Inferring Interpersonal Relations in Narrative Summaries

Shashank Srivastava
Machine Learning Department
ssrivastava@cmu.edu

November, 2016

Abstract

Background: Characterizing relationships between people is fundamental for the understanding of narratives. It can guide interpretation of narrative events, explain character actions and behavior and steer the reader's expectations. As such, it has direct value for applications such as machine reading, QA and summarization. However, the problem of exploring interpersonal relationships in narratives is considerably understudied.

Aim: We aim to address the problem of inferring the polarity of relationships between people in narrative summaries. We expect relationships between narrative characters to be reflected in their interactions with each other during the course of the narrative, as well as their relationships with others in the social community of the narrative. Thus, we attempt to develop a method for joint inference of cooperative and adversarial relationships between characters in a text that accounts for both these types of evidence.

Data: We process and create an annotated dataset of interpersonal relationships from a subset of 153 movie summaries from the CMU Movie Summaries dataset, consisting of 1044 sentiment annotations of interpersonal relationships.

Methods: We formulate the problem as a structured prediction for each narrative, where narrative characters represent nodes in a graph, and latent variables for edge labels represent polarities of their relationships. The polarity of the relationship between any pair of characters is inferred from the actions they do to each other, the narrative's bias in describing them, as well as structural features indicating their relations with other characters. This allows the model to use the structured perceptron framework for learning empirical affinities for structural regularities such as e.g., 'a friend of a friend is a friend', which have so far been studied from abstract perspectives of structural balance theory and social psychology. We also provide a clustering based extension to the model to make narrative-specific predictions, e.g. predict more love triangles in romantic stories than in love dramas. The approach does this by learning a common base model, while allowing additive cluster-specific weights.

Results: On a labeled dataset of movie summaries from Wikipedia, our structured models provide more than a 30% error-reduction over a competitive baseline that considers pairs of characters in isolation. Social networks extracted by the models are qualitatively meaningful. Many errors are due to pre-processing problems with the NLP pipeline.

1 Introduction

Understanding narratives requires the ability to interpret character intentions, desires and relationships. The importance of characters in narratives has been explored in works that focus on their roles and representations [Bamman et al., 2014, Valls-Vargas et al., 2014, Chambers, 2013], as against a plot-centric perspective of narratives as primarily sequences of events [Finlayson, 2012, Schank and Abelson, 1977, Chambers and Jurafsky, 2008]. However, while such approaches can identify characters types or personas [Bamman et al., 2013], they do not model *relationships* between characters in a narrative.

In this work, we address the problem of inferring cooperative and adversarial relationships between people in narrative summaries. Identifying character cooperation and conflict can guide interpretation of narrative events, explain character actions and behavior and steer the reader's expectation about the plot. As such, categorizing relationships can have value for applications such as machine reading, QA and document summarization. An example of such a task is the recently proposed story cloze task, that attempts to identify the most suitable ending to a story based on previous context [Mostafazadeh et al., 2016]. Figure 1 shows an example from this task, showing how characterizing interpersonal relations can help identify the correct ending.

Context	Right Ending	Wrong Ending
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.

Figure 1: Story cloze

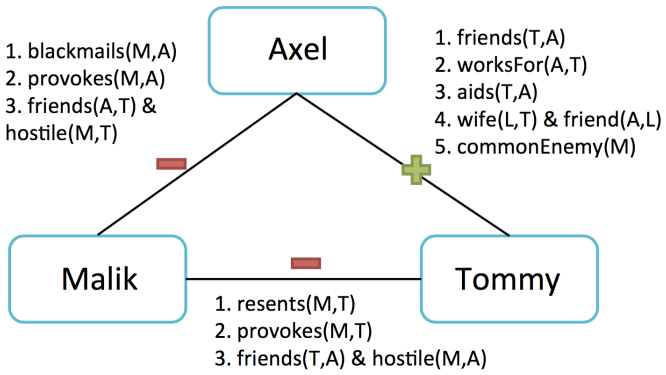
As a concrete example of interpersonal relationships in realistic narratives, let us consider the Wikipedia plot summary for the 1957 movie 'Edge of the City' in Figure 2a (condensed here for brevity). In this passage, the relations between the principal characters can be identified through a combination of cues, as seen in Figure 2b. For instance, one can infer that Alex (A) and Tommy (T) have a cooperative relationship through a combination of the following observations (among others): (1) T initially 'befriends' A, (2) A works for T, and its connotation that A is likely to cooperate with T , (3) T aids A in fights, (4) A is a friend of T's wife, (5) A and T have a common adversary. The ability to make such an inference can assist understanding of the text, as well as predict/justify future actions of A or T in the text (for example, this might explain why A later kills M in the story).

