Abstract

We present a new approach for document-level sentiment inference, where the goal is to predict directed opinions (who feels positively or negatively towards whom) for all entities mentioned in a text. To encourage more complete and consistent predictions, we introduce an ILP that jointly models (1) sentence- and discourse-level sentiment cues, (2) factual evidence about entity factions, and (3) global constraints based on social science theories such as homophily, social balance, and reciprocity. Together, these cues allow for rich inference across groups of entities, including for example that CEOs and the companies they lead are likely to have similar sentiment towards others. We evaluate performance on new, densely labeled data that provides supervision for all pairs, complementing previous work that only labeled pairs mentioned in the same sentence. Experiments demonstrate that the global model outperforms sentence-level baselines, by providing more coherent predictions across sets of related entities.

1 Introduction

Documents often present a complex web of facts and opinions that hold among the entities they describe. Consider the international relations story in Figure 1. Representatives from three countries form factions and create a network of sentiment. While some opinions are relatively directly stated (e.g., Russia criticizes Belarus), many others must be inferred based on the factual ties among entities (e.g., Moscow, Gryzlov, and Russia probably share the same sentiment towards other entities) and known social context (e.g., Russia probably dislikes Saakashvili since Russia criticized Belarus for supporting him). In this paper, we show that jointly reasoning about all of these factors can provide more complete and consistent document-level sentiment predictions.

More concretely, we present a global model for document-level entity-to-entity sentiment, i.e., who feels positively (or negatively) towards whom. Our goal is to make exhaustive predictions over all entity pairs, including those that require cross-sentence inference. We present a Integer Linear Programming (ILP) model that combines three complementary types of evidence: entity-pair sentiment classification, template-based faction extraction, and sentiment dynamics in social groups. Together, they allow for recovering more complete predictions of both the explicitly stated and im-
Figure 2: Entity subgraphs for the example in Figure 1: (a) shows explicitly stated sentiment, (b) shows faction relationships and (c) shows all edges for Georgia and its representative Saakashvili. Through Saakashvili’s relationship with Belarus, Georgia forms an alliance with Belarus, providing evidence for an inferred negative stance towards Russia. Green dotted edges represent positive sentiment, red are negative, and blue dashed lines show faction relationship.

explicit sentiment, while preserving consistency.

The sentiment dynamics in social groups, motivated by social science theories, are encoded as soft ILP constraints. They include a notion of homophily, that entities in the same group tend to have similar opinions (Lazarsfeld and Merton, 1954). For example, Figure 2b shows directed faction edges, where one entity is likely to agree with the other’s opinions. They also encode dyadic social constraints (i.e., the likely reciprocity of opinions (Gouldner, 1960)) and triadic social dynamics following social balance theory (Heider, 1946). For example, from Russia’s criticism on Belarus and Belarus’ positive attitude towards Saakashvili (in Figure 2a), we can infer that Russia is negative towards Saakashvili (in Figure 2c). When considered in aggregate, these constraints can greatly improve the consistency over the overall document-level predictions.

Our work stands in contrast to previous approaches in three aspects. First, we apply social dynamics motivated by social science theories to entity-entity sentiment analysis in unstructured text. In contrast, most previous studies focused on social media or dialogue data with overt social network structure when integrating social dynamics (Tan et al., 2011; Hu et al., 2013; West et al., 2014). Second, we aim to recover sentiment that can be inferred through partial evidence that spans multiple sentences. This complements prior efforts for accessing implied sentiment where the key evidence is, by and large, at the sentence level (Zhang and Liu, 2011; Yang and Cardie, 2013; Deng and Wiebe, 2015a). Finally, we present the first approach to model the relationship between factual and subjective relations.

We evaluate the approach on a newly gathered corpus with dense document-level sentiment labels in news articles. This data includes comprehensively annotated sentiment between all entity pairs, including those that do not appear together in any single sentence. Experiments demonstrate that the global model significantly improves performance over a pairwise classifier and other strong baselines. We also perform a detailed ablation and error analysis, showing cases where the global constraints contribute and pointing towards important areas for future work.

