LIMSI @ WMT'14 Medical Translation Task

Nicolas Pécheux[1,2], **Li Gong**[1,2], **Quoc Khanh Do**[1,2], **Benjamin Marie**[2,3], **Yulia Ivanishcheva**[2,4], **Alexandre Allauzen**[1,2], **Thomas Lavergne**[1,2], **Jan Niehues**[2], **Aurélien Max**[1,2], **François Yvon**[2]
Univ. Paris-Sud[1], LIMSI-CNRS[2]
B.P. 133, 91403 Orsay, France
Lingua et Machina[3], Centre Cochrane français[4]
{firstname.lastname}@limsi.fr

Abstract

This paper describes LIMSI's submission to the first medical translation task at WMT'14. We report results for English-French on the subtask of sentence translation from summaries of medical articles. Our main submission uses a combination of NCODE (n-gram-based) and MOSES (phrase-based) output and continuous-space language models used in a post-processing step for each system. Other characteristics of our submission include: the use of sampling for building MOSES' phrase table; the implementation of the vector space model proposed by Chen et al. (2013); adaptation of the POS-tagger used by NCODE to the medical domain; and a report of error analysis based on the typology of Vilar et al. (2006).

1 Introduction

This paper describes LIMSI's submission to the first medical translation task at WMT'14. This task is characterized by high-quality input text and the availability of large amounts of training data from the same domain, yielding unusually high translation performance. This prompted us to experiment with two systems exploring different translation spaces, the n-gram-based NCODE (§2.1) and an on-the-fly variant of the phrase-based MOSES (§2.2), and to later combine their output. Further attempts at improving translation quality were made by resorting to continuous language model rescoring (§2.4), vector space subcorpus adaptation (§2.3), and POS-tagging adaptation to the medical domain (§3.3). We also performed a small-scale error analysis of the outputs of some of our systems (§5).

2 System Overview

2.1 NCODE

NCODE implements the bilingual n-gram approach to SMT (Casacuberta and Vidal, 2004; Mariño et al., 2006; Crego and Mariño, 2006) that is closely related to the standard phrase-based approach (Zens et al., 2002). In this framework, the translation is divided into two steps. To translate a source sentence $\mathbf{f}$ into a target sentence $\mathbf{e}$, the source sentence is first reordered according to a set of rewriting rules so as to reproduce the target word order. This generates a word lattice containing the most promising source permutations, which is then translated. Since the translation step is monotonic, the peculiarity of this approach is to rely on the n-gram assumption to decompose the joint probability of a sentence pair in a sequence of *bilingual* units called *tuples*.

The best translation is selected by maximizing a linear combination of feature functions using the following inference rule:

$$\mathbf{e}^* = \underset{\mathbf{e},\mathbf{a}}{\operatorname{argmax}} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{f}, \mathbf{e}, \mathbf{a}) \qquad (1)$$

where K feature functions (f_k) are weighted by a set of coefficients (λ_k) and $\mathbf{a}$ denotes the set of hidden variables corresponding to the reordering and segmentation of the source sentence. Along with the n-gram translation models and target n-gram language models, 13 conventional features are combined: 4 *lexicon models* similar to the ones used in standard phrase-based systems; 6 *lexicalized reordering models* (Tillmann, 2004; Crego et al., 2011) aimed at predicting the orientation of the next translation unit; a "weak" distance-based *distortion model*; and finally a *word-bonus model* and a *tuple-bonus model* which compensate for the system preference for short translations. Features are estimated during the training phase. Training source sentences are first reordered so as to match

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 246–253,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

the target word order by unfolding the word alignments (Crego and Mariño, 2006). Tuples are then extracted in such a way that a unique segmentation of the bilingual corpus is achieved (Mariño et al., 2006) and n-gram translation models are then estimated over the training corpus composed of tuple sequences made of surface forms or POS tags. Reordering rules are automatically learned during the unfolding procedure and are built using part-of-speech (POS), rather than surface word forms, to increase their generalization power (Crego and Mariño, 2006).

