LIMSI@WMT'16: Machine Translation of News

Alexandre Allauzen[1], **Lauriane Aufrant**[1,2], **Franck Burlot**[1], **Elena Knyazeva**[1], **Ophélie Lacroix**[1], **Thomas Lavergne**[1], **Guillaume Wisniewski**[1], **François Yvon**[1]
[1]LIMSI, CNRS, Univ. Paris-Sud, Université Paris Saclay, 91 403 Orsay, France
[2] DGA, 60 boulevard du Général Martial Valin, 75 509 Paris, France
firstname.lastname@limsi.fr

Abstract

This paper describes LIMSI's submissions to the shared WMT'16 task "Translation of News". We report results for Romanian-English in both directions, for English to Russian, as well as preliminary experiments on reordering to translate from English into German. Our submissions use mainly NCODE and MOSES along with continuous space models in a post-processing step. The main novelties of this year's participation are the following: for the translation into Russian and Romanian, we have attempted to extend the output of the decoder with morphological variations and to use a CRF model to rescore this new search space; as for the translation into German, we have been experimenting with source-side pre-ordering based on a dependency structure allowing permutations in order to reproduce the target word order.

1 Introduction

This paper documents LIMSI's participation to the shared task of machine translation of news for three language pairs: English to Russian, Romanian-English in both directions and English to German. The reported experiments are mainly related to two challenging domains: inflection prediction and word order in morphologically rich languages.

In our systems translating from English into Romanian and Russian, we have attempted to address the difficulties that go along with translating into morphologically reach languages. First, a baseline system outputs sentences in which we reconsider the choices previously made for inflected words by generating their full paradigm. Second, a CRF model is expected to make better choices than the translation system.

For English to German, experiments are reported on the pre-ordering of the source sentence. Using the dependency structure of the sentence, the model predicts permutations of source words that lead to an order that is as close as possible to the right order in the target language.

2 System Overview

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

2.1 Data pre-processing and word alignments

All the English and Russian data have been cleaned by normalizing character encoding.

Tokenization for English text relies on in-house text processing tools (Déchelotte et al., 2008). For the Russian corpora, we used the `TreeTagger` tokenizer. For Romanian, we developed and used `tokro`, a rule-based tokenizer. After normalization of diacritics, it repeatedly applies 3 rules: (a) word splitting on slashes, except for url addresses, (b) isolation of punctuation characters from a predefined set (including quotes, parentheses and ellipses as triple dots) adjoined at the beginning or end of words (considering a few exceptions like 'Dr.' or 'etc.') and (c) clitic tokenization on hyphens, notably for 'nu', 'dă', 'și' and unstressed personal pronouns. The hyphen is kept on the clitic token. Multi-word expressions are not joined into a single token.

[1] `http://ncode.limsi.fr`
[2] `http://www.statmt.org/moses/`

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 239–245,
Berlin, Germany, August 11-12, 2016.

The parallel corpora were tagged and lemmatized using `TreeTagger` (Schmid, 1994) for English and Russian (Sharoff and Nivre, 2011). The same pre-processing was obtained for Romanian with the TTL tagger and lemmatizer (Tufiş et al., 2008). Having noticed many sentence alignment errors and out-of-domain parts in the Russian common-crawl parallel corpus, we have used a bilingual sentence aligner[3] and proceeded to a domain adaptation filtering using the same procedure as for monolingual data (see section 2.2). As a result, one third of the initial corpus has been removed. Apart from the russian wiki-headlines corpus, the systems presented below used all the parallel data provided by the shared task.

Word alignments were trained according to IBM model 4, using MGIZA.

2.2 Language modelling and domain adaptation

Various English, Romanian and Russian language models (LM) were trained on the in-domain monolingual corpora, a subset of the common-crawl corpora and the relevant side of the parallel corpora (for English, the English side of the Czech-English parallel data was used). We trained 4-gram LMs, pruning all singletons with `lmplz` (Heafield, 2011).

In addition to in-domain monolingual data, a considerable amount of out-of-domain data was provided this year, gathered in the common-crawl corpora. Instead of directly training an LM on these corpora, we extracted from them in-domain sentences using the Moore-Lewis (Moore and Lewis, 2010) filtering method, more specifically its implementation in `XenC` (Rousseau, 2013). As a result, the common-crawl sub-corpora we have used contained about 200M sentences for Romanian and 300M for Russian and English. Finally, we perform a linear interpolation of these models, using the SRILM toolkit (Stolcke, 2002).

