Part-of-Speech Driven Cross-Lingual Pronoun Prediction with Feed-Forward Neural Networks

Jimmy Callin, Christian Hardmeier, Jörg Tiedemann
Department of Linguistics and Philology
Uppsala University, Sweden
jimmy.callin.3439@student.uu.se
{christian.hardmeier, jorg.tiedemann}@lingfil.uu.se

Abstract

For some language pairs, pronoun translation is a discourse-driven task which requires information that lies beyond its local context. This sets up the task of trying to predict the correct pronoun given a source sentence and a target translation, where the translated pronouns have been replaced with placeholders. 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

Most modern statistical machine translation (SMT) systems 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 use in modern SMT still mostly assumes that sentences are independent of one another, and dependencies between sentences are simply ignored. While today's popular SMT systems could use features from previous sentences in the source text, translated sentences within a document have up to this point rarely been included.

Hardmeier and Frederico (2010) argue that SMT research has become mature enough to stop assuming sentence independence, and start to incorporate features beyond the sentence boundary. 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 (*er*, referring to the monkey), feminine (*sie*, referring to the banana), or neuter (*es*, 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 which would have made necessary translation features 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 Nagard and Koehn (2010), who 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, they claim the bad performance of the co-reference resolution seriously impacted the results negatively. They performed this as a post-processing step, which seems to be primarily for practical reasons since most popular SMT frameworks such as Moses (Koehn et al., 2007) do not provide previous target translations for use as features. Guillou et al. (2012)

tried a similar approach for English-Czech translation with little improvement even after factoring out major sources of error. They singled 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. This demonstrates there is necessary linguistic information in the target translation that simply is not available in the source. Hardmeier and Frederico (2010) extended 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 produced minor improvements for English-German.

In light of this, there have been attempts at considering pronoun translation a classification task separate from traditional machine translation. This could potentially lead to further insights into the nature of anaphora resolution. In this fashion a pronoun translation module could be treated as just another part of translation by discourse oriented machine translation systems, or as a post-processing step similarly to Guillou et al. (2012). Hardmeier et al. (2013b) introduced this task and presented 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 was later integrated into their document-level decoder Docent (Hardmeier et al., 2013a; Hardmeier, 2014, chapter 9).

3 Task setup

The goal of cross-lingual pronoun prediction is to accurately predict the correct missing pronoun in translated text. The pronouns in focus are *it* and *they*, where the word aligned phrases in the 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 English-French 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 neural network architecture set up in Hardmeier et al. (2013b), 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 final architecture as shown in fig. 1 uses both source context and translation context around the missing pronoun, by encoding a number of word embeddings n words to the left and m words to the right (hereby referred to as having a context window size of n+m).

The main difference in our model lies in avoiding using an external anaphora resolution system to collect antecedent features. 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 fig. 2. Wherever the alignments map to more than one word, only the left-most word in the phrase is used. We encode these nouns and articles as embeddings in the first input layer. This way, the order of each word is embedded which should approximate the distance from the missing pronoun. Additionally, we allow ourselves to look at the French context of the missing pronoun. While the automatically translated context might be too unreliable, French usage should be a better indicator for some of the classes, e.g. *ce* which is highly dependent being precedent of *est*. See fig. 3 for an example of context in source and translation as features.

Similarly to the original model in Hardmeier et al. (2013b), the neural network 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. Furthermore the dimensionality of the embeddings is increased from 20 to 50, since we saw minor improvements of the scores on the development set, with a hidden layer size of 50. To reduce training time and speed up convergence, we use $\tanh$ as activation function between the hidden layers (LeCun et al., 2012), in contrast to the