2.2 On-the-fly System (OTF)

We develop an alternative approach implementing an on-the-fly estimation of the parameter of a standard phrase-based model as in (Le et al., 2012b), also adding an inverse translation model. Given an input source file, it is possible to compute only those statistics which are required to translate the phrases it contains. As in previous works on on-the-fly model estimation for SMT (Callison-Burch et al., 2005; Lopez, 2008), we first build a suffix array for the source corpus. Only a limited number of translation examples, selected by deterministic random sampling, are then used by traversing the suffix array appropriately. A coherent translation probability (Lopez, 2008) (which also takes into account examples where translation extraction failed) is then estimated. As we cannot compute exactly an inverse translation probability (because sampling is performed independently for each source phrase), we resort to the following approximation:

$$p(\bar{f}|\bar{e}) = \min\left(1.0, \frac{p(\bar{e}|\bar{f}) \times freq(\bar{f})}{freq(\bar{e})}\right) \quad (2)$$

where the $freq(\cdot)$ is the number of occurrences of the given phrase in the whole corpus, and the numerator $p(\bar{e}|\bar{f}) \times freq(\bar{f})$ represents the predicted joint count of $\bar{f}$ and $\bar{e}$. The other models in this system are the same as in the default configuration of MOSES.

2.3 Vector Space Model (VSM)

We used the vector space model (VSM) of Chen et al. (2013) to perform domain adaptation. In this approach, each phrase pair $(\bar{f}, \bar{e})$ present in the phrase table is represented by a C-dimensional vector of TF-IDF scores, one for each sub-corpus, where C represents the number of sub-corpora (see Table 1). Each component $w_c(\bar{f}, \bar{e})$ is a standard TF-IDF weight of each phrase pair for the c^{th} sub-corpus. TF$(\bar{f}, \bar{e})$ is the raw joint count of $(\bar{f}, \bar{e})$ in the sub-corpus; the IDF$(\bar{f}, \bar{e})$ is the inverse document frequency across all sub-corpora.

A similar C-dimensional representation of the development set is computed as follows: we first perform word alignment and phrase pairs extraction. For each extracted phrase pair, we compute its TF-IDF vector and finally combine all vectors to obtain the vector for the develompent set:

$$w_c^{dev} = \sum_{j=0}^{J} \sum_{k=0}^{K} count_{dev}(\bar{f}_j, \bar{e}_k) w_c(\bar{f}_j, \bar{e}_k) \quad (3)$$

where J and K are the total numbers of source and target phrases extracted from the development data, respectively, and $count_{dev}(\bar{f}_j, \bar{e}_k)$ is the joint count of phrase pairs $(\bar{f}_j, \bar{e}_k)$ found in the development set. The similarity score between each phrase pair's vector and the development set vector is added into the phrase table as a VSM feature. We also replace the joint count with the marginal count of the source/target phrase to compute an alternative average representation for the development set, thus adding two VSM additional features.

2.4 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 discrete language models. As for our submitted translation systems to WMT'12 and WMT'13 (Le et al., 2012b; Allauzen et al., 2013), we take advantage of the recent 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 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 $n = 4$).

Additionally, continuous models can also be easily and efficiently adapted as in (Lavergne et al., 2011). Starting from a previously trained SOUL model, only a few more training epochs are

	Corpus	Sentences	Tokens (en-fr)	Description	wrd-lm	pos-lm
in-domain	COPPA	454 246	10-12M		-3	-15
	EMEA	324 189	6-7M		26	-1
	PATTR-ABSTRACTS	634 616	20-24M		22	21
	PATTR-CLAIMS	888 725	32-36M		6	2
	PATTR-TITLES	385 829	3-4M		4	-17
	UMLS	2 166 612	8-8M	term dictionary	-7	-22
	WIKIPEDIA	8 421	17-18k	short titles	-5	-13
out-of-domain	NEWSCOMMENTARY	171 277	4-5M		6	16
	EUROPARL	1 982 937	54-60M		-7	-33
	GIGA	9 625 480	260-319M		27	52
all parallel	all	17M	397-475M	concatenation	33	69
target-lm	medical-data		-146M		69	-
	wmt13-data		-2 536M		49	-
devel/test	DEVEL	500	10-12k	*khresmoi-summary*		
	LMTEST	3 000	61-69k	see Section 3.4		
	NEWSTEST12	3 003	73-82k	from WMT'12		
	TEST	1 000	21-26k	*khresmoi-summary*		