In particular, we note that cues (4) and (5) cannot be extracted from looking at the relation between A and T in isolation, but depend on their relations with others. Hence, such indirect structural cues can be significant for inference of character relationships. The aim of this work is to develop a model that takes a narrative text such as in Figure 2a as input, and returns a social network indicating relationship types such as in Figure 2b as output.

For the purpose of this work, we focus only on identifying the sentiment polarities of relationships (positive or friendly, as against negative or adversarial/hostile). This is a simple characterization, which doesn't acknowledge important facets of relationships such as asymmetry, or different types of relationships such as authority, family and romance. Secondly, our problem formulation assumes a fixed relation between a pair of characters

Young drifter Axel Nordmann goes to work in a gang of stevedores headed by Charlie Malik, a vicious bully, and is befriended by Tommy Tyler, who also supervises a stevedore gang. Malik resents blacks in positions of authority, and is antagonized when Axel goes to work for Tommy. Axel moves into Tommy's neighborhood and becomes friends with Tommy's wife Lucy. Axel is hiding something, and it emerges that he is a deserter from the United States Army. Malik is aware of that, and is extorting money from him. Malik frequently tries to provoke Tommy and Axel into fights, with Tommy coming to Axel's aid ...

(a)



(b)

Figure 2: (a) Sample summary extract for the 1957 movie 'Edge of the City'; and (b) inferred relationship polarities with supporting evidence

within a narrative. While this can be problematic since relationships can transform over time; in a wide range of examples, the assumption is reasonable. Even in complex narratives, relationships remain persistent within sub-parts. From a pragmatic perspective, these approximations serves as useful starting points for research.

The layout of this report is as follows: In section 3, we describe our formulation of the problem as a structured prediction, and describe our models for joint inference of interpersonal relationships. In section 4, we describe the text-based as well as the structural features used by our models in detail. In section 5, we then describe the dataset and data annotation process. In section 6, we present quantitative and qualitative evaluations of our models.

2 Related work

Most previous research on characterizing relationships between people has almost exclusively focused on dialogue or social network data. Such methods have explored aspects of relations such as power [Bramsen et al., 2011], address formality [Krishnan and Eisenstein, 2015] and sentiment [Hassan et al., 2012] in conversations. Recently, [Agarwal et al., 2014] studied the problem of parsing movie screenplays for extracting social networks. However, analysis of character relationships in narrative texts has largely been limited to simplistic schemes based

on counting character co-occurrences in quoted conversations [Elson et al., 2010] or social events [Agarwal et al., 2013].

In terms of approach, our use of structural triads as features is most closely related to [Krishnan and Eisenstein, 2015] who use an unsupervised joint probabilistic model of text and structure for the task of inducing formality from address terms in dialogue, and [Leskovec et al., 2010] who empirically analyze signed triads in social networks from a perspective of structural theories. Such social triads have previously been studied from perspectives of social psychology and networks [Heider, 1946, Cartwright and Harary, 1956].

Most of the material presented here is taken from [Srivastava et al., 2016]. More recently, other approaches have analyzed the evolution of specific relationships between two characters in a narrative [Chaturvedi et al., 2016], as well as unsupervised modeling of types of relationships beyond sentence polarities [Iyyer et al., 2016, Chaturvedi et al., 2017].

3 Relation classification as Structured Prediction

We formulate the problem of relation classification to allow arbitrary text-based and structural features. We consider the problem as a structured prediction, where we *jointly* infer the collective assignment of relations-labels for all pairs of characters in a document. Let $\mathbf{x}$ denote a narrative document for which we want to infer relationship structure $\boldsymbol{y}$. We could think of $\mathbf{x}$ as a graph with characters as nodes, and relationship predictions corresponding to edge-labels. We assume a supervised learning setting where we have labeled training set $\mathcal{T} := \{(\mathbf{x_1}, \mathbf{y_1}), ..., (\mathbf{x_m}, \mathbf{y_m})\}$. For each $\boldsymbol{x}$, we have a set of allowed assignments $\mathcal{Y}(\boldsymbol{x})$ (consisting of combinations of binary assignments to each edge-label in $\boldsymbol{x}$). Following standard approaches in structured classification, we consider linear classifiers of form:

$$h_{\boldsymbol{w}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})}{\operatorname{argmax}} \quad \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

