LIMSI @ WMT'13

Alexandre Allauzen[1,2], **Nicolas Pécheux**[1,2], **Quoc Khanh Do**[1,2], **Marco Dinarelli**[2], **Thomas Lavergne**[1,2], **Aurélien Max**[1,2], **Hai-Son Le**[3], **François Yvon**[1,2]
Univ. Paris-Sud[1] and LIMSI-CNRS[2]
rue John von Neumann, 91403 Orsay cedex, France
{firstname.lastname}@limsi.fr
Vietnamese Academy of Science and Technology[3], Hanoi, Vietnam
lehaison@ioit.ac.vn

Abstract

This paper describes LIMSI's submissions to the shared WMT'13 translation task. We report results for French-English, German-English and Spanish-English in both directions. Our submissions use n-code, an open source system based on bilingual n-grams, and continuous space models in a post-processing step. The main novelties of this year's participation are the following: our first participation to the Spanish-English task; experiments with source pre-ordering; a tighter integration of continuous space language models using artificial text generation (for German); and the use of different tuning sets according to the original language of the text to be translated.

1 Introduction

This paper describes LIMSI's submissions to the shared translation task of the Eighth Workshop on Statistical Machine Translation. LIMSI participated in the French-English, German-English and Spanish-English tasks in both directions. For this evaluation, we used n-code, an open source in-house Statistical Machine Translation (SMT) system based on bilingual n-grams[1], and continuous space models in a post-processing step, both for translation and target language modeling.

This paper is organized as follows. Section 2 contains an overview of the baseline systems built with n-code, including the continuous space models. As in our previous participations, several steps of data pre-processing, cleaning and filtering are applied, and their improvement took a non-negligible part of our work. These steps are summarized in Section 3. The rest of the paper is devoted to the novelties of the systems submitted this year. Section 4 describes the system developed for our first participation to the Spanish-English translation task in both directions. To translate from German into English, the impact of source pre-ordering is investigated, and experimental results are reported in Section 5, while for the reverse direction, we explored a text sampling strategy using a 10-gram SOUL model to allow a tighter integration of continuous space models during the translation process (see Section 6). A final section discusses the main lessons of this study.

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

2 System overview

n-code implements the bilingual n-gram approach to SMT (Casacuberta and Vidal, 2004; Mariño et al., 2006; Crego and Mariño, 2006). In this framework, translation is divided in two steps: a source reordering step and a (monotonic) translation step. Source reordering is based on a set of learned rewrite rules that non-deterministically reorder the input words. Applying these rules result in a finite-state graph of possible source reorderings, which is then searched for the best possible candidate translation.

2.1 Features

Given a source sentence $\mathbf{s}$ of I words, the best translation hypothesis $\hat{\mathbf{t}}$ is defined as the sequence of J words that maximizes a linear combination of feature functions:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t},\mathbf{a}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{a}, \mathbf{s}, \mathbf{t}) \right\} \quad (1)$$

where λ_m is the weight associated with feature function h_m and $\mathbf{a}$ denotes an alignment between source and target phrases. Among the feature functions, the peculiar form of the translation model constitutes one of the main difference between the n-gram approach and standard phrase-based systems.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 62–69,
Sofia, Bulgaria, August 8-9, 2013

In addition to the translation model (TM), *fourteen* feature functions are combined: a *target-language model*; four *lexicon models*; six *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. The four *lexicon models* are similar to the ones used in standard phrase-based systems: two scores correspond to the relative frequencies of the tuples and two lexical weights are estimated from the automatic word alignments. The weight vector λ is learned using the Minimum Error Rate Training framework (MERT) (Och, 2003) and BLEU (Papineni et al., 2002) measured on *nt09* (newstest2009) as the optimization criteria.

2.2 Translation Inference

During decoding, source sentences are represented in the form of word lattices containing the most promising reordering hypotheses, so as to reproduce the word order modifications introduced during the tuple extraction process. Hence, only those reordering hypotheses are translated and are introduced using a set of reordering rules automatically learned from the word alignments. Part-of-speech (POS) information is used to increase the generalization power of these rules. Hence, rewrite rules are built using POS, rather than surface word forms (Crego and Mariño, 2006).

