Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification

Swapna Somasundaran
Univ. of Pittsburgh
Pittsburgh, PA 15260
swapna@cs.pitt.edu

Galileo Namata
Univ. of Maryland
College Park, MD 20742
namatag@cs.umd.edu

Janyce Wiebe
Univ. of Pittsburgh
Pittsburgh, PA 15260
wiebe@cs.pitt.edu

Lise Getoor
Univ. of Maryland
College Park, MD 20742
getoor@cs.umd.edu

Abstract

This work investigates design choices in modeling a discourse scheme for improving opinion polarity classification. For this, two diverse global inference paradigms are used: a supervised collective classification framework and an unsupervised optimization framework. Both approaches perform substantially better than baseline approaches, establishing the efficacy of the methods and the underlying discourse scheme. We also present quantitative and qualitative analyses showing how the improvements are achieved.

1 Introduction

The importance of discourse in opinion analysis is being increasingly recognized (Polanyi and Zaenen, 2006). Motivated by the need to enable discourse-based opinion analysis, previous research (Asher et al., 2008; Somasundaran et al., 2008) developed discourse schemes and created manually annotated corpora. However, it was not known whether and how well these linguistic ideas and schemes can be translated into effective computational implementations.

In this paper, we first investigate ways in which an opinion discourse scheme can be computationally modeled, and then how it can be utilized to improve polarity classification. Specifically, the discourse scheme we use is from Somasundaran et al. (2008), which was developed to support a global, interdependent polarity interpretation. To achieve discourse-based global inference, we explore two different frameworks. The first is a supervised framework that learns interdependent opinion interpretations from training data. The second is an unsupervised optimization framework which uses constraints to express the ideas of coherent opinion interpretation embodied in the scheme. For the supervised framework, we use Iterative Collective Classification (ICA), which facilitates machine learning using relational information. The unsupervised optimization is implemented as an Integer Linear Programming (ILP) problem. Via our implementations, we aim to empirically test if discourse-based approaches to opinion analysis are useful.

Our results show that both of our implementations achieve significantly better accuracies in polarity classification than classifiers using local information alone. This confirms the hypothesis that the discourse-based scheme is useful, and also shows that both of our design choices are effective. We also find that there is a difference in the way ICA and ILP achieve improvements, and a simple hybrid approach, which incorporates the strengths of both, is able to achieve significant overall improvements over both. Our analyses show that even when our discourse-based methods bootstrap from noisy classifications, they can achieve good improvements.

The rest of this paper is organized as follows: we discuss related work in Section 2 and the discourse scheme in Section 3. We present our discourse-based implementations in Section 4, experiments in Section 5, discussions in Section 6 and conclusions in Section 7.

2 Related Work

Previous work on polarity disambiguation has used contextual clues and reversal words (Wilson et al., 2005; Kennedy and Inkpen, 2006; Kanayama and Nasukawa, 2006; Devitt and Ahmad, 2007; Sadamitsu et al., 2008). However, these do not capture discourse-level relations.

Researchers, such as (Polanyi and Zaenen, 2006), have discussed how the discourse structure can influence opinion interpretation; and previous work, such as (Asher et al., 2008; Somasundaran et al., 2008), have developed annotation schemes for interpreting opinions with discourse relations. However, they do not empirically demonstrate how automatic methods can use their ideas to improve polarity classification. In this work, we demonstrate concrete ways in which a discourse-based scheme can be modeled using global inference paradigms.

Joint models have been previously explored for other NLP problems (Haghighi et al., 2005; Moschitti et al., 2006; Moschitti, 2009). Our global inference model focuses on opinion polarity recognition task.

The biggest difference between this work and previous work in opinion analysis that use global inference methods is in the type of linguistic relations used to achieve the global inference. Some of the work is not related to discourse at all (e.g., lexical similarities (Takamura et al., 2007), morphosyntactic similarities (Popescu and Etzioni, 2005) and word-based measures like TF-IDF (Goldberg and Zhu, 2006)). Others use sentence cohesion (Pang and Lee, 2004), agreement/disagreement between speakers (Thomas et al., 2006; Bansal et al., 2008), or structural adjacency. In contrast, our work focuses on discourse-based relations for global inference. Another difference from the above work is that our work is over multi-party conversations.

Previous work on emotion and subjectivity detection in multi-party conversations has explored using prosodic information (Neiberg et al., 2006), combining linguistic and acoustic information (Raaijmakers et al., 2008) and combining lexical and dialog information (Somasundaran et al., 2007). Our work is focused on harnessing discourse-based knowledge and on interdependent inference.

