QA-It: Classifying Non-Referential *It* for Question Answer Pairs

Timothy Lee
Math & Computer Science
Emory University
Atlanta, GA 30322, USA
tlee54@emory.edu

Alex Lutz
Math & Computer Science
Emory University
Atlanta, GA 30322, USA
ajlutz@emory.edu

Jinho D. Choi
Math & Computer Science
Emory University
Atlanta, GA 30322, USA
jinho.choi@emory.edu

Abstract

This paper introduces a new corpus, QA-It, for the classification of non-referential *it*. Our dataset is unique in a sense that it is annotated on question answer pairs collected from multiple genres, useful for developing advanced QA systems. Our annotation scheme makes clear distinctions between 4 types of *it*, providing guidelines for many erroneous cases. Several statistical models are built for the classification of *it*, showing encouraging results. To the best of our knowledge, this is the first time that such a corpus is created for question answering.

1 Introduction

One important factor in processing document-level text is to resolve coreference resolution; one of the least developed tasks left in natural language processing. Coreference resolution can be processed in two steps, mention detection and antecedent resolution. For mention detection, the classification of the pronoun *it* as either referential or non-referential is of critical importance because the identification of non-referential instances of *it* is essential to remove from the total list of possible mentions (Branco et al., 2005; Wiseman et al., 2015).

Although previous work has demonstrated a lot of promise for classifying all instances of *it* (Boyd et al., 2005; Müller, 2006; Bergsma et al., 2008; Li et al., 2009), it is still a difficult task, especially when performed on social networks data containing grammatical errors, ambiguity, and colloquial language. In specific, we found that the incorrect classification of non-referential *it* was one of the major reasons for the failure of a question answering system handling social networks data. In this paper, we first introduce our new corpus, QA-It, sampled from the Yahoo! Answers corpus and manually annotated with 4 categories of *it*, referential-nominal, referential-others, non-referential, and errors. We also present statistical models for the classification of these four categories, each showing incremental improvements from one another.

The manual annotation of this corpus is challenging because the rhetoric used in this dataset is often ambiguous; consequently, the automatic classification becomes undoubtedly more challenging. Our best model shows an accuracy of ≈78%, which is lower than some of the results achieved by previous work, but expected because our dataset is much harder to comprehend even for humans, showing an inter-annotation agreement of ≈65%. However, we believe that this corpus provides an initiative to development a better coreference resolution system for the setting of question answering.

2 Related Work

The identification of non-referential *it*, also known as pleonastic *it*, has been studied for many years, starting with Hobbs (1978). Although most of these earlier approaches are not used any more, the rules they discovered have helped for finding useful features for later machine learning approaches. Evans (2001) used 35 features and memory-based learning to classify 7 categories of *it* using data sampled from the SUSANNE and BNC corpora. Boyd et al. (2005) took this approach and added 25 more features to identify 5 categories of *it*.

Müller (2006) classified 6 categories of *it* using spoken dialogues from the ICSI Meeting corpus. Bergsma et al. (2008) used n-gram models to identify *it* as either referential or non-referential. Li et al. (2009) used search queries to help classify 7 categories of *it*. Figure 2 shows how the annotation scheme for non-referential *it* has changed over time. Our approach differs from the recent work because we not only identify instances of *it* as either refer-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 132–137,
Berlin, Germany, August 7-12, 2016.

