The Kyoto University Cross-Lingual Pronoun Translation System

Raj Dabre
Graduate School of Informatics
Kyoto University, Japan
prajdabre@gmail.com

Yevgeniy Puzikov
Graduate School of Informatics
Kyoto University, Japan
puzikov@nlp.ist.i.kyoto-u.ac.jp

Fabien Cromieres
JST, Japan
Kyoto University
fabien@nlp.ist.i.kyoto-u.ac.jp

Sadao Kurohashi
Graduate School of Informatics
Kyoto University, Japan
kuro@i.kyoto-u.ac.jp

Abstract

In this paper we describe our system we designed and implemented for the cross-lingual pronoun prediction task as a part of WMT 2016. The majority of the paper will be dedicated to the system whose outputs we submitted wherein we describe the simplified mathematical model, the details of the components and the working by means of an architecture diagram which also serves as a flowchart. We then discuss the results of the official scores and our observations on the same.

1 Introduction

The cross-lingual pronoun prediction task in WMT 2016 is a lot more challenging than its 2015 counterpart (Hardmeier et al., 2015) since one cannot rely on solely the target side sentence due to loss of grammatical gender, number and person which is a consequence of lemmatization. As such looking at the source side sentence is quite essential. Since Deep Neural Networks (NN) are becoming increasingly popular and being shown to be extremely effective when it comes to many NLP tasks we decided to go for a full NN approach to see how far it can go. We refer to the shared task overview paper (Guillou et al., 2016) for details of the task and the various other submitted systems.

2 Our System

Here we describe in detail our system and give brief overviews of its variants.

2.1 Motivation

As mentioned earlier, we chose a purely neural network approach since many recent works have shown that NNs are extremely effective when it comes to NLP tasks and can produce results that are able to beat the state of art systems by a reasonable margin. (Mikolov et al., 2010) showed that the word embeddings obtained using a simple feed-forward neural network give better results for word similarity tasks compared to those given by the embeddings obtained using GLOVE(Pennington et al., 2014). Furthermore, (Devlin et al., 2014) have shown that using a Neural Network based Lexical Translation Model can help boost the quality of Statistical Machine Translation. (Bahdanau et al., 2014) showed that it is possible to perform end to end MT whose quality surpasses that of Moses(Koehn et al., 2007) by using a combination of Recurrent Neural Networks (RNNs) and dictionary based unknown word substitution.
In particular we wanted to test the capabilities of Recurrent Neural Networks augmented with an Attention Based Mechanism for this task. They are easy to design, implement and test due to the availability of NN frameworks like Chainer [1], Torch [2], Tensorflow [3] etc. Since Chainer provides a lot of useful functionality and enables rapid prototyping we decided to use it to implement our system.

2.2 System Description

Refer to Figure-1 for a simple overview of our pronoun translation system which we describe in detail below.
Consider that the input sentence (IN) is : *Cabin restaurants , as they 're known in the trade , are venues for forced prostitution* ., the lemmatized output sentence (OUT) is : *le " restaurant cabane " , comme REPLACE_PRON la appeler dans ce commerce , être du lieu de prostitution forcé* . and the pronoun to be predicted in place of REPLACE_PRON is *on*. The following must be no-

1. http ://docs.chainer.org
2. https ://github.com/torch/distro
3. http ://tensorflow.org

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 571–575,
Berlin, Germany, August 11-12, 2016.

