A Maximum Entropy Classifier for Cross-Lingual Pronoun Prediction

Dominikus Wetzel and **Adam Lopez** and **Bonnie Webber**
School of Informatics
University of Edinburgh
11 Crichton Street, Edinburgh
d.wetzel@ed.ac.uk, {alopez,bonnie}@inf.ed.ac.uk

Abstract

We present a maximum entropy classifier for cross-lingual pronoun prediction. The features are based on local source- and target-side contexts and antecedent information obtained by a co-reference resolution system. With only a small set of feature types our best performing system achieves an accuracy of 72.31%. According to the shared task's official macro-averaged F1-score at 57.07%, we are among the top systems, at position three out of 14. Feature ablation results show the important role of target-side information in general and of the resolved target-side antecedent in particular for predicting the correct classes.

1 Introduction

In this paper we focus on pronouns which pose a problem for machine translation (MT). Pronoun translation is challenging due to the fact that pronouns often refer to entities mentioned in a non-local context such as previous clauses or sentences. Furthermore, languages differ with respect to usage of pronouns, e.g. how they agree with their antecedent or whether source and target language exhibit similar patterns of pronoun usage. Since pronouns contribute an important part to the meaning of an utterance, the meaning can be changed considerably when wrongly resolved and translated.

This problem gained recent interest and work has been presented in annotating and analysing translations of pronouns in parallel corpora (Guillou et al., 2014) and MT systems focusing on translation of pronouns have been proposed (Hardmeier and Federico, 2010; Le Nagard and Koehn, 2010; Guillou, 2012; Hardmeier et al., 2014).

The DiscoMT 2015 shared task on pronoun translation (Hardmeier et al., 2015) calls for contributions to tackle this problem.We focus on the cross-lingual pronoun prediction subtask, which is set up as follows: the two English (source language) third-person subject pronouns *it* and *they* can be translated in a variety of ways into French. A common set of nine classes (*ce, cela, elle, elles, il, ils, on, ça*) is defined as possible translations including an extra class OTHER which groups together any less frequent translations, including *null*, noun translations, alignment errors. The source and target corpora both consist of human-created documents and therefore abstract away from additional difficulties that arise with noisy automatic translations.

Hardmeier et al. (2013) propose a neural-network-based approach for a similar cross-lingual pronoun prediction task. Their model jointly models anaphora resolution and pronoun prediction. Our approach builds on a maximum entropy (MaxEnt) classifier that incorporates various features based on the source pronoun and local source- and target-side contexts. Moreover, the target-side noun referent (i.e. the *antecedent*) of a pronoun is used and obtained with an automatic co-reference resolution system. Our system achieves high accuracy and performs third-best according to the official evaluation metric.

In Section 2 we present our MaxEnt classifier including a description of the features used. This is followed by Section 3 with experiments and evaluation. Furthermore, in Section 4 we discuss the results and in Section 5 we give concluding remarks.

2 Systems for Cross-Lingual Pronoun Prediction

2.1 Maximum Entropy Classification

A MaxEnt classifier can model multinomial dependent variables (discrete class labels) given a set of independent variables (i.e. observations).

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 115–121,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

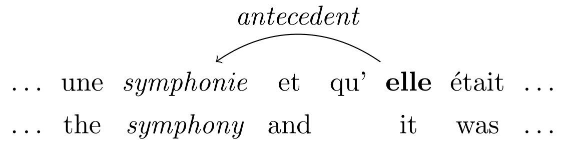
antecedent
... une symphonie et qu' elle était ...
... the symphony and it was ...

Figure 1: Antecedent of a pronoun within local context, which is also captured by a 5-gram language model.

Each observation is represented by a set of m features extracted from the observation. The m features can provide overlapping evidence, hence do not have to be independent of each other. The model consists of a function $f(x_i, y_i) \rightarrow \mathbb{R}^{m+1}$ that maps the i-th observation x and associated label y to a real valued vector. It also consists of a weight vector $\vec{\theta}$ of corresponding size, which contains the model parameters that are learned from the training data. The model is of the form

$$p(y|x) = \frac{exp\,\vec{\theta} \cdot f(x, y)}{Z(x)}$$

where $Z(x)$ is a normalizing factor ensuring valid probabilities.

2.2 Features

Local Context The local context around the source pronoun and target pronoun can contain the antecedent (cf. Figure 1) or other information, such as the inflection of a verb which can provide evidence for the gender or number of the target-side pronoun. Therefore, we include the tokens that are within a symmetric window of size 3 around the pronoun. We integrate this information as bag-of-words, but separate the feature space by source and target side vocabulary and whether the word occurs before or after the pronoun. Special BOS and EOS markers are included for contexts at the beginning or end of sentence, respectively. We neither remove stopwords nor normalize the tokens.

