A Document-Level SMT System with Integrated Pronoun Prediction

Christian Hardmeier
Uppsala University
Department of Linguistics and Philology
Box 635, 751 26 Uppsala, Sweden
first.last@lingfil.uu.se

Abstract

This paper describes one of Uppsala University's submissions to the pronoun-focused machine translation (MT) shared task at DiscoMT 2015. The system is based on phrase-based statistical MT implemented with the document-level decoder Docent. It includes a neural network for pronoun prediction trained with latent anaphora resolution. At translation time, coreference information is obtained from the Stanford CoreNLP system.

1 Introduction

One of Uppsala University's submissions to the pronoun-focused translation task at DiscoMT 2015 is a document-level phrase-based statistical machine translation (SMT) system integrating a neural network classifier for pronoun prediction. The system unites various contributions to discourse-level machine translation that we made during the last few years: The translation system uses our document-level decoder for phrase-based SMT, Docent (Hardmeier et al., 2012; Hardmeier et al., 2013a). The pronoun prediction network was first described by Hardmeier et al. (2013b), and its integration into the decoder by Hardmeier (2014, Chapter 9). In comparison to previous work, the size of the parallel training corpus has been reduced to be more consistent with the official data sets of the shared task. However, for practical reasons, we still use previously trained models that do not match the constraints of the official data sets exactly. Also, while the latent anaphora resolution approach of Hardmeier et al. (2013b) is used for training, allowing us to train our system without running anaphora resolution over the entire training corpus, we rely on coreference annotations generated with the Stanford CoreNLP toolkit (Lee et al., 2013) at test time, as we believe them to be more reliable.

2 MT setup

Owing to time constraints, the setup of our MT system is different from the official baseline provided by the shared task organisers. The system we use is a standard phrase-based SMT system with a phrase table trained on the TED, Europarl (v7) and News commentary (v9) corpora. The system has 3 language models (LMs). The main LM is a 6-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998), trained with KenLM (Heafield, 2011) on the TED, News commentary and News crawl corpora provided for the WMT 2014 shared task (Bojar et al., 2014) and the French Gigaword corpus, LDC2011T10. Additionally, we include a 4-gram bilingual LM (Niehues et al., 2011) and a 9-gram LM over Brown clusters (Brown et al., 1992). Both of these are trained with SRILM (Stolcke et al., 2011) using Witten-Bell smoothing (Witten and Bell, 1991) over a corpus consisting of TED, Europarl, News commentary and United Nations data. Unlike the official baseline, we do not use any lowercasing, recasing or truecasing steps in our training procedure. Instead, all our models are trained directly on the original text in the form in which it occurs in the corpus data. The phrase table is trained with the Moses toolkit (Koehn et al., 2007), and the feature weights of all the models except for the pronoun prediction classifier are optimised towards the BLEU score (Papineni et al., 2002) with the MERT algorithm (Och, 2003) as implemented in Moses.

To increase the effect of the pronoun prediction model, our system uses pronoun placeholders for the pronouns *il*, *elle*, *ils* and *elles* (Hardmeier, 2014, Chapter 9). In the phrase table and the main LM, these pronouns are substituted by four placeholders, LCPRONOUN-SG and UCPRONOUN-SG for upper- and lowercase *il* or *elle* and LCPRONOUN-PL and UCPRONOUN-PL for upper- and lowercase *ils* and

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 72–77,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

elles, respectively. This means that the translation probabilities and the main LM do not offer the system any help to select between the masculine and the feminine forms of the pronouns. The same is true of the Brown cluster LM, since the clustering algorithm automatically assigned the feminine and masculine pronouns to the same clusters. In the bilingual LM, no substitution was made, so this LM still contains information about pronoun choice.

At decoding time, we first run a pass of dynamic-programming beam search decoding with Moses, using only sentence-level models, to initialise the state of our document-level decoder, Docent. Then we add the pronoun prediction model and continue decoding with Docent for 2^{25} iterations. In Docent, we use the simulated annealing search algorithm with a geometric decay cooling schedule, starting at a temperature of 1 and reducing the temperature by a decay factor of 0.99999 at each accepted step. In addition to the *change-phrase-translation*, *swap-phrases* and *resegment* operations described by Hardmeier et al. (2012), we include a *crossover* operation that generates a new state by randomly picking complete sentences either from the current decoder state or from the best state encountered so far, and a *restore-best* operation that unconditionally jumps back to the best state encountered. The last two operations are necessary because simulated annealing accepts state changes with a certain probability even if they decrease the score, and after a sequence of accepted changes to the worse the decoder may get lost in unpromising regions of the search space.

