# Cross-Lingual Pronoun Prediction with Deep Recurrent Neural Networks v2.0

**Juhani Luotolahti**[1,2] **Jenna Kanerva**[1,2] **and Filip Ginter**[1]
[1]TurkuNLP Group, 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 system in the DiscoMT 2017 Shared Task on Cross-lingual Pronoun Prediction. Our entry builds on our last year's success, our system based on deep recurrent neural networks outperformed all the other systems with a clear margin. This year we investigate whether different pre-trained word embeddings can be used to improve the neural systems, and whether the recently published Gated Convolutions outperform the Gated Recurrent Units used last year.

## 1 Introduction

The DiscoMT 2017 Shared Task on Cross-lingual Pronoun Prediction (Loáiciga et al., 2017) concentrates on the difficult task of translating pronouns between languages. For example different gender marking between languages complicates the translation process. This shared task includes three languages and four translation directions: English-French, English-German, German-English and Spanish-English. In the target language side selected set of pronouns are substituted with `replace` token, 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, which makes even harder to model the target language. An example of an English-French sentence pair is given in Figure 1.

In this paper we describe the pronoun prediction system of the Turku NLP Group. Our system extends the last year's deep recurrent neural networks based system with word-level embeddings, two layers of Gated Recurrent Units (GRUs) and a softmax layer on top of it to make the final prediction (Luotolahti et al., 2016). This year we investigate whether pre-trained word embeddings improve the system performance compared to the random initialization used in the previous system. We also study whether the recently published Gated Convolution outperforms Gated Recurrent Units.

***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.

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. Also our pre-trained word embeddings are trained on the same data.

## 2 System Architecture

As in the previous year, our system is a deep neural network model reading context from both source and target side sentences around the focus pronoun. The most important change are the token-level embeddings, which are now pre-trained before training the full system. The system architecture itself is improved relative to the last year system by filtering from the data aligned pronouns that are too long, as these are alignment errors rather than actual pronouns. We also increase the size of the last dense neural network layer from 320 to 720 units, to address a possible bottleneck caused by excessive data compression. We also experiment with changing the basic network units from Gated Recurrent Units to Gated Convolutions. Otherwise the network and parameters are exactly the same, and are only shortly explained



*Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 63–66,
Copenhagen, Denmark, September 8, 2017. 

here. More information is provided in Luotolahti et al. (2016).

In both source and target side the context is read separately in left and right directions starting from the focus pronoun[1] or the `replace` token, so that the source side pronoun is always included in both right and left contexts, but the special `replace` token in the target side is not, as it does not provide any useful information. All words in the contexts are embedded and pushed through the layers of either GRU or Gated Convolutions, finally concatenating the vectors, along with the embedding vector for the aligned pronoun, for the last softmax layer, which makes the final prediction.

The systems tested can be divided into three categories, those with pre-trained embeddings, those using GRU as the basic network unit and those using convolutional neural networks as the basic unit. All systems were tested on the dev-set and the best two were chosen for submission. All systems use the same input data, basic structure of the system, and features. The context used by the systems is restricted to a single sentence, as this provided the best results last year and in preliminary experiments we were unable to obtain a consistent gain by expanding the context.

The systems using GRU as the basic network unit are listed in Table 1 as GRU, GRU_dropouts, GRU_Pronoun_Context, Mixed_Context and GRU_Word2Vec. Of these systems, GRU uses randomly initialized embeddings and is essentially our last year's system. GRU_dropouts is identical to the former system, but has dropouts of 0.5 added after every GRU layer to possibly improve generalization of the system. The three latter systems, GRU_Pronoun_Context, GRU_Mixed_Context and GRU_Word2Vec, all have identical architecture to the GRU system, but use pre-trained embeddings. The architecture of these systems is depicted in Figure 2.

