Anaphora Models and Reordering for Phrase-Based SMT

Christian Hardmeier Sara Stymne Jörg Tiedemann Aaron Smith Joakim Nivre
Uppsala University
Department of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

We describe the Uppsala University systems for WMT14. We look at the integration of a model for translating pronominal anaphora and a syntactic dependency projection model for English–French. Furthermore, we investigate post-ordering and tunable POS distortion models for English–German.

1 Introduction

In this paper we describe the Uppsala University systems for WMT14. We present three different systems. Two of them are based on the document-level decoder Docent (Hardmeier et al., 2012; Hardmeier et al., 2013a). In our English–French system we extend Docent to handle pronoun anaphora, and in our English–German system we add part-of-speech phrase-distortion models to Docent. For German–English we also have a system based on Moses (Koehn et al., 2007). Again the focus is on word order, this time by using pre- and post-reordering.

2 Document-Level Decoding

Traditional SMT decoders translate texts as bags of sentences, assuming independence between sentences. This assumption allows efficient algorithms for exploring a large search space based on dynamic programming (Och et al., 2001). Because of the dynamic programming assumptions it is hard to directly include discourse-level and long-distance features into a traditional SMT decoder.

In contrast to this very popular stack decoding approach, our decoder Docent (Hardmeier et al., 2012; Hardmeier et al., 2013a) implements a search procedure based on local search. At any stage of the search process, its search state consists of a complete document translation, making it easy for feature models to access the complete document with its current translation at any point in time. The search algorithm is a stochastic variant of standard hill climbing. At each step, it generates a successor of the current search state by randomly applying one of a set of state changing operations to a random location in the document, and accepts the new state if it has a better score than the previous state. The operations are to change the translation of a phrase, to change the word order by swapping the positions of two phrases or moving a sequence of phrases, and to resegment phrases. The initial state can either be initialized randomly, or be based on an initial run from Moses. This setup is not limited by dynamic programming constraints, and enables the use of the full translated target document to extract features.

3 English–French

Our English–French system is a phrase-based SMT system with a combination of two decoders, Moses (Koehn et al., 2007) and Docent (Hardmeier et al., 2013a). The fundamental setup is loosely based on the system submitted by Cho et al. (2013) to the WMT 2013 shared task. Our phrase table is trained on data taken from the News commentary, Europarl, UN, Common crawl and 10^9 corpora. The first three of these corpora were included integrally into the training set after filtering out sentences of more than 80 words. The Common crawl and 10^9 data sets were run through an additional filtering step with an SVM classifier, closely following Mediani et al. (2011). The system includes three language models, a regular 6-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) trained with KenLM (Heafield, 2011), a 4-gram bilingual language model (Niehues et al., 2011) with Kneser-Ney smoothing trained with KenLM and a 9-gram model over Brown clusters (Brown et al., 1992) with Witten-Bell smoothing (Witten and Bell, 1991) trained with SRILM (Stolcke, 2002).

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 122–129,
Baltimore, Maryland USA, June 26–27, 2014.

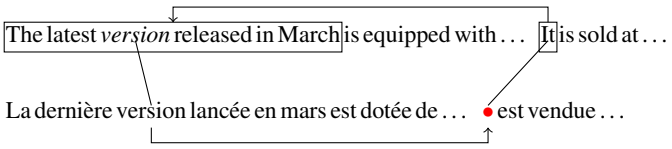

Figure 1: Pronominal Anaphora Model

Our baseline system achieved a cased BLEU score of 33.2 points on the newstest2014 data set. Since the anaphora model used in our submission suffered from a serious bug, we do not discuss the results of the primary submission in more detail.

3.1 Pronominal Anaphora Model

Our pronominal anaphora model is an adaptation of the pronoun prediction model described by Hardmeier et al. (2013b) to SMT. The model consists of a neural network that discriminatively predicts the translation of a source language pronoun from a short list of possible target language pronouns using features from the context of the source language pronouns and from the translations of possibly remote antecedents. The objective of this model is to handle situations like the one depicted in Figure 1, where the correct choice of a target-language pronoun is subject to morphosyntactic agreement with its antecedent. This problem consists of several steps. To score a pronoun, the system must decide if a pronoun is anaphoric and, if so, find potential antecedents. Then, it can predict what pronouns are likely to occur in the translation. Our pronoun prediction model is trained on both tasks jointly, including anaphora resolution as a set of latent variables. At test time, we split the network in two parts. The anaphora resolution part is run separately as a preprocessing step, whereas the pronoun prediction part is integrated into the document-level decoder with two additional feature models.

