An Error Analysis of Relation Extraction in Social Media Documents

Gregory Ichneumon Brown
University of Colorado at Boulder
Boulder, Colorado
browngp@colorado.edu

Abstract

Relation extraction in documents allows the detection of how entities being discussed in a document are related to one another (e.g. part-of). This paper presents an analysis of a relation extraction system based on prior work but applied to the J.D. Power and Associates Sentiment Corpus to examine how the system works on documents from a range of social media. The results are examined on three different subsets of the JDPA Corpus, showing that the system performs much worse on documents from certain sources. The proposed explanation is that the features used are more appropriate to text with strong editorial standards than the informal writing style of blogs.

1 Introduction

To summarize accurately, determine the sentiment, or answer questions about a document it is often necessary to be able to determine the relationships between entities being discussed in the document (such as part-of or member-of). In the simple sentiment example

Example 1.1: I bought a new car yesterday. I love the powerful engine.

determining the sentiment the author is expressing about the car requires knowing that the engine is a part of the car so that the positive sentiment being expressed about the engine can also be attributed to the car.

In this paper we examine our preliminary results from applying a relation extraction system to the J.D. Power and Associates (JDPA) Sentiment Corpus (Kessler et al., 2010). Our system uses lexical features from prior work to classify relations, and we examine how the system works on different subsets from the JDPA Sentiment Corpus, breaking the source documents down into professionally written reviews, blog reviews, and social networking reviews. These three document types represent quite different writing styles, and we see significant difference in how the relation extraction system performs on the documents from different sources.

2 Relation Corpora

2.1 ACE-2004 Corpus

The Automatic Content Extraction (ACE) Corpus (Mitchell, et al., 2005) is one of the most common corpora for performing relation extraction. In addition to the co-reference annotations, the Corpus is annotated to indicate 23 different relations between real-world entities that are mentioned in the same sentence. The documents consist of broadcast news transcripts and newswire articles from a variety of news organizations.

2.2 JDPA Sentiment Corpus

The JDPA Corpus consists of 457 documents containing discussions about cars, and 180 documents discussing cameras (Kessler et al., 2010). In this work we only use the automotive documents. The documents are drawn from a variety of sources, and we particularly focus on the 24% of the documents from the JDPA Power Steering blog, 18% from Blogspot, and 18% from LiveJournal.

The annotated mentions in the Corpus are single or multi-word expressions which refer to a particular real world or abstract entity. The mentions are annotated to indicate sets of mentions which constitute co-reference groups referring to the same entity. Five relationships are annotated between these entities: PartOf, FeatureOf, Produces, InstanceOf, and MemberOf. One significant difference between these relation annotations and those in the ACE Corpus is that the former are relations between sets of mentions (the co-reference groups) rather than between individual mentions. This means that these relations are not limited to being between mentions in the same sentence. So in Example 1.1, "engine" would be marked as a part of "car" in the JDPA Corpus annotations, but there would be no relation annotated in the ACE Corpus. For a more direct comparison to the ACE Corpus results, we restrict ourselves only to mentions within the same sentence (we discuss this decision further in section 5.4).

3 Relation Extraction System

3.1 Overview

The system extracts all pairs of mentions in a sentence, and then classifies each pair of mentions as either having a relationship, having an inverse relationship, or having no relationship. So for the PartOf relation in the JDPA Sentiment Corpus we consider both the relation "X is part of Y" and "Y is part of X". The classification of each mention pair is performed using a support vector machine implemented using libLinear (Fan et al., 2008).

To generate the features for each of the mention pairs a proprietary JDPA Tokenizer is used for parsing the document and the Stanford Parser (Klein and Manning, 2003) is used to generate parse trees and part of speech tags for the sentences in the documents.