3 The Pronoun Prediction Network

We model pronoun prediction with the feed-forward neural network classifier introduced by Hardmeier et al. (2013b). Its overall structure is shown in figure 1. To create input data for the network, we first generate a set of antecedent candidates for a given pronoun by running the pre-processing pipeline of the coreference resolution system BART (Versley et al., 2008). Each training example for our network can have an arbitrary number of antecedent candidates. Next, we prepare three types of features. *Anaphor context features* describe the source language (SL) pronoun (**P**) and its immediate context consisting of three words to its left (**L1** to **L3**) and three words to its right (**R1** to **R3**), encoded as one-hot vectors. *Antecedent features* (**A**) describe an antecedent candidate. Candidates are represented by the TL words aligned to the syntactic head of the source language markable noun phrase as identified by the Collins head finder (Collins, 1999), again represented as one-hot vectors. These vectors cannot be fed into the network directly because their number depends on the number of antecedent candidates and on the number of TL words aligned to the head word of each antecedent. Instead, they are averaged to yield a single vector per antecedent candidate. Finally, *anaphoric link vectors* (**T**) describe the relationship between an anaphor and a particular antecedent candidate. These vectors are generated by the feature extraction machinery in BART and include a standard set of features for coreference resolution (Soon et al., 2001; Uryupina, 2006) borrowed wholesale from a working coreference system.

In the forward propagation pass, the input word representations are mapped to a low-dimensional representation in an embedding layer (**E**). In this layer, the embedding weights for all the SL vectors (the pronoun and its 6 context words) are tied, so if two words are the same, they are mapped to the same lower-dimensional embedding regardless of their position relative to the pronoun. To process the information contained in the antecedents, the network first computes the link probability for each antecedent candidate. The anaphoric link features (**T**) are mapped to a hidden layer with logistic sigmoid units (**U**). The activations of the hidden units are then mapped to a single value, which functions as an element in an internal softmax layer over all antecedent candidates (**V**). This softmax layer assigns a probability $p_1 \dots p_n$ to each antecedent candidate. The antecedent feature vectors **A** are projected to lower-dimensional embeddings, weighted with their corresponding link probabilities and summed. The weighted sum is then concatenated with the source language embeddings in the **E** layer. The embedding of the antecedent word vectors is independent from that of the SL features since they refer to a different vocabulary.

In the next step, the entire **E** layer is mapped to another hidden layer (**H**), which is in turn connected to a binary output layer predicting the classes *il* and *elle* for the singular classifier and *ils* and *elles* for the plural classifier, respectively. The non-linearity of both hidden layers is the logistic sigmoid function. The dimensionality of the source and target language word embeddings is 50 in our setup, resulting in a total embedding layer