The features correspond to two copies of the neural network, one to handle the singular pronoun *it* and one to handle the plural pronoun *they*. Each network just predicts a binary distinction between two cases, *il* and *elle* for the singular network and *ils* and *elles* for the plural network. Unlike Hardmeier et al. (2013b), we do not use an OTHER category to capture cases that should not be translated with any of these options. Instead, we treat all other cases in the phrase table and activate the anaphora models only if one of their target pronouns actually occurs in the output.

To achieve this, we generate pronouns in two steps. In the phrase table training corpus, we replace all pronouns that should be handled by the classifier, i.e. instances of *il* and *elle* aligned to *it* and instances of *ils* and *elles* aligned to *they*, with special placeholders. At decoding time, if a placeholder is encountered in a target language phrase, the applicable pronouns are generated with equal translation model probability, and the anaphora model adds a score to discriminate between them.

To reduce the influence of the language model on pronoun choice and give full control to the anaphora model, our primary language model is trained on text containing placeholders instead of pronouns. Since all output pronouns can also be generated without the interaction of the anaphora model if they are not aligned to a source language pronoun, we must make sure that the language model sees training data for both placeholders and actual pronouns. However, for the monolingual training corpora we have no word alignments to decide whether or not to replace a pronoun by a placeholder. To get around this problem, we train a 6-gram placeholder language model on the target language side of the Europarl and News commentary corpora. Then, we use the Viterbi n-gram model decoder of SRILM (Stolcke, 2002) to map pronouns in the entire language model training set to placeholders where appropriate. No substitutions are made in the bilingual language model or the Brown cluster language model.

3.2 Dependency Projection Model

Our English–French system also includes a dependency projection model, which uses source-side dependency structure to model target-side relations between words. This model assigns a score to each dependency arc in the source language by considering the target words aligned to the head and the dependent. In Figure 2, for instance, there is an *nsubjpass* arc connecting *dominated* to *production*. The head is aligned to the target word *dominée*, while the dependent is aligned to the set $\{production, de\}$. The score is computed by a neural network taking as features the head and dependent words and their part-of-speech tags in the source language, the target word sets aligned to the head and dependent, the label of the dependency arc, the distance between the head and dependent word in the source language as well as the shortest distance between any pair of words in the aligned sets. The network is a binary classifier trained to discriminate positive examples extracted from human-made reference

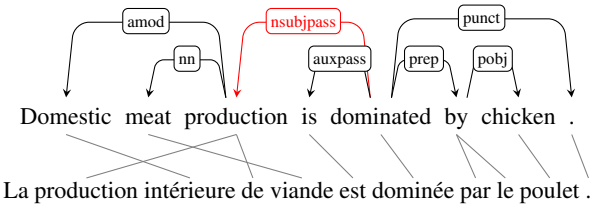

Figure 2: Dependency projection model

translations from negative examples extracted from n-best lists generated by a baseline SMT system.

4 English–German

For English–German we have two systems, one based on Moses, and one based on Docent. In both cases we have focused on word order, particularly for verbs and particles.

Both our systems are trained on the same data made available by WMT. The Common crawl data was filtered using the method of Stymne et al. (2013). We use factored models with POS tags as a second output factor for German. The possibility to use language models for different factors has been added to our Docent decoder. Language models include an in-domain news language model, an out-of-domain model trained on the target side of the parallel training data and a POS language model trained on tagged news data. The LMs are trained in the same way as for English–French. All systems are tuned using MERT (Och, 2003). Phrase-tables are filtered using entropy-based pruning (Johnson et al., 2007) as implemented in Moses. All BLEU scores are given for uncased data.

