AFRL-RI-RS-TR-2018-134

UNCOVERING MOTIVATIONS, STANCES AND ANOMALIES THROUGH PRIVATE-STATE RECOGNITION

CORNELL UNIVERSITY

MAY 2018

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2018-134 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
PETER ROCCI
Work Unit Manager

/ S /
JON S. JONES
Technical Advisor, Information Intelligence
Systems and Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MAY 2018	FINAL TECHNICAL REPORT	OCT 2012 – NOV 2017

4. TITLE AND SUBTITLE

UNCOVERING MOTIVATIONS, STANCES AND ANOMALIES THROUGH PRIVATE-STATE RECOGNITION

5a. CONTRACT NUMBER
FA8750-13-2-0015

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Claire Cardi, Rada Michalcea

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
R0WF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Cornell University
Office of Sponsored Programs
373 Pine Tree Rd
Ithaca, NY 14850-2820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIED
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2018-134

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We report on natural language processing research on the automatic identification of *private states* (e.g. opinions, beliefs, evaluations, judgments, feelings) in unstructured text. We focus on the development and evaluation of neural network-based methods for fine-grained opinion analysis methods for inferring private states that are not explicitly mentioned, and methods for determining private-state-aware semantic equivalence.

15. SUBJECT TERMS

Natural language processing (NLP), opinion, sentiment, semantic equivalence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Peter Rocci
U	U	U	UU	27	19b. TELEPHONE NUMBER (*Include area code*)

TABLE OF CONTENTS

Section **Page**

List(s) of Figures

List(s) of Tables

1.0 SUMMARY

This report describes the research done by Cornell University, the University of Michigan, and the University of Pittsburgh under the Darpa Deft program. In particular, it describes our investigations to develop natural language processing methods to automatically recognize opinions, beliefs, and other *private states* in text, including explicitly stated private states as well as those only implied in the text.

2.0 INTRODUCTION

The goal for our DEFT project was to bring to the DEFT program and the DEFT System the ability to uncover the underlying, emerging and evolving motivations of agents --- people, organizations, groups, countries --- by recognizing and analyzing the explicit and implicit private-state information buried in natural language text. We anticipated that this would require: (1) improved recognition and tracking of "private state" information in conversation or throughout a discourse and (2) the subsequent development of private-state-aware algorithms that could identify consistent vs. inconsistent private states of protagonists across multiple documents while also accounting for the event-based semantic relations in the text.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

Throughout our work, we relied primarily on linguistically informed statistical machine learning methods. The remainder of this section lays out the key problems that we studied: fine-grained opinion analysis (Section 3.1), belief and sentiment extraction for Chinese (Section 3.2), inference of private states (Section 3.3), and private-state-aware semantic equivalence (Section 3.4). For each, we describe the specific techniques developed or methods employed. Section 4 summarizes the results of each of these investigations.

3.1 Fine-grained opinion analysis

Private states --- mental and emotional states such as speculations, opinions, evaluations, sentiments, judgments and beliefs --- are fundamental to understanding people's motivations and to explaining the reasons for their actions. They are also ubiquitous in language: they appear in written and spoken language ranging from editorials, reviews, political argument, email and social media; to face-to-face meetings and informal conversations; to purportedly objective reports and news. In the DEFT program, we developed an increasingly sophisticated set of learning-based algorithms that aimed to improve the identification of private state information in text. Specifically, we were interested in *fine-grained opinion analysis*: the ability to recognize expressions of private states at the phrase level; to determine their *polarity*; to identify the opinion *holder* or *source*, i.e. the entity expressing the private state; and to identify the *target* of the opinion --- the object or topic of the private state.

In particular, we aimed to identify fine-grained opinions expressed *either explicitly or implicitly* in a text. Consider, for example, the following sentences:

1. John was glad that Mary saved Bill.
2. The international community seems to be tolerating the Israeli campaign of suppression against the Palestinians.

Explicitly mentioned opinions include:

- John has a positive attitude ("was glad") about Mary's saving Bill.
- The international community is mildly positive ("tolerating") of the Israeli suppression.

But there are also a number of opinion implicatures, i.e., implied private states. For sentence (1), for example: John is positive toward Bill, John is positive towards Mary (for saving Bill), John believes that Mary is positive towards Bill (because she saved him). For sentence (2): the writer is negative toward the Israeli campaign of suppression, toward the International Community (for tolerating it), and toward Israel; and the writer is positive toward the Palestinians. (There are other private-state implicatures here as well.) Note that each of these inferences is, of course, defeasible given the right context: John might hate Bill (sentence 1); he might just be happy that Mary's saving Bill means he won't have to attend another funeral. But it is precisely this deeper interpretation of subjective language that is fundamental for uncovering and characterizing the complex motivations of agents based on natural language text.

The subsections below describe our Deft research on explicit and implicit fine-grained opinion extraction.

3.1.1. A Joint Inference Approach. Fine-grained opinion extraction is a challenging task due to the complexity and variety of the language used to express opinions and their components. Most existing approaches tackle the extraction of opinion entities and opinion relations in a pipelined manner, where the interdependencies among different extraction stages are not captured. In particular, sequence labeling models have been successfully employed to identify opinion expressions and relation extraction techniques have been proposed to extract opinion holders and targets based on their linking relations to the opinion expressions. However, most existing work treats the extraction of different opinion entities and opinion relations in a pipelined manner: the interaction between different extraction tasks is not modeled jointly and error propagation is not considered. One exception is the work of [1] which proposed an Integer Linear Programming (ILP) approach to jointly identify opinion holders, opinion expressions and their IS-FROM linking relations, and demonstrated the effectiveness of joint inference. Their ILP formulation, however, does not handle *implicit linking relations*, i.e. opinion expressions with no explicit opinion holder; nor does it consider IS-ABOUT relations.

In [2] and [3], we propose instead an ILP model that jointly identifies opinion-related entities, including *opinion expressions*, *opinion targets* and *opinion holders* as well as the associated opinion linking relations, IS-ABOUT and IS-FROM. For each type of opinion relation, we also allow implicit (i.e. empty) arguments for cases when the opinion holder or target is not explicitly expressed in text. We model entity identification as a sequence tagging problem and relation extraction as binary classification. A joint inference framework is proposed to jointly optimize the predictors for different subproblems with constraints that enforce global consistency.

We hypothesized that the ambiguity in the extraction results would be reduced and thus, performance increased. For example, uncertainty w.r.t. the spans of opinion entities can adversely affect the prediction of opinion relations; and evidence of opinion relations might provide clues to guide the accurate extraction of opinion entities.

The ILP Model. First, we treat the task of *opinion entity* identification (i.e. opinion expression, opinion holder/source, opinion target) as a sequence labeling problem and employ conditional random fields (CRFs) to learn the probability of a sequence assignment **y** for a given sentence **x**. Through inference we can find the best sequence assignment for sentence **x** and recover the opinion entities according to the standard "IOB" encoding scheme. For *opinion-arg relation classification* (i.e. opinion IS-FROM and opinion IS-ABOUT), we construct candidates of opinion expressions and opinion arguments and consider each pair of an opinion candidate and an argument candidate as a potential opinion relation. Conceptually, all possible subsequences in the sentence are candidates. To filter out candidates that are less reasonable, we consider the opinion expressions and arguments obtained from the n-best predictions by CRFs. We also employ syntactic patterns from dependency trees to generate candidates. The features we use aim to capture (a) local properties of the candidate opinion expressions and arguments and (b) syntactic and semantic attributes of their relation. Similarly, we develop a classifier for *opinion-implicit-arg classification*, which decides whether an opinion candidate is linked to an implicit argument, i.e. no argument is mentioned.

