Committed Belief Annotation and Tagging

Mona T. Diab
CCLS
Columbia U.
mdiab@cs.columbia.edu

Lori Levin
LTI
CMU
lsl@cs.cmu.edu

Teruko Mitamura
LTI
CMU
teruko+@cs.cmu.edu

Owen Rambow
CCLS
Columbia U.
rambow@ccls.columbia.edu

Vinodkumar Prabhakaran
CS
Columbia U.

Weiwei Guo
CS
Columbia U.

Abstract

We present a preliminary pilot study of belief annotation and automatic tagging. Our objective is to explore semantic meaning beyond surface propositions. We aim to model people's cognitive states, namely their beliefs as expressed through linguistic means. We model the strength of their beliefs and their (the human) degree of commitment to their utterance. We explore only the perspective of the author of a text. We classify predicates into one of three possibilities: committed belief, non committed belief, or not applicable. We proceed to manually annotate data to that end, then we build a supervised framework to test the feasibility of automatically predicting these belief states. Even though the data is relatively small, we show that automatic prediction of a belief class is a feasible task. Using syntactic features, we are able to obtain significant improvements over a simple baseline of 23% F-measure absolute points. The best performing automatic tagging condition is where we use POS tag, word type feature AlphaNumeric, and shallow syntactic chunk information CHUNK. Our best overall performance is 53.97% F-measure.

1 Introduction

As access to large amounts of textual information increases, there is a strong realization that searches and processing purely based on surface words is highly limiting. Researchers in information retrieval and natural language processing (NLP) have long used morphological and (in a more limited way) syntactic analysis to improve access and processing of text; recently, interest has grown in relating text to more abstract representations of its propositional meaning, as witnessed by work on semantic role labeling, word sense disambiguation, and textual entailment. However, there are more levels to "meaning" than just propositional content. Consider the following examples, and suppose we find these sentences in the *New York Times*:[1]

(1) a. GM will lay off workers.

b. A spokesman for GM said GM will lay off workers.

c. GM may lay off workers.

d. The politician claimed that GM will lay off workers.

e. Some wish GM would lay of workers.

f. Will GM lay off workers?

g. Many wonder if GM will lay off workers.

If we are searching text to find out whether GM will lay off workers, all of the sentences in (1) con-

[1] In this paper, we concentrate on written communication, and we use the terms *reader* and *writer*. However, nothing in the approach precludes applying it to spoken communication.

tain the proposition LAYOFF(GM,WORKERS). However, the six sentences clearly allow us very different inferences about whether GM will lay off workers or not. Supposing we consider the *Times* a trustworthy news source, we would be fairly certain with (1a) and (1b). (1c) suggests the *Times* is not certain about the layoffs, but considers them possible. When reading (1d), we know that someone else thinks that GM will lay off workers, but that the *Times* does not necessarily share this belief. (1e), (1f), and (1g) do not tell us anything about whether anyone believes whether GM will lay off workers.

In order to tease apart what is happening, we need to refine a simple IR-ish view of text as a repository of propositions about the world. We use two theories to aid us. The first theory is that in addition to facts about the world (GM will or will not lay off workers), we have facts about people's cognitive states, and these cognitive states relate their bearer to the facts in the world. (Though perhaps there are only cognitive states, and no facts about the world.) Following the literature in Artificial Intelligence (Cohen and Levesque, 1990), we can model cognitive state as beliefs, desires, and intentions. In this paper, we are only interested in beliefs (and in distinguishing them from desires and intentions). The second theory is that communication is intention-driven, and understanding text actually means understanding the communicative intention of the writer. Furthermore, communicative intentions are intentions to affect the reader's cognitive state – his or her beliefs, desires, and/or intentions. This view has been worked out in the text generation and dialog community more than in the text understanding community (Mann and Thompson, 1987; Hovy, 1993; Moore, 1994).

In this paper we are interested in exploring the following: we would like to recognize what the text wants to make us believe about various people's cognitive states, including the speaker's. As mentioned, we are only interested in people's belief. In this view, the result of text processing is not a list of facts about the world, but a list of facts about different people's cognitive states.