Here, $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ is a feature vector that can depend on both the narrative document $\mathbf{x}$ and a relation-polarity assignment $\mathbf{y}$, $\boldsymbol{w}$ is a weight vector, and $\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ denotes a linear score indicating the goodness of the assignment. Finding the best assignment corresponds to the decoding problem, i.e. finding the highest scoring assignment under a given model. On the other hand, the model parameters $\boldsymbol{w}$ can be learnt using a voted structured perceptron training algorithm [Collins, 2002]. The structured perceptron updates can also be seen as stochastic sub-gradient descent steps minimizing the following structured hinge loss:

$$L(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}) := \max_{\boldsymbol{y'} \in \mathcal{Y}(\boldsymbol{x})} \boldsymbol{w}^T \big(\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y'}) - \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \big) \tag{2}$$

For our problem, we define the feature vector $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ as a concatenation of features based on text and structural components: $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) := \big(\boldsymbol{\phi}_{text}(\boldsymbol{x}, \boldsymbol{y}) \; \boldsymbol{\phi}_{struct}(\boldsymbol{x}, \boldsymbol{y}) \big)$. The text-based component can be defined by extending the traditional perceptron framework as $\boldsymbol{\phi}_{text}(\boldsymbol{x}, \boldsymbol{y}) := \sum_{x_e \in E(\boldsymbol{x})} y_e \boldsymbol{\phi}(x_e)$. Here $E(\boldsymbol{x})$ consists of the set of annotated character-pair relationships for the narrative text $\boldsymbol{x}$, $\boldsymbol{\phi}(x_e)$ denotes the text-based feature-representation for the character-pair (as described in Section 4.1), and y_e is the binary assignment label (± 1) for the pair in $\boldsymbol{y}$. On the other hand, our structural features $\boldsymbol{\phi}_{struct}(\boldsymbol{x}, \boldsymbol{y})$ focus on configurations of relationship

assignments of triads of characters, and are motivated in our discussion of transitive relations in Section 4.2 . We note that while this is not the case in the current work, structural features can also encode character attributes (such as age or gender) in conjunction with assignment labels $\boldsymbol{y}$.

3.1 Learning and Inference:

Structured perceptrons have been conventionally used for simple structured inputs (sequences and trees) and locally factoring models, which are amenable to efficient dynamic programming inference algorithms. This is because updates require inference over an exponentially large space (solving the decoding problem in Equation 1), and updates from inexact search can violate convergence properties. However, [Huang et al., 2012] show that exact search is not needed for convergence as long as we can guarantee updates on 'violations' only, i.e. it suffices to find a labeling assignment with higher score than the correct update. Additionally, edge labels are expected to be relatively sparse for our domain since character graphs in most narratives are not fully connected. Hence, the inference problem decomposes for relation-edges which are not parts of structural triangles, and the decoding problem can be exactly solved for nearly all of the movie summaries in our dataset.

Algorithm 1 Perceptron Training for Relations

```
Initialize w to 0
for iter : 1 to numEpochs do
    for j : 1 to m do
        Randomly choose i ∈ {1..m}
        ŷ ← Decode(x_i, w)
        if ( w^T φ(x_i, ŷ) ≥ w^T φ(x_i, y_i)) then
      w ← w + η(φ(x_i, y_i) − φ(x_i, y'))
        end if
    end for
end for
```

For inference on a new document where the edge relations are not known, decoding can proceed by initializing the narrative graph to high confidence edges from the text-based model only (character relationships firmly embedded in text), and appending single edges which complete triads. To avoid speculative inference of relations between character pairs that are ungrounded in the text, we only consider structural triads for which at least two edges are grounded in the text while decoding with the structural model.

3.2 Accounting for narrative types

The framework described in the previous section provides a simple model to incorporate text-based and structural features for relation classification. However, a shortcoming in the approach is that the model is agnostic to narrative types. Ideally, a model could allow differential weights to features depending on the narrative type. As speculative illustrations, 'Mexican standoffs' might be common in 'revenge/gangster' narratives, or family-relations

might be highly indicative of cooperation in children stories; and a model would ideally learn and leverage such regularities in the data.

