Target-focused Sentiment and Belief Extraction and Classification using CUBISM

Adam Dalton and **Morgan Wixted** and **Yorick Wilks**
Institute for Human and Machine Cognition, Ocala, FL

Meenakshi Alagesan and **Gregorios Katsios** and **Ananya Subburathinam** and **Tomek Strzalkowski**
State University of New York - University at Albany

Abstract

This document contains a brief description of the CUBISM belief and sentiment classification systems.

1 Introduction

The CUBISM system consists of belief and sentiment components that incorporate social aspects of dialogue. The motivation for the systems is that beliefs held by a source entity shape the sentiment towards target objects, sentiment then determines and predicts behavior, and finally, sentiment and belief are observable in language through interaction dynamics and semantic role modeling.

1.1 Sentiment

We describe a novel approach to automatic extraction of sentiment from natural language text, along with the sentiment holder and the sentiment target. We have adapted the affect calculus algorithm (ACA) [Str+14], originally designed to compute affect in metaphors. ACA combines information about syntactic and semantic structure of a sentence with base polarity values of words and phrases in order to estimate polarity and intensity of sentiment from the holder towards the target. Base polarity values for English words are obtained from automatically derived ANEW+ polarity lexicon [Sha+16]. Experimental results are very promising, including the best performance at 2014 TAC KBPSent Slot Filling evaluation. Further improvements reported here have doubled the systems performance accuracy as shown on reruns on 2014 KBP data and the KBP 2016 sentiment training data.

We focus here on extraction of individual instances of sentiment. Our system works on any type of text by attempting to instantiate the *sentiment triple*:

<sentiment holder, sentiment relation, sentiment target>

The *sentiment holder* and sentiment target are selected from among the entities and relations already annotated in the data. The speaker/writer is always a potential holder. In such case any entity, relation or event in the text attributed to this speaker/writer is a potential target unless it is in a segment expressly attributed to another source.

An agent/subject in each sentence/clause is a potential holder for reported sentiment. In this case the patient/object is a potential target.

The *sentiment relation* is the verb or other lexical item with Target as an argument.

The Holder and Target are restricted to ERE entities (or relations or events), and may come from Gold annotation or from automatically generated annotation.

The instantiated *sentiment triple* is passed to Affect Calculus for sentiment determination. Affect Calculus is described in detail in [Str+14]. AC works by first determining the type of sentiment relation (agentive, patientive or propertive), typically anchored at verbs or action nouns, that has the sentiment target as one of its arguments. The sentiment holder may be one of the other arguments, but it is more typically outside of the relation, reporting it.

The type of sentiment relation is determined based on the role of the sentiment target. For example, if the target entity is in the agent role (typically the subject of a sentence), then the relation is agentive; conversely, if the

This work was supported, in part, by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-12-2-0348 (Adam Dalton,Yorick Wilks, Meenakshi Alagesan, Gregorios Katsios, Ananya Subburathinam, and Tomek Strzalkowski), and (Morgan Wixted) through a summer internship funded through the Institute for Human and Machine Cognition.

target is in a patient role, the sentiment relation is patientive. A propertive relation is when a property of the target is described, typically in a unary relation often anchored at an adjective.

- **Propertive:** the way Target appears, looks, smells, sounds, feels, etc.
  - Examples: <heavy, harmful to, affordable
- **Agentive:** the way Target acts or affects other things
  - Examples: <crushing, helps, adapts,
- **Patientive:** the way to deal with Target or to affect it
  - Examples: <Navigate, fight, donate to,

Target role is determined by syntactic information obtained from a dependency parser. Here are some examples of sentiment target (GMO) occurring in different roles (in all cases below the sentiment holder is the writer):

- ***GMOs*** *pollute the environment.*
  - Relation type: Agentive
  - Relation is highly negative (1.85)
- *Also, certain* ***GMO's*** *are nutrient enriched, so that's an advantage.*
  - Relation type: Propertive
  - Relation is highly positive (7.7)
- *It is easier for farmers to grow* ***GMOs*** *with less loss.*
  - Relation type: Patientive
  - Relation is slightly positive (5.6)

In order to calculate an affect score (and thus sentiment) towards a target entity we first assign polarity and strength values to the extracted relations and their arguments based on the expanded Affective Norms in English (ANEW+) lexicon [Sha+16]. We then combine these scores based on the type of relation wrt. the target, using the Affect Calculus (Table 1). This way we obtain the value of sentiment towards the target in the <H, R, T>triple. It should be noted that the sentiment holder is determined by a separate process as it is often not part of the same relation.