This paper is part of an on-going research effort. The goals of this paper are to summarize a pilot annotation effort, and to present the results of initial experiments in automatically extracting facts about people's beliefs from open domain running text.

2 Belief Annotation

We have developed a manual for annotating belief, which we summarize here. For more detailed information, we refer to the cited works. In general, we are interested in the writer's intention as to making us believe that various people have certain beliefs, desires, and intentions. We simplify the annotation in two ways: we are only interetsed in beliefs, and we are only interested in the writer's beliefs. This is not because we think this is the only interesting information in text, but we do this in order to obtain a manageable annotation in our pilot study. Specifically, we annotate whether the writer intends the reader to interpret a stated proposition as the writer's strongly held belief, as a proposition which the writer does not believe strongly (but could), or as a proposition towards which the writer has an entirely different cognitive attitude, such as desire or intention. We do not annotate subjectivity (Janyce Wiebe and Martin, 2004; Wilson and Wiebe, 2005), nor opinion (for example: (Somasundaran et al., 2008)): the nature of the proposition (opinion and type of opinion, statement about interior world, external world) is not of interest. Thus, this work is orthogonal to the extensive literature on opinion detection. And we do not annotate truth: real-world (encyclopedic) truth is not relevant.

We have three categories:

- Committed belief (CB): the writer indicates in this utterance that he or she believes the proposition. For example, *GM has laid off workers*, or, even stronger, *We know that GM has laid off workers*.

  A subcase of committed belief concerns propositions about the future, such as *GM will lay off workers*. People can have equally strong beliefs about the future as about the past, though in practice probably we have stronger beliefs about the past than about the future.

- Non-committed belief (NCB): the writer identifies the propositon as something which he or she could believe, but he or she happens not to have a strong belief in. There are two subcases. First, there are cases in which the writer makes clear that the belief is not strong, for example by using a modal auxiliary:[2] *GM may lay off workers*. Second, in reported speech, the writer is not signaling to us what he or she believes about the reported speech: *The politician claimed that GM will lay off workers*. However, sometimes, we can use the speech act verb to infer the writer's attitude,[3] and we can use our own knowledge

[2] The annotators must distinguish epistemic and deontic uses of modals.

[3] Some languages may also use grammatical devices; for

to infer the writer's beliefs; for example, in *A GM spokesman said that GM will lay off workers*, we can assume that the writer believes that GM intends to lay off workers, not just the spokesman. However, this is not part of the annotation, and all reported speech is annotated as NCB. Again, the issue of tense is orthogonal.

- Not applicable (NA): for the writer, the proposition is not of the type in which he or she is expressing a belief, or could express a belief. Usually, this is because the proposition does not have a truth value in this world (be it in the past or in the future). This covers expressions of desire (*Some wish GM would lay of workers*), questions (Will GM lay off workers? or *Many wonder if GM will lay off workers*, and expressions of requirements (*GM is required to lay off workers* or *Lay off workers!*).

This sort of annotation is part of an annotation of all "modalities" that a text may express. We only annotate belief. A further complication is that these modalities can be nested: one can express a belief about someone else's belief, and one may be strong and the other weak (*I believe John may believe that GM will lay off workers*). At this phase, we only annotate from the perspective of the writer, i.e. what the writer of the text that is being annotated believes.

The annotation units (annotatables) are, conceptually, propositions as defined by PropBank (Kingsbury et al., 2002). In practice, annotators are asked to identify full lexical verbs (whether in main or embedded clauses, whether finite or non-finite). In predicative constructions (*John is a doctor/in the kitchen/drunk*), we ask them to identify the nominal, prepositional, or adjectival head rather than the form of *to be*, in order to also handle small clauses (*I think [John an idiot]*).

The interest of the annotation is clear: we want to be able to determine automatically from a given text what beliefs we can ascribe to the writer, and with what strengths he or she holds them. Across languages, many different linguistic means are used to denote this attitude towards an uttered proposition, including syntax, lexicon, and morphology. To our knowledge, no systematic empirical study exists for English, and this annotation is a step towards that goal.

