LIMSI @ WMT'12

Hai-Son Le[1,2], **Thomas Lavergne**[2], **Alexandre Allauzen**[1,2],
Marianna Apidianaki[2], **Li Gong**[1,2], **Aurélien Max**[1,2],
Artem Sokolov[2], **Guillaume Wisniewski**[1,2], **François Yvon**[1,2]
Univ. Paris-Sud[1] and LIMSI-CNRS[2]
rue John von Neumann, 91403 Orsay cedex, France
{firstname.lastname}@limsi.fr

Abstract

This paper describes LIMSI's submissions to the shared translation task. We report results for French-English and German-English in both directions. Our submissions use n-code, an open source system based on bilingual n-grams. In this approach, both the translation and target language models are estimated as conventional smoothed n-gram models; an approach we extend here by estimating the translation probabilities in a continuous space using neural networks. Experimental results show a significant and consistent BLEU improvement of approximately 1 point for all conditions. We also report preliminary experiments using an "on-the-fly" translation model.

1 Introduction

This paper describes LIMSI's submissions to the shared translation task of the Seventh Workshop on Statistical Machine Translation. LIMSI participated in the French-English and German-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]. The main novelty of this year's participation is the use, in a large scale system, of the continuous space translation models described in (Hai-Son et al., 2012). These models estimate the n-gram probabilities of bilingual translation units using neural networks. We also investigate an alternative approach where the translation probabilities of a phrase based system are estimated "on-the-fly" by sampling relevant examples, instead of considering the entire training set. Finally we also describe the use in a rescoring step of several additional features based on IBM1 models and word sense disambiguation information.

The rest of this paper is organized as follows. Section 2 provides an overview of the baseline systems built with n-code, including the standard translation model (TM). The continuous space translation models are then described in Section 3. As in our previous participations, several steps of data preprocessing, cleaning and filtering are applied, and their improvement took a non-negligible part of our work. These steps are summarized in Section 5. The last two sections report experimental results obtained with the "on-the-fly" system in Section 6 and with n-code in Section 7.

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 fea-

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

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 330–337,
Montréal, Canada, June 7-8, 2012.

ture 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 constitute one of the main difference between the n-gram approach and standard phrase-based systems. This will be further detailled in section 2.2 and 3.

In addition to the translation model, *fourteen* feature functions are combined: a *target-language model* (Section 5.3); four *lexicon models*; six *lexicalized reordering models* (Tillmann, 2004; Crego et al., 2011) aiming 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 weights vector λ is learned using a discriminative training framework (Och, 2003) (Minimum Error Rate Training (MERT)) using the *newstest2009* as development set and BLEU (Papineni et al., 2002) as the optimization criteria.

2.2 Standard n-gram translation models

n-gram translation models rely on a specific decomposition of the joint probability of a sentence pair $P(\mathbf{s}, \mathbf{t})$: a sentence pair is assumed to be decomposed into a sequence of L bilingual units called *tuples* defining a joint segmentation: $(\mathbf{s}, \mathbf{t}) = u_1, ..., u_L$[2]. In the approach of (Mariño et al., 2006), this segmentation is a by-product of source reordering obtained by "unfolding" initial word alignments.

In this framework, the basic translation units are *tuples*, which are the analogous of phrase pairs and represent a matching $u = (\overline{s}, \overline{t})$ between a source $\overline{s}$ and a target $\overline{t}$ phrase (see Figure 1). Using the n-gram assumption, the joint probability of a segmented sentence pair decomposes as:

$$P(\mathbf{s}, \mathbf{t}) = \prod_{i=1}^{L} P(u_i | u_{i-1}, ..., u_{i-n+1}) \quad (2)$$

During the training phase (Mariño et al., 2006), tuples are extracted from a word-aligned corpus (using MGIZA++[3] with default settings) in such a way that a unique segmentation of the bilingual corpus is achieved. A baseline n-gram translation model is then estimated over a training corpus composed of tuple sequences using modified Knesser-Ney Smoothing (Chen and Goodman, 1998).

2.3 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 they are introduced using a set of reordering rules automatically learned from the word alignments.

In the example in Figure 1, the rule [*prix nobel de la paix* $\rightsquigarrow$ *nobel de la paix prix*] reproduces the invertion of the French words that is observed when translating from French into English. Typically, 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).

3 SOUL translation models

A first issue with the model described by equation (2) is that the elementary units are bilingual pairs. As a consequence, the underlying vocabulary, hence the number of parameters, can be quite large, even for small translation tasks. Due to data sparsity issues, such model are bound to face severe estimation problems. Another problem with (2) is that the source and target sides play symmetric roles: yet, in decoding, the source side is known and only the target side must be predicted.

