Document-level Sentiment Inference with Social, Faction, and Discourse Context

Eunsol Choi **Hannah Rashkin** **Luke Zettlemoyer** **Yejin Choi**
Computer Science & Engineering
University of Washington
{eunsol,hrashkin,lsz,yejin}@cs.washington.edu

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 Saakhashvili 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.

Russia criticized *Belarus* for permitting *Georgian* President *Mikheil Saakhashvili* to appear on *Belorussian* television. "The appearance was an unfriendly step towards *Russia*," the speaker of *Russian* parliament *Boris Gryzlov* said. . . . *Saakhashvili* announced Thursday that *he* did not understand *Russia*'s claims. *Moscow* refused to have any business with *Georgia*'s president after the armed conflict in 2008 . . .

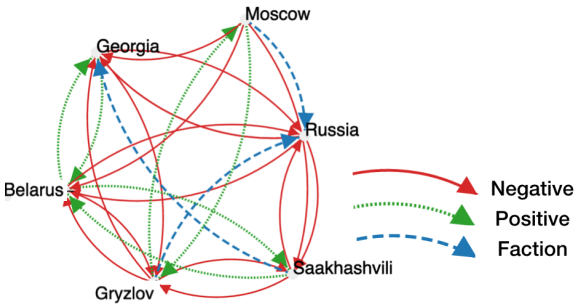


Figure 1: Example text excerpt paired with the document-level sentiment graph we aim to recover. The graph includes edges with direct textual support (e.g., from Russian to Belarus given the verb "criticized") as well as ones that must be inferred at the whole-document level (e.g., from Gryzlov to Saakhashvili given the web of relationships and opinions between them, Georgia, Russian, and Belarus).

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-

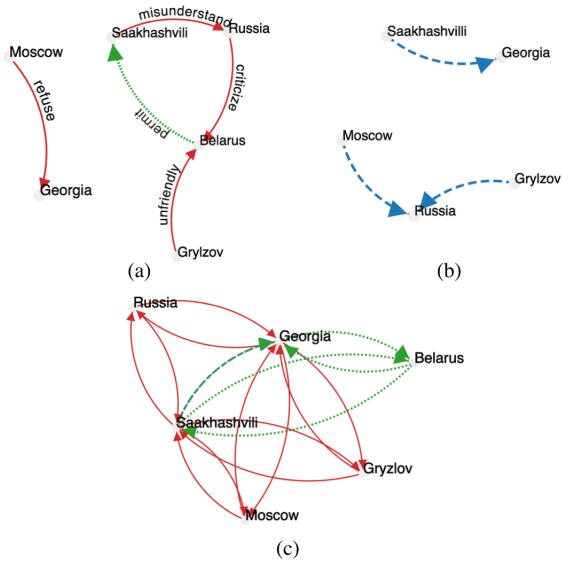


Figure 2: Entity subgraphs for the example in Figure 1: (a) shows explicitly stated sentiment, (b) shows faction relationships and (c) shows all edges for Georgia and its representative Saakhashvili. Through Saakhasvili's relationship with Belarus, Georgia forms an alliance with Belarus, providing evidence for an inferred negative stance towards Russia. Green dotted edges represent positive sentiment, red are negative, and blue dashed lines show faction relationship.

plicit 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 Saakhashvilli (in Figure 2a), we can infer that Russia is negative towards Saakhashvilli (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.[1] 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($e_i \rightarrow e_j$) $\in$ {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:

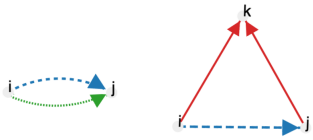
$$F = \psi_{social} + \psi_{fact} + \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij}$$

where F combines soft constraints (ψ_{social}, ψ_{fact} defined in detail in this section) with pairwise potentials ψ_{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.

Sentence	i	j
Canadian Prime Minister Harper. . .	Canada	Harper
. . . Reid, the Democratic leader. . .	Reid	Democratic
Goldman spokesman DuVally	Goldman	DuVally
. . . Djibouti, a key U.S. ally.	Djibouti	U.S.