We present a clustering-based extension to our structured model, which can incorporate features descriptive of the narrative text to infer regularities, and make content-based predictions. Let us surmise that the data consists of K natural clusters of narrative-types, with a specific structured model for each cluster (specified by weights $\boldsymbol{w}_k$). For each narrative text x, we associate a vector $f(x)$ that represents content and determines narrative type. Examples of such representations could be keywords for a document, genre information for a movie or novel, topic proportions of words in the text from a topic-model, etc. We model the membership of narrative $\boldsymbol{x}$ to the cluster c_k by a softmax logistic multinomial.

$$P(c = c_k; \boldsymbol{x}) = \frac{\exp(\lambda_k^T f(\boldsymbol{x}))}{\sum_{k'} \exp(\lambda_{k'}^T f(\boldsymbol{x}))} \tag{3}$$

From our observation of the loss objective for the structured perceptron in Equation 2, we can define the *expected* loss for a narrative text $(\boldsymbol{x}, \boldsymbol{y})$ under the clustering model as:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \sum_k \frac{\exp(\lambda_k^T f(\boldsymbol{x}))}{\sum_{k'} \exp(\lambda_{k'}^T f(\boldsymbol{x}))} \quad L(\boldsymbol{w}_k, \boldsymbol{x}, \boldsymbol{y}) \tag{4}$$

Then the overall objective loss over the training set $\mathcal{T}$ is:

$$\begin{aligned} J &= \sum_i \mathcal{L}(\boldsymbol{x}_i, \boldsymbol{y}_i) \\ &= \sum_i \sum_k \frac{\exp(\lambda_k^T f(\boldsymbol{x}_i)) L(\boldsymbol{w}_k, \boldsymbol{x}_i, \boldsymbol{y}_i)}{\sum_{k'} \exp(\lambda_{k'}^T f(\boldsymbol{x}_i))} \end{aligned} \tag{5}$$

We jointly minimize the overall objective through a block-coordinate descent procedure. This consists of a two-step alternating minimization of the objective w.r.t. the prediction model weights $\boldsymbol{w}_k$ and the clustering parameters λ_k, respectively. In the first step, we optimize the prediction model weights $\boldsymbol{w}_k$ while fixing the clustering parameters λ_k. This can be done by weighting the training examples for each cluster by their cluster membership; and invoking the structured perceptron procedure for each cluster. In the alternating step, we fix the predictions model weights; and update the clustering parameters using gradient descent:

$$\nabla_{\lambda_k} J = \sum_{i=1} \frac{\exp(\lambda_k^T f(\boldsymbol{x}_i))}{Z_i^2} \sum_{k'} \exp(\lambda_k'^T f(\boldsymbol{x}_i))(L(\boldsymbol{w}_k, \boldsymbol{x}_i, \boldsymbol{y}_i) - L(\boldsymbol{w}'_k, \boldsymbol{x}_i, \boldsymbol{y}_i)) f(\boldsymbol{x}_i)$$

This can be interpreted as a bootstrapping procedure, where given cluster assignments of points, we update the prediction model weights; and given losses from the prediction model, update data clusters parameters to reassign the most violating data-points. We note that the objective is non-convex due to the softmax, and hence different initializations of the procedure can lead to different solutions. However, since each sub-procedure decreases the objective value; the overall objective decreases for small enough step sizes. The procedure is summarized in Algorithm 2. For prediction, each narrative text is assigned to the most likely cluster with the clustering model.

Algorithm 2 Narrative-specific Model

Initialize λ_k to random vectors
repeat
Update perceptron weights: Train structured perceptron models for each cluster c_k, weighting training instance $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in $\mathcal{T}$ by $\frac{\exp(\lambda_k^T f(\boldsymbol{x}_i))}{\sum_{k'} \exp(\lambda_{k'}^T f(\boldsymbol{x}_i))}$
Update clustering model
for $i : 1$ to $numIter$ **do**
    for $k : 1$ to K **do** $\lambda_k \leftarrow \lambda_k - \mu \nabla_{\lambda_k} J$
    end for
end for

To efficiently use training data, we allow parameter-sharing across cluster-specific prediction models, drawing from methods in multi-task learning [Evgeniou and Pontil, 2004]. In particular, we model each $\boldsymbol{w}_k$ as composed of a shared base model, and additive cluster-specific weights:

$$\boldsymbol{w}_k = \boldsymbol{w}_0 + \boldsymbol{v}_k$$

Implementationally, we can do this by simply augmenting cluster-specific feature representations as follows: $\phi_k(\boldsymbol{x}, \boldsymbol{y}) = \Big((1-\alpha)\phi(\boldsymbol{x}, \boldsymbol{y}), \underbrace{\boldsymbol{0}, ..., \boldsymbol{0}}_{\text{k-1}}, \alpha\phi(\boldsymbol{x}, \boldsymbol{y}), \underbrace{\boldsymbol{0}, ..., \boldsymbol{0}}_{\text{K-k}}\Big)$ Here α is a hyper-parameter between 0 and 1, which specifies the weighting of the shared and cluster-specific models. $\alpha = 0$ negates clustering, and reduces the clustering model to the plain structured model without clustering. Conversely, $\alpha = 1$ implies no parameter sharing across clusters.

