Baseline Models for Pronoun Prediction and Pronoun-Aware Translation

Jörg Tiedemann
Department of Linguistics and Philology, Uppsala University
Department of Modern Languages, University of Helsinki
firstname.lastname@lingfil.uu.se

Abstract

This paper presents baseline models for the cross-lingual pronoun prediction task and the pronoun-focused translation task at DiscoMT 2015. We present simple yet effective classifiers for the former and discuss the impact of various contextual features on the prediction performance. In the translation task we rely on the document-level decoder Docent and a cross-sentence target language-model over selected words based on the parts-of-speech of the aligned source language words.

1 Introduction

The second workshop on discourse in machine translation (DiscoMT 2015) features a shared task on pronoun translation. Pronouns are difficult to translate due to their complex semantics. Anaphoric pronouns refer back to their antecedent and, therefore, have to agree with linguistic properties such as gender and number. The main problem for machine translation is that antecedents can be arbitrarily far away from the pronouns that refer back to them. This is not an issue if gender, number and other properties are preserved in translation and if these properties are marked in both languages. However, this is not always the case and for most language pairs there are various grammatical differences that need to be taken care of. A prototypical example is grammatical gender which is used in languages like German or French. Translations of inanimate nouns such as "the door" are assigned to a gender (feminine in the case of the German "die Tür") which is not derivable from the source. Hence, machine translation faces the problem to decide which pronoun to use in translations of "it" referring back to "the door". The task, however, is even more complex due to the frequent use of non-referential pronouns in constructions like "it is raining" where an equivalent pronoun may or may not appear in the translation. The shared task focuses on French translations of the third-person pronouns "it" and "they". The cross-lingual pronoun prediction task asks for the corresponding item in French (grouped into nine classes) for given English documents and their human-generated translations into French. The translation task requires complete translations of English documents to French and the evaluation emphasizes the translations of the two types of pronouns. The domain is translated TED talks. In the following, we first look at the prediction task and our classification approach. Thereafter, we discuss the translation model that we used in our submission (UU-TIEDEMANN).

2 Cross-Lingual Pronoun Prediction

In the pronoun prediction task, the system needs to return one of nine classes that correspond to the translation of "it" and "they" into French in given context. The classes include the pronouns *ce, cela, elle, elles, il, ils, on* and *ça* which are common translations of the given English pronouns, and another class (*OTHER*) that covers all other cases (including pleonastic uses and other cases that do not have any correspondence in French). English and French context is fully visible for the entire document with special place holders marking the space where the corresponding class is to be filled in. Note that the data (training and test data) is prepared using automatic word alignment and, therefore, includes noise.

In our submission, we were mainly interested in testing various baselines in order to test how far we can get with a rather poor feature model and minimal amounts of pre-processing. Hence, we do not attempt to run any kind of anaphora resolution to identify co-referential links nor any other kind of linguistic analyses that might help to resolve the ambiguities of the decision. We look at two types of features only:

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 108–114,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

Local context: Surrounding words in source and target language.

Preceding noun phrases: Preceding noun phrases in the close neighborhood have a good chance to represent antecedents of given pronouns. Assuming that they may be marked with the properties we require for disambiguation (number and gender) we extract simple features from them as additional features.

Our experiments are based on standard classifiers and we use existing implementations out of the box. We tested local classification models based on maximum entropy models, averaged perceptrons (using MegaM (Daumé III, 2004)) and linear SVMs (using liblinear (Fan et al., 2008)) but also a sequence model based on conditional random fields (using crf++ (Kudo, 2013)). In our initial experiments it turned out that liblinear produces significantly better results than any of the other tools and, therefore, we only report results from applying that software. In all experiments we use L2-loss SVC dual solvers which is the standard setting in liblinear. We did not perform any optimization of the regularization parameter C and we only use IWSLT14 for training.

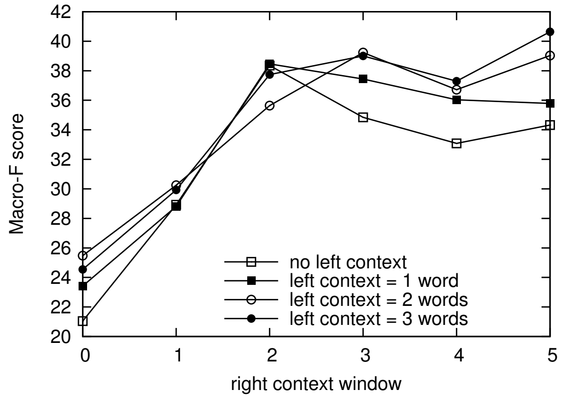

Figure 1: Various context windows in the source language (used as bag of words).