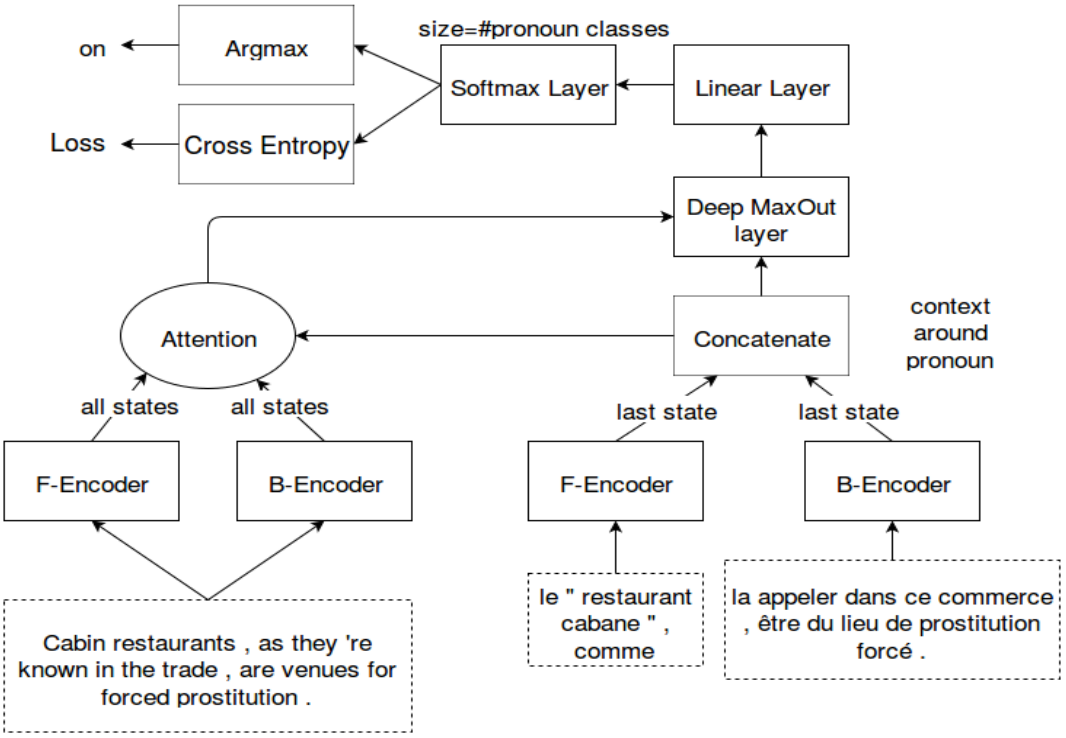

FIGURE 1 – The RNN Pronoun Translation System

ted :

- In the target sentence, *le " restaurant cabane " , comme* and *la appeler dans ce commerce , être du lieu de prostitution forcé .* represent the context before (left) and after (right) the pronoun respectively.
- In case the contexts contain other pronouns to be predicted then they are simply represented by a token called "PRON_PLACEHOLDER".
- If either of the contexts are empty (the pronoun is the first word of the sentence) we use a padding like "UNK" or "#".
- The memory cells used in the RNN encoders are GRUs and we do not consider stacked RNNs.
- The prefixes F and B represent forward (left to right) and backward (right to left) respectively and indicate the direction of the RNN encoding of the sentence. The encoders used for the source and target languages are separate.
- The size of the output of the Softmax layer is equal to the number of the pronoun classes in the target language.
- Unless mentioned otherwise, all the Neural network layers like Attention, Softmax, Linear and Deep Maxout are the same as the ones mentioned in (Bahdanau et al., 2014).

To predict the pronoun given the input sentence (IN) and the target side contexts (OUT-Left and OUT-Right) we perform the following steps :

1. FWD_ENC_SRC=F-Encoder(IN) and BWD_ENC_SRC=B-Encoder(IN). These are 2 sequences of RNN states with a forward and backward representation for each word.
2. FWD_ENC_TGT=Last(F-Encoder(OUT-Left)) and BWD_ENC_TGT=Last(B-Encoder(OUT-Right)). TGT_CONTEXT=Concatenate(FWD_ENC_TGT, BWD_ENC_TGT). We select the last states which represent left and right context. As mentioned before, the encoders for the source and target languages are separate and do not share parameters.
3. SRC_ATTENTION=Attention(FWD_ENC_SRC, BWD_ENC_SRC). This gives an attention vector which is a weighted average of the forward and backward RNN state sequences.
4. LOGITS=Linear(Maxout(SRC_ATTENTION, TGT_CONTEXT)).

These give the logits which represent the weights for each pronoun class.

5. LOSS=Softmax-Cross-Entropy(LOGITS) and PREDICTION=Argmax(LOGITS). The criterion for the prediction loss (on which backpropagation is done) is the Softmax Cross Entropy. The pronoun class which receives the maximum weight is output as the predicted class.

Apart from this, we do not do any post-editing of any sort. Thus the NN model tries to learn the following probability distribution :

$$P_\theta(REPLACE\_PRON|IN, OUT)$$

The optimization objective is simply to maximize the following likelihood function :

$$L_\theta = \prod_{\forall(PR,IN,OUT)\in T} P_\theta(PR|IN, OUT)$$

Where *PR* is the same as REPLACE_PRON, the pronoun to be predicted and T is the training set collecting all input, output and the label to be predicted. Note that OUT is decomposed as (OUT-Left,OUT-right).

