Cross-lingual Pronoun Prediction for English, French and German with Maximum Entropy Classification

Dominikus Wetzel
School of Informatics
University of Edinburgh
10 Crichton Street, Edinburgh
d.wetzel@ed.ac.uk

Abstract

We present our submission to the cross-lingual pronoun prediction (CLPP) shared task for English-German and English-French at the First Conference on Machine Translation (WMT16). We trained a Maximum Entropy (MaxEnt) classifier based on features from Wetzel et al. (2015), that we adapted to the new task and applied to a new language pair. Additional features such as n-grams of the pronoun context and prediction of NULL-translations proved helpful to a varying degree. Experiments with a sequence classifier over pronoun sequences did not show any improvements. Our submission is among the top three systems for English-French (61.62% macro-averaged recall) and in the middle range for English-German (48.72%) out of nine submissions.

1 Introduction

Translation of pronouns is a non-trivial task due to ambiguities in the source language (event pronouns, referential and non-referential uses) and due to diverging usage of pronouns between two languages (e.g. morphological differences including gender and number, pro-drop languages, preference of passive construction with expletive *it*). In the recent past there has been work on analysing these differences and various approaches to tackle the problem exist (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010; Guillou, 2012; Weiner, 2014; Hardmeier et al., 2014; Guillou et al., 2014) including the submissions to the CLPP shared task (Hardmeier et al., 2015).

This shared task is organized again this year (Guillou et al., 2016). In addition to the English-French language pair, it introduces data sets for English-German, as well as the inverse translation directions from French and German into English. The task is to predict a target-side pronoun from a closed set of classes for each subject-position 3rd person pronoun in the source language.

One of the major differences to the shared task from last year is the target-side data. It comes in the form of lemmatized tokens with their Part-of-Speech (POS) tag, instead of the full word forms. This makes the task more challenging, since agreement features of words surrounding a pronoun are no longer available. For example all the determiners are mapped to one generic form irrespective of their gender or number. One can also argue that it makes the task more realistic, when considering Statistical Machine Translation (SMT) as the driving goal. SMT systems do not necessarily produce the correct target-side surface word forms and approaches to pronoun translation should not rely on error-free translations of the relevant context. This change therefore helps with handling more noisy or underspecified input.

In this paper we focus on learning to predict translations of pronouns from English into French and German. The set of source pronouns (i.e. *it* and *they*) is the same for both language pairs. For French, the closed target classes are: *ce, elle, elles, il, ils, cela, on,* OTHER and for German they are: *er, sie, es, man,* OTHER.

We use a MaxEnt classification model to learn pronoun predictions. This work is based on findings in (Wetzel et al., 2015). We incorporate source- and target-side bag-of-words context window features based on tokens and POS tags, a target-side pronoun antecedent feature and a target-side Language Model (LM) feature. Furthermore, we focus on predicting cases where the source pronoun does not have a corresponding translation and is therefore aligned to a special

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 620–626,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

NONE token. We conduct additional experiments in an attempt to exploit the sequential character of coreference chains that contain pronouns by using linear-chain Conditional Random Fields (CRFs).

2 Related Work

The CLPP shared task from last year (Hardmeier et al., 2015) had eight contributions and a very strong baseline. The official macro-averaged F1 metric ranked the baseline highest, however in terms of accuracy, a few of the submission managed to perform better.

Tiedemann (2015) explores models for CLPP with the focus on using only simple features. The major simplification is that no coreference resolution is performed. Experiments on using a sequence model for classification are reported, which makes predictions based on previous classification choices. However, only a degradation of performance was observed. One possible reason for that is that not every preceding classification choice corresponds to a mention of the same entity, and hence should only influence the current choice if it does. This distinction was not captured by (Tiedemann, 2015). We also explore the usefulness of a sequence classifier, however our sequences are more informed in that they follow automatically resolved source-side coreference chains.

Pham and van der Plas (2015) train a Multi-Layer Perceptron. Features consist of word-embeddings of local context words, averaged word vectors of target-side antecedents of a pronoun obtained via automatic coreference chains from the source projected to the target side via word-alignments and additional vectors containing morphological information. They use a subset of the types of our features, however integration is via word-embeddings and training is based on Neural Networks. They could not find any improvements when including target-side antecedents via source side coreference chains.