There are several collective classification frameworks, including (Neville and Jensen, 2000; Lu and Getoor, 2003; Taskar et al., 2004; Richardson and Domingos, 2006; Bilgic et al., 2007). In this paper, we use an approach by (Lu and Getoor, 2003) which iteratively predicts class values using local and relational features. ILP has been used on other NLP tasks, e.g., (Denis and Baldridge, 2007; Choi et al., 2006; Roth and Yih, 2004). In this work, we employ ILP for modeling discourse constraints for polarity classification.

3 Discourse Scheme and Data

The scheme in Somasundaran et al. (2008) has been developed and annotated over the AMI meeting corpus (Carletta et al., 2005).[1] This scheme annotates opinions, their polarities (positive, negative, neutral) and their targets (a target is what the opinion is about). The targets of opinions are related via two types of relations: the *same* relation, which relates targets referring to the same entity or proposition, and the *alternative* relation, which relates targets referring to mutually exclusive options in the context of the discourse. Additionally, the scheme relates opinions via two types of *frame* relations: the *reinforcing* and *non-reinforcing* relations. The frame relations represent discourse scenarios: reinforcing relations exist between opinions when they contribute to the same overall stance, while non-reinforcing relations exist between opinions that show ambivalence.

The opinion annotations are text-span based, while in this work, we use Dialog Act (DA) based segmentation of meetings.[2] As the DAs are our units of classification, we map opinion annotations to the DA units as follows. If a DA unit contains an opinion annotation, the label is transferred upwards to the containing DA. When a DA contains multiple opinion annotations, each with a different polarity, one of them is randomly chosen as the label for the DA. The discourse relations existing between opinions are also transferred upwards, between the DAs containing each of these annotations. We recreate an example from Somasundaran et al. (2008) using DA segmentation in Example 1. Here, the speaker has a positive opinion towards the rubbery material for the TV remote.

(1) DA-1: ... this kind of rubbery material,
DA-2: *it's* a **bit more bouncy**,
DA-3: like you said they get chucked around a lot.
DA-4: A **bit more durable** and *that* can also be **ergonomic** and
DA-5: *it* kind of feels **a bit different from all the other remote controls**.

In the example, the individual opinion expressions (shown in bold) are essentially regarding the same thing – the rubbery material. Thus, the explicit targets (shown in *italics*), *it's*, *that*, and *it*, and the implicit target of **a bit more durable** are all linked

[1] The AMI corpus contains a set of scenario-based meetings where participants have to design a new TV remote prototype.

[2] DA segmentation is provided with the AMI corpus.

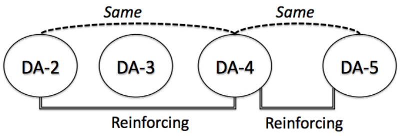

Figure 1: Discourse Relations between DA segments for Example 1.

with *same* target relations. Also, notice that the opinions reinforce a particular stance, i.e., a pro-rubbery-material stance. Thus, the scheme links the opinions via reinforcing relations. Figure 1 illustrates the corresponding discourse relations between the containing DA units.

4 Implementing the Discourse Model

The hypothesis in using discourse information for polarity classification is that the global discourse view will improve upon a classification with only a local view. Thus, we implement a local classifier to bootstrap the classification process, and then implement classifiers that use discourse information from the scheme annotations, over it. We explore two approaches for implementing our discourse-based classifier. The first is ICA, where discourse relations and the neighborhood information brought in by these relations are incorporated as features into the learner. The second approach is ILP optimization, which tries to maximize the class distributions predicted by the local classifier, subject to constraints imposed by discourse relations. Both classifiers thus accommodate preferences of the local classifier and for coherence with discourse neighbors.

4.1 Local Classifier

A supervised local classifier, *Local*, is used to provide the classifications to bootstrap the discourse-based classifiers.[3] It is important to make Local as reliable as possible; otherwise, the discourse relations will propagate misclassifications. Thus, we build Local using a variety of knowledge sources that have been shown to be useful for opinion analysis in previous work. Specifically, we construct features using polarity lexicons (used by (Wilson et al., 2005)), DA tags (used by (Somasundaran et al., 2007)) and unigrams (used by many researchers, e.g., (Pang and Lee, 2004)).

Note that, as our discourse-based classifiers attempt to improve upon the local classifications, Local is also a baseline for our experiments.

