Joint Extraction of Entities and Relations for Opinion Recognition

Yejin Choi and Eric Breck and Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY 14853
{ychoi,ebreck,cardie}@cs.cornell.edu

Abstract

We present an approach for the joint extraction of entities and relations in the context of opinion recognition and analysis. We identify two types of opinion-related entities — expressions of opinions and sources of opinions — along with the linking relation that exists between them. Inspired by Roth and Yih (2004), we employ an integer linear programming approach to solve the joint opinion recognition task, and show that global, constraint-based inference can significantly boost the performance of both relation extraction and the extraction of opinion-related entities. Performance further improves when a semantic role labeling system is incorporated. The resulting system achieves F-measures of 79 and 69 for entity and relation extraction, respectively, improving substantially over prior results in the area.

1 Introduction

Information extraction tasks such as recognizing entities and relations have long been considered critical to many domain-specific NLP tasks (e.g. Mooney and Bunescu (2005), Prager et al. (2000), White et al. (2001)). Researchers have further shown that *opinion-oriented information extraction* can provide analogous benefits to a variety of practical applications including product reputation tracking (Morinaga et al., 2002), opinion-oriented question answering (Stoyanov et al., 2005), and opinion-oriented summarization (e.g. Cardie et al. (2004), Liu et al. (2005)). Moreover, much progress has been made in the area of opinion extraction: it is possible to identify sources of opinions (i.e. the opinion holders) (e.g. Choi et al. (2005) and Kim and Hovy (2005b)), to determine the polarity and strength of opinion expressions (e.g. Wilson et al. (2005)), and to recognize propositional opinions and their sources (e.g. Bethard et al. (2004)) with reasonable accuracy. To date, however, there has been no effort to simultaneously identify arbitrary opinion expressions, their sources, and the relations between them. Without progress on the *joint extraction of opinion entities **and** their relations*, the capabilities of opinion-based applications will remain limited.

Fortunately, research in machine learning has produced methods for global inference and joint classification that can help to address this deficiency (e.g. Bunescu and Mooney (2004), Roth and Yih (2004)). Moreover, it has been shown that exploiting dependencies among entities and/or relations via global inference not only solves the joint extraction task, but often boosts performance on the individual tasks when compared to classifiers that handle the tasks independently — for semantic role labeling (e.g. Punyakanok et al. (2004)), information extraction (e.g. Roth and Yih (2004)), and sequence tagging (e.g. Sutton et al. (2004)).

In this paper, we present a global inference approach (Roth and Yih, 2004) to the extraction of opinion-related entities and relations. In particular, we aim to identify two types of entities (i.e. spans of text): entities that express opinions and entities that denote sources of opinions. More specifically, we use the term *opinion expression* to denote all direct expressions of subjectivity including opinions, emotions, beliefs, sentiment, etc., as well as all speech expressions that introduce subjective propositions; and use the term *source* to denote the person or entity (e.g. a re-

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 431–439,
Sydney, July 2006.

port) that holds the opinion.[1] In addition, we aim to identify the relations between opinion expression entities and source entities. That is, for a given opinion expression O_i and source entity S_j, we determine whether the relation $L_{i,j} \stackrel{\text{def}}{=} (S_j \text{ expresses } O_i)$ obtains, i.e. whether S_j is the source of opinion expression O_i. We refer to this particular relation as the *link* relation in the rest of the paper. Consider, for example, the following sentences:

S1. [*Bush*](1) intends(1) to curb the increase in harmful gas emissions and is counting on(1) the good will(2) of [*US industrialists*](2).

S2. By questioning(3) [*the Imam*](4)'s edict(4) [*the Islamic Republic of Iran*](3) made [*the people of the world*](5) understand(5)...

The underlined phrases above are opinion expressions and phrases marked with square brackets are source entities. The numeric superscripts on entities indicate link relations: a source entity and an opinion expression with the same number satisfy the link relation. For instance, the source entity *"Bush"* and the opinion expression *"intends"* satisfy the link relation, and so do *"Bush"* and *"counting on."* Notice that a sentence may contain more than one link relation, and link relations are not one-to-one mappings between sources and opinions. Also, the pair of entities in a link relation may not be the closest entities to each other, as is the case in the second sentence, between *"questioning"* and *"the Islamic Republic of Iran."*

We expect the extraction of opinion relations to be critical for many opinion-oriented NLP applications. For instance, consider the following question that might be given to a question-answering system:

- What is *the Imam*'s opinion toward *the Islamic Republic of Iran*?

Without in-depth opinion analysis, the question-answering system might mistake example S2 as relevant to the query, even though S2 exhibits the opinion of the Islamic Republic of Iran toward Imam, not the other way around.

