A BiLSTM-based System for Cross-lingual Pronoun Prediction

Sara Stymne, Sharid Loáiciga and **Fabienne Cap**
Department of Linguistics and Philology
Uppsala University
firstname.lastname@lingfil.uu.se

Abstract

We describe the Uppsala system for the 2017 DiscoMT shared task on cross-lingual pronoun prediction. The system is based on a lower layer of BiLSTMs reading the source and target sentences respectively. Classification is based on the BiLSTM representation of the source and target positions for the pronouns. In addition we enrich our system with dependency representations from an external parser and character representations of the source sentence. We show that these additions perform well for German and Spanish as source languages. Our system is competitive and is in first or second place for all language pairs.

1 Introduction

Cross-lingual pronoun prediction is a classification approach to directly estimate the translation of a pronoun, without generating a full translation of the segment containing the pronoun. The task is restricted to pronouns at subject positions only and it is defined as a "fill-in-the-gap-task": given an input text and a translation with placeholders, replace the placeholders with pronouns. Word alignment links of the placeholders to the source sentence are also given. This setting allows to analyze both the source and the target languages to create features, potentially providing the means to understand the different aspects involved in pronoun translation.

First formalized by Hardmeier (2014), the approach was introduced as a shared task at the DiscoMT 2015 Workshop (Hardmeier et al., 2015). In 2016, the shared task included more language pairs and lemmatized target data (Guillou et al., 2016). This year's edition (Loáiciga et al., 2017) also features lemmatized target data and it includes the Spanish-English language pair, which introduces pro-drops or null subjects to the task. These refer to omitted subject pronouns whose interpretation is recovered through the verb's morphology, as shown in Figure 1.

src	*me* ***ayudan*** *a ser escuchada*
	"me help$_{\textbf{3.Pers.Pl}}$ to be heard"
trg	**REPLACE** help me to be heard
pos	PRON VERB PRON PART AUX VERB
ref	**They** help me to be heard

Figure 1: Spanish-English example.

Given the success of neural networks for cross-lingual pronoun classification (Hardmeier et al., 2013; Luotolahti et al., 2016; Dabre et al., 2016), we wanted to explore this type of system architecture. Our system is based on BiLSTMs enhanced with information about the source pronoun, the pronoun's syntactic head dependency and character-level representations of the source words. Our system ranked first for English–German, with 10 percentage points of macro recall ahead of the second best team. For the other three language pairs, the system obtained the second best macro recall. In addition, our system reached the highest accuracy for three out of the four language pairs.

2 Related Work

Our system architecture draws inspiration from several sources, most prominently from the pronoun prediction system by Luotolahti et al. (2016) and the parser architecture by Kiperwasser and Goldberg (2016).

Luotolahti et al. (2016) built the winning system for the 2016 edition of this shared task. The system is based on two stack levels of GRU units and it relies almost uniquely on context. Other

than representations of the source pronouns, its input contains up to 50 tokens of context, reading away from the pronoun to be predicted, to the left and the right, both for the source and the target language. It uses a weighted loss which penalizes classification errors on low frequency classes. Our system mainly differs from this in that we use BiLSTM units reading from the sentence boundaries towards the pronoun and we rely on sampling strategies instead of weighting the losses.

Kiperwasser and Goldberg (2016) describe a dependency parser based on a BiLSTM layer representing the input sentence. The input to the BiLSTMs are word and POS-tag embeddings. Each word is then represented by the BiLSTM representation at this position, which forms a basis for both a graph-based and a transition-based parser. We use the same underlying BiLSTM layer for word representations, but in our case, we feed the representation of selected words to a pronoun classifier. de Lhoneux et al. (2017) describe several additions to this parser, including character embeddings as part of the word representation. Given their value to capture morphological information, we include character embeddings for the source language in our system.

