolingual data		2.2B-2.7B		834.7M-2.7B		-2.7B

Table 1: Statistical description of the training corpora

3.3 Reranking

The N-best reranking steps uses the following feature sets to find a better hypothesis among the 1,000-best hypotheses of the decoder:

- **IBM1**: IBM1 features (Hildebrand and Vogel, 2008);
- **POSLM**: 6-gram Witten-Bell smoothed POS LM trained with SRILM on all the monolingual news-discussions corpus;
- **SOUL**: Five features, one monolingual target language model and 4 translation models, see section 2.3 for details;
- **TagRatio**: ratio of translation hypothesis by number of source tokens tagged as verb, noun or adjective;
- **WPP**: count-based word posterior probability (Ueffing and Ney, 2007);

POS tagging is performed using the Stanford Tagger[7]. The reranking system is trained using the kb-mira algorithm (Cherry and Foster, 2012) implemented in MOSES.

3.4 Experimental results

For all French-English experiments, we used MOSES and NCODE with the default options, including lexicalized reordering models. Tuning is performed using kb-mira with default options on 200-best hypotheses.

Table 2 reports experimental results for filtering the bilingual data using MML before or after learning the word alignment step. Results for filtering are always lower when the word alignments are learnt only on the filtered data. The baseline system, which uses all the bilingual data, yields better performance than all our filtered systems, even though keeping only 25% of the bi-sentences, gives almost similar results. However, since there is no clear gain in filtering, we kept all the data without any MML filtering for the following experiments. The additional LM learned only on the in-domain data gives a slight improvement, +0.18 BLEU, for Fr-En, and a larger improvement for En-Fr (+0.66 BLEU, see Table 3).

Configuration		Fr-En
baseline		29.33
before	10%	28.63
	25%	29.09
	50%	28.96
after	10%	29.14
	25%	29.31
	50%	29.11

Table 2: Results (BLEU) for keeping the top 10%, 25% or 50% of the bi-sentences scored with MML, before and after word alignment. The baseline system uses all the bilingual data.

Configuration	Fr-En	En-Fr
w/o additional LM	29.15	29.56
w/ additional LM	29.33	30.22

Table 3: Results (BLEU) with and without the additional in-domain language model.

Table 4 reports the comparison between NCODE and MOSES. MOSES outperforms NCODE on our in-house test set using the 3-fold cross-validation procedure. However, when tuning on the complete development set and testing on the official test set, we observed a different result where NCODE outperforms MOSES for Fr-En (+0.69 BLEU), while MOSES remains the best choice for En-Fr (+0.74 BLEU). These differences between the results obtained with our dev/test configuration and the official ones may be due to the lack of tuning data when performing the 3-fold cross-validation, leaving only 1,000 sentences for tuning. Nonetheless, further investigations will be helpful to better understand these discrepancies.

Regarding reranking, results in Table 5 show that SOUL is the most useful feature and significantly improves translation performance when reranking a 1,000-best list generated by the decoder: we observe an improvement of nearly +0.9 BLEU for both translation directions. These re-

tences for French and English.

[7] http://nlp.stanford.edu/software/tagger.shtml

System	in-house test		official test	
	Fr-En	En-Fr	Fr-En	En-Fr
MOSES	29.33	30.22	32.16	35.74
NCODE	28.66	30.17	32.85	35.00

Table 4: Results (BLEU) for NCODE and MOSES on respectively the in-house and official test set.

Feature sets	Fr-En	En-Fr
baseline	29.33	30.22
+ **IBM1**	29.24	30.25
+ **POSLM**	29.45	30.28
+ **SOUL**	30.20	31.15
+ **TagRatio**	29.33	30.30
+ **WPP**	29.40	30.20
`all`	30.45	31.25

Table 5: Reranking results (BLEU) using different feature sets individually and their combination. For the `all` configurations these features are introduced during a reranking step.

sults can be further improved by adding more features during the reranking phase, with a final gain of +1.12 and +1.03 BLEU, for respectively Fr-En and En-Fr.

Our primary submissions for Fr-En and En-Fr use MOSES to generate n-best list, with phrase and reordering tables learned from all our bilingual data; the reranking step includes all the features presented in section 3.3.

4 Russian-English

Russian is a morphologically rich language characterized notably by a much more complex inflection system than English. This observation was the starting point of our work and led us to explore ways to process Russian in order to make it closer to English.

4.1 Preprocessing Russian

Inflections in Russian encode much more information than in English. For instance, while English adjectives are invariable, their Russian counterparts surface as twelve distinct word forms, expressing variations in gender (3), number (2) and case (6). Such a diversity of forms creates data sparsity issues, since many word forms are not observed in training corpora. When translating from Russian, the number of unknown words is accordingly high, making it impossible to translate many forms, even when they exist in the training corpus with a different inflection mark. Conversely, when translating into Russian, the system may not be able to generate the correct word form in a given context. Finally note that training translation models for such a language pair causes each English word to be typically paired with a lot of translations of low probability, corresponding to morphological variants on the Russian side.