3 Related Work

The work of Roser et al. (2006) is, in many respects, very similar to ours. In particular, they are concerned with extracting information about people's beliefs and the strength of these beliefs from text. However, their annotation is very different from ours. They extend the TimeML annotation scheme to include annotation of markers of belief and strength of belief. For example, in the sentence *The Human Rights Committee regretted that discrimination against women persisted in practice*, TimeML identifies the events associated with the verbs *regret* and *persist*, and then the extension to the annotation adds the mark that there is a "factive" link between the *regret* event and the *persist* event, i.e., if we regret something, then we assume the truth of that something. In contrast, in our annotation, we directly annotate events with their level of belief. In this example, we would annotate *persist* as being a committed belief of the Human Rights Committee (though in this paper we only report on beliefs attributed to the writer). This difference is important, as in the annotation of Roser et al. (2006), the annotator must analyze the situation and find evidence for the level of belief attributed to an event. As a result, we cannot use the annotation to *discover* how natural language expresses level of belief. Our annotation is more primitively semantic: we ask the annotators simply to annotate meaning (does X believe the event takes place), as opposed to annotating the linguistic structures which express meaning. As a consequence of the difference in annotation, we cannot compare our automatic prediction results to theirs.

Other related works explored belief systems in an inference scenario as opposed to an intentionality scenario. In work by (Ralf Krestel and Bergler, 2007; Krestel et al., 2008), the authors explore belief in the context of news media exploring reported speech where they track newspaper text looking for elements indicating evidentiality. The notion of belief is more akin to finding statements that support or negate specific events with different degrees of support. This is different from our notion of committed belief in this work, since we seek to make explicit the intention of the author or the speaker.

4 Our Approach

4.1 Data

We create a relatively small corpus of English manually annotated for the three categories: CB, NCB, NA. The data covers different domains and genres from newswire, to blog data, to email correspondence, to letter correspondence, to tran-

example, in German, the choice between indicative mood and subjunctive mood in reported speech can signal the writer's attitude.

scribed dialogue data. The data comprises 10K words of running text. 70% of the data was doubly annotated comprising 6188 potentially annotatable tokens. Hence we had a 4 way manual classification in essence between NONE, CB, NCB, and NA. Most of the confusions between NONE and CB from both annotators, for 103 tokens. The next point of disagreement was on NCB and NONE for 48 tokens.They disagreed on NCB and CB for 32 of the tokens. In general the interannotator agreements were high as they agreed 95.8% of the time on the annotatable and the exact belief classification.[4] Here is an example of a disagreement between the two annotators, *The Iraqi government has* **agreed** *to let Rep Tony Hall visit the country next week to assess a humanitarian crisis that has festered since the Gulf War of 1991 Hall's office said Monday.* One annotator deemed "agreed" a CB while the other considered it an NCB.

4.2 Automatic approach

Once we had the data manually annotated and revised, we wanted to explore the feasibility of automatically predicting belief states based on linguistic features. We apply a supervised learning framework to the problem of both identifying and classifying a *belief* annotatable token in context. This is a three way classification task where an annotatable token is tagged as one of our three classes: Committed Belief (CB), Non Committed Belief (NCB), and Not Applicable (NA). We adopt a chunking approach to the problem using an Inside Outside Beginning (IOB) tagging framework for performing the identification and classification of belief tokens in context. For chunk tagging, we use YamCha sequence labeling system.[5] YamCha is based on SVM technology. We use the default parameter settings most importantly the kernels are polynomial degree 2 with a c value of 0.5.

We label each sentence with standard IOB tags. Since this is a ternary classification task, we have 7 different tags: B-CB (Beginning of a committed belief chunk), I-CB (Inside of a committed belief chunk), B-NCB (Beginning of non committed belief chunk), I-NCB (Inside of a non committed belief chunk), B-NA (Beginning of a not applicable chunk), I-NA (Inside a not applicable chunk), and O (Outside a chunk) for the cases that are not annotatable tokens. As an example of the annotation, a sentence such as *H*all said he wanted to investigate reports from relief agencies that a quarter of Iraqi children may be suffering from chronic malnutrition. will be annotated as follows: {Hall_O said_B-CB he_O wanted_B-NCB to_B-NA investigate_I-NA reports_O from_O relief_O agencies_O that_O a_O quarter_O of_O Iraqi_O children_O may_O be_O suffering_B-NCB from_O chronic_O malnutrition_O.}

We experiment with some basic features and some more linguistically motivated ones.

