Enriching Morphologically Poor Languages for Statistical Machine Translation

Eleftherios Avramidis
e.avramidis@sms.ed.ac.uk

Philipp Koehn
pkoehn@inf.ed.ac.uk

School of Informatics
University of Edinburgh
2 Baccleuch Place
Edinburgh, EH8 9LW, UK

Abstract

We address the problem of translating from morphologically poor to morphologically rich languages by adding per-word linguistic information to the source language. We use the syntax of the source sentence to extract information for noun cases and verb persons and annotate the corresponding words accordingly. In experiments, we show improved performance for translating from English into Greek and Czech. For English–Greek, we reduce the error on the verb conjugation from 19% to 5.4% and noun case agreement from 9% to 6%.

1 Introduction

Traditional statistical machine translation methods are based on mapping on the lexical level, which takes place in a local window of a few words. Hence, they fail to produce adequate output in many cases where more complex linguistic phenomena play a role. Take the example of morphology. Predicting the correct morphological variant for a target word may not depend solely on the source words, but require additional information about its role in the sentence.

Recent research on handling rich morphology has largely focused on translating from rich morphology languages, such as Arabic, into English (Habash and Sadat, 2006). There has been less work on the opposite case, translating from English into morphologically richer languages. In a study of translation quality for languages in the Europarl corpus, Koehn (2005) reports that translating into morphologically richer languages is more difficult than translating from them.

There are intuitive reasons why generating richer morphology from morphologically poor languages is harder. Take the example of translating noun phrases from English to Greek (or German, Czech, etc.). In English, a noun phrase is rendered the same if it is the subject or the object. However, Greek words in noun phrases are inflected based on their role in the sentence. A purely lexical mapping of English noun phrases to Greek noun phrases suffers from the lack of information about its role in the sentence, making it hard to choose the right inflected forms.

Our method is based on factored phrase-based statistical machine translation models. We focused on preprocessing the source data to acquire the needed information and then use it within the models. We mainly carried out experiments on English to Greek translation, a language pair that exemplifies the problems of translating from a morphologically poor to a morphologically rich language.

1.1 Morphology in Phrase-based SMT

When examining parallel sentences of such language pairs, it is apparent that for many English words and phrases which appear usually in the same form, the corresponding terms of the richer target language appear inflected in many different ways. On a single word-based probabilistic level, it is then obvious that for one specific English word e the probability $p(f|e)$ of it being translated into a word f decreases as the number of translation candidates increase, making the decisions more uncertain.

Proceedings of ACL-08: HLT, pages 763–770,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

- ***English***: **The president, after reading the press review and the announcements, left his office**
- ***Greek***-1: **The president**[nominative], **after reading**[3^{rd}sing] **the press review**[accusative,sing] **and the announcements**[accusative,plur], **left**[3^{rd}sing] **his office**[accusative,sing]
- ***Greek***-2: **The president**[nominative], **after reading**[3^{rd}sing] **the press review**[accusative,sing] **and the announcements**[nominative,plur], **left**[3^{rd}plur] **his office**[accusative,sing]

Figure 1: Example of missing agreement information, affecting the meaning of the second sentence

One of the main aspects required for the fluency of a sentence is *agreement*. Certain words have to match in gender, case, number, person etc. within a sentence. The exact rules of agreement are language-dependent and are closely linked to the morphological structure of the language.

Traditional statistical machine translation models deal with this problems in two ways:

- The basic SMT approach uses the target language model as a feature in the argument maximisation function. This language model is trained on grammatically correct text, and would therefore give a good probability for word sequences that are likely to occur in a sentence, while it would penalise ungrammatical or badly ordered formations.
- Meanwhile, in phrase-based SMT models, words are mapped in chunks. This can resolve phenomena where the English side uses more than one words to describe what is denoted on the target side by one morphologically inflected term.

Thus, with respect to these methods, there is a problem when agreement needs to be applied on part of a sentence whose length exceeds the order of the of the target *n*-gram language model and the size of the chunks that are translated (see Figure 1 for an example).

