# POS Driven Cross-Lingual Pronoun Prediction with Feed-Forward Neural Networks

**Jimmy Callin**
Uppsala University
jimmy.callin@gmail.com

## Abstract

For some language pairs, pronoun translation is a discourse driven task. For cross-lingual pronoun prediction, we suggest a neural network based model using preceding nouns and determiners as features for suggesting antecedent candidates. Our model scores on par with similar models while having a simpler architecture.

## 1 Introduction

Translation requires context. Most modern statistical machine translation (SMT) systems do use context for translation; the meaning of a word is more often than not ambiguous, and can only be decoded through its usage. That said, context in modern SMT still mostly assumes that sentences are independent of each other, and sentence dependencies are simply ignored. While today's popular SMT systems could and in some cases do use source language features from previous sentences, previously translated sentences within a document has up to this point rarely been included. Hardmeier and Federico (2010) argue that SMT research has become mature enough to stop assuming sentence independence, and start to incorporate features beyond the sentence boundary. They demonstrate the difficulty of correctly determining the correct German pronoun given English as source language. While translating pronouns indeed can be unproblematic for certain language pairs, languages with gender-marked pronouns introduce certain difficulties, since the choice of pronoun is determined upon the gender of its antecedent. Picking the wrong third-person pronoun might seem like a relatively minor error, especially if present in an otherwise comprehensible translation, but could potentially produce misunderstandings. Take the following English sentences:

– The monkey ate the banana because *it* was hungry.
– The monkey ate the banana because *it* was ripe.
– The monkey ate the banana because *it* was tea-time.

*It* in each of these three cases reference something different, either the monkey, the banana, or the abstract notion of time. If we were to translate these sentences to German, we would have to consciously make decisions whether *it* should be in masculine (*der*, referring to the monkey), feminine (*die*, referring to the banana), or neuter (*das*, referring to the time) (Mitkov et al. 1995). While these examples use a local dependency, the antecedent of *it* could just as easily have been one or several sentences away and its translation would have been out of reach for sentence based SMT decoders.

## 2 Related work

Most of the work in anaphora resolution for machine translation has been done in the paradigm of rule-based MT, while the topic has gained little interest within SMT (Hardmeier and Federico 2010; Mitkov 1999). One of the first examples of using discourse analysis for pronoun translation in SMT was done by Le Nagard et al. (2010), where they use co-reference resolution to predict the antecedents in the source language as features in a standard SMT system. While they saw score improvements in pronoun prediction, the bad performance of the co-reference resolution seriously impacts the results negatively. They implement this in a two-step solution by running the co-reference resolver separately from the SMT decoder, and thereafter annotate the translated French pronouns with gender and quantity based on their predicted antecedent from which they once again translate using another MT system where phrase tables have been annotated in the same fashion. This

seems to be mostly for practical reasons, since most popular SMT frameworks such as Moses do not provide previous target translations for use as features (Koehn et al. 2007). Guillou (2012) tries a similar approach for English-Czech translation with little improvement even after factoring out major sources of error. They single out one possible reason for this, which is how a reasonable translation alternative of a pronoun's antecedent could affect the predicted pronoun, including the possibility of simply canceling out pronouns. E.g, *the u.s. , claiming some success in its trade* could be paraphrased as *the u.s. , claiming some success in trade diplomacy* without any loss in translation quality, while still affecting the score negatively.

In reality, there is necessary linguistic information in the target translation that simply is not available in the source. An English antecedent might in most cases be translated to a word with male gender, which is learned by the decoder. This means that in cases where the antecedent produce an alternative translation with a non-male gender, the decoder has no information available to condition upon this. It could potentially be possible to include a larger source context window around the source antecedent in hope of catching contexts where alternative translations are made, but in our opinion this would likely only increase data sparseness and overall be a non-optimal solution. Rather, you would like to have the translated antecedent available at decoding, which would mean that the SMT decoder in contrast to most decoders today should be able to handle dependencies across sentence boundaries. Hardmeier and Federico (2010) extends the phrase-based Moses decoder with a word dependency model based on existing co-reference resolution systems, by parsing the output of the decoder and catching its previous translations. Unfortunately they only produce minor improvements for English-German, and negative results for English-French (Hardmeier, Tiedemann, Saers, et al. 2011).

