LIMSI's Contribution to the WMT'16 Biomedical Translation Task

Julia Ive, François Yvon, Aurélien Max

LIMSI's Contribution to the WMT'16 Biomedical Translation Task

Julia Ive[1,2], **Aurélien Max**[1], **François Yvon**[1]
LIMSI, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France,[1]
Cochrane France, INSERM U1153, 75181 Paris, France[2]
{firstname.lastname}@limsi.fr

Abstract

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

1 Introduction

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

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

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

2 Systems Overview

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

2.1 Additional Parallel Data

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

[1] http://www.translatecochrane.fr/corpus

[2] http://statmt.org/wmt14/medical-task

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 469–476,
Berlin, Germany, August 11-12, 2016.

2.2 Additional Monolingual Data

As additional monolingual data we used the full French dataset provided by the organizers of the WMT'15 translation task.[3]

2.3 Preprocessing and Word Alignment

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

2.4 Language Models

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

2.5 SOUL

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

[3] `http://www.statmt.org/wmt15/translation-task.html`

2.6 Development and Test Sets

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

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

LIMSIDEV and LIMSITEST were used to respectively tune and test our main PBSMT systems. LIMSITEST was further split into LIMSIDEV2 and LIMSITEST2 for SOUL and system combination optimizations. Statistics for these datasets are in Table 1.

2.7 Evaluation Metrics

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

3 Baseline System

3.1 Details of System Building

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

3.2 Experiments and Results

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

Corpus	# Lines	# Tok., en	# Tok., fr
SCIELO-ABS	≈ 8 K	≈ 200 K	≈ 280 K
SCIELO-TITLES	≈ 700	≈ 10 K	≈ 14 K
MEDLINE-TITLES	≈ 600 K	≈ 7 M	≈ 8 M
COCHRANE-ABS	≈ 140 K	≈ 3 M	≈ 5 M
COCHRANE-TITLES	≈ 8 K	≈ 90 K	≈ 130 K
IN-DOMAIN-DATA	≈800 K	≈ 10 M	≈ 13 M
COPPA	≈ 454 K	≈ 10 M	≈ 12 M
EMEA	≈ 324 K	≈ 6 M	≈ 7 M
PATTR-ABS	≈ 635 K	≈ 20 M	≈ 24 M
PATTR-CLAIMS	≈ 889 K	≈ 32 M	≈ 36 M
PATTR-TITLES	≈ 386 K	≈ 3 M	≈ 4 M
UMLS	≈ 2 M	≈ 8 M	≈ 8 M
WIKIPEDIA	≈8 K	≈ 17 K	≈ 19 K
WMT'14 medical task	≈6 M	≈ 160 M	≈ 190 M
WMT'15 translation task			≈ 2.2 B

Set	# Lines	# Tok., en	# Tok., fr
LIMSIDEV-TITLES	100	1360	1834
LIMSIDEV-ABS	900	24367	30560
LIMSIDEV	1000	25727	32394
LIMSIDEV2-TITLES	50	686	943
LIMSIDEV2-ABS	450	13116	16261
LIMSIDEV2	500	13802	17204
LIMSITEST2-TITLES	50	738	915
LIMSITEST2-ABS	450	12487	15276
LIMSITEST2	500	13225	16191
LIMSITEST	1000	27027	33395

Table 1: Corpora used for training (left); development and test (right)

translation model scores (Le et al., 2012a). The baseline 300-best list was reranked according to the combination of all baseline features and the SOUL features. Reranking allowed us to obtain an improvement of +1.17 BLEU over our baseline system. The system tends to perform better on the LIMSITEST2-TITLES part than on the LIMSITEST2-ABS part. In the rest of this article, we focus our efforts on improving the translation quality of abstracts only.

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

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

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

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

4.1 Methodology

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

Building a confusion network implies several decisions:

