Feature Exploration for Cross-Lingual Pronoun Prediction

Sara Stymne
Uppsala University
Department of Linguistics and Philology
sara.stymne@lingfil.uu.se

Abstract

We explore a large number of features for cross-lingual pronoun prediction for translation between English and German/French. We find that features related to German/French are more informative than features related to English, regardless of the translation direction. Our most useful features are local context, dependency head features, and source pronouns. We also find that it is sometimes more successful to employ a 2-step procedure that first makes a binary choice between pronouns and *other*, then classifies pronouns. For the pronoun/other distinction POS n-grams were very useful.

1 Introduction

This paper reports results for the UU-Stymne system on the WMT 2016 pronoun prediction shared task. The task entails classifying which among a set of target pronouns, or *other* is the correct translation of a given source pronoun. There are tasks for two language pairs, English and German/French, in both directions.

An example is shown in (1), where we need to predict which German pronoun should be the translation of *It*, which in this case should be *er* since it refers to the masculine word *Saal* (*room*) in the previous sentence. Had the antecedent instead been the neuter *Zimmer*, the correct pronoun would have been *es*. The target words are lemmatized with coarse POS-tags, to better mimic the SMT task, in contrast to previous versions of this task where full forms were used. For full details of the task and training data, see the task overview paper (Guillou et al., 2016).

(1) **It** 's smaller than this . REPLACE_0 sein|VERB klein|ADJ als|CONJ dies|PRON hier|ADV .|.

We set out to establish the usefulness of a large number of features for this task in all translation directions, without any explicit use of anaphora resolution. We also investigate a 2-step classification procedure.

2 System

We followed Tiedemann (2015) by using linear SVMs implemented in LIBLINEAR (Fan et al., 2008). In all experiments we use L2-loss support vector classification with dual solvers and the 1-vs-rest strategy for multi-class classification. The regularization parameter C was optimized using grid search and cross-validation as implemented in LIBLINEAR. The results were quite stable for reasonable values of C, however, and in all cases we used values between 2^{-2} and 2^{-5}.

In most of our experiments we only used IWSLT training data, with 66K–92K pronoun examples, to train our classifier, since it contains TED talks like the dev and test sets. We perform final experiments where we investigate the usefulness of adding out-of-domain News data of similar size and much larger Europarl data. Due to space restrictions we will mainly give Macro-averaged Recall (Macro-R) scores, the official workshop metric, on the TED dev set. Macro-R gives the average recall for all classes and thus gives the same weight to rare classes as to common classes.

For some of our features we needed dependency trees and POS-tags for the source. We used Mate Tools to jointly tag and dependency parse (Bohnet and Nivre, 2012) the source text for sentences that contained pronoun examples. For all languages the output is a dependency parse tree and POS-tags, and for German and French it also gives morphological descriptions. For the target side we used the given POS-tags.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 609–615,
Berlin, Germany, August 11-12, 2016.

3 2-step classification

We had two approaches to classification, a standard classifier, which we will call 1-step, and a 2-step classifier. We noted that the *other* class often were quite different from the pronoun classes, since it is very diverse, and sometimes an artifact of alignment errors. This observation led us to design the 2-step system where we first trained a binary classifier to distinguish between *other* and all the pronoun classes grouped into one class. We then had a second classifier that only had to distinguish between the pronoun classes. For the training data we collected instances for the second classifier based on gold tags. At test time we used the results from the first classifier to feed the examples classified as pronouns to the second classifier.

4 Features

We explored a high number of features of different types, which will be described in this section. We did not explicitly attempt to model anaphora in any way, but tried to identify other types of features that could give indications of which pronoun translation to use. The main reasons why we decided not to use any anaphora software is that it is not readily available for all source languages, it is error prone, and it gave no clear improvements in the 2015 shared task. All our features are largely language independent; we did not design any specific features for a specific language pair.

The WMT 2016 shared task is a follow-up to the DiscoMT 2015 shared task on en-fr pronoun prediction. An important difference between these two tasks is that target full forms were given in 2015 and only lemma+POS in 2016. However, many of our features were inspired by the submissions to the 2015 shared task. For all feature groups below, we used special beginning and end of sentence markers when needed.

Source pronoun (SP) The source pronoun to be translated was added as a feature. We believe that this is an important feature since it restricts the possible translations. Source pronouns has been used before for cross-lingual pronoun prediction (Hardmeier et al., 2013; Wetzel et al., 2015).