In the above examples, the sentiment target was an entity; however, we should note here that the process works analogously for sentiment targets that are events or relations, which may be expressed by nominalizations or by embedded clauses, e.g., **Yanukovich was also negotiating with Russia and Belarus for a customs union,** *which left the EU negotiators confused.*

The goal of our system is to extract all instances of sentiment between any potential holder and a potential target mentioned in a text document (including the writer/author as a holder). To do so, we need to locate all possible holdertarget pairs and then apply the algorithm described above to extract sentiment between them. Our system does the following two steps to find all possible candidates from each sentence:

- To find sentiment from the speaker/writer towards the entities in their post or message or other text piece, we need to consider all the entities mentioned, but particularly these appearing in any of the three relation types described above. Other positions of the target entity may be possible (e.g., origin or destination of a motion relation), which are included in an extended version of AC [Str+14]; however, we did not incorporate these into our sentiment algorithm at this time.
- To find sentiment from a holder as reported by the speaker/writer, we look for entity pairs that (1) are in an agentpatient relation or (2) when one entity is explicitly reported as expressing opinion about a relation involving another entity or event.

1.2 Belief

The belief classifier operates over a graph constructed from the entity, relation, and event data provided for the task. Each post in discussion forum documents is also tagged with a dialogue act by implementing an approach inspired by dependency parsing [Wan+11]. Finally, nodes in the graph are assigned membership to a community [RAK07]on the assumption that authors who interact with have the same type of beliefs on similar event and relation types. Beliefs are then created for each event and relation and labeled using a Naive Bayes Classifier using training data developed for this task **??**). After initial labeling, edges are placed between beliefs that share either author, event type, or relation type. Relaxation labeling [AW06] is applied as a final step to refine the labels.

2 Sentiment extraction in 2016 TAC BEST evaluation

In this section we describe further details the process of sentiment extraction from sentences identified as containing possible sentiment mentions towards entities, relations, or events identified in the input data. This evaluation differs from the 2014 edition where the input text was not annotated in any way; however, a query specifying either the holder or the target was supplied along with the sentiment orientation sought. This restricted sentences of interest to those containing a reference to either target or

Relation type	Type 1 (propertive) Rel(Query)	Type 2 (agentive) Rel (Query, X)		Type 3 (patientive) Rel(X, Query)	
Relation/X		X $\geq$ neutral	X <neutral	X $\geq$ neutral	X <neutral
Positive	positive	positive	$\leq$ unsymp	Positive	$\leq$ sympat
Negative	negative	$\leq$ unsymp	$\geq$ sympat	$\leq$ sympat	$\geq$ sympat
Neutral	neutral	neutral	$\leq$ neutral	neutral	$\leq$ neutral

Table 1: A simple affect calculus specifies affect polarity towards a target as an argument of a affect carrying relation using a 5-point polarity scale [negative <unsympathetic <neutral <sympathetic <positive]. X is an optional second argument of the relation.

the holder (including the holder as writer); however, in 2016 version, all targets and holders need to be considered, and thus no query was given. Instead, a selection of entities, relations and events of interest were annotated on the input text, either by hand (the Gold condition) or by an automated system (the Predicted condition).

In the following, we detail this process step-by-step.

Determining Holder and Target

As already described above, all annotated entities were potential holders, and all entities, relations and events were potential targets.

Semantic Role decision

We create a dependency tree of each sentence selected in step 1, using a dependency parser (e.g., Marneffe et al. 2006) to determine if the conditions for the holder and the target are met. This is indeed the case for the example sentence shown below in Figure 1. The extracted entities and relations between them are then passed to the Affect Calculus module for computing the sentiment value.