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

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

Figure 2 illustrates the hypothesis alignment

system	LIMSITEST2	LIMSITEST2-TITLES	LIMSITEST2-ABS
MOSES	30.38	**49.42**	**29.20**
MOSES + SOUL	**31.55**	50.44	30.27

Table 2: Results (BLEU) for MOSES and MOSES + SOUL on the in-house test set

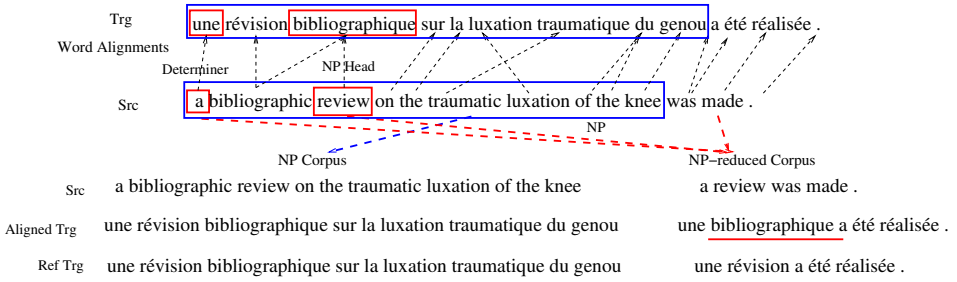

Figure 1: Extraction of NP and NP-reduced instances

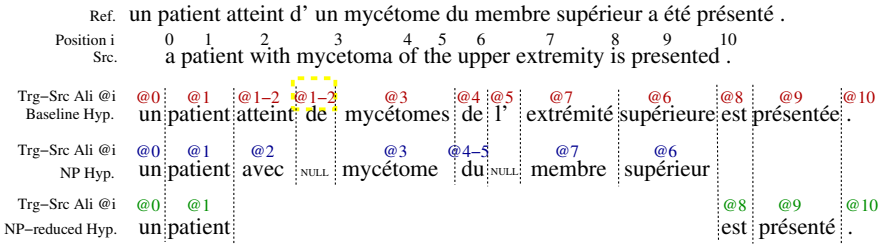

Figure 2: Source-based hypothesis alignment. @i denotes to the source index of a given target word.

procedure. We used source-target phrasal alignments produced by the decoder to assign unaligned words to several positions (see e.g. "de" highlighted in yellow on Figure 2). It may also happen that a target phrase in a partial hypothesis is longer than the corresponding baseline translation; in this case the backbone is extended as needed with `NULL` arcs.

3. Arc scores. Each arc labelled u receives a score equal to the posterior unigram probability $P(u|\varepsilon)$ of the system generating u at this position. $P(u|\varepsilon)$ is computed as in (de Gispert et al., 2013):

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

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

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

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

Each path is the CN is finally scored as follows:

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

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

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

4.2 Details of System Building

We used the SCIELO-ABS and COCHRANE-ABS corpora, as well as LIMSIDEV and LIMSITEST to create the `NP-SMT` and `NP-Reduced-SMT` training, development and test data. The `NP-SMT` source data contained the NPs extracted from the source side of all bitexts. The `NP-Reduced-SMT` target data contained the original source sentences with the NPs replaced by their heads (also preserving the associated article or possessive determiner) (Klein and Manning, 2003; Toutanova et al., 2003). The `NP-SMT` and `NP-Reduced-SMT` target data were created using the translations of the corresponding syntactic structures obtained from the `fast_align`

source-target word alignments, where the non-aligned words were not considered. The NP-SMT training corpus was enriched with the titles and glossary corpora data (SCIELO-TITLES, COCHRANE-TITLES, MEDLINE-TITLES, PATTR-TITLES, UMLS, WIKIPEDIA).

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

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

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

system	BLEU
NP-SMT	**27.47**
NP-SMT + SOUL	28.46
base MOSES (NPs)	**27.25**
base MOSES + SOUL (NPs)	28.33
NP-Reduced-SMT	**22.81**
NP-Reduced-SMT + SOUL	23.53
base MOSES (NP-reduced)	**23.16**
base MOSES + SOUL (NP-reduced)	24.04

Table 3: NP-SMT and NP-Reduced-SMT performance for LIMSITEST2-ABS.

4.3 Experiments and Results

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

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

n-gram	%
1-gram	11
2-gram	28
3-gram	39
4-gram	48

Table 4: Average % of new unique n-gram per CN-DECODING hypothesis (using 300-best lists) LIMSITEST2-ABS).

Results in Table 5 confirm that CN-DECODING is superior here to MOSES xml-mode (+2.06 BLEU for LIMSITEST2-ABS).

test set	MOSES base	MOSES + xml	CN-DECODING
LIMSIDEV2-ABS	32.38	29.59	32.84
LIMSITEST2-ABS	29.20	**26.79**	**28.85**

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

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

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

In this condition, the quality of CN-DECODING output reranked by SOUL shows an insignificant gain over the baseline MOSES + SOUL (+0.18 BLEU for LIMSITEST2-ABS, see Tables 6, 2). It seems that the CN-DECODING procedure allowed our system to locally choose "good" translation variants, in spite of the quality decrease that we observed for NP-Reduced-SMT hypotheses (see