1.2 Related Work

In one of the first efforts to enrich the source in word-based SMT, Ueffing and Ney (2003) used part-of-speech (POS) tags, in order to deal with the verb conjugation of Spanish and Catalan; so, POS tags were used to identify the pronoun+verb sequence and splice these two words into one term. The approach was clearly motivated by the problems occurring by a single-word-based SMT and have been solved by adopting a phrase-based model. Meanwhile, there is no handling of the case when the pronoun stays in distance with the related verb.

Minkov et al. (2007) suggested a post-processing system which uses morphological and syntactic features, in order to ensure grammatical agreement on the output. The method, using various grammatical source-side features, achieved higher accuracy when applied directly to the reference translations but it was not tested as a part of an MT system. Similarly, translating English into Turkish (Durgar El-Kahlout and Oflazer, 2006) uses POS and morph stems in the input along with rich Turkish morph tags on the target side, but improvement was gained only after augmenting the generation process with morphotactical knowledge. Habash et al. (2007) also investigated case determination in Arabic. Carpuat and Wu (2007) approached the issue as a Word Sense Disambiguation problem.

In their presentation of the factored SMT models, Koehn and Hoang (2007) describe experiments for translating from English to German, Spanish and Czech, using morphology tags added on the morphologically rich side, along with POS tags. The morphological factors are added on the morphologically rich side and scored with a 7-gram sequence model. Probabilistic models for using only source tags were investigated by Birch et al. (2007), who attached syntax hints in factored SMT models by having *Combinatorial Categorial Grammar* (CCG) *supertags* as factors on the input words, but in this case English was the target language.

This paper reports work that strictly focuses on translation from English to a morphologically richer language. We go one step further than just using easily acquired information (e.g. English POS or lemmata) and extract target-specific information from the source sentence context. We use syntax, not in

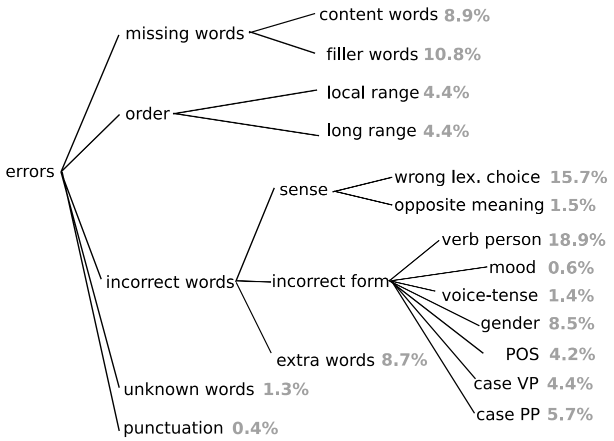

Figure 2: Classification of the errors on our English-Greek baseline system (ch. 4.1), as suggested by Vilar et al. (2006)

order to aid reordering (Yamada and Knight, 2001; Collins et al., 2005; Huang et al., 2006), but as a means for getting the "missing" morphology information, depending on the syntactic position of the words of interest. Then, contrary to the methods that added only output features or altered the generation procedure, we used this information in order to augment only the source side of a factored translation model, assuming that we do not have resources allowing factors or specialized generation in the target language (a common problem, when translating from English into under-resourced languages).

2 Methods for enriching input

We selected to focus on *noun cases agreement* and *verb person conjugation*, since they were the most frequent grammatical errors of our baseline SMT system (see full error analysis in Figure 2). Moreover, these types of inflection signify the constituents of every phrase, tightly linked to the meaning of the sentence.

2.1 Case agreement

The case agreement for nouns, adjectives and articles is mainly defined by the syntactic role that each noun phrase has. Nominative case is used to define the nouns which are the subject of the sentence, accusative shows usually the direct object of the verbs and dative case refers to the indirect object of bitransitive verbs.

Therefore, the followed approach takes advantage of syntax, following a method similar to *Semantic Role Labelling* (Carreras and Marquez, 2005; Surdeanu and Turmo, 2005). English, as morphologically poor language, usually follows a fixed word order (subject-verb-object), so that a syntax parser can be easily used for identifying the subject and the object of most sentences. Considering such annotation, a factored translation model is trained to map the word-case pair to the correct inflection of the target noun. Given the agreement restriction, all words that accompany the noun (adjectives, articles, determiners) must follow the case of the noun, so their likely case needs to be identified as well.