3 Features

In this section we motivate and describe the types of features we extract for learning the MaxEnt classifier and the CRF models. For a more detailed description of the features from last year, please refer to (Wetzel et al., 2015).

3.1 Context window

For each training instance, i.e. for each source pronoun for which we want a prediction, we extract a bag of words consisting of the ±3 tokens around the source pronoun. Additionally, we extract the tokens in the ±3 context window of the aligned target pronoun. The source-side feature consists of tokens in their full form, whereas the target-side feature uses the lemmatized tokens from the training data.

Additionally, we extract POS tags for these tokens. For the source side we automatically obtain POS tags with StanfordCoreNLP (Lee et al., 2013). For the target side the POS tags are provided as part of the training and test data.

A common strategy to improve linear classifiers is to include combinations of features so that the classifier can tune additional weights if predictive n-gram combinations provide useful information. Therefore, we experiment with combining the above context window features within each type. In addition to the unigram values, we extract n-gram values by concatenating adjacent tokens or POS tags.

All of the above features are extracted both from the source and the target side.

3.2 Pleonastic pronouns

Pleonastic pronouns are non-referential pronouns, i.e. they do not have an antecedent in the discourse. They behave differently compared to referential pronouns, e.g. grammatical agreement requirements do not exist. We use Nada (Bergsma and Yarowsky, 2011) to get an estimate if a particular pronoun is pleonastic and integrate this estimate directly as feature value into our classifier.

Furthermore, the Stanford deterministic coreference system (Lee et al., 2013), which we use in the feature described in Section 3.4, only has a very basic rule-based detection mechanism for pleonastic pronouns. Intuitively, Nada's estimates should therefore counterbalance erroneous handling in coreference resolution.

This feature is only applied on the source side.

3.3 Language Model prediction

LMs provide a probability of a sequence of words trained on large monolingual corpora and are used in SMT as a model to encourage fluency, i.e. producing typical target-language sentences. Wetzel et al. (2015) incorporated a LM feature based on the preceding 5-gram context of a target pronoun, by utilising the conditional probability $P(classLabel_5|w_1, w_2, w_3, w_4)$, where $classLabel$ is one of the class labels from the closed set of target classes, or the OTHER class, and w are the preceding words. This ignored any information following the pronoun, which could as well be indicative of the correct prediction. Therefore, we expand the feature to provide a rating for the entire sentence, i.e. $P(\langle s \rangle, w_1, ..., classLabel, ..., w_n, \langle /s \rangle)$, where n is the sentence length, and $\langle s \rangle$ and $\langle /s \rangle$ are sentence boundary markers.

The class label that produces the highest scoring sentence according to the LM is then used as a feature value in our classifier. To obtain such a prediction for the class labels that correspond to pronouns we can directly substitute the target-side pronoun placeholder with each class label when querying the LM.

The OTHER class requires special treatment, since it does not occur as such in the LM training data. We approximate the probability for this class in the same way as described in (Wetzel et al., 2015). We first collect frequencies of words that are tagged as OTHER from the training data. Then we query the LM with the top-n words as substitute for the placeholder. The highest scoring word within that group then competes as representative for OTHER against the probabilities of the rest of the class labels.

This feature is only applied on the target side.

3.4 Antecedent information

The antecedent feature proved useful in (Wetzel et al., 2015). Intuitively, if we know the closest target-side antecedent of a referential target-side pronoun, we have access to additional information such as grammatical gender and number. Both in German and French, the pronoun has to agree in gender and number with its antecedent. Furthermore, the fact whether we find an antecedent at all should be useful information as well, since it separates referential from non-referential cases.

We perform antecedent detection with the help of source-side coreference chains. We follow the source-side chain that contains the source pronoun of interest in reverse order (i.e. towards the beginning of the document) and check if the token that is aligned to the source-side mention head is a noun. If it is not, the search proceeds. The reason why we do not just search for the closest noun-antecedent on the source side and then take its projection is that nouns do not necessarily have to align to nouns, but could be aligned to NULL, pronouns, etc. We take the closest noun that we can find on the target side.