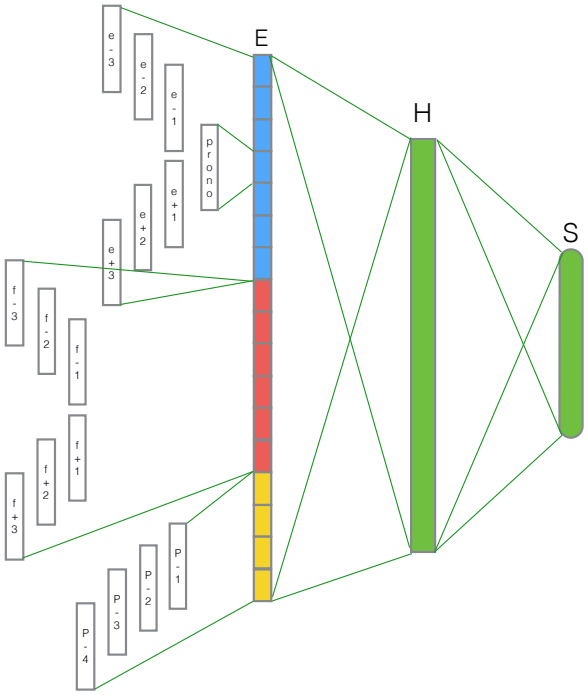

Figure 1: Neural network architecture. Blue embeddings (E) signifies source context, red target context, and yellow the preceding POS tags. The shown number of features is not equivalent with what is used in the final model.

sigmoid function in Hardmeier's model. 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. The model uses a layer-wise uniform random weight initialization as proposed by Glorot and Bengio (2010), where they show that neural network models using $\tanh$ as activation function generally perform better with a uniformally distributed random initialization within the interval $[-\frac{\sqrt{6}}{\sqrt{\text{fan}_{in}+\text{fan}_{out}}}, \frac{\sqrt{6}}{\sqrt{\text{fan}_{in}+\text{fan}_{out}}}]$, 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 optimal parameter settings. We observe 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 noticed no improvements (fig. 5). The neural network is implemented in Theano (Bergstra et al., 2010), and is publicly available on Github.[1]

Figure 2: 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.

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

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

Figure 3: Example of context used in the classification model, color coded according to their position in the neural network as illustrated in fig. 1.

5 Results

The results from the shared task are presented in table 1 and table 2. The best performing classes are *ce*, *ils*, and *other*, all reaching F1 scores over 80 percent. The less commonly occurring classes *elle* and *elles* perform significantly worse, espe-

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

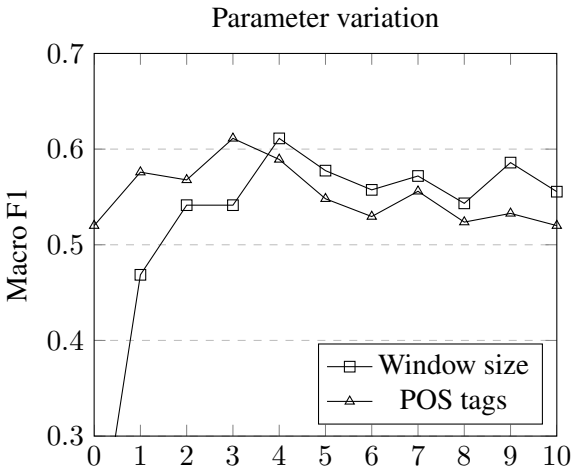

Figure 4: Parameter variation of window size and number of preceding POS tags. Window size is varied in a symmetrical fashion of n+n. When varying window size, 3 preceding POS tags are used. When varying number of POS tags, a window size of 4+4 is used.

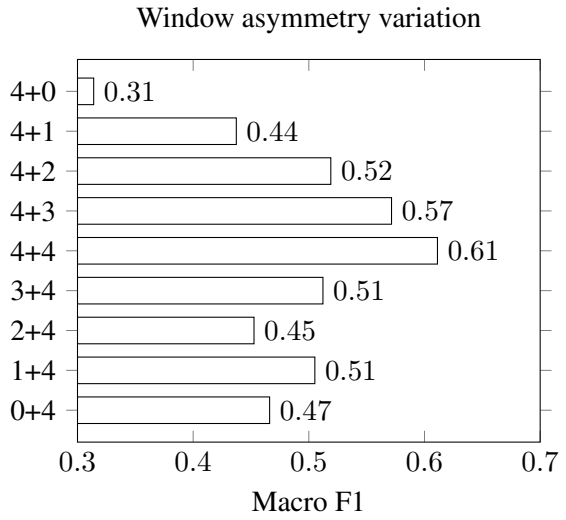

Figure 5: Parameter variation of window size asymmetry, where each label corresponds to *n+n*, where *n* is the context size in each direction.

cially recall-wise. The overall macro F1 score ends up being 55.3%.