CXT: Since we adopt a sequence labeling paradigm, we experiment with different window sizes for context ranging from $-/+2$ tokens after and before the token of interest to $-/+5$.

NGRAM: This is a character n-gram feature, explicity representing the first and last character ngrams of a word. In this case we experiment with up to $-/+4$ characters of a token. This feature allows us to capture implicitly the word inflection morphology.

POS: An important feature is the Part-of-Speech (POS) tag of the words. Most of the annotatables are predicates but not all predicates in the text are annotatables. We obtain the POS tags from the TreeTagger POS tagger tool which is trained on the Penn Treebank.[6]

ALPHANUM: This feature indicates whether the word has a digit in it or not or if it is a non alphanumeric token.

VerbType: We classify the verbs as to whether they are modals (eg. may, might, shall, will, should, can, etc.), auxilliaries (eg. do, be, have),[7] or regular verbs. Many of our annotatables occur in the vicinity of modals and auxilliaries. The list of modals and auxilliaries is deterministic.

Syntactic Chunk (CHUNK): This feature explicitly models the syntactic phrases in which our tokens occur. The possible phrases are shallow syntactic representations that we obtain from the TreeTagger chunker:[8] ADJC (Adjective Chunk), ADVC (Adverbial Chunk), CONJC (Conjunctional Chunk), INTJ (Interjunctional Chunk), LST (numbers 1, 2,3 etc), NC (Noun Chunk), PC (Prepositional Chunk), PRT (off,out,up etc), VC (Verb Chunk).

5 Experiments and Results

5.1 Conditions

Since the data is very small, we tested our automatic annotation using 5 fold cross validation

[4] This interannotator agreement number includes the NONE category.

[5] http://www.tado-chasen.com/yamcha

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] We realize in some of the grammar books auxilliaries include modal verbs.

[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

where 10% of the data is set aside as development data, then 70% is used for training and 20% for testing. The reported results are averaged over the 5 folds for the Test data for each of our experimental conditions.

Our baseline condition is using the tokenized words only with no other features (TOK). We empirically establish that a context size of $-/+3$ yields the best results in the baseline condition as evaluated on the development data set. Hence all the results are yielded from a CXT of size 3.

The next conditions present the impact of adding a single feature at a time and then combining them. It is worth noting that the results reflect the ability of the classifier to identify a token that could be annotatable and also classify it correctly as one of the possible classes.

5.2 Evaluation Metrics

We use $F_{\beta=1}$ (F-measure) as the harmonic mean between (P)recision and (R)ecall. All the presented results are the F-measure. We report the results separately for the three classes CB, NCB, and NA as well as the overall global F measure for any one condition averaged over the 5 folds of the TEST data set.

5.3 Results

In Table 1 we present the results yielded per condition including the baseline TOK and presented for the three different classes as well as the overall F-measure.

All the results yielded by our experiments outperform the baseline TOK. We highlight the highest performing conditions in Table 1: TOK+AlphaNum+POS +CHUNK, TOK+AN+POS and TOK+POS. Even though all the features independently outperform the baseline TOK in isolation, POS is the single most contributing feature. The least contributing factor independently is the AlphaNumeric feature AN. However combining AN with character Ngram NG yields better results than using each of them independently. We note that adding NG to any other feature combination is not helpful, in fact it seems to add noise rather than signal to the learning process in the presence of more sophisticated features such as POS or syntactic chunk information. Adding the verbtype VT explicitly as a feature is not helpful for all categories, it seems most effective with CB. As mentioned earlier we deterministically considered all modal verbs to be modal. This might not be the case for all modal auxilliaries since some of them are used epistemically while others deontically, hence our feature could be introducing an element of noise. Adding syntactic chunk information helps boost the results by a small margin from 53.5 to 53.97 F-measure. All the results seem to suggest the domination of the POS feature and it's importance for such a tagging problem. In general our performance on CB is the highest, followed by NA then we note that NCB is the hardest category to predict. Examining the data, NCB has the lowest number of occurrence instances in this data set across the board in the whole data set and accordingly in the training data, which might explain the very low performance. Also in our annotation effort, it was the hardest category to annotate since the annotation takes more than the sentential context into account. Hence a typical CB verb such as "believe" in the scope of a reporting predicate such as "say" as in the following example *Mary said he believed the suspect with no qualms*. The verb *believed* should be tagged NCB however in most cases it is tagged as a CB. Our syntactic feature CHUNK helps a little but it does not capture the overall dependencies in the structure. We believe that representing deeper syntactic structure should help tremendously as it will model these relatively longer dependencies.