Loáiciga (2015) reports that pronoun prediction benefits from syntactic features when using a Maximum Entropy classifier. Similarly, but using an SVM classifier, Stymne (2016) provides evidence in favor of including information about dependency heads for pronoun classification, especially for the source languages German and French. We followed these findings and included head dependency information into our current system.

3 Data and Evaluation

We use only the training data provided by the shared task (Loáiciga et al., 2017).[1] For development data, we concatenate all available development data for each language pair. Test data is the official shared task test data. For training data we either concatenate all available training data, or use only the in-domain IWSLT data, which contains TED talks. In addition, we perform experiments with a very simple domain adaptation technique in the spirit of Zoph et al. (2016), but applying it to different domains instead of to different languages. We first train models on all available data, then continue training these models for additional epochs using only in-domain IWSLT data.

While the source side sentences are regular inflected words, the target side sentences are given as lemmas with POS-tags. In order to utilize richer representations for the source side we tag and parse the source data. For English and German we use Mate Tools (Bohnet and Nivre, 2012) and for Spanish we use UD-Pipe (Straka et al., 2016). To achieve a flat representation, we represent each source word by its word form, POS-tag and the dependency label for its head (e.g. *woman*|*NOUN*|*SBJ*, *false*|*JJ*|*NMOD*). After parsing, all input words and lemmas are lowercased, and all numerals are replaced by a single token.

3.1 Sampling

One of the inherent difficulties of the task is the imbalance in the distribution of the classes. Every language pair is different, but in general the OTHER class is large in comparison to all other classes, and masculine pronouns are more frequent than feminine pronouns. The feminine plural pronouns is one of the most extreme cases, since they are only used whenever their referent points to a group containing exclusively feminine members.

During training, we sample the sentences to use in each epoch, in order to handle the imbalance in the data, which in addition also reduces the memory needed to handle all training data. For each epoch we use a small proportion of the training data that we randomly sample by selecting each sentence based on a different probability for each pronoun class. In case a sentence has several pronoun instances, we use the probability of the rarest class in the sentence.[2] We use several sampling schemes. **Equal sampling** optimizes macro-recall, it accommodates an equal number of instances for each pronoun class per sample. In case a class has fewer instances than required, all available instances for that class are used. **Proportional sampling** optimizes accuracy by sampling based on the class proportions in the development data. We also investigated an **offline sampling** scheme, which is similar to proportional sampling. In this case the sample has the same distribution of classes as the development data and also the same size. Because the sample size is small, this

[1] See also https://www.idiap.ch/workshop/DiscoMT/shared-task.

[2] Using all pronoun instances of a sentence improves training efficiency, but at the cost of making the sample proportions less precise.

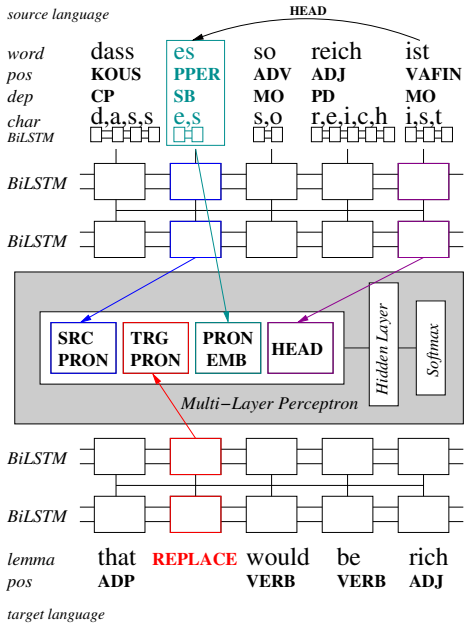

Figure 2: System architecture overview.

sampling method requires training for many more epochs. In order to have exact sample proportions, rather than the inexact proportions from choosing each example with a specified probability, we precompute and store the samples in this scheme.

3.2 Evaluation

We give results on two metrics, macro-recall (macro-R) and accuracy. Macro-R is the official shared task metric. It gives the average recall for each pronoun class, thus giving the same importance to rare classes as to common classes. We also give unofficial accuracy scores, to give a more balanced view of system performance. All scores are given on both the official test data and dev data.

