Cross-Lingual Pronoun Prediction with Deep Recurrent Neural Networks

Juhani Luotolahti[⋆,1] **Jenna Kanerva**[⋆,1,2] **and Filip Ginter**[1]
[1]Department of Information Technology, University of Turku, Finland
[2]University of Turku Graduate School (UTUGS), Turku, Finland
mjluot@utu.fi jmnybl@utu.fi figint@utu.fi

Abstract

In this paper we present our winning system in the WMT16 Shared Task on Cross-Lingual Pronoun Prediction, where the objective is to predict a missing target language pronoun based on the target and source sentences. Our system is a deep recurrent neural network, which reads both the source language and target language context with a softmax layer making the final prediction. Our system achieves the best macro recall on all four language pairs. The margin to the next best system ranges between less than 1pp and almost 12pp depending on the language pair.

1 Introduction

Automatic translation of pronouns across languages can be seen as an subtask of the full machine translation. In the pronoun translation task the special challenge is posed by anaphora resolution as well as differing gender marking in different languages. The WMT16 Shared Task on Cross-Language Pronoun Prediction strives to seek for methods to address this particular problem (Guillou et al., 2016).

This shared task includes two language pairs, English-French and English-German, and both translation directions, so in total four different source-target pairs must be considered. In the target language side selected set of pronouns are substituted with `replace`, and the task is then to predict the missing pronoun. Furthermore, the target side language is not given as running text, but instead in lemma plus part-of-speech tag format. This is to mimic the representation which many standard machine translation systems produce and to complicate the matter of standard language modeling. An example of an English-French sentence pair is given in Figure 1. Furthermore, the training data as provided by the organizers of the the task includes automatically produced word-level alignments between the source and the target language.

Source: *That 's how* ***they*** *like to live .*
Target: *ce|PRON être|VER comme|ADV cela|PRON que|PRON* ***REPLACE_3*** *aimer|VER vivre|VER .|.*

Figure 1: An example sentence from the English to French training data, where the REPLACE_3 is a placeholder for the word to be predicted.

In this paper we describe the pronoun prediction system of the Turku NLP Group. Our system is a deep recurrent neural network with word-level embeddings, two layers of Gated Recurrent Units (GRUs) and a softmax layer on top of it to make the final prediction. The network uses both source and target contexts to make the prediction, and no additional data or tools are used beside the data provided by the organizers. The system has the best macro recall score in the official evaluation on all four language pairs.

2 Related work

This shared task is a spiritual successor to an earlier cross-lingual pronoun prediction shared task (Hardmeier et al., 2015). The systems submitted to the earlier task provide us with a good view of the recent related work on the problem. The earlier task received altogether six system description papers. The organizers identify two main approaches used by the participants. Teams UEDIN (Wetzel et al., 2015) and MALTA (Pham and van der Plas, 2015) explicitly tried to resolve anaphoras in the text and using the information to

⋆ Both authors contributed equally to this work.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 596–601,
Berlin, Germany, August 11-12, 2016.