Inspired by Roth and Yih (2004), we model our task as global, constraint-based inference over separately trained entity and relation classifiers. In particular, we develop three base classifiers: two sequence-tagging classifiers for the extraction of opinion expressions and sources, and a binary classifier to identify the link relation. The global inference procedure is implemented via integer linear programming (ILP) to produce an optimal and coherent extraction of entities and relations.

Because many (60%) opinion-source relations appear as predicate-argument relations, where the predicate is a verb, we also hypothesize that semantic role labeling (SRL) will be very useful for our task. We present two baseline methods for the joint opinion-source recognition task that use a state-of-the-art SRL system (Punyakanok et al., 2005), and describe two additional methods for incorporating SRL into our ILP-based system.

Our experiments show that the global inference approach not only improves relation extraction over the base classifier, but does the same for individual entity extractions. For source extraction in particular, our system achieves an F-measure of 78.1, significantly outperforming previous results in this area (Choi et al., 2005), which obtained an F-measure of 69.4 on the same corpus. In addition, we achieve an F-measure of 68.9 for link relation identification and 82.0 for opinion expression extraction; for the latter task, our system achieves human-level performance.[2]

2 High-Level Approach and Related Work

Our system operates in three phases.

Opinion and Source Entity Extraction We begin by developing two separate token-level sequence-tagging classifiers for opinion expression extraction and source extraction, using linear-chain Conditional Random Fields (CRFs) (Lafferty et al., 2001). The sequence-tagging classifiers are trained using only local syntactic and lexical information to extract each type of entity without knowledge of any nearby or neighboring entities or relations. We collect n-best sequences from each sequence tagger in order to boost the recall of the final system.

Link Relation Classification We also develop a relation classifier that is trained and tested on all pairs of opinion and source entities extracted from the aforementioned n-best opinion expression and source sequences. The relation classifier is modeled using Markov order-0 CRFs(Lafferty

[1] See Wiebe et al. (2005) for additional details.

[2] Wiebe et al. (2005) reports human annotation agreement for opinion expression as 82.0 by F1 measure.

et al., 2001), which are equivalent to maximum entropy models. It is trained using only local syntactic information potentially useful for connecting a pair of entities, but has no knowledge of nearby or neighboring extracted entities and link relations.

Integer Linear Programming Finally, we formulate an integer linear programming problem for each sentence using the results from the previous two phases. In particular, we specify a number of soft and hard constraints among relations and entities that take into account the confidence values provided by the supporting entity and relation classifiers, and that encode a number of heuristics to ensure coherent output. Given these constraints, global inference via ILP finds the optimal, coherent set of opinion-source pairs by exploiting mutual dependencies among the entities and relations.

While good performance in entity or relation extraction can contribute to better performance of the final system, this is not always the case. Punyakanok et al. (2004) notes that, in general, it is better to have high recall from the classifiers included in the ILP formulation. For this reason, it is not our goal to directly optimize the performance of our opinion and source entity extraction models or our relation classifier.

The rest of the paper is organized as follows. Related work is outlined below. Section 3 describes the components of the first phase of our system, the opinion and source extraction classifiers. Section 4 describes the construction of the link relation classifier for phase two. Section 5 describes the ILP formulation to perform global inference over the results from the previous two phases. Experimental results that compare our ILP approach to a number of baselines are presented in Section 6. Section 7 describes how SRL can be incorporated into our global inference system to further improve the performance. Final experimental results and discussion comprise Section 8.

Related Work The definition of our source-expresses-opinion task is similar to that of Bethard et al. (2004); however, our definition of opinion and source entities are much more extensive, going beyond single sentences and propositional opinion expressions. In particular, we evaluate our approach with respect to (1) a wide variety of opinion expressions, (2) explicit and implicit[3] sources, (3) multiple opinion-source link relations per sentence, and (4) link relations that span more than one sentence. In addition, the link relation model explicitly exploits mutual dependencies among entities and relations, while Bethard et al. (2004) does not directly capture the potential influence among entities.

Kim and Hovy (2005b) and Choi et al. (2005) focus only on the extraction of sources of opinions, without extracting opinion expressions. Specifically, Kim and Hovy (2005b) assume a priori existence of the opinion expressions and extract a single source for each, while Choi et al. (2005) do not explicitly extract opinion expressions nor link an opinion expression to a source even though their model implicitly learns approximations of opinion expressions in order to identify opinion sources. Other previous research focuses only on the extraction of opinion expressions (e.g. Kim and Hovy (2005a), Munson et al. (2005) and Wilson et al. (2005)), omitting source identification altogether.