2 A Document-level Sentiment Model

Given a news document $d$, and named entities $e_1,...,e_n$ in $d$, where each entity $e_i$ has mentions $m_{i1} \cdots m_{ik}$, the task is to decide directed sentiment between all pairs of entities. We predict the directed sentiment from $e_i$ to $e_j$ at the document level, i.e., sent($e_i \rightarrow e_j$) $\in \{\text{positive, unbiased, negative}\}$, for all $e_i, e_j \in d$ where $i \neq j$, assuming that sentiment is consistent within the document.

We introduce a document-level ILP that includes base models and soft social constraints. ILP has been used successfully for a wide range of NLP tasks (Roth and Yih, 2004), perhaps because they easily support incorporating different types of global constraints. We use two base models: (1) a learned pairwise sentiment classifier (Sec 3.1) that combines sentence- and discourse-level features to make predictions for each entity pair and (2) a pattern-based faction extractor (Sec 3.2) that detects alliances among a subset of the entities.

The ILP is solved by maximizing:

$$F = \psi_{social} + \psi_{fact} + \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij}$$

where $F$ combines soft constraints ($\psi_{social}, \psi_{fact}$ defined in detail in this section) with pairwise potentials $\psi_{ij}$ defined as:

---

1 All data will be made publicly available. You can browse it at http://homes.cs.washington.edu/~eunsol/project_page/acl16, and download it from the author’s webpage.
<table>
<thead>
<tr>
<th>Sentence</th>
<th>$i$</th>
<th>$j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Prime Minister Harper...</td>
<td>Canada</td>
<td>Harper</td>
</tr>
<tr>
<td>... Reid, the Democratic leader...</td>
<td>Reid</td>
<td>Democratic</td>
</tr>
<tr>
<td>Goldman spokesman DuVally</td>
<td>Goldman</td>
<td>DuVally</td>
</tr>
<tr>
<td>... Djibouti, a key U.S. ally.</td>
<td>Djibouti</td>
<td>U.S.</td>
</tr>
</tbody>
</table>

(a) Detection examples

Figure 3: An example sentiment inference from faction relationships. Pairs in factions are encouraged to share opinions, and to be positive towards other tied entities. On the right, sentiment edges can be both positive or both negative.

$$\psi_{ij} = \phi_{pos_{ij}} \cdot pos_{ij} + \phi_{neg_{ij}} \cdot neg_{ij} + \phi_{neu_{ij}} \cdot neu_{ij}$$

Each potential $\psi_{ij}$ includes the sentiment classifier scores ($\phi_{pos}$, $\phi_{neg}$, $\phi_{neu}$) with binary variables $pos_{ij}$, $neg_{ij}$ and $neu_{ij}$ where, for example, $neg_{ij}=1$ indicates that $e_i$ is negative towards $e_j$. Decision variables $pos_{ij}$ and $neu_{ij}$ are defined analogously for positive and neutral opinion. Finally, we introduce a hard constraint:

$$\forall i,j pos_{ij} + neg_{ij} + neu_{ij} = 1$$

to ensure a single prediction is made per pair.

2.1 Inference with factions

Our first soft ILP constraint $\psi_{fact}$ models that fact that entities in supportive social relations tend to share similar sentiment toward others (Lazarsfeld and Merton, 1954), and are often positive towards each other. For now, we assume access to a base extractor to provide such faction relations (Sec. 3.2 provides details of our pattern-based extractor). Figure 3a illustrates sample detections.

