Annotation Tags for Social Media Content

Kamala Sai Theja Nimmagadda

June 25, 2020

1 Introduction

Annotation of social media content has been a challenging task for researchers for a long time. This is especially true in case of microblog text as they account to a lot of disambiguous content. This report deals with two methods that are used to process said microblog text from twitter. The microblog's (tweets) from twitter pose a significant challenge as tweets are generally very small, averaging over 180 characters per tweet and are very noisy with very short utterances. The presence of hashtags, URL's, abbreviated phrases(e.g. 'ROFL' : rolled on the floor laughing, 'LOL' : laugh out loud etc) and grammatical mistakes make the annotation of microblog data an incredibly difficult task, when compared to news text or articles. As such, conventional NLP tools fail at annotating tweets and microblog text.

2 TwitIE

The first method performs annotation tasks like NER(Named Entity Recognition), semi-automatically using a plugin of GATE[2](General Architecture for Text Engineering) called TwitIE[1]. TwitIE is an open source Information Extraction pipeline for microblog text. For longer texts, NER methods typically have an accuracy of around 90%, but this drops to 30-50% on tweets. TwitIE ensures higher accuracy by accounting for generic differences between textual data in newswire text to microblog text.

TwitIE re-uses the sentence splitter and name gazetteer from ANNIE, a pre-packaged general purpose IE pipeline in GATE unmodified, with some of it's own components like TwitIE tokeniser and Twitter adapted model of Stanford POS tagger. The process starts from collection of corpora from Twitter and outputs a XML dump of IE annotations.

2.1 Tweet Import

The data in Twitter is delivered from Twitter API in JSON format. Format_Twitter plugin in GATE codebase automatically converts this data from twitter into fully annotated GATE document. Tweet objects from same JSON are separated with blank lines.

2.2 Language Identification

The TwitIE system then tries to identify the language using TextCat[3], an N-gram based textual categorization algorithm used to identify languages of given text by building a n-gram of given textual data and comparing it with language profile or fingerprint. It is also possible to train the TextCat[4] on new language by using Fingerprint Generation PR, available in Language_Identification plugin.

2.3 Tokenisation

TwitIE tokeniser is an improvisation over ANNIE's tokeniser and follows Ritter's tokenisation scheme[5]. It treats abbreviations and URL's as one token each and makes Hashtags and user mentions as two tokens each covered in annotation hashtag. Emoticons are handled in separate modules as imformation regarding them is rarely needed.

2.4 Normalisation

TwitIE normaliser uses two separate spelling-correction dictionaries, one generic and other specifically catered to social media posts. The latter contains entries that are most prominent in online conversations. The spelling dictionary corrects these entries using levenshtein distance and double-metaphone distance.

2.5 POS Tagger

Generic POS taggers fail on microblog posts as they tend to be very noisy and short. TwitIE uses an adapted Stanford POS tagger[6] trained on tweets tagged with PTB(Penn TreeBank) tagset. It also pre-tags automatically with regular expressions and ANNIE's gazetteer lookup list to improve the token accuracy for tweets.

2.6 Named Entity Recognition

TwitIE's NER is observed to out-perform ANNIE and Stanford's generic POS tagger and even Ritter's NER algorithm in support of everything mentioned above. However, F1 score significantly drops when compared to aforementioned systems when processing longer news content due to false categorisation.

TwitIE is specifically developed to handle microblog's. This can be attributed to generic longer news text being close to proper written documentation when compared to tweets which are casual texts and are more close to interactive conversations. As such a large gap can be observed between NER performance on microblog's to news content in account of insufficient context and noise. But this can be marked as one of the earlier steps to process microblog data which accounts major discourse.

3 PDTB-style Annotation of Shallow Discourse Relations

PDTB[8](Penn Discourse TreeBank) is a large scale data corpus annotated with discourse structure and semantic related information. This method aims at applying PDTB-style[7] annotations to Twitter conversations as well as conducting a Inter-Annotator Agreement Study. While deeper usage of PDTB-syle annotations are limited to news content, a shallow approach of PDTB-style annotation for discourse structure can also be adapted to microblog text.

