A Linear Baseline Classifier for Cross-Lingual Pronoun Prediction

Jörg Tiedemann
Department of Modern Languages
University of Helsinki

Abstract

This paper presents baseline models using linear classifiers for the pronoun translation task at WMT 2016. We explore various local context features and include history features of potential antecedents extracted by means of a simple PoS-matching strategy. The results show the difficulties of the task in general but also represent valuable baselines to compare other more-informed systems with. Our experiments reveal that the predictions of English correspondences for given ambiguous pronouns in French and German is easier than the other way around. This seems to verify that predictions, which need to follow more complex agreement constraints, require more reliable information about the referential links of the tokens to be inserted.

1 Introduction

This short system paper describes the baseline classifier we have submitted to the shared task on cross-lingual pronoun prediction at WMT 2016. The goal of the submission is to provide yet another baseline that is slightly more informed than the language model baseline provided by the organisers otherwise. In the following, we will briefly discuss the model and our feature engineering efforts. Thereafter, we discuss the results for each language pair and conclude.

2 The Model

Our model follows the setup of our submissions from last year to the same task at the workshop on discourse in machine translation (Tiedemann, 2015; Hardmeier et al., 2015). Again, we apply a linear SVM classifier out-of-the-box using liblinear (Fan et al., 2008) with its L2-loss SVC dual solver without any dedicated optimisation of regularisation parameters. This year, we did not experiment with alternative classifiers and rely on our positive experience from our previous experiments. Similar to our previous submission, we explore various context windows in source and target language and optimise the feature model in a brute-force manner on the provided development data.

The scenario is slightly different from the previous year. First of all, there is an additional language pair and the reverse direction for both language pairs is also explored. The four sub-tasks have different complexity as they cover different sets of target classes and different types of phenomena. However, we do not treat the language pairs differently and run our training procedures in a language-independent mode using the same kind of feature extraction for all of them. A difference is also that we can rely on the provided coarse-grained PoS labels in the target language as another source of information. However, we cannot make use of the inflectional information in the target language as the data sets are now lemmatised. This is a serious handicap for the system as the morphological features disambiguate the choice very well as we have seen last year.

We played with various variants of the feature model trying to systematically study the impact of certain extraction methods on classification performance. The following extraction parameters are explored:

- Source language context before the pronoun in question
- Source language context after the pronoun in question

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 616–619,
Berlin, Germany, August 11-12, 2016.

	he	she	it	they	you	this	these	there	OTHER	
he	30	0	1	0	0	0	0	0	0	31
she	0	11	3	5	1	0	0	0	1	21
it	5	2	95	8	0	1	0	2	1	114
they	2	2	6	61	4	0	0	1	2	78
you	0	1	2	11	89	0	0	0	3	106
this	0	1	7	0	0	2	0	1	2	13
these	0	0	0	0	0	0	0	0	0	0
there	0	0	4	0	0	0	0	12	0	16
OTHER	1	0	8	7	12	0	0	0	76	104
-SUM-	38	17	126	92	106	3	0	16	85	

Table 1: The confusion matrix for German–English. Columns represent the predicted classes.

- Target language context before the placeholder token
- Target language context after the placeholder token
- Bag-of-word context versus context marked by relative position
- Lowercasing versus original casing
- Separate PoS and word features versus concatenated word/PoS features versus both types (separate and concatenated)

All of those features only explore local context which was quite successful in the previous year especially the local context in the target language. The new edition with lemmatised data, however, requires additional knowledge to make basic decisions that would otherwise work with local features. Last year, we included history features that list target language tokens aligned to preceding determiners and their local context as part of the potential antecedents that could determine pronoun choice based on gender and number agreement. The impact of these features was not very significant. However, with lemmatised data those features become more interesting.

We rely on the same procedure, simply including a fixed number of previous items without employing any kind of coreference resolution or deep linguistic analyses. However, this time we can rely on PoS labels to select the items we would like to include. Assuming that simple noun-phrases are common antecedents we define a pattern for matching PoS labels in prior context (determiners, nouns and proper nouns):

```
(DET|NOUN|NAM|NOM|PRON)
```

Furthermore, assuming that the nearest noun phrases have the highest likelihood to represent the referenced item, we extract the n closest words that match the pattern above. n is another parameter that we explore in tuning the model.

3 The Results

After running various combinations of parameters we ended up with settings that work best on the development data. First of all, lowercasing did not help but made things slightly worse. Adding relative position information to the context features also seems to work, so we always applied this method. Splitting tokens into separate features for lemma and PoS is also beneficial but additional keeping the concatenated variant has a positive effect.

We tested different sizes of the context window by varying the number of tokens before and after the source language pronoun and before and after the target language place-holder between zero and five in all combinations. Table 3 lists the final settings that gave the highest macro-averaged recall value on the development data.

We can see that the local context is rather small and the system does not seem to benefit from adding more data from surrounding context that is further away than 3-4 tokens. Note that we use position information for each token extracted from the context as discussed above. This worked slightly better than a bag-of-words approach that suffers less from data sparseness.

We also tried to optimise the number of antecedent candidate features coming from the history based on the PoS matching approach described earlier. We tried up to ten candidates but our models performed best with only a few of them in the feature model. In particular, we used four candidates for French–English and two candidates for all other language pairs. Using more confused the system and the performance on development data went down.