Table 1: Parallel corpora used in this work, along with the number of sentences and the number of English and French tokens, respectively. Weights (λ_k) from our best NCODE configuration are indicated for each sub-corpora's bilingual word language model (wrd-lm) and POS factor language model (pos-lm).

needed on a new corpus in order to adapt the parameters to the new domain.

3 Data and Systems Preparation

3.1 Corpora

We use all the available (constrained) medical data extracted using the scripts provided by the organizers. This resulted in 7 sub-corpora from the medical domain with distinctive features. As out-of-domain data, we reuse the data processed for WMT'13 (Allauzen et al., 2013).

For pre-processing of medical data, we closely followed (Allauzen et al., 2013) so as to be able to directly integrate existing translation and language models, using in-house text processing tools for tokenization and detokenization steps (Déchelotte et al., 2008). All systems are built using a "true case" scheme, but sentences fully capitalized (plentiful especially in PATTR-TITLES) are previously lowercased. Duplicate sentence pairs are removed, yielding a sentence reduction up to 70% for EMEA. Table 1 summarizes the data used along with some statistics after the cleaning and pre-processing steps.

3.2 Language Models

A medical-domain 4-gram language model is built by concatenating the target side of the parallel data and all the available monolingual data[1], with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1996), using the SRILM (Stolcke, 2002) and KENLM (Heafield, 2011) toolkits. Although more similar to term-to-term dictionaries, UMLS and WIKIPEDIA proved better to be included in the language model. The large out-of-domain language model used for WMT'13 (Allauzen et al., 2013) is additionaly used (see Table 1).

3.3 Part-of-Speech Tagging

Medical data exhibit many peculiarities, including different syntactic constructions and a specific vocabulary. As standard POS-taggers are known not to perform very well for this type of texts, we use a specific model trained on the Penn Treebank and on medical data from the MedPost project (Smith et al., 2004). We use Wapiti (Lavergne et al., 2010), a state-of-the-art CRF implementation, with a standard feature set. Adaptation is performed as in (Chelba and Acero, 2004) using the out-of-domain model as a prior when training the in-domain model on medical data. On a medical test set, this adaptation leads to a 8 point reduction of the error rate. A standard model is used for WMT'13 data. For the French side, due to the lack of annotaded data for the medical domain, corpora are tagged using the TreeTagger (Schmid, 1994).

[1] Attempting include one language model per sub-corpora yielded a significant drop in performance.

3.4 Proxy Test Set

For this first edition of a Medical Translation Task, only a very small development set was made available (DEVEL in Table 1). This made both system design and tuning challenging. In fact, with such a small development set, conventional tuning methods are known to be very unstable and prone to overfitting, and it would be suboptimal to select a configuration based on results on the development set only.[2] To circumvent this, we artificially created our own internal test set by randomly selecting 3 000 sentences out from the 30 000 sentences from PATTR-ABSTRACTS having the lowest perplexity according to 3-gram language models trained on both sides of the DEVEL set. This test set, denoted by LMTEST, is however highly biaised, especially because of the high redundancy in PATTR-ABSTRACTS, and should be used with great care when tuning or comparing systems.

3.5 Systems

NCODE We use NCODE with default settings, 3-gram bilingual translation models on words and 4-gram bilingual translation factor models on POS, for each included corpora (see Table 1) and for the concatenation of them all.

OTF When using our OTF system, all in-domain and out-of-domain data are concatenated, respectively. For both corpora, we use a maximum random sampling size of 1 000 examples and a maximum phrase length of 15. However, all sub-corpora but GIGA[3] are used to compute the vectors for VSM features. Decoding is done with MOSES[4] (Koehn et al., 2007).

