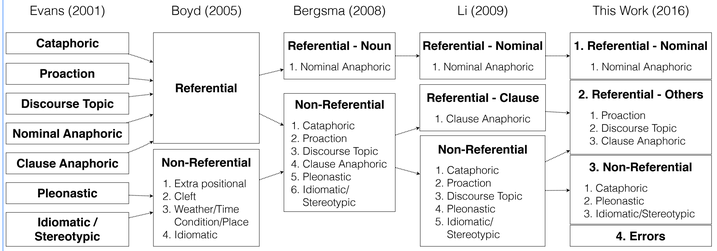
Classifying Non-Referential It for Question Answer Pairs

Timothy Lee, Alex Lutz, and Jinho D. Choi
Department of Mathematics and Computer Science, Emory University

Introduction

- Classifying non-referential it is an important task in coreference resolution
- But no such attempts in this classification has been done for question answer pairs
- The style of English used in question answer documents differs greatly from standard documents such as the Wall Street Journal
- Our task is to classify non-referential it for question answer pairs and introduce the dataset



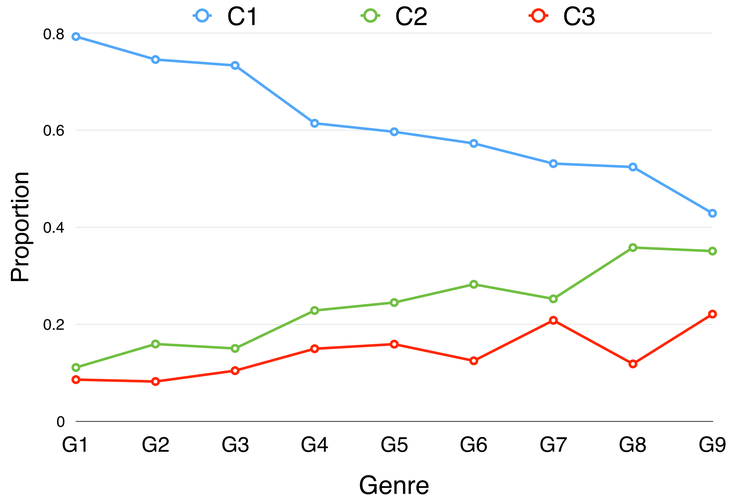
- Initially Evans classified it into seven categories which effectively captures all types of it
- We propose a more efficient set of rules to classify it for coreference specifically for question answer pairs

Type	Description
C1	Anaphoric instances of it which refer to nouns, noun phrases, or gerunds e.g. John bought a book. It was about space.
C2	Anaphoric instances of it which do not refer to nominals such as proaction, clause anaphoras, and discourse topic. e.g. Always use tools correctly. If it feels very awkward, stop.
C3	Contains the most common instances of pleonastic it including extraposition, cleft, atmospheric, and idiomatic e.g. It is sunny outside.
C4	Non-pronoun forms of it including disfluencies, abbreviations, and misspellings e.g. Why did my email move it self?

Corpus Analytics

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

- "Politics and Government" and "Society and Culture" had the highest proportion of non-referential instances due to their abstract ideas
- "Computers and Internet" and "Science and Mathematics" had the most referential-nominal cases because these dealt with tangible objects



- One large problem while annotating was resolving ambiguous references of *it*
- If it were $1,700.00 … and let *it* go but for $170,000...
- Here, it can be either idiomatic, or refer to the "post dated cheque" or the "process of receiving the post dated cheque"
- Contextual information proved vital to solve this ambiguity
- Q: Regarding *IT*, what are the fastest ways of getting superich? A: …. with maintenece or service of systems or with old programming languages.
- Since Yahoo! Answers isn't standard english, *IT* could have been capitalized for emphasis or to mean Information Technology. With the help of contextual information, such ambiguity could be avoided

Results

Experimental Setup

- Used stochastic adaptive gradient descent with mini-batch and L1 regularization
- Tested new sets of features including brown clusters, word embeddings, and dependency derived relationships

The following features were used to classify instances of it:

- POS and dependency of current word
- POS and lemma for dependency head of current word
- POS and lemma for succeeding token and POS of 2nd succeeding token
- POS of succeeding token with lemma of 2nd succeeding token
- POS of 1st and 2nd succeeding token with lemma of 3rd succeeding token

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		-

- m0 only used the baseline features
- m1 uses additional features based on the relative position of it, the relative distance from preceding noun, and relative position of sentence within document
- m2 discard the annotations for errors
- m3 merges referential-nominal and referential-other during the training set
- m4 merges referential-nominal and referential-other during the evaluation set

Conclusion

- We introduced a new corpus from Yahoo! Answers which classified instances of it into four categories
- Using a mixture of old and new features, we were able to achieve promising results despite a challenging dataset
- In the future, we plan to increase the size of our dataset by adding more genres from Yahoo! Answers
- We plan to use a recurrent neural network with our dataset in the future to see if that will yield better results