We introduce a binary variable $tie_{ij}$, where $tie_{ij}=1$ denotes an extracted faction relationship. These variables are tied to the variables regarding sentiment via the variables

$$tie_{same_{ijk}} = tie_{ij} \land pos_{ik} \land pos_{jk}$$

$$tie_{diff_{ijk}} = tie_{ij} \land pos_{ik} \land neg_{jk}$$

$$\text{itself}_{ij} = tie_{ij} \land pos_{ij} - tie_{ij} \land neg_{ij}$$

which are used in the following objective term:

$$\psi_{fact} = \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_{diff} \cdot \text{itself}_{ij} + \sum_{k=1}^{n} (\alpha_{fact} \cdot (tie_{same_{ijk}} - tie_{diff_{ijk}})))$$

This formulation enables the model to predict implicit sentiment by jointly considering factual and sentiment relations among other entity pairs, essentially drawing a connection between sentiment analysis and information extraction. Figure 3 visualizes this inference pattern.

2.2 Inference with sentiment relations

We also include constraints $\psi_{social}$ in the objective that model social balance and reciprocity.

**Balance theory constraints:** Social balance theory (Heider, 1946) models the sentiment dynamics in an interpersonal network. In particular, in balanced states, entities on positive terms have similar opinions towards other entities and those on negative terms have opposing opinions. We introduce a set of variables to capture this insight: for example, the case where $e_i$ is positive towards $e_j$ is shown below (analogous when negative).

$$\text{pos}_{same_{ijk}} = pos_{ij} \land pos_{ik} \land pos_{jk}$$

$$\text{pos}_{diff_{ijk}} = pos_{ij} \land neg_{ik} \land pos_{jk}$$

and add the term $\psi_{bl}$ to $\psi_{social}$.

$$\psi_{bl} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} (\alpha_{bl} \cdot (\text{pos}_{same_{ijk}} + \text{neg}_{diff_{ijk}})$$

A visualization of these constraints is in Figure 4.
<table>
<thead>
<tr>
<th></th>
<th>Faction</th>
<th>Balance</th>
<th>Reciprocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS</td>
<td>57%</td>
<td>64%</td>
<td>73%</td>
</tr>
<tr>
<td>NEG</td>
<td>60%</td>
<td>61%</td>
<td>78%</td>
</tr>
</tbody>
</table>

Table 1: Percentage of labels where each constraint holds. For example, positive on reciprocity means when \( pos(e_i, e_j) \) is true, 73% of times \( pos(e_j, e_i) \) is also true.

**Reciprocity constraint:** Reciprocity of sentiment has been recognized as a key aspect of social stability (Johnston, 1916; Gouldner, 1960). To model reciprocity among the real world entities, we introduce variables:

\[
    r_{\text{same}ij} = pos_{ij} \land pos_{ji} \land neg_{ij} \land neg_{ji}
\]

\[
    r_{\text{diff}ij} = pos_{ij} \land neg_{ji} \land neg_{ij} \land pos_{ji}
\]

\[
    \psi_r = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_r(r_{\text{same}ij}) + \alpha_{badr}(r_{\text{diff}ij})
\]

and add the term \( \psi_r \) to the \( \psi_{\text{social}} \).

### 2.3 Discussion

While many studies exist on homophily, social balance, and reciprocity, no prior work has reported quantitative analysis on the sentiment dynamics among the real world entities that appear in unstructured text. Thus we report the data statistics based on the development set in Table 1. We find that the global constraints hold commonly but are not universal, motivating the use of soft constraints (see Sec. 6).

### 3 Pairwise Base Models

The global model in Sec. 2 uses two base models, one for pairwise sentiment classification and the other for detecting faction relationships.

#### 3.1 Sentiment Classifier

The entity-pair classifier considers a holder entity \( e_i \), its mentions \( m_{i1} \cdots m_{ip} \), a target entity \( e_j \), its mentions \( m_{j1} \cdots m_{jp} \), and document \( d \). It predicts \( \text{sent}(e_i \rightarrow e_j) \in \{\text{positive, unbiased, negative}\} \). The input is plain text and no gold labels are assumed; entity detection, dependency parse and co-reference resolution are automatic, and include common nouns and pronoun mentions (details in Sec. 4.1). We trained separate classifiers for pairs that co-occur in a sentence and those that do not, using a linear class-weighted SVM classifier with crowd-sourced data described in Sec. 4.2.