4.2 Iterative Collective Classification

We use a variant of ICA (Lu and Getoor, 2003; Neville and Jensen, 2000), which is a collective classification algorithm shown to perform consistently well over a wide variety of relational data.

Algorithm 1 ICA Algorithm

```
for each instance $i$ do {bootstrapping}
  Compute polarity for $i$ using local attributes
end for
repeat {iterative}
  Generate ordering $I$ over all instances
  for each $i$ in $I$ do
    Compute polarity for $i$ using local and relational attributes
  end for
until Stopping criterion is met
```

ICA uses two classifiers: a local classifier and a *relational classifier*. The local classifier is trained to predict the DA labels using only the local features. We use Local, described in Section 4.1, for this purpose. The relational classifier is trained using the local features, and an additional set of features commonly referred to as *relational features*. The value of a relational feature, for a given DA, depends on the polarity of the discourse neighbors of that DA. Thus, the relational features incorporate discourse and neighbor information; that is, they incorporate the information about the frame and target relations in conjunction with the polarity of the discourse neighbors. Intuitively, our motivation for this approach can be explained using Example 1. Here, in interpreting the ambiguous opinion **a bit different** as being positive, we use the knowledge that it participates in a reinforcing discourse, and that all its neighbors (e.g., **ergonomic**, **durable**) are positive opinions regarding the same thing. On the other hand, if it had been a non-reinforcing discourse, then the polarity of **a bit different**, when viewed with respect to the other opinions, could have been interpreted as negative.

Table 1 lists the relational features we defined for our experiments where each row represents a

[3] Local is supervised, as previous work has shown that supervised methods are effective in opinion analysis. Even though this makes the final end-to-end system with the ILP implementation semi-supervised, note that the discourse-based ILP part is itself unsupervised.

Percent of neighbors with polarity type a related via frame relation f'
Percent of neighbors with polarity type a related via target relation t'
Percent of neighbors with polarity type a related via frame relation f and target relation t
Percent of neighbors with polarity type a and same speaker related via frame relation f'
Percent of neighbors with polarity type a and same speaker related via target relation t'
Percent of neighbors with polarity type a related via a frame relation or target relation
Percent of neighbors with polarity type a related via a reinforcing frame relation or *same* target relation
Percent of neighbors with polarity type a related via a non-reinforcing frame relation or alt target relation
Most common polarity type of neighbors related via a *same* target relation
Most common polarity type of neighbors related via a reinforcing frame relation and *same* target relation

Table 1: Relational features: $a \in$ {non-neutral (i.e., positive or negative), positive, negative}, $t \in$ {same, alt}, $f \in$ {reinforcing, non-reinforcing}, $t' \in$ {same or alt, same, alt}, $f' \in$ {reinforcing or non-reinforcing, reinforcing, non-reinforcing}

set of features. Features are generated for all combinations of a, t, t', f and f' for each row. For example, one of the features in the first row is *Percent of neighbors with polarity type positive, that are related via a reinforcing frame relation.* Thus, each feature for the relational classifier identifies neighbors for a given instance via a specific relation (f, t, f' or t', obtained from the scheme annotations) and factors in their polarity values (a, obtained from the classifier predictions from the previous round). This adds a total of 59 relational features to the already existing local features.

ICA has two main phases: the bootstrapping and iterative phases. In the bootstrapping phase, the polarity of each instance is initialized to the most likely value given only the local classifier and its features. In the iterative phase, we create a random ordering of all the instances and, in turn, apply the relational classifier to each instance where the relational features, for a given instance, are computed using the most recent polarity assignments of its neighbors. We repeat this until some stopping criterion is met. For our experiments, we use a fixed number of 30 iterations, which has been found to be sufficient in most data sets for ICA to converge to a solution.

The pseudocode for the algorithm is shown in Algorithm 1.

4.3 Integer Linear Programming

First, we explain the intuition behind viewing discourse relations as enforcing constraints on polarity interpretation. Then, we explain how the constraints are encoded in the optimization problem.