4.1 Pre-Ordered Alignment and Post-Ordered Translation

The use of syntactic reordering as a separate preprocessing step has already a long tradition in statistical MT. Handcrafted rules (Collins et al., 2005; Popović and Ney, 2006) or data-driven models (Xia and McCord, 2004; Genzel, 2010; Rottmann and Vogel, 2007; Niehues and Kolss, 2009) for *preordering* training data and system input have been explored in numerous publications. For certain language pairs, such as German and English, this method can be very effective and often improves the quality of standard SMT systems significantly. Typically, the source language is reordered to better match the syntax of the target language when translating between languages that exhibit consistent word order differences, which are difficult to handle by SMT systems with limited reordering capabilities such as phrase-based models. Preordering is often done on the entire training data as well to optimize translation models for the pre-ordered input. Less common is the idea of *post-ordering*, which refers to a separate step after translating source language input to an intermediate target language with corrupted (source-language like) word order (Na et al., 2009; Sudoh et al., 2011).

In our experiments, we focus on the translation from English to German. Post-ordering becomes attractive for several reasons: One reason is the common split of verb-particle constructions that can lead to long distance dependencies in German clauses. Phrase-based systems and n-gram language models are not able to handle such relations beyond a certain distance and it is desirable to keep them as connected units in the phrase translation tables. Another reason is the possible distance of finite and infinitival verbs in German verb phrases that can lead to the same problems described above with verb-particle constructions. The auxiliary or modal verb is placed at the second position but the main verb appears at the end of the associated verb phrase. The distances can be arbitrarily long and long-range dependencies are quite frequent. Similarly, negation particles and adverbials move away from the inflected verb forms in certain constructions. For more details on specific phenomena in German, we refer to (Collins et al., 2005; Gojun and Fraser, 2012). Pre-ordering, i.e. moving English words into German word order does not seem to be a good option as we loose the connection between related items when moving particles and main verbs away from their associated elements. Hence, we are interested in reordering the target language German into English word order which can be beneficial in two ways: (i) Reordering the German part of the parallel training data makes it possible to improve word alignment (which tends to prefer monotonic mappings) and subsequent phrase extraction which leads to better translation models. (ii) We can explore a two-step procedure in which we train a phrase-based SMT model for translating English into German with English word order first (which covers many long-distance relations locally) and then apply a second system that moves words into place according to correct German syntax (which may involve long-range distortion).

For simplicity, we base our experiments on hand-

crafted rules for some of the special cases discussed above. For efficiency reasons, we define our rules over POS tag patterns rather than on full syntactic parse trees. We rely on TreeTagger and apply rules to join verbs in discontinuous verb phrases and to move verb-finals in subordinate clauses, to move verb particles, adverbials and negation particles. Table 1 shows two examples of reordered sentences together with the original sentences in English and German. Our rules implement rough heuristics to identify clause boundaries and word positions. We do not properly evaluate these rules but focus on the down-stream evaluation of the MT system instead.

It is therefore dangerous to extrapolate from short-term trends.
Daher ist es gefährlich, aus kurzfristigen Trends Prognosen abzuleiten.
Daher ist gefährlich es, abzuleiten aus kurzfristigen Trends Prognosen.

The fall of Saddam ushers in the right circumstances.
Der Sturz von Saddam leitet solche richtigen Umstände ein.
Der Sturz von Saddam ein leitet solche richtigen Umstände.

Table 1: Two examples of pre-ordering outputs. The first two lines are the original English and German sentences and the third line shows the reordered sentence.

We use three systems based on Moses to compare the effect of reordering on alignment and translation. All systems are case-sensitive phrase-based systems with lexicalized reordering trained on data provided by WMT. Word alignment is performed using fast_align (Dyer et al., 2013). For tuning we use newstest2011. Additionally, we also test parallel data from OPUS (Tiedemann, 2012) filtered by a method adopted from Mediani et al. (2011).

To contrast our baseline system, we trained a phrase-based model on parallel data that has been aligned on data pre-ordered using the reordering rules for German, which has been restored to the original word order after word alignment and before phrase extraction (similar to (Carpuat et al., 2010; Stymne et al., 2010)). We expect that the word alignment is improved by reducing crossings and long-distance links. However, the translation model as such has the same limitations as the baseline system in terms of long-range distortions. The final system is a two-step model in which we apply translation and language models trained on pre-ordered target language data to perform the first step, which also includes a reordered POS language model. The second step is also treated as a translation problem as in Sudoh et al. (2011), and in our case we use a phrase-based model here with lexicalized reordering and a rather large distortion limit of 12 words. Another possibility would be to apply another rule set that reverts the misplaced words to the grammatically correct positions. This, however, would require deeper syntactic information about the target language to, for example, distinguish main from subordinate clauses. Instead, our model is trained on parallel target language data with the pre-ordered version as input and the original version as output language. For this model, both sides are tagged and a POS language model is used again as one of the target language factors in decoding. Table 2 shows the results in terms of BLEU scores on the newstest sets from 2013 and 2014.