Local context (LCS, LCT) For these features we considered the source words surrounding the source pronoun and the lemmas+POS-tags surrounding the target pronouns. We included up to 3 words before and 3 words after the pronouns. We tried both to use bag-of-words models for words before and after the pronoun, and to encode the position of each word. Local context features were the core of the best submitted system for the 2015 shared task (Tiedemann, 2015) and were also used in many other submissions.

Preceding nouns (NN) The nouns preceding a pronoun are potential antecedents to the pronoun, and are therefore included. The target side of the shared task data included POS-tags, so there we used the four preceding nouns, including proper names, possibly going across sentence boundaries, but not crossing document boundaries. For the source side we had only parsed the sentences that contains pronouns. Because of this, we did not include cross sentence instances of source nouns, so we only included up to four previous nouns within the sentence, which meant that we often had 0 or just a few nouns on the source side. Since the source contains full forms, we also included some morphological information for these nouns, we added a feature for each POS-tag extended with morphology for number, and gender for proper names. Finally we added a feature indicating how many previous nouns there were in the sentence.

Preceding nouns or NPs have previously been used for this task with differing results. Callin et al. (2015) used up to four preceding nouns and determiners. Wetzel et al. (2015) also used preceding noun tokens, however they were identified by co-reference resolution. A difference from 2015 is that this year there are no determiners or other words on the target side that carries information such as gender, since it is lemmatized.

Target POS n-grams (POS) To generalize from lemmas, we included target POS-tags. We used n-grams of POS-tags for words surrounding the pronoun position in the target language. Using the abbreviation *b* for words before the pronoun and *a* for words after, we included the following n-gram windows: 3b, 1b, 2b+2a, 1b+1a, 1a, 3a.

POS-tags were used in several 2015 systems (Callin et al., 2015; Loáiciga, 2015; Wetzel et al., 2015), with either positive results or no separate results shown in the paper. They all used single tags, though, not POS n-grams.

Target extended POS n-grams (EPOS) The tag sets in the data are coarse-grained, with the 12 universal POS-tags (Petrov et al., 2012) for En-

glish and German, and a set of 15 POS-tags for French. To compensate somewhat for this, we also included n-grams using an extended tag set where we use the identity of the 100 most common lemmas in the training data in addition to the POS tags. As an example, *be-VERB-all* and *can-make-the* are two EPOS options for *VERB-VERB-DET*. We use the same n-gram windows as for standard POS n-grams.

As far as we know, no one has used this particular extension of POS-tags for this task. However, several teams successfully used fine-grained morphological target tags last year, for instance Pham and van der Plas (2015), which was not possible this year, given the lemma+POS representation.

Dependency head of pronouns (DEP) For each source pronoun we identified its dependency head, based on the parse from Mate Tools. As features we used the head word and the label. In addition we used the POS-tag of the head for English, which distinguished between tenses and third person. For French and German we added morphological information about number and person to the POS-tag of the head. We also used indicator features for common verb suffixes that we thought were informative about tense, person and number: *s* and *d* for English, *en* and *t* for German and *e*, *nt* and *[ˆn]t* for French.

To find potential dependency heads in the target, we followed the alignment links from the word identified as dependency head in the source. For any aligned words that were POS-tagged as a verb, we included the lemma as a feature. We restricted this feature to verbs, since we believe they are most informative with regard to the pronoun, and to reduce noise from the automatic alignment.

The only work we are aware of that used syntactic feature for cross-lingual pronoun prediction is Loáiciga (2015), who parsed the target and used the dependency label of the pronoun, which only had a small impact on the results. This differed from our use since we used the dependency head, and parsed the source text and projected this information to the target through word alignments.

Target language model features (LM) For this group we included language model scores from the baseline system provided by the shared task (Guillou et al., 2016). This system uses a target language LM to score the target pronouns and 11–22 other high-frequency words, and not using a word, *NONE*. The language model we used was also provided for the shared task, a large 5-gram model trained using KenLM (Heafield, 2011) on the workshop data and monolingual News data (Guillou et al., 2016). There is a penalty for the *NONE* case, which we set to -2, which was the best value from the 2015 shared task (Hardmeier et al., 2015), and that we found to give good results for all language pairs, as shown in Table 1. Note that this LM used lemmatized data, which gave a much worse performance than the full form LM from 2015, which had .584 MACRO-F (Hardmeier et al., 2015), compared to .342 on lemmas.

