Translating Pronouns with Latent Anaphora Resolution

Christian Hardmeier Jörg Tiedemann Joakim Nivre
Dept. of Linguistics and Philology
Uppsala University, Sweden
firstname.lastname@lingfil.uu.se

Abstract

We discuss the translation of anaphoric pronouns in statistical machine translation from English into French. Pronoun translation requires resolving the antecedents of the pronouns in the input, a classic discourse processing problem that is usually approached through supervised learning from manually annotated data. We cast cross-lingual pronoun prediction as a classification task and present a neural network architecture that incorporates the links between anaphors and potential antecedents as latent variables, allowing us to train the classifier on parallel text without explicit supervision for the anaphora resolver. We demonstrate that our approach works just as well for classification as using an external coreference resolver whereas its impact in a practical translation experiment is more limited.

1 Introduction

Much of the success of statistical machine translation (SMT) has been achieved by training fairly simple models on large amounts of data consisting mostly of parallel text, which is available in abundant quantities. The standard SMT models are very local and have difficulties with non-local linguistic phenomena because they cannot easily pick up the relevant information in a document-wide context. By contrast, the supervised methods popular in discourse modelling have other limitations. Systems like coreference resolvers or discourse parsers are often trained on relatively small data sets with elaborate annotations. In a practical machine translation (MT) task, the annotations may not provide exactly the information required for MT (task mismatch) or the texts available for training may not be representative of the texts encountered at test time (domain mismatch).

In this paper, we present results from our research on modelling pronominal anaphora in SMT [1, 2], a task that requires anaphora resolution, i. e., identifying the antecedents of anaphoric pronouns in the input texts. We first approach the problem as a classification task and predict the translation of pronouns in the context of a given reference translation. In this setting, the links between anaphoric pronouns and their potential antecedents can be modelled as latent variables in a neural network, which allows us to avoid using annotated data for training the anaphora resolver and to train the classifier on parallel text only. We demonstrate that this weakly supervised approach works just as well as using a separately trained anaphora resolution system. By incorporating the classifier into an end-to-end SMT system, we find evidence that our method works well for training, but the anaphoric links found by the model are too weak to improve translations at test time.

2 Pronoun Prediction with Latent Anaphora Resolution

The overall setup of the pronoun prediction task is shown in Fig. 1. We are given an English discourse containing a pronoun along with its French translation and automatically computed word alignments. We focus on the four English third-person subject pronouns *he*, *she*, *it* and *they*. The pronoun *it*,

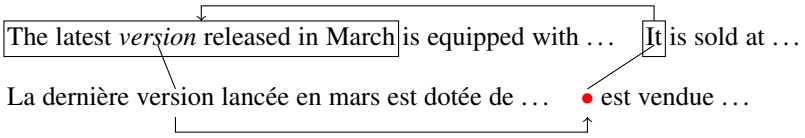

Figure 1: Task setup

unlike the other pronouns, can also be an object pronoun, which adds some of noise to our data sets. The output of the classifier is a multinomial distribution over six classes: Four classes correspond to the four pronouns *il*, *elle*, *ils* and *elles*. These are the masculine and feminine singular and plural forms of the third person subject pronoun, respectively. One class corresponds to the impersonal pronoun *ce* or *c'*, which occurs in some very frequent constructions such as *c'est* 'it is'. The elided form *c'* is used when the following word starts with a vowel. A sixth class OTHER indicates that none of these pronouns was used.

The fundamental input features of the classifier (the words making up the anaphor and the antecedent) as well as the output labels (the translation of the anaphor) are readily available in parallel text data and do not require additional annotation. While the anaphoric links must be resolved in order to predict the translation reliably, this information is not part of the output of the classifier. Instead of resolving the links explicitly, we can assign a probability to each potential link and marginalise over the entire set of potential antecedents. The parallel text used to supervise the classifier gives us some information about the link probabilities since knowing the gender of the translation of a pronoun limits the set of possible antecedents to those whose translation is morphologically compatible with the target language (TL) pronoun. The link probabilities can therefore be treated as latent variables in the classification model.