To address this issue, we decided to normalize Russian by replacing all case marks by the corresponding nominative inflection: this applies to nouns, adjectives and pronouns. For these word types, the case information is thus lost, but the gender and number marks are preserved.

4.2 Predicting Case Marks

When translating into Russian, the normalization scheme described above is not well suited because of its lossy reduction of Russian word forms. Its use therefore requires a post processing step which aims to recover the inflected forms from the output of the SMT system. Since normalization essentially removes the case information, this last step consists in predicting the right case for a given normalized word before generating the correctly inflected form.

For this purpose, we designed a cascade of Conditional Random Fields (CRFs) models. A first model predicts POS tags, which are then used by a second model to predict the gender and number information. A last model is then used to infer the case from this information. POS, gender and number prediction are used to disambiguate the normalized words, which is necessary to generate the correct word forms. All predictions were performed considering only the target side output, meaning that no information from the source was used. The first two models use standard features for POS tagging as described in (Lavergne et al., 2010). The last one (for case prediction) additionally contains features testing the presence of a verb or a preposition in the close vicinity of the word under consideration.

4.3 Experimental results

Standard NCODE and MOSES configurations with lexicalized reordering models were used for all the English-Russian and Russian-English experiments. Alignments in both directions were computed with normalized Russian. The models were tuned with `kb-mira` using 300-best lists.

The results reported in Table 6 show a similar trend for NCODE and MOSES in both translation directions. Note that MOSES outperforms NCODE (+0.72 BLEU) for Ru-En task. Using normalized Russian as the source language allows us to achieve a slight gain of +0.4 over the baseline for both systems. Moreover, the addition of SOUL models yields a further improvement of 1.1 BLEU score (see Table 7). The English-into-normalized-Russian task has been performed for the sake of comparison, to assess the gain we could expect if we were able to always predict the right case for the normalized Russian output. The comparison of BLEU scores between translating directly into Russian and producing an intermediate normalized Russian shows differences of 3.15 BLEU for NCODE and 3.44 BLEU for MOSES. These scores represent an upper-bound that unfortunately we were not able to reach with our post-processing scheme.

System	MOSES	NCODE
Baseline	26.85	26.02
+ Normalized Ru	27.27	26.44
+ SOUL		27.28

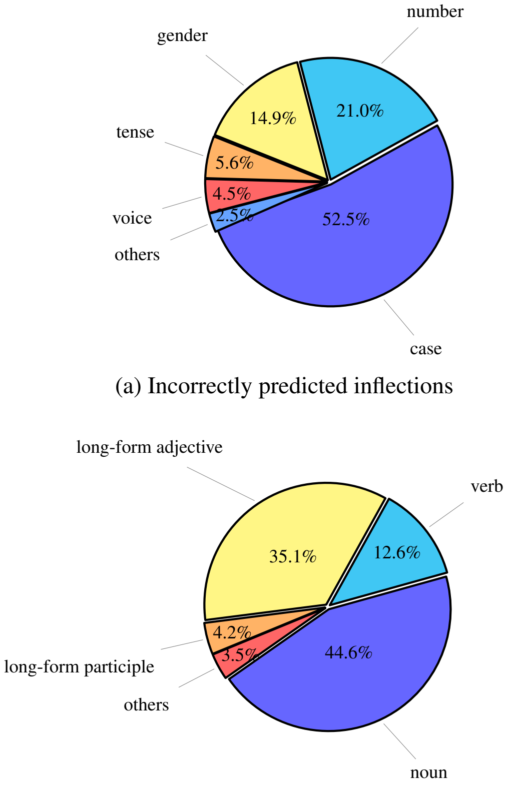
Table 6: Results (BLEU) for Russian-English with NCODE and MOSES on the official test.

System	MOSES	NCODE
Baseline	22.91	22.97
+ SOUL		24.08
En-Rx	26.35	26.12
En-Rx-Ru	19.99	19.88

Table 7: Results (BLEU) for English-Russian (Rx stands for normalized Russian) with NCODE and MOSES on the official test. The score for En-Rx was obtained over the normalized test.

4.4 Error Analysis

As Russian is a morphologically rich language, which has many features not observed in the English language, we conducted a simple error analysis to better understand the possible morphological mistakes made by our NCODE baseline. We used METEOR to automatically align the outputs with the original references at the word level, discarding multiple alignment links. About 56.3% of the words in the NCODE output have a counterpart in the human references, which is consistent with the BLEU unigram precision (53.3%). Among those, 85.4% are identical and 9.8% are different but share a common lemma. This last situation happens when our system fails to predict the correct form. The remaining 4.8% (different word forms with no common lemma), correspond either to synonyms or to METEOR alignment errors. Figure 1 also suggests that, within the 9.8% word form errors, most morphological errors are related to case prediction. Figure 1 displays detailed results split by POS. Results for MOSES or when rescoring NCODE outputs with SOUL are very similar.