SOUL Given the computational cost of computing n-gram probabilities with neural network models, we resort to a reranking approach. In the following experiments, we use 10-gram SOUL models to rescore 1 000-best lists. SOUL models provide *five* new features: a target language model score and four translation scores (Le et al., 2012a).

We reused the SOUL models trained for our participation to WMT'12 (Le et al., 2012b). Moreover, target language models are adapted by running 6 more epochs on the new medical data.

[2]This issue is traditionally solved in Machine Learning by folded cross-validation, an approach that would be too prohibitive to use here.

[3]The GIGA corpus is actually very varied in content.

[4]`http://www.statmt.org/moses/`

System Combination As NCODE and OTF differ in many aspects and make different errors, we use system combination techniques to take advantage of their complementarity. This is done by reranking the concatenation of the 1 000-best lists of both systems. For each hypothesis within this list, we use two global features, corresponding either to the score computed by the corresponding system or 0 otherwise. We then learn reranking weights using Minimum Error Rate Training (MERT) (Och, 2003) on the development set for this combined list, using only these two features (SysComb-2). In an alternative configuration, we use the two systems without the SOUL rescoring, and add instead the five SOUL scores as features in the system combination reranking (SysComb-7).

Evaluation Metrics All BLEU scores (Papineni et al., 2002) are computed using cased `multi-bleu` with our internal tokenization. Reported results correspond to the average and standard deviation across 3 optimization runs to better account for the optimizer variance (Clark et al., 2011).

4 Experiments

4.1 Tuning Optimization Method

MERT is usually used to optimize Equation 1. However, with up to 42 features when using SOUL, this method is known to become very sensitive to local minima. Table 2 compares MERT, a batch variant of the Margin Infused Relaxation Algorithm (MIRA) (Cherry and Foster, 2012) and PRO (Hopkins and May, 2011) when tuning an NCODE system. MIRA slightly outperforms PRO on DEVEL, but seems prone to overfitting. However this was not possible to detect before the release of the test set (TEST), and so we use MIRA in all our experiments.

	DEVEL	TEST
MERT	$47.0_{\pm 0.4}$	$44.1_{\pm 0.8}$
MIRA	$47.9_{\pm 0.0}$	$44.8_{\pm 0.1}$
PRO	$47.1_{\pm 0.1}$	$45.1_{\pm 0.1}$

Table 2: Impact of the optimization method during the tuning process on BLEU score, for a baseline NCODE system.

4.2 Importance of the Data Sources

Table 3 shows that using the out-of-domain data from WMT'13 yields better scores than only using the provided medical data only. Moreover, combining both data sources drastically boosts performance. Table 1 displays the weights (λ_k) given by NCODE to the different sub-corpora bilingual language models. Three corpora seems particulary useful: EMEA, PATTR-ABSTRACTS and GIGA. Note that several models are given a negative weight, but removing them from the model surprisingly results in a drop of performance.

	DEVEL	TEST
medical	$42.2_{\pm 0.1}$	$39.6_{\pm 0.1}$
WMT'13	$43.0_{\pm 0.1}$	$41.0_{\pm 0.0}$
both	$48.3_{\pm 0.1}$	$45.4_{\pm 0.0}$

Table 3: BLEU scores obtained by NCODE trained on medical data only, WMT'13 data only, or both.

4.3 Part-of-Speech Tagging

Using the specialized POS-tagging models for medical data described in Section 3.3 instead of a standart POS-tagger, a 0.5 BLEU points increase is observed. Table 4 suggests that a better POS tagging quality is mainly beneficial to the reordering mechanism in NCODE, in contrast with the POS-POS factor models included as features.

Reordering	Factor model	DEVEL	TEST
std	std	$47.9_{\pm 0.0}$	$44.8_{\pm 0.1}$
std	spec	$47.9_{\pm 0.1}$	$45.0_{\pm 0.1}$
spec	std	$48.4_{\pm 0.1}$	$45.3_{\pm 0.1}$
spec	spec	$48.3_{\pm 0.1}$	$45.4_{\pm 0.0}$

Table 4: BLEU results when using a standard POS tagging (std) or our medical adapted specialized method (spec), either for the reordering rule mechanism (Reordering) or for the POS-POS bilingual language models features (Factor model).