4.3.1 Discourse Constraints on Polarity

The discourse relations between opinions can provide coherence constraints on the way their polarity is interpreted. Consider a discourse scenario in which a speaker expresses multiple opinions regarding the same thing, and is reinforcing his stance in the process (as in Example 1). The set of individual polarity assignments that is most coherent with this global scenario is the one where all the opinions have the same (*equal*) polarity. On the other hand, a pair of individual polarity assignments most consistent with a discourse scenario where a speaker reinforces his stance via opinions towards alternative options, is one with opinions having mutually *opposite* polarity. For instance, in the utterance "Shapes **should** be *curved*, **nothing** *square-like*", the speaker reinforces his pro-curved stance via his opinions about the alternative shapes: *curved* and *square-like*. And, we see that the first opinion is positive and the second is negative. Table 2 lists the discourse relations (target and frame relation combinations) found in the corpus, and the likely polarity interpretation for the related instances.

Target relation + Frame relation	Polarity
same+reinforcing	equal (e)
same+non-reinforcing	opposite (o)
alternative+reinforcing	opposite (o)
alternative+non-reinforcing	equal (e)

Table 2: Discourse relations and their polarity constraints on the related instances.

4.3.2 Optimization Problem

For each DA instance i in a dataset, the local classifier provides a class distribution $[p_i, q_i, r_i]$, where p_i, q_i and r_i correspond to the probabilities that i belongs to positive, negative and neutral categories, respectively. The optimization problem is formulated as an ILP minimization of the objective function in Equation 1.

$$-1 \times \sum_{i} (p_i x_i + q_i y_i + r_i z_i) + \sum_{i,j} \epsilon_{ij} + \sum_{i,j} \delta_{ij} \quad (1)$$

where the x_i, y_i and z_i are binary *class variables* corresponding to positive, negative and neutral classes, respectively. When a class variable is 1, the corresponding class is chosen. Variables ϵ_{ij} and δ_{ij} are binary *slack variables* that correspond to the discourse constraints between two distinct DA instances i and j. When a given slack variable is 1, the corresponding discourse constraint is violated. Note that the objective function tries to achieve two goals. The first part ($\sum_i p_i x_i + q_i y_i + r_i z_i$) is a maximization that tries to choose a classification for the instances that maximizes the probabilities provided by the local classifier. The second part ($\sum_{i,j} \epsilon_{ij} + \sum_{i,j} \delta_{ij}$) is a minimization that tries to minimize the number of slack variables used, that is, minimize the number of discourse constraints violated.

Constraints in Equations 2 and 3 listed below impose binary constraints on the variables. The constraint in Equation 4 ensures that, for each instance i, only one class variable is set to 1.

$$x_i \in \{0,1\}, y_i \in \{0,1\}, z_i \in \{0,1\} \ , \ \forall i \qquad (2)$$

$$\epsilon_{ij} \in \{0,1\}, \delta_{ij} \in \{0,1\} \ , \ \forall i \neq j \qquad (3)$$

$$x_i + y_i + z_i = 1 \ , \ \forall i \qquad (4)$$

We pair distinct DA instances i and j as ij, and if there exists a discourse relation between them, they can be subject to the corresponding polarity constraints listed in Table 2. For this, we define two binary discourse-constraint constants: the *equal-polarity* constant, e_{ij} and the *opposite-polarity* constant, o_{ij}. If a given DA pair ij is related by either a same+reinforcing relation or an alternative+non-reinforcing relation (rows 1, 4 of Table 2), then $e_{ij} = 1$; otherwise it is zero. Similarly, if it is related by either a same+non-reinforcing relation or an alternative+reinforcing relation (rows 2, 3 of Table 2), then $o_{ij} = 1$. Both e_{ij} and o_{ij} are zero if the instance pair is unrelated in the discourse.

For each DA instance pair ij, equal-polarity constraints are applied to the polarity variables of i (x_i, y_i) and j (x_j, y_j) via the following equations:

$$|x_i - x_j| \leq 1 - e_{ij} + \epsilon_{ij} \ , \ \forall i \neq j \qquad (5)$$

$$|y_i - y_j| \leq 1 - e_{ij} + \epsilon_{ij} \ , \ \forall i \neq j \qquad (6)$$

$$-(x_i + y_i) \leq -l_i \ , \ \forall i \qquad (7)$$

When $e_{ij} = 1$, the Equation 5 constrains x_i and x_j to be of the same value (both zero or both one). Similarly, Equation 6 constrains y_i and y_j to be of the same value. Via these equations, we ensure that the instances i and j do not have the opposite polarity when $e_{ij} = 1$. However, notice that, if we use just Equations 5 and 6, the optimization can converge to the same, non-polar (neutral) category. To guide the convergence to the same polar (positive or negative) category, we use Equation 7. Here $l_i = 1$ if the instance i participates in one or more discourse relations. When $e_{ij} = 0$, x_i and x_j (and y_i and y_j), can take on assignments independently of one another. Notice that both constraints 5 and 6 are relaxed when $\epsilon_{ij} = 1$; thus, x_i and x_j (or y_i and y_j) can take on values independently of one another, even if $e_{ij} = 1$.