For this purpose we use a syntax parser to acquire the syntax tree for each English sentence. The trees are parsed depth-first and the cases are identified within particular "sub-tree patterns" which are manually specified. We use the sequence of the nodes in the tree to identify the syntactic role of each noun phrase.

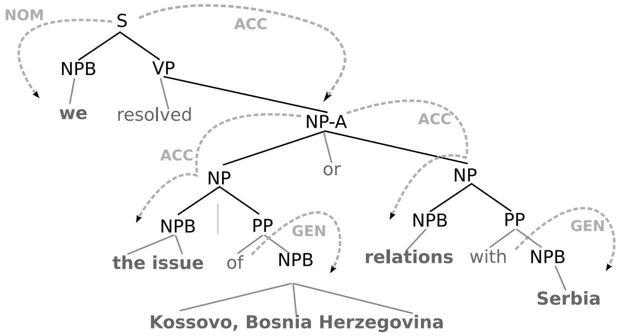

Figure 3: Case tags are assigned on depth-first parse of the English syntax tree, based on sub-tree patterns

To make things more clear, an example can be seen in figure 3. At first, the algorithm identifies the subtree "S-(NPB-VP)" and the *nominative* tag is applied on the NPB node, so that it is assigned to the word "we" (since a pronoun can have a case). The example of accusative shows how cases get transferred to nested subtrees. In practice, they are recursively transferred to every underlying noun phrase (NP) but not to clauses that do not need this information (e.g. prepositional phrases). Similar rules are applied for covering a wide range of node sequence patterns.

Also note that this method had to be targetoriented in some sense: we considered the target language rules for choosing the noun case in every prepositional phrase, depending on the leading preposition. This way, almost all nouns were tagged and therefore the number of the factored words was increased, in an effort to decrease sparsity. Similarly, cases which do not actively affect morphology (e.g. dative in Greek) were not tagged during factorization.

2.2 Verb person conjugation

For resolving the verb conjugation, we needed to identify the person of a verb and add this piece of linguistic information as a tag. As we parse the tree top-down, on every level, we look for two discrete nodes which, somewhere in their children, include the verb and the corresponding subject. Consequently, the node which contains the subject is searched recursively until a subject is found. Then, the person is identified and the tag is assigned to the node which contains the verb, which recursively bequeaths this tag to the nested subtree.

For the subject selection, the following rules were applied:

- The verb person is directly connected to the subject of the sentence and in most cases it is directly inferred by a personal pronoun (I, you etc). Therefore, since this is usually the case, when a pronoun existed, it was directly used as a tag.
- All pronouns in a different case (e.g. *them, myself*) were were converted into nominative case before being used as a tag.
- When the subject of the sentence is not a pronoun, but a single noun, then it is in third person. The POS tag of this noun is then used to identify if it is plural or singular. This was selectively modified for nouns which despite being in singular, take a verb in plural.
- The gender of the subject does not affect the inflection of the verb in Greek. Therefore, all three genders that are given by the third person pronouns were reduced to one.

In Figure 4 we can see an example of how the person tag is extracted from the subject of the sentence and gets passed to the relative clause. In particular, as the algorithm parses the syntax tree, it identifies the sub-tree which has NP-A as a head and includes the WHNP node. Consequently, it recursively browses the preceding NPB so as to get the subject of the sentence. The word "aspects" is found, which has a POS tag that shows it is a plural noun. Therefore, we consider the subject to be of the third person in plural (tagged by *they*) which is recursively passed to the children of the head node.