To put this idea into practice, we design a feed-forward neural network for the classification task. Its overall structure is shown in Fig. 2. To create input data for the network, we first generate a set of antecedent candidates for a given pronoun by running the preprocessing pipeline of the coreference resolution system BART [3]. 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 [4], 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 [5, 6] 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 softmax output layer (**S**) with 6 outputs representing the classes *ce*, *elle*, *elles*, *il*, *ils* and OTHER. The non-linearity of both hidden layers is the logistic sigmoid function. The dimensionality of the source and target language word embeddings is 20 in our setup, resulting in a total embedding layer

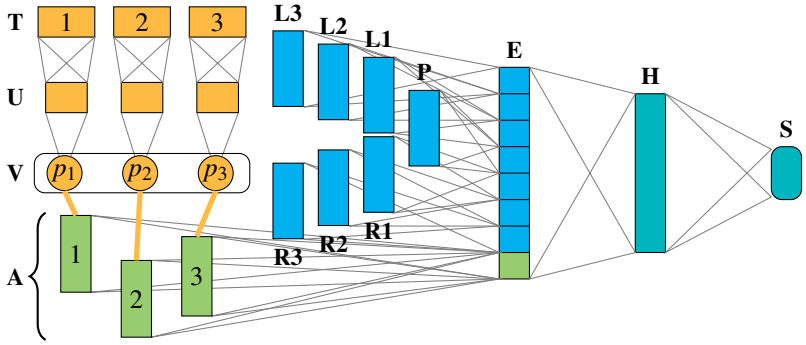

Figure 2: Neural network with latent anaphora resolution

	TED talks						*News commentary*					
	external coref (Accuracy: 0.700)			*latent coref* (Accuracy: 0.696)			*external coref* (Accuracy: 0.576)			*latent coref* (Accuracy: 0.597)		
	P	R	F	P	R	F	P	R	F	P	R	F
ce	0.634	0.747	0.686	0.618	0.722	0.666	0.477	0.344	0.400	0.419	0.368	0.392
elle	0.756	0.617	0.679	0.754	0.548	0.635	0.498	0.401	0.444	0.547	0.460	0.500
elles	0.679	0.319	0.434	0.737	0.340	0.465	0.565	0.116	0.193	0.539	0.135	0.215
il	0.719	0.591	0.649	0.718	0.629	0.670	0.655	0.626	0.640	0.623	0.719	0.667
ils	0.663	0.940	0.778	0.652	0.916	0.761	0.570	0.834	0.677	0.596	0.783	0.677
OTHER	0.743	0.678	0.709	0.741	0.682	0.711	0.567	0.573	0.570	0.614	0.544	0.577

Table 1: Classification results

size of 160, and the size of the last hidden layer is set to 50. In experiments with larger layer sizes, we obtained similar, but no better results. 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.

3 Experimental Results and Discussion

We run experiments with two different test sets. The TED data set consists of around 2.6 million tokens of lecture subtitles released in the WIT^3 corpus [7]. We extract 71,131 training examples from this corpus. The examples are randomly partitioned into a training set of 56,905 examples and a validation set and a test set of 7,113 examples each. The official WIT^3 development and test sets are not used in our classifier experiments because we want to reserve some held-out data for MT experiments. The News commentary data set is version 6 of the parallel News commentary corpus released as a part of the WMT 2011 training data [8]. It contains around 2.8 million tokens of news text and yields 31,090 data points, which are randomly split into 28,090 training examples and validation and test sets of 1,500 examples each.

In Table 1, we compare experimental results from two systems. The system labelled *latent coref* refers to the system described in the previous section. The system named *external coref* is similar, but instead of modelling anaphora resolution as latent variables, it relies on the external coreference resolution system BART [3]. It uses a classifier similar to the one in Fig. 2, but the **V** layer is replaced by probability estimates derived from the machine learning component of BART, which is trained on the coreference-annotated *ACE02-npaper* corpus [9], and the **T** and **U** layers are absent.

Evaluation results are presented in terms of overall accuracy and precision (P), recall (R) and F-score (F) for the individual output classes. The overall accuracy is heavily dominated by the performance on the most frequent output classes. A simple majority class baseline achieves accuracies of 0.622 and 0.555 on TED and News commentary data, respectively. Majority choices are generally handled well by SMT systems, even without specific pronoun handling capacities, so to determine how much value a classifier will add to an MT system, it is more interesting to look at the performance of the

(a) BLEU scores

Corpus	News	TED	
Anaphora res.	predicted	predicted	gold
Baseline	0.2439	0.3086	
With classifier	0.2440	0.3088	0.3089

(b) Human evaluation

Baseline	*with classifier*			
	News/pred.		*TED/gold*	
	–	+	–	+
–	3	17	18	29
+	13	2	7	10

Table 2: SMT results

more infrequent output classes. The pronoun *elles* is particularly interesting. It is the feminine plural of the third person subject pronoun, and it usually corresponds to the English pronoun *they*. It is a marked choice which is used instead of the masculine default *ils* only if the antecedent is exclusively comprised of linguistic elements of feminine grammatical gender, making up no more than 3–4 % of the training set. The distinction between *ils* and *elles* cannot be predicted from the English source pronoun or its context; making correct predictions requires knowledge about the antecedent of the pronoun. Because of these special properties of *elles*, the performance on this class is a good indicator of how well a classifier can represent relevant knowledge about pronominal anaphora as opposed to overfitting to source contexts or acting on prior assumptions about class frequencies.