Next, the opposite-polarity constraints are applied via the following equations:

$$|x_i + x_j - 1| \leq 1 - o_{ij} + \delta_{ij} \ , \ \forall i \neq j \qquad (8)$$

$$|y_i + y_j - 1| \leq 1 - o_{ij} + \delta_{ij} \ , \ \forall i \neq j \qquad (9)$$

In the above equations, when $o_{ij} = 1$, x_i and x_j (and y_i and y_j) take on opposite values; for example, if $x_i = 1$ then $x_j = 0$ and vice versa. When $o_{ij} = 0$, the variable assignments are independent of one another. This set of constraints is relaxed when $\delta_{ij} = 1$.

In general, in our ILP formulation, notice that if an instance does not have a discourse relation to any other instance in the data, its classification is unaffected by the optimization. Also, as the underlying discourse scheme poses constraints only on the interpretation of the polarity of the related instances, discourse constraints are applied only to the polarity variables x and y, and not to the neutral class variable, z. Finally, even though slack variables are used, we discourage the ILP system from indiscriminately setting the slack variables to 1 by making them a part of the objective function that is minimized.

5 Experiments

In this work, we are particularly interested in improvements due to discourse-based methods. Thus, we report performance under three conditions: over only those instances that are related via discourse relations (*Connected*), over instances not related via discourse relations (*Singletons*), and over all instances (*All*).

The annotated data consists of 7 scenario-based, multi-party meetings from the AMI meeting corpus. We filter out very small DAs (DAs with fewer than 3 tokens, punctuation included). This gives us a total of 4606 DA instances, of which 1935 (42%) have opinion annotations. For our experiments, the DAs with no opinion annotations as well as those with neutral opinions are considered as neutral. Table 3 shows the class distributions in the data for the three conditions.

	Pos	Neg	Neutral	Total
Connected	643	343	81	1067
Singleton	553	233	2753	3539
All	1196	576	2834	4606

Table 3: Class distribution over connected, single and all instances.

5.1 Classifiers

Our first baseline, *Base*, is a simple distribution-based classifier that classifies the test data based on the overall distribution of the classes in the training data. However, in Table 3, the class distribution is different for the Connected and Singleton conditions. We incorporate this in a smarter baseline, *Base-2*, which constructs separate distributions for connected instances and singletons. Thus, given a test instance, depending on whether it is connected, Base-2 uses the corresponding distribution to make its prediction.

The third baseline is the supervised classifier, *Local*, described in Section 4.1. It is implemented using the SVM classifiers from the Weka toolkit (Witten and Frank, 2002).[4] Our supervised discourse-based classifier, ICA from Section 4.2, also uses a similar SVM implementation for its relational classifier. We implement our ILP approach from Section 4.3 using the optimization toolbox from Mathworks (http://www.mathworks.com) and GNU Linear Programming Kit.

We observed that the ILP system performs better than the ICA system on instances that are connected, while ICA performs better on singletons. Thus, we also implemented a simple hybrid classifier (HYB), which selects the ICA prediction for classification of singletons and the ILP prediction for classification of connected instances.

5.2 Results

We performed 7-fold cross validation experiments, where six meetings are used for training and the seventh is used for testing the supervised classifiers (Base, Base-2, Local and ICA). In the case of ILP, the optimization is applied to the output of Local for each test fold. Table 4 reports the accuracies of the classifiers, averaged over 7 folds.

First, we observe that Base performs poorly over connected instances, but performs considerably better over singletons. This is expected as the overall majority class is neutral and the singletons are more likely to be neutral. Base-2, which incorporates the differentiated distributions, performs substantially better than Base. Local achieves an overall performance improvement over Base and Base-2 by 23 percentage points and 9 percentage points, respectively. In general, Local outperforms Base for all three conditions ($p < 0.001$), and Base-2 for the Singleton and All conditions ($p < 0.001$). This overall improvement in Local's accuracy corroborates the utility of the lexical, unigram and DA based features for polarity detection in this corpus.