There have also been previous efforts to simultaneously extract entities and relations by exploiting their mutual dependencies. Roth and Yih (2002) formulated global inference using a Bayesian network, where they captured the influence between a relation and a pair of entities via the conditional probability of a relation, given a pair of entities. This approach however, could not exploit dependencies between relations. Roth and Yih (2004) later formulated global inference using integer linear programming, which is the approach that we apply here. In contrast to our work, Roth and Yih (2004) operated in the domain of factual information extraction rather than opinion extraction, and assumed that the exact boundaries of entities from the gold standard are known a priori, which may not be available in practice.

3 Extraction of Opinion and Source Entities

We develop two separate sequence tagging classifiers for opinion extraction and source extraction, using linear-chain Conditional Random Fields (CRFs) (Lafferty et al., 2001). The sequence tagging is encoded as the typical 'BIO' scheme.[4] Each training or test instance represents a sentence, encoded as a linear chain of tokens and their

[3] *Implicit* sources are those that are not explicitly mentioned. See Section 8 for more details.

[4] 'B' is for the token that **b**egins an entity, 'I' is for tokens that are **i**nside an entity, and 'O' is for tokens **o**utside an entity.

associated features. Our feature set is based on that of Choi et al. (2005) for source extraction[5], but we include additional lexical and WordNet-based features. For simplicity, we use the same features for opinion entity extraction and source extraction, and let the CRFs learn appropriate feature weights for each task.

3.1 Entity extraction features

For each token x_i, we include the following features. For details, see Choi et al. (2005).

word: words in a [-4, +4] window centered on x_i.

part-of-speech: POS tags in a [-2, +2] window.[6]

grammatical role: grammatical role (subject, object, prepositional phrase types) of x_i derived from a dependency parse.[7]

dictionary: whether x_i is in the opinion expression dictionary culled from the training data and augmented by approximately 500 opinion words from the MPQA Final Report[8]. Also computed for tokens in a [-1, +1] window and for x_i's parent "chunk" in the dependency parse.

semantic class: x_i's semantic class.[9]

WordNet: the WordNet hypernym of x_i.[10]

4 Relation Classification

We also develop a maximum entropy binary classifier for opinion-source *link* relation classification. Given an opinion-source pair, O_i-S_j, the relation classifier decides whether the pair exhibits a valid link relation, $L_{i,j}$. The relation classifier focuses only on the syntactic structure and lexical properties between the two entities of a given pair, without knowing whether the proposed entities are correct. Opinion and source entities are taken from the n-best sequences of the entity extraction models; therefore, some are invariably incorrect.

From each sentence, we create training and test instances for all possible opinion-source pairings that do not overlap: we create an instance for $L_{i,j}$ only if the span of O_i and S_j do not overlap.

For training, we also filter out instances for which neither the proposed opinion nor source entity overlaps with a correct opinion or source entity per the gold standard. This training instance filtering helps to avoid confusion between examples like the following (where entities marked in bold are the gold standard entities, and entities in square brackets represent the n-best output sequences from the entity extraction classifiers):

(1) [**The president**]_s_1 walked away from [the meeting]_o_1, [[**revealing**]_o_2 **his disappointment**]_o_3 with the deal.

(2) [The monster]_s_2 walked away, [revealing]_o_4 a little box hidden underneath.

For these sentences, we construct training instances for $L_{1,1}$, $L_{1,2}$, and $L_{1,3}$, but not $L_{2,4}$, which in fact has very similar sentential structure as $L_{1,2}$, and hence could confuse the learning algorithm.

4.1 Relation extraction features

The training and test instances for each (potential) link $L_{i,j}$ (with opinion candidate entity O_i and source candidate entity S_j) include the following features.

opinion entity word: the words contained in O_i.

phrase type: the syntactic category of the constituent in which the entity is embedded, e.g. NP or VP. We encode separate features for O_i and S_j.

grammatical role: the grammatical role of the constituent in which the entity is embedded. Grammatical roles are derived from dependency parse trees, as done for the entity extraction classifiers. We encode separate features for O_i and S_j.

position: a boolean value indicating whether S_j precedes O_i.

distance: the distance between O_i and S_j in numbers of tokens. We use four coarse categories: adjacent, very near, near, far.

dependency path: the path through the dependency tree from the head of S_j to the head of O_i. For instance, 'subj↑verb' or 'subj↑verb↓obj'.

voice: whether the voice of O_i is passive or active.

syntactic frame: key intra-sentential relations between O_i and S_j. The syntactic frames that we use are:

- [E_1:role]_[distance]_[E_2:role], where distance $\in$ {adjacent, very near, near, far}, and E_i:role is the grammatical role of E_i. Either E_1 is an opinion entity and E_2 is a source, or vice versa.
- [E_1:phrase]_[distance]_[E_2:phrase], where E_i:phrase is the phrasal type of entity E_i.