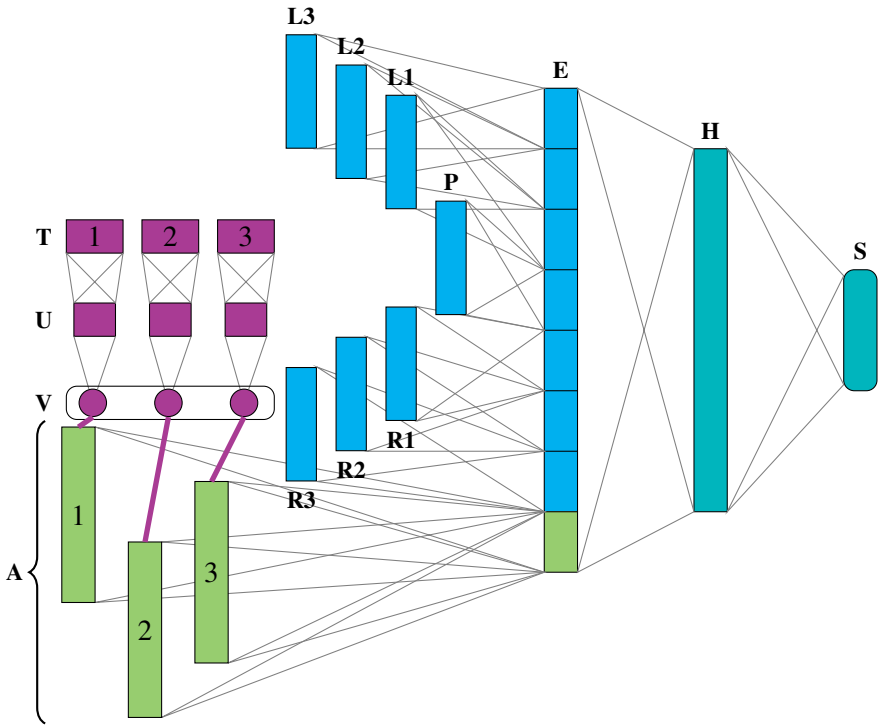

Figure 1: Neural network with latent anaphora resolution

size of 400, and the size of the last hidden layer is set to 150. The network was regularised with an ℓ_2 penalty that was set using grid search over a held-out development set. The network is trained with the RMSPROP algorithm with cross-entropy as the training objective. The gradients are computed using backpropagation. Note that the number of weights in the network is the same for all training examples even though the number of antecedent candidates varies because all weights related to antecedent word features and anaphoric link features are shared between all antecedent candidates. The model is trained on the entirety of the TED corpus enriched with examples from the 10^9 corpus. We reserve a random sample of 10 % of the TED part of the training data as a validation set. Training is run for 300 epochs, and the model used for testing is the one that achieves the best classification accuracy on the validation set.

As earlier experiments suggested that the latent anaphora resolution method integrated in the pronoun prediction network, though useful for training, may not be sufficient for good performance at test time, we decided to use annotations created with an external coreference resolution system when translating the test set. Coreference links were generated with the Stanford CoreNLP software[1] (Lee et al., 2013). The output of the anaphora resolver is deterministic and clusters the mention in the document into a number of coreference sets. We transform these clusters into links by selecting, for each anaphoric pronoun, the closest preceding mention in the same coreference set that is realised as a full noun phrase (rather than another pronoun), if such a mention exists, or the closest mention in the same set otherwise. This leaves us with (at most) a single antecedent per pronoun, so the **V** layer of the neural network is trivially reduced to a single element with probability one, and the **T** and **U** layers are not used at all at test time.

4 Results and Discussion

When considering the outcome of the shared task, we first notice that the performance of our system in terms of BLEU scores (Papineni et al., 2002), with a score of 32.6 %, is several points below that of the systems based on the officially provided baseline, which range around 37 %.[2] It seems likely that this difference, which is confirmed by other automatic

[1] We are grateful to Liane Guillou for providing us with ready-made CoreNLP annotations of the DiscoMT test set.

[2] For a presentation and discussion of the complete shared task methodology and results, we refer the reader to the shared task overview paper (Hardmeier et al., 2015).

	This system					*Baseline*
	Precision		R_{max}		F_{max}	F_{max}
ce	29/ 35	(0.829)	32/ 45	(0.711)	0.765	0.832
ça/cela	9/ 10	(0.900)	22/ 60	(0.367)	0.521	0.631
elle	3/ 9	(0.333)	3/ 20	(0.150)	0.207	0.452
elles	3/ 3	(1.000)	4/ 15	(0.267)	0.421	0.436
il	7/ 43	(0.163)	11/ 19	(0.579)	0.254	0.522
ils	45/ 54	(0.833)	45/ 48	(0.938)	0.882	0.900
on	0/ 0	(n/a)	0/ 0	(n/a)	n/a	n/a
Micro-average	96/154	(0.623)	96/177	(0.542)	0.580	0.699

Accuracy with OTHER: 122/210 = 0.581 (Baseline: 0.676)
Accuracy without OTHER: 96/183 = 0.525 (Baseline: 0.630)
6 bad translations (Baseline: 9)

Table 1: Manual evaluation results for the UU-HARDMEIER system

metrics, is mainly due to differences in the underlying SMT baseline, and the result suggests that we should reconsider the baseline to be used in future experiments. At the same time, it is worth pointing out that the SMT system described in our earlier work (Hardmeier, 2014) used a considerably larger phrase table than our DiscoMT system. It included, in addition to the News commentary and the Europarl corpora, a large amount of data from the Common crawl, United Nations and 10^9 corpora from the WMT shared tasks, and we expect that a system with the full phrase table would reach a higher performance than the one presented here.