We also include as features, the *Part-of-Speech (POS) tags* in a 3-word window to each side of source and target pronouns. This gives some abstraction from the lexical surface form. For the source side we use the POS tags from Stanford CoreNLP (Manning et al., 2014) mapped to universal POS tags (Petrov et al., 2012). For the target side we use coarse-grained tags provided by Morfette (Chrupała et al., 2008).[1]

Language Model Prediction We include a target-side *Language Model (LM) prediction* as a feature for the classifier. A 5-gram LM is queried by providing the preceding four context words followed by one of the eight target-side pronouns that the class labels represent. The pronoun that has the highest prediction probability is the feature that we include in the training data. The ninth class OTHER requires special treatment, since it represents all other tokens that were observed in the aligned data and thus does not itself appear in the LM training data. To get an accurate prediction probability for this aggregate class one would have to iterate over the entire vocabulary V (excluding the other eight pronouns) and find the most likely token. Since this would require a huge amount of LM queries ($|V|\times$ number of training instances) we approximated this search by taking the 40 most frequent tokens that are observed in the training data in the position which was labelled as OTHER. The highest prediction probability is then used to compete with the probabilities of the other explicit classes. Once the most likely prediction is determined we included the predicted class label as feature.

Target-side Antecedent The *target-side noun antecedent* of the pronoun determines the morphological features the pronoun has to agree with, i.e. *number* and *gender*. We use the source-side co-reference resolution system provided by Stanford CoreNLP (Lee et al., 2013) to determine the co-reference chains in each document of the training data. We then project these chains to the target side via word-alignments (cf. Figure 2). The motivation to obtain target-side co-reference chains in that way is three-fold. First, the target side of the training data is missing most of the target-side pronouns since it is the task to predict them. Therefore, relevant parts of co-reference chains are missing and the place-holders for these pronouns will introduce noise to the resolution system. Secondly, we have a statistical machine translation (SMT) scenario in mind as an application for cross-lingual pronoun prediction. Applying a co-reference system to the noisy SMT output of already translated parts of the document is subjecting the system to much noisier data than it was originally developed for. Thirdly, resources

[1] https://github.com/gchrupala/morfette

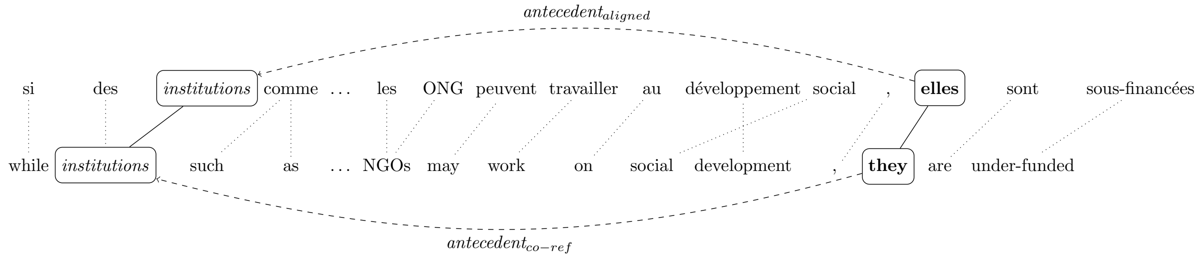

Figure 2: The $antecedent_{co-ref}$ of *they* on the English sentence (source language) is determined with a co-reference resolution system. The target-side $antecedent_{aligned}$ is obtained by following the word alignment links. In the shared task, the target pronoun *elles* has to be predicted.

and tools for automatic co-reference resolution are more easily available for English than for French.

Given the target-side co-reference chains in a document, we consider the chain the target-side pronoun is assigned to and greedily search for the closest noun token in the chain in the preceding context. This mention is included in the training data for the classifier as lexical feature. In addition, we extract morphological features from the noun (i.e. number and gender) by automatically analyzing the target-side sentences with Morfette.[2] In cases where the pronoun was not assigned to a co-reference chain, a special indicator feature was used. In addition, the word alignment can align one source token to multiple target tokens. We searched for the first noun in the aligned tokens and considered this to be the representative head antecedent of the given pronoun. If no noun could be found with this method, we resorted to taking the best representative antecedent of the source chain as determined by the Stanford co-reference system and took the aligned token as the relevant target-side antecedent. In this case *null* alignments are also possible and a special indicator feature is used for that.

Pleonastic Pronouns *Pleonastic pronouns* are a class of pronouns that do not have a referent in the discourse, e.g. in "*It* is raining". Their surface form in English is indistinguishable from referential forms. Nada (Bergsma and Yarowsky, 2011) is a tool that provides confidence estimates for pronouns whether they are referential.[3] We include these estimates as an additional feature. This should provide information especially for the French class labels that can be used as pleonastic pronouns, e.g. "*il* pleut (it is raining)" or "*ça* fait mal (it hurts)".

In addition, the rule-based detection of pleonastic pronouns is only basic in the Stanford co-reference system (Lee et al., 2013). However since they do not have a referent, they cannot be part of a co-reference chain. Therefore, we expect this feature to also counteract wrong decisions by the co-reference resolution system to a certain degree. Since Nada only provides estimates for *it*, we do not have such a feature for pleonastic uses of the other source pronoun of the task *they*.

2.3 Classifier Types

We trained classifiers in two different setups. The first setup provides all our extracted features as training data to one MaxEnt classifier, including the source pronoun as additional feature for each training instance (from now on referred to as the ALLINONE system). The second setup splits the training data into the two source pronoun cases (*it* and *they*) and trains a separate classifier for each of them (POSTCOMBINED system).

3 Experiments and Evaluation

3.1 Data

The shared task provides three corpora that can be used for training. The Europarl7 corpus, the NewsCommentary9 corpus and the IWSLT14 corpus which are transcripts of planned speech, i.e. TED talks. Only the latter two corpora come with natural text boundaries. Since these boundaries are necessary for co-reference resolution, we did not use the Europarl corpus. The test data contains 1105 classification instances within a total of

[2] Morfette's performance is quite robust and can handle sentences that contain $REPLACE\_xx$ tokens, which are the placeholders for target-side pronouns that have to be predicted. A comparison of the performance on the original sentences and the sentences with the $REPLACE\_xx$ tokens showed only minor differences.

[3] https://code.google.com/p/nada-nonref-pronoun-detector/

	fine Mac-F1	**coarse** Acc
BASELINE	58.40 (1)	68.42 (8)
ALLINONE	57.07 (3)	74.84 (6)
POSTCOMBINED	54.96 (7)	74.03 (7)

Table 1: Official performance on the test data. Ranks according to each metric are given in parenthesis out of 14 submitted systems (including multiple submissions per submitter and the baseline).

2093 sentences in twelve TED talk documents.

3.2 Classifier

We extract features from the training and test set and use Mallet (McCallum, 2002) to train the MaxEnt classifier.[4] The variance for regularizing the weights is set to 1 (default setting).

For the LM component of our system we use the baseline model provided for the pronoun translation subtask. This is a 5-gram modified Kneser-Ney LM trained with KenLM (Heafield, 2011).[5]

3.3 Evaluation Metrics

The official evaluation metric for the shared task is the macro-averaged F-score over all prediction classes (Mac-F1). Since this metric favours systems that perform equally well on all classes, the task puts emphasis on handling low-frequency classes well instead of only getting the frequent classes right. In addition to scores with the official metric we also report overall accuracy (Acc), i.e. the ratio between the correctly predicted classes and all test instances.

The evaluation script of the shared task provides results for the official fine-grained class separation with nine classes. It also provides a coarse-grained separation where some of the class labels are merged. Results reflect the fine-grained distinction except where stated.

3.4 Results on the Test Set

Table 1 shows the official results on the test set together with the respective ranks out of 14 submitted systems. Table 2 and Table 3 provide the per-class precision, recall and F1, overall accuracy, and overall macro-averaged F-score. Table 4 shows results of our feature ablation experiments.

	Prec	Recall	F1
ce	77.78	87.50	82.35
cela	25.00	18.52	21.28
elle	51.79	34.94	41.73
elles	85.00	33.33	47.89
il	50.00	59.62	54.39
ils	76.84	91.25	83.43
on	63.64	37.84	47.46
ça	62.69	41.18	49.70
OTHER	80.95	90.48	85.45
Macro-averaged	63.74	54.96	**57.07**
Accuracy			**72.31**

Table 2: Performance of **ALLINONE** classifier on the **test** set.

	Prec	Recall	F1
ce	78.05	86.96	82.26
cela	9.52	7.41	8.33
elle	49.06	31.33	38.24
elles	80.00	31.37	45.07
il	51.54	64.42	57.26
ils	75.79	90.00	82.29
on	61.90	35.14	44.83
ça	64.29	44.12	52.33
OTHER	80.00	88.52	84.04
Macro-averaged	61.13	53.25	**54.96**
Accuracy			**71.40**

Table 3: Performance of **POSTCOMBINED** classifier on the **test** set.

4 Discussion

Confusion Matrices Table 5 and Table 6 present confusion matrices on the test set. Divergences from strong diagonal values in both tables derive in part from gender-choice errors. In addition, the morphological number of the personal pronouns is almost perfectly predicted in all cases. The OTHER class causes quite a few confusions, which is not surprising since it aggregates a heterogeneous set of possible source pronoun translations. We expect a more detailed distinction in this group to lead to better systems in general.