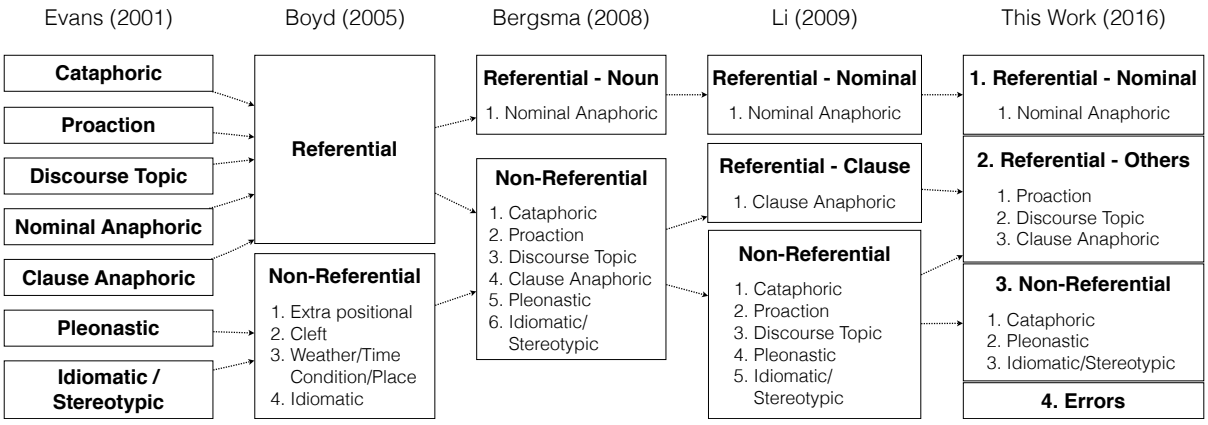

Figure 1: The chronicles of non-referential *it* annotation schemes.

ential or not, but also categorize whether referential *it* refers to a nominal or others, providing coreference resolution systems more valuable information. Furthermore, our corpus includes more colloquial language, which makes it harder to disambiguate different categories of *it*.

3 Data Collection

We inspected several corpora (e.g., Amazon product reviews, Wikipedia, New York Times, Yahoo! Answers), and estimated the maximum likelihood of non-referential *it* in each corpus. After thorough inspection, the Yahoo! Answers and the Amazon product reviews were found to contain the highest numbers of *it*; however, an overwhelming percentage of *it* in the Amazon product reviews was referential. On the other hand, the Yahoo! Answers showed great promise with over 35% instances of non-referential and referential-others *it*. Thus, question answer pairs were uniformly sampled from 9 genres in the Yahoo! Answers corpus:

[1]Computers and Internet, [2]Science and Mathematics, [3]Yahoo! Products, [4]Education and Reference, [5]Business and Finance, [6]Entertainment and Music, [7]Society and Culture, [8]Health, [9]Politics and Government

These genres contained the highest numbers of *it*. Each question answer pair was then ranked by the number of tokens it contained, ranging from 0 to 20, 20 to 40, all the way from 200 to 220, to see the impact of the document size on the classification of *it*. It is worth mentioning that our annotation was done on the document-level whereas annotations from most of the previous work were done on the sentence-level. While training our annotators, we confirmed that the contextual information was vital in classifying different categories of *it*.

4 Annotation Scheme

Instances of *it* are grouped into 4 categories in our annotation scheme; referential-nominal, referential-others, non-referential, and errors (Figure 2). Some of these categories are adapted from Evans (2001) who classified *it* into 7 categories; their categories captured almost every form of *it*, thus linguistically valuable, but a simpler scheme could enhance the annotation quality, potentially leading to more robust coreference resolution.

Boyd et al. (2005) focused on the detection of non-referential *it*, and although their scheme was effective, they did not distinguish referents that were nominals from the others (e.g., proaction, clause, discourse topic), which was not as suited for coreference resolution. Bergsma et al. (2008) attempted to solve this issue by defining that only instances of *it* referent to nominals were referential. Li et al. (2009) further elaborated above rules by adding referential-clause; their annotation scheme is similar to ours such that we both make the distinction between whether *it* refers to a nominal or a clause; however, we include proaction and discourse topic to referential-others as well as cataphoric instances to non-referential.

Our aim is to generate a dataset that is useful for a coreference system to handle both nominal and non-nominal referents. With our proposed scheme, it is up to a coreference resolution system whether or not to handle the referential-others category, including clause, proaction, and discourse topic, during the process of mention detection. Furthermore, the errors category is added to handle non-pronoun cases of *it*. Note that we only consider referential as those that do have antecedents. If the pronoun is cataphoric, it is categorized as non-referential.

Genre	Doc	Sen	Tok	C_1	C_2	C_3	C_4	C_*
1. Computers and Internet	100	918	11,586	222	31	24	3	280
2. Science and Mathematics	100	801	11,589	164	35	18	3	220
3. Yahoo! Products	100	1,027	11,803	176	36	25	3	240
4. Education and Reference	100	831	11,520	148	55	36	2	241
5. Business and Finance	100	817	11,267	139	57	37	0	233
6. Entertainment and Music	100	946	11,656	138	68	30	5	241
7. Society and Culture	100	864	11,589	120	57	47	2	226
8. Health	100	906	11,305	142	97	32	0	271
9. Politics and Government	100	876	11,482	99	81	51	0	231
Total	900	7,986	103,797	1,348	517	300	18	2,183

Table 1: Distributions of our corpus. Doc/Sen/Tok: number of documents/sentences/tokens. $C_{1..4}$: number of *it*-instances in categories described in Sections 4.1, 4.2, 4.3, and 4.4.

