Discovery of Events in the Europarl Languages

Author1, Author2, Author3
Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com
{author1, author5, author9}@abc.org

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

Depending on the context, the English pronoun *it* can express anaphoric reference with a nominal entity antecedent or with a non-nominal antecedent such as an event. It can also be used pleonastically. The examples 1 to 3 below illustrate these different readings. Nominal coreference has been studied extensively, but work on the automatic recognition of non-nominal anaphora is scarce, as are annotated data sets.

In this paper, we evaluate the potential of multilingual parallel data to create artificial training data for the classification of different readings of 'it', by exploiting the hypothesis that languages have different preferences to encode the competing readings. Multilingual parallel data serves thus as a cheap supervision signal.

While developing the method, 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

   The infectious disease that's killed more humans than any other is **malaria**. **It**'s carried in the bites of infected mosquitos.

   Jene Krankheit, die mehr Leute als jede andere umgebracht hat, ist Malaria gewesen. ***Sie*** *wird über die Stiche von infizierten Moskitos übertragen.*

2. EVENT READING

   But I think **if we lost everyone with Down syndrome**, **it** would be a catastrophic loss.

   Aber, wenn wir alle Menschen mit Down-Syndrom verlören, wäre ***das*** *ein katastrophaler Verlust.*

3. PLEONASTIC READING

   And **it** seemed to me that there were three levels of acceptance that needed to take place.

   Und ***es*** *schien, dass es drei Stufen der Akzeptanz gibt, die alle zum Tragen kommen mussten.*

2. Related Work

The study of reference has mostly focused on nominal expressions and their relationships of coreference. Therefore, the biggest annotation efforts in the field of coreference resolution have also focused on nominal coreference. Ontonotes (Pradhan and Xue, 2009), the largest and most used corpus today, for instance, only includes verbs if "they can be co-referenced with an existing noun phrase" according to its guidelines.

It follows that most anaphora and coreference resolution systems focus on nominal reference. Before current state of art systems which work in a end-to-end fashion, these systems needed to explicitly do 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.

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'. Last, (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) has 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, reports 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 used the corpus Europarl (Koehn, 2005) v8 as found in the OPUS collection (Tiedemann, 2012). OPUS also includes parsed and sentence-level and word-level alignments files for the Europarl corpus. We used all 15 languages paired with English as the source language. The

languages are German, Finnish, Swedish, Italian, Latvian, Dutch, Hungarian, Polish, Slovenian, Portuguese, Slovak, Romanian, Estonian, and Spanish.
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 accross all languages. This already reduced the data from 2,039,537 segments to 286,053.

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 all the 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 was not a pronoun (we have POS information from the parsing files), we would search for a pronoun translation within the window range.

4. We aim at creating English data in which the instances of 'it' are annotated as 'entity', 'event' or 'pleonastic'. While the three readings use the same pronoun in English, we rely on the assumption that they have at least partially different realizations in the other languages.

   For the expletives, 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).

   Taking advantage of the parallel data, we decided to use French as a seed language, and 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*. From 69,431 'it' pronouns, we labeled 22,574 instances, corresponding to approximately 30% (Tables 1 and 2).

5. Last, translations from the other 14 languages than French are used as features for the classification task. Each line in Figure 1 represents a feature vector.

English	French	Class
it	*elle/il*	entity
it	*cela/ça/ceci*	event
it	–	pleonastic

Table 1: Summary of the translation assumptions for labeling the classes.

Label		
Entity	Event	Pleonastic
11,607	877	10,252

Table 2: Resulting distribution after automatic labeling.

A manual analysis of a sample of 600 instances reveals that the main problem seems to be the large number of examples that cannot be labeled (column 'Unknown'). In addition, there is a natural imbalance in the classes (nominal and pleonastic are more common than events in previous work) that seems to be accentuated by the automatic labeling. Concerning the quality of the annotation, it can be seen in Table 3 that the automatic labeling achieves approximately 20% accuracy. We believe that this is mainly due to the combination of two factors: word-alignment issues and many different translations different from the assumptions we made by using French as the seed language.

	Entity	Event	Pleonastic	Unknown
Entity	65	2	0	250
Event	5	3	0	34
Pleonastic	41	1	65	134

Table 3: Manual evaluation of a sample of 600 instances.

4. Classification Experiments

We used 22,554 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 generated data tor 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
15,787	6,767	22,554

Table 4: Data set split for the classification experiments.

Although the results using the automatic labels seem reasonable (Table 4.), when using the same model to predict the manually annotated sample of 600 instances, we see

Features													
DE	FI	SV	IT	LV	NL	HU	PL	SL	PT	SK	RO	ET	ES
dies	*on*	*det*	*e*	*tas*	*het*	*hogy*	*pre*	*je*	empty	*'*	empty	*on*	*desgracia*
das	*puhua*	*det*	*-*	*ir*	*sprake*	*lehetsége*	*sa*	*prav*	*soar*	*'*	*în*	empty	*puede*
es	*ei*	empty	*non*	*nepietiks*	empty	*nem*	empty	*zgolj*	empty	*'*	*nu*	*vaeste*	*no*

Figure 1: Exemplification of the extracted features.

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.

Automatically annotated data

MaxEnt	Precision	Recall	Accuracy
it-Entity	0.69	0.74	0.69
it-Event	0.48	0.14	(4,669/6,767)
it-Pleonastic	0.69	0.67	

Manually annotated sample

MaxEnt	Precision	Recall	Accuracy
Entity	0.53	0.99	0.53
Event	0.0	0.0	(318/600)
Pleonastic	0.50	0.02	

Table 5: Classification results using a Maximum Entropy classifier.

To address the problem of the uneven distribution of the classes, in a second experiment, we used bootstrap with resampling in order to achieve the same number of examples per class.

Event	**Entity**	**Pleonastic**
11,496	11,496	11,496

Table 6: 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.

Oversampling of the event class

MaxEnt	Precision	Recall	Accuracy
Entity	0.73	0.68	0.80
Event	0.91	1.0	(8,277/10,347)
Pleonastic	0.74	0.72	

Table 7: Classification results using bootstrap resampling to achieve an even distribution of the classes.

5. Discussion and Conclusion

The experiments presented in the previous section seem to suggest that relying on translations as features for the different readings of 'it' is a good method but only for a particular type of 'event' that has a natural low frequency. Indeed, our method only produces labels for about 30% of the total amount of pronouns 'it' and within this, 'event' has the lowest absolute frequency.

Further analysis from the output of a decision tree classifier on the same data partition also suggests that this type of events are easily discernible. 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 non-nominal uses of 'it', through developing this method we found that the task is hard because there are many potential cases.

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. The French translation, however, prefers a translation with a full lexical noun phrase *ces questions* (these questions) for the same referential relationship.

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 rep-

```
 -- 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`.

resentative of the non-nominal coreference phenomenon, however, we found that only a very specific 'event' reading is discernible using our approach.

6. References

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.

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.

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.

Pradhan, S. S. and Xue, N. (2009). OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado, May. 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.