What is it? Disambiguating the different readings of the pronoun ‘it’

Sharid Loáiciga
Uppsala University
Dept. of Linguistics & Philology
Uppsala, Sweden
sharid.loaiciga@lingfil.uu.se

Liane Guillou
University of Edinburgh
School of Informatics
Scotland, United Kingdom
lguillou@inf.ed.ac.uk

Christian Hardmeier
Uppsala University
Dept. of Linguistics & Philology
Uppsala, Sweden
christian.hardmeier@lingfil.uu.se

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 Loáiciga et al. (2016) have also considered disambiguating the function of the pronoun ‘it’ in the interest of improving pronoun translation into different languages.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1325–1331
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

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.

Data set	Event	Anaphoric	Pleonastic	Total
Training	504	779	221	1,504
Dev	157	252	92	501
Test	169	270	62	501
Total	830	1,301	375	2,506

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 Loáiciga 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 a 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/mate-tools/downloads/list`

	Dev-set				Test-set			
MC BASELINE	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.539	1	0.700	(252/501)	0.503	1	0.669	(270/501)
				0.503				0.539
LM BASELINE	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.613	0.290	0.394	(166/501)	0.732	0.263	0.387	(163/501)
it-pleonastic	0.169	0.523	0.255	0.331	0.139	0.694	0.231	0.325
it-event	0.459	0.287	0.353		0.521	0.290	0.373	
MAXENT	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.685	**0.758**	**0.719**	(326/501)	0.716	**0.756**	**0.735**	(344/501)
it-pleonastic	**0.884**	**0.543**	**0.633**	0.651	**0.750**	0.726	**0.738**	0.687
it-event	**0.545**	0.541	**0.543**		**0.564**	0.521	0.542	
RNN-GOLD	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.544	0.560	0.552	(221/501)	0.595	0.659	0.626	(250/501)
it-pleonastic	0.274	0.217	0.242	0.441	0.177	0.177	0.177	0.499
it-event	0.355	0.382	0.368		0.436	0.361	0.394	
RNN-SILVER	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	0.661	0.611	0.635	(286/501)	0.706	0.552	0.620	(286/501)
it-pleonastic	0.725	0.402	0.517	0.571	0.542	0.516	0.529	0.571
it-event	0.438	0.605	0.508		0.455	0.621	0.525	
RNN-COMBINED	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
it-anaphoric	**0.697**	0.492	0.577	(280/501)	**0.794**	0.530	0.636	(315/501)
it-pleonastic	0.633	**0.543**	0.585	0.559	0.582	**0.742**	0.652	0.629
it-event	0.434	**0.675**	0.529		0.520	**0.746**	**0.613**	

Table 2: Comparison of baselines and classification results.

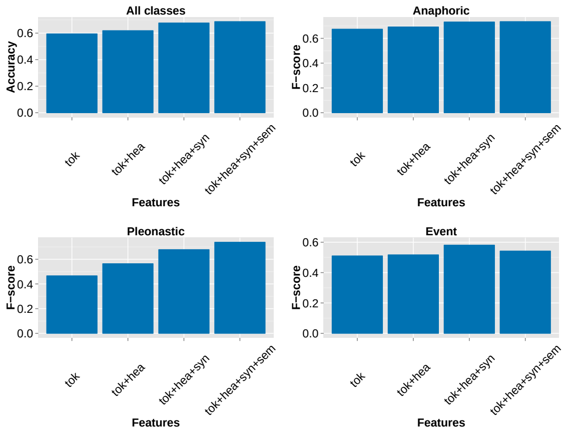

Figure 1: Feature ablation – MAXENT system.

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 event*ness*, 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 on word-level embeddings of size 90, two layers of Gated

Reference relationship	MaxEnt	Rnn-Combined
(1) NP antecedent in previous 2 sentences *e.g. The infectious disease that's killed more humans than any other is **malaria**. It's carried in the bites of infected mosquitos, and **it**'s probably our oldest scourge.*	***(191/248)** **0.770**	(136/248) 0.548
(2) VP antecedent in previous 2 sentences *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**. It just doesn't do **it** well enough.*	(25/38) 0.658	**(27/38)** **0.711**
(3) NP or VP antecedent further away in the text (not in snippet) *e.g. It has spread. It has more ways to evade attack than we know. **It**'s a shape-shifter, for one thing.*	(28/47) 0.596	(28/47) 0.596
(4) Sentential or clausal antecedent e.g. ***Pension systems have a hugely important economic and social role and are affected by a great variety of factors. It*** *has been reflected in EU policy on pensions, which has become increasingly comprehensive over the years.*	(52/88) 0.591	***(66/88)** **0.750**
(5) Pleonastic constructions *e.g. And **it** seemed to me that there were three levels of acceptance that needed to take place.*	(43/59) 0.729	(42/59) 0.728
(6) Ambiguous between event and anaphoric *e.g. Today, multimedia is a desktop or living room experience, because the apparatus is so clunky . **It** will change dramatically with small, bright, thin, high-resolution displays.*	(3/12) 0.250	**(7/12)** **0.583**
(7) Ambiguous between event and pleonastic *e.g. 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.*	**(2/5)** **0.400**	(1/5) 0.200
(8) Annotation errors *e.g. Youth unemployment is particularly worrying in **it** 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.*	(0/4) –	(0/4) –

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 χ^2 test.

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.

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

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.

Adriane Boyd, Whitney Gegg-Harrison, and Donna K. Byron. 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.

Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL 2002, pages 80–87, Philadelphia. Association for Computational Linguistics.

Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Liane Guillou. 2015. Automatic post-editing for the DiscoMT pronoun translation task. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, DiscoMT 2015, pages 65–71, Lisbon, Portugal. Association for Computational Linguistics.

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

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, 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 2014, pages 3191–3198, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT11, pages 187–197, Edinburgh, UK. Association for Computational Linguistics.

Kailang Jiang, Giuseppe Carenini, and Raymond Ng. 2016. Training data enrichment for infrequent discourse relations. In *Proceedings the 26th International Conference on Computational Linguistics: Technical Papers*, COLING 2016, pages 2603–2614, Osaka, Japan. The COLING 2016 Organizing Committee.

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

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation*, WMT10, pages 258–267, Uppsala, Sweden. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL 2011, pages 28–34, Portland, Oregon. Association for Computational Linguistics.

Timothy Lee, Alex Lutz, and Jinho D. Choi. 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.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2016. It-disambiguation and source-aware language models for cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, WMT16, pages 581–588, Berlin, Germany. Association for Computational Linguistics.

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 2014, pages 674–681, Reykjavik, Iceland. European Language Resources Association (ELRA).

Juhani Luotolahti, Jenna Kanerva, and Filip Ginter. 2016. Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, WMT16, pages 596–601, Berlin, Germany. Association for Computational Linguistics.

Christopher Manning and Dan Klein. 2003. MaxEnt models, conditional estimation, and optimization without magic. Tutorial at HLT-NAACL and 41st ACL conferences.

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

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 2004, pages 233–239, Geneva, Switzerland. Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2013. Translation of "It" in a deep syntax framework. In *Proceedings of the Workshop on Discourse in Machine Translation*, DiscoMT 2015, pages 51–59, Sofia, Bulgaria. Association for Computational Linguistics.

Princeton University. 2010. WordNet.

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.