In what follows, we describe three different types of features we developed: dependency features, document features, and quotation features. Many of the features test the overall sentiment of a set of words (e.g., the complete document, a dependency path, or a quotation). In each case, we define the sentiment label for the text to be positive if it contains more words that appear in the positive sentiment lexicon than that appear in the negative one (and similarly for the negative label). We used MPQA sentiment lexicon (Wilson et al., 2005) for our study, which contains 2,718 positive and 4,912 negative lexicons.

**Dependency Features** We consider all dependency paths between the head word of \( e_i \) and \( e_j \) in each sentence, and aggregate over all co-occurring sentences. The features compute: (1) The sentiment label of the path containing dobj and nsubjrev, up to length three if the path contains sentiment lexicon words (e.g., Olympic hero Skah accuses Norway over custody battle.) (2) The sentiment label of the path \( e_i \uparrow \text{nsubj} \downarrow \text{ccomp} \downarrow \text{nsubj} \downarrow e_j \), when it exists (e.g., McCully said any action against Henry is a matter entirely for TVNZ) (3) The sentiment label of path when the path does not contain any named entity (e.g., Nobel winner, Shirin Ebadi) (4) An indicator for the link nmod:against.

**Document Features** Previous work has shown that notions related to salience (e.g., proximity to sentiment words) can help to detect sentiment targets (Ben-Ami et al., 2014). In our data, we found that an entity’s occurrence pattern is highly indicative of being involved in sentiment, for example the most frequently mentioned entity is 3.4 times more likely to be polarized and an entity in the headline is two times more likely to be polarized.

Pairwise features include the NER type of \( e_i \) and \( e_j \) and the percentage of sentences they co-occur in. We also use features indicating whether \( e_i \) and \( e_j \) (1) are mentioned in the headline and (2) appear only once in the document. When they are the two most frequent entities, we add the document sentiment label as a feature. For entity pairs that do not appear together in any sentence, we also include the rank of holder and target in terms of overall number of mentions in the document.

**Quotation Features** Quotations often involve subjective opinions towards prominent entities in news articles. Thus we include document-level
features encoding this intuition. For example, the sentence “We’re pleased to put this behind us,” said Michael DuVally implies positive sentiment from DuVally. We extract direct quotations using regular expressions. We include the sentiment label of the direct quotation from the speaker to the entities in it, excluding entities that appear less than three times in the document. We add the sentiment label of the quotation as a feature to (speaker, the most frequent entity) pair as well.

To extract indirect quotations, we follow studies (Bethard et al., 2004; Lu, 2010) and use a list of 20 verbs indicating speech events (e.g., say, speak, and announce) to detect direct quotations and their opinion holders. We then add the sentiment label of words connected to e_j via a dependency path of length up to two that also includes the subject of quotation verb to e_j (e.g. Hassanal said that cooperation between Brunei and China were fruitful). We also include an indicator feature for whether e_i is the subject of the quotation verb.

3.2 Faction Detector

We use a simple pattern-based detector that extracts a faction relationship between a pair of entities if the dependency path between them either:
1. contains only one link of modifier or compound label (nmod, nmod : poss, amod, nn, or compound).
2. or contains less than three links and has a possessive or appositive label (poss or appos).

Example extractions for this approach, which we adopted for its simplicity and the fact that it works reasonably well in practice, are shown in Figure 3a. On average we detect 1.7 ties per document on a small development set with roughly 30% recall and 60% precision. Improving performance and adding more relation types is an important area for future work.^{2}

4 Data

We collected new datasets that densely label sentiment among entities in news articles, including: 208 documents, 2,226 sentences, and 15,185 entity pair labels. It complements existing datasets such as MPQA which provides rich annotations at the sentence-level (Deng and Wiebe, 2015b) and the recent KBP challenge which provides sparse annotations at the corpus-level (Ellis et al., 2014), by providing document-level annotations for all entity pairs (see Sec. 7 for discussion).