	newstest2013	newstest2014
baseline	19.3	19.1
pre	19.4	19.3
post	18.6	18.7
baseline+OPUS	19.5	19.3
pre+OPUS	19.5	19.3
post+OPUS	19.7	18.8

Table 2: BLEU4 scores for English-German systems (w/o OPUS): Standard phrase-based (*baseline*); phrase-based with pre-ordered parallel corpus used for word alignment (*pre*); two-step phrase-based with post-reordering (*post*)

The results show that pre-ordering has some effect on word alignment quality in terms of supporting better phrase extractions in subsequent steps. Our experiments show a consistent but small improvement for models trained on data that have been prepared in this way. In contrast, the two-step procedure is more difficult to judge in terms of automatic metrics. On the 2013 newstest data we can see another small improvement in the setup that includes OPUS data but in most cases the BLEU scores go down, even below the baseline. The short-comings of the two-step procedure are obvious. Separating translation and reordering in a pipeline adds the risk of error propagation. Furthermore, reducing the second step to single-best translations is a strong limitation and using phrase-based models for the final reordering procedure is probably not the wisest decision. However, manual inspections reveals that many interesting phenomena can be handled even with this simplistic setup.

Table 3 illustrates this with a few selected outcomes of our three systems. They show how verb-particle constructions with long-range distortion

reference	Schauspieler Orlando Bloom **hat sich** zur Trennung von seiner Frau , Topmodel Miranda Kerr , **geäußert** .
baseline	Schauspieler Orlando Bloom **hat** die Trennung von seiner Frau , Supermodel Miranda Kerr .
pre-ordering	Schauspieler Orlando Bloom **hat angekündigt** , die Trennung von seiner Frau , Supermodel Miranda Kerr .
post-ordering	Schauspieler Orlando Bloom **hat** seine Trennung von seiner Frau **angekündigt** , Supermodel Miranda Kerr .
reference	Er **gab** bei einer früheren Befragung den Kokainbesitz **zu** .
baseline	Er **gab** den Besitz von Kokain in einer früheren Anhörung .
pre-ordering	Er **räumte** den Besitz von Kokain in einer früheren Anhörung .
post-ordering	Er **räumte** den Besitz von Kokain in einer früheren Anhörung **ein** .
reference	Borussia Dortmund **kündigte** daraufhin harte Konsequenzen **an** .
baseline	Borussia Dortmund **kündigte an** , es werde schwere Folgen .
pre-ordering	Borussia Dortmund **hat angekündigt** , dass es schwerwiegende Konsequenzen .
post-ordering	Borussia Dortmund **kündigte an** , dass es schwere Folgen **geben werde** .

Table 3: Selected translation examples from the newstest 2014 data; the human *reference* translation; the *baseline* system, *pre-ordering* for word alignment and two-step translation with *post-ordering*.

such as "räumte ... ein" can be created and how discontinuous verb phrases can be handled ("hat ... angekündigt") with the two-step procedure. The model is also often better in producing verb finals in subordinate clauses (see the final example with "geben werde"). Note that many of these improvements do not get any credit by metrics like BLEU. For example the acceptable expression "räumte ein" which is synonymous to "gab zu" obtains less credit then the incomplete baseline translation. Interesting is also to see the effect of pre-ordering when used for alignment only in the second system. The first example in Table 3, for example, includes a correct main verb which is omitted in the baseline translation, probably because it is not extracted as a valid translation option.

4.2 Part-of-Speech Phrase-Distortion Models

Traditional SMT distortion models consist of two parts. A distance-based distortion cost is based on the position of the last word in a phrase, compared to the first word in the next phrase, given the source phrase order. A hard distortion limit blocks translations where the distortion is too large. The distortion limit serves to decrease the complexity of the decoder, thus increasing its speed.

In the Docent decoder, the distortion limit is not implemented as a hard limit, but as a feature, which could be seen as a soft constraint. We showed in previous work (Stymne et al., 2013) that it was useful to relax the hard distortion limit by either using a soft constraint, which could be tuned, or removing the limit completely. In that work we still used the standard parametrization of distortion, based on the positions of the first and last words in phrases.