2.3 NCODE

NCODE implements the bilingual n-gram approach to SMT (Casacuberta and Vidal, 2004; Crego and Mariño, 2006b; Mariño et al., 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 **f** into a target sentence **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.

$$e^* = \arg\max_{\mathbf{e},\mathbf{a}} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{f}, \mathbf{e}, \mathbf{a})$$

where K feature functions (f_k) are weighted by a set of coefficients (λ_k) and **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 the target word order by unfolding the word alignments (Crego and Mariño, 2006a). 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, 2006a).

2.4 Continuous-space models

Neural networks, working on top of conventional n-gram back-off language models, have been introduced in (Bengio et al., 2003; Schwenk et al., 2006) as a potential means to improve conventional language models. More recently, these techniques have been applied to statistical machine translation in order to estimate continuous-space translation models (CTMs) (Schwenk et al., 2007; Le et al., 2012a; Devlin et al., 2014).

[3] *Bilingual Sentence Aligner*, available at `http://research.microsoft.com/apps/catalog/`

As in our previous participations (Le et al., 2012b; Allauzen et al., 2013; Pécheux et al., 2014; Marie et al., 2015), we take advantage of the proposal of (Le et al., 2012a). Using a specific neural network architecture, the *Structured OUtput Layer* (SOUL), it becomes possible to estimate n-gram models that use large output vocabulary, thereby making the training of large neural network language models feasible both for target language models (Le et al., 2011) 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 $n = 4$). Initialization is an important issue when optimizing neural networks. For CTMs, a solution consists in pre-training monolingual n-gram models. Their parameters are then used to initialize bilingual models.

Given the computational cost of computing n-gram probabilities with neural network models, a solution is to resort to a two-pass approach: the first pass uses a conventional system to produce a k-best list (the k most likely hypotheses); in the second pass, probabilities are computed by continuous-space models for each hypothesis and added as new features. For this year evaluation, we used the following models: one continuous target language model and four CTMs as described in (Le et al., 2012a).

For English to Russian and Romanian to English, the models have the same architecture:

- words are projected into a 500-dimensional vector space;
- the feed-forward architecture includes two hidden layers of size 1000 and 500;
- the non-linearity is a sigmoid function;

All models are trained for 20 epochs, then the selection relies on the perplexity measured on a validation set. For CTMs, the validation sets are sampled from the parallel training data.

3 Experiments

For all our experiments, the MT systems are tuned using the `kb-mira` algorithm (Cherry and Foster, 2012) implemented in MOSES, including the reranking step. POS tagging is performed using the `TreeTagger` (Schmid, 1994) for English and Russian (Sharoff and Nivre, 2011), and TTL (Tufiş et al., 2008) for Romanian.

3.1 Development and test sets

Since only one development set was provided for Romanian, we split the given development set into two equal parts: newsdev-2016/1 and newsdev-2016/2. The first part was used as development set while the second part was our internal test set.

The Russian development and test sets we have used consisted in shuffled sentences from newstest 2012, 2013 and 2014. Tests were also performed on newstest-2015.

3.2 Hidden-CRF for inflection prediction

In morphologically rich languages, each single lemma corresponds to a large number of word forms that are not all observed in the training data. A traditional statistical translation system can not generate a non-observed form. On the other hand, even if a form has been seen at training time, it might be hard to use it in a relevant way if its frequency is low, which is a common phenomena, since the number of singletons in Romanian and Russian corpora is a lot higher than in English corpora. In such a situation, surface heuristics are less reliable.

In order to address this limitation, we tried to extend the output of the decoder with morphological variations of nouns, pronouns and adjectives. Therefore, for each word in the output baring one of these PoS-tags, we introduced all forms in the paradigm as possible alternatives. The paradigm generation was performed for Russian using pymorphy, a dictionary implemented as a Python module.[4] For Romanian, we used the crawled (thus sparse) lexicon introduced in (Aufrant et al., 2016).

Once the outputs were extended, we used a CRF model to rescore this new search space. The CRF can use the features of the MT decoder, but can also include morphological or syntactic features in order to estimate output scores, even for words that were not observed in the training data.

In the Russian experiment, oracle scores show that a maximum gain of 6.3 BLEU points can be obtained if the extension is performed on the full search space and 2.3 BLEU points on 300-best output of the NCODE decoder. The full search space, while being more promising, proved to be too large to be handled by the CRF, so the following experiments were performed on the 300-best output.