The results of our system in the official manual evaluation are shown in Table 1. In the manual evaluation, 210 instances of the English pronouns *it* and *they* were annotated with correct pronouns in the context of the MT output. The table displays the class-specific evaluation metrics for each of the pronoun types in the human evaluation, two accuracy scores including and excluding the OTHER label and the number of examples labelled BAD TRANSLATION by the human annotators. The primary metric of the shared task evaluation is the "Accuracy with OTHER" score, which corresponds to the total proportion of matching examples in the annotated sample. The "Accuracy without OTHER" score is computed over the subset of the examples not annotated with OTHER only. The class-specific scores include a standard precision score in combination with a modified recall score named R_{max} that accounts for the fact that every example potentially has multiple correct annotations, as well as an F_{max} score defined as the harmonic mean of these two quantities. A more detailed description of and rationale for the scores can be found in the shared task overview paper (Hardmeier et al., 2015). For comparison, the table also includes the scores of the official baseline system, which happens to be the top-ranked system in the evaluation.

In terms of pronoun translation accuracy, our model ends up in the middle field of the participants with rank 4 out of 7 (including the baseline). The class-specific scores are consistently below the baseline, in particular for the singular pronouns *il* and *elle*. The masculine pronoun *il* seems to suffer from serious overgeneration, which leads to a very low precision score. The instances of *elle* that the system generated, by contrast, are both too few and mostly wrong. On the whole, the results are rather disappointing, especially since our earlier results with this model (Hardmeier, 2014) had resulted in slightly positive findings. In those experiments, however, we had used oracle annotations of pronoun coreference instead of the automatic CoreNLP annotations used here, and even in that setting, the improvement was very modest.

The results of the shared task suggest that in both the pronoun prediction and the pronoun-focused translation task, it is very hard to beat the baseline systems. In both baseline systems, the *n*-gram model is the only context-sensitive source of information for pronoun choice, and it seems that it is surprisingly difficult to improve pronoun prediction or translation by exploiting additional information despite the obvious and well-known shortcomings of the *n*-gram approach. Future work must show whether this is due to the *n*-gram model's extraordinary capacity for making guesses about remote context by analysing local context, as certain findings suggest (Hardmeier, 2014, 137–138), or just to the fact that our incomplete understanding of the

problem leads us to design bad predictors that are easily beaten by a somewhat sophisticated baseline. By using placeholders in the phrase table and the main LM, we explicitly disable the *n*-gram model for pronoun prediction in our system. It seems likely that this, in conjunction with the fact that our prediction model does not appear to deliver the performance required for improved pronoun translation, is one of the reasons contributing to the lower scores we achieve.

Acknowledgements

This work was supported by the Swedish Research Council under project 2012-916 *Discourse-Oriented Statistical Machine Translation*. The experiments were run on 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, the University of Oslo IT-department.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore (Maryland, USA).

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.

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

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

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 Association for Computational Linguistics: System Demonstrations*, 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).

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, Lisbon (Portugal).

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*, volume 15 of *Studia Linguistica Upsaliensia*. Acta Universitatis Upsaliensis, Uppsala.

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, UK).

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open source toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics: Demonstration session*, pages 177–180, Prague (Czech Republic).

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics*, 39(4):885–916.

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, UK).

Franz Josef Och. 2003. Minimum error rate training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo (Japan).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia (Pennsylvania, USA).

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa (Hawaii, USA).

Olga Uryupina. 2006. Coreference resolution with and without linguistic knowledge. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC-2006)*, pages 893–898, Genoa (Italy).

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of the ACL-08: HLT Demo Session*, pages 9–12, Columbus (Ohio, USA).

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.