What is it? Disambiguating the different readings of the pronoun 'it'

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract

In this paper, we address the problem of predicting one of three functions for the English pronoun ‘it’: anaphoric, event reference or pleonastic. This disambiguation is valuable in the context of machine translation and coreference resolution. We present experiments using a MAXENT classifier trained on gold-standard data and self-training experiments of an RNN trained on silver-standard data, annotated using the MAXENT classifier. Lastly, we report on an analysis of the strengths of these two models.

1 Introduction

We address the problem of disambiguating the English pronoun ‘it’, which may function as a pleonastic, anaphoric, or event reference pronoun. As an anaphoric pronoun, ‘it’ corefers with a noun phrase (called the antecedent), as in example (1):

(1) I have a bicycle. It is red.

Pleonastic pronouns, in contrast, do not refer to anything but are required to fill the subject position in many languages, including English, French and German:

(2) It is raining.

Event reference pronouns are anaphoric, but instead of referring to a noun phrase, they refer to a verb, verb phrase, clause or even an entire sentence, as in example (3):

(3) He lost his job. It came as a total surprise.

We propose the identification of the three usage types of it, namely anaphoric, event reference, and pleonastic, with a single system. We present several classification experiments which rely on information from the current and previous sentences, as well as on the output of external tools.

2 Related Work

Due to its difficulty, proposals for the identification and the subsequent resolution of abstract anaphora (i.e., event reference) are scarce (Eckert and Strube, 2000; Byron, 2002; Navarretta, 2004; Müller, 2007). The automatic detection of instances of pleonastic ‘it’, on the other hand, has been addressed by the non-referential ‘it’ detector NADA (Bergsma and Yarowsky, 2011), and also in the context of several coreference resolution systems, including the Stanford sieve-based coreference resolution system (Lee et al., 2011).

The coreference resolution task focuses on the resolution of nominal anaphoric pronouns, de facto grouping our event and pleonastic categories together and discarding both of them. The coreference resolution task can be seen as a two-step problem: mention identification followed by antecedent identification. Identifying instances of pleonastic ‘it’ typically takes place in the mention identification step. The recognition of event reference ‘it’ is, however, to our knowledge not currently included in any such systems, although from a linguistic point of view, event instances are also referential (Boyd et al., 2005). As suggested by Lee et al., (2016), it would be advantageous to incorporate event reference resolution in the second step.

In the context of machine translation, work by Le Nagard and Koehn (2010); Novák et al. (2013); Guillou (2015) and Loúiciga et al. (2016) have also considered disambiguating the function of the pronoun ‘it’ in the interest of improving pronoun translation into different languages.
3 Disambiguating ‘it’

3.1 Labeled Data
The ParCor corpus (Guillou et al., 2014) and DiscoMT2015.test dataset (Hardmeier et al., 2016) were used as gold-standard data. Under the ParCor annotation scheme, which was used to annotate both corpora, pronouns are manually labeled according to their function: anaphoric, event reference, pleonastic, etc. For all instances of ‘it’ in the corpora, we extracted the sentence-internal position of the pronoun, the sentence itself, and the two previous sentences. All examples were shuffled before the corpus was divided, ensuring a balanced distribution of the classes (Table 1).

The pronouns ‘this’ and ‘that’, when used as event reference pronouns, may often be used interchangeably with the pronoun ‘it’ (Guillou, 2016). We therefore automatically substituted all instances of event reference ‘this’ and ‘that’ with ‘it’ to increase the number of training examples.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Event</th>
<th>Anaphoric</th>
<th>Pleonastic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>504</td>
<td>779</td>
<td>221</td>
<td>1,504</td>
</tr>
<tr>
<td>Dev</td>
<td>157</td>
<td>252</td>
<td>92</td>
<td>501</td>
</tr>
<tr>
<td>Test</td>
<td>169</td>
<td>270</td>
<td>62</td>
<td>501</td>
</tr>
<tr>
<td>Total</td>
<td>830</td>
<td>1,301</td>
<td>375</td>
<td>2,506</td>
</tr>
</tbody>
</table>

Table 1: Distribution of classes in the data.

3.2 Baselines
We provide two different baselines (MC and LM BASELINE in Table 2). The first is a setting in which all instances are assigned to the majority class it-anaphoric. The second baseline system is a 3-gram language model built using KenLM (Heafield, 2011) and trained on a modified version of the annotated corpus in which every instance of ‘it’ is concatenated with its function (e.g. ‘it-event’). At test time, the ‘it’ position is filled with each of the three it-function labels in turn, the language model is queried, and the highest scoring option is chosen.

3.3 Features
We designed features to capture not only the token context, but also the syntactic and semantic context preceding the pronouns and, where appropriate, their antecedents/referents, as well as the pronoun head. We used the output of the POS tagger and dependency parser of Bohnet et al. (2013)\(^1\), and of the TreeTagger lemmatizer (Schmid, 1994) to extract the following information for each training example:

**Token context (tok)** 1. Previous three tokens and next two tokens. This includes words, punctuation and the tokens in the previous sentence when the ‘it’ occupies the first position of the current sentence. 2. Lemmas of the next two tokens.