Null penalty	de-en	fr-en	en-de	en-fr
0	.361	.337	.344	.406
-2	.389	.388	.358	.411

Table 1: Macro-R for workshop baseline.

The baseline system can output marginal probabilities for each pronoun or alternative word and *NONE*, giving all options larger than 0.001. We used these probabilities as feature values for each word. In addition we had features giving the highest scoring word, always and if it had a probability over 0.85; the highest probability for any pronoun, any other word, and other or *NONE*. We also had a feature for the number of options given, i.e. how many words that had a probability higher than 0.001. Target language model features were used by Wetzel et al. (2015) with mixed results.

Alignment, position, and length (APL) We used a set of features related to position, sentence length and alignments, both on instance level and sentence level. We believe that this could both give some indication about pronouns, and about how close a translation the target is. We are not aware of these features being used for this task before.

The position of the pronoun in the sentence likely plays some importance to its identity. Thus we added as features the relative position of the source and target pronouns in the sentence, the difference in relative position, and three indicator features for the target and/or the source pronoun being in a sentence initial position.

We also included some features based on word alignments. For the pronouns we indicated how many words they were aligned to in the other language, which we believe can be useful especially for identifying non-pronoun translations, which are likely noisier than pronoun translations. In addition we added two features for the total number

of alignments in the sentence, normalized by the length of the source and target sentence, respectively. On the sentence level, we also included the length ratio between the source and target sentence where the pronouns occurred.

5 Results

We performed most experiments for the 1-step classifier. For all experiments up to section 5.3 we use only IWSLT as training data. To start with we investigated whether it was best to use true-cased or lower-cased features. We did not try this individually for the different feature groups, instead we made this choice for all features for a language pair. Overall, for de-en true-casing leads to a clear improvement, which we believe is mainly caused by the *sie* pronoun, which is spelled with a capital *S* in the meaning *you* (polite), and with a lower-case *s* in the meaning *she* or *they*. For the other languages the difference is quite small, but we choose to use true-case when English is the target language, and lower-case otherwise.

5.1 Feature groups

To assess how useful each feature group is we first ran experiments using a single feature group at a time. Table 2 shows the results for individual features, all features combined and for features only from the source or target language. An interesting pattern is that features from German and French give better results than features from English, regardless of translation direction. Both for the grouped source and target features, and for local context with only source or target the best results for into English is when using source features, and from English using target features. For de-en using source features only is nearly as good as using all features. In French and German we have to distinguish between pronouns based on the gender of the antecedent, for which target features are clearly useful. English, though, does not have grammatical gender, and cannot benefit in this way from the target features.

The best individual feature group when used on its own is always local context, followed in most cases by dependency features. For en-fr we got relatively good results for EPOS and LM features. The extended EPOS-tags were clearly better than the coarse POS-tags. Using only nouns was clearly not useful, and performed even worse than APL, alignment, position and length, which we did not expect to be very informative on its own. It is interesting to see that classification only by the source pronoun, a single feature, give similar results to many of the feature groups with a high number of features, which indicates its importance. While no individual group is close to the performance of all features, several feature groups are better than the baseline system.

Group	de-en	fr-en	en-de	en-fr
All	.640	.597	.389	.583
+source	.636	.560	.345	.337
+target	.414	.368	.360	.494
+SP	.370	.371	.353	.244
+LC	.518	.561	.404	.560
+LCS	.514	.475	.339	.340
+LCT	.389	.365	.367	.456
+NN	.150	.155	.207	.161
+POS	.272	.278	.327	.300
+EPOS	.362	.353	.380	.450
+DEP	.449	.418	.369	.375
+LM	.382	.331	.338	.421
+APL	.178	.208	.276	.172

Table 2: Macro-R for individual feature groups and source and target features

Group	de-en	fr-en	en-de	en-fr
All	.640	.597	.389	.583
−SP	.536	.498	.358	.562
−LC	.639	.583	.375	.580
−LCS	.638	.592	.379	.577
−LCT	.638	.601	.389	.582
−NN	.643	.610	.375	.582
−POS	.640	.592	.386	.579
−EPOS	.649	.586	.400	.576
−DEP	.617	.589	.377	.580
−LM	.652	.674	.457	.599
−APL	.634	.595	.386	.580

Table 3: Feature ablation study. Macro-R with individual feature groups removed.