Turning to the discourse-based classifiers, ICA, ILP and HYB, all of these perform better than Base and Base-2 for all conditions. ICA improves over Local by 9 percentage points for Connected, 3 points for Singleton and 4 points for All. ILP's improvement over Local for Connected and All is even more substantial: 28 percentage points and 6 points, respectively. Notice that ILP has the same performance as Local for Singletons, as the discourse constraints are not applied over unconnected instances. Finally, HYB significantly outperforms Local under all conditions. The significance levels of the improvements over Local are highlighted in Table 4. These improvements also signify that the underlying discourse scheme is effective, and adaptable to different implementations.

Interestingly, ICA and ILP improve over Local in different ways. While ILP sharply improves the performance over the connected instances, ICA shows relatively modest improvements over both connected and singletons. ICA's improvement over singletons is interesting because it indicates that, even though the features in Table 1 are focused on discourse relations, ICA utilizes them to learn the classification of singletons too.

Comparing our discourse-based approaches, ILP does significantly better than ICA over connected instances ($p < 0.001$), while ICA does significantly better than ILP over singletons ($p <$

[4] We use the SMO implementation, which, when used with the logistic regression, has an output that can be viewed as a posterior probability distribution.

	Base	Base-2	Local	ICA	ILP	HYB
Connected	24.4	47.56	46.66	**55.64**	**75.07**	**75.07**
Singleton	51.72	63.23	75.73	78.72	75.73	78.72
All	45.34	59.46	68.72	**73.31**	**75.35**	**77.72**

Table 4: Accuracies of the classifiers measured over Connected, Singleton and All instances. Performance significantly better than Local are indicated in **bold** for $p < 0.001$ and underline for $p < 0.01$.

0.01). However, there is no significant difference between ICA and ILP for the All condition. The HYB classifier outperforms ILP for the Singleton condition ($p < 0.01$) and ICA for the Connected condition ($p < 0.001$). Interestingly, over all instances (the All condition), HYB also performs significantly better than *both* ICA ($p < 0.001$) and ILP ($p < 0.01$).

5.3 Analysis

Amongst our two approaches, ILP performs better, and hence we further analyze its behavior to understand how the improvements are achieved. Table 5 reports the performance of ILP and Local for the precision, recall and f-measure metrics (averaged over 7 test folds), measured separately for each of the opinion categories. The most prominent improvement by ILP is observed for the recall of the polar categories under the Connected condition: 40 percentage points for the positive class, and 29 percentage points for the negative class. The gain in recall is not accompanied by a significant loss in precision. This results in an improvement in f-measure for the polar categories (24 points for positive and 16 points for negative). Also note that, by virtue of the constraint in Equation 7, ILP does not classify any connected instance as neutral; thus the precision is NaN, recall is 0 and the f-meaure is NaN. This is indicated as * in the Table.

The improvement of ILP for the All condition, for the polar classes, follows a similar trend for recall (18 to 21 point improvement) and f-measure (9 to 13 point improvement). In addition to this, the ILP has an *overall improvement in precision over Local*. This may seem counterintuitive, as in Table 5, ILP's precision for connected nodes is similar to, or lower than, that of Local. This is explained by the fact that, while going from connected to overall conditions, Local's polar predictions increase by threefold (565 to 1482), but its *correct* polar predictions increase by only twofold (430 to 801). Thus, the ratio of change in the total polar predictions to the correct polar predictions is 3 : 2. On the other hand, while polar predictions by ILP increase by only twofold (1067 to 1984), its *correct* polar predictions increase by 1.5 times (804 to 1175). Here, the ratio of change in the total polar predictions to the correct polar predictions is 4 : 3, a smaller ratio.

Gold	Local			
	Pos	Neg	Neut	Total
Pos	551	113	532	1196
Neg	121	250	205	576
Neut	312	135	2387	2834
Total	984	498	3124	4606
Gold	ILP			
	Pos	Neg	Neut	Total
Pos	817	157	222	1196
Neg	147	358	71	576
Neut	358	147	2329	2834
Total	1322	662	2622	4606

Table 6: Contingency table over all instances.

The contingency table (Table 6) shows how Local and ILP compare against the gold standard annotations. Notice here, that even though ILP makes more polar guesses as compared to Local, a greater proportion of the ILP guesses are correct. The number of non-diagonal elements are much smaller for ILP, resulting in the accuracy improvements seen in Table 4.