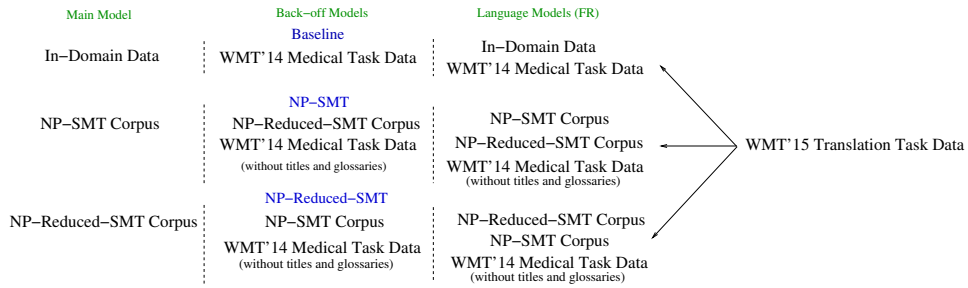

Figure 3: Data used in systems building

Table 3).

test set	LIMSITEST2-ABS
CN-DECODING	30.01
CN-DECODING + SOUL	**30.45**

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

4.4 Observations and Further Improvements

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

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

5 Conclusions

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

In our experiments, we developed an approach aimed at mitigating the syntactic complexity which is a characteristic of a medical scientific publications. Our solution exploits the potential of phrase-based Statistical Machine Translation. We combined the output of the PBSMT system, trained to translate entire source sentences, with the outputs of specialized PBSMT systems, trained to translate syntactically defined subparts of the source sentence: complex noun phrases on the one hand, simplified sentences on the other hand. The combination was performed using confusion network decoding and showed small improvements over a strong baseline when the output of CN decoding is reranked using SOUL. In our future work, we plan to improve the extraction procedure for the reduced systems, as well as to separately improve their performance. For the `NP-SMT` system, this could be achieve by digging additional resources such as comparable corpora.

Acknowledgments

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

References

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

Eleftherios Avramidis, Maja Popović, and Aljoscha Burchardt. 2015. DFKI's experimental hybrid MT system for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 66–73, Lisbon, Portugal, September.

Stanley F. Chen and Joshua T. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, Santa Cruz, US.

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

Marta Costa-Jussà, Mireia Farrús, José Mariño, and José Fonollosa. 2012. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and Informatics*, 31(2).

Adrià de Gispert, Graeme W. Blackwood, Gonzalo Iglesias, and William Byrne. 2013. N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27(2):85–114.

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

Dun Deng, Nianwen Xue, and Shiman Guo. 2015. Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 1–9, Denver, Colorado, USA, June.

Jinhua Du, Yanjun Ma, and Andy Way. 2009. Source-side context-informed hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

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

Markus Freitag, Matthias Huck, and Hermann Ney. 2014a. Jane: Open source machine translation system combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden, April.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014b. EU-BRIDGE MT: Combined machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 105–113, Baltimore, Maryland, USA, June.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of WMT*, Edinburgh, Scotland.

Sanjika Hewavitharana, Alon Lavie, and Vogel Stephan. 2007. Experiments with a noun-phrase driven statistical machine translation system. In *Conference Proceedings: the 11th Machine Translation Summit*, pages 247–253, Copenhagen, Denmark, September.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 254–261.

Julia Ive, Aurélien Max, François Yvon, and Philippe Ravaud. 2016. Diagnosing high-quality statistical machine translation using traces of post-edition operations. In *Proceedings of the LREC 2016 Workshop: Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, Portorož, Slovenia, May.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July.

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 ACL Demo*, Prague, Czech Republic.

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

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

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

Hongzheng Li, Kai Zhao, Yun Hu, Renfen Zhu, and Yaohong Jin. 2015. A hybrid system for chinese-english patent machine translation. In *Conference Proceedings: the 15th Machine Translation Summit*, pages 52–67, Miami, Florida, USA, November.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400.

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

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, Philadelphia, US.

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

Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186, Columbus, Ohio, June.

Antti-Veikko Rosti, Xiaodong He, Damianos Karakos, Gregor Leusch, Yuan Cao, Markus Freitag, Spyros Matsoukas, Hermann Ney, Jason Smith, and Bing Zhang. 2012. Review of hypothesis alignment algorithms for mt system combination via confusion network decoding. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 191–199, Montréal, Canada, June.

Andreas 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.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Junguo Zhu, Muyun Yang, Sheng Li, and Tiejun Zhao. 2013. Repairing incorrect translation with examples. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 967–971, Nagoya, Japan, October. Asian Federation of Natural Language Processing.