Exploiting Cross-Lingual Hints to Discover Event Pronouns

Sharid Loáiciga[1]*, Christian Hardmeier[2] & Asad Sayeed[3]
[1] CoLab Potsdam, Department of Linguistics, University of Potsdam
[2] Department of Linguistics and Philology, Uppsala University
[3] CLASP, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
sharid.loaiciga@gmail.com, christian.hardmeier@lingfil.uu.se, asad.sayeed@gu.se

Abstract

Non-nominal co-reference is much less studied than nominal coreference, partly because of the lack of annotated corpora. In this paper, we have explored the possibility to exploit parallel multilingual corpora as a means of cheap supervision for the task of it-disambiguation. We found that only a very specific 'event' reading is discernible using our approach.

Keywords: 'it', reference, Europarl corpus

1. Introduction

Nominal coreference has been studied extensively, but work on the automatic recognition of non-nominal anaphora is scarce, as are annotated data sets. Among the challenges of non-nominal anaphora is the difficulty to characterize the large variance of antecedent types, which often include clauses, sentences, and even paragraphs. Here we focus on the English pronoun *it* and its capacity to function as anaphor for nominal entity and non-nominal event antecedents, and as a pleonastic token. The examples 1 to 3 below illustrate these different readings using English passages from the Europarl corpus and their French parallel translations.

In this paper, we evaluate the potential of multilingual parallel data as a form of supervision for the classification of different readings of the English pronoun *it*. We explore the hypothesis that languages have different strategies and preferences to encode referential relationships, and that these differences surface as systematic patterns in multilingual parallel data. Therefore, the competing readings of the pronoun *it* should present different strategies of translation across languages.

We present a method for creating artificial training data for the classification of three different readings of *it*: entity, event or pleonastic. We found that the 'event' reading can be easily predicted, as the languages studied here have a similar strategy for their translation. Despite this, 'event' uses of the pronoun *it* are not enough to generalize to other types of non-nominal reference. Deictic uses in particular, are expressed very differently and are therefore difficult to normalize.

1. ENTITY READING

   Madam President, I have been deluged with messages from growers from all over the south-east of England who regard this proposal as near catastrophic. **It** will result, they tell me, in smaller crops and in higher prices.

   Madame la Présidente, j'ai été assailli de messages de cultivateurs en provenance de tout le sud-est de l'Angleterre, qui considèrent cette proposition comme une quasi-catastrophe. ***Elle*** *entraînera, me disent-ils, une baisse de les rendements agricoles et une augmentation des prix.*

2. EVENT READING

   The European Parliament has always taken a vigorous stance against racism and ethnic intolerance. I appeal to you, as Members of this House, to do **it** once again and support our written declaration condemning Turkish racism against Bulgarians.

   Le Parlement européen a toujours pris des positions véhémentes contre le racisme et l'intolérance ethnique. Je fais appel à vous, en tant que membres de cette Assemblée, pour que vous ***le*** *fassiez à nouveau, et que vous souteniez notre déclaration écrite condamnant le racisme turc à l'égard des Bulgares.*

3. PLEONASTIC READING

   Since the beginning of October 2008 I have been trying to get speaking time in the one -minute contributions and I am pleased that I have finally succeeded. **It** is interesting that Mr Rogalski has been allowed to speak three times in the meantime.

   Depuis le début d'octobre 2008, j'ai essayé d'obtenir un temps de parole dans le cadre des interventions d'une minute et je suis heureux d'avoir finalement réussi. ***Il*** *est intéressant que M. Rogalski ait été autorisé à prendre la parole trois fois dans l'intervalle.*

2. Related Work

Reference to non-nominal antecedents has largely been a niche area in NLP research, which is extensively surveyed in a recent article by Kolhatkar et al. (2018). The most extensive annotation efforts in the field of coreference resolution have focused on nominal coreference. OntoNotes (Pradhan et al., 2013), the largest and most frequently used corpus for training coreference resolution systems, for instance, only includes verbs if "they can be co-referenced with an existing noun phrase" according to its guidelines. Corpora with a richer annotation of event pronouns exist, but are much smaller. The most important resource is the ARRAU corpus (Poesio et al., 2018), whose size

*Work completed while the first author was affiliated at the Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

amounts to about 20% of version 5 of OntoNotes. ParCorFull (Lapshinova-Koltunski et al., 2018) also contains annotations of event pronouns.

The scarcity of manually annotated resources has led to the use of artificial training data for the resolution of non-nominal anaphora. Kolhatkar et al. (2013) study the resolution of anaphoric shell nouns such as 'this issue' or 'this fact' by exploiting cataphoric instances such as 'the fact that…'. Marasovic et al. (2017) construct training examples based on specific patterns of verbs governing embedded sentences. As far as we know, the use of multilingual data for automatic data creation is novel in our work.