6 Discussion

Results indicate that the model performs on par with previously suggested models (Hardmeier et al., 2013b), 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 of its presence. This is supported by the large performance gains from 4+0 to 4+1 in fig. 5, since *est* usually follows *ce*. Singular and plural classes rarely get confused, due to them being predicated on the English pronoun which marks *it* or *they*. The classes of feminine gender does not perform as well, especially recall-wise, although this was to be expected since the only information from which to infer its antecedent is ordered 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 there is simply too little data to realistically create usable embeddings, except for a few reoccurring circumstances.

A somewhat interesting example of what POS tags might cause is:

	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.

... which is the history of who invented games ... and they would be so immersed in playing the dice games ...
... l' histoire de qui a inventé le jeu et pourquoi ...
__ seraient si concentrés sur leur jeu de dés ...

This is one of the few instances where *ils* has been misclassified as *elles*. Since this classification only happens when using at least three preceding POS tags, it is likely there is something happening with the antecedent candidates. The third determiner is *the (history)*, and points to *histoire* which is a noun of feminine gender. It is likely the classifier has learned this connection and has put too much weight into it.

	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.

The extra number of features as well as the increase in embedding dimensionality makes the training and prediction slightly slower, but since the training still is done in less than an hour, and testing does not take longer 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 found no significant improvements beyond the final parameter settings when run on the development set, as seen in fig. 4. Figure 5 makes it clear that a symmetric window size is beneficial, while we are not as sure of why this is the case. Right context seems to be more important than left context, which could be due to the fact that 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 contributes to the classification, given a target context. While the English context is nice to have, since you cannot be entirely certain of the translation quality in the target language, intuitively all necessary linguistic information for inferring the correct pronoun should be available in the target translation. After all, the gender of a pronoun is not dependent on whatever source language you translate from, as long as you have found its antecedent. If the source text still were found useful, all English word embeddings could be pre-trained on a large number of translation examples and through this process learn the most probable cross-linguistic gender. In the same manner, gender aware French word embeddings would hypothetically increase the score as well.

7 Conclusion

In this work, we develop a cross-lingual pronoun prediction classifier based on a feed-forward neural network. The model is heavily inspired by Hardmeier et al. (2013b), 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 in the original paper, since this causes large performance overhead.

We find out that the model indeed performs 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 et al.2010] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

[Broscheit et al.2010] Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. Bart: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 104–107. Association for Computational Linguistics.

[Glorot and Bengio2010] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of

training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256.

[Guillou2012] Liane Guillou. 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 Computational Linguistics*, EACL '12, pages 1–10. Association for Computational Linguistics.

[Hardmeier and Federico2010] Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 283–289.

[Hardmeier et al.2013a] Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013a. Docent: A document-level decoder for phrase-based statistical machine translation. In *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 193–198. Association for Computational Linguistics.

[Hardmeier et al.2013b] Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013b. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391.

[Hardmeier2014] Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University.

[Koehn et al.2007] 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.

[Koehn2005] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

[Le Nagard and Koehn2010] Ronan Le Nagard 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, pages 252–261. Association for Computational Linguistics.

[LeCun et al.2012] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 2012. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

[Mitkov et al.1995] Ruslan Mitkov, 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*, pages 5–7.

[Mitkov1999] Ruslan Mitkov. 1999. Introduction: Special issue on anaphora resolution in machine translation and multilingual nlp. *Machine translation*, 14(3):159–161.

[Och2003] FJ Och. 2003. *Giza++ software*. Internal report, RWTH Aachen University.