(a) Detection examples



(b) Visual representation of common inference patterns.

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.

$$\psi_{ij} = \phi_{pos_{ij}} \cdot \texttt{pos}_{ij} + \phi_{neg_{ij}} \cdot \texttt{neg}_{ij} + \phi_{neu_{ij}} \cdot \texttt{neu}_{ij}$$

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

$$\forall i, j \; \texttt{pos}_{ij} + \texttt{neg}_{ij} + \texttt{neu}_{ij} = 1$$

to ensure a single prediction is made per pair.

2.1 Inference with factions

Our first soft ILP constraint ψ_{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 $\texttt{tie}_{ij}$, where $\texttt{tie}_{ij}$=1 denotes an extracted faction relationship. These variables are tied to the variables regarding sentiment via the variables

$$\begin{aligned}
\texttt{tie_same}_{ijk} &= \texttt{tie}_{ij} \wedge \texttt{pos}_{ik} \wedge \texttt{pos}_{jk} \\
&\quad + \texttt{tie}_{ij} \wedge \texttt{neg}_{ik} \wedge \texttt{neg}_{jk} \\
\texttt{tie_diff}_{ijk} &= \texttt{tie}_{ij} \wedge \texttt{pos}_{ik} \wedge \texttt{neg}_{jk} \\
&\quad + \texttt{tie}_{ij} \wedge \texttt{neg}_{ik} \wedge \texttt{pos}_{jk} \\
\texttt{itself}_{ij} &= \texttt{tie}_{ij} \wedge \texttt{pos}_{ij} - \texttt{tie}_{ij} \wedge \texttt{neg}_{ij}
\end{aligned}$$

which are used in the following objective term:

$$\psi_{fact} = \sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_{itself} \cdot \texttt{itself}_{ij} + \sum_{k=1}^{n}(\alpha_{fact} \cdot (\texttt{tie\_same}_{ijk} - \texttt{tie\_diff}_{ijk})))$$

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.

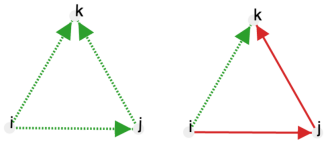


Figure 4: Balance Theory Constraints. When i is positive towards j, sharing same sentiment towards k define a balanced state. When i is negative towards j, differing opinions towards k define a balanced state.

2.2 Inference with sentiment relations

We also include constraints ψ_{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).

$$\begin{aligned}
\texttt{pos_same}_{ijk} &= \texttt{pos}_{ij} \wedge \texttt{pos}_{ik} \wedge \texttt{pos}_{jk} \\
&\quad + \texttt{pos}_{ij} \wedge \texttt{neg}_{ik} \wedge \texttt{neg}_{jk} \\
\texttt{pos_diff}_{ijk} &= \texttt{pos}_{ij} \wedge \texttt{neg}_{ik} \wedge \texttt{pos}_{ik} \\
&\quad + \texttt{pos}_{ij} \wedge \texttt{pos}_{ik} \wedge \texttt{neg}_{ik}
\end{aligned}$$

and add the term ψ_{bl} to ψ_{social}.

$$\psi_{bl} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(\alpha_{bl} \cdot (\texttt{pos\_same}_{ijk} + \texttt{neg\_diff}_{ijk}) + \alpha_{bad_{bl}} \cdot (\texttt{pos\_diff}_{ijk} + \texttt{neg\_same}_{ijk}))$$

A visualization of these constraints is in Figure 4.