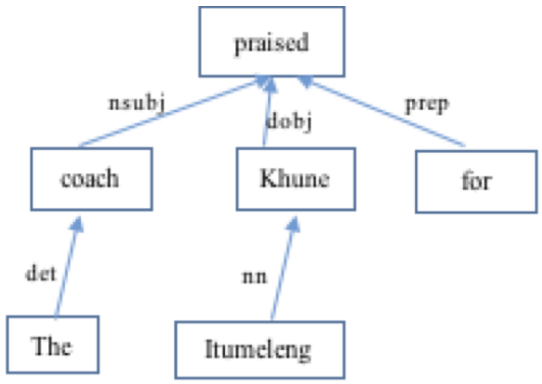

Figure 1. Sentiment-carrying relation "praise" in a sentence dependency structure

Applying Affect Calculus

In order to accurately calculate sentiment towards the sentiment target, we first look up base affect associated with the relation between H, T and any other arguments involved (the X argument). This is done by consulting the expanded ANEW+ lexicon (Shaikh et al, 2016). In our example, the relation "praise" returns 7.24 (strong positive); the entity "Itumeleng Khune", as expected, is not listed in ANEW+ so no base score is returned, which is thus assumed to be neutral. Since "praise" is a patientive relation wrt. the target, according to the AC Table 1, the overall sentiment from the holder to the target is positive.

Reported/nested relations

A common case of nested relations is an indirect sentiment report, as shown in Figure 2. Here the sentiment target is an argument of a relation that is embedded in another relation where the sentiment holder appears. A common embedding relation are reporting verbs such as "say", "claim", "announce" etc., which are neutral and can be treated the same way as quoted speech. (We should note that sentiment loaded embedding relations are also possible, e.g., "is afraid that", "is proud to announce that", etc., which are handled as sentiment towards events or relations or their arguments.) When the embedding relation is a neutral reporting verb, the relation from the complement is simply used in affect calculation. For the sentence in Figure 2 the system skips over the reporting relation, "said", and takes the propertive relation "perfect" involving the sentiment target "Altman".

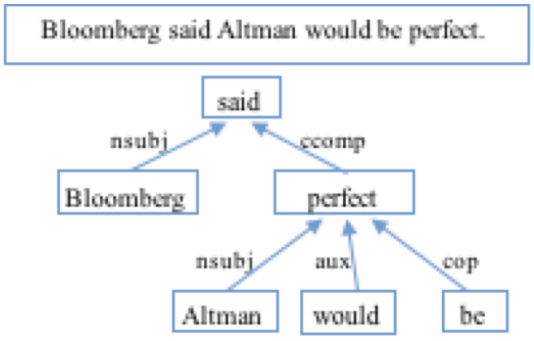

Figure 2. A sentence contains report sentiment

Correcting for negation

Negation can affect the meaning or overturn the polarity of an entire sentence. In our system, negation detection relies on the dependency type, neg, which is negation modifier. If the relation has neg as one of its children, we negate the affect score of the relation. For example, in the sentence, "I dont like NY", the dependency type between "n't" and "like" is neg; consequently, the negated affect score of "like" is used. To calculate the negated affect score, we simply subtract the base score (7.52 for "like") from 9 (which is the maximum score in ANEW+). Thus, the affect score for "don't like" is $9 - 7.52 = 1.48$, which is very strong negative score, (ANEW+ sentiment scores range from 1 (very negative) to 9 (very positive) on a continuous scale; the interval (4.5 to 5.5) is typically considered neutral.)

2.1 Types of Runs submitted

For all of the run submissions the run never accessed the web and the confidence values were based on the Affective Norms of Words (ANEW) lexicon valence scores. The higher the valence score in ANEW, the higher our confidence in the answer. The ANEW lexicon has valence score of words ranging from 1-9, 1 being more negative and 9 being more positive.