Corpus	en-de	en-fr
NC9	63.72	25.12
IWSLT15	68.55	31.25
TEDdev	60.00	34.31

Table 1: Percentage of NONE within the OTHER class.

Since the target side only contains lemma information, where all gender- or number-specific information has been removed from nouns (or merged to the same token for e.g. determiners), we cannot apply a morphological tagger to give us this information. Therefore, we resort to a simpler method and look up the most frequent gender for a given lemma in a lexicon. We only experiment with this feature on the English-German task.

All of the above features are extracted from the target side (with the help of source-side annotation).

3.5 Predicting NONE

Source pronouns do not necessarily have a counterpart in the target language. These cases are recorded in the training data with NONE labels and occur very frequently (cf. Table 1). However, they are not part of the official set of class labels and mapped to the OTHER class for training and testing. If we know that a source pronoun does not have a translation, then this might be useful in an SMT scenario, where a feature function could score phrases higher that do not contain target-side pronouns. For CLPP our expectation is that it should help to improve prediction performance for the very heterogeneous OTHER class.

For training the classifiers we therefore first map all NONE cases from OTHER to NONE, train with the above features and map the final predictions back to OTHER before evaluation.

3.6 Pronoun prediction in a sequence

The MaxEnt classifier makes the assumption that the translation of the pronoun is only dependent on the source and target contexts and the antecedent

Sequence length	%
1	74.45
2	15.34
3	5.34
4	2.21
5	1.08

Table 2: Percentage of sequence lengths up to 5 in the English-German training data (IWSLT15 and NC9) for the ALLINONE setup.

Gender	Frequency
Masculine	20878
Feminine	21221
Neuter	12894
Total	54993

Table 3: Number of nouns with gender information in the raw Zmorge lexicon (zmorge-lexicon-20150315) for German.

it refers to (for referential pronouns). This ignores the fact that pronouns are part of a longer chain of co-referring expressions, among them other pronouns.

Therefore, we first prepare the training and test data such that all pronoun instances that belong to the same coreference chain form one training or testing sequence. We then train a linear-chain CRF with the same features as given above instead of a MaxEnt classifier to predict an optimal sequence of target pronouns, rather than making each prediction independently of the other pronouns. This way, typical patterns of pronoun sequences can be learnt, which might help with the prediction. Table 2 gives the distribution of sequence lengths.

4 Experiments

We first describe the experimental setup of our systems, then briefly describe the data we used and provide information about feature and parameter settings. Finally, we report our results on development and test data.

4.1 Systems

We use Mallet (McCallum, 2002) for training the MaxEnt classifiers and CRF models. For the MaxEnt classifier we use the default settings. For the CRF we train *three-quarter* order models (i.e. one weight for each ⟨feature, label⟩ pair, and one for each ⟨current label, previous label⟩ pair) and only allow label transitions that have been observed in the training data.

In all experiments, we have two setups. The POSTCOMBINED setup, where we split the training and test data for each source pronoun into separate sets, train separate classifiers and combine the predictions after classification. And the ALLINONE setup, where we do not split the data.

The systems marked with *initial* consist of the context window features, the pleonastic pronoun feature, the LM feature and the antecedent information (without gender information). We use *fGender* to refer to the gender feature, *3-gram window* to refer to the n-grams from the context window and *fNone* to refer to the NONE-prediction feature. Systems marked with *sequence* are the CRF models. We submit the best performing system according to the official macro-averaged recall measure on the development set for each language pair as primary test set submission.

The official BASELINE uses LM predictions similarly to our LM feature. Additionally, it attempts to find the optimal predictions for a sentence, if there are multiple pronouns that have to be predicted. It has a NULL penalty parameter that determines the influence of not predicting a pronoun at all. For a more detailed description, please refer to the shared task paper (Guillou et al., 2016).

4.2 Data

For training, we only extract information from the IWSLT15 and NewsCommentary (NC9) corpus. We do not employ the provided Europarl corpus, as it does not come with predefined document boundaries other than parliamentary sessions of a complete day. For development, we use the TEDdev set. For the final submission on the official test set we include TEDdev in the training data.