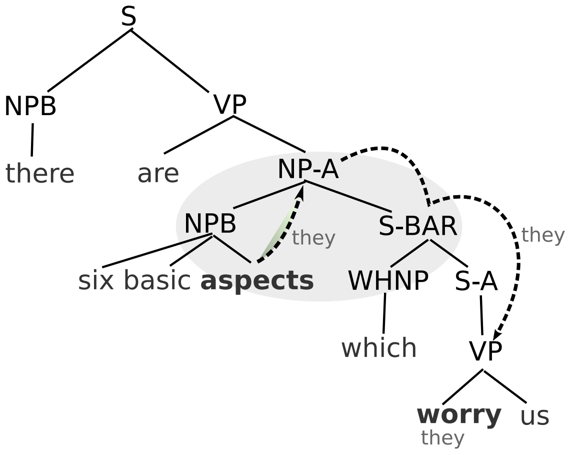

Figure 4: Applying person tags on an English syntax tree

3 Factored Model

The factored statistical machine translation model uses a log-linear approach, in order to combine the several components, including the language model, the reordering model, the translation models and the generation models. The model is defined mathematically (Koehn and Hoang, 2007) as following:

$$p(\mathbf{f}|\mathbf{e}) = \frac{1}{Z}\exp\sum_{i=1}^{n}\lambda_i h_i(\mathbf{f},\mathbf{e}) \quad (1)$$

where λ_i is a vector of weights determined during a tuning process, and h_i is the feature function. The feature function for a translation probability distribution is

$$h_T(\mathbf{f}|\mathbf{e}) = \sum_j \tau(\bar{e}_j, \bar{f}_j) \quad (2)$$

While factored models may use a generation step to combine the several translation components based on the output factors, we use only source factors;

therefore we don't need a generation step to combine the probabilities of the several components.

Instead, factors are added so that both words and its factor(s) are assigned the same probability. Of course, when there is not 1-1 mapping between the *word+factor* splice on the source and the inflected word on the target, the well-known issue of sparse data arises. In order to reduce these problems, decoding needed to consider alternative paths to translation tables trained with less or no factors (as Birch et al. (2007) suggested), so as to cover instances where a word appears with a factor which it has not been trained with. This is similar to back-off. The alternative paths are combined as following (fig. 5):

$$h_T(\mathbf{f}|\mathbf{e}) = \sum_j h_{T_{t(j)}}(\overline{e}_j, \overline{f}_j) \qquad (3)$$

where each phrase j is translated by one translation table $t(j)$ and each table i has a feature function h_{T_i}. as shown in eq. (2).

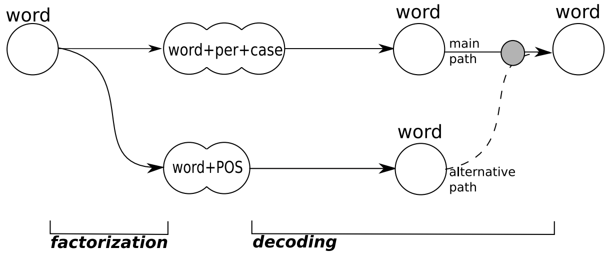

Figure 5: Decoding using an alternative path with different factorization

4 Experiments

This preprocessing led to annotated source data, which were given as an input to a factored SMT system.

4.1 Experiment setup

For testing the factored translation systems, we used Moses (Koehn et al., 2007), along with a 5-gram SRILM language model (Stolcke, 2002). A Greek model was trained on 440,082 aligned sentences of Europarl v.3, tuned with *Minimum Error Training* (Och, 2003). It was tuned over a development set of 2,000 Europarl sentences and tested on two sets of 2,000 sentences each, from the Europarl and a News Commentary respectively, following the specifications made by the ACL 2007 2nd Workshop on SMT[1]. A Czech model was trained on 57,464 aligned sentences, tuned over 1057 sentences of the News Commentary corpus and and tested on two sets of 964 sentences and 2000 sentences respectively.

The training sentences were trimmed to a length of 60 words for reducing perplexity and a standard lexicalised reordering, with distortion limit set to 6. For getting the syntax trees, the latest version of Collins' parser (Collins, 1997) was used. When needed, part-of-speech (POS) tags were acquired by using Brill's tagger (Brill, 1992) on v1.14. Results were evaluated with both BLEU (Papineni et al., 2001) and NIST metrics (NIST, 2002).