[5] We omit only the extraction pattern features.

[6] Using GATE: `http://gate.ac.uk/`

[7] Provided by Rebecca Hwa, based on the Collins parser: `ftp://ftp.cis.upenn.edu/pub/mcollins/PARSER.tar.gz`

[8] `https://rrc.mitre.org/pubs/mpqaFinalReport.pdf`

[9] Using SUNDANCE: (`http://www.cs.utah.edu/řiloff/publications.html#sundance`)

[10] `http://wordnet.princeton.edu/`

- [E_1:phrase]_[E_2:headword], where E_2 must be the opinion entity, and E_1 must be the source entity (i.e. no lexicalized frames for sources). E_1 and E_2 can be contiguous.
- [E_1:role]_[E_2:headword], where E_2 must be the opinion entity, and E_1 must be the source entity.
- [E_1:phrase]_NP_[E_2:phrase] indicates the presence of specific syntactic patterns, e.g. 'VP_NP_VP' depending on the possible phrase types of opinion and source entities. The three phrases do not need to be contiguous.
- [E_1:phrase]_VP_[E_2:phrase] (See above.)
- [E_1:phrase]_[wh-word]_[E_2:phrase] (See above.)
- Src_[distance]_[x]_[distance]_Op, where $x \in$ {by, of, from, for, between, among, and, have, be, will, not,], ", ... }.

When a syntactic frame is matched to a sentence, the bracketed items should be instantiated with particular values corresponding to the sentence. Pattern elements without square brackets are constants. For instance, the syntactic frame '[E_1:phrase]_NP_[E_2:phrase]' may be instantiated as 'VP_NP_VP'. Some frames are lexicalized with respect to the head of an opinion entity to reflect the fact that different verbs expect source entities in different argument positions (e.g. SOURCE *blamed* TARGET vs. TARGET *angered* SOURCE).

5 Integer Linear Programming Approach

As noted in the introduction, we model our task as global, constraint-based inference over the separately trained entity and relation classifiers, and implement the inference procedure as binary integer linear programming (ILP) ((Roth and Yih, 2004), (Punyakanok et al., 2004)). ILP consists of an objective function which is a dot product between a vector of variables and a vector of weights, and a set of equality and inequality constraints among variables. Given an objective function and a set of constraints, LP finds the optimal assignment of values to variables, i.e. one that minimizes the objective function. In binary ILP, the assignments to variables must be either 0 or 1. The variables and constraints defined for the opinion recognition task are summarized in Table 1 and explained below.

Entity variables and weights For each opinion entity, we add two variables, O_i and $\bar{O}_i$, where $O_i = 1$ means to extract the opinion entity, and

Objective function f $= \sum_i (w_{o_i} O_i) + \sum_i (\bar{w}_{o_i} \bar{O}_i)$ $+ \sum_j (w_{s_j} S_j) + \sum_j (\bar{w}_{s_j} \bar{S}_j)$ $+ \sum_{i,j} (w_{l_{i,j}} L_{i,j}) + \sum_i (\bar{w}_{l_{i,j}} \bar{L}_{i,j})$
$\forall i,\ O_i + \bar{O}_i = 1$ $\forall j,\ S_j + \bar{S}_j = 1$ $\forall i, j,\ L_{i,j} + \bar{L}_{i,j} = 1$
$\forall i,\ O_i = \sum_j L_{i,j}$ $\forall j,\ S_j + A_j = \sum_i L_{i,j}$ $\forall j,\ A_j - S_j \leq 0$
$\forall i, j, i < j,\ X_i + X_j = 1, X \in \{S, O\}$

Table 1: Binary ILP formulation

$\bar{O}_i = 1$ means to discard the opinion entity. To ensure coherent assignments, we add equality constraints $\forall i,\ O_i + \bar{O}_i = 1$. The weights w_{o_i} and $\bar{w}_{o_i}$ for O_i and $\bar{O}_i$ respectively, are computed as a negative conditional probability of the span of an entity to be extracted (or suppressed) given the labelings of the adjacent variables of the CRFs:

$$w_{o_i} \stackrel{\text{def}}{=} -\text{P}\ (x_k, x_{k+1}, ..., x_l | x_{k-1}, x_{l+1})$$
$$\text{where } x_k = \text{'B'}$$
$$\& \ x_m = \text{'I' for } m \in [k+1, l]$$
$$\bar{w}_{o_i} \stackrel{\text{def}}{=} -\text{P}\ (x_k, x_{k+1}, ..., x_l | x_{k-1}, x_{l+1})$$
$$\text{where } x_m = \text{'O' for } m \in [k, l]$$

where x_i is the value assigned to the random variable of the CRF corresponding to an entity O_i. Likewise, for each source entity, we add two variables S_j and $\bar{S}_j$ and a constraint $S_j + \bar{S}_j = 1$. The weights for source variables are computed in the same way as opinion entities.