2.3 Training and Testing

We only used the IWSLT corpus for each language pair for training and the corresponding TEDdev corpus as the development set. We refer to the shared task overview paper for the corpora details. We simply process the corpora to convert it into the format (as in figure-1) which our system accepts. No other kind of preprocessing or annotation in terms of anaphora resolution is performed. No external/extra corpus was used. Our objective was to see how far a pure Neural Network system could go. We use the following neural network parameters/vector dimensions.

— Vocabulary size : 600000 (which is enough to cover all words in the training data and more than 99.5% of the words in the development and test set)
— Source and target words embedding size : 100
— Source and target GRU cell output size : 200
— Attention Module Hidden layer size : 200
— Maxout output size : 150
— Minibatch size : 80 (80 pronouns predicted per batch)
— Weight decay : 0.000001 (for regularization)
— Optimization algorithm : ADAM (Kingma and Ba, 2014)

Additionally we tried with embedding and other layer sizes 5 times the above but they had very little effect. Moreover, the reduced dimensionality gave smaller models and allowed for faster training. As an early stopping criterion we evaluate our model every 50 iterations (4000 predictions) on the development set and save it only if its performance on the development set improves over the previous evaluation. We give the results of the evaluation of the test set pronoun translations for the various languages in the following section.

2.4 Results and Discussion

Refer to Table-1 for the official scores for all language pairs. The official score is the Macro Averaged R score. In general our system secured 2nd rank in 3 out of 4 language pairs with respect to R-score and 1st rank in 2/4 language pairs with respect to the Accuracy. Based on our preliminary evaluations our system performs well on the non-rare classes. Based on the confusion matrices obtained on the results, we noted that pronoun classes that rarely occurred in the training corpus (and equivalently in the development and text corpus) had very low classification accuracy and hence contributed to reduced R-scores. Another interesting observation is that although our accuracies were high, the R-score was not which is a further indicator that our system simply does not learn to classify the rare pronouns accurately.

If one takes a look at the language pairs then it is interesting to note that when German is the target language our system has the worst performance but is almost on par with the best system when it is the source language. We believe that since we use both the input and output sentences for the pronoun prediction and that German is a morphologically rich language our system is able to leverage the morphological richness through the attention mechanism. It is also evident that only using the target side sentence to predict the pronoun (like the baseline system does) will not be very helpful since the pronoun depends on information such as gender, number and person information (which is removed as a result of lemmatization) of the word that it refers to.

As a side note we would like to point out that we evaluated our system every 50 iterations and recorded the scores at each stage. In case of German-

Language Pair	R-score	Accuracy	Rank	Difference wrt Best System
German-English	73.17%	80.33%	2/6	-0.74%
English-German	52.50%	71.28%	2/9	-11.91%
French-English	65.63%	82.93%	2/5	-7.4%
English-French	62.44%	70.51%	3/9	-3.26%

TABLE 1 – The R-Scores and Accuracies on the test sets for all language pairs

English we observed that we had overfitted on the development set and during a previous iteration the R-score on the test set was 58.37%. This clearly indicates that if the development set is different from the test set then overfitting can have undesirable consequences. One way of avoiding overfitting is reducing the size of the NN (in terms of the sizes of layers and embeddings) which cannot be really verified in our case since it needs a grid search on all possible NN sizes which in turn needs a lot of time and/or a large number of GPUs which we lacked. However, as we have mentioned before, a five-fold reduction in parameter space did not hurt the performance and hence it would be interesting to find out the smallest model (in terms of number of parameters) that can still have high performance.

3 Conclusion

We have reported our Recurrent Neural Network based pronoun classification (or translation) system in sufficient detail along with the official scores. Overall we have secured second place in the competition inspite of a simple RNN system which uses a very small amount of data (IWSLT only) for training without any additional pre/post processing involving coreference resolution. In the future, we would like to work on leveraging larger corpora and coreference resolution so as to address the rare pronoun classes. We would also like to conduct a proper grid search so as to determine the best embedding and layer sizes. Finally we would like to investigate into ensemble systems where we train a bunch of RNN systems for the same language pair and then use a simple scheme like max-voting to overcome the problem of models that have overfitted on the development set and those that may have inferior performance possibly due to reasons such as model initialization.

Acknowledgments

The first two authors would like to thank MEXT (Japan Government) for the scholarship they receive. We would also like to thank John Richardson for a number of tips with respect to the neural network parameters we chose for our system. We also thank the organizers and the reviewers for their efforts and helpful reviews.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1370–1380, Baltimore, Maryland, June. Association for Computational Linguistics.

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.

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*, pages 1–16, Lisbon, Portugal, September. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980.*

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive*

poster and demonstration sessions, pages 177–180. Association for Computational Linguistics.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pages 1045–1048.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.