Table 3 shows the results of an ablation study, where we removed one feature group at the time from the full set of features. Here we see that several features are not useful in combination with the other features, and improve the results when removed. The biggest improvement is seen when removing the LM features, even for en-fr where they had quite a good performance on their own. This is interesting since the LM is the most important knowledge source for pronoun translation in an SMT system. We believe that the lemmatized target has too little information for these features to be useful. It is always better to use the target context words directly in the classifier than to use the LM features derived from the target context. Removing the noun features improves results somewhat for into English. As expected, the source

Type	de-en	fr-en	en-de	en-fr
All, position	.640	.597	.389	.583
All, BOW	.654	.561	.396	.567
Best, position	.656	.609	.392	.581
Best, BOW	.656	.579	.397	.557
Window T	3+3	3+3	2+2	2+3
Window S	1+3	3+1	2+2	2+3

Table 4: Macro-R with different local context, and the best window sizes

pronouns are important also in combination with the other features, and gives the biggest score drop when removed. For most of the feature groups the score difference is quite small when removed.

5.2 Final feature sets

In the above experiments we used a local context window of 3 words before and 3 words after for both source and target context. In order to improve results we first tried all combinations of target windows, from 1 to 3 words, and with this window set, all source window sizes. We also tried using positions for the context words and compared to using bags-of-words for words before and after the pronoun. Table 4 shows the best windows. Changing the window sizes led to improvements for all language pairs, except en-fr, for which it, however, improved Macro-R from .563 to .573 on the DiscoMT15 test set. There is no clear pattern of which window size that is most useful across languages. For the best windows positional features were better or similar to bag-of-words features, whereas the results were conflicting with the full context window. These results were similar to Tiedemann (2015). We decided to use positional features with the best context windows.

Finally, we tried to remove combinations of the least useful feature groups, on the systems with optimized local context. Unfortunately, due to time constraints, we had not done the full ablation tests before submission time, and failed to notice the advantage of removing the LM and NN feature groups. We thus only tried to remove sets of other less promising features for the submitted systems. The results with removed features are shown in Table 5. For the final submitted systems we used the full feature sets for en-fr and fr-en, removed EPOS for en-de and removed alignment features for de-en. This led to an improvement for en-de but for de-en we have the same score as before. When trying to remove further feature groups we had large improvements for all language pairs except de-en when also removing the LM and NN feature groups. We call this system *Final*.

System	de-en	fr-en	en-de	en-fr
Submitted	.656	.609	.411	.581
Final	.653	.675	.455	.619

Table 5: Macro-R for systems with removed sets of feature groups

Corpus	de-en	fr-en	en-de	en-fr	en-fr (D)
I	.656	**.609**	**.411**	.581	.572
IN	**.654**	.578	.379	.558	.581
IE	.627	.586	.377	.559	.582
INE	.632	.564	.395	.572	.581
INE-16	.630	.572	.377	**.557**	**.584**

Table 6: Macro-R with different combinations of training data with the feature set from the submitted system (I=IWSLT, N=News, E=Europarl), -16 means filtering away features occurring less than 16 times in the training data. (D) is for results on the DiscoMT15 set. The training data used in the submitted 1-step systems are marked in bold.

5.3 Training data

In this section we investigate the effect of adding more training data to IWSLT that was used in previous experiments. Table 6 shows the results. In most cases adding more training data led to considerably worse results on the TED dev set. For de-en, though, adding News gave similar Macro-R, and an improvement of accuracy from .853 to .873, which made us choose this option for our submitted system. For en-fr, on the DiscoMT15 dev set the results were better with more data.