4 System Description

Our system is a neural network architecture with a multi-layer perceptron (MLP) classifier fed with BiLSTM (Hochreiter and Schmidhuber, 1997) representations of tokens, which in turn are based on embeddings for word forms, lemmas, POS-tags, dependency labels, and character representations. The system is depicted in Figure 2.

Each token in the target sentence is represented as the concatenation of an embedding of its lemma and its POS-tag. Each source token is represented as the concatenation of embeddings for the input word, POS-tag, dependency label, and a character representation based on a separate character BiLSTM, reading the sequence of characters in the token. Character representations were only used in the source, since we believe that they can capture morphology, which is not meaningful for the lemmatized target sentence. All embeddings are initialized randomly. The source and target token representations are then fed to a separate two-level BiLSTM that reads the sentence backwards and forwards. No cross-sentence information is used.

Parameter	Value
Word embedding dimensions	100
Lemma embedding dimensions	100
POS-tag embedding dimensions	10
Dep label embedding dimensions	15
Character embedding dimensions	12
Character BiLSTM dimensions	100
BiLSTM Layers	2
BiLSTM hidden dimensions	200
BiLSTM output dimensions	200
Hidden units in MLP	100
α (for word dropout)	0.25
LSTM dropout	0.33

Table 1: Hyper-parameter values.

On top of this architecture we have an MLP, using `tanh` for activation, that for each pronoun instance takes as input the BiLSTM representation of the target pronoun, the source pronoun, the dependency head word of the source pronoun, and in addition takes the token representation of the source pronoun. For Spanish-English, we did not use the dependency head word, since the source pronoun is already encoded in a verb, because of pro-drop, see Figure 1. The MLP consists of this input layer, a hidden layer and a softmax output layer, representing all pronoun classes for the given target language.

We use dropout on all LSTMs. In addition, we use the word dropout of Iyyer et al. (2015) for words and lemmas, where we randomly replace a word with the UNKNOWN token with a frequency inversely proportional to the word frequency. Moreover, we replace all words occuring only once in the training data with the UNKNOWN token. Table 1 shows the values of the hyper parameters used in the system. We did not perform any optimization of hyper parameter values. Our system is implemented using DyNet (Neubig et al., 2017), and re-uses code from Kiperwasser and Goldberg (2016) and de Lhoneux et al. (2017).

We train the full model jointly, using a log loss on the final pronoun classification and Adam (Kingma and Ba, 2015) as the optimizer. Training the BiLSTMs as part of the full classification

System	de-en		en-de		en-fr		es-en	
	mac-R	acc	mac-R	acc	mac-R	acc	mac-R	acc
All components	0.67	0.84	0.47	0.73	0.51	0.73	0.72	0.83
No char emb	0.65	0.83	0.47	0.73	0.52	0.73	0.69	0.82
No dep emb	0.67	0.84	0.47	0.74	0.53	0.74	0.70	0.83
No pos+dep emb	0.67	0.82	0.46	0.72	0.51	0.72	0.68	0.82
No dep emb/head	0.57	0.80	0.47	0.74	0.53	0.74	–	–
No pron emb (MLP)	0.65	0.81	0.46	0.73	0.53	0.74	0.69	0.82
None of the above	0.49	0.73	0.46	0.72	0.50	0.71	0.67	0.81

Table 2: Development results with different system settings, training with IWSLT data, and proportional sampling. Scores are Macro-R and accuracy.