2.3 SOUL rescoring

Neural networks, working on top of conventional n-gram back-off language models (BOLMs), have been introduced in (Bengio et al., 2003; Schwenk et al., 2006) as a potential means to improve discrete language models (LMs). As for our last year participation (Le et al., 2012c), we take advantage of the recent proposal of Le et al. (2011). Using a specific neural network architecture (the *Structured OUtput Layer* or SOUL model), it becomes possible to estimate n-gram models that use large vocabulary, thereby making the training of large neural network LMs (NNLMs) feasible both for target language models and translation models (Le et al., 2012a). We use the same models as last year, meaning that the SOUL rescoring was used for all systems, except for translating into Spanish. See section 6 and (Le et al., 2012c) for more details.

3 Corpora and data pre-processing

Concerning data pre-processing, we started from our submissions from last year (Le et al., 2012c) and mainly upgraded the corpora and the associated language-dependent pre-processing routines. We used in-house text processing tools for the tokenization and detokenization steps (Déchelotte et al., 2008). Previous experiments have demonstrated that better normalization tools provide better BLEU scores: all systems are thus built using the "true-case" scheme.

As German is morphologically more complex than English, the default policy which consists in treating each word form independently is plagued with data sparsity, which severely impacts both training (alignment) and decoding (due to unknown forms). When translating from German into English, the German side is thus normalized using a specific pre-processing scheme (Allauzen et al., 2010; Durgar El-Kahlout and Yvon, 2010) which aims at reducing the lexical redundancy by (i) normalizing the orthography, (ii) neutralizing most inflections and (iii) splitting complex compounds. All parallel corpora were POS-tagged with the TreeTagger (Schmid, 1994); in addition, for German, fine-grained POS labels were also needed for pre-processing and were obtained using the RFTagger (Schmid and Laws, 2008).

For Spanish, all the availaible data are tokenized using FreeLing[2] toolkit (Padró and Stanilovsky, 2012), with default settings and some added rules. Sentence splitting and morphological analysis are disabled except for *del* $\rightarrow$ *de el* and *al* $\rightarrow$ *a el*. Moreover, a simple "true-caser" based on uppercase word frequency is used, and the specific Spanish punctuation signs "¿" and "¡" are removed and heuristically reintroduced in a post-processing step. All Spanish texts are POS-tagged also using Freeling. The EAGLES tag set is however simplified by truncating the category label to the first two symbols, in order to reduce the sparsity of the reordering rules estimated by n-code.

For the CommonCrawl corpus, we found that many sentences are not in the expected language. For example, in the French side of the French-English version, most of the first sentences are in English. Therefore, foreign sentence pairs are filtered out with a MaxEnt classifier that uses n-grams of characters as features (n is between 1 and 4). This filter discards approximatively 10%

[2]http://nlp.lsi.upc.edu/freeling/

of the sentence pairs. Moreover, we also observe that a lot of sentence pairs are not translation of each other. Therefore, an extra sentence alignment step is carried out using an in-house implementation of the tool described in (Moore, 2002). This last step discards approximately 20% of the corpus. For the Spanish-English task, the same filtering is applied to all the available corpora.

4 System development for the Spanish-English task

This is our first participation to the Spanish-English translation task in both directions. This section provides details about the development of n-code systems for this language pair.

4.1 Data selection and filtering

The CommonCrawl and UN corpora can be considered as very noisy and out-of-domain. As described in (Allauzen et al., 2011), to select a subset of parallel sentences, trigram LMs were trained for both Spanish and English languages on a subset of the available News data: the Spanish (resp. English) LM was used to rank the Spanish (resp. English) side of the corpus, and only those sentences with perplexity above a given threshold were selected. Finally, the two selected sets were intersected. In the following experiments, the filtered versions of these corpora are used to train the translation systems unless explicitly stated.