4.1 Document Preprocessing

All-pair annotation can be expensive, as there are \(N^2\) pairs to annotate for each document with \(N\) entities. We determined that it would be more cost efficient to cover a large number of short documents than a small number of very long documents. We therefore selected articles with less than eleven entities from KBP and less than fifteen from MPQA and took the first 15 sentences for annotation. We used Stanford CoreNLP (Manning et al., 2014) for sentence splitting, part-of-speech tagging, named entity recognition, co-reference resolution and dependency parsing. We discarded entities of type date, duration, money, time and number and merged named entities using several heuristics, such as merging acronyms, merging named entity of person type with the same last name (e.g., Tiger Woods to Woods). We merged names listed as alias in when there is an exact match from Freebase. We included all mentions in a co-reference chain with the named entity, discarding chains with more than one entity. The corpus statistics are shown in Table 2.

4.2 Sentiment Data Collection

We annotated data using two methods: freelancers ($7.6 per article on average) covering all entity pairs and crowd-sourcing ($1.6 per article on average) covering a subset of entity pairs.

Evaluation Dataset We provide exhaustive annotations covering all pairs for the evaluation set. We hired freelancers from UpWork,\(^3\) after examining performance on five documents. They labeled entity pairs with one of the following classes.

**POS**: positive towards the target.

**NOTNEG**: positive or unbiased towards the target.

---

\(^{2}\)We experimented with using relations from an external knowledge base (Freebase), but KB sparsity and entity linking errors posed major challenges.

\(^{3}\)https://www.upwork.com

<table>
<thead>
<tr>
<th>Document count</th>
<th>KBP</th>
<th>MPQA</th>
<th>Crowdsourced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. sentence count</td>
<td>10.0</td>
<td>12.7</td>
<td>14.8</td>
</tr>
<tr>
<td>Avg. entity count</td>
<td>7.9</td>
<td>10.6</td>
<td>8.8</td>
</tr>
<tr>
<td>Avg. mentions / entity</td>
<td>3.6</td>
<td>2.7</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 2: Corpus Statistics
Here, we introduced the NotPOS and NotNEG classes to mark more subjective cases where we expect agreement might be lower. For example, one assigned NotPOS to sentiment(Goldman, FINRA), The FINRA said Goldman lacked adequate procedures to ... and another assigned NotNEG to sentiment(Macalintal, Arroyo) in the next example. ... Arroyo’s election lawyer, Romulo Macalintal. Arguments could be made for NEG or POS, respectively, but the decision is inherently subjective and requires careful reading.\(^4\)

We also asked annotators to mark the label as inferred when not explicitly stated but implied from the context or world knowledge. Allowing for inferred labels and finer-grained labels encouraged annotators to capture implicit sentiment. For each judgement, we acquired two labels. Inter-annotator agreement, in Table 4, is high for the relaxed metrics, confirming our intuitions about the annotation control. We collected labels from three annotators randomly presented test questions for quality control. We collected labels from three annotators for each entity pair, and considered labels when at least two agreed. The resulting annotation contains total 2,995 labels on 914 documents, 682 positive, 836 negative and 474 without sentiment, which we discarded.

4.3 Insights Into Data

This data supports the study of sentiment-laden entity pairs across sentence boundaries and inferred labels among entities, as we show here.

4.3.1 Sentiment Beyond Sentence Boundary

Approximately 25% of polarized sentiment labels are between entities that do not co-occur\(^7\) in a sentence (see Table 5). For example, in the article

\(^4\) In the construction of MPQA3.0 dataset, entity-entity/event sentiment corpus, even with iterative expert annotation, 31% of disagreements are caused by negligence.