4.2 Results

	BLEU		NIST	
set	devtest	test07	devtest	test07
baseline	18.13	18.05	5.218	5.279
person	18.16	18.17	5.224	5.316
pos+person	18.14	18.16	5.259	5.316
person+case	18.08	18.24	5.258	5.340
*altpath:*POS	18.21	18.20	5.285	5.340

Table 1: Translating English to Greek: Using a single translation table may cause sparse data problems, which are addressed using an alternative path to a second translation table

We tested several various combinations of tags, while using a single translation component. Some combinations seem to be affected by sparse data problems and the best score is achieved by using both person and case tags. Our full method, using both factors, was more effective on the second testset, but the best score in average was succeeded by using an alternative path to a POS-factored translation table (table 1). The NIST metric clearly shows a significant improvement, because it mostly measures difficult n-gram matches (e.g. due to the long-distance rules we have been dealing with).

[1] see http://www.statmt.org/wmt07 referring to sets *dev2006* (tuning) and *devtest2006*, *test2007* (testing)

4.3 Error analysis

In *n*-gram based metrics, the scores for all words are equally weighted, so mistakes on crucial sentence constituents may be penalized the same as errors on redundant or meaningless words (Callison-Burch et al., 2006). We consider agreement on verbs and nouns an important factor for the adequacy of the result, since they adhere more to the semantics of the sentence. Since we targeted these problems, we conducted a manual error analysis focused on the success of the improved system regarding those specific phenomena.

system	verbs	errors	missing
baseline	311	19.0%	7.4%
single	295	4.7%	5.4%
alt.path	294	5.4%	2.7%

Table 2: Error analysis of 100 test sentences, focused on verb person conjugation, for using both person and case tags

system	NPs	errors	missing
baseline	469	9.0%	4.9%
single	465	6.2%	4.5%
alt. path	452	6.0%	4.0%

Table 3: Error analysis of 100 test sentences, focused on noun cases, for using both person and case tags

The analysis shows that using a system with only one phrase translation table caused a high percentage of missing or untranslated words. When a word appears with a tag with which it has not been trained, that would be considered an unseen event and remain untranslated. The use of the alternative path seems to be a good solution.

step	parsing	tagging	decoding
VPs	16.7%	25%	58.3%
NPs	39.2%	21.7%	39.1%
avg	31.4%	22.9%	45.7 %

Table 4: Analysis on which step of the translation process the agreement errors derive from, based on manual resolution on the errors of table 3

The impact of the preprocessing stage to the errors may be seen in table 4, where errors are tracked back to the stage they derived from. Apart from the decoding errors, which may be attributed to sparse data or other statistical factors, a large part of the errors derive from the preprocessing step; either the syntax tree of the sentence was incorrectly or partially resolved, or our labelling process did not correctly match all possible sub-trees.

4.4 Investigating applicability to other inflected languages

The grammatical phenomena of noun cases and verb persons are quite common among many human languages. While the method was tested in Greek, there was an effort to investigate whether it is useful for other languages with similar characteristics. For this reason, the method was adapted for Czech, which needs agreement on both verb conjugation and 9 noun cases. Dative case was included for the indirect object and the rules of the prepositional phrases were adapted to tag all three cases that can be verb phrase constituents. The Czech noun cases which appear only in prepositional phrases were ignored, since they are covered by the phrase-based model.

	BLUE		NIST	
set	devtest	test	devtest	test
baseline	12.08	12.34	4.634	4.865
person+case *altpath*:POS	11.98	11.99	4.584	4.801
person *altpath*:word	12.23	12.11	4.647	4.846
case *altpath*:word	12.54	12.51	4.758	4.957

Table 5: Enriching source data can be useful when translating from English to Czech, since it is a morphologically rich language. Experiments shown improvement when using factors on noun-cases with an alternative path

In Czech, due to the small size of the corpus, it was possible to improve metric scores only by using an alternative path to a bare word-to-word translation table. Combining case and verb tags worsened the results, which suggests that, while applying the method to more languages, a different use of the attributes may be beneficial for each of them.