[4] `http://pymorphy.readthedocs.io/`

In order to train this model, we split the parallel data in two parts. The first (largest) part was used to train the translation system baseline. The second part was used for the training of the hidden CRF. First, the source side was translated with the baseline system, then the resulting output was extended (paradigm generation). References were obtained by searching for oracle translations in the augmented output. Models were trained using in-house implementation of hidden CRF (Lavergne et al., 2013) and used features from the decoder as well as additional ones: unigram and bigram of words and POS-tags; number, gender and case of the forms and of the surrounding ones; and information about nearest prepositions and verbs.

3.3 Experimental results

The experimental results were not conclusive, as, in the best configuration for Russian our model achieved the same results as the baseline and slightly worsened the NCODE+SOUL system (see Table 1).

System	MOSES	NCODE
Baseline	22.91	23.05
Baseline + SOUL		23.75
Baseline + Hidden-CRF		23.03
Baseline + SOUL + Hidden-CRF		23.46

Table 1: Results (BLEU) for English-Russian with NCODE and MOSES on the official test.

	System	MOSES	NCODE
En-Ro	Baseline	23.98	24.15
	Baseline + Hidden-CRF		23.68
Ro-En	Baseline	30.41	29.90
	Baseline + SOUL		30.60

Table 2: Results (BLEU) for English:Romanian with NCODE and MOSES on the official test.

As for Romanian (Table 2), our model performed worse than for Russian. We assume that this must be partly due to the sparsity of the lexicon used for Romanian, with which we only generated partial paradigms, as opposed to full paradigms for Russian.

3.4 Reordering experiments for English to German

NCODE translates a sentence by first re-ordering the source sentence and then monotonically decoding it. Reorderings of the source sentence are compactly encoded in a permutation lattice generated by iteratively applying POS-based reordering rules extracted from the parallel data.

In this year's WMT evaluation campaign we investigated ways to improve the re-ordering step by re-implementing the approach proposed by (Lerner and Petrov, 2013). This approach aims at taking advantage of the dependency structure of the source sentence to predict a permutation of the source words that is as close as possible to a correct syntactic word order in the target language: starting from the root of the dependency tree a classifier is used to recursively predict the order of a node and all its children. More precisely, for a family[5] of size n, a multiclass classifier is used to select the best ordering of this family among its $n!$ permutations. A different classifier is trained for each possible family size.

Predicting the best re-ordering These experiments were only performed for English to German translation. The source sentences were PoS-tagged and dependency parsed using the MATEPARSER (Bohnet and Nivre, 2012) trained on the UDT v2.0. The parallel source and target sentences were aligned in both directions with FASTALIGN (Dyer et al., 2013) and these alignments were merged with the intersection heuristic.[6]

The training set used to learn the classifiers is generated as follows: during a depth-first traversal of each source sentence, an example is extracted from a node if each child of this node is aligned with exactly one word in the target sentence. In this case, it is possible, by following the alignment links, to extract the order of the family members in the target language. An example is therefore a permutation of n members (1 head and its $n-1$ children).

In practice, we did not extract training examples from families having more than 8 members[7] and train 7 classifiers (one binary classifier for the family made of a head and a single dependent and 6 multi-class classifiers able to discriminate between up to 5 040 classes). Our experiments used

[5]Following (Lerner and Petrov, 2013), we call family a head in a dependency tree and all its children.

[6]Preliminary experiments with the gdfa heuristic showed that the symmetrization heuristic has no impact on the quality of the predicted pre-ordering.

[7]Families with more than 8 members account for less than 0.5% of the extracted examples.

VOWPAL WABBIT, a very efficient implementation of the logistic regression capable to handle a large number of output classes.[8] The features used for training are the same as those proposed by (Lerner and Petrov, 2013): word forms, PoS-tags, relative positions of the head, children, their siblings and the gaps between them, etc.

Building permutation lattices In order to mitigate the impact of errouneously predicted word preorderings, we propose to build lattices of permutations rather than using just one reordering of the source sentence. This lattice includes the two best predicted permutations *at each node*.

It is built as follows: starting from an automaton with a single arc between the initial state and the final state labeled with the ROOT token, each arc is successively substituted by two automata describing two possible re-orderings of the token t corresponding to this arc label and its children in the dependency tree. Each of these automata has $n + 1$ arcs corresponding to the n children of t in the dependency tree and t itself that appear in the predicted order. The weight of the first arc is defined as the probability predicted by the classifier; all other arcs have a weight of 0.

MT experiments We report preliminary results for pre-ordering. All the source side of training data is reordered using the method described above. Then, the reordered source side, along with the target side, are considered as the new parallel training data on which a new NCODE system is trained (including new word alignment, tuple extraction, ...). For tuning and test steps, the learned classifiers are used to generate a permutation lattice that will be decoded.