4 Features

In this section, we outline the text-based and structural features used by our classification models. The text-based features make use of existing linguistic and semantic resources, whereas the structural features are based on counts of specific signed social triads, which can be enumerated for any assignment.

4.1 Text-based cues

These features aim to identfiy relationships between pairs of characters in isolation. These are based on resources such as sentiment lexicons, syntactic and semantic parsers, distributional word-representations, and multi-word-expression dictionaries, and are engineered to capture:

- Overall polarities of interactions between characters (from text-spans between coreferent mentions) based on lexical and phrasal-level polarity clues.
- Semantic connotations of actions one agent does to the other, actions they share as agents or patients, and how often they act as a team.
- Character co-occurrences in semantic frames that evoke positivity, negativity or social relationship .
- Character similarity based on whether they are described by similar adjectives; and the narrative sentiment of adverbs describing their actions.
- Existence of familial relations between characters.

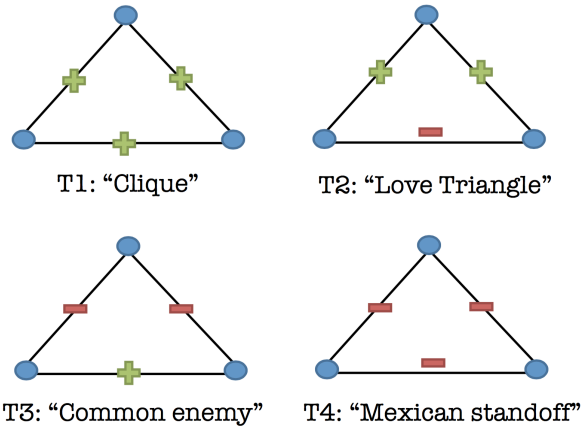


Figure 3: Primary triadic structural features. '+' signs indicate cooperative and '–' indicate adversarial relationships

We base our entity-tracking on the Stanford Core-NLP system; and augment the computation of all sentiment features with basic negation handling. Based on such features extracted for each character pair, relationship characterization can be treated as supervised classification (with $y = \pm 1$ corresponding to cooperative or adversarial relations). Our baseline unstructured approach uses only these features with a logistic regression model.

Feature details: Texts are initially processed with the Stanford Core NLP system to identify personal named entities, and obtain dependency parses. As basic post-processing, we mark coreferent text-spans with their corresponding personal entities, using basic string-matching and pronominal resolution from the Stanford coreference system. For enumerating actions by/on two characters of interest, we identify verbs in sentences for which the characters occurred in a agent or a patient role (identified using 'nsubj' or 'agent'; and 'dobj' or 'nsubjpass' dependencies respectively). We extend this for conjoint verbs, identified by the 'conj' relation (e.g., 'A shot and injured B'). The dependency relation 'neg' is used to determine negation status of verbs.

Given a pair of characters, we identify the set of sentences which contain mentions of both (identified by coreferent text-spans). For this set, we extract the arithmetic means, maximum and cumulative sums for the following sentence-level cues as text-based features (whenever meaningful):

1. *Are Team:* This models if the two characters participate in acts together. It is a binary feature indicating if the two characters were both agents (or patients) of a verb in a sentence e.g., 'Malik provokes Tommy and Axel'.
2. *Acts Together:* These features count the number of actions with positive and negative connotations that either character (in an agent role) does to the other (in a patient role). There are three variants based on different word connotation resources, viz., semantic lexicons for subjectivity clues [Feng et al., 2013], sentiment [Liu et al., 2005] and prior-polarity [Wilson et al., 2005] of verbs. The feature does not fire for neutral verbs. e.g, 'Malik blackmails Axel'.
3. *Surrogate Acts Together:* Coverage for the above features suffers from limitations of NLP

processing tools. e.g., In 'On being provoked by Malik, Tommy...' , Tommy is not a direct patient of the verb. These features extend coverage to verbs which have either of the characters as the agent or the patient in sentences that did not contain any other character apart from the two of interest.