Our first batch of experiments considers various sizes of source language context. Figure 1 illustrates the impact of source language features with increasing window sizes using tokens to the left and to the right. The figure shows that context to the right seems to be more important than left-side context. Windows larger than 2 words seem to be sufficient but overall, the performance is not satisfactory.

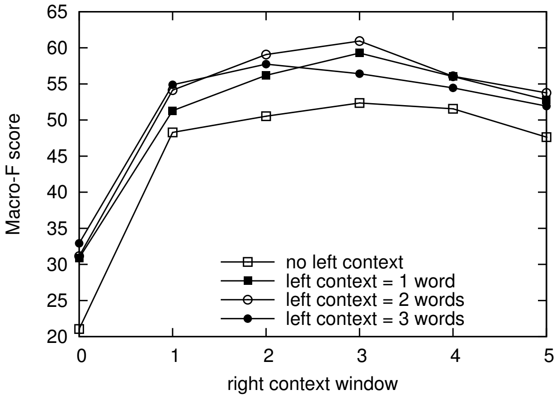

Figure 2: Various context windows in the target language (used as bag of words).

It is to be expected that target language context is more informative for the classifier decision. Figure 2 demonstrates this using the same setup as in the experiments with source language context. The overall performance in terms of macro F-scores is much higher now but similar to source language context, tokens to the right seem to be more informative for classification decisions. Small window sizes are preferred as well and the optimal performance on development data is achieved for two words to the left and three words to the right.

system	macro F	accuracy
bag-of-words		
trg2+3, 1 det	61.67	79.79
trg2+3, 2 det	61.97	79.52
trg2+3, 3 det	57.85	79.25
trg2+3, 4 det	58.54	78.98
trg2+3, 5 det	55.42	78.85
position-sensitive		
trg2+3, det 1	60.82	81.79
trg2+3, det 2	57.78	80.59
trg2+3, det 3	57.45	80.72
trg2+3, det 4	56.91	80.32
trg2+3, det 5	57.01	80.46

Table 1: Classifiers with tokens aligned to English determiners in previous context as extra features besides target language context (2 words before and 3 words after).

The results above use bag-of-words models that do not make any difference between the positions of the contextual words within the selected window. We also ran experiments with features marked with their positions relative to the predicted item but the outcome was rather inconclusive. In our next setup, we present both, position-sensitive models and bag-of-word models. The main difference in the feature model is, otherwise, the addition of long-distance contextual information. Assuming that preceding noun-phrases in the close neighborhood are good candidates of antecedents that may be marked with gender and number, we extract French tokens that are linked to English determiners and demonstratives from previous context. In order to make our approach completely independent from external tools we simply specify a fixed list of common determiners: *a, an, the, those, this, these* and *that*. The corresponding French tokens are taken from the given word alignments. Table 1 lists the classifier performances with these additional features in terms of macro F-scores and overall accuracy. We can see that the determiner information adds information that leads to modest improvements but only if one or two items are considered. We can also see that there is a discrepancy between macro F-scores and accuracy with respect to the use of positional information. Bag-of-word models produce higher F-scores for small windows but lower overall accuracy than position-sensitive models. For our final experiments, we rely on position-sensitive models assuming that macro F-scores are less stable than accuracy especially also considering the differences in class distributions between development and test set.

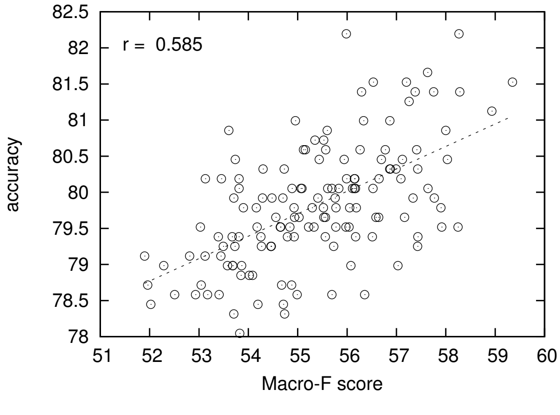

Figure 3: Correlation between macro F-score and accuracy for various context windows in source and target language (development set).

The correlation between macro F-score and accuracy is further shown in Figure 3. The plot shows the relation between these two metrics for various context windows in source and target language. From the plot we can see that there certainly is a correlation between overall accuracy and macro F-score but that this correlation is not as strong as one might expect especially with respect to these quite homogenous features.