4.3 Features and parameters

For the LM feature, we take the provided trained models from the shared task, which are 5-gram modified Kneser-Ney LMs that work on lemmatized text. We use KenLM (Heafield, 2011) for obtaining probabilities. As proxy for the OTHER class we use the top 35 words for German, and the top 70 for French.

	Mac-R	Acc
BASELINE	34.35	42.81
ALLINONE-initial	39.24	56.14
+ fGender	40.00	57.37
+ fGender, 3-gram window	41.21	57.72
+ fGender, 3-gram win, fNone	40.86	58.77
ALLINONE-sequence-initial	35.67	54.91

Table 4: System performance in percent for English-German on the development data set.

For gender detection of German antecedents we use the lexicon from Zmorge (Sennrich and Kunz, 2014). Gender distribution of nouns is given in Table 3. When a noun has multiple genders in the lexicon, we take the most frequent one for that noun.

The different parameters such as context window size were taken from our findings of the previous year (Wetzel et al., 2015). The n-grams of the context window are extracted for n=1..3 including beginning- and end-of-sentence markers if necessary.

4.4 Results

The results on the development set are given in Table 4 for English-German and in Table 5 for English-French. The final results including the ranks on the official test set of the shared task are given in Table 6.

The initial systems in each language-pair perform much better than the baseline, which is especially noticeable in English-French. Adding the gender feature to the English-German classifier shows some good improvements in performance, thereby confirming the usefulness of adding gender information.

The additional feature that predicts NONE as possible translation is helpful for the English-French pair. Results on English-German showed a decrease in performance with respect to macro-averaged recall. This decrease is surprising, especially considering the much larger frequency of NONE in the German data set (cf. Table 1).

5 Discussion

In general, performance is considerably lower for English-German compared to English-French, despite the former having a much smaller set of class

	Mac-R	Acc
BASELINE	40.63	49.73
ALLINONE-initial	52.25	69.98
+ 3-gram window	54.68	73.36
+ 3-gram window, fNone	57.34	74.25
ALLINONE-sequence-initial	49.27	64.65

Table 5: System performance in percent for English-French on the development data set.

	en-de		en-fr	
	Mac-R	Acc	Mac-R	Acc
ALLINONE	48.72$_5$	66.32$_6$	61.62$_4$	71.31$_3$
POSTCOMBINED	47.75	64.75	59.83	68.63
BASELINE-1	n/a	n/a	50.85	53.35
BASELINE-2	47.86	54.31	n/a	n/a

Table 6: Official shared task results. Ranks of our primary submission are given in subscripts with a total of nine submissions for each language pair.

	er	sie	es	man	OTHER	*Total*
er	4/4	2/2	3/8	·	6/1	15
sie	3/2	73/100	11/15	3/·	34/7	124
es	2/·	9/4	61/85	2/·	27/12	101
man	·	·/1	2/4	1/1	5/2	8
OTHER	2/1	11/17	7/16	·	115/101	135
Total	11/7	95/124	84/128	6/1	187/123	383

	ce	elle	elles	il	ils	cela	on	OTHER	*Total*
ce	58/60	·	·	6/6	·/1	1/·	·	3/1	68
elle	2/2	10/9	2/·	5/8	·/1	2/3	·	2/·	23
elles	2/·	2/·	3/6	·	15/17	1/·	·/1	2/1	25
il	5/6	1/6	·	43/43	2/1	4/3	2/2	4/·	61
ils	·	·	9/7	·	54/63	·	·	8/1	71
cela	·	3/1	·	8/7	·	15/20	1/1	4/2	31
on	·	·	·	·/1	2/4	·	6/4	1/·	9
OTHER	1/3	1/·	·	4/7	1/·	1/1	·/2	77/72	85
Total	68/71	17/16	14/13	66/72	74/87	24/27	9/10	101/77	373

Table 7: Confusion matrices for the ALLINONE classifier on the English-German (top) and English-French (bottom) test set. Row labels are gold labels and column labels are labels as they were classified. Dots represent zeros. Numbers to the left represent our shared task submissions, numbers to the right are for the results when we removed the LM feature from these submissions.

	en-de		en-fr	
	Mac-R	Acc	Mac-R	Acc
ALLINONE	48.72	66.32	61.62	71.31
− fAntecedent	46.24	64.23	61.89	71.85
− fLM	55.76	75.98	63.03	74.26

Table 8: Feature ablation results on the test set when removing the antecedent or LM feature from our submitted systems.

labels to choose from. One reason for that might be that in the former setting, the OTHER class is even more heterogeneous than in French, and taking apart this class to the same degree as in the English-French data sets might be beneficial.