3.2 Features

We used Zhou et al.'s lexical features (Zhou et al., 2005) as the basis for the features of our system similar to what other researchers have done (Chan and Roth, 2010). Additional work has extended these features (Jiang and Zhai, 2007) or incorporated other data sources (e.g. WordNet), but in this paper we focus solely on the initial step of applying these same lexical features to the JDPA Corpus.

The Mention Level, Overlap, Base Phrase Chunking, Dependency Tree, and Parse Tree features are the same as Zhou et al. (except for using the Stanford Parser rather than the Collins Parser). The minor changes we have made are summarized below:

- **Word Features**: Identical, except rather than using a heuristic to determine the head word of the phrase it is chosen to be the noun (or any other word if there are no nouns in the mention) that is the least deep in the parse tree. This change has minimal impact.
- **Entity Types**: Some of the entity types in the JDPA Corpus indicate the type of the relation (e.g. CarFeature, CarPart) and so we replace those entity types with "Unknown".
- **Token Class**: We added an additional feature (TC12+ET12) indicating the Token Class of the head words (e.g. Abbreviation, DollarAmmount, Honorific) combined with the entity types.
- **Semantic Information**: These features are specific to the ACE relations and so are not used. In Zhou et al.'s work, this set of features increases the overall F-Measure by 1.5.

4 Results

4.1 ACE Corpus Results

We ran our system on the ACE-2004 Corpus as a baseline to prove that the system worked properly and could approximately duplicate Zhou et al.'s results. Using 5-fold cross validation on the newswire and broadcast news documents in the dataset we achieved an average overall F-Measure of 50.6 on the fine-grained relations. Although a bit lower than Zhou et al.'s result of 55.5 (Zhou et al., 2005), we attribute the difference to our use of a different tokenizer, different parser, and having not used the semantic information features.

4.2 JDPA Sentiment Corpus Results

We randomly divided the JDPA Corpus into training (70%), development (10%), and test (20%) datasets. Table 1 shows relation extraction results of the system on the test portion of the corpus. The results are further broken out by three different source types to highlight the differences caused

Relation	**All Documents**			LiveJournal			Blogspot			JDPA		
	P	**R**	**F**	P	R	F	P	R	F	P	R	F
FEATURE OF	44.8	42.3	43.5	26.8	35.8	30.6	44.1	40.0	42.0	59.0	55.0	56.9
MEMBER OF	34.1	10.7	16.3	0.0	0.0	0.0	36.0	13.2	19.4	36.4	13.7	19.9
PART OF	46.5	34.7	39.8	41.4	17.5	24.6	48.1	35.6	40.9	48.8	43.9	46.2
PRODUCES	51.7	49.2	50.4	05.0	36.4	08.8	43.7	36.0	39.5	66.5	64.6	65.6
INSTANCE OF	37.1	16.7	23.0	44.8	14.9	22.4	42.1	13.0	19.9	30.9	29.6	30.2
Overall	46.0	36.2	**40.5**	27.1	22.6	**24.6**	45.2	33.3	**38.3**	53.7	46.5	**49.9**

Table 1: Relation extraction results on the JDPA Corpus test set, broken down by document source.

	LiveJournal	Blogspot	JDPA	ACE
Tokens Per Sentence	**19.2**	18.6	16.5	19.7
Relations Per Sentence	1.08	1.71	**2.56**	0.56
Relations Not In Same Sentence	33%	30%	**27%**	0%
Training Mention Pairs in One Sentence	58,452	54,480	**95,630**	77,572
Mentions Per Sentence	**4.26**	**4.32**	4.03	3.16
Mentions Per Entity	**1.73**	1.63	1.33	2.36
Mentions With Only One Token	**77.3%**	73.2%	61.2%	56.2%

Table 2: Selected document statistics for three JDPA Corpus document sources.

by the writing styles from different types of media: LiveJournal (livejournal.com), a social media site where users comment and discuss stories with each other; Blogspot (blospot.com), Google's blogging platform; and JDPA (jdpower.com's Power Steering blog), consisting of reviews of cars written by JDPA professional writers/analysts. These subsets were selected because they provide the extreme (JDPA and LiveJournal) and average (Blogspot) results for the overall dataset.