\(^5\) LDC2014E13:TAC2014KBP English Corpus

\(^6\) http://www.crowdflower.com

\(^7\) This is an estimate due to co-reference resolution errors.
with headline ‘Russia heat, smog trigger health problems’,

... “We never care to work with a future perspective in mind,” Alexei Skripkov of the Federal Medical and Biological Agency said. “It’s a big systemic mistake.”

Skripkov never appears together with Russia in any sentence, but he manifests negative sentiment towards it. When a document revolves around a theme (in this example Russia), sentiment is often directed to it without being explicitly mentioned.

**Inferred sentiment** Annotators marked labels as inferred frequently, especially on less polarized sentiment (see Table 6). Various clues led to sentiment inference. For example, in the following document, we can read Sam Lake’s positive attitude towards Paul Auster from his ‘citing’ action:

Ask most video-game designers about their inspirations ... Sam Lake cites Paul Auster’s “The Book of Illusions”

Sentiment can also be inferred through reasoning over another entity.

The U.N. imposed an embargo against Eritrea for helping insurgents opposed to the Somali government.

By considering relations with Eritrea, we can infer U.N. would be positive towards Somalia.

### 5 Experimental Setup

#### Data and Metrics

We randomly split the densely labeled KBP document set, using half as a test data and half as a development data. One half of the development set was used to tune hyper parameters,\(^8\) and the other for error analysis and ablations. After development, we ran on the test sets composed of KBP documents and MPQA documents. For MPQA we did not create a separate development set and reserved all of the relatively modest amount of data for a more reliable test set. For the pairwise classifier, we report development results using five-fold cross validation on the training data.

We report micro-averaged precision, recall, and F-measure for both sentiment labels.

#### Comparison Systems

We compare performance to two simple baselines and two adaptations of existing sentiment classifiers. The baselines include our base pairwise classifier (Pair) and randomly assigning labels according to their empirical distribution (Random).

The first existing method adaptation (Sentence) uses the publicly released sentence-level RNN sentiment model from Socher et al (2013). For each entity pair, we collect sentiment labels from sentences they co-occur in and assign a positive label if a positive-labeled sentence exists, negative if there exists more than one sentence with a negative label and no positives.\(^9\)

We also report a proxy for doing similar aggregation over a state-of-the-art entity-entity sentiment classifier. Here, because we added our new labels to the original KBP and MPQA3.0 annotations, we can simply predict the union of the original gold annotations using mention string overlap to align the entities (KM_Gold). This provides a reasonable upper bound on the performance of any extractor trained on this data.\(^10\)

#### Implementation Details

We use CPLEX\(^11\) to solve the ILP described in Sec. 2. For computational efficiency and to avoid erroneous propagation, soft constraints associated with reciprocity and balance theory are introduced only on pairs for which a high-precision classifier assigned polarity. For the pairwise classifier, we use a class-weighted linear SVM.\(^12\) We include annotated pairs, and randomly sample negative examples from pairs without a label in the crowd-sourced training dataset. We made two versions of pairwise classifiers by tuning weight on polarized classes and negative sampling ratio by grid search. One is tuned for high precision to be used as a base classifier for ILP (ILP base), and the other is tuned for the best F1 (Pairwise).\(^13\)

### 6 Results

Table 7 shows results on the evaluation datasets. The global model achieves the best F1 on both labels. All systems do significantly better than the random baseline but, overall, we see that entity-entity sentiment detection is challenging, require-

\(^8\) We used the following values \((\alpha_f, \beta_{bad}, \alpha_{self}, \alpha_{faction}, \alpha_{bd}, \alpha_{badbd}) = (0.7, -0.8, 0.4, 0.5, 0.1, -0.5).\)

\(^9\) Due to domain difference, the system predicted negative labels more (73% of sentences were classified as negative).

\(^10\) We consider this gold evaluation a direct proxy for the recent work Deng and Wiebe (2015a), which is the most related recent entity-entity sentiment model trained on the gold data whose predictions we are evaluating against.