4.1 Referential - Nominal

This category is for anaphoric instances of *it* that clearly refer to nouns, noun phrases, or gerunds. This is the standard use of *it* that is already being referenced by coreference resolution models today.

4.2 Referential - Others

This category is for any anaphoric instances of *it* that do not refer to nominals. Some anaphora referents could be in the form of proaction, clause anaphoras, or discourse topic (Evans, 2001). Most coreference resolution models do not handle these cases, but as they still have anaphora referents, it would be valuable to indicate such category for the future advance of a coreference resolution system.

4.3 Non-Referential

This category is for any extraposition, clefts, and pronouns that do not have referent. This also includes cataphora (Evans, 2001). Our distinction of non-referential *it* is similar to the one made by Boyd et al. (2005), except that we do not include weather, condition, time, or place in this category because it would often be helpful to have those instances of *it* be referential:

> What time is it now in Delaware US?
> *It* would be approximately 9:00 am.

Many could argue that the second instance of *it* is non-referential for the above example. But when context is provided, it would be more informative to have *it* refer to "the time now in Delaware US" for coreference resolution. If *it* is simply marked as non-referential, we would essentially be losing the context that the time in Delaware is 9:00 am. Although this does not appear many times in our corpus, it is important to make this distinction based on the context because without the context, this instance of *it* would be simply marked as non-referential.

4.4 Errors

This category includes any uses of a non-pronoun form of *it* including IT (Information Technology), disfluencies, and ambiguous *it* in book/song titles.

> When you leave a glass of water sitting around for a couple hours or so , do bubbles form *it it*

In the example above, the two instances of *it* serves no purpose and cannot be identified as a potential misspelling of another word. This category is not present in any of the previous work, but due to the nature of our corpus as mentioned in difficulties, it is included in our annotation scheme.

5 Corpus Analytics

5.1 Annotation Difficulties

The Yahoo! Answers contains numerous grammatical errors, ambiguous references, disfluency, fragments, and unintelligible question and answer pairs, all of which contributes to difficulties in annotation. Ambiguous referencing had been problematic throughout the annotation and sometimes an agreement was hard to reach between annotators:

> After selling mobile phones, I got post dated cheques ($170,000). But he closed office and bank account. help me?... That's a lot of money to just let go. If it were $1,700.00 then I might just whoop his a** and let ***it*** go but for $170,000... are you kidding?...

Here, *it* can be either idiomatic, or refer to the "post dated cheque" or the "process of receiving the post dated cheque" such that disambiguating its category is difficult even with the context. There were more of such cases where we were not certain if the referent was referential-nominal, referential-others, or idiomatic; in which case, the annotators were instructed to use their best intuition to categorize.

5.2 Inter-Annotation Agreement

All instances of *it* were double annotated by students trained in both linguistics and computer science. Adjudication was performed by the authors of this paper. For the inter-annotator agreement, our annotation gave the Cohans Kappa score of 65.25% and the observed proportionate agreement score of 81.81%.

5.3 Analysis By Genre

The genre has a noticeable influence on the relative number of either referential or non-referential instances of *it*. The genres with the lowest percentage of referential-nominal are "Society and Culture" and "Politics and Government". These genres also contain the most abstract ideas and thoughts within the question and answer pairs. The genres which contain the most number of referential-nominal are "Computers and Internet", "Science and Mathematics", and "Yahoo! Products". This makes sense because in each of these categories, the questions and answers deal with specific, tangible objects such as "pressing a button on the computer to uninstall software". Overall, the more abstract the questions and answers get, the more likely it is to use non-referential *it* or referential-others.

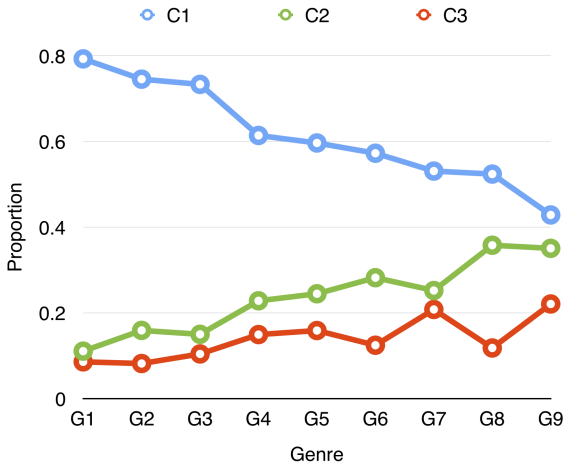

Figure 2: The proportion of referential-nominal for each genre. C1..3: the first 3 categories in Section 4, G1..9: the 9 genres in Table 1.