4.2 Spanish language model

To train the language models, we assumed that the test set would consist in a selection of recent news texts and all the available monolingual data for Spanish were used, including the Spanish Gigaword, Third Edition. A vocabulary is first defined by including all tokens observed in the News-Commentary and Europarl corpora. This vocabulary is then expanded with all words that occur more than 10 times in the recent news texts (LDC-2007-2011 and news-crawl-2011-2012). This procedure results in a vocabulary containing 372k words. Then, the training data are divided into 7 sets based on dates or genres. On each set, a standard 4-gram LM is estimated from the vocabulary using absolute discounting interpolated with lower order models (Kneser and Ney, 1995; Chen and Goodman, 1998). The resulting LMs are then linearly interpolated using coefficients chosen so as to minimise the perplexity evaluated on the development set (*nt08*).

	Corpora	*BLEU*	
		dev nt11	*test nt12*
es2en	N,E	30.2	33.2
	N,E,C	30.6	33.7
	N,E,U	30.3	33.6
	N,E,C,U	30.6	33.7
	N,E,C,U (nf)	30.7	33.6
en2es	N,E	32.2	33.3
	N,E,C,U	32.3	33.6
	N,E,C,U (nf)	32.5	33.9

Table 1: BLEU scores achieved with different sets of parallel corpora. All systems are baseline n-code with POS factor models. The following shorthands are used to denote corpora, : "N" stands for News-Commentary, "E" for Europarl, "C" for CommonCrawl, "U" for UN and (nf) for non filtered corpora.

4.3 Experiments

All reported results are averaged on 3 MERT runs. Table 1 shows the BLEU scores obtained with different corpora setups. We can observe that using the CommonCrawl corpus improves the performances in both directions, while the impact of the UN data is less important, especially when combined with CommonCrawl. The filtering strategy described in Section 4.2 has a slightly positive impact of +0.1 BLEU point for the Spanish-to-English direction but yields a 0.2 BLEU point decrease in the opposite direction.

For the following experiments, all the available corpora are therefore used: News-Commentary, Europarl, filtered CommonCrawl and UN. For each of these corpora, a bilingual n-gram model is estimated and used by n-code as one individual model score. An additionnal TM is trained on the concatenation all these corpora, resulting in a total of 5 TMs. Moreover, n-code is able to handle additional "factored" bilingual models where the source side words are replaced by the corresponding lemma or even POS tag (Koehn and Hoang, 2007). Table 2 reports the scores obtained with different settings.

In Table 2, *big* denotes the use of a wider context for n-gram TMs ($n = 4, 5, 4$ instead of $3, 4, 3$ respectively for word-based, POS-based and lemma-based TMs). Using POS factored

	Condition	*BLEU*	
		dev nt11	*test nt12*
es2en	base	30.3	33.5
	pos	30.6	33.7
	big-pos	30.7	33.7
	big-pos-lem	30.7	33.8
en2es	base	32.0	33.4
	pos	32.3	33.6
	big-pos	32.3	33.8
	big-pos-pos+	32.2	33.4

Table 2: BLEU scores for different configuration of factored translation models. The *big* prefix denotes experiments with the larger context for n-gram translation models.

models yields a significant BLEU improvement, as well as using a wider context for n-gram TMs. Since Spanish is morphologically richer than English, lemmas are introduced only on the Spanish side. An additionnal BLEU improvement is achieved by adding factored models based on lemmas when translating from Spanish to English, while in the opposite direction it does not seem to have any clear impact.

For English to Spanish, we also experimented with a 5-gram target factored model, using the whole morphosyntactic EAGLES tagset, (*pos+* in Table 2), to add some syntactic information, but this, in fact, proved harmful.

As several tuning sets were available, experiments were carried out with the concatenation of *nt09* to *nt11* as a tuning data set. This yields an improvement between 0.1 and 0.3 BLEU point when testing on *nt12* when translating from Spanish to English.