5 Analysis

Overall the system is not performing as well as it does on the ACE-2004 dataset. However, there is a 25 point F-Measure difference between the LiveJournal and JDPA authored documents. This suggests that the informal style of the LiveJournal documents may be reducing the effectiveness of the features developed by Zhou et al., which were developed on newswire and broadcast news transcript documents.

In the remainder of this section we look at a statistical analysis of the training portion of the JDPA Corpus, separated by document source, and suggest areas where improved features may be able to aid relation extraction on the JDPA Corpus.

5.1 Document Statistic Effects on Classifier

Table 2 summarizes some important statistical differences between the documents from different sources. These differences suggest two reasons why the instances being used to train the classifier could be skewed disproportionately towards the JDPA authored documents.

First, the JDPA written documents express a much larger number of relations between entities. When training the classifier, these differences will cause a large share of the instances that have a relation to be from a JDPA written document, skewing the classifier towards any language clues specific to these documents.

Second, the number of mention pairs occurring within one sentence is significantly higher in the JDPA authored documents than the other documents. This disparity is even true on a per sentence or per document basis. This provides the classifier with significantly more negative examples written in a JDPA written style.

LiveJournal		**Blogspot**		**JDPA**	
Mention Phrase	%	Mention Phrase	%	Mention Phrase	%
car	6.2	it	8.1	features	2.4
Maybach	5.6	car	2.1	vehicles	1.6
it	3.7	its	2.0	its	1.4
it's	1.7	cars	2.0	Journey	1.3
Maybach 57 S	1.5	Hyundai	2.0	car	1.2
It	1.2	vehicle	1.5	2 T Sport	1.2
mileage	1.1	one	1.5	G37	1.2
its	1.1	engine	1.5	models	1.1
engine	0.9	power	1.1	engine	1.1
57 S	0.9	interior	1.1	It	1.1
Total: 23.9%		**Total: 22.9%**		**Total: 13.6%**	

Table 3: Top 10 phrases in mention pairs whose relation was incorrectly classified, and the total percentage of errors from the top ten.

5.2 Common Errors

Table 3 shows the mention phrases that occur most commonly in the incorrectly classified mention pairs. For the LiveJournal and Blogspot data, many more of the errors are due to a few specific phrases being classified incorrectly such as "car", "Maybach", and various forms of "it". The top four phrases constitute 17% of the errors for LiveJournal and 14% for Blogspot. Whereas the JDPA documents have the errors spread more evenly across mention phrases, with the top 10 phrases constituting 13.6% of the total errors.

Furthermore, the phrases causing many of the problems for the LiveJournal and Blogspot relation detection are generic nouns and pronouns such as "car" and "it". This suggests that the classifier is having difficulty determining relationships when these less descriptive words are involved.

5.3 Vocabulary

To investigate where these variations in phrase error rates comes from, we performed two analyses of the word frequencies in the documents: Table 4 shows the frequency of some common words in the documents; Table 5 shows the frequency of a select set of parts-of-speech per sentence in the document.

Word	Percent of All Tokens in Documents			
	LiveJournal	Blogspot	JDPA	ACE
car	**0.86**	0.71	0.20	0.01
I	**1.91**	1.28	0.24	0.21
it	**1.42**	0.97	0.23	0.63
It	0.33	0.27	**0.35**	0.09
its	**0.25**	0.18	0.22	0.19
the	**4.43**	**4.60**	3.54	4.81

Table 4: Frequency of some common words per token.