5.4 Analysis By Document Size

The document size shows a small influence on the categorization of *it*. The document group with the most instances of non-referential *it* is the smallest in size with a total number of tokens between 0 and 20. The rest of the document groups contains fewer instances of non-referential *it* although the differences are not as large as expected.

Document Size	C_1	C_2	C_3	C_4	C_*
0–20	21	60	20	0	101
20–40	14	84	33	0	131
40–60	27	100	33	1	161
60–80	24	129	42	2	197
100–120	29	132	56	2	219
120–140	28	148	53	3	232
140–160	32	163	68	2	265
160–180	28	158	74	6	266
180–200	43	190	70	0	303
200–220	54	184	68	2	308

Table 2: Distributions of our data for each document size.

5.5 Importance of Contextual Information

In certain cases, context is mandatory in determining the category of *it*:

> Q: Regarding ***IT***, what are the fastest ways of getting superich?
> A: Find something everyone will need and then patent it. It could be anything that would do with or about computers. Look at RIM and the struggle it is now facing. With good maketing ANY enhancement or a new design could be worth millions. However, the biggest path to being rich is with maintenece or service of systems or with old programming languages.

For the first instance of *it*, if the annotators are only given the question, they possibly categorize it as referential-nominal or referential-others. However, we can confirm from further reading the context that *it* refers to the IT, "Information Technology".

6 Experiments

6.1 Corpus

Table 4 shows the distributions of our corpus, split into training (70%), development (10%), and evaluation (20%) sets. A total of 1,500, 209, and 474 instances of *it* is found in each set, respectively.

Model	Development Set					Evaluation Set				
	ACC	C_1	C_2	C_3	C_4	ACC	C_1	C_2	C_3	C_4
M_0	72.73	82.43	35.48	57.14	0.00	74.05	82.65	49.20	71.07	0.00
M_1	**73.21**	82.56	50.00	62.50	0.00	74.68	82.93	53.14	73.33	0.00
M_2	73.08	82.56	49.41	60.00	-	**75.21**	83.39	51.23	73.95	-
M_3	76.44	82.31	64.75		-	77.14	82.26	67.87		-
M_4	**76.92**	83.45	61.90		-	**78.21**	83.39	68.32		-

Table 3: Accuracies achieved by each model (in %). ACC: overall accuracy, $C_{1..4}$: F1 scores for 4 categories in Section 4. The highest accuracies are highlighted in bold.

All data are tokenized, sentence segmented, part-of-speech tagged, lemmatized, and dependency parsed by the open-source NLP toolkit, NLP4J (Choi and Palmer, 2012; Choi and McCallum, 2013).[1]

Set	Doc	Sen	Tok	C_1	C_2	C_3	C_4
TRN	630	5,650	72,824	927	353	209	11
DEV	90	787	10,348	139	42	27	1
TST	180	1,549	20,625	282	122	64	6

Table 4: Distributions of our data splits.

6.2 Feature Template

For each token w_i whose lemma is either *it* or *its*, features are extracted from the template in Table 5. w_{i-k} and w_{i+k} are the k'th preceding and succeeding tokens of w_i, respectively. $h(w_i)$ is the dependency head of w_i. The joint features in line 2 are motivated by the rules in Boyd et al. (2005). For instance, with a sufficient amount of training data, features extracted from [w_{i+1}.p + w_{i+2}.m] should cover all rules such as [*it* + verb + *to/that/what/etc*]. Three additional features are used, the relative position of w_i within the sentence S_k (rpw; $w_i \in S_k$), the relative distance of w_i from the nearest preceding noun w_j (rdw; $w_j \in S_k$), and the relative position of S_k within the document D (rps; $S_k \in D$):

$$rpw = i/t \quad , t = \text{\# of tokens in } S_k.$$
$$rdw = |i-j|/t \quad , t = \text{\# of tokens in } S_k.$$
$$rps = k/d \quad , d = \text{\# of sentences in } D.$$

w_i.p, $w_{i\pm1}$.p, $w_{i\pm2}$.p, $h(w_i)$.p, $w_{i\pm1}$.m, $h(w_i)$.m
w_{i+1}.p + w_{i+2}.m, w_{i+1}.p + w_{i+2}.p + w_{i+3}.m
w_i.d , $h(w_i)$.dm

Table 5: Feature template used for our experiments. p: part-of-speech tag, m: lemma, d: dependency label, dm: set of dependents' lemmas.