Our Docent decoder, however, always provides us with a full target translation that is step-wise improved, which means that we can apply distortion measures on the phrase-level without resorting to heuristics, which, for instance, are needed in the case of the lexicalized reordering models in Moses (Koehn et al., 2005). Because of this it is possible to use phrase-based distortion, where we calculate distortion based on the order of phrases, not on the order of some words. It is possible to parametrize phrase-distortion in different ways. In this work we use the phrase-distortion distance and a soft limit on the distortion distance, to mimic the word-based distortion. In our experiments we always set the soft limit to a distance of four phrases. In addition we use a measure based on how many crossings a phrase order gives rise to. We thus have three phrase-distortion features.

As captured by lexicalized reordering models, different phrases have different tendencies to move. To capture this to some extent, we also decided to add part-of-speech (POS) classes to our models. POS has previously successfully been used in pre-reordering approaches (Popović and Ney, 2006; Niehues and Kolss, 2009). The word types that are most likely to move long distances in English–German translation are verbs and particles. Based on this observation we split phrases into two classes, phrases that only contains verbs and particles, and all other phrases. For these two groups we use separate phrase-distortion features, thus having a total of six part-of-speech phrase-distortion features. All of these features are soft, and are optimized during tuning.

In our system we initialize Docent by running Moses with a standard distortion model and lexicalized reordering, and then continuing the search with Docent including our part-of-speech phrase-distortion features. Tuning was done separately for the two components, first for the Moses component, and then for the Docent component initialized by

reference	Laut Dmitrij Kislow von der Organisation "Pravo na oryzhie" **kann** man eine Pistole vom Typ Makarow für 100 bis 300 Dollar **kaufen**.
baseline	Laut Dmitry Kislov aus der Rechten zu Waffen, eine Makarov Gun-spiele **erworben werden können** für 100-300 Dollar.
POS+phrase	Laut Dmitry Kislov von die Rechte an Waffen, eine Pistole Makarov für 100-300 Dollar **erworben werden können**.
reference	Die Waffen **gelangen** über mehrere Kanäle auf den Schwarzmarkt.
baseline	Der "Schwarze" Markt der Waffen ist wieder **aufgefüllt** über mehrere Kanäle.
POS+phrase	Der "Schwarze" Markt der Waffen durch mehrere Kanäle wieder **aufgefüllt ist**.
reference	Mehr Kameras **könnten** möglicherweise das Problem **lösen**...
baseline	Möglicherweise **könnte** das Problem **lösen**, eine große Anzahl von Kameras...
POS+phrase	Möglicherweise, eine große Anzahl von Kameras **könnte** das Problem **lösen**...

Table 4: Selected translation examples from the newstest2013 data; the human *reference* translation; the *baseline* system (Moses with lexicalized reordering) and the system with a *POS+phrase* distortion model.

Moses with lexicalized reordering with its tuned weights. We used newstest2009 for tuning. The training data was lowercased for training and decoding, and recasing was performed using a second Moses run trained on News data. As baselines we present two Moses systems, without and with lexicalized reordering, in addition to standard distortion features.

Table 5 shows results with our different distortion models. Overall the differences are quite small. The clearest difference is between the two Moses baselines, where the lexicalized reordering model leads to an improvement. With Docent, both the word distortion and phrase distortion without POS do not help to improve on Moses, with a small decrease in scores on one dataset. This is not very surprising, since lexical distortion is currently not supported by Docent, and the distortion models are thus weaker than the ones implemented in Moses. For our POS phrase distortion, however, we see a small improvement compared to Moses, despite the lack of lexicalized distortion. This shows that this distortion model is actually useful, and can even successfully replace lexicalized reordering. In future work, we plan to combine this method with a lexicalized reordering model, to see if the two models have complementary strengths. Our submitted system uses the POS phrase-distortion model.

System	Distortion	newstest2013	newstest2014
Moses	word	19.4	19.3
Moses	word+LexReo	19.6	19.6
Docent	word	19.5	19.6
Docent	phrase	19.5	19.6
Docent	POS+phrase	19.7	19.7

Table 5: BLEU4 scores for English–German systems with different distortion models.