4.4 Submitted systems

For both directions, the submitted systems are trained on all the available training data, the corpora CommonCrawl and UN being filtered as described previously. A word-based TM and a POS factored TM are estimated for each training set. To translate from Spanish to English, the system is tuned on the concatenation of the *nt09* to *nt11* datasets with an additionnal 4-gram lemma-based factored model, while in the opposite direction, we only use *nt11*.

	dev nt09	*test nt11*
en2de	15.43	15.35
en-mod2de	15.06	15.00

Table 3: BLEU scores for pre-ordering experiments with a n-code system and the approach proposed by (Neubig et al., 2012)

5 Source pre-ordering for English to German translation

While distorsion models can efficiently handle short range reorderings, they are inadequate to capture long-range reorderings, especially for language pairs that differ significantly in their syntax. A promising workaround is the source pre-ordering method that can be considered similar, to some extent, to the reordering strategy implemented in n-code; the main difference is that the latter uses one deterministic (long-range) reordering on top of conventional distortion-based models, while the former only considers one single model delivering permutation lattices. The pre-ordering approach is illustrated by the recent work of Neubig et al. (2012), where the authors use a discriminatively trained ITG parser to infer a single permutation of the source sentence.

In this section, we investigate the use of this pre-ordering model in conjunction with the bilingual n-gram approach for translating English into German (see (Collins et al., 2005) for similar experiments with the reverse translation direction). Experiments are carried out with the same settings as described in (Neubig et al., 2012): given the source side of the parallel data (*en*), the parser is estimated to modify the original word order and to generate a new source side (*en-mod*); then a SMT system is built for the new language pair (*en-mod* $\rightarrow$ *de*). The same reordering model is used to reorder the test set, which is then translated with the *en-mod* $\rightarrow$ *de* system.

Results for these experiments are reported in Table 3, where *nt09* and *nt11* are respectively used as development and test sets. We can observe that applying pre-ordering on source sentences leads to small drops in performance for this language pair.

To explain this degradation, the histogram of token movements performed by the model on the pre-ordered training data is represented in Figure 1. We can observe that most of the movements are in the range $[-4, +6]$ (92% of the total occur-

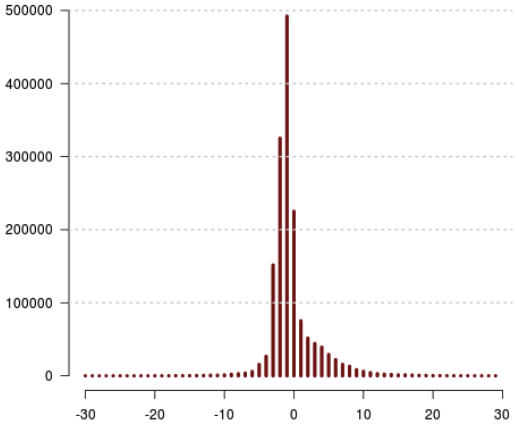

Figure 1: Histogram of token movement size versus its occurrences performed by the model *Neubig* on the source english data.

rences), which can be already taken into account by the standard reordering model of the baseline system. This is reflected also by the following statistics: surprisingly, only 16% of the total number of sentences are changed by the pre-ordering model, and the average sentence-wise Kendall's τ and the average displacement of these small parts of modified sentences are, respectively, 0.027 and 3.5. These numbers are striking for two reasons: first, English and German have in general quite different word order, thus our experimental condition should be somehow similar to the English-Japanese scenario studied in (Neubig et al., 2012); second, since the model is able to perform pre-ordering basically at any distance, it is surprising that a large part of the data remains unmodified.

6 Artificial Text generation with SOUL

While the context size for BOLMs is limited (usually up to 4-grams) because of sparsity issues, NNLMs can efficiently handle larger contexts up to 10-grams without a prohibitive increase of the overall number of parameters (see for instance the study in (Le et al., 2012b)). However the major bottleneck of NNLMs is the computation cost during both training and inference. In fact, the prohibitive inference time usually implies to resort to a two-pass approach: the first pass uses a conventional BOLM to produce a k-best list (the k most likely translations); in the second pass, the probability of a NNLM is computed for each hypothesis, which is then added as a new feature before the k-best list is reranked. Note that to produce the k-best list, the decoder uses a beam search strategy to prune the search space. Crucially, this pruning does not use the NNLMs scores and results in potentially sub-optimal k-best-lists.