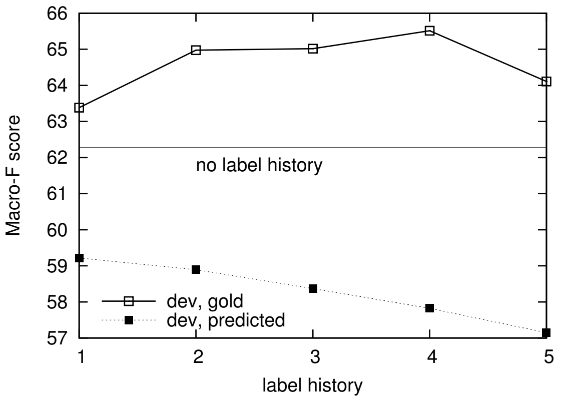

Figure 4: Using label history features: Oracle scores with gold label features and predicted labels as features besides position-sensitive local context (src1+2 and trg1+2) and tokens aligned to English determiners (det 2) tested on the development set.

Another strategy that we explored is the use of local dependencies between predicted labels. Intuitively, it should be important to know about previously used pronouns to predict the next ambiguous one. Referential pronouns are often included in larger coreferential chains and refer back to the same entity in the discourse. This fact can be exploited by sequence labeling techniques that incorporate target dependencies. However, our results with CRF that include markovian dependencies on predicted labels were quite disappointing and fall far behind the results obtained with local predictions using liblinear. Therefore, we also added a model with history features that include previous labels as additional features in local predictions. Training such models is straightforward with fully visible data sets as the ones given in the pronoun prediction task. The main problem is that the model needs to handle noisy predicted labels at testing time where gold labels of previous decisions are not available. Figure 4 plots the score obtained with history features on the DiscoMT development set. The oracle scores using gold history labels from the development set shows the capacity of these features. They significantly push the performance with over five point gains in macro F-score. Dependencies up to four labels in history seem to be beneficial. However, using a simplistic approach to incorporate predicted labels at testing time results in drastic drops leading to scores below the models without history features. These results are rather discouraging and we did not try to improve the history-based models by common techniques such as training with predicted labels using jackknifing approaches. This could, however, be interesting to explore in future work.

class	precision	recall	F
ce	80.28	92.93	86.15
cela	25.00	22.22	23.53
elle	45.65	25.30	32.56
elles	66.67	27.45	38.89
il	49.26	64.42	55.83
ils	74.50	93.12	82.78
on	70.83	45.95	55.74
ça	66.22	48.04	55.68
OTHER	88.83	91.32	90.06
micro avg	74.21	74.21	74.21
macro avg	63.03	56.75	57.91

Table 2: Final classifier result on the DiscoMT test set (submission UU-TIED).

Finally, in our submitted system we, therefore, applied a local classifier without history features and target context only. We used two words before and three words after from the local context and target language words linked to the closest source language determiner from previous context regardless of distance. Furthermore, we added the word that follows next to those linked words in the target language to add yet another feature that may help the classifier to predict gender and number correctly. The final results of this model applied to the official test set is shown in Table 2. The scores show that we cannot achieve the same quality on test data as we have seen on the development data. This is certainly to be expected but the drop is quite significant (both in macro F-score and in overall accuracy). Still, our system is the highest ranked submission according to the official macro average F-score. However, it is below the baseline model (58.4%) but significantly outperforms the baseline in overall accuracy (74.2% versus 66.3%).

The system works surprisingly well in recognizing OTHER cases and also the frequent demonstrative pronoun "ce" as well as the masculine plural "ils" works reasonably well. Most problems can be found in the predictions of the female pronouns "elle" and "elles" but also the confusion between "cela" and "ça" is noticeable. For further details of the individual mistakes done by the classifier, please look at the confusion matrix in Table 3. Here, we can see that "il" is very often misclassified as "ce" and "elle", and "elle" is often tagged as "il" – important ambiguous cases that DiscoMT tries to focus on. Looking at these results, we can conclude that the final model is only modestly successful and further work needs to be done to improve prediction quality.