Before the breakthrough of neural end-to-end systems in coreference resolution (Lee et al., 2017), coreference resolvers needed to do explicit mention classification in order to exclude non-referential mentions before any resolution was attempted. In this context, the pronoun 'it' has been targeted, as many of its uses are non-referential. Evans (2001) proposes the classification of the pronoun 'it' into seven classes using contextual features. Boyd et al. (2005) report similar results of around 80% accuracy using more complex syntactic patterns. Bergsma and Yarowsky (2011) describe a system for identifying non-referential pronouns using web n-gram features, however without accounting explicitly for event reference.

The many uses of 'it' are also particularly relevant in dialog texts, where event reference is much more common than in news data. In this context, Müller (2007) proposes a disambiguation of 'it' together with the deictic pronouns 'this' and 'that'. Finally, Lee et al. (2016) create a corpus for it-disambiguation in question answering, a domain close to dialog. It is worth noting that current coreference resolution systems are not trained to manage dialog data.

More recently, Loáiciga et al. (2017) proposed a semi-supervised setup based on a combination of syntactic and semantic features used in a two-step classification approach where a maximum entropy classifier is used first and a recurrent recursive network (RNN) after. Yaneva et al. (2018), on the other hand, report on experiments using features from eye gaze that prove to be more effective than any of the other types of features reported in previous works.

3. Method

We worked with the corpus Europarl (Koehn, 2005) v8 as provided in the OPUS collection (Tiedemann, 2012). OPUS includes parsed, sentence-level and word-level alignments files, as well as a toolbox for corpus processing (Aulamo et al., 2020). We used all 15 languages paired with English as the source language. The languages are German, Spanish, Estonian, Finnish, Hungarian, Italian, Latvian, Dutch, Polish, Portuguese, Romanian, Slovak and Slovenian.

The overall method is as follows:

1. Europarl is a parallel corpus of translations between the language pairs, but the amount of data from one language to another varies. Therefore, we began by extracting only the set of common sentences across all languages. This already reduced the data from 2,039,537 segments to 281,346.

2. Next, we relied on the English parsed files to identify all instances of the pronoun *it*.

3. We then used the word-level alignment files to extract the aligned translation in each of the target languages.

   Word alignment is not perfect. One-to-one correspondences are unstable for particles and other small word forms, in particular if they depend on verbs and might be translated by just one verb form, virtually disappearing then from the translation. Pronouns in particular, depending on the language, might not be translated for instance if the language is a pro-drop one, or they might be translated as a full nominal phrase, because the language has a different use of pronouns.

   For improving the quality of the word alignments, we used a window of -3 and +3 tokens before and after the position of the aligned token. This means that if the translated token is not a pronoun (we have POS information from the parsing files), we search for a pronoun translation within the window range.

4. To label the English instances of *it* as 'entity', 'event' or 'pleonastic' we use French as a seed language.

   We consider all instances translated with the neutral demonstrative pronouns *cela*, *ceci* or *ça* as events. In French, these pronoun are typically used to reference proposition or phrases.

   For the entity nominal case, we took the French translations *elle* and *il*.

   Last, for the 'pleonastic' readings, we took all instances of *it* analyzed as expletives in the parsed files. These files have been processed using universal dependencies v2.0 (UDPipe parser, models from 2017-08-01), which includes the dedicated dependency relation `expl` (Bouma et al., 2018).

   From 69,126 *it* pronouns, we labeled 22,615 instances, corresponding to approximately 30% (Table 1).

English	**French**	**Class**	**Instances**
it	*elle/il*	entity	11,483
it	*cela/ça/ceci*	event	910
it	–	pleonastic	10,222

Table 1: Summary of the translation assumptions and the total examples annotated automatically.

5. The translations from the other 14 languages that are not French are used as features for a classification task (Section 4.). We present an example in Figure 1, where each line represents a feature vector.