If we inspect the translations, most of the differences between the Moses baseline and the system with POS+phrase distortion are actually due to lexical choice. Table 4 shows some examples where there are word order differences. The result is quite mixed with respect to the placement of verbs. In the first example, both systems put the verbs together but in different positions, instead of splitting them like the reference suggests. In the second example, our system erroneously put the verbs at the end, which would be fine if the sentence had been a subordinate clause. In the third example, the baseline system has the correct placement of the auxiliary "könnte", while our system is better at placing the main verb "lösen". In general, this indicates that our system is able to support long-distance distortion as it is needed in certain cases but sometimes overuses this flexibility. A better model would certainly need to incorporate syntactic information to distinguish main from subordinate clauses. However, this would add a lot of complexity to the model.

5 Conclusion

We have described the three Uppsala University systems for WMT14. In the English–French system we extend our document-level decoder Docent (Hardmeier et al., 2013a) to handle pronoun anaphora and introduced a dependency projection model. In our two English–German system we explore different methods for handling reordering, based on Docent and Moses. In particular, we look at post-ordering as a separate step and tunable POS phrase distortion.

Acknowledgements

This work forms part of the Swedish strategic research programme eSSENCE. We also acknowledge the use of the Abel cluster, owned by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR) and operated by the Department for Research Computing at USIT, under project `nn9106k`. Finally, we would also like to thank Eva Pettersson, Ali Basirat, and Eva Martinez for help with human evaluation.

References

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English statistical machine translation by reordering post-verbal subjects for alignment. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 178–183, Uppsala, Sweden.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard University, Cambridge, Massachusetts, USA.

Eunah Cho, Thanh-Le Ha, Mohammed Mediani, Jan Niehues, Teresa Herrmann, Isabel Slawik, and Alex Waibel. 2013. The Karlsruhe Institute of Technology translation systems for the WMT 2013. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 104–108, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 531–540, Ann Arbor, Michigan, USA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the NAACL: Human Language Technologies*, pages 644–648, Atlanta, Georgia, USA.

Dmitriy Genzel. 2010. Automatically learning source-side reordering rules for large scale machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 376–384, Beijing, China.

Anita Gojun and Alexander Fraser. 2012. Determining the placement of German verbs in English–to–German SMT. In *Proceedings of the 13th Conference of the EACL*, pages 726–735, Avignon, France.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013a. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the ACL, Demonstration session*, pages 193–198, Sofia, Bulgaria.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013b. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 967–975, Prague, Czech Republic.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

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 ACL, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Mohammed Mediani, Eunah Cho, Jan Niehues, Teresa Herrmann, and Alex Waibel. 2011. The KIT English–French translation systems for IWSLT 2011. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 73–78, San Francisco, California, USA.

Hwidong Na, Jin-Ji Li, Jungi Kim, and Jong-Hyeok Lee. 2009. Improving fluency by reordering target constituents using MST parser in English-to-Japanese phrase-based SMT. In *Proceedings of MT Summit XII*, pages 276–283, Ottawa, Ontario, Canada.

Jan Niehues and Muntsin Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 206–214, Athens, Greece.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider context by using bilingual language models in machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 198–206, Edinburgh, Scotland. Association for Computational Linguistics.

Franz Josef Och, Nicola Ueffing, and Hermann Ney. 2001. An efficient A* search algorithm for Statistical Machine Translation. In *Proceedings of the ACL 2001 Workshop on Data-Driven Machine Translation*, pages 55–62, Toulouse, France.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 160–167, Sapporo, Japan.

Maja Popović and Hermann Ney. 2006. POS-based reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1278–1283, Genoa, Italy.

Kay Rottmann and Stephan Vogel. 2007. Word reordering in statistical machine translation with a POS-based distortion model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 171–180, Skövde, Sweden.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the Seventh International Conference on Spoken Language Processing*, pages 901–904, Denver, Colorado, USA.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2010. Vs and OOVs: Two problems for translation between German and English. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 183–188, Uppsala, Sweden.

Sara Stymne, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Tunable distortion limits and corpus cleaning for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 225–231, Sofia, Bulgaria.

Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2011. Post-ordering in statistical machine translation. In *Proceedings of MT Summit XIII*, pages 316–323, Xiamen. China.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey.

Ian H. Witten and Timothy C. Bell. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094.

Fei Xia and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 508–514, Geneva, Switzerland.