	←——— classified as	———→								
	ce	cela	elle	elles	il	ils	on	ça	other	sum
ce	171	1	0	0	7	1	0	2	2	184
cela	1	6	3	0	0	0	0	11	6	27
elle	9	2	21	0	28	5	3	6	9	83
elles	1	0	0	14	2	32	0	0	2	51
il	12	2	15	1	67	0	3	2	2	104
ils	0	0	0	5	1	149	0	0	5	160
on	2	1	0	0	11	4	17	2	0	37
ça	4	11	6	0	15	2	0	49	15	102
other	13	1	1	1	5	7	1	2	326	357
sum	213	24	46	21	136	200	24	74	367	

Table 3: Confusion matrix

3 Pronoun-Focused Translation

The pronoun-focused translation task at DiscoMT requires a full machine translation system. Our submission uses a phrase-based model with one additional document-level feature function that captures long-distance relations spanning over arbitrarily long distances within a given document and its translation. We use Docent (Hardmeier et al., 2013), a document-level decoder that supports such feature functions and test our model on the DiscoMT test set.

3.1 Document-Level Decoding

The common strategy to decode phrase-based SMT models is to use a beam search algorithm based on dynamic programming and incremental hypotheses expansion (Koehn, 2010). This approach is very efficient and successful for local features such as context-independent translation options of word sequences and n-gram-based language models. Long-distance dependencies on the target language are impossible to incorporate which makes it difficult to account for coreferential relations over arbitrary spans in order to resolve, for example, ambiguities in the translation of anaphoric pronouns. Docent implements a different decoding strategy that starts with a complete translation hypotheses of an entire document applying local changes to improve the translation according to the model it uses (Hardmeier et al., 2012). The algorithm is a stochastic variant of standard hill climbing and at each step, the decoder generates a successor of the current translation by randomly applying one of a set of state-changing operations at a random location in the document. Operations include changing the translation of a phrase, swapping positions of two phrases, moving a phrase and re-segmenting phrases. The decoder is non-deterministic but has been shown to be quite stable at least with standard features commonly used in phrase-based SMT. The decoder can be initialized using a randomly generated translation of the entire document based on translation options from the phrase table or using the beam-search decoder implemented in Moses (Koehn et al., 2007). More details about the decoder and document-level feature models can be found in (Hardmeier, 2014).

3.2 Selected Word Language Models

For the purpose of DiscoMT, we implemented a feature function that can handle n-gram language models over selected words. These n-grams can easily cross sentence boundaries within a given document d but otherwise they use the same approach as any other Markovian language model:

$$p_{swlm}(d) = p(w_{s1})p(w_{s2}|w_{s1})..p(w_{sn}|w_{sn-k+1}..w_{sn-1})$$

The *selected* words $w_{s1}..w_{sn}$ can be found using various criteria. The selection can be based on part-of-speech labels or other annotation or properties such as word length. Depending on the chosen criteria, only a small subset of words may be selected and the distance between them can be arbitrary long within the limits of the document. One problematic issue in the machine translation setup where arbitrary strings can be generated is that such a language model prefers hypotheses that include as few elements as possible if corresponding n-gram probabilities are sufficiently high. This is a typical behavior of any n-gram language model and penalty features are commonly used to penalize short hypotheses. Another possibility is to base the selection process on the given source language string which is given and fixed and to obtain the target language tokens through word alignment. In this way, the feature function includes a similar number of factors (small differences are due to different word alignment types) for each hypotheses and additional penalty features can be avoided. This is especially useful for our document-level decoder in which tuning of feature weights is not very stable.

The strategy that we like to explore in the pronoun-focused translation task is to make use of the relation between subsequent pronouns and context words that may indicate anaphoric agreement constraints such as gender and number. For this, we implemented an n-gram language model over words that are linked to English pronouns and determiners and used this feature function as the only additional long-distance feature besides standard sentence-level phrase-based SMT features. We tagged the English part of the DiscoMT training data (Europarl, IWSLT15 and News Commentary v9) with HunPos and a model trained on the Universal Dependency Treebank v1 (McDonald et al., 2013) using the coarse universal PoS tag set of Petrov et al. (2012). From the tagged corpus and the alignments to their French translations, we extracted the linked French tokens for selected words using the provided word alignments and, finally, trained a 7-gram language model with modified Kneser-Ney smoothing using KenLM (Heafield et al., 2013) from that data set.