(a) Incorrectly predicted inflections

(b) Word form errors *wrt* POS

Figure 1: Distribution of mispredictions for NCODE outputs, according to the mispredicted inflection (a) and their POS (b).

5 Translating Finnish into English

This is our first attempt to translate from Finnish to English. The provided development set contains only 1,500 parallel sentences. Therefore all the results are computed using a two-fold cross validation. The baseline system is a conventional phrase-based system built with the MOSES toolkit. Experimental results are in Table 8. The first two

Configuration	dev.	test
Baseline	13.2	12.8
+ large LM	16.1	15.7
+ Morph. segmentation	16.2	15.9

Table 8: BLEU scores for the Finnish to English translation task, obtained with different configurations after a two-fold cross-validation.

lines give the BLEU scores obtained with a basic tokenization of the Finnish side. When the English LM is only estimated on the parallel data, the system achieves a BLEU score of 12.8, while using a LM estimated on all the available monolingual data yields a 1.8 BLEU point improvement.

Finnish is a synthetic language that employs extensive regular agglutination. This peculiarity implies a large variety of word forms and, again, severe sparsity issues. For instance, we observed on the available parallel training data 860K different Finnish forms for 37.3M running words and only 2M sentences. Among these forms, more than half are hapax. For comparison purposes, we observed in English 208K word forms for 51.7M running words. To address this issue, we have tried to reduce the number of forms in the Finnish part of the data. For that purpose, we use Morfessor [8] to perform an unsupervised morphological segmentation. The new Finnish corpus therefore contains 67K types for 77M running words. With this new version, we obtain only a slight improvement of 0.2 BLEU point. We assume that the Finnish data was over-segmented and that a better tradeoff can be found with an extensive tuning of Morfessor.

6 Discussion and Conclusion

This paper described LIMSI's submissions to the shared WMT'15 translation task. We reported results for French-English, Russian-English in both direction, as well as for Finnish-into-English. Our submissions used NCODE and MOSES along with continuous space translation models in a post-processing step. Most of our efforts for this years participation were dedicated to domain adaptation and more importantly to explore different strategies when translating from and into a morphologically rich language.

For French-English, we experimented adaptation using only monolingual data that represents the targeted text, *i.e* news-discussions. Our attempt to filter the available parallel corpora did not bring any gain, while the use of an additional language model estimated on news-discussions yielded slight improvement.

When translating from Russian into English, small improvements were observed with a tailored normalization of Russian. This normalization was designed to reduce the number of word forms and to make it closer to English. However, experiments in the other direction were disappointing. While the first step that translates from English to the normalized version of Russian showed positive results, the second step designed to recover Russian inflected forms failed. This failure may be related to the cascade of statistical models, working solely on the target side. However, the reasons need to be better understood with a more detailed study.

To translate from Finnish into English, we explored the use of unsupervised morphological segmentation. Our attempt to reduce the number of forms on the Finnish side did not significantly change the the BLEU score. This under-performance can be explained by an over-segmentation of the Finnish data, and maybe a better tradeoff can be found with a more adapted segmentation strategy.

We finally reiterate our past observations that continuous space translation models used in a post-processing step always yielded significant improvements across the board.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This work has been partly funded by the European Unions Horizon 2020 research and innovation programme under grant agreement No. 645452 (QT21).

[8] https://github.com/aalto-speech/morfessor

References

Alexandre Allauzen, Nicolas Pécheux, Quoc Khanh Do, Marco Dinarelli, Thomas Lavergne, Aurélien Max, Hai-son Le, and François Yvon. 2013. LIMSI @ WMT13. In *Proceedings of WMT*, Sofia, Bulgaria.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*, Edinburgh, Scotland.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of NAACL-HLT*, Montréal, Canada.

Josep M. Crego, François Yvon, and José B. Mariño. 2011. N-code: an open-source bilingual N-gram SMT toolkit. *Prague Bulletin of Mathematical Linguistics*, 96.

Daniel Déchelotte, Gilles Adda, Alexandre Allauzen, Olivier Galibert, Jean-Luc Gauvain, Hélène Maynard, and François Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proceedings of NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL*, Atlanta, Georgia.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL*, Sofia, Bulgaria.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of AMTA*, Honolulu, Hawa.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *Proceedings of WMT*, Montréal, Canada.

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 *Proceedings of the ACL Demo*, Prague, Czech Republic.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings of ACL*, Uppsala, Sweden.

Hai-Son Le, Ilya Oparin, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. 2011. Structured output layer neural network language model. In *Proceedings of ICASSP*, Prague, Czech Republic.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012a. Continuous space translation models with neural networks. In *Proceedings of NAACL-HLT*, Montréal, Canada.

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.

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.

Holger Schwenk, Daniel Déchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL*, Morristown, US.

Nicola Ueffing and Hermann Ney. 2007. Word-Level Confidence Estimation for Machine Translation. *Computational Linguistics*, 33.