The inference goal is to find the optimal prediction for both opinion entity identification and opinion relation extraction. For this we employ an Integer Linear Program whose objective function is defined as a linear combination of the potentials from different predictors to balance the contribution of two components: opinion entity identification and opinion relation extraction subject to a set of task-specific constraints:

- Constraint 1: Uniqueness. Each entity span must be assigned a single type.
- Constraint 2: Non-overlapping. No two entity spans i and j may overlap.
- Constraint 3: Consistency between the opinion-arg and opinion-implicit-arg classifiers. For an opinion candidate i, if it is predicted to have an implicit argument, then no argument candidate should form a relation with i.
- Constraint 4: Consistency between the opinion-arg classifier and the opinion entity extractor. Suppose an argument candidate j in relation k is assigned an argument label by the entity extractor, then there must exist some opinion candidates that associate with j.
- Constraint 5: Consistency between the opinion-implicit-arg classifier and opinion entity extractor. When an opinion candidate i is predicted to associate with an implicit argument in relation k, then there should be no opinion holder.

Our ILP approach for joint extraction of the individual elements of an opinion frame produced state-of-the-art results for the widely used MPQA data set developed for this task. Results are described in Section 4.1.

3.1.2 Context and Sentiment. We next addressed issues in assigning sentiment values (e.g. positive, negative) to spans of text. In addition to fine-grained opinion extraction, identifying

sentiment is crucial for many opinion-mining applications such as opinion summarization, opinion question answering and opinion retrieval. Accordingly, extracting sentiment at the fine-grained level (e.g. at the sentence- or phrase-level) has received increasing attention recently due to its challenging nature and its importance in supporting these opinion analysis tasks.

In this research, we focused on the task of sentence-level sentiment classification in online reviews. Typical approaches to the task employ supervised machine learning algorithms with rich features and take into account the interactions between words to handle compositional effects such as polarity reversal. Still, these methods can encounter difficulty when the sentence on its own does not contain strong enough sentiment signals (due to the lack of statistical evidence or the requirement for background knowledge). Consider the following review for example,

1. Hearing the music in real stereo is a true revelation. 2. You can feel that the music is no longer constrained by the mono recording. 3. In fact, it is more like the players are performing on a stage in front of you ...

Existing feature-based classifiers may be effective in identifying the positive sentiment of the first sentence due to the use of the word "revelation", but they could be less effective in the last two sentences due to the lack of explicit sentiment signals. However, if we examine these sentences within the discourse context, we can see that: the second sentence expresses sentiment towards the same aspect – "the music" – as the first sentence; the third sentence expands the second sentence with the discourse connective "In fact". These discourse-level relations help indicate that sentences 2 and 3 are likely to have positive sentiment as well.

The importance of discourse for sentiment analysis has become increasingly recognized. Most existing work considers discourse relations between adjacent sentences or clauses and incorporates them as constraints or features in classifiers. Very little work has explored long-distance discourse relations for sentiment analysis. [4] defines coreference relations on opinion targets and applies them to constrain the polarity of sentences. However, the discourse relations were obtained from manual fine-grained annotations (which are not generally readily available) and implemented as hard constraints on polarity.

In response, in this part of our Deft work, we proposed in [5] a sentence-level sentiment classification method that can (1) incorporate rich discourse information at both local and global levels; (2) encode discourse knowledge as soft constraints during learning; (3) make use of unlabeled data to enhance learning. Specifically, we used a standard Conditional Random Field (CRF) model as the learner for sentence-level sentiment classification, and incorporated rich discourse and lexical knowledge as soft constraints into the learning of CRF parameters via Posterior Regularization (PR) [6]. As a framework for structured learning with constraints, PR has been successfully applied to many structural NLP tasks, but our work was the first to explore PR for sentiment analysis. Unlike most previous work, we explored a rich set of structural constraints that cannot be naturally encoded in the feature-label form, and showed that such constraints can improve the performance of the CRF model.

Table 1. Summarization of posterior constraints for sentence-level sentiment classification

Types	Description and Examples	Inter-sentential
Lexical patterns	The sentence containing a polar lexical pattern w tends to have the polarity indicated by w. Example lexical patterns are *annoying, hate, amazing, not disappointed, no concerns, favorite, recommend.*	
Discourse Connectives (clause)	The sentence containing a discourse connective c which connects its two clauses that have opposite polarities indicated by the lexical patterns tends to have neutral sentiment. Example connectives are *while, although, though, but.*	
Discourse Connectives (sentence)	Two adjacent sentences which are connected by a discourse connective c tends to have the same polarity if c indicates a *Expansion* or *Contingency* relation, e.g. *also, for example, in fact, because* ; opposite polarities if c indicates a *Comparison* relation, e.g. *otherwise, nevertheless, however.*	✓
Coreference	The sentences which contain coreferential entities appeared as targets of opinion expressions tend to have the same polarity.	✓
Listing patterns	A series of sentences connected via a listing tend to have the same polarity.	✓
Global labels	The sentence-level polarity tends to be consistent with the document-level polarity.	✓

Table 1 above summarizes the various posterior constraints employed. See Section 4.1 for a description of the results.

3.1.3 Neural Network Methods. The bulk of our research on the extraction of explicitly mentioned fine-grained opinions involved the development of neural network techniques. Our Deft research in this area was among the very first such work in the field.

The	committee	,	as	usual	,	has
O	O	O	B_ESE	I_ESE	O	B_DSE
refused	to	make	any	statements	.	
I_DSE	I_DSE	I_DSE	I_DSE	I_DSE	O	

Figure 1. Example sentence with labels of DSEs and ESEs

In one line of work [8], we focused on the use of recurrent neural networks for the detection of opinion expressions — both *direct subjective expressions* (DSEs) and *expressive subjective expressions* (ESEs) as defined in [7]. DSEs consist of explicit mentions of private states or speech events expressing private states; and ESEs consist of expressions that indicate sentiment, emotion, etc., without explicitly conveying them. An example sentence shown in Figure 1 in which the DSE "has refused to make any statements" explicitly expresses an opinion holder's attitude and the ESE "as usual" indirectly expresses the attitude of the writer.

Opinion extraction has often been tackled as a sequence labeling problem in previous work (including our own). This approach views a sentence as a sequence of tokens labeled using the conventional BIO tagging scheme: B indicates the beginning of an opinion-related expression, I is used for tokens inside the opinion-related expression, and O indicates tokens outside any opinion-related class. The example sentence in Table 1 shows the appropriate tags in the BIO scheme. For instance, the ESE "as usual" results in the tags B ESE for "as" and I ESE for "usual".

Variants of **conditional random field (CRF)** approaches have been successfully applied to opinion expression extraction using this token-based view: the state-of-the-art approach at the

time of this work was the semiCRF, which relaxes the Markovian assumption inherent to CRFs and operates at the phrase level rather than the token level, allowing the incorporation of phrase-level features [9]. The success of the CRF- and semiCRF-based approaches, however, hinges critically on access to an appropriate feature set, typically based on constituent and dependency parse trees, manually crafted opinion lexicons, named entity taggers and other preprocessing components.