3.1 A word factored translation model

To overcome these issues, the n-gram probability in equation (2) can be factored by decomposing tuples

[2] From now on, $(\mathbf{s}, \mathbf{t})$ thus denotes an *aligned* sentence pair, and we omit the alignment variable $\mathbf{a}$ in further developments.

[3] `http://www.kyloo.net/software/doku.php`

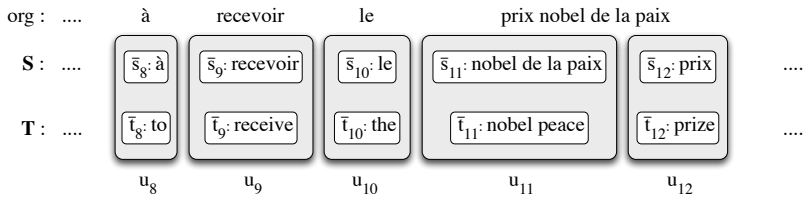

Figure 1: Extract of a French-English sentence pair segmented into bilingual units. The original (*org*) French sentence appears at the top of the figure, just above the reordered source $\mathbf{s}$ and target $\mathbf{t}$. The pair $(\mathbf{s}, \mathbf{t})$ decomposes into a sequence of L bilingual units (*tuples*) $u_1, ..., u_L$. Each tuple u_i contains a source and a target phrase: $\bar{s}_i$ and $\bar{t}_i$.

in two parts (source and target), and by taking words as the basic units of the n-gram TM. This may seem to be a regression with respect to current state-of-the-art SMT systems, as the shift from the word-based model of (Brown et al., 1993) to the phrase-based models of (Zens et al., 2002) is usually considered as a major breakthrough of the recent years. Indeed, one important motivation for considering phrases was to capture local context in translation and reordering. It should however be emphasized that the decomposition of phrases into words is only re-introduced here as a way to mitigate the parameter estimation problems. Translation units are still pairs of *phrases*, derived from a bilingual segmentation in tuples synchronizing the source and target n-gram streams. In fact, the estimation policy described in section 4 will actually allow us to take into account *larger contexts* than is possible with conventional n-gram models.

Let s_i^k denote the k^{th} word of source tuple $\bar{s}_i$. Considering the example of Figure 1, s_{11}^1 denotes the source word *nobel*, s_{11}^4 the source word *paix*. We finally denote $h^{n-1}(t_i^k)$ the sequence made of the $n-1$ words preceding t_i^k in the target sentence: in Figure 1, $h^3(t_{11}^2)$ thus refers to the three words context *receive the nobel* associated with t_{11}^2 *peace*. Using these notations, equation (2) is rewritten as:

$$P(\mathbf{a}, \mathbf{s}, \mathbf{t}) = \prod_{i=1}^{L} \Big[\prod_{k=1}^{|\bar{t}_i|} P\big(t_i^k | h^{n-1}(t_i^k), h^{n-1}(s_{i+1}^1)\big) \times \prod_{k=1}^{|\bar{s}_i|} P\big(s_i^k | h^{n-1}(t_i^1), h^{n-1}(s_i^k)\big) \Big] \quad (3)$$

This decomposition relies on the n-gram assumption, this time at the word level. Therefore, this model estimates the joint probability of a sentence pair using two sliding windows of length n, one for each language; however, the moves of these windows remain synchronized by the tuple segmentation. Moreover, the context is not limited to the current phrase, and continues to include words from adjacent phrases. Using the example of Figure 1, the contribution of the target phrase $\bar{t}_{11} =$ *nobel, peace* to $P(\mathbf{s}, \mathbf{t})$ using a 3- gram model is:

$$P\big(\text{nobel}|[\text{receive, the}], [\text{la, paix}]\big) \times P\big(\text{peace}|[\text{the, nobel}], [\text{la, paix}]\big).$$

A benefit of this new formulation is that the vocabularies involved only contain words, and are thus much smaller that tuple vocabularies. These models are thus less at risk to be plagued by data sparsity issues. Moreover, the decomposition (3) now involves two models: the first term represents a TM, the second term is best viewed as a reordering model. In this formulation, the TM only predicts the target phrase, given its source and target contexts.

$$P(\mathbf{s}, \mathbf{t}) = \prod_{i=1}^{L} \Big[\prod_{k=1}^{|\bar{s}_i|} P\big(s_i^k | h^{n-1}(s_i^k), h^{n-1}(t_{i+1}^1)\big) \times \prod_{k=1}^{|\bar{t}_i|} P\big(t_i^k | h^{n-1}(s_i^1), h^{n-1}(t_i^k)\big) \Big] \quad (4)$$

4 The principles of SOUL

In section 3.1, we defined a n-gram translation model based on equations (3) and (4). A major difficulty with such models is to reliably estimate their parameters, the numbers of which grow exponentially with the order of the model. This problem is aggravated in natural language processing due to

the well-known data sparsity issue. In this work, 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 handle large vocabulary language modeling tasks. This approach was experimented last year for target language models only and is now extended to translation models. More details about the SOUL architecture can be found in (Le et al., 2011), while its extension to translation models is more precisely described in (Hai-Son et al., 2012).