	Faction	Balance	Reciprocity
POS	57%	64%	73%
NEG	60%	61%	78%

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:

$$\texttt{r\_same}_{ij} = \texttt{pos}_{ij} \wedge \texttt{pos}_{ji} + \texttt{neg}_{ij} \wedge \texttt{neg}_{ji}$$
$$\texttt{r\_diff}_{ij} = \texttt{pos}_{ij} \wedge \texttt{neg}_{ji} + \texttt{neg}_{ij} \wedge \texttt{pos}_{ji}$$
$$\psi_r = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_r(\texttt{r\_same}_{ij}) + \alpha_{bad_r}(\texttt{r\_diff}_{ij})$$

and add the term ψ_r to the ψ_{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 $m_{j1} \cdots m_{jq}$, and document d. It predicts sent($e_i \rightarrow e_j$) $\in$ {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 `nsubj_rev`, 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 $e_i \uparrow$ `nsubj` $\downarrow$ `ccomp` $\downarrow$ `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).

	KBP	MPQA	Crowdsourced
Document count	154	54	914
Avg. sentence count	10.0	12.7	14.8
Avg. entity count	7.9	10.6	8.8
Avg. mentions / entity	3.6	2.7	3.5

Table 2: Corpus Statistics

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`

Label	KBP	MPQA
POS	3.93	3.52
NOT NEG	5.73	8.06
UNBIASED	44.64	91.04
NOT POS	2.73	6.70
NEG	2.27	2.94

Table 3: Sentiment Label Statistics. Each count represents the average number per document.

UNB: unbiased towards the target
NOTPOS: negative or unbiased towards the target.
NEG: negative towards the target.

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 ambiguity of the NOTNEG and NOTPOS labels.

For experiments, we combine the fine grained labels as follows: POS or NEG is assigned when both marked it as such. When only one of the annotators marked it, we assigned the weaker sentiment (POS to NOTNEG, NEG to NOTPOS). NOTNEG and NOTPOS are assigned when either annotator marked it without 'Inferred' label. When the labels contradict in polarity or the labels are inferred weaker sentiment, UNB was assigned.

Crowdsourced Dataset We also randomly selected news articles from the Gigaword corpus,[5] and collected labels to train the base sentiment classifier (Sec. 3.1). We designed a pipelined approach, with three steps:

1. Document selection: Is there sentiment among entities in this document?
2. Entity selection: (1) Select all entities **holding** sentiment towards any other entities., and (2) Select all entities which are the **target** of sentiment by any other entity.
3. Sentiment label collection: Choose the sentiment A has towards B, from {Positive, No Sentiment, Negative}

We used CrowdFlower,[6] where annotators were 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.

	Exact	Strict	Relaxed
Positive	0.35	0.54	0.67
Negative	0.50	0.64	0.74

Table 4: Inter-annotator Agreement. Cohen's kappa score: Exact counts only exact matches, Strict counts allows NOT NEG labels to match POS, and Relaxed allows NOT NEG to match POS or UNBIASED (analogously for negative).

	POS	NOT NEG	NOT POS	NEG
KBP	25%	29%	30%	28%
MPQA	35%	49%	46%	50%

Table 5: Percentage of entity pairs that do not co-occur in a sentence.

	POS	NOTNEG	NOTPOS	NEG
KBP	70%	94%	88%	58%
MPQA	68%	74%	83%	66%

Table 6: Percentage of labels marked as inferred.

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.

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 macro-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 CPLEX4[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, requir-

[8] We used the following values (α_r, α_{bad_r}, α_{itself}, $\alpha_{faction}$, α_{bl}, $\alpha_{bad_{bl}}$) = (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.

	Development Set (KBP)						KBP						MPQA					
	Positive			Negative			Positive			Negative			Positive			Negative		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
KM_Gold	90.9	2.5	4.8	93.8	8.6	15.8	93.9	4.3	8.3	93.5	6.6	12.4	61.5	1.3	2.5	90.0	5.2	9.8
Random	16.6	13.1	14.7	4.9	4.0	4.4	13.3	12.7	13.0	10.1	6.9	8.2	10.9	15.4	12.8	8.9	6.7	7.7
Sentence	**60.0**	16.3	25.7	21.7	**43.1**	28.8	40.9	20.6	27.4	21.0	31.4	25.2	18.9	3.7	6.2	16.7	18.2	17.4
Pairwise	47.3	36.9	41.4	25.6	36.8	30.2	36.2	**35.5**	35.9	27.6	**41.2**	33.1	**28.7**	23.0	25.6	**23.2**	16.3	19.2
Global	58.2	**37.9**	**45.9**	**37.2**	35.1	**36.1**	**45.5**	32.7	**38.1**	**34.6**	36.8	**35.7**	25.2	**29.3**	**27.1**	17.6	**24.4**	**20.4**

Table 7: Performance on the evaluation datasets: including implicit and explicit sentiment.