4.4 Development and Proxy Test Sets

In Table 5, we assess the importance of domain adaptation via tuning on the development set used and investigate the benefits of our internal test set.

Best scores are obtained when using the provided development set in the tuning process. Using NEWSTEST12 as development set unsurprisingly leads to poor results, as no domain adaptation is carried out. However, using LMTEST does not result in much better TEST score. We also note a positive correlation between DEVEL and TEST. From the first three columns, we decided to use the DEVEL data set as development set for our submission, which is *a posteriori* the right choice.

DEVEL	LMTEST	NEWSTEST12	TEST
$\boxed{48.3_{\pm 0.1}}$	$46.8_{\pm 0.1}$	$26.2_{\pm 0.1}$	$45.4_{\pm 0.0}$
$41.8_{\pm 0.2}$	$\boxed{48.9_{\pm 0.1}}$	$18.5_{\pm 0.1}$	$40.1_{\pm 0.1}$
$39.8_{\pm 0.1}$	$37.4_{\pm 0.2}$	$\boxed{29.0_{\pm 0.1}}$	$39.0_{\pm 0.3}$

Table 5: Influence of the choice of the development set when using our baseline NCODE system. Each row corresponds to the choice of a development set used in the tuning process, indicated by a surrounded BLEU score.

Table 6: Contrast of our two main systems and their combination, when adding SOUL language (LM) and translation (TM) models. Stars indicate an adapted LM. BLEU results for the best run on the development set are reported.

	DEVEL	TEST
NCODE	48.5	45.2
+ SOUL LM	49.4	45.7
+ SOUL LM*	49.8	45.9
+ SOUL LM + TM	50.1	47.0
+ SOUL LM*+ TM	50.1	47.0
OTF	46.6	42.5
+ VSM	46.9	42.8
+ SOUL LM	48.6	44.0
+ SOUL LM*	48.4	44.2
+ SOUL LM + TM	49.6	44.8
+ SOUL LM*+ TM	49.7	44.9
SysComb-2	50.5	46.6
SysComb-7	50.7	46.5

4.5 NCODE vs. OTF

Table 6 contrasts our different approaches. Preliminary experiments suggest that OTF is a comparable but cheaper alternative to a full MOSES system.[5] We find a large difference in performance,

[5] A control experiment for a full MOSES system (using a single phrase table) yielded a BLEU score of 45.9 on DEVEL and 43.2 on TEST, and took 3 more days to complete.

	extra	*missing*		*incorrect*					*unknown*		
	word	content	filler	disamb.	form	style	term	order	word	term	**all**
syscomb	4	13	20	47	62	8	18	21	1	11	**205**
OTF+VSM+SOUL	4	4	31	44	82	6	20	42	3	12	**248**

Table 7: Results for manual error analysis following (Vilar et al., 2006) for the first 100 test sentences.

NCODE outperforming OTF by 2.8 BLEU points on the TEST set. VSM does not yield any significant improvement, contrarily to the work of Chen et al. (2013); it may be the case all individual subcorpus are equally good (or bad) at approximating the stylistic preferences of the TEST set.

4.6 Integrating SOUL

Table 6 shows the substantial impact of adding SOUL models for both baseline systems. With only the SOUL LM, improvements on the test set range from 0.5 BLEU points for NCODE system to 1.2 points for the OTF system. The adaptation of SOUL LM with the medical data brings an additional improvement of about 0.2 BLEU points.

Adding all SOUL translation models yield an improvement of 1.8 BLEU points for NCODE and of 2.4 BLEU points with the OTF system using VSM models. However, the SOUL adaptation step has then only a modest impact. In future work, we plan to also adapt the translation models in order to increase the benefit of using in-domain data.