In light of this, there have been attempts at considering pronoun translation as a classification task separate from traditional machine translation. This could could potentially lead to further insights into the nature of anaphora resolution. In this fashion a pronoun translation module could potentially be treated as just another part of translation by discourse oriented machine translation systems, or as a post-processing step similarly to Guillou (2012). Hardmeier, Tiedemann, and Nivre (2013) introduce this task and present a feed-forward neural network model using features from an external anaphora resolution system, BART (Broscheit et al. 2010), to infer the pronoun's antecedent candidates and use the aligned words in the target translation as input. This model is later integrated into their document-level decoder Docent (Hardmeier 2014; Hardmeier, Stymne, et al. 2013).

## 3 Task setup

The goal of this task is to accurately predict the correct pronoun in a French translation. The pronouns in focus are *she*, *he*, *it*, and *they*, where the word aligned phrases in the French translation have been replaced by placeholders. The word alignment is included, and was automatically produced by GIZA++ (Och 2003). We are also aware of document boundaries within the corpus. The corpus is a set of three different parallel corpora gathered from three separate domains: transcribed TED talks, Europarl (Koehn 2005) with transcribed proceedings from the European parliament, and a set of news text. Test data is a collection of transcribed TED talks, in total 12 documents containing 2093 sentences with a total of 1105 classification problems, with a similar development set.

## 4 Method

Inspired by the the neural network architecture set up in Hardmeier, Tiedemann, and Nivre (2013), we similarly propose a feed-forward neural network with a layer of word embeddings as well as an additional hidden layer for learning abstract feature representations. The model is trained using stochastic gradient descent with mini-batches and L2 regularization. Cross-entropy is used as a cost function, with a softmax output layer. The main difference in our model lies in avoiding using antecedent features as gathered by an external anaphora resolution system. Rather, to simplify the model we try to simply look at the four closest previous nouns and determiners in English, and use the corresponding aligned French nouns and articles in the model, as illustrated in figure 1. Wherever the alignments map to more than one word, only the furthermost left of the words in the phrase is used.

Additionally, we allow ourselves to look at the


Figure 1: An English POS tagger is used to find nouns and articles in preceding utterances, while the word alignments determine which French words are to be used as features.

French context of the missing pronoun. While this restricts the potential usage of our model as a part of the translation process, French usage should be a better indicator for some of the classes, e.g. *ce* which is highly dependent being precedent of *est*. See figure 2 for an example of context in source and translation as features.

**<S> <S> <S> it expresses our view** of how we…

**<S> <S> <S> __ exprime notre manière** d' aborder …

Figure 2: Example of context used in the classification model, color coded according to their position in the neural network as illustrated in figure 3.

Furthermore the dimensionality of the embeddings are increased from 20 to 50, since this was empirically shown to somewhat improve the scores on the development set, with a hidden layer size of 50. To reduce training time and faster find convergence, we use $\tanh$ as activation function between the hidden layers, in contrast to the sigmoid function in the original paper (LeCun et al. 2012). To avoid overfitting, early stopping is introduced where the training stops if no improvements have been found within a number of iterations. This usually results in a training time of 130 epochs, when run on TED data from IWSLT.

The model uses a uniform random weight initialization according to Glorot et al. (2010), where they show that neural network models using $\tanh$ as activation function generally perform better with an initialization within the interval $uniform[-\frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}, \frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}]$, where $fan_{in}$ and $fan_{out}$ are number of inputs and number of hidden units respectively.