4. *Adverb Connotations:* This feature models the narrative's overall bias in describing either character's actions by summing the semantic connotations of adverbs that modify their joint(or surrogate) acts. e.g., 'Tommy nobly befriends Axel'. Positive adverbs count as +1, negative as -1. Uses the same connotation resources as 2.
5. *Character similarity:* Models similarity of two characters as the average pairwise similarity of adjectives describing each (where lexical similarity is given by the cosine similarity of embeddings from [Collobert et al., 2011]). This is computed for the entire document, and not at the per-sentence level.
6. *Lexical sentiment:* These features count the number of positive and negative sentiments words or multi-word phrases in spans between mentions of the two characters using sentiment lexicons (similar to 2). For multi-word phrases (identified from a list of MWEs), we use a dictionary to map these to single words if possible, and look for these words in connotation lexicons. e.g., 'kick the bucket' maps to 'die'. This helps with phrases like 'fell in love', where 'fell' has a negative connotation by itself.
7. *Relation keywords:* This feature indicates presence of keywords denoting familial ties between characters ('father', 'wife', etc.) in spans between character mentions.
8. *Frame semantic:* These are based on Framenet-style parses of the sentence from the Semafor parser [Das et al., 2014]. We compiled lists of frames associated with: (i) positive or (ii) negative connotations (e.g., frames like *cause hurt* or *rescue*), (iii) personal or professional relationships (e.g., *subordinates and superiors* or *forming relationships*). Three binary features denote frame evocation for each of these lists.

4.2 Structural cues

As motivated earlier, relationships between people can also be inferred from their relationships with others in a narrative. Our thesis is that a joint inference model that incorporates both structure and text would perform better than one that considers pairwise relations in isolation. In some domains, observed relations between entities can directly imply unknown relations among others dues to natural orderings. For example, temporal relations among events yield natural transitive constraints. For the current task; such constraints do not apply. While structural regularities like 'a friend of a friend is a friend' might be prevalent, these configurations are not logically entailed; and affinities for such structural regularities must be directly learnt from the observed data.

In Figure 3, we characterize the primary triadic structural features that we use in our models, along with our informal appellations for them. The values of the four structural features for a narrative document $\mathbf{x}$ and relation polarity assignment $\mathbf{y}$ are simply the number of such configurations in the assignment, and are easily computed. Empirical affinities for such configurations, as reflected in corresponding weights can then be learnt from the data.

5 Dataset

We processed the CMU Movie Summary corpus, which is a collection of 42,306 movie plot summaries extracted from Wikipedia, along with aligned meta-data [Bamman et al., 2013]; and set up an online annotation task using BRAT [Stenetorp et al., 2012]. We use Stanford Core NLP annotations and basic post-processing to identify personal entities in each text.

Annotators could choose pairs of characters in the text, and characterize a directed relationship between them on an ordinal five-point scale as 'Hostile', 'Adversarial', 'Neutral', 'Cooperative' or 'Friendly'. A sample annotation for a very short summary is shown in Figure 4. This resulted in a dataset of 153 movie summaries, consisting of 1044 character relationship annotations.[1]

Figure 4: Sample annotation for a very short summary

For evaluation, we aggregated 'hostile' and 'adversarial' edge-labels, and 'friendly' and 'cooperative' edge-labels to have two classes (neutral annotations were sparse, and ignored in the evaluation). Of these, 58% of the relations were classified as cooperative or friendly, while 42% were hostile or adversarial. The estimated annotator agreement for the collapsed classes on a small subset of the data was >0.95 (Cohen's Kappa: 0.89).

6 Evaluation and Analysis

In this section, we discuss quantitative and qualitative evaluation of our methods. First, we make an ablation study to assess the relative importance of families of text-based features. We then make a comparative evaluation of our methods in recovering gold-standard annotations on a held-out test set of movie summaries. We qualitatively analyze the performance of the model, and discuss common sources of errors.

[1]Most relations were annotated symmetrically. For relations with asymmetric labels, we 'averaged' the annotations in the two directions to get the annotation for the relation.