Distributed representation learners provide a different approach to learning in which latent features are modeled as distributed dense vectors of hidden layers. A **recurrent neural network (RNN)** is one such learner that can operate on sequential data of variable length, which means it can also be applied as a sequence labeler. Moreover, **bidirectional RNNs** incorporate information from preceding as well as following tokens while advances in word embedding induction had enabled more effective training of RNNs by allowing a lower dimensional dense input representation and hence, more compact networks. Finally, deep recurrent networks, a type of RNN with multiple stacked hidden layers, had been shown to naturally employ a temporal hierarchy with multiple layers operating at different time scales: lower levels capture short term interactions among words; higher layers reflect interpretations aggregated over longer spans of text. When applied to natural language sentences, such hierarchies have been suggested to better model the multi-scale language effects that are emblematic of natural languages, as suggested by previous results.

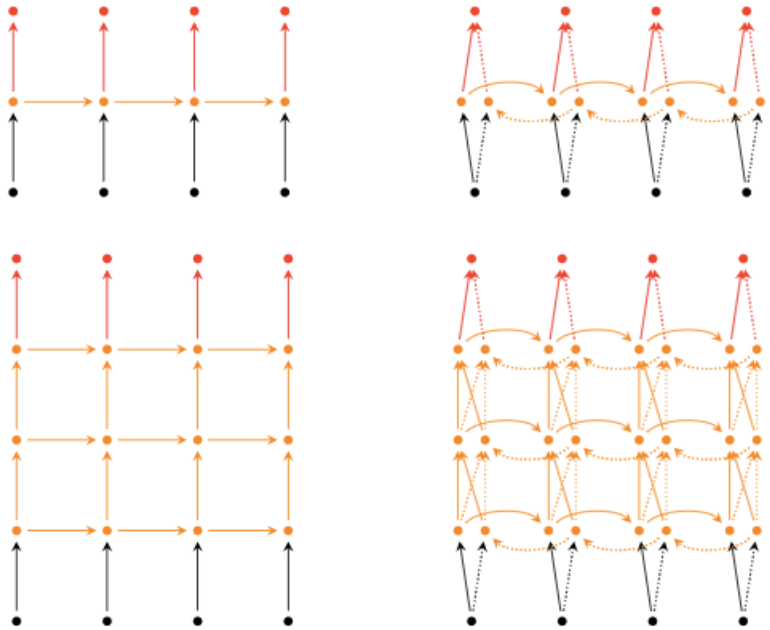


Figure 2. Variations of recurrent neural networks

Figure 2 shows different varieties of recurrent neural networks. Each black, orange and red node denotes an input, hidden or output layer, respectively. Solid and dotted lines denote the connections of forward and backward layers, respectively. The top half of the figure depicts shallow unidirectional (on the left) and bidirectional (on the right) RNN. At the bottom, we show a 3-layer deep unidirectional (left) and bidirectional (right) RNN.

In general, we expected that the deep RNNs would show the most improvement over shallow RNNS for ESEs — phrases that implicitly convey subjectivity. Existing research has shown that these are harder to identify than direct expressions of subjectivity (DSEs): they are variable in length and involve terms that, in many (or most) contexts, are neutral with respect to sentiment and subjectivity. As a result, models that do a better job interpreting the context should be better at disambiguating subjective vs. non-subjective uses of phrases involving common words (e.g. "as usual", "in fact"). Whether or not deep RNNs would be powerful enough to outperform the state-of-the-art semiCRF was unclear, especially if the semiCRF is given access to the distributed word representations (embeddings) employed by the deep RNNs. In addition, the semiCRF has access to parse tree information and opinion lexicons, neither of which is available to the deep RNNs.

Results and findings are presented in Section 4.1.

In a similar line of work, we explored **deep recursive neural networks** for fine-grained opinion extraction [10]. Recursive neural networks comprise a class of architecture that can operate on structured input. They had been previously successfully applied to model compositionality in natural language using parse-tree-based structural representations. Even though these architectures are deep in structure, they lacked the capacity for hierarchical representation that exists in conventional deep feed-forward networks as well as in previously investigated deep recurrent neural networks.

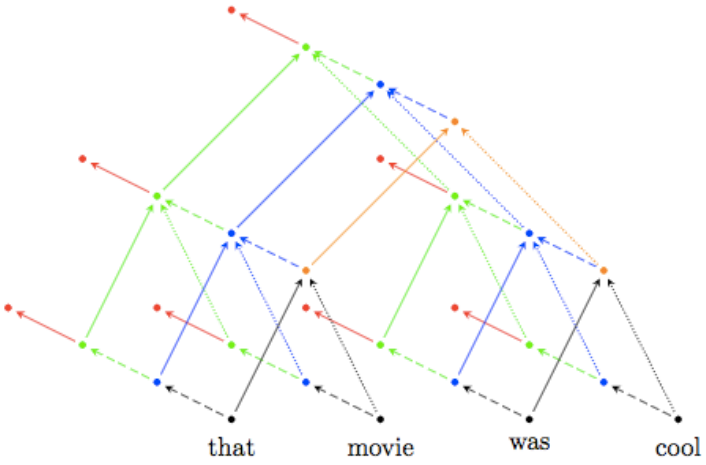


Figure 3. Operation of a 3-layer deep recursive neural network (Red and black points denote output and input vectors, respectively; other colors denote intermediate memory representations. Connections denoted by the same color-style combination are shared (i.e. share the same set of weights).)

In [10], we introduced a new architecture — a **deep recursive neural network (deep RecursNN)** — constructed by stacking multiple recursive layers. In particular, the a deep RecursNN is essentially a deep feedforward neural network with an additional structural

processing within each layer (see Figure 3). During forward propagation, information travels through the structure within each layer (because of the recursive nature of the network, weights regarding structural processing are shared). In addition, every node in the structure (i.e. in the parse tree) feeds its own hidden state to its counterpart in the next layer. This can be seen as a combination of feedforward and recursive nets. In a shallow recursive neural network, a single layer is responsible for learning a representation of composition that is both useful and sufficient for the final decision. In a deep recursive neural network, a layer can learn some parts of the composition to apply, and pass this intermediate representation to the next layer for further processing for the remaining parts of the overall composition.

We evaluated the proposed model on the task of fine-grained sentiment classification and showed that deep RecursNNs outperformed their associated shallow counterparts that employed the same number of parameters. Furthermore, our approach outperformed previous baselines on the sentiment analysis task, achieving new state-of-the-art results. See Section 4.1 for more information on the results.

In a third research effort [11], we developed and investigated the **multiplicative recurrent neural network** as a general model for compositional meaning in language, and evaluated it on the same task of fine-grained sentiment analysis. We found that these models performed comparably or better than standard additive recurrent neural networks and outperformed matrix-space models on the MPQA corpus. They yielded comparable results to structural deep models (like our deep recursive networks) on the Stanford Sentiment Treebank (without the need for generating parse trees as was required by earlier approaches).

All of the neural network methods above are described in detail in the PhD dissertation of Ozan Irsoy [12].

In follow-on work [13], we investigated the use of potentially more powerful neural architectures for opinion extraction. In particular, we looked at **deep bi-directional LSTMs** (long short-term memory recurrent networks). Our goal was to use LSTMs to jointly extract opinion entities (sources, targets, opinion expressions) as well as the IS-FROM and IS-ABOUT relations that connect them. This was the first such attempt at this task using a deep learning approach.