\(^11\) http://tinyurl.com/joccfqy

\(^12\) http://scikit-learn.org/

\(^13\) We use 10 as the weights for the polarized classes. Pairwise and base classifier for MPQA sampled 4%, base classifier for KBP sampled 10% of unlabeled pairs.
Table 7: Performance on the evaluation datasets: including implicit and explicit sentiment.

<table>
<thead>
<tr>
<th></th>
<th>Development Set (KBP)</th>
<th>KBP</th>
<th>MPQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>KM_Gold</td>
<td>90.9</td>
<td>2.3</td>
<td>4.8</td>
</tr>
<tr>
<td>Random</td>
<td>16.6</td>
<td>13.1</td>
<td>14.7</td>
</tr>
<tr>
<td>Sentence</td>
<td>25.2</td>
<td>19.4</td>
<td>17.0</td>
</tr>
<tr>
<td>Pairwise</td>
<td>34.5</td>
<td>32.0</td>
<td>32.3</td>
</tr>
<tr>
<td>Global</td>
<td>58.2</td>
<td>37.9</td>
<td>45.9</td>
</tr>
</tbody>
</table>

Table 8: ILP constraints ablation study.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>ILP base</td>
<td>56.7</td>
<td>25.2</td>
</tr>
<tr>
<td>+ Rec.</td>
<td>53.5</td>
<td>30.0</td>
</tr>
<tr>
<td>+ Balance</td>
<td>49.6</td>
<td>30.4</td>
</tr>
<tr>
<td>+ Faction</td>
<td>58.9</td>
<td>30.2</td>
</tr>
</tbody>
</table>

Table 9: Pairwise classifier feature ablation study.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>All</td>
<td>34.5</td>
<td>39.7</td>
</tr>
<tr>
<td>- Depend.</td>
<td>32.9</td>
<td>32.1</td>
</tr>
<tr>
<td>- Doc.</td>
<td>32.6</td>
<td>41.0</td>
</tr>
<tr>
<td>- Quotation</td>
<td>33.6</td>
<td>39.5</td>
</tr>
</tbody>
</table>

Table 10: Error Analysis on the development set.

- Sentiment expression detection error 21.0%
- Missing world knowledge 19.3%
- Named entity detection error 17.5%
- Co-reference failure 14.8%
- Propagation error 12.3%
- Missing faction 7.0%

The global model also improves performance over the pairwise classifier (Pairwise) for both datasets, but we see very different behavior due to the different sentiment label distributions (see Table 3). The KBP data has many fewer unbiased pairs and many mistakes are from choosing the wrong polarity. For the pairwise classifier 17% of all predictions were assigned the opposite polarity. After the global inference, it is reduced to 11%, contributing to the gain in overall precision. For MPQA the base classifier has a more challenging detection task, due to relatively large amount of the unbiased pairs. Here, the best base classifier misses many pairs and the global model helps to fill in some of these gaps in recall.

In both cases, the document-level model often propagates correct labels by detecting easier, explicit expressions. For example, given the sentence “Buphavanh said Laos creates favorable conditions for Vietnamese companies,” the base classifier detected positive sentiment from Buphavanh to Vietnam, but not between Vietnam and Laos. By detecting the fact that Buphavanh is the prime minister of Laos, it infers the extra sentiment pairs.

We also did ablation studies to measure the contributions of different components. Table 8 shows ablations of each soft constraint. The fact constraint is the most helpful, improving both precision and recall for both labels. The reciprocity and social balance constraints tend to improve recall at the cost of precision. Table 9 shows ablations of the base classifier features. All features are helpful, with dependency features most helpful for positive labels, and quotation and document-level features more with negatives.