Systems GatedConv_1, GatedConv_2 and GatedConv_Mixed_Context use all convolutional neural networks as their basic unit. Of these the last, GatedConv_Mixed_Context, uses the same pre-trained embeddings as the Mixed_Context system. All of these systems use stacked gated convolutional layers as a replacement to stacked GRUs. Gated convolutional networks have lately been demonstrated to offer comparable performance to recurrent neural networks (Dauphin et al., 2016). GatedConv_1 uses two layers of gated linear units and both GatedConv_2 and GatedConv_Mixed_Context use four layers, all convolutional systems use convolution width of 10 and 90 units. For more details on the gated convolutional architectures, refer to Dauphin et al. (2016). The architecture of the network for convolutional systems is identical to the GRU ones, except we have replaced GRU layers with convolutional layers. The convolutional layers are gated, in practice we the output of a gated convolutional layer is an elementwise product between a linear convolutional layer and a convolutional layer with sigmoid activation function, both convolutional layers receiving the same input.

[1] 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.

### 2.1 Word Embeddings

Word embeddings are trained on the official training data provided by the organizers having approximately 60 million words per language, which is relatively small for training regular word2vec (Mikolov et al., 2013) style word embeddings. In addition to the regular word2vec embeddings we train two alternative word embedding models with the training task geared towards this particular pronoun prediction task. Firstly, instead of a sliding window of words we define the context for a source word to be all pronouns in the counterpart target sentence. In other words, instead of predicting nearby words, we modify word2vec to predict target sentence pronouns. This way, similar embeddings are given to source-side words which associate with similar pronouns on the target side, which we expect to be a good pre-training strategy for pronoun prediction. This pretraining method we refer to as the *pronoun context*. Secondly, we extend the pronoun context method with the standard skip-gram context, i.e. predicting all target sentence pronouns as well as words nearby in the linear order. Since the shared task training data includes also word alignments, we use a union of skip-gram contexts on the source side and the target side. Therefore, in this *mixed context* method, for every source word, word2vec is used to predict the target sentence pronouns, the source sentence context words, and the target sentence context lemmas.

The word embeddings are trained using

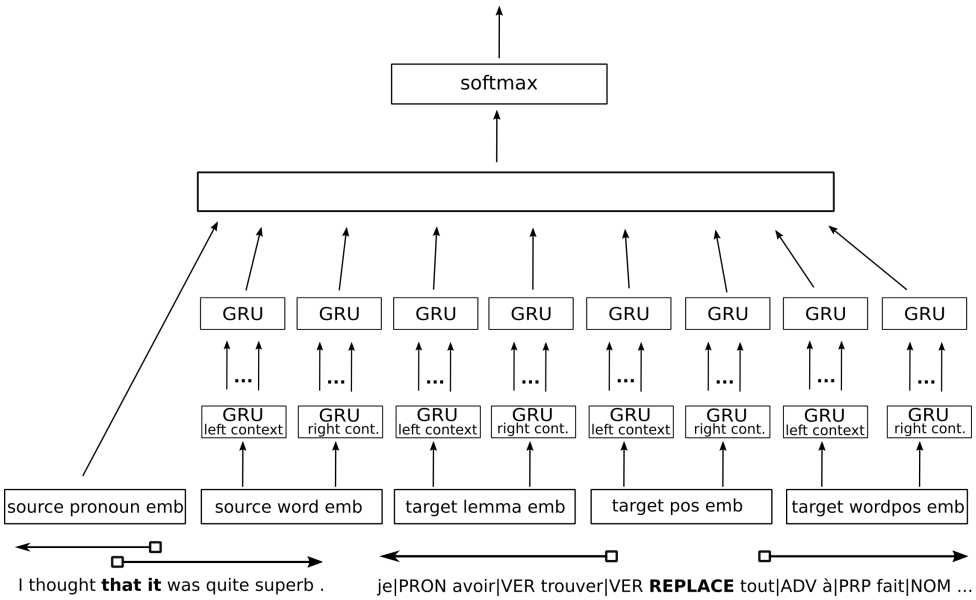


Figure 2: GRU architecture

word2vec[2] and word2vecf[3] softwares by Mikolov et al. (2013) and Levy and Goldberg (2014) respectively, the latter supporting arbitrary contexts for word2vec style embedding learning. All embeddings are trained using the full training data, i.e. also sentences without training examples for the pronoun prediction task and no other data is used. All word embeddings use 90-dimensional vectors, and are trained using the skip-gram architecture with negative sampling and 10 training iterations.