System	de-en		en-de		en-fr		es-en	
	mac-R	acc	mac-R	acc	mac-R	acc	mac-R	acc
All components	0.65	0.84	0.48	0.76	0.47	0.65	0.56	0.65
No char emb	0.59	0.77	0.47	0.75	0.48	0.67	0.55	0.66
No dep emb	0.63	0.81	0.48	0.76	0.45	0.66	0.54	0.62
No pos+dep emb	0.62	0.78	0.46	0.73	0.46	0.65	0.48	0.50
No dep emb/head	0.54	0.74	0.50	0.80	0.46	0.66	–	–
No pron emb (MLP)	0.63	0.80	0.48	0.76	0.46	0.67	0.56	0.65
None of the above	0.51	0.71	0.46	0.73	0.47	0.64	0.44	0.47

Table 3: Test results with different system settings, training with IWSLT data, and proportional sampling. Scores are Macro-R and accuracy.

instead of training them separately allows them to adapt better to the pronoun classification task. We use no mini-batching, so in order to stabilize the system to some extent, we follow Kiperwasser and Goldberg (2016) and only update the parameters after collecting several non-zero losses, in our case, 25. In all cases we choose the best epoch based on the average of macro-R and accuracy on the development data. We believe that using both metrics for choosing the best epoch will give us a system that can predict rare classes well, while not sacrificing the overall accuracy across classes.

5 Experiments and Results

First we performed experiments to evaluate the different components of our network, using only IWSLT data. These experiments are run for 100 epochs with proportional sampling and 10% of the training data in each epoch. Table 2 shows the results on development data and Table 3 shows the results on test data. We can note a marked difference in performance for English as a target language on the one hand, and English as a source language on the other hand, which interestingly mirrors previous results with an SVM classifier (Stymne, 2016). With German or Spanish as the source, nearly all the components are useful, and discarding them all results in a large performance drop on both metrics. Using the source pronoun head in the MLP was highly useful for German, but not used for Spanish, where the source pronoun is already encoded in the verb. When English is the source language, we see little effect of any component; some of them even hurt performance slightly. The **all** system did give slightly better scores than the **none** system even in this direction, though, so we decided to use the all components system for all languages in our submission.

For our main experiments, we used all training data and different sampling schemes. For the equal and proportional sampling schemes we used samples containing 10% of the data and ran the system for 72 hours, which resulted in 36–66 epochs, depending on the language pair and sampling scheme. When domain adaptation is used, we ran an additional 100 epochs with the same settings but only IWSLT data, as a final step. For offline sampling, we precomputed 500 samples per training file, and ran 860–1204 epochs.

Tables 4 and 5 shows the results of these experiments for development and test data. Using all data and proportional sampling improves over using only IWSLT, but to different degrees for the different language pairs. Overall we see that for several language pairs the scores are quite different on dev and test data. For English–German, macro-R on test is higher, which can be explained by the missing rare class *man* in the test data. For German–English macro-R is lower on test, which can be explained by our system failing to predict

Sampling	DA	de-en		en-de		en-fr		es-en	
		mac-R	acc	mac-R	acc	mac-R	acc	mac-R	acc
Equal	no	0.80	0.81	0.64	0.72	0.64	0.75	0.75	0.82
Equal	**yes**	0.81	0.86	0.62	0.73	0.66	0.77	0.79	0.82
Proportional	no	0.69	0.85	0.48	0.76	0.58	0.75	0.71	0.83
Proportional	*yes*	0.71	0.87	0.51	0.75	0.60	0.76	0.72	0.84
Offline	no	0.67	0.83	0.49	0.73	0.59	0.76	0.70	0.83

Table 4: Final development results on all training data with different types of sampling, with and without domain adaptation (DA). Scores are Macro-R and accuracy.

Sampling	DA	de-en		en-de		en-fr		es-en	
		mac-R	acc	mac-R	acc	mac-R	acc	mac-R	acc
Equal	no	0.65	0.78	0.73	0.76	0.64	0.69	0.59	0.64
Equal	**yes**	0.69	0.85	0.78	0.79	0.64	0.70	0.59	0.68
Proportional	no	0.66	0.85	0.62	0.79	0.53	0.65	0.58	0.66
Proportional	*yes*	0.67	0.85	0.62	0.79	0.50	0.65	0.56	0.62
Offline	no	0.66	0.83	0.59	0.74	0.48	0.65	0.51	0.65
Shared task baseline	–	0.38	0.54	0.54	0.55	0.37	0.48	0.34	0.37

Table 5: Final test results on all training data with different types of sampling, with and without domain adaptation (DA). The last line shows the official shared task baseline scores. Scores are Macro-R and accuracy.

the very few instances of two rare classes. For Spanish–English, the scores on both metrics are overall lower for all classes in test, for which we can see no clear explanation.