A manual analysis of a sample of 600 instances confirms that a big drawback of the method is the large number of examples for which a label cannot be determined, as shown in the column 'Unknown' in Table 2 (these examples are not

Features													
DE	ES	ET	FI	HU	IT	LV	NL	PL	PT	RO	SK	SL	SV
empty	*idea*	*seda*	*empty*	*képeznie*	*essenza*	*es*	*dit*	*dodać*	*adaug*	*že*	*empty*	*empty*	*detta*
du	*usted*	*sa*	*empty*	*te*	*l'*	*empty*	*u*	*empty*	*empty*	*eşti*	*ty*	*empty*	*du*
empty	*señor*	*ja*	*empty*	*.*	*-*	*empty*	*ik*	*cohn-bendit*	*cohn-bendit*	*îi*	*a*	*gospod*	*sluta*
empty	*que*	*juhataja*	*siirtämisestä*	*úr*	*presidente*	*empty*	*de*	*!*	*é*	*,*	*je*	*predsednik*	*det*
empty	*es*	*üksluine*	*ne*	*dolog*	*in*	*tas*	*empty*	*co*	*empty*	*ce*	*spôsobom*	*govoriti*	*allt*

Figure 1: Exemplification of the extracted translations of English *it* used as input features features in the classificaton experiments.

counted in our 22,615 labeled examples reported above). As for the examples that are labeled, the main problem is the annotation of pleonastics as nominals. Since we take pleonastic from the parsing annotation, these are therefore undetected expletive constructions by the parser that get labeled as nominals by our assumption French *il* $\rightarrow$ entity. In addition, there is a natural imbalance in the classes, with nominal and pleonastic instances being largely more frequent than events.

Concerning the quality of the annotation, it can be seen in Table 2 that the automatic labeling achieves approximately 30% accuracy overall (133/600) and 70% accuracy if only successfully labeled examples are considered (133/189). A closer inspection of the 'unknown' labels reveals that these are mostly due to many divergent translation from the assumptions we made by using French as the seed language.

	Automatic label $\rightarrow$			
Gold $\downarrow$	Entity	Event	Pleonastic	Unknown
Entity	56	5	0	259
Event	5	6	0	23
Pleonastic	45	1	71	129

Table 2: Manual evaluation of a sample of 600 instances.

4. Classification Experiments

We used the 22,615 generated examples in a classification setting. All the experiments were completed using the implementations of the `scikit-learn` library, including their `train_test_split` function.

In a first experiment, we use the extracted translations with the split in Table 3 to predict one of the three automatically generated labels: 'entity', 'event' or 'pleonastic'. We report results using a Maximum Entropy classifier, although replication experiments using a SVM and a Naive Bayes classifier yielded very similar results.

Train	**Test**	**Total**
15,887	6,728	22,615

Table 3: Data set split for the classification experiments.

Although the results using the automatic labels seem reasonable (Table 4), when applying the same model to predict the manually annotated sample of 600 instances, we see a dramatic decrease in performance, in particular for the 'event' class. As mentioned before, this class has a natural low frequency, which makes it more difficult to predict in itself, with only 6 examples accurately labeled in the manual sample.

Automatically annotated data

MaxEnt	Precision	Recall	Accuracy
it-Entity	0.70	0.75	0.70
it-Event	0.44	0.15	(4,710/6,728)
it-Pleonastic	0.70	0.68	

Manually annotated sample

MaxEnt	Precision	Recall	Accuracy
Entity	0.55	0.84	0.54
Event	0.0	0.0	(318/600)
Pleonastic	0.50	0.22	

Table 4: Classification results using a Maximum Entropy classifier.

To determine whether the imbalance in the data is a factor preventing the model to learn, in a second experiment, we used bootstrap with resampling in order to achieve the same number of examples per class. The data distribution for this experiment is given in Table 5.

Event	**Entity**	**Pleonastic**
11,377	11,377	11,377

Table 5: Equal distribution of the classes for the experiment with oversampling.

In this second scenario, we obtained a comparable performance for the 'entity' and 'pleonastic' classes, and almost perfect scores for the 'event' class (Table 6).

5. Discussion and Conclusion

The experiments presented in the previous section suggest that relying on translations as features for the different readings of 'it' is a good method for the cases of *it* that are captured by the seed language assumptions, as these cases also present a pronoun translation in the other languages. These represent about 30% of the total amount of pronouns 'it'

Oversampling of the event class

MaxEnt	Precision	Recall	Accuracy
Entity	0.73	0.67	0.80
Event	0.92	0.99	(8,277/10,347)
Pleonastic	0.73	0.74	

Table 6: Classification results using bootstrap resampling to achieve an even distribution of the classes.

```
 -- see_et<=0.5
|   |--- é_pt<=0.5
|   |   |--- tas_lv<=0.5
|   |   |   |--- to_pl<=0.5
|   |   |   |   |--- este_ro<=0.5
|   |   |   |   |   |--- ez_hu<=0.5
|   |   |   |   |   |   |--- es_es<=0.5
|   |   |   |   |   |   |   |--- den_sv<=0.5
|   |   |   |   |   |   |   |   |--- je_sk<=0.5
|   |   |   |   |   |   |   |   |   |--- to_sk<=0.5
|   |   |   |   |   |   |   |   |   |   |--se_fi<=0.5
```

Figure 2: Output of a decision tree classifier. The leaves have the form `pronoun_language`.

and unfortunately they do not seem to generalize to the rest of the cases.