Run 1. In the first run, the neutral sentiment range is 4.0 to 6.0. If the ANEW score is less than 4.0, we would assign negative sentiment to the appropriate elements of the sentiment triple (specifically the sentiment relation and its arguments other than the target). If the score was greater than 6.0, a positive sentiment was assigned. Neutral sentiment was used for all items that had no entries in ANEW+. This run includes results for the languages English, Spanish and Chinese. Additionally, for the NW subset of the English language, the sentiment triples which contain a possessive relation are excluded from the process. The DF subset of the English language, as well as all sets from the other two languages, do not make this exception.

Run 2. This run includes results for the English language only, and the rules for the neutral range are the same as in the previous one. The previous exception regarding the presence of possessives in the sentiment triple does not apply.

Run 3. This run includes results for the English language only, and the rules for the neutral range are the same as in the previous ones. The exception regarding the presence of possessives in the sentiment triple applies to both NW and DF subsets.

2.2 Spanish and Chinese sentiment

Our system also works on Spanish and Chinese texts. The algorithm for sentiment extraction is the same as for English, except for the use of language specific sentiment lexicons. Spanish version of ANEW+ (or ANSW+) has been derived from the English ANEW+ using Wordnet-based translation. A subsequent validation, including human validation and comparison against an existing Spanish affective norms lexicon (Redondo et al., 2007) has shown strong correlation with human judgment (Shaikh et al, 2016). The Chinese version of ANEW+ (or ANCW+) came from two resources, Chinese sentiment vocabulary (VSA)[1] and Translation from English ANEW+. VSA is a Chinese Lexicon that contains 8942 positive and negative sentiment entries accessed by human experts. The major issue of VSA is the size limitation, therefore, we employed Bing API to translate English ANEW+ into Chinese with scores averaged for many to one translations. From these two resources, we totally collected 75241 sentiment entries for Chinese ANEW+. It is undergoing validation through the same process as we used for Spanish.

The Chinese texts were parsed with the dependency parser[2] included in the Stanford CoreNLP[3] toolkit. Spanish texts were parsed using the dependency parser[4] that was included in the FreeLing[5] toolkit.

3 Belief extraction in 2016 TAC BEST evaluation

Our approach on belief relied on the notion that the mental state strength and strategy (stated vs reported) is, at least in part, a function of the source and target. That is, a given individual or collective will discuss events and relations in a consistent way depending on the roles and corresponding entities. For this reason, we populated a Neo4j[6] graph database with the content of both the provided richERE annotations and data provided in the source documents. We used the development set provided by the task to attach belief nodes to event and relation arguments, and used that information to train a naive bayes classifier.

Determining Holder and Target

The targets of beliefs were determined to consistently be the arguments that made up the triggers and roles of relations and events. In order to determine the source

[1] http://www.keenage.com/html/e_index.html

[2] http://nlp.stanford.edu/software/nndep.shtml

[3] http://stanfordnlp.github.io/CoreNLP/

[4] https://www.researchgate.net/publication/28167462_TXALA_un_analizador_libre_de_dependencias_para_el_castellano

[5] http://nlp.lsi.upc.edu/freeling/node/1

[6] https://neo4j.com/

of belief via the offset information available, we developed an XML parsing expression grammar using Treetop [7]. Treetop was able to provide offsets consistent with the ERE annotations provided, allowing our system to perform an entity-mention lookup in the graph database. Reported belief sources were not implemented at this time, but future work will incorporate dependency parsing and semantic role labeling to achieve that aim.

3.1 Types of Runs submitted

One run was submitted to the 2016 BeSt evaluation and it consisted of a Naive Bayes Classifier trained on the following feature set.

From the richERE annotations:

- Event Type
- Event Subtype
- Event Realis
- Trigger Content
- Argument Role
- Argument Realis
- Relation Type
- Relation Subtype
- Relation Realis
- Relation Type
- Relation Relation Arg Content
- Relation Arg 1 Content
- Relation Arg 2 Content

We also included the the text snippet surrounding the arguments as a feature.

Weka was used to implement the Naive Bayes Classifier. The features consisting of Trigger Content, the Snippet, and Arg Contents were implemented as string attributes. The remaining features were implemented as nominal attributes.

After the evaluation we implemented the forum dialogue act parser described in [Wan+11]. The dataset provided for that task was a technical support discussion board, and the dialogue acts described were could potentially be informative to the task of mental state classification. Their selection of features is also likely to be a good first step for developing a dialogue act parser for determining belief types.