Relation variables and weights For each link relation, we add two variables $L_{i,j}$ and $\bar{L}_{i,j}$, and a constraint $L_{i,j} + \bar{L}_{i,j} = 1$. By the definition of a link, if $L_{i,j} = 1$, then it is implied that $O_i = 1$ and $S_j = 1$. That is, if a link is extracted, then the pair of entities for the link must be also extracted. Constraints to ensure this coherency are explained in the following subsection. The weights for link variables are based on probabilities from the binary link classifier.

Constraints for link coherency In our corpus, a source entity can be linked to more than one opinion entity, but an opinion entity is linked to only

one source. Nonetheless, the majority of opinion-source pairs involve one-to-one mappings, which we encode as hard and soft constraints as follows:

For each opinion entity, we add an equality constraint $O_i = \sum_j L_{i,j}$ to enforce that only one link can emanate from an opinion entity. For each source entity, we add an equality constraint and an inequality constraint that together allow a source to link to at most two opinions: $S_j + A_j = \sum_i L_{i,j}$ and $A_j - S_j \leq 0$, where A_j is an auxiliary variable, such that its weight is some positive constant value that suppresses A_j from being assigned to 1. And A_j can be assigned to 1 only if S_j is already assigned to 1. It is possible to add more auxiliary variables to allow more than two opinions to link to a source, but for our experiments two seemed to be a reasonable limit.

Constraints for entity coherency When we use n-best sequences where $n > 1$, proposed entities can overlap. Because this should not be the case in the final result, we add an equality constraint $X_i + X_j = 1,\ X \in \{S, O\}$ for all pairs of entities with overlapping spans.

Adjustments to weights To balance the precision and recall, and to take into account the performance of different base classifiers, we apply adjustments to weights as follows.

1) We define six coefficients c_x and $\bar{c}_x$, where $x \in \{O, S, L\}$ to modify a group of weights as follows.
   $\forall i, x,\ w_{x_i} := w_{x_i} * c_x$;
   $\forall i, x,\ \bar{w}_{x_i} := \bar{w}_{x_i} * \bar{c}_x$;
   In general, increasing c_x will promote recall, while increasing $\bar{c}_x$ will promote precision. Also, setting $c_o > c_s$ will put higher confidence on the opinion extraction classifier than the source extraction classifier.
2) We also define one constant c_A to set the weights for auxiliary variable A_i. That is, $\forall i,\ w_{A_i} := c_A$.
3) Finally, we adjust the confidence of the link variable based on n-th-best sequences of the entity extraction classifiers as follows.
   $\forall i,\ w_{L_{i,j}} := w_{L_{i,j}} * d$
   where $d \stackrel{\text{def}}{=} 4/(3 + min(m, n))$, when O_i is from an m-th sequence and S_j is from a n-th sequence.[11]

[11] This will smoothly degrade the confidence of a link based on the entities from higher n-th sequences. Values of d decrease as $4/4,\ 4/5,\ 4/6,\ 4/7$....

6 Experiments–I

We evaluate our system using the NRRC Multi-Perspective Question Answering (MPQA) corpus that contains 535 newswire articles that are manually annotated for opinion-related information. In particular, our gold standard opinion entities correspond to *direct subjective expression* annotations and *subjective speech event* annotations (i.e. speech events that introduce opinions) in the MPQA corpus (Wiebe et al., 2005). Gold standard source entities and link relations can be extracted from the *agent* attribute associated with each opinion entity. We use 135 documents as a development set and report 10-fold cross validation results on the remaining 400 documents in all experiments below.

We evaluate entity and link extraction using both an *overlap* and *exact* matching scheme.[12] Because the exact start and endpoints of the manual annotations are somewhat arbitrary, the overlap scheme is more reasonable for our task (Wiebe et al., 2005). We report results according to both matching schemes, but focus our discussion on results obtained using overlap matching.[13]

We use the Mallet[14] implementation of CRFs. For brevity, we will refer to the opinion extraction classifier as CRF-OP, the source extraction classifier as CRF-SRC, and the link relation classifier as CRF-LINK. For ILP, we use Matlab, which produced the optimal assignment in a matter of few seconds for each sentence. The weight adjustment constants defined for ILP are based on the development data.[15]

The link-nearest baselines For baselines, we first consider a *link-nearest* heuristic: for each opinion entity extracted by CRF-OP, the link-nearest heuristic creates a link relation with the closest source entity extracted by CRF-SRC. Recall that CRF-SRC and CRF-OP extract entities from n-best sequences. We test the link-nearest heuristic with $n = \{1, 2, 10\}$ where larger n will boost recall at the cost of precision. Results for the

[12] Given two links $L_{1,1} = (O_1, S_1)$ and $L_{2,2} = (O_2, S_2)$, exact matching requires the spans of O_1 and O_2, and the spans of S_1 and S_2, to match exactly, while overlap matching requires the spans to overlap.