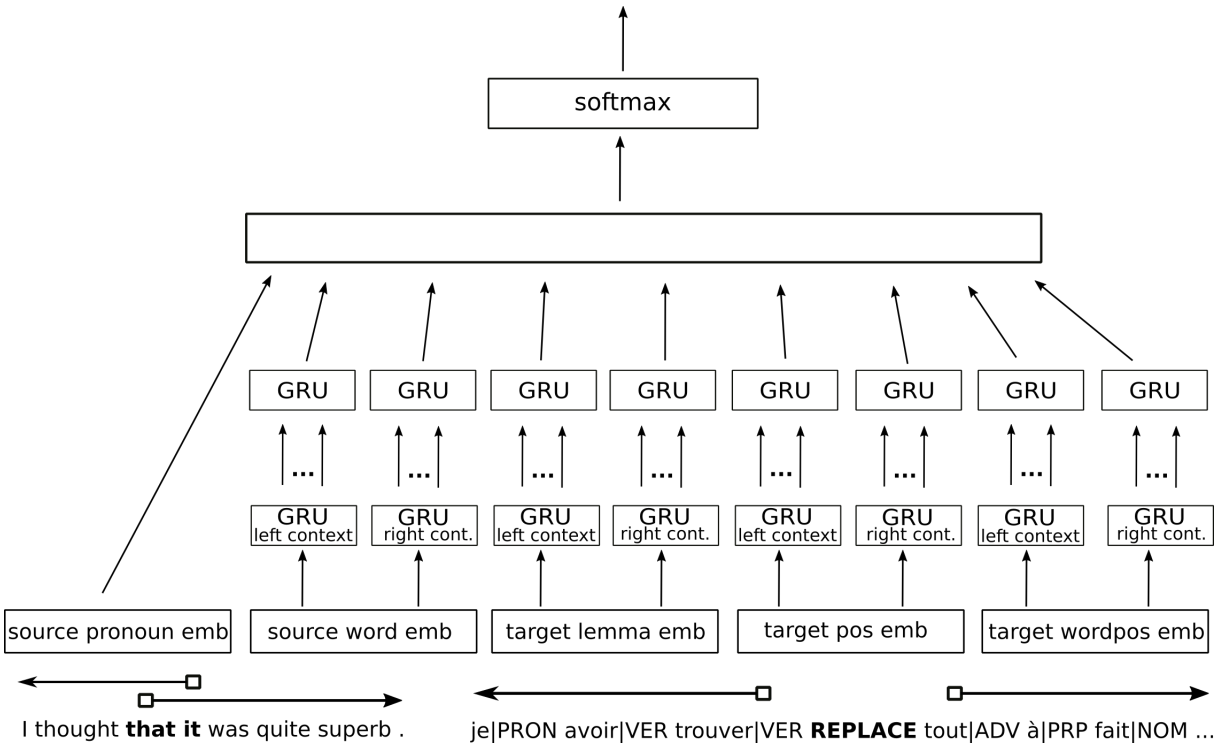

Figure 2: The architecture of our recurrent neural network system.

help predict the pronoun.

Other teams relied more on the context, for example UU-Tiedemann (Tiedemann, 2015) used a linear SVM with features from the context of the pronoun. IDIAP (Luong et al., 2015) went on to use a naive-bayes classifier with features from contextual noun-phrases. WHATELLES (Callin et al., 2015) used a neural network approach with features from preceding noun-phrases.

It is to be noted that the last year's task was won by a language model baseline, provided by the organizers. Our system fits the second category of systems, those relying on the context to predict the pronoun. None of the systems participating in the shared task seem to be using explicit sequence classification approaches.

3 Network

3.1 Architecture

Our system is a deep recurrent neural network model with learned token-level embeddings, two layers of Gated Recurrent Units (GRUs), a dense network layer with rectified linear unit (ReLU) activation, and a softmax layer. Our network architecture is described in Figure 2.

The first layer on the bottom of the Figure 2 illustrates how source and target contexts are read. On the target side the context is read in the left and right direction starting from the `replace` token, and the `replace` token itself is not included in the context window. As the training data includes word-level alignments between the source and target language, we are able to identify the source language counterpart for the missing pronoun. This pronoun is used as a starting point for source context reading to both left and right direction the same way as in the target side. However, in the source side the aligned pronoun is always included in both context windows. If the `replace` token is aligned to multiple source side words (the pronoun to be translated can be considered as a multi-word expression), reading the right-side context always starts from the left-most alignment, and vice versa.

Starting from the input of the network, our system has five sets of 90-dimensional embedding matrices; embeddings for source language words, separate embeddings for the target language lemmas, part-of-speech tags and combination of lemmas and part-of-speech tags. In addition we have separate embeddings for source language pronouns aligned with the unknown target pronoun. Context windows are then sequences of indices for these different token-level embeddings, except the aligned source language pro-

noun, which is always just one index as the tokens are concatenated if the alignments refers to multiple source language words. Thus, the network has a total of nine inputs, two different directions for each set of context embeddings, and the aligned source language pronoun. As we do not use external data sources, these embeddings are randomly initialized.

Once the sequence of context words are turned into embeddings, they are given to the first layer of GRUs, which output is given as a sequence to the second layer of GRUs. The second GRU layer then reads the input sequence and outputs the last vector produced, i.e. a fixed-length representation of the input sequence. In all GRUs we use 90-dimensional internal representation.

All these eight products of the recurrent layers, are concatenated together with the embedding for the aligned source language pronoun and given to a 256-dimensional dense neural network layer, with ReLU activation function[1]. This vector is then fed to a layer with softmax activation and an output for each possible output pronoun to make the final prediction.

While our model relies on learned embeddings instead of predefined set of features, a process similar to feature engineering takes place while designing the system architecture. The design choices were made in a greedy manner and mostly the system was built additively, testing new features and adding the promising ones to the final system. Since not all design choice combinations were properly tested during the system development, we include a short evaluation of different settings in Section 4.1.

3.2 Training the system

Only the training data provided by the shared task organizers is used to train our system. The data is based on three different datasets, the Europarl dataset (Koehn, 2005), news commentary corpora (IWSLT15, NCv9), and the TED corpus[2]. We used the whole TED corpus only as development data, and thus our submitted systems are trained on the union of Europarl and news commentary texts, which are randomly shuffled on document level. The total size of training data for each source–target pair is approximately 2.4M sentences, having 590K–760K training examples depending on the pair. The vocabulary sizes, when training with the full training data are listed in Table 1. The large number of aligned pronouns for French-English and English-French language pairs is because of the alignments for the pronoun were often multiple token in length.