Finally, the official scores obtained using our

	ce	elle	elles	il	ils	cela	on	OTHER	
ce	57	1	0	5	2	1	0	2	68
elle	4	8	0	8	1	1	0	1	23
elles	1	1	2	0	20	1	0	0	25
il	1	11	0	40	2	5	2	0	61
ils	0	0	9	4	56	0	0	2	71
cela	0	4	0	9	0	14	0	4	31
on	0	0	2	0	2	0	5	0	9
OTHER	0	2	1	1	0	3	3	75	85
-SUM-	63	27	14	67	83	25	10	84	

Table 2: The confusion matrix for English–French. Columns represent the predicted classes.

	source		target	
language	before	after	before	after
Eng–Ger	1	0	4	4
Ger–Eng	1	4	3	4
Eng–Fre	1	3	1	3
Fre–Eng	3	1	3	4

Table 3: Final context windows used for each language pair.

	Macro-averaged			
language	precision	recall	F1	accuracy
Eng–Ger	60.43	44.69	45.24	65.80
Ger–Eng	75.05	69.76	70.02	77.85
Eng–Fre	57.11	57.50	56.99	68.90
Fre–Eng	70.54	62.98	63.72	78.96

Table 4: The official results of our submitted systems.

submitted systems are listed in 4. There is quite some variation in the quality of our classifiers. Especially English–German is quite poor, in particular in terms of macro-averaged recall, which is used as the official score of the campaign. The reason for this is not entirely clear but the confusion matrix presented below give some ideas about the situation.

3.1 English – German

The task for English–German includes only five target classes but seems (at least for our classifier) to be the hardest case. Our macro-averaged recall score is far below the other language pairs, which suggests that the model does not work well for small classes. The confusion matrix in Table 5 illustrates this as well. Recall for "er" and "man" is zero in both cases and this effects the official score significantly. The confusion between the more common classes "sie" and "es" with "OTHER" is also striking. The overall accuracy is also the worst among all language pairs considering that this sub-task has the lowest number of target classes involved.

	er	sie	es	man	OTHER	
er	0	3	10	0	2	15
sie	1	89	26	0	8	124
es	2	7	77	0	15	101
man	0	1	6	1	0	8
OTHER	1	26	23	0	85	135
-SUM-	4	126	142	1	110	

Table 5: The confusion matrix for English–German. Columns represent the predicted classes.

3.2 German – English

The results for German–English look much more promising. The overall accuracy is almost 78%, which is quite successful for a classification task with nine target classes. The confusion matrix in Table 1 shows the distribution of predicted labels and the model picks up the signals quite well for all classes. Even smaller classes like "she" and "there" work pretty well and we believe that the local context is again most informative for those decisions. The scores for "this" with its very few examples cause some problems for the macro-averaged recall score and "she" is also more frequently misclassified than bigger classes. Besides those issues, we are quite satisfied with the result for this language pair.

3.3 English – French

Similar to English–German, English–French also seems to be a harder case. The overall accuracy is in the same range as for English–German, slightly above, but now for eight classes, which is harder. The confusion matrix in Table 2 shows the frequent misclassifications for "elle" and "cela" and especially "elles", which is classified as "ils" in most of the cases. Even other classes show quite some confusion and the overall score is much below predicting pronoun translations in the other direction as we will see below.

	he	she	it	they	this	these	there	OTHER	
he	22	0	6	0	0	0	0	4	32
she	0	15	3	0	0	0	0	0	18
it	6	5	35	3	1	0	3	4	57
they	0	0	1	77	0	0	0	2	80
this	0	0	1	1	0	0	1	0	3
these	0	0	0	2	1	1	0	0	4
there	0	0	0	1	0	0	46	1	48
OTHER	4	1	5	10	1	0	2	63	86
-SUM-	32	21	51	94	3	1	52	74	

Table 6: The confusion matrix for French–English. Columns represent the predicted classes.

3.4 French – English

French–English is the best performing language pair in terms of overall accuracy. However, the macro-averaged scores are significantly below the scores for German–English; still a lot better than the predictions from English to the other two languages. The biggest problem appears in the small classes "this" and "these" but this effects the overall accuracy only little. Another class that seems more difficult is "it" with its around 64% F1 score and "he" is not much better. However, overall the model performs rather well for this language pair condering the limited information that is available to the classifier.

4 Conclusions

This paper presents baseline classifiers for the pronoun translation task at WMT 2016. Our linear classifier uses local context features and antecedent candidates from a simple PoS-based matching procedure. The results are satisfactory especially for the predictions of pronoun correspondences in English. This seems to be a simpler task than guessing the correct translations of the ambiguous English third-person pronouns into French and German with their grammatical gender and corresponding agreement problems. Our model shows that simple classifiers without further linguistic pre-processing can be used to obtain decent baseline scores in this difficult task. However, the prediction quality is still rather low and its use in machine translation or other cross-lingual applications remains to be seen.

References

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, 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*, pages 1–16, Lisbon, Portugal, September. Association for Computational Linguistics.

Jörg Tiedemann. 2015. Baseline models for pronoun prediction and pronoun-aware translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 108–114, Lisbon, Portugal, September. Association for Computational Linguistics.