POS	POS Occurrence Per Sentence			
	LiveJournal	Blogspot	JDPA	ACE
NN	2.68	3.01	**3.21**	2.90
NNS	0.68	0.73	**0.85**	1.08
NNP	0.93	1.41	**1.89**	1.48
NNPS	0.03	0.03	0.03	0.06
PRP	**0.98**	0.70	0.20	0.57
PRP$	**0.21**	0.18	0.07	0.20

Table 5: Frequency of select part-of-speech tags.

We find that despite all the documents discussing cars, the JDPA reviews use the word "car" much less often, and use proper nouns significantly more often. Although "car" also appears in the top ten errors on the JDPA documents, the total percentage of the errors is one fifth of the error rate on the LiveJournal documents. The JDPA authored documents also tend to have more multi-word mention phrases (Table 2) suggesting that the authors use more descriptive language when referring to an entity. 77.3% of the mentions in LiveJournal documents use only a single word while 61.2% of mentions JDPA authored documents are a single word.

Rather than descriptive noun phrases, the LiveJournal and Blogspot documents make more use of pronouns. LiveJournal especially uses pronouns often, to the point of averaging one per sentence, while JDPA uses only one every five sentences.

5.4 Extra-Sentential Relations

Many relations in the JDPA Corpus occur between entities which are not mentioned in the same sentence. Our system only detects relations between mentions in the same sentence, causing about 29% of entity relations to never be detected (Table 2).

The LiveJournal documents are more likely to contain relationships between entities that are not mentioned in the same sentence. In the semantic role labeling (SRL) domain, extra-sentential arguments have been shown to significantly improve SRL performance (Gerber and Chai, 2010). Improvements in entity relation extraction could likely be made by extending Zhou et al.'s features across sentences.

6 Conclusion

The above analysis shows that at least some of the reason for the system performing worse on the JDPA Corpus than on the ACE-2004 Corpus is that many of the documents in the JDPA Corpus have a different writing style from the news articles in the ACE Corpus. Both the ACE news documents, and the JDPA authored documents are written by professional writers with stronger editorial standards than the other JDPA Corpus documents, and the relation extraction system performs much better on professionally edited documents. The heavy use of pronouns and less descriptive mention phrases in the other documents seems to be one cause of the reduction in relation extraction performance. There is also some evidence that because of the greater number of relations in the JPDA authored documents that the classifier training data could be skewed more towards those documents.

Future work needs to explore features that can address the difference in language usage that the different authors use. This work also does not address whether the relation extraction task is being negatively impacted by poor tokenization or parsing of the documents rather than the problems being caused by the relation classification itself. Further work is also needed to classify extra-sentential relations, as the current methods look only at relations occurring within a single sentence thus ignoring a large percentage of relations between entities.

Acknowledgments

This work was partially funded and supported by J. D. Power and Associates. I would like to thank Nicholas Nicolov, Jason Kessler, and Will Headden for their help in formulating this work, and my thesis advisers: Jim Martin, Rodney Nielsen, and Mike Mozer.

References

Chan, Y. S. and Roth D. *Exploiting Background Knowledge for Relation Extraction*. Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. *LIBLINEAR: A library for large linear classification*. Journal of Machine Learning Research 9(2008), 1871-1874. 2008.

Gerber, M. and Chai, J. *Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1583-1592. 2010.

Jiang, J. and Zhai, C.X. *A systematic exploration of the feature space for relation extraction*. In The Proceedings of NAACL/HLT. 2007.

Kessler J., Eckert M., Clark L., and Nicolov N.. *The ICWSM 2010 JDPA Sentiment Corpus for the Automotive Domain* International AAAI Conference on Weblogs and Social Media Data Challenge Workshop. 2010.

Klein D. and Manning C. *Accurate Unlexicalized Parsing*. Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430. 2003.

Mitchell A., et al. *ACE 2004 Multilingual Training Corpus*. Linguistic Data Consortium, Philadelphia. 2005.

Zhou G., Su J., Zhang J., and Zhang M. *Exploring various knowledge in relation extraction*. Proceedings of the 43rd Annual Meeting of the ACL. 2005.