**Pronoun head (hea)** 3. Head word and its lemma. Most of the time the head word is a verb. 4. If the head verb is copular, we include its complement head and not the verb itself (for the verbs be, appear, seem, look, sound, smell, taste, feel, become and get). 5. Whether the head word takes a ‘that’ complement (verbs only). 6. Tense of head word (verbs only), computed as described by Loaiciga et al. (2014).

**Syntactic context (syn)** 7. Whether a ‘that’ complement appears in the previous sentence. 8. Closest NP head to the left and to the right. 9. Presence or absence of extraposed sentential subjects as in ‘So it’s difficult to attack malaria from inside malarious societies, [...]’. 10. Closest adjective to the right.

**Semantic context (sem)** 11. 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. 12. Likelihood of head word taking an event subject (verbs only). An estimate of the likelihood of a verb taking an event subject was computed over the Annotated English Gigaword v.5 corpus (Napoles et al., 2012). We considered two cases favouring event subjects that 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. 13. Non-referential probability assigned to the instance of ‘it’ by NADA (Bergsma and Yarowsky, 2011).

3.4 MaxEnt
The MaxEnt classifier is trained using the Stanford Maximum Entropy package (Manning and Klein, 2003) with all of the features described above. We also experimented with other features and options. For features 1 and 2, a window

\[^1\]We used the pre-trained models for English that are available online https://code.google.com/p/
of three tokens showed a degradation in performance. For feature 8, adding one of the 26 WordNet (Princeton University, 2010) types of nouns had no effect. The feature combination of noun and adjectives to the left or right also had no effect. Feature ablation tests revealed that while combining all features is beneficial for the prediction of the anaphoric and pleonastic classes, the same is not true for the event class. In particular, the inclusion of semantic features, which we designed as indicators of eventness, appears to be harmful (Figure 1).

### 3.5 Unlabeled Data

Given the small size of the gold-standard data, and with the aim of gaining insight from unstructured and unseen data, we used the MAXENT classifier to label additional data from the pronoun prediction shared task at WMT16 (Guillou et al., 2016). This new silver-standard training corpus comprises 1,101,922 sentences taken from the Europarl (3,752,440 sentences), News (344,805 sentences) and TED talks (380,072 sentences) sections of the shared task training data.

### 3.6 RNN

Our second system is a bidirectional recurrent neural network (RNN) which reads the context words and then makes a decision based on the representations that it builds. Concretely, it consists of word-level embeddings of size 90, two layers of Gated
Recurrent Units (GRUs) of size 90 as well, and a final softmax layer to make the predictions. The network uses a context window of 50 tokens both to the left and right of the ‘it’ to be predicted. The features described above are also fed to the network in the form of one-hot vectors. The system uses the `adam` optimizer and the categorical cross-entropy loss function. We chose this architecture following the example of Luotolahti et al. (2016), who built a system for the related task of cross-lingual pronoun prediction.

4 Discussion

We report all of the results in Table 2. **MaxEnt** and RNN-Gold are trained on the gold-standard data only. RNN-Silver is trained on the silver-standard data (annotated using the MaxEnt classifier). RNN-Combined is trained on both the silver-standard and gold-standard data.

The MaxEnt and RNN models show improvements, albeit small for the it-event class, over the baseline systems. Since they are trained on the same gold-standard data, one would expect RNN-Gold to perform similarly to MaxEnt. However, in the case of the RNN-gold, the 50 tokens window may actually not have enough words to be filled with, because the gold-standard data is composed of the sentence with the it-pronoun and the three previous sentences, which in addition tend to be short. For the RNN-Silver system this is not a problem, since the sentences of interest have not been taken out of their original context, fully exploiting the RNN capacity to learn the entirety of the context window they are presented with, even if the data is noisy. As expected, RNN-Combined performs better than RNN-Gold and RNN-Silver. Although it does not perform overwhelmingly better than MaxEnt, there are gains in precision for the it-anaphoric class, and in recall for the it-pleonastic and it-event classes, suggesting that the system benefits from the inclusion of gold-standard data.

### Table 3: Accuracy scores of the systems in different portions of the test-set. For each category, we test whether MaxEnt is better or worse than RNN-Combined. A * indicates significance at $p < 0.001$ using McNemar’s $\chi^2$ test.