The integration of SOUL models for large SMT tasks is carried out using a two-pass approach: the first pass uses conventional back-off n-gram translation and language models to produce a k-best list (the k most likely translations); in the second pass, the probability of a m-gram SOUL model is computed for each hypothesis and the k-best list is accordingly reordered. In all the following experiments, we used a context size for SOUL of $m = 10$, and used $k = 300$. The two decompositions of equations (3) and (4) are used by introducing 4 scores during the rescoring step.

5 Corpora and data pre-processing

Concerning data pre-processing, we started from our submissions from last year (Allauzen et al., 2011) and mainly upgraded the corpora and the associated language-dependent pre-processing routines.

5.1 Pre-processing

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 in "true-case". Compared to last year, the pre-processing of utf-8 characters was significantly improved.

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 (described in (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 RF-Tagger (Schmid and Laws, 2008).

5.2 Bilingual corpora

As for last year's evaluation, we used all the available parallel data for the German-English language pair, while only a subpart of the French-English parallel data was selected. Word alignment models were trained using all the data, whereas the translation models were estimated on a subpart of the parallel data: the UN corpus was discarded for this step and about half of the French-English Giga corpus was filtered based on a perplexity criterion as in (Allauzen et al., 2011)).

For French-English, we mainly upgraded the training material from last year by extracting the new parts from the common data. The word alignment models trained last year were then updated by running a forced alignment [4] of the new data. These new word-aligned data was added to last year's parallel corpus and constitute the training material for the translation models and feature functions described in Section 2. Given the large amount of available data, three different bilingual n-gram models are estimated, one for each source of data: News-Commentary, Europarl, and the French-English Giga corpus. These models are then added to the weighted mixture defined by equation (1). For German-English, we simply used all the available parallel data to train one single translation models.

5.3 Monolingual corpora and language models

For the monolingual training data, we also used the same setup as last year. For German, all the training data allowed in the constrained task were divided into several sets based on dates or genres: News-Commentary, the news crawled from the Web grouped by year, and Europarl. For each subset, a standard 4-gram LM was estimated using interpolated Kneser-Ney smoothing (Kneser and Ney,

[4] The forced alignment step consists in an additional EM iteration.

1995; Chen and Goodman, 1998). The resulting LMs are then linearly combined using interpolation coefficients chosen so as to minimize the perplexity of the development set. The German vocabulary is created using all the words contained in the parallel data and expanded to reach a total of 500k words by including the most frequent words observed in the monolingual News data for 2011.

For French and English, the same monolingual corpora as last year were used[5]. We did not observe any perplexity decrease in our attempts to include the new data specifically provided for this year's evaluation. We therefore used the same language models as in (Allauzen et al., 2011).

6 "On-the-fly" system

We also developped an alternative approach implementing "on-the-fly" estimation of the parameter of a standard phase-based model, using Moses (Koehn et al., 2007) as the decoder. Implementing on-the-fly estimation for n-code, while possible in theory, is less appealing due to the computational cost of estimating a smoothed language 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 compute a suffix array for the source corpus. This further enables to consider only a subset of translation examples, which we select by deterministic random sampling, meaning that the sample is chosen randomly with respect to the full corpus but that the same sample is always returned for a given value of sample size, hereafter denoted N. In our experiments, we used $N = 1,000$ and computed from the sample and the word alignments (we used the same tokenization and word alignments as in all other submitted systems) the same translation[6] and lexical reordering models as the standard training scripts of the Moses system.

Experiments were run on the data sets used for WMT English-French machine translation evaluation tasks, using the same corpora and optimization procedure as in our other experiments. The only notable difference is our use of the Moses decoder instead of the n-gram-based system. As shown in Table 1, our on-the-fly system achieves a result (31.7 BLEU point) that is slightly worst than the n-code baseline (32.0) and slightly better than the equivalent Moses baseline (31.5), but does it much faster. Model estimation for the test file is reduced to 2 hours and 50 minutes, with an additional overhead for loading and writing files of one and a half hours, compared to roughly 210 hours for our baseline systems under comparable hardware conditions.

7 Experimental results

7.1 n-code with SOUL

Table 1 summarizes the experimental results submitted to the shared translation for French-English and German-English in both directions. The performances are measured in terms of BLEU on *newstest2011*, last year's test set, and this year's test set *newstest2012*. For the former, BLEU scores are computed with the NIST script *mteva-v13.pl*, while we provide for *newstest2012* the results computed by the organizers [7]. The *Baseline* results are obtained with standard n-gram models estimated with back-off, both for the bilingual and monolingual target models. With standard n-gram estimates, the order is limited to $n = 4$. For instance, the n-code French-English baseline achieves a 0.5 BLEU point improvement over a Moses system trained with the same data setup in both directions.