An observation has been made that the distribution of discourse relations and connectives in twitter conversations, strictly differ from distribution in generic PDTB text[9]. In particular, twitter conversations hold more CONTINGENCY and less COMPARISON and TEMPORAL relations.

The annotations are made in two phases. The first deals with explicit connectives whose arguments are from a single tweet and the second handles connectives whose Arg1 is located in previous tweet. This approach leverages the property of PDTB to not make strong commitments as to comprehensive structure of discourse.

3.1 Annotation Phase I

Raw data corpus was extracted from Twitter stream. This is used to obtain conversations that are linked to each other using the reply-to relation to form a tree structure with origin of tweet as root node. A single thread i.e., the longest path from root node to one of the leaf nodes is chosen and everything else is discarded. Thus, we annotate the explicit connectives whose arguments are in a single thread or within a single tweet using explicit connectives from PDTB corpus. This can also be helped to identify few new connectives from spoken conversations.

3.2 Annotation Phase II

An Inter-Annotator Agreement study is conducted to verify for exact and partial agreement of selected text spans over Arg1 and Arg2 at all levels of sense assignment. A one character difference at the end and beginning of text spans is accounted to leave room for punctuation. In partial matching, along with exact matches, text spans with overlapping tokens were considered. Upon manual inspection, the agreement for exact argument spans for Arg1 were particularly low in regard to determination of utterance and clause breaks in Twitter. This is the result of presence of social media specific items like emoticons and hashtags.

3.3 Analysis: Twitter vs. WSJ

3.3.1 Qualitative Analysis

Unlike WSJ, which majorly consists of news content, Tweets can be attributed close to interactive conversations. As such, twitter texts are often fragmented and incomplete leading to encounters with nouns and noun phrases standing in for propositions. Twitter also contains a wide range of acronyms used in casual expressions and online conversations. Therefore, a huge difference was observed in the type of prominent connectives encountered in PDTB-style annotation of Twitter data and generic PDTB.

3.3.2 Quantitative Analysis

A quantitative analysis performed on annotations revealed that explicit discourse relations are frequent occurrence in Twitter data. While EXPANSION class tags are similar in both Twitter data and generic PDTB, twitter has a lot more CONTINGENCY and less COMPARISON and TEMPORAL relation tags, which are often observed frequently in generic PDTB on news data.

The PDTB-style annotation of shallow discourse relations reveals that twitter data, despite being in written format are more closely related to spoken conversations, which allows for a deeper study on discourse structure of multi-party interactions.

4 Conclusion

Annotation of social media data has always proven a challenging task. May it be the huge stream of data flowing in everyday or the noise present in the data, it always shows itself as the hardest problem in natural language processing. Twitter data specifically poses a greater challenge due to it's extremely small size and presence of discourse. However, these steps towards processing discourse which is close to human interaction can prove itself to be the important milestone in achieving cognitive intelligence.

5 REFERENCES

1. K. Bontcheva, L. Derczynski, A. Funk, M. Greenwood, D. Maynard, N. Aswani. 2013, "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text"
2. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. 2013, "GATE: an Architecture for Development of Robust HLT Applications"
3. W. Cavnar and J. Trenkle. 1994, "N-Gram-Based Text Categorization"
4. S. Carter, W. Weerkamp and M. Tsagkias. 2013, "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text"
5. A. Ritter, S. Clark, Mausam and O. Etzioni. 2011, "Named Entity Recognition in Tweets: An Experimental Study"
6. K. Toutanova, D. Klein, C. Manning and Y. Singer. 2003, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network"
7. T. Scheffler, B. Aktas, D. Das, M. Stede. 2019, "Annotating Shallow Discourse Relations in Twitter Conversations"
8. R. Prasad, A. Lee, N. Dinesh, E. Miltsakaki, G. Campion, A. Joshi, B. Webber. 2008, "Penn Discourse Treebank Version 2.0"
9. I. Rehbein. 2016, "The role of discourse relations in persuasive texts"