5 Conclusion

In this paper we have shown how SMT performance can be improved, when translating from English into morphologically richer languages, by adding linguistic information on the source. Although the source language misses morphology attributes required by the target language, the needed information is inherent in the syntactic structure of the source sentence. Therefore, we have shown that this information can be easily be included in a SMT model by preprocessing the source text.

Our method focuses on two linguistic phenomena which produce common errors on the output and are important constituents of the sentence. In particular, noun cases and verb persons are required by the target language, but not directly inferred by the source. For each of the sub-problems, our algorithm used heuristic syntax-based rules on the statistically generated syntax tree of each sentence, in order to address the missing information, which was consequently tagged in by means of word factors. This information was proven to improve the outcome of a factored SMT model, by reducing the grammatical agreement errors on the generated sentences.

An initial system using one translation table with additional source side factors caused sparse data problems, due to the increased number of unseen word-factor combinations. Therefore, the decoding process is given an *alternative path* towards a translation table with less or no factors.

The method was tested on translating from English into two morphologically rich languages. Note that this may be easily expanded for translating from English into many morphologically richer languages with similar attributes. Opposed to other factored translation model approaches that require target language factors, that are not easily obtainable for many languages, our approach only requires English syntax trees, which are acquired with widely available automatic parsers. The preprocessing scripts were adapted so that they provide the morphology attributes required by the target language and the best combination of factors and alternative paths was chosen.

Acknowledgments

This work was supported in part under the EuroMatrix project funded by the European Commission (6th Framework Programme). Many thanks to Josh Schroeder for preparing the training, development and test data for Greek, in accordance to the standards of ACL 2007 2nd Workshop on SMT; to Hieu Hoang, Alexandra Birch and all the members of the Edinburgh University SMT group for answering questions, making suggestions and providing support.

References

Birch, A., Osborne, M., and Koehn, P. 2007. CCG Supertags in factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.

Brill, E. 1992. A simple rule-based part of speech tagger. *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155.

Callison-Burch, C., Osborne, M., and Koehn, P. 2006. Re-evaluation the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. The Association for Computer Linguistics.

Carpuat, M. and Wu, D. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Czech Republic.

Carreras, X. and Marquez, L. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of 9th Conference on Computational Natural Language Learning (CoNLL)*, pages 169–172, Ann Arbor, Michigan, USA.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th conference on Association for Computational Linguistics*, pages 16–23.

Collins, M., Koehn, P., and Kučerová, I. 2005. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Morristown, NJ, USA. Association for Computational Linguistics.

Durgar El-Kahlout, i. and Oflazer, K. 2006. Initial explorations in english to turkish statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 7–14, New York City. Association for Computational Linguistics.

Habash, N., Gabbard, R., Rambow, O., Kulick, S., and Marcus, M. 2007. Determining case in Arabic: Learning complex linguistic behavior requires complex linguistic features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1084–1092.

Habash, N. and Sadat, F. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAAC L, Companion Volume: Short Papers*, pages 49–52, New York City, USA. Association for Computational Linguistics.

Huang, L., Knight, K., and Joshi, A. 2006. Statistical syntax-directed translation with extended domain of locality. *Proc. AMTA*, pages 66–73.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit*, 5.

Koehn, P. and Hoang, H. 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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 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. Association for Computational Linguistics.

Minkov, E., Toutanova, K., and Suzuki, H. 2007. Generating complex morphology for machine translation. In *ACL 07: Proceedings of the 45th Annual Meeting of the Association of Computational linguistics*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

NIST 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2001. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Stolcke, A. 2002. SRILM-an extensible language modeling toolkit. *Proc. ICSLP*, 2:901–904.

Surdeanu, M. and Turmo, J. 2005. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of 9th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–224, Ann Arbor, Michigan, USA.

Ueffing, N. and Ney, H. 2003. Using pos information for statistical machine translation into morphologically rich languages. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 347–354, Morristown, NJ, USA. Association for Computational Linguistics.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy.

Yamada, K. and Knight, K. 2001. A syntax-based statistical translation model. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.