With the large training data we have a very high number of features, between 263K and 563K for the different language pairs for the submitted feature sets. We tried two ways of reducing the number of features: by filtering features that occurred with a low frequency in the training data and by filtering features that had a low model score in the SVM training. When using only IWSLT data we saw little effect of either type of filtering. When training with all data we had some improvements by filtering, with the best results using frequencies. We tried many different values for filtering and overall we had good results by removing features occurring 16 times or less, but as shown in Table 6 results were mixed across language pairs and test data. Using this filtering reduced the number of features to between 31K and 55K, a reduction of around 90%. The final combination of training

System	Test set				Dev set			
	de-en	fr-en	en-de	en-fr	de-en	fr-en	en-de	en-fr
Submitted Primary (2-step)	.592	.364	.521	.654	.651	.606	.426	.592
Without bug	.702	.620	–	–				
Submitted Secondary (1-step)	.608	.341	.489	.607	.654	.609	.411	.557
Without bug	.715	.629	–	–				
Final 1-step (IWSLT)	.735	.615	.490	.616	.653	.675	.455	.619
Final 1-step (all training data)	.733	.685	.503	.613	.632	.622	.455	.608

Table 7: Macro-R for submitted system, and best systems trained after submission time,using IWSLT and all data for training.

data and filtering used for the submitted systems are shown in bold in Table 6.

5.4 2-step Classification

For the 2-step classification we needed to train two classifiers, one for the binary pronoun–*other* distinction and one for the distinction between the different target pronouns. For the first classifier we chose classifiers that gave high precision and reasonable recall on the *other* class from the 1-step classifiers. Across language pairs the best results we saw before submission was to either use only the POS or EPOS feature groups, or all features. In addition we tried using either IWSLT or all training data for this classifier. We had the best results using the following feature sets and data for the first binary classifier:

- de-en: all data, all features,
- fr-en: IWSLT, all features
- en-de: all data, POS
- en-fr: IWSLT, EPOS

Overall we tended to get better precision for the *other* class using (E)POS and better recall using all features. The fact that (E)POS-patterns gave a high precision, indicates that the *other* class tends to occur in different contexts than pronouns.

For the pronoun classifier we used the full feature set and only experimented with using either IWSLT or all data for training. We had the best results with all training data for en-fr and with IWSLT for the other language pairs, similar to the results for 1-step classification. The results for the 2-step classifier are shown in Table 7, labeled as primary. We choose to submit the 2-step classifier as our primary system since it performed best on the dev data for from English, and only slightly worse in the other direction. We believe that there is room for similar improvements with the 2-step classifier as with the 1-step classifier with more careful feature engineering. We leave this for future work.

5.5 Final results

Table 7 shows our submitted and final results on the TED dev set and on the WMT 2016 official test set. For the submitted system we unfortunately had a bug in the feature extraction for de-en and fr-en, which severely affected the scores, so for these systems we also show scores with the bug corrected. For the dev set we see that we could considerably improve the submitted scores by more careful feature engineering for all language pairs except de-en, but that we had worse or equal results for this feature set with large training data.

For the test set the primary 2-step system was better than the 1-step system only for translation from English. The final feature set helped mainly for de-en, which it did not on the dev set. For en-de and en-fr the final 1-step system did not beat the submitted 2-step system, as it did for the dev set. Adding more training data gave improvements or nearly equal scores for all language pairs. The discrepancy of the results between the dev and test sets could partly be explained by the different distribution of pronouns, especially for the rare classes that are important for Macro-R. It is also likely that our classifier has over-fitted somewhat to our dev data. In the workshop our best submitted system ended up in 2nd place for en-fr, which had the highest number of submissions.

6 Conclusion

We described the UU-Stymne system for the WMT shared task on cross-lingual pronoun prediction. We used linear SVMs with a high number of features, the most successful being local context, especially in German and French, source pronouns, and dependency heads. For the binary choice between pronoun and *other* we found part-of-speech patterns highly useful.

Acknowledgments

This work was supported by the Swedish strategic research programme eSSENCE.

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*, pages 1455–1465, Jeju Island, Korea.

Jimmy Callin, Christian Hardmeier, and Jörg Tiedemann. 2015. Part-of-speech driven cross-lingual pronoun prediction with feed-forward neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 59–64, Lisbon, Portugal.

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.

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 (WMT16)*, Berlin, Germany.

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

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.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

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*, pages 78–85, Lisbon, Portugal.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Ngoc-Quan Pham and Lonneke van der Plas. 2015. Predicting pronouns across languages with continuous word spaces. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 101–107, Lisbon, Portugal.

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.

Dominikus Wetzel, Adam Lopez, and Bonnie Webber. 2015. A maximum entropy classifier for cross-lingual pronoun prediction. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 115–121, Lisbon, Portugal.