[13] Wiebe et al. (2005) also reports the human annotation agreement study via the overlap scheme.

[14] Available at http://mallet.cs.umass.edu

[15] $c_o = 2.5, \bar{c}_o = 1.0, c_s = 1.5, \bar{c}_s = 1.0, c_L = 2.5, \bar{c}_L = 2.5, c_A = 0.2$. Values are picked so as to boost recall while reasonably suppressing incorrect links.

	Overlap Match			Exact Match		
	r(%)	p(%)	f(%)	r(%)	p(%)	f(%)
NEAREST-1	51.6	71.4	59.9	26.2	36.9	30.7
NEAREST-2	60.7	45.8	52.2	29.7	19.0	23.1
NEAREST-10	66.3	20.9	31.7	28.2	00.0	00.0
SRL	59.7	36.3	45.2	32.6	19.3	24.2
SRL+CRF-OP	45.6	83.2	58.9	27.6	49.7	35.5
ILP-1	51.6	80.8	63.0	26.4	42.0	32.4
ILP-10	64.0	72.4	68.0	31.0	34.8	32.8

Table 2: Relation extraction performance

NEAREST-n : link-nearest heuristic w/ n-best
SRL : all V-A0 frames from SRL
SRL+CRF-OP : all V-A0 filtered by CRF-OP
ILP-n : ILP applied to n-best sequences

link-nearest heuristic on the full source-expresses-opinion relation extraction task are shown in the first three rows of table 2. NEAREST-1 performs the best in overlap-match F-measure, reaching 59.9. NEAREST-10 has higher recall (66.3%), but the precision is really low (20.9%). Performance of the opinion and source entity classifiers will be discussed in Section 8.

SRL baselines Next, we consider two baselines that use a state-of-the-art SRL system (Punyakanok et al., 2005). In many link relations, the opinion expression entity is a verb phrase and the source entity is in an agent argument position. Hence our second baseline, SRL, extracts all verb(V)-agent(A0) frames from the output of the SRL system and provides an upper bound on recall (59.7%) for systems that use SRL in isolation for our task. A more sophisticated baseline, SRL+CRF-OP, extracts only those V-A0 frames whose verb overlaps with entities extracted by the opinion expression extractor, CRF-OP. As shown in table 2, filtering out V-A0 frames that are incompatible with the opinion extractor boosts precision to 83.2%, but the F-measure (58.9) is lower than that of NEAREST-1.

ILP results The ILP-n system in table 2 denotes the results of the ILP approach applied to the n-best sequences. ILP-10 reaches an F-measure of 68.0, a significant improvement over the highest performing baseline[16], and also a substantial improvement over ILP-1. Note that the performance of NEAREST-10 was much worse than that

[16]Statistically significant by paired-t test, where $p < 0.001$.

	Overlap Match			Exact Match		
	r(%)	p(%)	f(%)	r(%)	p(%)	f(%)
ILP-1	51.6	80.8	63.0	26.4	42.0	32.4
ILP-10	64.0	72.4	68.0	31.0	34.8	32.8
ILP+SRL-f-1	51.7	81.5	63.3	26.6	42.5	32.7
ILP+SRL-f-10	65.7	72.4	68.9	31.5	34.3	32.9
ILP+SRL-fc-10	64.0	73.5	68.4	28.4	31.3	29.8

Table 3: Relation extraction with ILP and SRL

ILP-n : ILP applied to n-best sequences
ILP+SRL-f-n : ILP w/ SRL features, n-best
ILP+SRL-fc-n : ILP w/ SRL features, and SRL constraints, n-best

of NEAREST-1, because the 10-best sequences include many incorrect entities whereas the corresponding ILP formulation can discard the bad entities by considering dependencies among entities and relations.[17]

7 Additional SRL Incorporation

We next explore two approaches for more directly incorporating SRL into our system.

Extra SRL Features for the Link classifier We incorporate SRL into the link classifier by adding extra features based on SRL. We add boolean features to check whether the span of an SRL argument and an entity matches exactly. In addition, we include **syntactic frame** features as follows:

- [E_1:srl-arg]_[E_2:srl-arg], where E_i:srl-arg indicates the SRL argument type of entity E_i.
- [E_1.srl-arg]_[E_1:headword]_[E_2:srl-arg], where E_1 must be an opinion entity, and E_2 must be a source entity.