6.1 Sampling texts with SOUL

In language modeling, a language is represented by a corpus that is approximated by a n-gram model. Following (Sutskever et al., 2011; Deoras et al., 2013), we propose an additionnal approximation to allow a tighter integration of the NNLM: a 10-gram NNLM is first estimated on the training corpus; texts then are sampled from this model to create an artificial training corpus; finally, this artificial corpus is approximated by a 4-gram BOLM.

The training procedure for the SOUL NNLM is the same as the one described in (Le et al., 2012c). To sample a sentence from the SOUL model, first the sentence length is randomly drawn from the empirical distribution, then each word of the sentence is sampled from the 10-gram distribution estimated with the SOUL model.

The convergence of this sampling strategy can be evaluated by monitoring the perplexity evolution *vs.* the number of sentences that are generated. Figure 2 depicts this evolution by measuring perplexity on the *nt08* set with a step size of 400M sampled sentences. The baseline BOLM (*std*) is estimated on all the available training data that consist of approximately 300M of running words. We can observe that the perplexity of the BOLM estimated on sampled texts (*generated texts*) decreases when the number of sample sentences increases, and tends to reach slowly the perplexity of the baseline BOLM. Moreover, when both BOLMs are interpolated, an even lower perplexity is obtained, which further decreases with the amount of sampled training texts.

6.2 Translation results

Experiments are run for translation into German, which lacks a GigaWord corpus. An artificial corpus containing 3 billions of running words is first generated as described in Section 6.1. This corpus is used to estimate a BOLM with standard settings, that is then used for decoding, thereby approximating the use of a NNLM during the first pass. Results reported in Table 4 show that adding generated texts improves the BLEU scores even when the SOUL model is added in a rescoring step. Also note that using the LM trained on the sampled corpus yields the same BLEU score that using the standard LM.

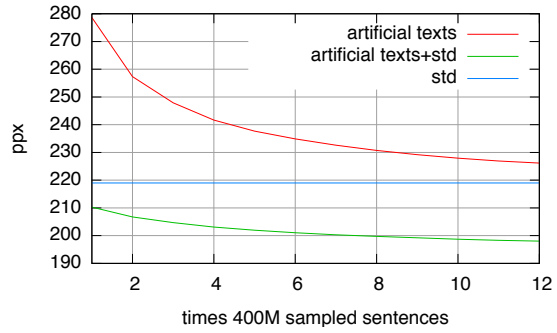

Figure 2: Perplexity measured on *nt08* with the baseline LM (*std*), with the LM estimated on the sampled texts (*generated texts*), and with the interpolation of both.

Therefore, to translate from English to German, the submitted system includes three BOLMs: one trained on all the monolingual data, one on artificial texts and a third one that uses the freely available *deWack* corpus[3] (1.7 billion words).

target LM	*BLEU*	
	dev nt09	test nt10
base	15.3	16.5
+genText	15.5	16.8
+SOUL	16.4	17.6
+genText+SOUL	16.5	17.8

Table 4: Impact of the use of sampled texts.

7 Different tunings for different original languages

As shown by Lembersky et al. (2012), the original language of a text can have a significant impact on translation performance. In this section, this effect is assessed on the French to English translation task. Training one SMT system per original language is impractical, since the required information is not available for most of parallel corpora. However, metadata provided by the WMT evaluation allows us to split the development and test sets according to the original language of the text. To ensure a sufficient amount of texts for each condition, we used the concatenation of newstest corpora for the years 2008, 2009, 2011, and 2012, leaving *nt10* for testing purposes.