In previous studies using only in-domain data has provided competitive performance (Tiedemann, 2015; Callin et al., 2015), and as Europarl can be seen as out-of-the-domain data, in Section 4 we compare the performance of our system when trained using only in-domain data.

Since the main metric in the official evaluation is macro recall, our primary submission is trained to optimize that. This is done by weighting the loss of the training examples relative to the frequencies of the classes, so that misclassifying a rare class is seen by the network as more serious mistake than misclassifying a common class. This scheme produces outputs with more emphasis on rare classes, rather than going after the most common ones. The contrastive submission is trained in the standard way, where each example is seen as equal.

In both our submissions exactly the same system architecture is used for all four language pairs, and no language-dependent optimization was carried out. However, the number of epochs used in training differs, and the prediction performance on the development set was used to decide the optimal number of epochs for each language pair.

The system was implemented in Keras (Chollet, 2015), and trained and developed on the CSC cluster[3] of NVidia Tesla 40KT GPUs. Only one GPU was used to train a single network. Depending on the settings of the network and training data size a single training epoch took 25 minutes to an hour, and all networks were trained in 9 hours. Usually the performance of the network peaked within the first 5 training epochs when evaluated on the development set, and most often reached performance very close to the maximum within three training epochs. All networks were evaluated on the development set after each training epoch, and the model with the highest macro recall was selected for evaluation.

The practical, time-wise, predictive performance of our system is reasonable and doesn't require the use of a GPU. Predicting a test set for an individual language pair takes on a 6-core Intel

[1] Dense layer with tanh activation was also tested, but ReLU turned out to give better results.

[2] http://www.ted.com

[3] www.csc.fi

	Target POS	Target Word	Target Word-POS	Source Word	Aligned Pronouns	Pronouns
de-en	15	170,484	181,531	539,980	9	9
en-de	15	446,645	454,175	198,244	6	5
fr-en	34	171,633	182,763	220,204	18,960	8
en-fr	39	158,755	179,299	199,774	7174	8

Table 1: The vocabulary sizes of the models

	De-En		**En-De**		**Fr-En**		**En-Fr**	
Architecture	Macro R	Micro F	Macro R	Micro F	Macro R	Micro F	Macro R	Micro F
primary	**73.91**	75.36	64.41	71.54	**72.03**	80.79	65.70	70.51
no stacking	65.63	75.98	61.84	73.37	68.84	77.74	70.00	74.26
only in-domain	59.18	75.36	50.72	66.06	57.80	74.09	58.09	65.15
short context	61.29	73.50	**65.66**	71.80	65.84	79.59	69.27	70.51
cross-sentence	60.76	70.81	46.91	49.61	60.46	78.05	61.33	69.17
contrastive	72.60	**80.54**	58.39	72.85	66.54	**85.06**	61.46	72.39
no stacking	65.35	79.30	59.71	**76.76**	61.23	81.71	**70.88**	**77.75**

Table 2: Macro recall and micro F-score for all our system combinations evaluated on the test set. In the **primary** section, the systems are trained to optimize macro recall, and in the **contrastive** section, the systems are optimized without preference towards rare classes. In **no stacking**, only one layer of GRUs is used. **Only in-domain** refers to a version where the Europarl data was not used in training, and **short-context** refers to a version in which the context window was set to 5. **Cross-sentence** refers to a version where the context was expanded also beyond the current sentence.

Xeon CPU 1m 55s, of which 9 seconds is used for prediction and the rest for loading model weights and building the network.

4 Results

In the official test evaluation results our primary system has the best score across all language pairs (see Table 3). In two language pairs, German–English and English–French, we have a modest improvement over the second best system. However, in the other two language pairs, the margin is substantial, 11.9pp for the English–German pair and 6.4pp for the French–English pair. When we look closer into class frequencies and system predictions, it can be seen that in these two pairs our system benefits especially much from predicting small classes relatively well.

In our primary submission, the system was optimized towards macro-averaged recall whereas in our contrastive submission standard training metrics were used. Therefore the prediction accuracy is better in our contrastive submission than it is in the primary submission by 1.3pp–5.2pp depending on the language pair, but at the same time macro recall decreases by 1.3pp–6.0pp. Yet, in the two language pairs with a wide margin to other teams, our contrastive system still achieves better macro recall than any other system. For per-language scores for both our submissions, see rows *TurkuNLP* for primary and *TurkuNLP cont* for contrastive in Table 3.