We also calculated a confusion matrix for the different classes. The majority of the errors are identification errors where an annotatable is considered an O class as opposed to one of the 3 relevant classes. This suggests that identifying the annotatable words is a harder task than classification into one of the three classes, which is consistent with our observation from the interannotator disagreements where most of their disagreements were on the annotatable tokens, though a small overall number of tokens, 103 tokens out of 6188, it was the most significant disagreement category. We find that for the TOK+POS condition, CBs are mistagged as un-annotatable O 55% of the time. We find most of the confusions between NA and CB, and NCB and CB, both cases favoring a CB tag.

6 Conclusion

We presented a preliminary pilot study of belief annotation and automatic tagging. Even though the data is relatively tiny, we show that automatic prediction of a belief class is a feasible task. Using syntactic features, we are able to obtain significant improvements over a simple baseline of 23% F-measure absolute points. The best performing automatic tagging condition is where we use POS tag, word type feature AlphaNumeric, and shallow syntactic chunk information CHUNK. Our best overall performance is 53.97% F-measure.

	CB	NA	NCB	Overall F
TOK	25.12	41.18	13.64	30.3
TOK+NG	33.18	42.29	5	34.25
TOK+AN	30.43	44.57	12.24	33.92
TOK+AN+NG	37.17	42.46	9.3	36.61
TOK+POS	54.8	59.23	13.95	**53.5**
TOK+NG+POS	43.15	50.5	**22.73**	44.35
TOK+AN+POS	54.79	58.97	22.64	**53.54**
TOK+NG+AN+POS	43.09	54.98	18.18	45.91
TOK+POS+CHUNK	55.45	57.5	15.38	52.77
TOK+POS+VT+CHUNK	53.74	57.14	14.29	51.43
TOK+AN+POS+CHUNK	55.89	**59.59**	22.58	**53.97**
TOK+AN+POS+VT+CHUNK	**56.27**	58.87	12.9	52.89

Table 1: Final results averaged over 5 folds of test data using different features and their combinations: NG is NGRAM, AN is AlphaNumeric, VT is verbtype

In the future we are looking at ways of adding more sophisticated deep syntactic and semantic features using lexical chains from discourse structure. We will also be exploring belief annotation in Arabic and Urdu on a parallel data collection since these languages express evidentiality in ways that differ linguistically from English. Finally we will explore ways of automatically augmenting the labeled data pool using active learning.

Acknowledgement

This work was supported by grants from the Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

References

Philip R. Cohen and Hector J. Levesque. 1990. Rational interaction as the basis for communication. In Jerry Morgan Philip Cohen and James Allen, editors, *Intentions in Communication*. MIT Press.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63:341–385.

Rebecca Bruce Matthew Bell Janyce Wiebe, Theresa Wilson and Melanie Martin. 2004. Learning subjective language. In *Computational Linguistics, Volume 30 (3)*.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn Tree-Bank. In *Proceedings of the Human Language Technology Conference*, San Diego, CA.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28–30.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. Technical Report ISI/RS-87-190, ISI.

Johanna Moore. 1994. *Participating in Explanatory Dialogues*. MIT Press.

René Witte Ralf Krestel and Sabine Bergler. 2007. Processing of Beliefs extracted from Reported Speech in Newspaper Articles. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, September 27–29.

Saurí Roser, Marc Verhagen, and James Pustejovsky. 2006. Annotating and Recognizing Event Modality in Text. In FLAIRS 2006, editor, *In Proceedings of the 19th International FLAIRS Conference*, Melbourne Beach, Florida, May 11-13.

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 801–808, Manchester, UK, August. Coling 2008 Organizing Committee.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, Michigan, June. Association for Computational Linguistics.