Five different development sets have been created to tune five different systems. Experimental results are reported in Table 7 and show a drastic

[3] http://wacky.sslmit.unibo.it/doku.php

original language	baseline	adapted tuning
cz	22.31	23.83
en	36.41	39.21
fr	31.61	32.41
de	18.46	18.49
es	30.17	29.34
all	29.43	30.12

Table 5: BLEU scores for the French-to-English translation task measured on nt10 with systems tuned on development sets selected according to their original language (*adapted tuning*).

improvement in terms of BLEU score when translating back to the original English and a significant increase for original text in Czech and French. In this year's evaluation, Russian was introduced as a new language, so for sentences originally in this language, the baseline system was used. This system is used as our primary submission to the evaluation, with additional SOUL rescoring step.

8 Conclusion

In this paper, we have described our submissions to the translation task of WMT'13 for the French-English, German-English and Spanish-English language pairs. Similarly to last year's systems, our main submissions use n-code, and continuous space models are introduced in a post-processing step, both for translation and target language modeling. To translate from English to German, we showed a slight improvement with a tighter integration of the continuous space language model using a text sampling strategy. Experiments with pre-ordering were disappointing, and the reasons for this failure need to be better understood. We also explored the impact of using different tuning sets according to the original language of the text to be translated. Even though the gain vanishes when adding the SOUL model in a post-processing step, it should be noted that due to time limitation this second step was not tuned accordingly to the original language. We therefore plan to assess the impact of using different tuning sets on the post-processing step.

Acknowledgments

This work was partially funded by the French State agency for innovation (OSEO), in the Quaero Programme.

References

Alexandre Allauzen, Josep M. Crego, İlknur Durgar El-Kahlout, and François Yvon. 2010. LIMSI's statistical translation systems for WMT'10. In *Proc. of the Joint Workshop on Statistical Machine Translation and MetricsMATR*, pages 54–59, Uppsala, Sweden.

Alexandre Allauzen, Gilles Adda, Hélène Bonneau-Maynard, Josep M. Crego, Hai-Son Le, Aurélien Max, Adrien Lardilleux, Thomas Lavergne, Artem Sokolov, Guillaume Wisniewski, and François Yvon. 2011. LIMSI @ WMT11. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 309–315, Edinburgh, Scotland, July. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *JMLR*, 3:1137–1155.

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

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard Un iversity.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan.

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, Franois Yvon, and Jos 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, Hlne Maynard, and Franois Yvon. 2008. LIMSI's statistical translation systems for WMT'08. In *Proc. of the NAACL-HTL Statistical Machine Translation Workshop*, Columbus, Ohio.

Anoop Deoras, Tomáš Mikolov, Stefan Kombrink, and Kenneth Church. 2013. Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model. *Speech Communication*, 55(1):162 – 177.

Ilknur Durgar El-Kahlout and Franois Yvon. 2010. The pay-offs of preprocessing for German-English Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and Franois Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 251–258.

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 and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876.

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'11*, pages 5524–5527.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012a. Continuous space translation models with neural networks. In *NAACL '12: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Hai-Son Le, Alexandre Allauzen, and François Yvon. 2012b. Measuring the influence of long range dependencies with neural network language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 1–10, Montréal, Canada.

Hai-Son Le, Thomas Lavergne, Alexandre Allauzen, Marianna Apidianaki, Li Gong, Aurélien Max, Artem Sokolov, Guillaume Wisniewski, and François Yvon. 2012c. Limsi @ wmt12. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 330–337, Montréal, Canada.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Comput. Linguist.*, 38(4):799–825, December.

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.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, Tiburon, CA, USA. Springer-Verlag.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853, Jeju Island, Korea, July. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proc. of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784, Manchester, UK, August.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Holger Schwenk, Daniel Déchelotte, and Jean-Luc Gauvain. 2006. Continuous space language models for statistical machine translation. In *Proc. COLING/ACL'06*, pages 723–730.

Ilya Sutskever, James Martens, and Geoffrey Hinton. 2011. Generating text with recurrent neural networks. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 1017–1024, New York, NY, USA, June. ACM.

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