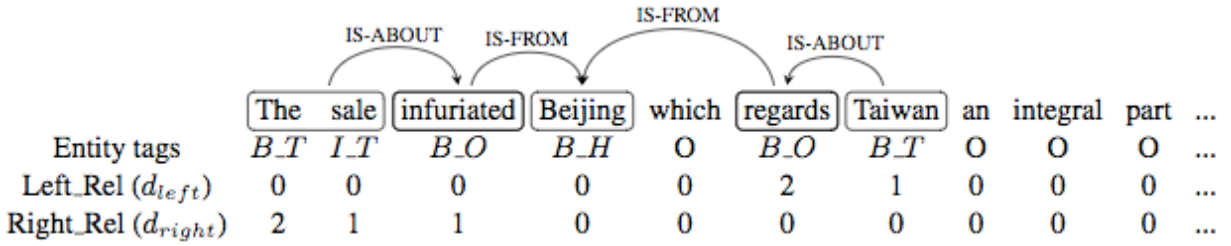


Figure 4. Gold standard annotation for an example sentence from MPQA dataset
(O represents the 'Other' tag in the BIO scheme.)

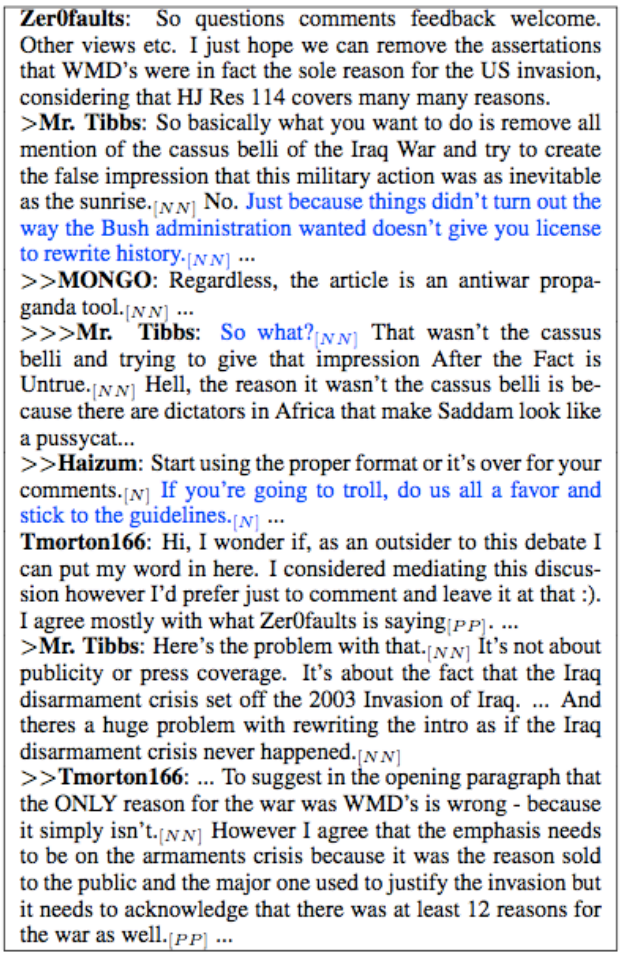
To handle relation identification within the LSTM's sequence-tagging framework, we encoded relation information as part of the token-level tagging scheme as shown in Figure 4, which provides an example of an annotated sentence from the dataset: boxes denote opinion entities and opinion relations are shown by arcs. We interpret these relations arcs as directed — from an opinion expression towards an opinion holder, and from an opinion target towards an opinion expression. We pre-process these relation arcs to obtain the left-relation distances (d left) and right-relation distances (d right) as shown. For each word in an entity, we find its distance to the nearest word in the related entity. These distances become our relation tags. The entity tags are interpreted using the BIO scheme, also shown in the figure. Our RLL model jointly models the entity tags and relation tags. At inference time, these entity tags and relation tags are used together to determine IS-FROM and IS-ABOUT relations.

For opinion entity extract we found, perhaps surprisingly, that standard LSTMs were not competitive with the existing state-of-the-art CRF+ILP joint inference approach [3], performing below even the standalone sequence-tagging CRF. Incorporating sentence-level and a novel relation-level optimization, however, allows the LSTM to identify opinion relations and to perform within 1–3% of the state-of-the-art joint model for opinion entities and the IS-FROM relation; and to perform as well as the state-of-the-art for the ISe-ABOUT relation — all without access to opinion lexicons, parsers and other preprocessing components required for the feature-rich CRF+ILP approach.

In the standard token-level log-likelihood computation of LSTMs, dependencies between the tags in a tag sequence are discarded. We can improve upon this using a modified, **sentence-level log-likelihood (SLL)** formulation (adapted from Collobert's) to incorporate these dependencies. In particular, we introduce a transition score $[A]_{i,j}$ for jumping from tag *i* to tag *j* of adjacent words in the tag sequence and learn the transition scores during training. This sentence-level log-likelihood improves performance, but it cannot be directly used for modeling relations between non-adjacent words in the sentence. Only when we extended it to also model relations between non-adjacent words does the LSTM obtain results within 1-3% of the (existing state-of-the-art) ILP-based joint model on opinion entities, within 3% for the IS-FROM relation, and comparably for IS-ABOUT relation.

We next improved upon the LSTM-based approach in the work reported in [14], where we proposed a cleaner, more efficient LSTM-based model. In particular, the LSTM-based formulation of [13] explicitly encoded the distance between the head of entities and the associated opinion relation labels. The output space of the model was quadratic in size of the entity and relation label set; in addition, that work did not specifically identify the relation type. Unfortunately, adding the relation type would make the output label space very sparse, making it difficult for the model to learn. In the new work [14], we presented an **attention-based recurrent neural network** for joint extraction of entity mentions and relations. (In this work, we applied the techniques to general entity and relation extraction --- as in the ACE evaluations --- rather than focusing on opinion entity and opinion relation extraction.) The attention framework replaced the distance-based encoding of relations used in [13]. Results for the investigation are described in Section 4.1.

3.1.4. Opinion analysis applications. In conjunction with the above research on fine-grained opinion analysis, we also investigated its application in a number of applications.



Zer0faults: So questions comments feedback welcome. Other views etc. I just hope we can remove the assertations that WMD's were in fact the sole reason for the US invasion, considering that HJ Res 114 covers many many reasons.
>**Mr. Tibbs**: So basically what you want to do is remove all mention of the cassus belli of the Iraq War and try to create the false impression that this military action was as inevitable as the sunrise.[NN] No. Just because things didn't turn out the way the Bush administration wanted doesn't give you license to rewrite history.[NN] ...
>>**MONGO**: Regardless, the article is an antiwar propaganda tool.[NN] ...
>>>**Mr. Tibbs**: So what?[NN] That wasn't the cassus belli and trying to give that impression After the Fact is Untrue.[NN] Hell, the reason it wasn't the cassus belli is because there are dictators in Africa that make Saddam look like a pussycat...
>>**Haizum**: Start using the proper format or it's over for your comments.[N] If you're going to troll, do us all a favor and stick to the guidelines.[N] ...
Tmorton166: Hi, I wonder if, as an outsider to this debate I can put my word in here. I considered mediating this discussion however I'd prefer just to comment and leave it at that :). I agree mostly with what Zer0faults is saying[PP]. ...
>**Mr. Tibbs**: Here's the problem with that.[NN] It's not about publicity or press coverage. It's about the fact that the Iraq disarmament crisis set off the 2003 Invasion of Iraq. ... And theres a huge problem with rewriting the intro as if the Iraq disarmament crisis never happened.[NN]
>>**Tmorton166**: ... To suggest in the opening paragraph that the ONLY reason for the war was WMD's is wrong - because it simply isn't.[NN] However I agree that the emphasis needs to be on the armaments crisis because it was the reason sold to the public and the major one used to justify the invasion but it needs to acknowledge that there was at least 12 reasons for the war as well.[PP] ...