4.1 Feature evaluation

We ran a small study of different system settings to evaluate our design choices. Results are shown in Table 2, where the performance is evaluated on the official test set. In the test set evaluation our primary system gives the highest score on two language pairs, but loses to another system setting in other two language pairs. Overall, the primary system still performs best on average when measured on macro recall.

As stated in Section 3.1, both our submissions are based on a version of the network with stacked GRU units. In preliminary studies, the stacked approach increased the prediction performance and this holds on the test set for all language pairs except English-French. While on average the stacked system performs 2.4pp better on macro recall, on the English-French pair the non-stacked model performs 4.3pp better.

Another important feature is the size of the context window. In previous work a rather small context was noted to work relatively well (Tiedemann,

System	Macro Recall			
	De–En	En–De	Fr–En	En-Fr
TurkuNLP	**73.91**	**64.41**	**72.03**	**65.70**
TurkuNLP cont	72.60	58.39	66.54	61.46
UKYOTO	73.17*	52.50*	65.63*	62.44
limsi				59.32
UHELSINKI	69.76	44.69	62.98	57.50
UU-Hardmeier		50.36		60.63
uedin		48.72		61.62
UUPSALA	59.56	47.43	62.65	48.92
UU-Stymne	59.28	52.12	36.44	65.35*
baseline–x	44.52	47.86	42.96	50.85
CUNI	60.42	28.26		
UU-Cap		41.61		
baseline–0	42.15	38.53	38.38	46.98
Idiap				36.36

Table 3: Scores for all primary systems and our contrastive system on the official test set evaluation sorted by the average score across language pairs. For each language pair the best score is bolded and the second best is marked with a star (our contrastive submission is not taken into account).

2015; Callin et al., 2015). However, in our submission systems the maximum size of the context was set to 50, and in our development experiments radically shorter context sizes hurt the prediction performance of our system. However, in test set evaluation both language pairs with English as the source language seem to benefit from shorter context, especially English-French pair which scores 3.6pp higher in macro recall than our primary system, but also loses to the version with longer context without stacking by 0.73pp in macro recall. Other language pairs benefit from larger context (see *short context* in Table 2).

In addition, we evaluate allowing the context window to extend beyond the current sentence boundary. The maximum context size is always 50, although when restricted to within one sentence, it naturally rarely reaches it. In our primary and contrastive submissions, the context was limited to include only the current sentence, and the results using the context beyond the sentence are in the row *cross-sentence* in Table 2. We can observe that no language pair seems to benefit from a larger context on the test set.

As mentioned earlier, the Europarl dataset can be considered as out-of-the-domain data. The *in-domain* row in Table 2 refers to an experiment where Europarl was discarded from the training data and thus the system was trained only on in-domain data. Naturally, the amount of training data is then much smaller, the data size drops from 2.4M sentences to approx. 400K sentences. This hurts the performance on all language pairs, indicating that our method benefits from a lot of training data and might be indicative of its ability to generalize to other domains.

5 Conclusion

In this paper we presented our system for the cross-lingual pronoun prediction shared task. Our system is based on recurrent neural networks and token-level embeddings of the source and target languages, and is trained without any external data. Our system fared well in the shared task, having the highest macro recall in all language pairs. Our results suggest sequence classification and recurrent neural networks to be an approach worthy of consideration when tackling the problem. It is also worth noting that our system is wholly language-agnostic and demonstrates that an approach with very little custom-built features can have a good performance on the task.

As the system is trained only using the official training data without any external tools, it would be interesting to test whether pre-trained token-level embeddings would increase its performance. Additionally, pre-training the network with monolingual data could be considered.

Our system is openly available at `https://github.com/TurkuNLP/smt-pronouns`.

Acknowledgments

This work was supported by the Kone Foundation. Computational resources were provided by CSC – IT Center for Science.

References

Jimmy Callin, Christian Hardmeier, and Jörg Tiedemann. 2015. Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64. Association for Computational Linguistics.

Franois Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

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)*. 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. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Ngoc-Quang Luong, Lesly Miculicich Werlen, and Andrei Popescu-Belis. 2015. Pronoun translation and prediction with or without coreference links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 95–100. Association for Computational Linguistics.

Ngoc-Quan Pham and Lonneke van der Plas. 2015. Predicting pronouns across languages with continuous word spaces. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 101–107. Association for Computational Linguistics.

Jörg Tiedemann. 2015. Baseline models for pronoun prediction and pronoun-aware translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 108–114. Association for Computational Linguistics.

Dominikus Wetzel, Adam Lopez, and Bonnie Webber. 2015. A maximum entropy classifier for crosslingual pronoun prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 115–121. Association for Computational Linguistics.