Neural Machine Translation for Cross-Lingual Pronoun Prediction

Sebastien Jean* and **Stanislas Lauly*** and **Orhan Firat*** and **Kyunghyun Cho**
Department of Computer Science
Center for Data Science
New York University
* *Both authors contributed equally*

Abstract

In this paper we present our systems for the DiscoMT 2017 cross-lingual pronoun prediction shared task. For all four language pairs, we trained a standard attention-based neural machine translation system as well as three variants that incorporate information from the preceding source sentence. We show that our systems, which are not specifically designed for pronoun prediction and may be used to generate complete sentence translations, generally achieve competitive results on this task.

1 Introduction

Given a source document and its corresponding partial translation, the goal of the DiscoMT 2017 cross-lingual pronoun prediction shared task (Loáiciga et al., 2017) is to correctly replace the missing pronouns, choosing among a small set of candidates. In this paper, we propose and evaluate models on four sub-tasks: En-Fr, En-De, De-En and Es-En.

We consider the use of attention-based neural machine translation systems (Bahdanau et al., 2014) for pronoun prediction and investigate the potential for incorporating discourse-level structure by integrating the preceding source sentence into the models. More specifically, instead of modeling the conditional distribution $p(Y|X)$ over translations given a source sentence, we explore different networks that model $p(Y|X, X_{-1})$, where X_{-1} is the previous source sentence. The proposed larger-context neural machine translation systems are inspired by recent work on larger-context language modeling (Wang and Cho, 2016) and multi-way, multilingual neural machine translation (Firat et al., 2016).

2 Baseline: Attention-based Neural Machine Translation

An attention-based translation system (Bahdanau et al., 2014) is composed of three parts: encoder, decoder, and attention model.

The source sentence $X = (x_1, x_2, \ldots, x_{T_x})$ is encoded into a set of annotation vectors $\{h_1, h_2, \ldots, h_{T_x}\}$. To do so, we use a bidirectional recurrent network (Schuster and Paliwal, 1997) with a gated recurrent unit (GRU, Cho et al., 2014; Hochreiter and Schmidhuber, 1997).

The decoder, composed of a GRU f topped by a one hidden layer MLP g, models the conditional probability of the target sentence word y_i knowing the previous words and the source sentence $\mathbf{x}$.

$$p(y_i|y_1, ..., y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i) \quad (1)$$

s_i is the RNN hidden state for time i, and c_i is a distinct context vector used to predict y_i.

$$s_i = f(s_{i-1}, y_{i-1}, c_i) \quad (2)$$

The computation of the context vector c_i depends on the previous decoder hidden state and on the sequence of annotations $(h_1, ..., h_{T_x})$, where each h_j is a representation of the whole source sentence with a focus on the j^{th} word. c_i is a weighted sum of the annotations.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (4)$$

* This work was done during his visit to NYU. Now at Google (orhanf@google.com).

Proceedings of the Third Workshop on Discourse in Machine Translation, pages 54–57,
Copenhagen, Denmark, September 8, 2017.

$$e_{ij} = a(s_{i-1}, y_{i-1}, h_j) \quad (5)$$

where e_{ij} is the attention model score, which represents how well the output at time i aligns with the input around time j.

3 Larger-Context Neural Machine Translation

As the antecedent needed to correctly translate a pronoun may be in a different sentence (inter-sentential anaphora) (Guillou et al., 2016), we added the previous sentence as a auxiliary input to the neural machine translation system, using an additional encoder and attention model. Similarly to the source sentence encoding, we apply a bidirectional recurrent network to generate context annotation vectors $\left\{h_1^c, \ldots, h_{T_c}^c\right\}$.

The additional attention model differs slightly from the original one by integrating the current source representation c_i as a new input, so that the context vector depends on the currently attended source words. As such, this attention model takes as input the previous target symbol, the previous decoder hidden state, the context annotation vectors as well as the source vector from the main attention model. That is, the unnormalized alignment scores are computed as

$$e_{ij}^c = a(s_{i-1}, y_{i-1}, h_j, c_i) \quad (6)$$

Similarly to the source vector c_i, the time-dependent context vector c_i^c is also a weighted sum, this time of the context annotation vectors. With this new information, we explored three different approaches.

3.1 Simple Context Model (SCM)

For the first approach, we simply use the context representation c_i^c as a additional input to the decoder GRU and the prediction function g.

$$s_i = f(s_{i-1}, y_{i-1}, c_i, c_i^c) \quad (7)$$

$$p(y_i|y_1, ..., y_{i-1}, \mathbf{x}, \mathbf{x}^c) = g(y_{i-1}, s_i, c_i, c_i^c) \quad (8)$$

3.2 Double-Gated Context Model (DGCM)

Our second approach is very similar to the first with the exception that, for both functions f and g, distinct gates (g_1 and g_2) are applied to the context representation c_i^c. Similar context-modulating gates were previously used by (Wang et al., 2017).

$$s_i = f(s_{i-1}, y_{i-1}, c_i, g_1 \odot c_i^c) \quad (9)$$

$$p(y_i|y_1, ..., y_{i-1}, \mathbf{x}, \mathbf{x}^c) = g(y_{i-1}, s_i, c_i, g_2 \odot c_i^c) \quad (10)$$

Each gate has its own set of parameters and depends on the previous target symbol, the current source representation and the decoder hidden state, at time $i-1$ for g_1 and i for g_2.