In the following experiments, we use only news-commentary and Europarl datasets as parallel training data; the development and test sets are, respectively, newstest-2014 and newstest-2015.

These preliminary experiments show a significant decrease in BLEU score which deserves closer investigations. This performance drop is more important when more reordering paths ("2-best" in Table 3) are proposed to the MT system. A similar trend was also observed when using a dependency-based model only to predict the reordering lattices for a system trained on raw data and without the pre-ordering step.

As shown in Table 4, in a large majority of cases

[8] http://hunch.net/~vw/.

Baseline system		
	dev	test
rule-based	19.4	18.5
Dependency-based pre-ordering		
	dev	test
1-best	18.5	17.7
2-best	18.2	17.2

Table 3: Translation results for pre-ordering on the English to German translation task

the members of a family have the same order in the source and in the target languages, a trend that is probably amplified by our instance extraction strategy. Dealing with skewed classes is a challenging problem in machine learning and it is not surprising that the performance of the classifier is rather low for the minority classes (see results in Table 4). It is interesting to note that the standard rule-based approach does not suffer from the class imbalance problem as all re-orderings observed in the training data are considered without taking into account their probability.

4 Discussion and Conclusion

This paper described LIMSI's submission to the shared WMT'16 task "Translation of News". We reported results for English-Romanian in both directions and for English into Russian, as well as English into German for which we have investigated pre-ordering of the source sentence. Our submissions used NCODE and MOSES along with continuous space translation models in a post-processing step. Most of our efforts this year were dedicated to the main difficulties of morphologically rich languages: word order and inflection prediction.

For the translation from English into Romanian and Russian, the generation of the paradigm of inflectional words and choice of the right word form using a CRF did not give any improvement over the baseline in our experimental conditions. The reason may be due to the fact that we did not only expect that the CRF would make a better choice than the baseline system regarding word inflection, we also assumed that these morphological predictions would help to make right decisions regarding lexical choices and word order. This was our motivation to run such a decoding extension

size	% mono.	prec.	prec. mono.	prec. non-mono.
2	85.6	88.2	97.6	31.5
3	71.3	79.0	95.5	37.6
4	62.0	74.3	95.9	38.8
5	51.8	68.4	91.8	43.2
6	41.9	53.4	81.8	32.8
7	46.2	14.7	18.3	11.6
8	25.0	7.5	12.1	6.0

Table 4: % of family that have the same order in English and German (% mono.), overall prediction performance (prec.) as well as precision for monotonic and non-monotonic reordering.

over the n-best hypotheses made by the baseline system: the CRF is then supposed to make decisions that go beyond word inflection, since it returns a single best translation. Presumably, the resulting search space turned out to be too complex for our CRF model to make relevant choices. We plan in the nearest future to address this issue by exploring a way to rely on the CRF for inflection prediction only.

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

5 Acknowledgments

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

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.

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Cross-lingual and supervised models for morphosyntactic annotation: a comparison on Romanian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

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

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

Josep M. Crego and José B. Mariño. 2006a. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20.

Joseph M. Crego and José B. Mariño. 2006b. Improving statistical mt by coupling reordering and decoding. *Machine translation*, 20(3):199–215, Jul.

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

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

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland, June. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameteri-

zation of ibm model 2. In *Proceedings of NAACL*, Atlanta, Georgia.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL Demo*, Prague, Czech Republic.

Thomas Lavergne, Alexandre Allauzen, and François Yvon. 2013. A fully discriminative training framework for statistical machine translation (un cadre d'apprentissage intégralement discriminant pour la traduction statistique) [in french]). In *Proceedings of TALN 2013 (Volume 1: Long Papers)*, pages 450–463, Les Sables d'Olonne, France, June. ATALA.

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.

Uri Lerner and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 513–523, Seattle, Washington, USA, October. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, (100):73–82.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of ICNMLP*, Manchester, England.

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

Holger Schwenk, Marta R. Costa-jussa, and Jose A. R. Fonollosa. 2007. Smooth bilingual n-gram translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 430–438, Prague, Czech Republic, June. Association for Computational Linguistics.

Serge Sharoff and Joakim Nivre. 2011. The proper place of men and machines in language technology processing russian without any linguistic knowledge. In *Russian Conference on Computational Linguistics*.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Tufiş, Radu Ion, Ru Ceauşu, and Dan Ştefănescu. 2008. Racai's linguistic web services. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. In *25th German Conf. on Artificial Intelligence (KI2002)*, pages 18–32, Aachen, Germany, September. Springer Verlag.