3.2 Spanish and Chinese sentiment

No changes were required to run our belief system on other languages.

[7] `http://treetop.rubyforge.org/index.html`

4 Evaluation and Results

In Table 2 and 3, we show the respective performances of the sentiment and belief systems for the runs submitted in 2016 evaluation and the follow up belief run that included dialogue acts.

All sentiment runs had similar performance, with Run 3 achieving hte best precision and Runs 1 and 2 having better recall. Overall, the system performs best on English documents, however precision is highest in Spanish. Table 4 shows performance on predicted input condition. As expected the results are significantly lower due to the errors introduced by entity, event and relation annotation.

Experimenting with the range of neutral scores, so as to exclude spurious answers, while also including as many valid responses is a critical piece that we will work in the future.

5 Discussion and Future Work

One clear piece of future work is to determine the best range of values to consider in the neutral zone from the range of valence scores in ANEW lexicon. Using an optimized range will maximize performance. Furthermore, this value appears to vary by context and genre of text, and data driven optimization may be appropriate.

The belief approach was extremely reliant on the features provided by the richERE. Going forward, we will attempt to increase the usage of information about the source of beliefs. One aspect we will be experimenting with the the ”believability” or ”trustworthiness” of the source. Group membership and status within the group will also be evaluated. Finally, we will review the dialogue acts to determine if a new set of acts could be informative in determining mental state and dependency parsing can be used to attribute beliefs to reported on sources.

References

[AW06] Ralitsa Angelova and Gerhard Weikum. “Graph-based text classification: learn from your neighbors”. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2006, pp. 485–492.

		En Gold Run 1	En Gold Run 2	En Gold Run 3	Sp Gold Run 4	Ch Gold Run 5
Precision	DF	0.139	0.139	0.147	0.167	0.067
	NW	0.047	0.046	0.047	0.078	0.019
Recall	DF	0.166	0.166	0.155	0.043	0.093
	NW	0.019	0.021	0.019	0.014	0.054
F-score	DF	0.151	0.151	0.151	0.068	0.078
	NW	0.027	0.029	0.027	0.024	0.028

Table 2: Performance of ”gold” condition sentiment runs submitted during BeST evaluation.

		En Gold Run 1	En Gold Run 2	En Gold Run 3	Sp Gold Run 4	Ch Gold Run 5
Precision	DF	0.682	0.682	0.76	0.57	0.687
	NW	0.814	0.814	-	0.608	0.751
Recall	DF	0.591	0.591	0.533	0.50	0.672
	NW	0.547	0.547	-	0.404	0.514
F-score	DF	0.633	0.633	627	0.50	0.679
	NW	0.654	0.654	-	0.486	0.610

Table 3: Performance of ”gold” condition belief runs. Run 1 was submitted during BeST evaluation. Runs 2 and 3 were scored once the evaluation gold standard was released

		En Pred Run 1	En Pred Run 2	En Pred Run 3	Sp Pred Run 4	Ch Pred Run 5
Precision	DF	-	-	-	0.136	0.032
	NW	-	-	-	0.042	0.03
Recall	DF	-	-	-	0.004	0.011
	NW	-	-	-	0.001	0.028
F-score	DF	-	-	-	0.007	0.016
	NW	-	-	-	0.002	0.029

Table 4: Performance of ”predicted” condition sentiment runs. These are the same runs as in table 2 except applied to different input condition. English runs were not scored.

[RAK07] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. “Near linear time algorithm to detect community structures in large-scale networks”. In: *Physical review E* 76.3 (2007), p. 036106.

[Wan+11] Li Wang et al. “Predicting thread discourse structure over technical web forums”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 13–25.

[Str+14] Tomek Strzalkowski et al. “Computing affect in metaphors”. In: *Proceedings of the Second Workshop on Metaphor in NLP*. 2014, pp. 42–51.

[Sha+16] Samira Shaikh et al. “ANEW+: Automatic Expansion and Validation of Affective Norms of Words Lexicons in Multiple Languages”. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC16)*. 2016.