3.3 Combined Context Model (CCM)

The last method first combines the source and context representations into a vector d_i through a multi-layer perceptron. As in the second approach, the context is also gated.

$$d_i = \mathbf{W}_3 \Big(\tanh(\mathbf{W}_1 c_i + \mathbf{W}_2(g_1 \odot c_i^c)) \Big) \quad (11)$$

$$s_i = f(s_{i-1}, y_{i-1}, d_i) \quad (12)$$

$$p(y_i|y_1, ..., y_{i-1}, \mathbf{x}, \mathbf{x}^c) = g(y_{i-1}, s_i, d_i) \quad (13)$$

4 Pronoun prediction task

The DiscoMT 2017 pronoun prediction task serves as a platform to improve pronoun prediction. We are provided source documents and their lemmatized translations for four language pairs: En-Fr, En-De, De-En and Es-En. In each translation, some sentences have one or more pronouns substituted by the placeholder "REPLACE". For each of these tokens, we must select the correct pronoun among a small set of candidates.

There are respectively 8, 5, 9 and 7 target classes for En-Fr, En-De, De-En and Es-En. For example, in the case of En-Fr, the task is concentrated on the translation of "it" and "they". The possible target classes are:

	Baseline	SCM	DGCM	CCM
En-Fr	67.9	66.2	**68.9**	64.5
En-De	58.2	57.1	**59.0**	57.6
De-En	70.9	70.3	72.4	**72.8**
Es-En	69.9	**77.1**	70.8	72.3

Table 1: Validation macro-average recall (in %) for cross-lingual pronoun prediction.

	Baseline	SCM	DGCM	CCM	Best
En-Fr	58.1	52.2	62.3	52.1	66.9
En-De	60.9	63.2	61.3	59.5	78.4
De-En	63.3	63.8	64.8	65.5	69.2
Es-En	58.9	56.1	58.7	56.4	58.9

Table 2: Test macro-average recall (in %) for cross-lingual pronoun prediction. The "Best" column displays the highest score across all primary and contrastive submissions to the DiscoMT 2017 shared task (Loáiciga et al., 2017).

- **ce, elle, elles, il, ils, cela, on, OTHER.**

Although only a subset of the data has context dependencies, it is not difficult to find such instances. The following set of sentences taken from the En-Fr development data is a good example:

- **Context:** *So the idea is that accurate perceptions are fitter perceptions .*
- **Source:** *They give you a survival advantage .*

And here are the source sentence translation with the missing token and the corresponding target:

- **Translation:** *REPLACE vous donner un avantage en terme de survie .*
- **Target:** elles

In this example, "REPLACE" should be the translation of the word "They", which refers to "perceptions" in the previous sentence. This is important because in French, "perceptions" is feminine. Correctly choosing a good pronoun here can only be done confidently with contextual information.

5 Experimental settings

To train our models, which are fully differentiable, we use the Adadelta optimizer (Zeiler, 2012). Word embeddings have dimensionality 620, decoder and source encoder RNNs have 1000-dimensional hidden representations, and the context encoder RNN hidden states are of size 620. As the source and context annotations are the concatenation of the forward and backward encoder hidden states, their dimensionality are 2000 and 1240 respectively. The models are regularized with 50% Dropout (Pham et al., 2014) applied to all RNN inputs and on the decoder hidden layer preceding the softmax.

Pronouns are predicted using a modified beam search where the beam is expanded only at the "REPLACE" placeholders, and is otherwise constrained to the reference. The beam size is set to the number of pronoun classes, so that our approach is equivalent to exhaustive search for sentences with a single placeholder. Models for which beam search lead to the highest validation macro-average recall were selected and submitted for the shared task. The baselines were also sent as contrastive submissions.

6 Results

Table 1 and 2 respectively present validation and test results across all language pairs for the models described in sections 2 and 3. Amongst the four models we evaluated on the test sets, a different one performs best for each language pair. Nevertheless, the DGCM model is the most consistent, always ranking second or first amongst our systems. Moreover, it beats the baseline on all tasks except Es-En, which it trails by a marginal 0.2%.

Our models, which don't leverage the given part-of-speech tags and external alignments, are generally competitive with the best submissions (Loáiciga et al., 2017). For Es-En, our contrastive submission achieves the best performance. As for En-Fr and De-En, our systems obtain a macro-average recall within 5% of the winners. Finally, the relatively poor performance of our models for En-De is due to their incapacity at correctly predicting the rare pronoun 'er'. Indeed, the

recall of 0/8 for that class greatly affects the results.

7 Conclusion

In this paper, we have presented our systems for the DiscoMT 2017 cross-lingual pronoun prediction shared task. We have explored various ways of incorporating discourse context into neural machine translation. Even if the DGCM model often achieves better performance than the baseline by taking in account the previous sentence, we believe there is still important progress to be made. In order to improve further, we may need to better understand the impact of context by carefully analyzing the behaviour of our models.

Acknowledgments

This work was supported by Samsung Electronics ("Larger-Context Neural Machine Translation" and "Next Generation Deep Learning: from pattern recognition to AI"). KC thanks Google (Faculty Award 2016), NVIDIA (NVAIL), Facebook, eBay and TenCent for their generous support.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL*.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 wmt shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany. Association for Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT shared task on cross-lingual pronoun prediction. In *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, DiscoMT-EMNLP17.

Vu Pham, Théodore Bluche, Christopher Kermorvant, and Jérôme Louradour. 2014. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*. IEEE, pages 285–290.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on* 45(11):2673–2681.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. *CoRR* abs/1704.04347. http://arxiv.org/abs/1704.04347.

Tian Wang and Kyunghyun Cho. 2016. Larger-context language modelling. In *ACL*.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .