It-disambiguation and source-aware language models for cross-lingual pronoun prediction

Sharid Loáiciga, Liane Guillou & Christian Hardmeier

University of Geneva, Brainn Wave Technologies Limited, Uppsala University
sharid.loaiciga@unige.ch, liane@brainnwave.com, christian.hardmeier@lingfil.uu.se

1. Introduction

In this paper we introduce a classifier to disambiguate the various uses of the pronoun "it". We then integrate the classifier into an n-gram model for pronoun prediction. This system was built for the recent WMT 2016 shared task on cross-lingual pronoun prediction (Guillou et al., 2016).

Both pronouns "it" and "they" perform multiple functions in text, and disambiguation is required if they are to be translated correctly into other languages (Guillou, 2016). The pronoun "they" is typically used as an anaphoric pronoun, but may also be used generically, for example in "**They** say it always rains in Scotland". The pronoun "it" may be used as an anaphoric, pleonastic or event reference pronoun. Examples of these pronoun functions are provided in Figure 1.

anaphoric	I have a **bicycle**. **It** is red.
pleonastic	**It** is raining.
event	He lost his job. **It** came as a total surprise.

Figure 1: Examples of different pronoun functions

Anaphoric pronouns corefer with a noun phrase (i.e. the *antecedent*). *Pleonastic* pronouns, in contrast, do not refer to anything but are required to fill the subject position in many languages. *Event reference* pronouns may refer to a verb, verb phrase, clause or even an entire sentence.

In this work, we focus on the French translation of the English pronoun "it". However, the method of disambiguating the function of a source language pronoun is not limited to this case; other pronouns may require disambiguation for different language pairs, i.e., where a pronoun in the source language requires different target-language translations depending on its function. In the case of French, for example, anaphoric "it" may be translated with the third-person singular pronouns "il" [masc.] and "elle" [fem.], or with a non-gendered demonstrative such as "cela". The French pronoun "ce" may function as both an event reference and a pleonastic pronoun, but "il" is used only as a pleonastic pronoun.

2. Disambiguating "it"

The ParCor corpus (Guillou et al., 2014) and *DiscoMT2015.test* dataset (Hardmeier et al., 2016) were used to train the classifier. Under the ParCor annotation scheme, which was used to annotate both corpora, pronouns are labelled according to their function. For all instances of "it" labelled as anaphoric, pleonastic or event reference, the sentence-internal position of the pronoun and the sentence itself are extracted. The data was divided into 1,504 instances for training, and 501 each for the development and test sets. All sentences were shuffled before the corpus was divided, promoting a balanced distribution of the classes (Table 1).

Data	*it-*			
Set	Event	Anaphoric	Pleonastic	Total
Training	504	779	221	1504
Dev	157	252	92	501
Test	169	270	62	501
Total	830	1301	375	2506

Table 1: Distribution of classes in the training data

2.1 Baselines

For development and comparison we built two different baselines. One is a 3-gram language model built using KenLM (Heafield, 2011) and trained on a modified version of the annotated corpus in which every "it" is concatenated with its type (e.g. *it-event*). For testing, the "it" position is filled with each of the three *it-labels* and the language model is queried. Table 3 presents the results of this baseline using 14-fold cross-validation and a single held-out test set. The motivation for the choice of the number of folds is threefold. First, we wanted to respect document boundaries; second, we aimed for a fair proportion of the three classes in all folds; and, lastly, we tried to lessen the variance given the relatively small size of the corpus.

The second baseline is a setting in which all instances of the test set are set to the majority class *it-anaphoric*. The majority class baseline for the 14-fold cross-validation is equivalent to setting all of the labels in the corpus to *it-anaphoric*.

2.2 Classification Experiments and Results

All classifiers were trained using the Stanford Maximum Entropy package (Manning and Klein, 2003). To extract most of our features, we parse the corpus with the joint part-of-speech tagger and dependency parser of Bohnet et al. (2013). Additionally, the corpus was lemmatised using the TreeTagger lemmatiser (Schmid, 1994). For each training example, we extract the following information:

1. Previous three tokens. This includes words and punctuation. It also includes the tokens in the previous sen-

	14-fold cross-validation			Test-set		
	Precision	Recall	F1	Precision	Recall	F1
it- anaphoric	0.599	0.248	0.350	0.732	0.262	0.387
it- pleonastic	0.152	0.621	0.244	0.139	0.694	0.231
it- event	0.528	0.277	0.363	0.521	0.290	0.373

Table 2: N-gram baseline for the classification of the three types of “it”.

	14-fold cross-validation			Test-set		
	Precision	Recall	F1	Precision	Recall	F1
it- anaphoric	0.519	1	0.683	0.539	1	0.700

Table 3: Majority class baseline for the classification of the three types of “it”.

tence when the “it” occupies the first position of the current sentence.

2. Next two tokens

3. Lemmas of the next two tokens

4. Head word. As the task is limited to subject “it” and “they”, most of the time the head word is a verb.

5. Whether the head word takes a ‘that’ complement (verbs only)

6. Tense of head word (verbs only). This is computed using the rules described in Loáiciga et al. (2014).

7. Presence of ‘that’ complement in previous sentence. A binary feature which follows Navarretta (2004)’s conclusion (for Danish) that a particular demonstrative pronoun (*dette*) is often used to refer to the last mentioned situation in the previous sentence, often expressed in a subordinated clause.

8. Predications head. This refers to the predicative complements of the verbs *be*, *appear*, *seem*, *look*, *sound*, *smell*, *taste*, *feel*, *become* and *get*.

9. Closest noun phrase (head) to the left

10. Closest noun phrase (head) to the right

11. Presence of a cleft construction. A binary feature which refers to constructions containing adjectives which trigger extraposed sentential subjects as in ‘*So it’s difficult to attack malaria from inside malarious societies, [...].*

12. Closest adjective to the right

13. VerbNet selectional restrictions of the verb. VerbNet (Kipper et al., 2008) specifies 36 types of argument that verbs can take. We limited ourselves to the values of ‘abstract’, ‘concrete’ and ‘unknown’.

14. Lemma of the head word

15. Likelihood of head word taking an event subject (verbs only). An estimate of the likelihood of a verb taking a event subject was computed over the Annotated English Gigaword v.5 corpus (Napoles et al., 2012). We considered two cases where an event subject appears often and may be identified by exploiting the parse annotation of the Gigaword corpus. The first case is when the subject is a gerund and the second case is composed of “this” pronoun subjects.

16. NADA probability. The probability that the non-referential “it” detector, NADA (Bergsma and Yarowsky, 2011), assigns to the instance of “it”.

A comparison of the baselines (Table 3) and the classification results (Table 4) shows that predicting event reference pronouns is a complex problem. A manual inspection of the results shows that discriminating between anaphoric and event reference instances of “it” is indeed a very subtle process. Determining the presence or the lack of a specific (np-like) antecedent requires the understanding of the complete coreference chain. The 3-gram baseline appears to be biased towards the pleonastic class, as suggested by its high precision and very low recall for the event and anaphoric classes and the opposite situation for the pleonastic class. While our own classifier is more balanced, it achieves only moderate results with the event class. Compared to both of the baselines, it shows only a very small improvement.

It is worth noting that from the 2,031 segments composing the annotated corpus, 349 (17%) contain co-occurrences of between 2 and 7 “it” pronouns within the same segment. We experimented with including the previous *it-label*, when there are several within the same sentence, as an additional feature and obtained important gains in performance. It can be seen in the *w/ oracle feature* section of Table 4 that performance is improved for the anaphoric and event classes but not for the pleonastic class. This outcome is explained by the fact that both anaphoric and event reference pronouns are intrinsically referential and therefore potentially part of a bigger coreferential chain including several pronouns. Pleonastic pronouns, on the other hand, are syntactically required but do not corefer.

3. Source-Aware Language Model with Disambiguation Labels

The pronoun prediction part of our model is based on an n-gram model over target lemmas. In addition to the pure target lemma context, our model also has access to the identity of the source language pronoun, which, in the absence of number inflection on the target words, provides valuable

	14-folds cross-validation			Test-set		
w/o oracle feature	Precision	Recall	F1	Precision	Recall	F1
it- anaphoric	0.627	0.741	0.676	0.716	0.756	0.735
it- pleonastic	0.692	0.565	0.613	0.750	0.726	0.738
it- event	0.579	0.475	0.519	0.564	0.521	0.542
w/ oracle feature	Precision	Recall	F1	Precision	Recall	F1
it- anaphoric	0.632	0.750	0.683	0.729	0.785	0.756
it- pleonastic	0.660	0.581	0.607	0.705	0.694	0.699
it- event	0.596	0.502	0.542	0.611	0.538	0.572

Table 4: Classification results of the three types of "it" using cross-validation and a single test set.

Source:	**It** 's got these fishing lures on the bottom .
Target lemmas:	**REPLACE_0** avoir ce leurre de płche au-dessous .
Solution:	*ils*
LM training data:	It **REPLACE_*ils*** avoir ce leurre de płche au-dessous .
LM test data:	It **REPLACE** avoir ce leurre de płche au-dessous .

Figure 2: Data for the source-aware language model. In the WMT 2016 shared task data, REPLACE items substitute the target pronouns to be predicted.

information about the number marking of the pronouns in the source and opens a way to inject the output of the pronoun type classifier into the system.

Our first source-aware language model is an n-gram model trained on an artificial corpus generated from the target lemmas of the parallel training data (Figure 2). Before every REPLACE tag occurring in the data (in the WMT 2016 shared task data, REPLACE items substitute the target pronouns to be predicted), we insert the source pronoun aligned to the tag (without lowercasing or any other processing). The alignment information attached to the REPLACE tag in the shared task data files is stripped off. In the training data, we instead add the pronoun class to be predicted. The n-gram model used for this component is a 6-gram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998) trained with the KenLM toolkit (Heafield, 2011).

We used the classifier described in Section 2.2 to annotate all instances of "it" in the source-language data which were mapped to a REPLACE item according to the alignment provided. Afterwards, a second new source-aware language model is trained. In this way, instead of the sentence *'It s got these fishing lures on the bottom .'* presented in Figure 2, the system receives the labelled input *'it-anaphoric s got these fishing lures on the bottom .'* All of the data provided for the shared-task was used in training this system.

3.1 Results

The macro-averaged recall (official metric of the WMT 2016 shared task) obtained is 57.03%. This is slightly lower than the score of 59.84% which was obtained by the system without the "it" labels (Table 5). However, some pronouns present better scores using the system with the it-labels than the system without them. Precision, in particular, is higher. This outcome is expected for the pronoun *cela*, which is the French neuter demonstrative pronoun frequently used for event reference. However, there are also gains in precision for *on*, *elles* and *ils* while maintaining recall.

w/o "it" labels Macro R: 59.84%			
Pronoun	Precision	Recall	F1
ce	**89.66**	76.47	82.54
elle	**40.00**	60.87	48.28
elles	27.27	12.00	16.67
il	**63.24**	70.49	66.67
ils	67.82	83.10	74.68
cela	76.47	41.94	54.17
on	36.36	44.44	40.00
OTHER	**88.37**	89.41	88.89
w/ "it" labels Macro R: 57.03%			
Pronoun	Precision	Recall	F1
ce	89.09	72.06	79.67
elle	31.25	43.48	36.36
elles	**30.77**	16.00	21.05
il	54.43	70.49	61.43
ils	**69.41**	83.10	75.64
cela	**86.67**	41.94	56.52
on	**40.00**	44.44	42.11
OTHER	85.71	84.71	85.21

Table 5: Source-aware language model with and without "it" disambiguation labels.

In order to further investigate the impact of the disambiguation of "it" on the prediction task, we isolated the cases where the French pronouns are translations of "it". We relied on the alignment information from the shared-task data to separate the French translations of "it" and "they". Once the "it" gold set was obtained, we computed precision, recall and f-score in the usual manner (Table 6).

This second evaluation shows that the improvements obtained for *cela* and *on* are legitimately due to the "it" disambiguation labels. While other classes do not show the same gain in performance, a manual analysis reveals somewhat fewer incoherence errors. For instance, the system with the *w/o "it" labels*

Pronoun	Precision	Recall	F1
ce	90.74	79.03	84.48
elle	43.75	63.64	51.85
elles	0	0	0
il	64.06	70.69	67.21
cela	76.47	41.94	54.17
on	33.33	75.00	46.15
OTHER	85.71	88.89	87.27

w/ "it" labels

Pronoun	Precision	Recall	F1
ce	90.00	72.58	80.36
elle	33.33	45.45	38.46
elles	0	0	0
il	54.67	70.69	61.65
cela	**86.67**	41.94	**56.52**
on	**37.5**	75.00	50.00
OTHER	83.33	83.33	83.33

Table 6: Source-aware language model with and without "it" disambiguation labels. Evaluation on subset of data.

labels classifies more often *il* as *elle* and not *il* as *on* than that system without the labels.

While our results are modest, they point towards an improvement in the general quality of pronoun translation. However, better results on the task of distinguishing between anaphoric and event reference realisations of "it" are needed. In our opinion, accurate disambiguation of the pronoun "it" has the potential to help NLP applications such as Machine Translation and Coreference Resolution.

References

Shane Bergsma and David Yarowsky. 2011. Nada: A robust system for non-referential pronoun detection. In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lecture Notes in Artificial Intelligence, pages 12–23. Springer, Faro, Portugal.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Computer Science Group, Harvard University, Cambridge (Mass.).

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 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 3191–3198, Reykjavik, Iceland. 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, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Christian Hardmeier, Jörg Tiedemann, Preslav Nakov, Sara Stymne, and Yannick Versely. 2016. DiscoMT 2015 Shared Task on Pronoun Translation. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, UK, July. Association for Computational Linguistics.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation Journal*, 42(1):21–40.

Sharid Loáiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-french verb phrase alignment in europarl for tense translation modeling. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC'14, pages 674–681, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christopher Manning and Dan Klein. 2003. Optimization, MaxEnt models, and conditional estimation without magic. In *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX, pages 95–100, Montreal, Canada. Association for Computational Linguistics.

Costanza Navarretta. 2004. Resolving individual and abstract anaphora in texts and dialogues. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING'04, pages 233–239, Geneva, Switzerland. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.