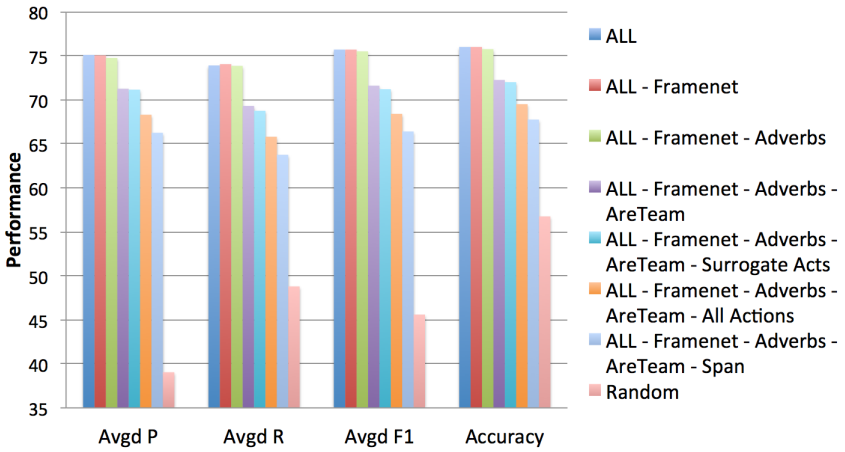


Figure 5: Ablation study for text feature families

6.1 Feature ablation:

Figure 5 shows the average 5-fold cross-validation performance of major feature families of text features on the training set. We note that Frame-semantic features and adverbial connotations of character actions do not add to model performance. This is perhaps because both these families of features were sparse. Additionally, frame-semantic parses were observed to have frequent errors in frame evocation, and frame element assignment. On the other hand, we observe that joint participation in actions (as agent or patient) is a strong indicator of cooperative relationships. In particular, incorporating these (*'Are Team'*) features was seen to improve both precision and recall for the cooperative class; while not degrading recall for the non-cooperative class. Further, while ignoring sentiment and connotation features for surrogate action features results in marginal degradation in performance; the most useful features are seen to be sentiment and connotation features for actions where characters occur in SVO roles (*'Acts Together'* features); and overall sentiment characterizations for words and phrases in spans of text between character mentions (span based *'Lexical sentiment'* features).

6.2 Structured vs unstructured models:

We now analyze the performance of our proposed models; and evaluate the importance of structural features to our models. In our experiments, we found the structured models to consistently outperform text-based models. We tune values of hyper-parameters, i.e. number of training epochs for the structured perceptron (10), the weighting parameter for the clustering model (α=0.8), and the number of clusters (K=2) through 5-fold cross validation on training data to optimize the averaged F1 score.[2] Table 1 compares the performance of the models on our held-out test set of 27 movie summaries (comprising about 20% of the all annotated character relations). For the structured models, reported results are averages over 10 initializations.

[2]For representations $f(x)$ of movie summaries, we use genre keywords from the metadata for movies (provided with dataset) and the average of text-feature vectors for all character pairs

We observe that the structured perceptron model for relations (SPR) notably outperforms the text-only model trained with logistic regression (LR). These results are consistent with our cross-validation findings, and vindicate our hypothesis that structural features can improve inference of character relations. Further, we observe that the narrative-specific model (with $K = 2$) slightly outperforms the structured perceptron model.

	Avg P	**Avg R**	**Avg F$_1$**	**Acc**
Naive (majority class)	0.269	0.520	0.355	0.520
LR	0.702	0.697	0.699	0.701
SPR	0.794	0.793	0.793	0.792
SPR +Narrative types	0.806	0.804	0.805	0.804

Table 1: Test set results for relation classification

Let us consider the affinities for structural features learnt by the model. Over 10 runs of SPR, the average weights were: $w_{\texttt{clique}} = -2.79$, $w_{\texttt{lovetriangle}} = -0.84$, $w_{\texttt{commonenemy}} = 10.26$ and $w_{\texttt{mexicanstandoff}} = -5.49$. From the perspective of structural balance, the social configurations `lovetriangle` and `mexicanstandoff` are inherently unstable. Hence, the learnt affinity for the configuration `lovetriangle` seems higher than expected. This is unsurprising, however, if we consider the domain of the data (movies), where it might be a common plot element. We also note that the 'friend of a friend is a friend' maxim is not supported by the feature weights (even though it is a stable configuration), and hence a model based on this as a hard transitive constraint could be expected to perform poorly.

6.3 Cluster analysis:

We briefly analyze a particular run of the clustering model for $K = 2$. In Figure 6, we plot the overall feature weights ($\boldsymbol{w}_k$) for a run (we plot 8 features with the highest weights from the text model, and the primary structural features). We note that the two clusters are reasonably well delineated; and thus clustering is meaningful. For this run, Cluster 1 appears correlated with higher weights of positive polarity features. Cluster 2 appears less differentiated in terms of structural features than Cluster 1 or the non-clustering structured model.