6 Examples and Discussion

The results in Table 4 show that Local, which provides the classifications for bootstrapping ICA and ILP, predicts an incorrect class for more than 50% of the connected instances. Methods starting with noisy starting points are in danger of propagating the errors and hence worsening the performance. Interestingly, in spite of starting with so many bad classifications, ILP is able to achieve a large performance improvement. We discovered that, given a set of connected instances, even when Local has only one correct guess, ILP is able to use this to rectify the related instances. We illustrate this situation in Figure 2, which reproduces the connected DAs for Example 1. It shows the classifications

	Positive		Negative		Neutral	
	Local	ILP	Local	ILP	Local	ILP
Connected-Prec	78.1	78.2	71.9	69.8	12.1	*
Connected-Recall	45.3	**86.3**	44.1	**73.4**	62.8	
Connected-F1	56.8	**81.5**	54.0	**70.7**	18.5	
All-Prec	56.2	**61.3**	52.3	*54.6*	76.3	**88.3**
All-Recall	46.6	**67.7**	44.3	**62.5**	83.9	81.5
All-F1	50.4	**64.0**	46.0	**57.1**	79.6	84.6

Table 5: Precision, Recall, Fmeasure for each Polarity category. Performance significantly better than Local are indicated in **bold** ($p < 0.001$), underline ($p < 0.01$) and *italics* ($p < 0.05$). The * denotes that ILP does not retrieve any connected node as neutral.

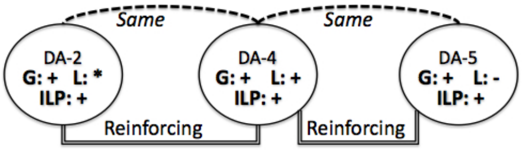

Figure 2: Discourse Relations and Classifications for Example 1.

for each DA from the gold standard (G), the Local classifier (L) and the ILP classifier (ILP). Observe that Local predicts the correct positive class (+) for only DA-4 (the DA containing **bit more durable** and **ergonomic**). Notice that these are clear cases of positive evaluation. It incorrectly predicts the polarity of DA-2 (containing **bit more bouncy**) as neutral (*), and DA-5 (containing **a bit different from all the other remote controls**) as negative (-). DA-2 and DA-5 exemplify the fact that polarity classification is a complex and difficult problem: being bouncy is a positive evaluation in this particular discourse context, and may not be so elsewhere. Thus, naturally, lexicons and unigram-based learning would fail to capture this positive evaluation. Similarly, "being different" could be deemed negative in other discourse contexts. However, ILP is able to arrive at the correct predictions for all the instances. As the DA-4 is connected to both DA-2 and DA-5 via a discourse relation that enforces an equal-polarity constraint (same+reinforcing relation of row 1, Table 2), both of the misclassifications are rectified. Presumably, the incorrect predictions made by Local are low confidence estimates, while the predictions of the correct cases have high confidence, which makes it possible for ILP to make the corrections.

We also observed the propagation of the correct classification for other types of discourse relations, for more complex types of connectivity, and also for conditions where an instance is not directly connected to the correctly predicted instance. The meeting snippet below (Example 2) and its corresponding DA relations (Figure 3) illustrate this. This example is a reinforcing discourse where the speaker is arguing for the number keypad, which is an alternative to the scrolling option. Thus, he argues against the scrolling, and argues for entering the number (which is a capability of the number keypad).

(2) D-1: I reckon you're **gonna have to have** a *number keypad* anyway for the amount of channels these days,
D-2: You **wouldn't want to** just have to *scroll through* all the channels to get to the one you want
D-3: You **wanna** *enter just the number of it* , if you know it
D-4: I reckon **we're gonna have to have** a *number keypad* anyway

In Figure 3, we see that, DA-2 is connected via an alternative+reinforcing discourse relation to each of its neighbors DA-1 and DA-3, which encourages the optimization to choose a class for it that is opposite to DA-1 and DA-3. Notice also, that even though Local predicts only DA-4 correctly, this correct classification finally influences the correct choice for all the instances, including the remotely connected DA-2.