### 2.2 Data and Training

The training data provided by the organizers is based on three different datasets, the Europarl dataset (Koehn, 2005), news commentary corpora (IWSLT15, NCv9), and the TED corpus[4]. We used the whole TED corpus only as development data, and thus our submitted systems and word embeddings are trained on the union of Europarl and news commentary texts, which are randomly shuffled. The total size of training data for each source–target pair is approximately 2.2–2.4 million sentences, having 590K–800K training examples depending on the pair.

Since the main metric in the official evaluation is macro recall, our submission is trained to optimize this metric. This is achieved by weighting the loss of the training examples inversely proportional to the frequencies of the classes, so that misclassifying a rare class is a more serious error than misclassifying a common class. This scheme produces outputs with a higher emphasis on rare classes. This scheme yielded very good results last year, giving more than 4 percent point improvement on average.

Exactly the same system architecture is used for all four language pairs, and no language-dependent optimization was carried out. This makes our system fully language-agnostic. The only difference is the number of epochs used in training, set for each language pair separately using the prediction performance on the development set.

## 3 Results

Table 1 shows our system variants evaluated on the test data. In general, the recurrent systems seem to be performing better than the convolutional systems. However, since due to time restrictions we were unable to perform a specialized hyper-parameter search on any of the systems, only tentative conclusions can be made. Further, all systems seem to generally benefit from the pre-trained input vectors, with the excpeption of plain word2vec. Pre-trained embeddings with context which includes pronoun information perform better than plain word2vec pre-training and random initialization. Adding dropouts also improved performance on the test set, which was not

[2] https://github.com/tmikolov/word2vec
[3] https://github.com/BIU-NLP/word2vecf
[4] http://www.ted.com

| | En-De | De-En | En-Fr | Es-En | Average | Rank |
|---|---|---|---|---|---|---|
| GRU | 52.22 | 56.79 | 53.65 | 45.51 | 52.22 | 7 |
| GRU_dropouts | 49.44 | 64.25 | 56.05 | 54.63 | 56.09 | 5 |
| GRU_Pronoun_Context | 61.66 | 69.21 | 64.74 | 58.78 | 63.60 | 2 |
| GRU_Mixed_Context | **68.95** | 68.88 | **66.89** | **58.82** | **65.89** | **1** |
| GRU_Word2Vec | 42.91 | 45.98 | 48.49 | 49.67 | 46.76 | 8 |
| GatedConv_1 | 43.57 | 59.22 | 60.37 | 52.29 | 53.86 | 6 |
| GatedConv_2 | 45.77 | 69.35 | 58.02 | 52.4 | 56.39 | 4 |
| GatedConv_Mixed_Context | 46.64 | **68.91** | 61.53 | 58.78 | 58.97 | 3 |

Table 1: Test set results of the variants of the system tested against the test sets.

visible in the development set results.

It is to be noted that the systems performed worse on the test data than the development data, indicating overfitting to the development data, but their relative strength remained roughly the same with all top three systems utilizing embedding pretraining based on the task, with the only exception being that system with dropouts performed better than without, which is fitting because dropouts should reduce overfitting. Also, surprisingly word2vec embedding initialization performed worse than random initialization.

Compared to systems submitted for the task, our system performed fairly well. For language pairs German − English and English − French our systems, when measured with macro recall, the official task metric, our system received the best scores among the submitted systems, and for language pair Spanish - English second best scores by 0.05 percent points. This is in contrast to language pair English - German in which our system received second best score, but the difference to the winning system is almost 10 percent points.

## 4 Conclusions

In this paper we presented our improved system for cross-lingual pronoun prediction shared task. We included pre-trained word embeddings as well as evaluated the performance of Gated Convolutions compared to Gated Recurrent Units as basic units of our deep network. On the development set, we found that the Gated Recurrent Units outperform the Gated Convolution and that pretraining the embeddings in a task-specific fashion outperforms the vanilla word2vec method.

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

## Acknowledgments

This work was supported by the Kone Foundation and the Finnish Academy. Computational resources were provided by CSC – IT Center for Science, Finland.

## References

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2016. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083* .

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

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*.

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.

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2016. Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, pages 596–601. http://www.aclweb.org/anthology/W/W16/W16-2353.pdf.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.