**Error Analysis** We manually analyzed errors on 20 articles from the development set (Table 10). Our system failed when there were sentiment words not in the lexicon, or negated sentiment words. Capturing subtle sentiment expressions beyond sentiment lexicon should improve the performance. Preprocessing, as a whole, was the largest source of error. It includes co-reference failure and named entity error. Co-reference mistakes happen as a result of not resolving pronouns, referring expressions, as well as named entities co-references (e.g., Financial Industry Regulatory Authority to FINRA), or erroneously merging them. Lengthy quotations or nested mentions triggered co-reference error, affecting mostly recall. Named entity errors includes incorrect named...
entity detection (e.g., pro-Israel) and mention detection boundary errors. For example, we detected negative sentiment from Mexico to Pakistan from *Mexico condemns Pakistan series suicide bomb attacks*. While actual sentiment is positive. Finally, the ILP propagates sentiment labels erroneously at times. Our constraints often hold among entities of the same type, but are less predictive among entities of different types. For example, when a person supports a peace treaty, the treaty does not have sentiment towards him/her. For future work refining constraints based on entity type should help performance.

7 Related Work

Sentiment Inference Our sentiment inference task is related to the recent KBP sentiment task,\(^{14}\) in that we aim to find opinion target and holder. While we study the complete document-level analysis over all entity pairs, the KBP task is formulated as query-focused retrieval of entity sentiment from a large pool of potentially relevant documents. Thus, their annotations focus only on query entities and relatively sparse compared to ours (see Sec. 6). Another recent dataset is MPQA 3.0 (Deng and Wiebe, 2015b), which captures various aspects of sentiment. Their sentiment pair annotations are only at the sentence-level and are therefore much sparser than we provide (see Sec. 6) for entity-entity relation analysis.

Several recent studies focused on various aspects of implied sentiment (Greene and Resnik, 2009; Mohammad and Turney, 2010; Zhang and Liu, 2011; Feng et al., 2013; Deng and Wiebe, 2014; Deng et al., 2014). Deng and Wiebe (2015a) in particular introduced sentiment implicature rules relevant for sentence-level entity-entity sentiment. Our work contributes to these recent efforts by presenting a new model and dataset for document-level sentiment inference over all entity pairs.

Document-level Analysis Stoyanov and Claire (2011) also studied document-level sentiment analysis based on fine-grained detection of directed sentiment. They aggregate sentence-level detections to make document-level predictions, while our we model global coherency among entities and can discover implied sentiment without direct sentence-level evidence. In the event extraction domain, previous research showed the effectiveness of jointly considering multiple sentences. Yang and Mitchell (2016) proposed joint extraction of entities and events with the document context, improving on the event extraction. Most work focuses on events, while we primarily study sentiment relations.

Social Network Analysis While many previous studies considered the effect of social dynamics for social media analysis, most relied on an explicitly available social network structure or considered dialogues and speech acts for which opinion holders are given (Tan et al., 2011; Hu et al., 2013; Li et al., 2014; West et al., 2014; Krishnan and Eisenstein, 2015). Compared to the recent work that focused on relationships among fictional characters in movie summaries and stories (Chaturvedi et al., 2016; Srivastava et al., 2016; Iyyer et al., 2016), we consider a broader types of named entities on news domains.

8 Conclusion

We presented an approach to interpreting sentiment among entities in news articles, with global constraints provided by social, faction and discourse context. Experiments demonstrated that the approach can infer implied sentiment and point toward potential directions for future work, including joint entity detection and incorporation of more varied types of factual relationships.

Acknowledgments

This research was supported in part by the NSF (IIS-1252835, IIS-1408287, IIS-1524371), DARPA under the DEFT program through the AFRL (FA8750-13-2-0019), an Allen Distinguished Investigator Award, and a gift from Google. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082. The authors thank the members of UW NLP group for discussions and support. We also thank the anonymous reviewers for insightful comments. Finally, we thank the annotators from the CrowdFlower and UpWork.

\(^{14}\)http://www.nist.gov/tac/2014/KBP/Sentiment
References


Fritz Heider. 1946. Attitudes and cognitive organization. The Journal of psychology, 21(1).