Extra SRL Constraints for the ILP phase We also incorporate SRL into the ILP phase of our system by adding extra constraints based on SRL. In particular, we assign very high weights for links that match V-A0 frames generated by SRL, in order to force the extraction of V-A0 frames.

[17]A potential issue with overlap precision and recall is that the measures may drastically overestimate the system's performance as follows: a system predicting a single link relation whose source and opinion expression both overlap with every token of a document would achieve 100% overlap precision and recall. We can ensure this does not happen by measuring the average number of (source, opinion) pairs to which each correct or predicted pair is aligned (excluding pairs not aligned at all). In our data, this does not exceed 1.08, (except for baselines), so we can conclude these evaluation measures are behaving reasonably.

		Opinion			Source			Link		
		r(%)	p(%)	f(%)	r(%)	p(%)	f(%)	r(%)	p(%)	f(%)
Before ILP	CRF-OP/SRC/LINK with 1 best	76.4	88.4	81.9	67.3	81.9	73.9	60.5	50.5	55.0
	merged 10 best	95.7	31.2	47.0	95.3	24.5	38.9	N/A		
After ILP	ILP-SRL-f-10	75.1	82.9	78.8	80.6	75.7	78.1	65.7	72.4	68.9
	ILP-SRL-f-10 $\cup$ CRF-OP/SRC with 1 best	82.3	81.7	82.0	81.5	73.4	77.3	N/A		

Table 4: Entity extraction performance (by overlap-matching)

8 Experiments–II

Results using SRL are shown in Table 3 (on the previous page). In the table, ILP+SRL-f denotes the ILP approach using the link classifier with the extra SRL 'f'eatures, and ILP+SRL-fc denotes the ILP approach using both the extra SRL 'f'eatures and the SRL 'c'onstraints. For comparison, the ILP-1 and ILP-10 results from Table 2 are shown in rows 1 and 2.

The F-measure score of ILP+SRL-f-10 is 68.9, about a 1 point increase from that of ILP-10, which shows that extra SRL features for the link classifier further improve the performance over our previous best results.[18] ILP+SRL-fc-10 also performs better than ILP-10 in F-measure, although it is slightly worse than ILP+SRL-f-10. This indicates that the link classifier with extra SRL features already makes good use of the V-A0 frames from the SRL system, so that forcing the extraction of such frames via extra ILP constraints only hurts performance by not allowing the extraction of non-V-A0 pairs in the neighborhood that could have been better choices.

Contribution of the ILP phase In order to highlight the contribution of the ILP phase for our task, we present 'before' and 'after' performance in Table 4. The first row shows the performance of the individual CRF-OP, CRF-SRC, and CRF-LINK classifiers before the ILP phase. Without the ILP phase, the 1-best sequence generates the best scores. However, we also present the performance with merged 10-best entity sequences[19] in order to demonstrate that using 10-best sequences without ILP will only hurt performance. The precision of the merged 10-best sequences system is very low, however the recall level is above 95% for both CRF-OP and CRF-SRC, giving an upper bound for recall for our approach. The third row presents results after the ILP phase is applied for the 10-best sequences, and we see that, in addition to the improved link extraction described in Section 7, the performance on source extraction is substantially improved, from F-measure of 73.9 to 78.1. Performance on opinion expression extraction decreases from F-measure of 81.9 to 78.8. This decrease is largely due to *implicit* links, which we will explain below. The fourth row takes the union of the entities from ILP-SRL-f-10 and the entities from the best sequences from CRF-OP and CRF-SRC. This process brings the F-measure of CRF-OP up to 82.0, with a different precision-recall break down from those of 1-best sequences without ILP phase. In particular, the recall on opinion expressions now reaches 82.3%, while maintaining a high precision of 81.7%.

	Overlap Match			Exact Match		
	r(%)	p(%)	f(%)	r(%)	p(%)	f(%)
DEV.CONF	65.7	72.4	68.9	31.5	34.3	32.9
NO.CONF	63.7	76.2	69.4	30.9	36.7	33.5

Table 5: Relation extraction with ILP weight adjustment. (All cases using ILP+SRL-f-10)

Effects of ILP weight adjustment Finally, we show the effect of weight adjustment in the ILP formulation in Table 5. The DEV.CONF row shows relation extraction performance using a weight configuration based from the development data. In order to see the effect of weight adjustment, we ran an experiment, NO.CONF, using fixed default weights.[20] Not surprisingly, our weight adjustment tuned from the development set is not the optimal choice for cross-validation set. Nevertheless, the weight adjustment helps to balance the precision and recall, i.e. it improves recall at the

[18] Statistically significant by paired-t test, where $p < 0.001$.