4.7 System Combination

Table 6 shows that performing the system combination allows a gain up to 0.6 BLEU points on the DEVEL set. However this gain does not transfer to the TEST set, where instead a drop of 0.5 BLEU is observed. The system combination using SOUL scores showed the best result over all of our other systems on the DEVEL set, so we chose this (*a posteriori* sub-obtimal) configuration as our main system submission.

Our system combination strategy chose for DEVEL about 50% hypotheses among those produced by NCODE and 25% hypotheses from OTF, the remainder been common to both systems. As expected, the system combination prefers hypotheses coming from the best system. We can observe nearly the same distribution for TEST.

5 Error Analysis

The high level of scores for automatic metrics encouraged us to perform a detailed, small-scale analysis of our system output, using the error types proposed by Vilar et al. (2006). A single annotator analyzed the output of our main submission, as well as our OTF variant. Results are in Table 7.

Looking at the most important types of errors, assuming the translation hypotheses were to be used for rapid assimilation of the text content, we find a moderate number of unknown terms and incorrectly translated terms. The most frequent error types include missing fillers, incorrect disambiguation, form and order, which all have some significant impact on automatic metrics. Comparing more specifically the two systems used in this small-scale study, we find that our combination (which reused more than 70% of hypotheses from NCODE) mostly improves over the OTF variant on the choice of correct word form and word order. We may attribute this in part to a more efficient reordering strategy that better exploits POS tags.

6 Conclusion

In this paper, we have demonstrated a successful approach that makes use of two flexible translation systems, an n-gram system and an on-the-fly phrase-based model, in a new medical translation task, through various approaches to perform domain adaptation. When combined with continuous language models, which yield additional gains of up to 2 BLEU points, moderate to high-quality translations are obtained, as confirmed by a fine-grained error analysis. The most challenging part of the task was undoubtedly the lack on an internal test to guide system development. Another interesting negative result lies in the absence of success for our configuration of the vector space model of Chen et al. (2013) for adaptation. Lastly, a more careful integration of medical terminology, as provided by the UMLS, proved necessary.

7 Acknowledgements

We would like to thank Guillaume Wisniewski and the anonymous reviewers for their helpful comments and suggestions.

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 the Workshkop on Statistical Machine Translation*, pages 62–69, Sofia, Bulgaria.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(6):1137–1155.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of ACL*, Ann Arbor, USA.

Francesco Casacuberta and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy classifier: Little data can help a lot. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Stanley F. Chen and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 310–318, Santa Cruz, NM.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in statistical machine translation. In *Proceedings of ACL*, Sofia, Bulgaria.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation : Controlling for Optimizer Instability. In *Better Hypothesis Testing for Statistical Machine Translation : Controlling for Optimizer Instability*, pages 176–181, Portland, Oregon.

Josep M. Crego and José B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215.

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*, 96:49–58.

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 *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1352–1362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Reinhard Kneser and Herman Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'95*, pages 181–184, Detroit, MI.

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 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Thomas Lavergne, Hai-Son Le, Alexandre Allauzen, and François Yvon. 2011. LIMSI's experiments in domain adaptation for IWSLT11. In Mei-Yuh Hwang and Sebastian Stüker, editors, *Proceedings of the heigth International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

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*, pages 5524–5527.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012a. Continuous space translation models with neural networks. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 39–48, Montréal, Canada, June. Association for Computational Linguistics.

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 the Seventh Workshop on Statistical Machine Translation*, pages 330–337, Montréal, Canada.

Adam Lopez. 2008. Tera-Scale Translation Models via Pattern Matching. In *Proceedings of COLING*, Manchester, UK.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrick Lambert, José A.R. Fonollosa, and Marta R. Costa-Jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, USA, July. 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.

Holger Schwenk, Daniel Dchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 723–730, Morristown, NJ, USA. Association for Computational Linguistics.

L. Smith, T. Rindflesch, and W. J. Wilbur. 2004. Medpost: a part of speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904, Denver, Colorado, September.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 101–104.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *LREC*, Genoa, Italy.

Richard Zens, Franz Joseph Och, and Herman Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *KI-2002: Advances in artificial intelligence*, volume 2479 of *LNAI*, pages 18–32. Springer Verlag.