<table>
<thead>
<tr>
<th>Reference Relationship</th>
<th>MaxEnt</th>
<th>RNN-Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) NP antecedent in previous 2 sentences</td>
<td><em>(191/248)</em></td>
<td>(136/248)</td>
</tr>
<tr>
<td>e.g. The infectious disease that’s killed more humans than any other is <em>malaria</em>. It’s carried in the bites of infected mosquitoes, and it’s probably our oldest scourge.</td>
<td>0.770</td>
<td>0.548</td>
</tr>
<tr>
<td>(2) VP antecedent in previous 2 sentences</td>
<td>(25/38)</td>
<td><em>(27/38)</em></td>
</tr>
<tr>
<td>e.g. And there’s hope in this next section, of this brain section of somebody else with M.S., because what it illustrates is, amazingly, the brain can repair itself. <em>It</em> just doesn’t do it well enough.</td>
<td>0.658</td>
<td>0.711</td>
</tr>
<tr>
<td>(3) NP or VP antecedent further away in the text (not in snippet)</td>
<td>(28/47)</td>
<td>(28/47)</td>
</tr>
<tr>
<td>e.g. <em>It</em> has spread. <em>It</em> has more ways to evade attack than we know. <em>It’s</em> a shape-shifter, for one thing.</td>
<td>0.596</td>
<td>0.596</td>
</tr>
<tr>
<td>(4) Sentential or clausal antecedent</td>
<td>(52/88)</td>
<td><em>(66/88)</em></td>
</tr>
<tr>
<td>e.g. <em>Pension systems have a hugely important economic and social role and are affected by a great variety of factors</em>. <em>It</em> has been reflected in EU policy on pensions, which has become increasingly comprehensive over the years.</td>
<td>0.591</td>
<td>0.750</td>
</tr>
<tr>
<td>(5) Pleonastic constructions</td>
<td>(43/59)</td>
<td>(42/59)</td>
</tr>
<tr>
<td>e.g. <em>And it seemed to me that there were three levels of acceptance that needed to take place.</em></td>
<td>0.729</td>
<td>0.728</td>
</tr>
<tr>
<td>(6) Ambiguous between event and anaphoric</td>
<td>(3/12)</td>
<td><em>(7/12)</em></td>
</tr>
<tr>
<td>e.g. <em>Today, multimedia is a desktop or living room experience, because the apparatus is so clunky. <em>It</em> will change dramatically with small, bright, thin, high-resolution displays.</em></td>
<td>0.250</td>
<td>0.583</td>
</tr>
<tr>
<td>(7) Ambiguous between event and pleonastic</td>
<td>(2/5)</td>
<td><em>(1/5)</em></td>
</tr>
<tr>
<td>e.g. <em>I did some research on how much it cost, and I just became a bit obsessed with transportation systems. And it began the idea of an automated car.</em></td>
<td>0.400</td>
<td>0.200</td>
</tr>
<tr>
<td>(8) Annotation errors</td>
<td>(0/4)</td>
<td>(0/4)</td>
</tr>
<tr>
<td>e.g. <em>Youth unemployment is particularly worrying in <em>it</em> context, as the lost opportunity for jobless young people to develop professional skills is likely to translate into lower productivity and lower earnings over a longer period of time.</em></td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
With the two-fold goal of gaining a better understanding of the difficulties of the task and strengths of the systems, we re-classified the test set in a stratified manner. We present the systems with seven scenarios reflecting the different types of reference relationships observed in the corpora (Table 3). Our scenarios are exhaustive, thus some only have few examples. The analysis reveals that the MAXENT is a better choice for nominal reference (case (1), mostly it-anaphoric) whereas the RNN-COMBINED system is better at identifying difficult antecedents such as cases (4) and (6). RNN-COMBINED performs slightly better at detecting verbal antecedents, case (2), while both systems perform similarly at learning pleonastic instances (5) or when the antecedent is not in the snippet (3). Finally, we found 4 instances of annotation errors (8). These correspond to some of the automatically substituted cases of ‘this’/‘that’ with ‘it’, for which the ‘this’/‘that’ should not have been marked as a pronoun by the human annotator in the first place. Case (8) is not taken into account in the evaluation.

Taking the complete test set, we found that the MAXENT system performs better than the RNN-COMBINED system in absolute terms ($\chi^2 = 50.8891, p < 0.001$), but this is because case (1) is the most frequent one, which is also the case the MAXENT system is strongest at.

5 Conclusions and Future Work

We have shown that distinguishing between nominal anaphoric and event reference realizations of ‘it’ is a complex task. Our results are promising, but there is room for improvement. The self-training experiment demonstrated the benefit of combining gold-standard and silver-standard data.

We also found that the RNN-COMBINED system is better at handling difficult and ambiguous referring relationships, while the MAXENT performed better for the nominal anaphoric case, when the antecedent is close. Since the two models have different strengths, in future work we plan to enrich the training data with re-training instances from the silver data where the two systems agree, in order to reduce the amount of noise, following the example of Jiang et al. (2016).

Ultimately, we aim towards integrating the it-prediction system within a full machine translation pipeline and a coreference resolution system. In the first case, the different translations of pronoun ‘it’ can be constrained according to their function. In the second case, the performance of a coreference resolution system vs a modified version using the three-way distinction can be measured.

Acknowledgments

This work was supported by the Swedish Research Council under project 2012-916 Discourse-Oriented Statistical Machine Translation. We used computing resources on the Abel cluster, owned by the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR) and on the Taito-CSC cluster provided through the Nordic Language Processing Laboratory (NLPL) (www.nlpl.eu). SL was also supported by the Swiss National Science Foundation under grant no. P1GEP1_161877.

References


Portugal. Association for Computational Linguistics.