Performance between development and test sets varies greatly despite similar class label distributions (except for a much smaller amount of OTHER instances in the English-French test set). To a certain degree this is expected, however the big changes in performance suggest that there are other differences in the data sets which are worth exploring.

Training a MaxEnt classifier where we substitute our LM feature with predictions from the shared task baseline performed slightly worse. This suggests that a simpler LM feature is sufficient when included in the classifier, and that joint prediction of multiple target pronouns within one sentence is not necessary. However, we did not tune the NULL penalty of the baseline model.

The confusion matrix for English-German in Table 7 (top-left) shows that OTHER is overpredicted, which might explain the overall lower performance of the system compared to other participants. Furthermore, *es* and *sie* are confused by our classifier. For English-French in Table 7 (bottom-left) one can observe that the biggest confusion is between gender in plural pronouns (i.e. *elles* and *ils*). This might be because we did not include any explicit gender information as feature. As above, the OTHER class is also very confused over all cases.

Similarly to our findings from last year, the POSTCOMBINED setup scored consistently worse on the test sets (and only once slightly better on the development set). This provides evidence, that splitting the training data according to source pronouns is counterproductive. Furthermore, it might even be worse for the inverse prediction tasks, since there are a lot more source pronouns, hence making the available data even sparser.

The lemmatization of the French data merges singular and plural forms of *il* into one lemma, similarly for *elle*. The baseline which uses the LM trained on the lemmatized data is therefore never able to predict the plural forms of these two pronouns, resulting in zero precision and recall. This is confirmed by the corresponding confusion matrix. This might also have an indirect impact on the performance of our classifiers, since they use LM prediction as a feature.

Feature ablation experiments shown in Table 8 revealed that the antecedent feature is helpful for English-German, but not for English-French. One possible explanation for this might be that we do not have gender information of the antecedent in French and only adding the antecedent itself might not be sufficient.

Additional ablation experiments showed that the LM feature in fact hurts performance. Removing this feature gives a boost in performance, which brings our systems to the second place (first for accuracy) for English-German and to the third place (second for accuracy) for English-French. This contradicts findings from experiments we conducted for last year's shared task, where adding baseline predictions, which are very similar to our LM feature, greatly improved results. An explanation for this behaviour could be that the LM this year was trained on lemmatized text and therefore performs much worse than when trained on original data. Confusion matrices for these results are given in Table 7 (numbers to the right). For both language pairs we are now underpredicting OTHER, however gaining accuracy on the classes representing pronouns.

6 Conclusion

We experimented with MaxEnt classifiers for CLPP applied to English-German and English-French. Some of the features are only useful for one of the two language pairs. Adding LM predictions considerably worsened performance, which is contrary to experiments performed on last year's shared task. Modelling pronoun sequences with CRFs did not prove useful at all.

The greatly varying degree of performance between development and test sets relativizes any findings of the shared task, and it should be further investigated what the cause of that is.

Acknowledgments

This research has been funded through the European Union's Horizon 2020 research and innovation programme under grant agreement 644402 (HimL).

References

Shane Bergsma and David Yarowsky. 2011. NADA: A robust system for non-referential pronoun detection. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 12–23, Faro, Portugal, October.

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie Webber. 2014. Parcor 1.0: A parallel pronoun-coreference corpus to support statistical mt. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France, April. Association for Computational Linguistics.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, Aaron Smith, and Joakim Nivre. 2014. Anaphora models and reordering for phrase-based smt. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 122–129, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

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. `http://www.idiap.ch/workshop/DiscoMT/shared-task`.

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

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden, July. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

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, September. Association for Computational Linguistics.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A german morphological lexicon extracted from wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

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.

Jochen Weiner. 2014. Pronominal anaphora in machine translation. Master's thesis, Karlsruhe Institute of Technology, January.

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, September. Association for Computational Linguistics.