[1] https://github.com/emorynlp/nlp4j

It is worth mentioning that we experimented with features extracted from brown clusters (Brown et al., 1992) and word embeddings (Mikolov et al., 2013) trained on the Wikipedia articles, which did not lead to a more accurate result. It may be due to the different nature of our source data, Yahoo! Answers. We will explore the possibility of improving our model by facilitating distributional semantics trained on the social networks data.

6.3 Machine Learning

A stochastic adaptive gradient algorithm is used for statistical learning, which adapts per-coordinate learning rates to exploit rarely seen features while remaining scalable (Duchi et al., 2011). Regularized dual averaging is applied for ℓ_1 regularization, shown to work well with ADAGRAD (Xiao, 2010). In addition, mini-batch is applied, where each batch consists of instances from k-number of documents. The following hyperparameters are found during the development and used for all our experiments: the learning rate $\eta = 0.1$, the mini-batch boundary $k = 5$, the regularization parameter $\lambda = 0.001$.

6.4 Evaluation

Table 3 shows the accuracies achieved by our models. M_0 is the baseline model using only the features extracted from Table. M_1 uses the additional features of rpw, rdw, and rps in Section 6.2. The additional features show robust improvements on both the development and the evaluation sets. Notice that the F1 score for C_4 (errors) is consistently 0; this is not surprising given the tiny amount of training instances C_4 has. M_2 is experimented on datasets where annotations for C_4 are discarded. A small improvement is shown for M_2 on the evaluation set but not on the development set, where only 1 instance of C_4 is found.

M_3 and M_4 aim to classify instances of *it* into 2 classes by merging C_2 and C_3 during either train-

ing (M_3) or evaluation (M_4). Training with 3 categories and merging the predicted output into 2 categories during evaluation (M_4) gives higher accuracies than merging the gold labels and training with 2 categories (M_3) in our experiments.

7 Conclusion

This paper introduces a new corpus called, QA-It, sampled from nine different genres in the Yahoo! Answers corpus and manually annotated with four categories of *it*.[2] Unlike many previous work, our annotation is done on the document-level, which is useful for both human annotators and machine learning algorithms to disambiguate different types of *it*. Our dataset is challenging because it includes many grammatical errors, ambiguous references, disfluency, and fragments. Thorough corpus analysts are provided for a better understanding of our corpus. Our corpus is experimented with several statistical models. Our best model shows an accuracy of 78%; considering the challenging nature of our corpus, this is quite encouraging. Our work can be useful for those who need to perform coreference resolution for question answering systems.

In the future, we will double the size of our annotation so we can train a better model and have a more meaningful evaluation. We are also planning on developing a recurrent neural network model for the classification of *it*.

References

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Distributional Identification of Non-Referential Pronouns. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*, ACL'08, pages 10–18.

Adriane Boyd, Whitney Gegg-Harrison, and Donna 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.

António Branco, Tony McEnery, and Ruslan Mitkov, editors. 2005. *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*. John Benjamins Publishing Company.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–480.

[2] https://github.com/emorynlp/qa-it

Jinho D. Choi and Andrew McCallum. 2013. Transition-based Dependency Parsing with Selectional Branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 1052–1062.

Jinho D. Choi and Martha Palmer. 2012. Fast and Robust Part-of-Speech Tagging Using Dynamic Model Selection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL'12, pages 363–367.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *The Journal of Machine Learning Research*, 12(39):2121–2159.

Richard Evans. 2001. Applying Machine Learning Toward an Automatic Classification of *It*. *Literary and Linguistic Computing*, 16(1):45–57.

Jerry R. Hobbs. 1978. Resolving Pronoun References. *Lingua*, 44:331–338.

Yifan Li, Petr Musílek, Marek Reformat, and Loren Wyard-Scott. 2009. Identification of Pleonastic It Using the Web. *Journal Of Artificial Intelligence Research*, 34:339–389.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26*, NIPS'13, pages 3111–3119.

Christoph Müller. 2006. Automatic detection of nonreferential it in spoken multi-party dialog. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'06, pages 49–56.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL'15, pages 1416–1426.

Lin Xiao. 2010. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *Journal of Machine Learning Research*, 11:2543–2596.