[19] If an entity E_i extracted by the ith-best sequence overlaps with an entity E_j extracted by the jth-best sequence, where $i < j$, then we discard E_j. If E_i and E_j do not overlap, then we extract both entities.

[20] To be precise, $c_x = 1.0, \bar{c}_x = 1.0$ for $x \in \{O, S, L\}$, but $c_A = 0.2$ is the same as before.

cost of precision. The weight adjustment is more effective when the gap between precision and recall is large, as was the case with the development data.

Implicit links A good portion of errors stem from the *implicit* link relation, which our system did not model directly. An implicit link relation holds for an opinion entity without an associated source entity. In this case, the opinion entity is linked to an *implicit* source. Consider the following example.

- Anti-Soviet hysteria was firmly oppressed.

Notice that opinion expressions such as *"Anti-Soviet hysteria"* and *"firmly oppressed"* do not have associated source entities, because sources of these opinion expressions are not explicitly mentioned in the text. Because our system forces each opinion to be linked with an explicit source entity, opinion expressions that do not have explicit source entities will be dropped during the global inference phase of our system. Implicit links amount to 7% of the link relations in our corpus, so the upper bound for recall for our ILP system is 93%. In the future we will extend our system to handle implicit links as well. Note that we report results against a gold standard that includes implicit links. Excluding them from the gold standard, the performance of our final system ILP+SRL-f-10 is 72.6% in recall, 72.4% in precision, and 72.5 in F-measure.

9 Conclusion

This paper presented a global inference approach to jointly extract entities and relations in the context of opinion oriented information extraction. The final system achieves performance levels that are potentially good enough for many practical NLP applications.

Acknowledgments We thank the reviewers for their many helpful comments and Vasin Punyakanok for running our data through his SRL system. This work was supported by the Advanced Research and Development Activity (ARDA), by NSF Grants IIS-0535099 and IIS-0208028, and by gifts from Google and the Xerox Foundation.

References

S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou and D. Jurafsky 2004. Automatic Extraction of Opinion Propositions and their Holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text.*

R. Bunescu and R. J. Mooney 2004. Collective Information Extraction with Relational Markov Networks. In *ACL*.

C. Cardie, J. Wiebe, T. Wilson and D. Litman 2004. Low-Level Annotations and Summary Representations of Opinions for Multi-Perspective Question Answering. *New Directions in Question Answering.*

Y. Choi, C. Cardie, E. Riloff and S. Patwardhan 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *HLT-EMNLP*.

S. Kim and E. Hovy 2005. Automatic Detection of Opinion Bearing Words and Sentences. In *IJCNLP*.

S. Kim and E. Hovy 2005. Identifying Opinion Holders for Question Answering in Opinion Texts. In *AAAI Workshop on Question Answering in Restricted Domains*.

J. Lafferty, A. K. McCallum and F. Pereira 2001 Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*.

B. Liu, M. Hu and J. Cheng 2005 Opinion Observer: Analyzing and Comparing Opinions on the Web. In *WWW*.

R. J. Mooney and R. Bunescu 2005 Mining Knowledge from Text Using Information Extraction. In *SIGKDD Explorations*.

S. Morinaga, K. Yamanishi, K. Tateishi and T. Fukushima 2002. Mining product reputations on the Web. In *KDD*.

M. A. Munson, C. Cardie and R. Caruana. 2005. Optimizing to arbitrary NLP metrics using ensemble selection. In *HLT-EMNLP*.

J. Prager, E. Brown, A. Coden and D. Radev 2000. Question-answering by predictive annotation. In *SIGIR*.

V. Punyakanok, D. Roth and W. Yih 2005. Generalized Inference with Multiple Semantic Role Labeling Systems (Shared Task Paper). In *CoNLL*.

V. Punyakanok, D. Roth, W. Yih and D. Zimak 2004. Semantic Role Labeling via Integer Linear Programming Inference. In *COLING*.

D. Roth and W. Yih 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *CoNLL*.

D. Roth and W. Yih 2002. Probabilistic Reasoning for Entity and Relation Recognition. In *COLING*.

V. Stoyanov, C. Cardie and J. Wiebe 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *HLT-EMNLP*.

C. Sutton, K. Rohanimanesh and A. K. McCallum 2004. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. In *ICML*.

M. White, T. Korelsky, C. Cardie, V. Ng, D. Pierce and K. Wagstaff 2001. Multi-document Summarization via Information Extraction In *HLT*.

J. Wiebe and T. Wilson and C. Cardie 2005. Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation, volume 39, issue 2-3*.

T. Wilson, J. Wiebe and P. Hoffmann 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *HLT-EMNLP*.