Figure 5. **Example discussion from the Wikipedia talk page for article "Iraq War"** where editors discuss about the correctness of the information in the opening paragraph. (We only show some sentences that are relevant for demonstration. Other sentences are omitted by ellipsis. Names of editors are in bold. ">" is an indicator for the reply structure, where turns starting with > are response for most previous turn that with one less >. We use "NN", "N", and "PP" to indicate "strongly disagree", "disagree", and "strongly agree". Sentences in blue are examples whose sentiment is hard to detect by an existing lexicon.)

In [15], we studied the problem of agreement and disagreement detection in online discussions. Sentence-level agreement and disagreement detection for this domain is challenging in its own right due to the dynamic nature of online conversations, and the less formal, and usually very emotional language used. As an example, consider a snippet of discussion from the Wikipedia Talk page for article "Iraq War" where editors argue on the correctness of the information in the opening paragraph (Figure 5). "So what?" should presumably be tagged as a negative sentence as should the sentence "If you're going to troll, do us all a favor and stick to the guidelines." We hypothesized that these, and other, examples will be difficult for the tagger unless the context

surrounding each sentence is considered and in the absence of a sentiment lexicon tuned for conversational text.

As a result, we investigated isotonic Conditional Random Fields (isotonic CRFs) as a sequential model for making predictions at the sentence- or segment-level. The method relied on an automatically constructed lexicon socially-tuned for the online discussion format that was bootstrapped from existing general-purpose sentiment lexicons to improve performance.

In related work [16], we investigated the task of online dispute detection (employing our isotonic CRF-based approach of [15] described just above) and proposed a sentiment analysis solution to the problem. In particular, we aimed to identify the sequence of sentence-level sentiments expressed during a discussion and to use them as features in a classifier that predicts the DISPUTE/NON-DISPUTE label for the discussion as a whole.

Results for the dispute detection work are described in Section 4.1 and in more detail in [17].

3.2 Belief and Sentiment Extraction for Chinese

All of the above research on fine-grained opinion extraction (Section 3.1) was undertaken for English. We also developed methods for sentiment (i.e. opinion) extraction for Chinese through our participation in the Text Analysis Conference (TAC) at NIST in 2016 [18] and 2017 [19]. (We also participated in the English evaluation in 2016 [18] and in the Coldstart KB evaluation in 2017 [20].) In both years, the tasks involved the identification of *beliefs* as well as sentiment and were evaluated on both newswire and discussion forum data. Our approach to the evaluation was to investigate learning-based and heuristic knowledge-based methods, both of which evolved directly from our existing work on opinion analysis in English. Heuristic-based methods were included to handle some of the idiosyncracies of the data sets.

3.3 Private-state-aware Inference

We developed a computational framework for fine-grained opinion analysis that operates across a paragraph or document (rather than at the sentence level) and captures explicitly stated opinions and sentiment as well as attitudes that can be inferred from actions and events. The system integrated the opinion analysis components (for English) referenced above and employed probabilistic soft logic models for the task that could recognize and resolve conflicting vs. reinforcing attitudes among entities. In addition, it operated at the entity- rather than the mention-level --- to reflect the shift in focus of the Deft program towards a knowledge-based view for NLP. A more in-depth description is provided next.

Our research on fine-grained opinion analysis described in the sections above is span-based in the sense that opinions are identified via spans of subjective text and associated with sources and targets that are **mentions** of an entity in the text. In contrast, our EMNLP 2015 research [21] contributed to **entity/event-level sentiment analysis**. In addition, it brought together a collection of our Deft research on the problem of **opinion inference** [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. Let us consider an example from the MPQA opinion annotated corpus.

Ex(1) When the Imam (may God be satisfied with him 1) issued the fatwa against 2 Salman Rushdie for insulting 3 the Prophet (peace be upon him 4), the countries that are so-called 5 supporters of human rights protested against 6 the fatwa.

Figure 6. Example from the MPQA corpus

There are several sentiment expressions annotated in the sample sentence from the MPQA corpus shown in Figure 6. In the first clause, the writer is positive toward Imam and Prophet as expressed by *may God be satisfied with him* (1) and *peace be upon him* (4), respectively. Imam is negative toward Salman Rushdie and the insulting event, as revealed by the expression *issued the fatwa against* (2). And Salman Rushdie is negative toward Prophet, as revealed by the expression *insulting* (3). In the second clause, the writer is negative toward the countries, as expressed by *so-called* (5). And the countries are negative toward fatwa, as revealed by the expression *protested against* (6). Using the source and the target, we summarize the positive opinions above in a set P, and the negative opinions above in another set N. Thus, P contains {(writer, Imam), (writer, Prophet)}, and N contains {(Imam, Rushdie), (Imam, insulting), (Rushdie, Prophet), (writer, countries), (countries, fatwa)}.

An (ideal) explicit sentiment analysis system is expected to extract the above sentiments expressed by (1)-(6). However, there are many more sentiments communicated by the writer but not expressed via explicit expressions. First, Imam is positive toward the Prophet, because Rushdie insults the Prophet and Imam is angry that he does so. Second, the writer is negative toward Rushdie, because the writer is positive toward the Prophet but Rushdie insults him! Also, the writer is probably positive toward the fatwa since it is against Rushdie. Third, the countries are probably negative toward Imam, because the countries are negative toward fatwa and it is Imam who issued the fatwa. Thus, the set P should also contain {(Imam, Prophet), (writer, fatwa)}, and the set N should also contain {(writer, Rushdie), (countries, Imam)}. These opinions are not directly

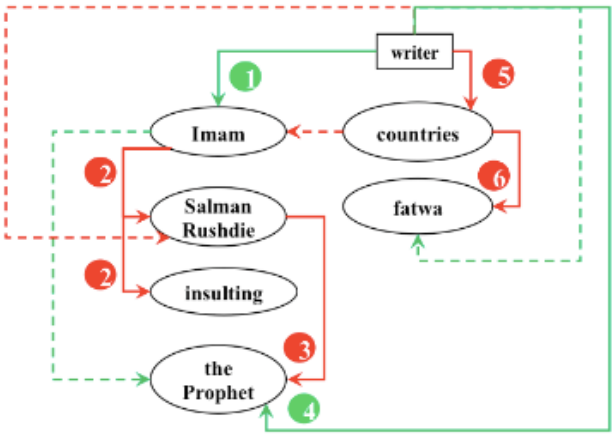


Figure 7: Explicit and implicit sentiments in Ex(1)

expressed, but are **inferred** by a human reader. Note that the inferences are **conversational implicatures**: they are defeasible and may not go through in context.

The explicit and implicit sentiments are summarized in Figure 7, where each green line represents a positive sentiment and each red line represents a negative sentiment. The solid lines are explicit sentiments and the dashed lines are implicit sentiments.

In this work, we aimed to detect sentiments such as those in P and N, where the sources are entities (or the writer) and the targets are entities and events. We proposed a general set of inference rules than in our previous work on this topic.