Since the model uses a fixed context window size for English and French, as well as a fixed number of preceding nouns and articles, we need to find out the optimal parameter settings. We empirically find that a parameter setting of 4+4 context window for English and French, with 3 preceding nouns and articles each perform well. Figure 4 showcases how window size and number of preceding POS tags affect the performance outcome on the development set. We also looked into asymmetric window sizes, but according to figure 5 we noticed no improvements.

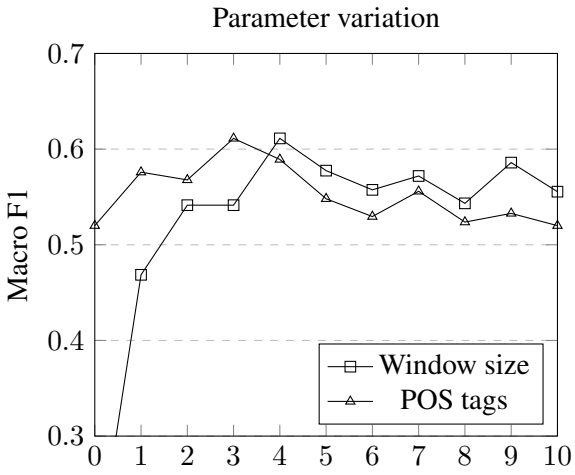


Figure 4: Parameter variation of window size and number of preceding POS tags. Window size is varied in a symmetrical fashion of n+n. When varying window size, 3 preceding POS tags are used. When varying number of POS tags, a window size of 4+4 is used.

The final architecture as shown in figure 3 was implemented in Theano (Bergstra et al. 2010), and is publicly available on Github[1].

## 5 Results

The results from the shared task are presented in table 1 and table 2. Best performing classes are *ce*, *ils*, and *other*, while the less commonly occurring classes *elle* and *elles* have precision comparable to other classes. Their recall, though, is significantly worse. The overall macro F1 score ends up being 55.3%.

## 6 Discussion

Results indicate that the model performs on par with previously suggested models, while having a simpler architecture. Classes highly dependent on local context, such as *ce*, perform especially well, which is likely due to *est* being a good indicator for its presence. This is supported by the large performance gains from 4+0 to 4+1 in figure 5,

[1] http://github.com/jimmycallin/whatelles

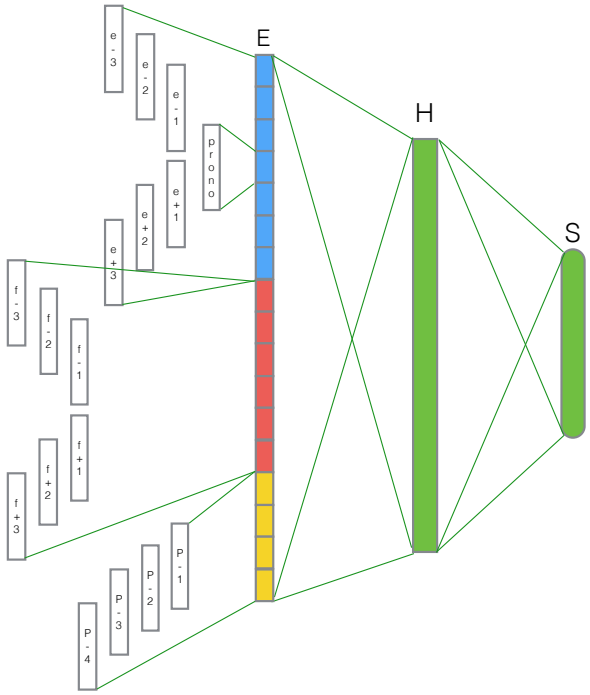


Figure 3: Neural network architecture. Blue embeddings (E) signifies source context, red target context, and yellow the preceding POS tags. The shown number of features is not equivalent with what is used in the final model.