In the results of Table 1, we find that the overall accuracy of the two systems compared are similar. For the TED talks, the accuracy drops by a mere 0.4 %, whereas it improves by 2.1 % for the News commentaries. Clearly, the removal of the coreference-annotated data from the training process has no major negative effect on performance. What is interesting is that the performance on the pronoun *elles* actually improves in both systems, with a gain of 3.1 % F-score for the TED talks and 2.2 % for News. We interpret this as evidence for the effectiveness of our latent anaphora resolution approach.

The ultimate goal of our research is to use the pronoun prediction classifier as a feature model in SMT. To test its performance, we integrate it in an English–French SMT system based on the document-level decoder Docent [10, 11]. Owing to space constraints, we cannot give a detailed description of these experiments. Our SMT system is trained on data from WMT, enriched by TED talks from WIT^3 in the TED case. An initial translation is created with the sentence-level decoder Moses [12] with a set of standard baseline models. Then it is processed with the document-level hill climbing algorithm in Docent using the pronoun prediction classifier in addition to the baseline models. In the standard condition labelled *predicted* in Table 2, we resolve pronouns with our classifier as described in the previous section. For TED data, there is a test set with manually resolved pronouns [13], which allows us to run an oracle experiment exploiting gold-standard anaphoric links instead of using the **T**, **U** and **V** layers for anaphora resolution. Even in this *gold* condition, the classifier is trained without manual coreference annotations. Manual annotations are only used at test time.

In terms of BLEU scores [14], the impact is negligible (Table 2a). Since BLEU is known to be unreliable for evaluating pronoun translation [15], we have done a simple human evaluation on a sample of the data in the *News/predicted* and the *TED/gold* conditions. The annotators are asked to determine the correct pronoun in the context of the actual machine translation. A subset of this evaluation is presented in Table 2b. The contingency table shows the counts of pronouns matching (+) or conflicting with (–) the human annotation for our SMT system and the baseline in cases where the pronouns generated by the two systems are different. In the News experiment with predicted anaphora resolution, the impact of the model is quite small and not statistically significant. In the TED experiment with oracle links, however, the improvement we observe is significant at a 0.01 level in Liddell's test [16]. We conclude that our latent anaphora resolution method is sufficiently effective for classifier training, but that the quality of its output is still insufficient to use it at test time.

In sum, our experimental results suggest that parallel texts contain valuable information about pronominal anaphora that can be exploited as a form of weak supervision to train an anaphora resolution component. Evaluated as a classifier, latent anaphora resolution performs comparably to using an external coreference resolution system trained on manually annotated data, which is remarkable. When the classifier is integrated into an end-to-end SMT system, its performance is insufficient to improve the MT output. There is evidence that the problematic step is anaphora resolution at test time, while training the classifier without annotated data seems to work fairly well. We see these results as encouraging and expect that they can be further optimised, e. g., by improving the structure of the classifier or by exploiting the existing out-of-domain data sets with manual annotations in addition to the in-domain parallel data in the training process.

References

[1] Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 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), October 2013. Association for Computational Linguistics.

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

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

[4] Michael Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

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

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

[7] Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento (Italy), May 2012.

[8] Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh (Scotland, UK), July 2011. Association for Computational Linguistics.

[9] Alexis Mitchell, Stephanie Strassel, Mark Przybocki, J. K. Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. *ACE-2 Version 1.0*. Linguistic Data Consortium, Philadelphia, 2003. LDC2003T11.

[10] Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 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), July 2012. Association for Computational Linguistics.

[11] Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 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), August 2013. Association for Computational Linguistics.

[12] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 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), 2007.

[13] Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. ParCor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of the Tenth Language Resources and Evaluation Conference (LREC'14)*, Reykjavík (Iceland), 2014.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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), 2002. ACL.

[15] Christian Hardmeier. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11, 2012.

[16] F. D. K. Liddell. Simplified exact analysis of case-referent studies: Matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health*, 37(1):82–84, 1983.