With the rules in hand, we encoded them in a probabilistic soft logic (PSL) framework. We chose PSL [35] because it is designed to have efficient inference and, as similar methods in Statistical Relational Learning do, it allows probabilistic models to be specified in first-order logic, an expressive and natural way to represent if-then rules, and it supports joint prediction. Joint prediction is critical for our task because it involves multiple, mutually constraining ambiguities (the source, polarity, and target). Thus, this work aimed at detecting both implicit and explicit sentiments expressed by an entity to ward another entity/event (i.e. an eTarget) within the sentence.

Results are briefly described in Section 4.1.

3.4 Private-state-aware Semantic Equivalence

As part of the *SEM evaluation we built a machine learning framework that leveraged our previous work on corpus and knowledge-based methods of text similarity, while also generating and incorporating opinion aware features [36]. This system is able to model not only semantic similarity across texts, but also the similarity of the opinions expressed in these texts.

4.0 RESULTS and DISCUSSION

In this section we describe the results of the investigations described in Section 3.0.

4.1 Fine-grained Opinion Analysis

4.1.1 A Joint Approach. We evaluated the Integer Linear Programming approach proposed in [2] and described in **section 3.1.1** using a standard corpus for fine-grained opinion analysis (the MPQA corpus [7]) and demonstrated that our model outperforms by a significant margin traditional baselines that do not employ joint inference for extracting opinion entities and different types of opinion relations.

More specifically, we compared our approach to several pipeline baselines. Each extracts opinion entities first using the same CRF employed in our approach, and then predicts opinion relations on the opinion entity candidates obtained from the CRF prediction. Three relation extraction techniques were used in the baselines:

- **Adj**: This baseline system links each argument candidate to its nearest opinion candidate. Arguments that do not link to any opinion candidate are discarded. This is also used as a strong baseline in [1].
- **Syn**: Links pairs of opinion and argument candidates that present prominent syntactic patterns.
- **RE**: Predicts opinion relations by employing the opinion-arg classifier and opinion-implicit-arg classifier. First, the opinion-arg classifier identifies pairs of opinion and argument candidates that form valid opinion relations, and then the opinion-implicit-arg classifier is used on the remaining opinion candidates to further identify opinion expressions without explicit arguments.

We report results using opinion entity candidates from the best CRF output and from the merged 10-best CRF output. The motivation of merging the 10-best output is to increase recall for the pipeline methods.

Table 2. Performance on opinion entity extraction using overlap and exact matching metrics (The top table uses overlap and the bottom table uses exact. Two-tailed t-test results are shown on F1 measure for our method compared to the other baselines (statistical significance is indicated with * ($p < 0.05$), ** ($p < 0.005$)).)

	Opinion Expression			Opinion Target			Opinion Holder		
Method	P	R	F1	P	R	F1	P	R	F1
CRF	82.21	66.15	73.31	73.22	48.58	58.41	72.32	49.09	58.48
CRF+Adj	82.21	66.15	73.31	80.87	42.31	55.56	75.24	48.48	58.97
CRF+Syn	82.21	66.15	73.31	81.87	30.36	44.29	78.97	40.20	53.28
CRF+RE	83.02	48.99	61.62	85.07	22.01	34.97	78.13	40.40	53.26
Joint-Model	71.16	77.85	**74.35***	75.18	57.12	**64.92****	67.01	66.46	**66.73****
CRF	66.60	52.57	58.76	44.44	29.60	35.54	65.18	44.24	52.71
CRF+Adj	66.60	52.57	58.76	49.10	25.81	33.83	68.03	43.84	53.32
CRF+Syn	66.60	52.57	58.76	50.26	18.41	26.94	74.60	37.98	50.33
CRF+RE	69.27	40.09	50.79	60.45	15.37	24.51	75	38.79	51.13
Joint-Model	57.39	62.40	**59.79***	49.15	38.33	**43.07****	62.73	62.22	**62.47****

Table 2 shows the results of opinion entity identification using both overlap and exact metrics. We compare our approach with the pipeline baselines and CRF (the first step of the pipeline). We can see that our joint inference approach significantly outperforms all the baselines in F1 measure on extracting all types of opinion entities. In general, by adding the relation extraction step, the pipeline baselines are able to improve precision over the CRF but fail at recall. CRF+Syn and CRF+Adj provide the same performance as CRF, since the relation extraction step only affects the results of opinion arguments. By incorporating syntactic information, CRF+Syn provides better precision than CRF+Adj on extracting arguments at the expense of recall. This indicates that using simple syntactic rules would mistakenly filter many correct relations. By using binary classifiers to predict relations, CRF+RE produces high precision on opinion and target extraction but also results in very low recall. Using the exact metric, we observe the same general trend in the results as the overlap metric. The scores are lower since the metric is much stricter.

Table 3. Performance on opinion relation extraction using the overlap metric

	IS-ABOUT			IS-FROM		
Method	**P**	**R**	**F1**	**P**	**R**	**F1**
CRF+Adj	73.65	37.34	49.55	70.22	41.58	52.23
CRF+Syn	76.21	28.28	41.25	77.48	36.63	49.74
CRF+RE	78.26	20.33	32.28	74.81	37.55	50.00
CRF+Adj-merged-10-best	25.05	61.18	35.55	30.28	62.82	40.87
CRF+Syn-merged-10-best	41.60	45.66	43.53	48.08	54.03	50.88
CRF+RE-merged-10-best	51.60	33.09	40.32	47.73	54.40	50.84
Joint-Model	64.38	51.20	**57.04****	64.97	58.61	**61.63****

Table 3 shows the results of opinion relation extraction using the overlap metric. We compare our approach with pipelined baselines in two settings: one employs relation extraction on 1-best output of CRF (top half of table) and the other employs the merged 10-best output of CRF (bottom half of table). We can see that in general, using merged 10-best CRF outputs boosts the recall while sacrificing precision. This is expected since merging the 10-best CRF outputs favors candidates that are believed to be more accurate by the CRF predictor. If CRF makes mistakes, the mistakes will propagate to the relation extraction step. The poor performance on precision further confirms the error propagation problem in the pipeline approaches. In contrast, our joint-inference method successfully boosts the recall while maintaining reasonable precision. This demonstrates that joint inference can effectively leverage the advantage of individual predictors and limit error propagation.

4.1.2 Context and Sentiment. We evaluated the context-aware CRF-based learning method for sentiment analysis described in Section 3.1.2 using two standard product review datasets. Experimental results showed that our model outperformed state-of-the-art methods in both the supervised and semi-supervised settings. We also found that discourse knowledge is highly useful for improving sentence-level sentiment classification.

4.1.3 Neural Network Methods. Overall, our work on neural network methods for opinion extraction were impressive in that they generally achieved comparable performance to non-neural techniques (that had been carefully developed over a decade or so) **without** the need for parsers or sentiment lexicon development and without extensive feature engineering.

Our research in [8] explored an application of deep bidirectional RNNs to the task of opinion expression extraction. (See Section 3.1.3.). Our experiments using the proposed architectures to identify direct subjective expressions (DSEs) and expressive subjective expressions (ESEs) using the standard MPQA corpus first confirmed that the (shallow) bidirectional RNN outperformed a (shallow) unidirectional RNN for both DSE and ESE recognition. When adding depth to the bidirectional RNN, we further found, for both DSE and ESE detection, that 3-layer RNNs provided the best results on our data set for large hidden layers and that 2, 3 and 4-layer RNNs showed equally good performance for smaller networks. Adding additional layers degraded performance.