	Positive			Negative		
	P	R	F1	P	R	F1
ILP base	56.7	25.2	34.9	36.9	27.6	31.6
+ Reci.	53.5	30.0	38.4	33.9	33.9	33.9
+ Balance	49.6	30.4	37.7	32.0	32.8	32.4
+ Faction	58.9	30.2	39.9	37.6	33.9	35.6

Table 8: ILP constraints ablation study.

	Positive			Negative		
	P	R	F1	P	R	F1
All	34.5	39.7	36.9	35.7	37.6	36.6
- Depend.	32.9	32.1	32.5	31.7	38.5	34.8
- Doc.	32.6	41.0	35.8	39.4	23.8	28.0
- Quotation	33.6	39.5	36.3	34.5	34.6	34.6

Table 9: Pairwise classifier feature ablation study.

ing identification of holders, targets, and sentiment jointly. While the numbers are not directly comparable, the best performing system for KBP 2014 sentiment task achieved F1 score of 25.7.

The first row (KM_Gold) shows the comparison against gold annotations from different datasets, highlighting the differences between the task definitions. Our annotations are much more dense, while KBP focuses on specific query entities and MPQA has a much broader focus with less emphasis on covering all entity pairs. The high precision suggests that all of the approaches agree when considering the same entity pairs.

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 unbaised 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.

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%

Table 10: Error Analysis on the development set.

We also did ablation studies to measure the contributions of different components. Table 8 shows ablations of each soft constraint. The faction 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

Zvi Ben-Ami, Ronen Feldman, and Binyamin Rosenfeld. 2014. Entities' sentiment relevance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the National Conference on Artificial Intelligence*.

Lingjia Deng and Janyce Wiebe. 2014. Sentiment propagation via implicature constraints. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.

Lingjia Deng and Janyce Wiebe. 2015a. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lingjia Deng and Janyce Wiebe. 2015b. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*.

Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. 2014. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *Proceedings of International Conference on Computational Linguistics*.

Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 2014. Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results. In *Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology*, pages 17–18.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alvin W. Gouldner. 1960. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2).

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, June. Association for Computational Linguistics.

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

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the ACM international conference on Web search and data mining*. ACM.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of North American Association for Computational Linguistics*.

G. A. Johnston. 1916. *International Journal of Ethics*, 26(2).

Vinodh Krishnan and Jacob Eisenstein. 2015. "You're Mr. Lebowski, I'm The Dude": Inducing address term formality in signed social networks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Paul F Lazarsfeld and Robert K Merton. 1954. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18:18–66.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Bin Lu. 2010. Identifying opinion holders and targets with dependency parser in chinese news texts. In *Proceedings of the NAACL HLT Student Research Workshop*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10.

D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of Conference on Natural Language Learning*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the National Conference on Artificial Intelligence*.

Veselin Stoyanov and Claire Cardie. 2011. Automatically Creating General-Purpose Opinion Summaries from Text. In *Proceedings of Recent Advances in Natural Language Processing*.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of Knowldege Discovery and Data Mining*.

Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *In the Proceedings of Transactions of the Association for Computational Linguistics*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing*.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of Association for Computational Linguistics*.

Bishan Yang and Tom Mitchell. 2016. Joint extraction of events and entities within a document context. In *North American Association for Computational Linguistics*.

Lei Zhang and Bing Liu. 2011. Identifying noun product features that imply opinions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.