We expected to see a trade-off between macro-R and accuracy for the equal sampling compared with the other sampling methods, like for Luotolahti et al. (2016) who used weighted loss. For the dev data we see clearly higher macro-R with equal sampling, but, less of a difference for accuracy. For the test data with domain adaption, though, scores on both metrics are either better or similar with equal sampling compared to the other sampling methods. This means that the system with equal sampling performs strongly on both metrics, contrary to our expectations, making it clearly the best choice for this task. We believe that one partial reason for this could be that we choose the best epoch based on the average of the two metrics.

Domain adaptation improved the results slightly in most cases on dev data. On the test data, we also saw improvements or stable results in most cases, the exceptions being proportional sampling for English-French and Spanish-English, where we saw a small drop in results. We also note that all of our systems are considerably better than the shared task LM-based baseline (Loáiciga et al., 2017), shown in Table 5, on both metrics.

For our shared task submission we used the system with equal sampling and domain adaptation as our primary system, **bold** in Table 4, since it had the best macro-R scores on the development set. We used the system with proportional sampling with domain adaptation as our secondary system, *italic* in Table 4. Our systems perform well in the shared task, achieving first and second places for both macro-R and accuracy in all cases. Our primary systems have high scores on both macro-R and accuracy, in contrast to most other systems in the shared task.

6 Conclusions

We have presented the Uppsala system for the 2017 DiscoMT shared task on cross-lingual pronoun prediction. It is a neural network with BiLSTMs as backbone representations of words and lemmas. We show that for German and Spanish as source languages it is useful to add information from characters, POS-tags and dependencies, whereas this has little effect for English as a source language. We define effective sampling schemes to optimize macro-R and accuracy. Our primary systems have high scores on both macro-R and accuracy, when we use sampling schemes with an equal distribution of classes, and choose the best epoch based on the average of macro-R and accuracy. We also show that simple domain adaptation where we train on only in-domain data in the last epochs can improve results. Our system has the highest or second highest score for both macro-R and accuracy for all language pairs in the official evaluation.

Acknowledgments

We would like to thank Eliyahu Kiperwasser and Miryam de Lhoneux for sharing their code and for valuable discussions. SL was supported by the Swedish Research Council under project 2012-916 *Discourse Oriented Statistical Machine Translation*. FC was funded by a VINNMER Marie Curie Incoming Grant within VINNOVAs Mobility for Growth programme. Computations were completed in the Taito-CSC cluster in Helsinki through NeIC-NLPL (www.nlpl.eu).

References

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea, pages 1455–1465.

Raj Dabre, Yevgeniy Puzikov, Fabien Cromieres, and Sadao Kurohashi. 2016. The Kyoto university cross-lingual pronoun translation system. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 571–575.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017. From raw text to universal dependencies – look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada.

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*. Berlin, Germany, pages 525–542.

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

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*. Lisbon, Portugal, pages 1–16.

Christian Hardmeier, 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*. Seattle, Washington, USA, pages 380–391.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, pages 1681–1691.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, California, USA.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics* 4:313–327.

Sharid Loáiciga. 2015. Predicting pronoun translation using syntactic, morphological and contextual features from parallel data. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Lisbon, Portugal, pages 78–85.

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*. Berlin, Germany, pages 596–601.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Kevin Duh Trevor Cohn, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, pages 4290–4297.

Sara Stymne. 2016. Feature exploration for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 609–615.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, USA, pages 1568–1575.