Table 4. Comparison of Deep RNNs to state-of-the-art (semi)CRF baselines for DSE and ESE detection

	Model		Precision		Recall		F1	
			Prop.	Bin.	Prop.	Bin.	Prop	Bin.
DSE	CRF		74.96*	82.28*	46.98	52.99	57.74	64.45
	semiCRF		61.67	69.41	**67.22***	73.08*	64.27	71.15*
	CRF	+vec	**74.97***	**82.43***	49.47	55.67	59.59	66.44
	semiCRF	+vec	66.00	71.98	60.96	68.13	63.30	69.91
	Deep RNN	3 100	65.56	69.12	66.73*	**74.69***	**66.01***	**71.72***
ESE	CRF		56.08	68.36	42.26	51.84	48.10	58.85
	semiCRF		45.64	69.06	58.05	64.15	50.95	66.37*
	CRF	+vec	**57.15***	69.84*	44.67	54.38	50.01	61.01
	semiCRF	+vec	53.76	**70.82***	52.72	61.59	53.10	65.73
	Deep RNN	3 100	52.04	60.50	**61.71***	**76.02***	**56.26***	**67.18***

In comparison to state-of-the-art (at the time) baselines (see Table 4), we showed that our deep bi-directional RNNs outperformed previous CRF-based state-of-the-art baselines, including the semiCRF model.

As described above in Section 3.1.3, we also developed (in [10]) **deep recursive neural network** models (deep RecursNN). To evaluate their performance, we applied deep recursive neural networks to the task of fine-grained sentiment detection on the Stanford Sentiment Treebank (SST). SST includes a supervised sentiment label for every node in the binary parse tree, not just at the root (sentence) level. This is especially important for deep learning, since it allows a richer supervised error signal to be backpropagated across the network, potentially alleviating vanishing gradients associated with deep neural networks.

We found that our deep recursive neural networks outperform shallow recursive nets of the same size in the fine-grained sentiment prediction task on the Stanford Sentiment Treebank. Furthermore, our models outperformed multiplicative recursive neural network variants, achieving new state-of-the-art performance on the task. We also conducted qualitative experiments that suggested that each layer handles a different aspect of compositionality, and representations at each layer capture different notions of similarity.

In [13] and [14], we tackled the **joint extraction of entities and relations**. Our experiments on both the MPQA opinion corpus [13] and the Automatic Content Extraction (ACE) entity+relation corpus [14] showed that the LSTM-based approaches performed comparably to the state-of-the-art non-neural methods without access to parsers and lexicons. When incorporating **attention mechanisms** [14], experiments on ACE showed that our model significantly outperformed feature-based joint models and performed within 1% on entity mentions and 2% on relations when compared to the state-of-the-art neural approach (of Miwa and Bansal) that relied on dependency parse trees.

4.1.4 Opinion analysis applications. In Section 3.1.4, we described our work on opinion analysis applications that involve noisy, informal text. In [15], we developed an isotonic CRF-

based model for agreement and disagreement detection in online discussions. We evaluated our agreement and disagreement tagging model on two disparate online discussion corpora – Wikipedia Talk pages and online debates and showed that the model outperformed the state-of-the-art approaches in both datasets. For example, our isotonic CRF model achieved F1 scores of 0.74 and 0.67 for agreement and disagreement detection in comparison to a linear chain CRF that obtained 0.58 and 0.56 for the discussions on Wikipedia Talk pages.

In [16], we looked specifically at the use of sentiment analysis for dispute detection. We evaluated dispute detection approaches on a corpus that we created consisting of Wikipedia Talk page disputes. We found that classifiers that rely on our sentiment tagging features outperform those that do not. The best model achieves a very promising F1 score of 0.78 and an accuracy of 0.80.

4.2 Belief and Sentiment Extraction for Chinese

Our systems obtained top scores for Chinese BeSt (Belief and Sentiment) at TAC 2016 and TAC 2017. Results are described in our TAC 2016 [18] and TAC 2017 reports [19]. Our Chinese BeSt system was also employed in the TAC 2017 Cold Start KB task as part of the Tinkerbell team [20]. Through these evaluations we found that a combination of heuristics and learning-based methods worked best due largely to data set creation idiosyncracies.

4.3 Private-state-aware Inference

Ours was the first system to perform sentiment/opinion analysis that includes inferred as well as explicit sentiment. In an evaluation on the MPQA corpus, we found that fine-grained opinion analysis improves if inferred attitudes are taken into account.

4.4 Private-state-aware Semantic Equivalence

Our system [36] was the first to incorporate private-state originating signals to expand the traditional approach to textual similarity that only looks at the semantic layer. We proposed that true text similarity also entails the similarity of the opinions expressed in those texts. Through the evaluations performed as part of the *SEM2013 Textual Similarity Task, we learned that similarity should be modeled in a pipelined framework, where only if texts meet the more basic semantic similarity requirement, they should also be represented through their private-state content.

5 CONCLUSIONS

We succeeded in improving the recognition and, to some extent, the tracking of explicit and implicit private state information: in one line of work, we substantially improved the state-of-the-art in **fine-grained opinion analysis** within and across sentences; in another, we extended our methods for **belief and sentiment extraction for Chinese**; and finally, we created and implemented the first unified framework for **private state inference** in which the goal is to infer private state information that follows from actions or events. We developed a **private-state-aware semantic equivalence** detection system.

Not described above is our participation and organization of community-wide evaluation efforts relevant to the Deft program. We organized **SemEval-2014** Task 10: Multilingual semantic textual similarity [37], **SemEval-2015** Task 2: Semantic textual similarity (English, Spanish and pilot on interpretability) [38], **SemEval-2016** Task 1: Semantic Textual Similarity - Monolingual and Cross-lingual Evaluation [39], and **TAC 2014** Sentiment Knowledge Base Population [40]. We were part of the organizing team for **TAC 2016** Belief/Sentiment Knowledge Base Population [41] and **TAC 2017** Source-and-Target Belief and Sentiment Evaluation [42].

We participated in the **SemEval-2014 Cross-level Similarity** task [43], the **TAC 2014 Sentiment Knowledge Base Population** task [44], the **TAC 2016 Belief/Sentiment Knowledge Base Population** task [18], the **TAC 2017 Source-and-Target Belief and Sentiment Evaluation** task for Chinese [19], and the **TAC 2017 Cold Start KB** (as part of the Tinkerbell system) [20].

6.0 REFERENCES

[1] Yejin Choi, Eric Breck and Claire Cardie. Joint Extraction of Entities and Relations for Opinion Recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP*), 2006.

[2] Bishan Yang. Extracting Opinions and Events from Text: Joint Inference Approaches. *Ph.D. Thesis*, Cornell University, February 2016.

[3] Bishan Yang and Claire Cardie. Joint Inference for Fine-grained Opinion Extraction. *Annual Meeting of the Association for Computational Linguistics* (ACL), 2013.

[4] Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. Discourse Level Opinion Interpretation. *The 22nd International Conference on Computational Linguistics (COLING-2008)*, 2008.

[5] Bishan Yang and Claire Cardie. Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (ACL), 2014.

[6] Kuzman Ganchev, Joao Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *The Journal of Machine Learning Research*, 99:2001–2049.