[4] http://mallet.cs.umass.edu/

[5] http://kheafield.com/code/kenlm/

	allInOne		postCombined	
	Mac-F1	Acc	Mac-F1	Acc
all features	57.07	72.31	54.96	71.40
all w/o antecedent features	51.59	70.14	54.15	71.13
all w/o nada	50.86	69.86	54.84	71.40
all w/o morph	54.62	71.67	54.33	71.40
all w/o language model	54.83	71.13	55.32	71.59
only src features	34.81	55.20	34.41	54.84
only tgt features	55.05	71.49	54.82	71.31

Table 4: Feature ablation for both types of classifiers on the **test** set.

classified as →	ce	cela	elle	elles	il	ils	on	ça	OTHER	*Total*
ce	161	0	1	1	11	0	0	3	7	184
cela	0	5	2	0	4	0	0	9	7	27
elle	8	1	29	0	21	3	2	5	14	83
elles	2	0	0	17	0	28	0	0	4	51
il	12	1	12	0	62	1	4	2	10	104
ils	1	0	0	1	0	146	0	0	12	160
on	2	0	3	1	5	4	14	2	6	37
ça	6	12	7	0	18	0	1	42	16	102
OTHER	15	1	2	0	3	8	1	4	323	357
Total	207	20	56	20	124	190	22	67	399	1105

Table 5: Confusion matrix for the **allInOne** classifier on the **test** set. Row labels are gold labels and column labels are labels as they were classified.

classified as →	ce	cela	elle	elles	il	ils	on	ça	OTHER	*Total*
ce	160	0	2	0	11	1	0	3	7	184
cela	0	2	1	1	5	0	0	8	10	27
elle	10	0	26	0	23	3	3	6	12	83
elles	2	0	1	16	0	28	0	0	4	51
il	9	1	10	1	67	1	2	2	11	104
ils	0	0	0	2	0	144	0	1	13	160
on	2	0	5	0	6	4	13	2	5	37
ça	5	14	6	0	14	0	1	45	17	102
OTHER	17	4	2	0	4	9	2	3	316	357
Total	205	21	53	20	130	190	21	70	395	1105

Table 6: Confusion matrix for the **postCombined** classifier on the **test** set. Row labels are gold labels and column labels are labels as they were classified.

Feature Ablation In order to investigate the usefulness of the different types of features, we performed a feature ablation. When removing all features that are related to the antecedent of the target pronoun we need to predict, i.e. the antecedent itself and its number and gender, we observe a considerable drop in performance for both evaluation metrics. This is according to our expectations, since number and gender are strong cues for most of the classes. The antecedent token itself also provides enough information to the classifier to make a positive impact on the results .

When removing all features related to the target side we can observe a consistent drop in performance over all sets and classifiers.[6] This result shows the important influence the target language has on the translation of a source pronoun. Removing the source-side features does not have a strong impact on the results, which is consistent again over all settings. Both results taken together strongly indicate that the target-side features are much more important than the source-side features.

Classifier Types The overall results show a consistent preference for the ALLINONE classifier over the POSTCOMBINED one. The difference in performance seems to be mostly influenced by the fact that splitting the training data into two separate sets for the POSTCOMBINED setting also results in much smaller data sizes for each of the individual classifiers. Our feature ablation results show that particular features are useful for the former classifier, but useless or even harmful for the latter. This instability might be due to the fact that the POSTCOMBINED classifier has to learn from much smaller data sets. Incorporating more training data from the Europarl corpus could alleviate this problem and would make it possible to determine whether these differences persist.

Language Model The mixed results for the usefulness of the LM features prompt for a further investigation of how to integrate the LM. Currently we base the LM predictions on the preceding n-gram of the target pronoun. However, it is also conceivable for this task to query the LM with n-grams that are within a sliding window of tokens containing the target pronoun. Furthermore, there is a small mismatch between the trained LM which has been trained on truecased data and the preceding tokens we have from the shared task data where the case was not modified. If this difference is eliminated we expect more accurate LM predictions, which should then in turn provide more accurate features for the classifiers.

Additionally, our LM feature currently predicts OTHER with a fairly high frequency of around 80% (followed by *il* with around 15%). This might be another reason why some classifiers work better without this feature, since this distribution does not match the observed distribution of target pronouns in the training data.

5 Conclusion

We presented a MaxEnt classifier that can determine the French translation of the English 3rd person subject pronouns with fairly high accuracy and performs among the top systems that have been submitted for this task. The classifier only uses a small set of feature types. Target-side features contribute most to the classification quality. Potentially non-local target-side antecedent features obtained via a source-side co-reference system and projected to the target via word alignments provide useful information as well.

Acknowledgments

This paper has received funding from 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.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

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 Eval-*

[6] Features related to the target side are the LM, the target side context windows (lexical tokens and POS tags), the antecedent of the target pronoun (lexical token and morphological features).

uation (LREC'14), Reykjavik, Iceland, May. European Language Resources Association (ELRA).

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, 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, October. Association for Computational Linguistics.

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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).