7 Conclusions and Future Work

This work focuses on the first step to ascertain whether discourse relations are useful for improving opinion polarity classification, whether they can be modeled and what modeling choices can be used. To this end, we explored two distinct paradigms: the supervised ICA and the unsupervised ILP. We showed that both of our approaches are effective in exploiting discourse relations to

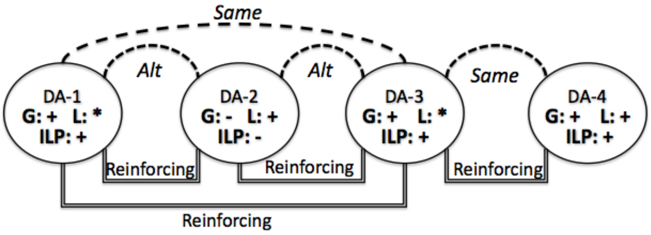

Figure 3: Discourse Relations and Classifications for Example 2.

significantly improve polarity classification. We found that there is a difference in how ICA and ILP achieve improvements, and that combining the two in a hybrid approach can lead to further overall improvement. Quantitatively, we showed that our approach is able to achieve a large increase in recall of the polar categories without harming the precision, which results in the performance improvements. Qualitatively, we illustrated how, even if the bootstrapping process is noisy, the optimization and discourse constraints effectively rectify the misclassifications. The improvements of our diverse global inference approaches indicate that discourse information can be adapted in different ways to augment and improve existing opinion analysis techniques.

The automation of the discourse-relation recognition is the next step in this research. The behavior of ICA and ILP can change, depending on the automation of discourse level recognition. The implementation and comparison of the two methods under full automation is the focus of our future work.

Acknowledgments

This research was supported in part by the Department of Homeland Security under grant N000140710152 and NSF Grant No. 0746930. We would also like to thank the anonymous reviewers for their helpful comments.

References

N. Asher, F. Benamara, and Y. Mathieu. 2008. Distilling opinion in discourse: A preliminary study. *COLING-2008*.

M. Bansal, C. Cardie, and L. Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *COLING-2008*.

M. Bilgic, G. M. Namata, and L. Getoor. 2007. Combining collective classification and link prediction. In *Workshop on Mining Graphs and Complex Structures at the IEEE International Conference on Data Mining*.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meetings corpus. In *Proceedings of the Measuring Behavior Symposium on "Annotating and measuring Meeting Behavior"*.

Y. Choi, E. Breck, and C. Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *EMNLP 2006*.

P. Denis and J. Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL 2007*.

A. Devitt and K. Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL 2007*.

A. B. Goldberg and X. Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*.

A. Haghighi, K. Toutanova, and C. Manning. 2005. A joint model for semantic role labeling. In *CoNLL*.

H. Kanayama and T. Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP-2006*, pages 355–363, Sydney, Australia.

A. Kennedy and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Q. Lu and L. Getoor. 2003. Link-based classification. In *Proceedings of the International Conference on Machine Learning (ICML)*.

A. Moschitti, D. Pighin, and R. Basili. 2006. Semantic role labeling via tree kernel joint inference. In *CoNLL*.

A. Moschitti. 2009. Syntactic and semantic kernels for short text pair categorization. In *EACL*.

D. Neiberg, K. Elenius, and K. Laskowski. 2006. Emotion recognition in spontaneous speech using gmms. In *INTERSPEECH 2006 ICSLP*.

J. Neville and D. Jensen. 2000. Iterative classification in relational data. In *In Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20. AAAI Press.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACl 2004*.

L. Polanyi and A. Zaenen, 2006. *Contextual Valence Shifters*. Computing Attitude and Affect in Text: Theory and Applications.

A.-M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT-EMNLP 2005*.

S. Raaijmakers, K. Truong, and T. Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *EMNLP*.

M. Richardson and P. Domingos. 2006. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136.

D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *Proceedings of CoNLL-2004*, pages 1–8. Boston, MA, USA.

K. Sadamitsu, S. Sekine, and M. Yamamoto. 2008. Sentiment analysis based on probabilistic models using inter-sentence information. In *LREC'08*.

S. Somasundaran, J. Ruppenhofer, and J. Wiebe. 2007. Detecting arguing and sentiment in meetings. In *SIGdial Workshop on Discourse and Dialogue 2007*.

S. Somasundaran, J. Wiebe, and J. Ruppenhofer. 2008. Discourse level opinion interpretation. In *Coling 2008*.

H. Takamura, T. Inui, and M. Okumura. 2007. Extracting semantic orientations of phrases from dictionary. In *HLT-NAACL 2007*.

B. Taskar, M. Wong, P. Abbeel, and D. Koller. 2004. Link prediction in relational data. In *Neural Information Processing Systems*.

M. Thomas, B. Pang, and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *EMNLP 2006*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT-EMNLP 2005*.

I. H. Witten and E. Frank. 2002. Data mining: practical machine learning tools and techniques with java implementations. *SIGMOD Rec.*, 31(1):76–77.