The feature function implemented in Docent caches the target word sequence aligned to selected source language words and updates the language model score each time the hypotheses is modified and the chain is effected by the modification. Similar to the interface of the standard language model implemented in Docent, we only consider the context window that is defined by the model to allow efficient computation of the feature. The model can easily be adjusted to other word selections using parameters in the configuration file. In our case, we use a regular expression to specify the PoS labels that need to be considered:

```
<model type="selected-pos-lm" id="splm">
<p name="lm-file">/path/to/lm.kenlm</p>
<p name="selected-pos-regex">^DET|PRON$</p>
```

We did not attempt to properly tune the corresponding weight for this feature function and fixed it to a rather arbitrary value of 0.2 which seemed to perform reasonably well on development data. Table 5 lists the BLEU scores of our models with and without the additional pronoun-oriented language model. The table includes also a model that contains a language model over pronouns only (without including determiners in the context). We can see that our modified models are slightly below the baseline model in overall BLEU which is most probably due to inappropriate tuning of the

	P		R_{min}		F_{min}	R_{max}		F_{max}
ce	38/50	0.760	38/51	0.745	0.752	41/51	0.804	0.781
cela	7/8	0.875	7/47	0.149	0.255	20/47	0.426	0.573
elle	8/12	0.667	8/19	0.421	0.516	8/19	0.421	0.516
elles	3/4	0.750	3/15	0.200	0.316	5/15	0.333	0.462
il	7/23	0.304	7/22	0.318	0.311	13/22	0.591	0.402
ils	45/53	0.849	45/48	0.938	0.891	45/48	0.938	0.891
on	0/0	n/a	0/0	n/a	n/a	0/0	n/a	n/a
All pronouns	108/150	0.720	108/170	0.635	0.675			
Other	27/ 47	0.574	27/ 27	1.000	0.730			

13 instances marked as "bad translations"

	accuracy			automatic evaluation			MT scores			
	+ other	- other	pron-F	P	R	F	BLEU	NIST	TER	METEOR
Baseline	0.676	0.630	0.699	0.371	0.361	0.366	37.18	8.04	46.74	60.05
Proposed	0.643	0.590	0.675	0.386	0.353	0.369	36.92	8.02	46.93	59.92

Table 4: Official results of the pronoun-focused translation task.

additional feature weight.

system	BLEU
baseline	0.4000
+PRON-LM	0.3982
+DET+PRON-LM	0.3969

Table 5: Translation with and without pronoun language model on development data. PRON uses words linked to English pronouns and DET+PRON includes words linked to determiners as well.

In order to test our models on the specific task of translating pronouns in context, we also performed automatic evaluations of the translations we obtained for the development set. Table 6 lists the results for the three models using the evaluation approach of Hardmeier and Federico (2010). We can see that both augmented models improve the overall F1 scores mainly due to an increase in precision. The model that includes target language words linked to determiners performs best at least according to our automatic evaluation and, therefore, we selected this model as our primary submission. The differences are, however, very small and the manual evaluation of the test set translations revealed that our model could not even beat the phrase-based baseline without a pronoun-specific model. The official results of the translation task are shown in Table 4. We can see that the proposed system still scores slightly better than the baseline mode with the automatic evaluation but it is clearly below the baseline according to the manual evaluation.

	Precision	Recall	F1
baseline			
it	0,3616	0,3712	0,3663
they	0,6641	0,7227	0,6922
TOTAL	0,5000	0,5270	0,5131
+PRON-LM			
it	0,3827	0,3545	0,3681
they	0,6800	0,7143	0,6967
TOTAL	0,5237	0,5140	0,5188
+DET+PRON-LM			
it	0,3793	0,3679	0,3735
they	0,6867	0,7185	0,7023
TOTAL	0,5213	0,5233	0,5223

Table 6: Automatic evaluation of translated pronouns using the development set and its reference translation.

4 Conclusions

This paper presents the results of simple but efficient baseline classifiers that predict translations of pronouns in given context. Our experiments look at varying contexts and show that small windows of target language context are very effective. Adding information from potential antecedents leads to modest improvements. We also present a language model over pronouns and determiners integrated in document-level decoding of phrase-based machine translation. The model is promising according to automatic evaluation but manual inspection reveals that it does not lead to better translations of the selected ambiguous pronouns.

References

Hal Daumé III. 2004. Notes on CG and LM-BFGS optimization of logistic regression. Paper and implementation available at `http://www.umiacs.umd.edu/~hal/megam/`.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of IWSLT*, pages 283–289.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of EMNLP-CONLL*, pages 1179–1190.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings ACL: System Demonstrations*, pages 193–198, Sofia, Bulgaria, August.

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

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of ACL*, pages 690–696.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL*, pages 177–180.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Taku Kudo. 2013. CRF++: Yet Another CRF toolkit. http://taku910.github.io/crfpp/. v0.58.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of ACL*, pages 92–97.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of LREC*, pages 2089–2096.