Further analysis from the output of a decision tree classifier on the same data partition confirms this finding. As shown in Figure 2, the top leaves in the tree all contain equivalent translations of either ‘it’ or ‘this’, pronouns associated with ‘entity’ and ‘event’ respectively.

Although we originally sought to identify systematic translations patterns indicative of non-nominal uses of ‘it’, through developing this method we found that apart from the pronoun-to-pronoun translation pattern, there is too much variability in the data.

Take for instance the following example:

ENGLISH Madam President , Commissioners , can I say to you that less than a year ago we were debating in this Chamber what we were going to do about global food security , and was there enough food in the world , and we were terribly worried about **it**.

FRENCH *Madame la Présidente , Mesdames et Messieurs les Commissaires , permettez -moi de vous rappeler qu’ il y a moins d’ un an , nous débattions en cette Assemblée de la manière de traiter la sécurité alimentaire mondiale , de la question de savoir si l’ on produisait suffisamment de nourriture à l’ échelle mondiale , et nous étions extrêmement préoccupés par* ***ces questions*** .

In the example the English pronoun ‘it’ refers to all what has previously been mentioned in the long sentence, a typical ‘event’ reading of the pronoun. The French translation, however, prefers a translation with a full lexical noun phrase *ces questions* (these questions) for the same referential relationship. This is a particular case of a shell-noun (Kolhatkar et al., 2013) and we believe that a multilingual study of this phenomenon might be a potential next logical step.

The task could be approached semantically by identifying all abstract nouns referencing actions, nominalizations or eventualities in the text. Or one could decide to focus on particular syntactic configurations as Marasovic et al. (2017).

Non-nominal co-reference is much less studied than nominal coreference, partly because of the lack of annotated corpora. In this paper, we have explored the possibility to exploit parallel multilingual corpora as a means of cheap supervision for the task of it-disambiguation. Since pronoun *it* has many potential uses or readings, we took it a as representative of the non-nominal coreference phenomenon, however, we found that only a very specific case is discernible using our approach.

6. Acknowledgements

Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930.

7. References

Aulamo, M., Sulubacak, U., Virpioja, S., and Tiedemann, J. (2020). OpusTools and Parallel Corpus Diagnostics. In *Proceedings of the 25th International Conference on Language Resources and Evaluation (LREC’20)*, Marseille, France. European Language Resources Association (ELRA).

Bergsma, S. and Yarowsky, D. (2011). NADA: A robust system for non-referential pronoun detection. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium*, volume 7099 of *Lecture Notes in Computer Science*, pages 12–23, Faro, Portugal, October. Springer.

Bouma, G., Hajic, J., Haug, D., Nivre, J., Solberg, P. E., and Øvrelid, L. (2018). Expletives in universal dependency treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26, Brussels, Belgium. Association for Computational Linguistics.

Boyd, A., Gegg-Harrison, W., and Byron, D. K. (2005). Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, Michigan. Association for Computational Linguistics.

Evans, R. (2001). Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45–57.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, MT Summit X, pages 79–86, Phuket, Thailand.

Kolhatkar, V., Zinsmeister, H., and Hirst, G. (2013). Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 300–310, Seattle, Washington, USA, October. Association for Computational Linguistics.

Kolhatkar, V., Roussel, A., Dipper, S., and Zinsmeister, H. (2018). Survey: Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612, September.

Lapshinova-Koltunski, E., Hardmeier, C., and Krielke, P. (2018). ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of 11th Language Resources and Evaluation Conference*, pages 423–428, Miyazaki, Japan. European Language Resources Association (ELRA). to appear.

Lee, T., Lutz, A., and Choi, J. D. (2016). QA-It: classifying non-referential it for question answer pairs. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 132–137, Berlin, Germany. Association for Computational Linguistics.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Loáiciga, S., Guillou, L., and Hardmeier, C. (2017). What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 1336–1342, Copenhagen, Denmark. Association for Computational Linguistics.

Marasovic, A., Born, L., Opitz, J., and Frank, A. (2017). A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark. Association for Computational Linguistics.

Müller, C. (2007). Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823, Prague, Czech Republic. Association for Computational Linguistics.

Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., and Zinsmeister, H. (2018). Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana, June. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Yaneva, V., Ha, L. A., Evans, R., and Mitkov, R. (2018). Classifying referential and non-referential it using gaze. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4896–4901, Brussels, Belgium. Association for Computational Linguistics.