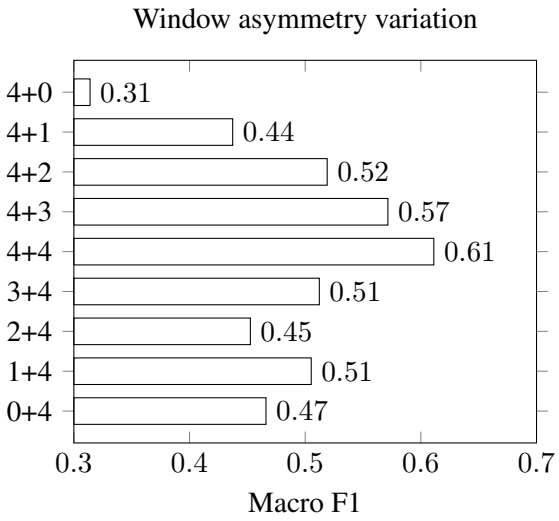


Figure 5: Parameter variation of window size asymmetry, where each label corresponds to *n*+*n*, where *n* is the context size in each direction.

| | Precision | Recall | F1 |
|---|---|---|---|
| ce | 0.8291 | 0.8967 | 0.8616 |
| cela | 0.7143 | 0.6202 | 0.6639 |
| elle | 0.5000 | 0.2651 | 0.3465 |
| elles | 0.6296 | 0.3333 | 0.4359 |
| il | 0.5161 | 0.6154 | 0.5614 |
| ils | 0.7487 | 0.9312 | 0.8301 |
| other | 0.8450 | 0.8579 | 0.8514 |
| Macro | 0.5816 | 0.5495 | 0.5530 |
| Micro | 0.7213 | 0.7213 | 0.7213 |

Table 1: Precision, recall, and F1-score for all classes. Micro score is the overall classification score, while macro is the average over each class. The latter scoring method is used for increasing the importance of classes with fewer instances.

since *est* usually follows *ce*. Singular and plural classes rarely get mixed up, probably due to its condition upon the English pronoun which marks *it* or *they*, making this prediction clear. The classes of feminine gender does not perform as well, especially not recall-wise, although this was to be expected since the only real information from which to infer its antecedent is distance from the pronoun in focus. It is apparent that the model has a bias towards making majority class predictions, especially given the low number of wrong predictions on the *elle* and *elles* classes relative to *il* and *ils*. The high recall of *ils* is explained by this phenomenon as well. An additional hypothesis is that it is simply too little data to realistically create usable embeddings, except for in a few reoccurring

| | ce | cela | elle | elles | il | ils | other | sum |
|---|---|---|---|---|---|---|---|---|
| ce | 165 | 3 | 0 | 1 | 8 | 1 | 6 | 184 |
| cela | 5 | 80 | 4 | 1 | 21 | 0 | 18 | 129 |
| elle | 7 | 10 | 22 | 2 | 22 | 2 | 18 | 83 |
| elles | 0 | 0 | 0 | 18 | 0 | 31 | 3 | 51 |
| il | 11 | 7 | 9 | 0 | 64 | 1 | 12 | 104 |
| ils | 1 | 0 | 0 | 5 | 0 | 149 | 5 | 160 |
| other | 10 | 12 | 9 | 1 | 9 | 15 | 338 | 394 |
| sum | 199 | 112 | 44 | 27 | 124 | 199 | 400 | |

Table 2: Confusion matrix of class predictions. Row signifies actual class according to gold standard, while column represents predicted class according to the classifier.

circumstances.

The extra number of features as well as the increase in embedding dimensionality makes the training and prediction somewhat slower, but since the training still is done in less than an hour, and predicting the test data does not take more than a few seconds, it is still good enough for general usage. Furthermore, the implementation is made in such a way that further performance increases are to be expected if you run it on CUDA compatible GPU with minor changes.