[7] J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

[8] Ozan Irsoy and Claire Cardie. Opinion Mining with Deep Recurrent Neural Networks. In *Proceedings of EMNLP* 2014.

[9] Bishan Yang and Claire Cardie. 2012. Extracting opinion expressions with semi-markov conditional random fields. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

[10] Ozan Irsoy and Claire Cardie. Deep Recursive Neural Networks for Compositionality in Language. In *Proceedings of NIPS* 2014.

[11] Ozan Irsoy and Claire Cardie. Modeling Compositionality with Multiplicative Recurrent Neural Networks. *Proceedings of ICLR* 2015.

[12] Ozan Irsoy. Deep sequential and structural neural models of compositionality. *PhD Thesis*, Cornell University, 2017.

[13] Arzoo Katiyar and Claire Cardie. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL), 2016.

[14] Arzoo Katiyar and Claire Cardie. Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (ACL), 2017.

[15] Lu Wang and Claire Cardie. Improving Agreement and Disagreement Identification in Online Discussions with A Socially-Tuned Sentiment Lexicon. *Proceedings of the ACL Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (WASSA), 2014.

[16] Lu Wang and Claire Cardie. A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (ACL), short paper, 2014.

[17] Lu Wang. Summarization and Sentiment Analysis for Understanding Socially-Generated Content. *Ph.D. Thesis*, Cornell University, February 2016.

[18] Vlad Niculae, Kai Sun, Xilun Chen, Yao Cheng, Xinya Du, Esin Durmus, Arzoo Katiyar and Claire Cardie. Cornell Belief and Sentiment System at TAC 2016.
Proceedings of the 2016 Text Analysis Conference, TAC 2016.

[19] Kai Sun and Claire Cardie. Cornell Belief and Sentiment System at TAC 2017.
Proceedings of the 2017 Text Analysis Conference, TAC 2017.

[20] Mohamed Al-Badrashiny1, Jason Bolton, Arun Tejavsi Chaganty, Kevin Clark, Craig Harman, Lifu Huang, Matthew Lamm, Jinhao Lei, Di Lu, Xiaoman Pan, Ashwin Paranjape, Ellie Pavlick, Haoruo Peng, Peng Qi, Pushpendre Rastogi, Abigail See, Kai Sun, Max Thomas, Chen-Tse Tsai, Hao Wu, Boliang Zhang, Chris Callison-Burch, Claire Cardie, Heng Ji, Christopher Manning, Smaranda Muresan, Owen C. Rambow, Dan Roth, Mark Sammons, Benjamin Van Durme. TinkerBell: Cross-lingual Cold-Start Knowledge Base Construction.
Proceedings of the 2017 Text Analysis Conference, TAC 2017.

[21] Lingjia Deng and Janyce Wiebe. Joint Prediction for Entity/Event-Level Sentiment Analysis using Probabilistic Soft Logic Models. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[22] Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. Benefactive/Malefactive Event and Writer Attitude Annotation. *Annual Meeting of the Association for Computational Linguistics* (short paper), 2013.

[23] Janyce Wiebe and Lingjia Deng (2014). An account of opinion implicatures. *arXiv:1404.6491v1* [cs.CL]

[24] Banea, C., R. Mihalcea, and J. Wiebe. Sense-level subjectivity in a multilingual setting. *Computer Speech and Language. 28(1):7–19.* 2014.

[25] Lingjia Deng, Janyce Wiebe and Yoonjung Choi. Joint Inference and Disambiguation of Implicit Sentiments via Implicature Constraints. *Proceedings of COLING*, 2014.

[26] Lingjia Deng and Janyce Wiebe. Sentiment Propagation via Implicature Constraints. *Meeting of the European Chapter of the Association for Computational Linguistics*, 2014.

[27] Yoonjung Choi, Janyce Wiebe and Lingjia Deng. Lexical Acquisition for Opinion Inference: A Sense-Level Lexicon of Benefactive and Malefactive Events. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2014.

[28] Lingjia Deng and Janyce Wiebe. An Investigation for Implicatures in Chinese: Implicatures in Chinese and in English are similar! *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2014.

[29] Janyce Wiebe and Lingjia Deng. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2014.

[30] Lingjia Deng and Janyce Wiebe. MPQA 3.0: Entity/Event-Level Sentiment Corpus. *Proceedings of NAACL-HLT,* 2015. (short paper)

[31] Lingjia Deng and Janyce Wiebe. Recognizing Opinion Sources Based on A New Categorization of Opinion Types. *Proceedings of the 25th International Joint Conference on Artificial Intelligence,* 2016.

[32] Lingjia Deng and Janyce Wiebe. How can NLP tasks mutually benefit sentiment analysis? A holistic approach to sentiment analysis. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2016.

[33] Yoonjung Choi, Janyce Wiebe, Rada Mihalcea. Coarse-grained +/-effect word sense disambiguation for implicit sentiment analysis. IEEE Transactions on Affective Computing, 2017.

[34] Lingjia Deng. Entity/event-level sentiment detection and inference. *PhD dissertation*, University of Pittsburgh, 2016.

[35] Bach-2015. Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2015. Hinge-loss markov random fields and probabilistic soft logic. *arXiv:1505.04406* [cs.LG].

[36] Banea, C., Choi, Yoonjung, Deng, L., Hassan, S., Mohler, M., Yang, B., Cardie, C., Mihalcea, R., and Wiebe, J. CPN-CORE: A Text Semantic Similarity System Infused with Opinion Knowledge. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: *Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. pp. 221-228. Association for Computational Linguistics, 2013.

[37] Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, J. Wiebe. Task 10: Multilingual semantic textual similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, 2014.

[38] Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, G. Rigau, L. Uria, J. Wiebe. Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval), 2015.

[39] Agirre, E., C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, and J. Wiebe. SemEval-2016 Task 1: Semantic textual similarity - Monolingual and cross-lingual evaluation. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, 2016.

[40] Overview of the TAC 2014 Knowledge Base Population Evaluation: Sentiment Slot Filling. Claire Cardie, Carmen Banea, Rada Mihalcea and Janyce Wiebe. *Text Analysis Conference (TAC)*, 2014.

[41] Owen Rambow, Meenakshi Alagesan, Michael Arrigo, Daniel Bauer, Claire Cardie, Adam Dalton, Hoa Dang, Mona Diab, Greg Dubbin, Jason Duncan, Gregorios Katsios, Axinia Radeva, Tomek Strzalkowski, Jennifer Tracey. The 2016 TAC KBP BeSt Evaluation. *Proceedings of the 2016 Text Analysis Conference, TAC 2016*.

[42] Owen Rambow, Mohamed Al-Badrashiny, Meenakshi Alagesan, Michael Arrigo, Daniel Bauer, Claire Cardie, Adam Dalton, Mona Diab, Greg Dubbin, Gregorios Katsios, Axinia Radeva, Tomek Strzalkowski, Jennifer Tracey. The 2017 TAC KBP BeSt Evaluation. *Proceedings of the 2017 Text Analysis Conference, TAC 2017*.

[43] Carmen Banea, Di Chen, Rada Mihalcea, Claire Cardie, Janyce Wiebe. SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, 2014.

[44] X Chen, A Katiyar, X Yan, L Wang, C Banea, Y Choi, L Deng, C Cardie, R Mihalcea and J Wiebe. CornPittMich Sentiment Slot-Filling System at TAC 2014. *Proceedings of TAC* 2014.