From Table 1, it can be observed that adding the SOUL models (translation models and target language model) consistently improves the baseline, with an increase of 1 BLEU point. Contrastive experiments show that the SOUL target LM does not bring significant gain when added to the SOUL translation models. For instance, a gain of 0.3 BLEU point is observed when translating from French to English with the addition of the SOUL target LM. In the other translation directions, the differences are negligible.

[5] The fifth edition of the English Gigaword (LDC2011T07) was *not* used.

[6] An approximation is used for $p(f|e)$, and *coherent* translation estimation is used; see (Lopez, 2008).

[7] All results come from the official website: `http://matrix.statmt.org/matrix/`.

Direction	*System*	BLEU	
		test2011	*test2012**
en2fr	Baseline	32.0	28.9
	+ SOUL TM	33.4	29.9
	on-the-fly	31.7	28.6
fr2en	Baseline	30.2	30.4
	+ SOUL TM	31.1	31.5
en2de	Baseline	15.4	16.0
	+ SOUL TM	16.6	17.0
de2en	Baseline	21.8	22.9
	+ SOUL TM	22.8	23.9

Table 1: Experimental results in terms of BLEU scores measured on the newstest2011 and newstest2012. For newstest2012, the scores are provided by the organizers.

7.2 Experiments with additional features

For this year's evaluation, we also investigated several additional features based on IBM1 models and word sense disambiguation (WSD) information in rescoring. As for the SOUL models, these features are added after the n-best list generation step.

In previous work (Och et al., 2004; Hasan, 2011), the IBM1 features (Brown et al., 1993) are found helpful. As the IBM1 model is asymmetric, two models are estimated, one in both directions. Contrary to the reported results, these additional features do not yield significant improvements over the baseline system. We assume that the difficulty is to add information to an already extensively optimized system. Moreover, the IBM1 models are estimated on the same training corpora as the translation system, a fact that may explain the redundancy of these additional features.

In a separate series of experiments, we also add WSD features calculated according to a variation of the method proposed in (Apidianaki, 2009). For each word of a subset of the input (source language) vocabulary, a simple WSD classifier produces a probability distribution over a set of translations[8]. During reranking, each translation hypothesis is scanned and the word translations that match one of the proposed variant are rewarded using an additional score. While this method had given some small gains on a smaller dataset (IWSLT'11), we did not observe here any improvement over the baseline system. Additional analysis hints that (i) most of the proposed variants are already covered by the translation model with high probabilities and (ii) that these variants are seldom found in the reference sentences. This means that, in the situation in which only one reference is provided, the hypotheses with a high score for the WSD feature are not adequately rewarded with the actual references.

[8]The difference with the method described in (Apidianaki, 2009) is that no sense clustering is performed, and each translation is represented by a separate weighted source feature vector which is used for disambiguation

8 Conclusion

In this paper, we described our submissions to WMT'12 in the French-English and German-English shared translation tasks, in both directions. As for our last year's participation, our main systems are built with n-code, the open source Statistical Machine Translation system based on bilingual n-grams. Our contributions are threefold. First, we have experimented a new kind of translation models, where the bilingual n-gram distribution are estimated in a continuous space with neural networks. As shown in past evaluations with target language model, there is a significant reward for using this kind of models in a rescoring step. We observed that, in general, the continuous space translation model yields a slightly larger improvement than the target translation model. However, their combination does not result in an additional gain.

We also reported preliminary results with a system "on-the-fly", where the training data are sampled according to the data to be translated in order to train contextually adapted system. While this system achieves comparable performance to our baseline system, it is worth noticing that its total training time is much smaller than a comparable Moses system. Finally, we investigated several additional features based on IBM1 models and word sense disambiguation information in rescoring. While these methods have sometimes been reported to help improve the results, we did not observe any improvement here over the baseline system.

Acknowledgment

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.

Marianna Apidianaki. 2009. Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 77–85, Athens, Greece, March. Association for Computational Linguistics.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 255–262, Ann Arbor, Michigan, June. Association for Computational Linguistics.

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.

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.

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

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.

Hai-Son, Alexandre Allauzen, and François Yvon. 2012. 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*.

Saša Hasan. 2011. *Triplet Lexicon Models for Statistical Machine Translation*. Ph.D. thesis, RWTH Aachen University.

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.

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.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 505–512, Manchester, UK, August. Coling 2008 Organizing Committee.

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, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA,

May 2 - May 7. 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.

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.

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

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI '02: Proceedings of the 25th Annual German Conference on AI*, pages 18–32, London, UK. Springer-Verlag.