While three separate training data collections were available, we only found interesting results when using data from the same domain as the test data, i.e. transcribed TED talks. To overcome the skew class distribution, attempts were made at oversampling the less frequent classes from Europarl, but unfortunately this only led to performance loss on the development set. The model does not seem to generalize well from other types of training data such as Europarl or news text, despite Europarl being transcribed speech as well. This is an obvious shortcoming of the model.

We tried several alterations in parameter settings for context window and POS tags, and there were no significant improvements found beyond the final parameter settings when run on the development set, as seen in figure 4. Figure 5 makes it clear that a symmetric window size is beneficial, while the reason for why is not made as clear. It does seem like right context is more important than left context, which could be due to pronouns in their role as subjects largely appears early in sentences, making left context nothing but sentence start markers. In future work, it would be interesting to look into how much source context actually contribute to the classification, given a target context. The English context is indeed nice to have, since you cannot be entirely certain of the translation quality in the target language, but intuitively all necessary linguistic information should only be available in the target language. If source language were to be used, each English word embedding could perhaps be pre-trained on a large number of translation examples hopefully learning the most probable cross-linguistic gender. Gender aware French word embeddings would hypothetically increase the score as well, if not more.

## 7 Conclusion

In this work, we have been developing a cross-lingual pronoun prediction classifier based on a feed-forward neural network. The model was heavily inspired by (Hardmeier, Tiedemann, and Nivre 2013), while trying to simplify the architecture by simply using preceding nouns and determiners for coreference resolution rather than using features from an anaphora extractor such as BART, as set up in the original paper, since these usually includes large performance overhead.

We find out that the model indeed perform on par with similar models, while being easier to train. There are some expected drops in performance for the less common classes heavily dependent on finding the correct anecedent. We discuss probable causes for this, as well as possible solutions using pretrained embeddings on larger amounts of training data.

## References


Bergstra, James et al. (2010). "Theano: a CPU and GPU Math Expression Compiler". In: *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Austin, TX.

Broscheit, Samuel et al. (2010). "BART: A multilingual anaphora resolution system". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 104–107.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *International conference on artificial intelligence and statistics*, pp. 249–256.

Guillou, Liane (2012). "Improving Pronoun Translation for Statistical Machine Translation". In: *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computa-*

*tional Linguistics*. EACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–10.

Hardmeier, Christian (2014). "Discourse in Statistical Machine Translation". In:

Hardmeier, Christian and Marcello Federico (2010). "Modelling Pronominal Anaphora in Statistical Machine Translation". eng. In: pp. 283–289.

Hardmeier, Christian, Sara Stymne, et al. (2013). "Docent: A document-level decoder for phrase-based statistical machine translation". In: *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics); 4-9 August 2013; Sofia, Bulgaria*. Association for Computational Linguistics, pp. 193–198.

Hardmeier, Christian, Jörg Tiedemann, and Joakim Nivre (2013). "Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 380–391.

Hardmeier, Christian, Jörg Tiedemann, Markus Saers, et al. (2011). "The Uppsala-FBK Systems at WMT 2011". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. WMT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 372–378.

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *MT summit*. Vol. 5, pp. 79–86.

Koehn, Philipp 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*. Association for Computational Linguistics, pp. 177–180.

Le Nagard, Ronan and Philipp Koehn (2010). "Aiding Pronoun Translation with Co-reference Resolution". In: *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. WMT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 252–261.

LeCun, Yann A. et al. (2012). "Efficient backprop". In: *Neural networks: Tricks of the trade*. Springer, pp. 9–48.

Mitkov, Ruslan (1999). "Introduction: Special issue on anaphora resolution in Machine Translation and Multilingual NLP". In: *Machine translation* 14.3, pp. 159–161.

Mitkov, Ruslan, Sung-kwon Choi R, and All Sharp (1995). "Anaphora resolution in Machine Translation". In: *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 5–7.

Och, FJ (2003). *Giza++ software*. Internal report, RWTH Aachen University.