6.4 Qualitative evaluation

We observe that relation characterizations for character pairs are reasonable for most narrative texts in the test set. Figure 7 shows labels inferred by the model for two well-known movies in the test set. Further, analysis of highest contributing evidences that lead to predictions indicated that the model usually provides meaningful justifications for relationship characterization in terms of narrative acts, or implied relations from structural features.
Error analysis revealed that mismatched coreference labelings are the most common source of errors for the model. Secondly, in some cases, the text-based features mistakenly identify negative sentiments due to our coarse model of sentiment. For example, consider the following segment for the movie *Smokin' Aces 2*: *'Baker drags the wounded Redstone to the "spider trap" ... used to safeguard people'*. Here, the model mistakenly predicts the relation between

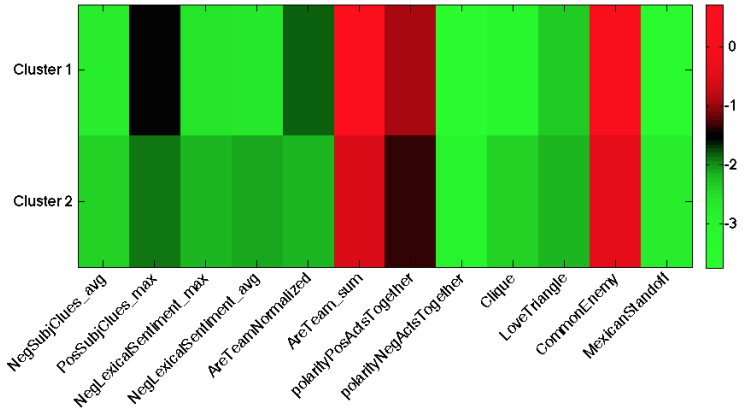


Figure 6: Heatmap of cluster-weights

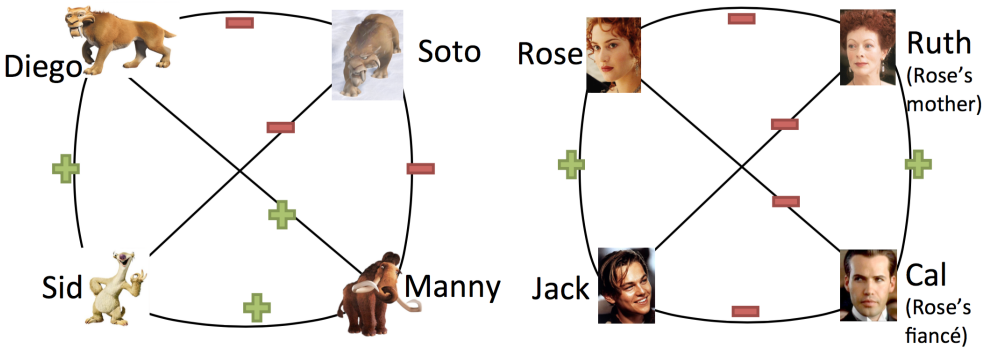


Figure 7: Inferred relationships for movies 'Titanic' (1997) and 'Ice Age' (2002)

Redstone and Baker as adversarial because of the negative connotation of 'drag', inspite of other structural evidence (common adversaries).

7 Future Work

We have presented a framework for automatically inferring interpersonal cooperation and conflict in narrative summaries. While our testbed was movie summaries, the framework could potentially apply to other domains of texts with social narratives, such as news stories and literary fiction. Our clustering framework provides a natural approach for such domain adaptation. In the future, the framework could be extended to handle nuanced relation categorizations, asymmetric and evolving relationships. Conceptually, a natural extension would be to use predictions about character relations to infer subtle character attributes such as agenda, intentions and goals.

References

[Agarwal et al., 2014] Agarwal, A., Balasubramanian, S., Zheng, J., and Dash, S. (2014). Parsing screenplays for extracting social networks from movies. *EACL 2014*, pages 50–58.

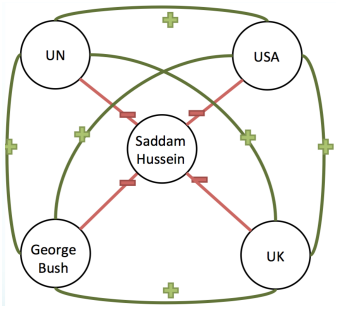


Figure 8: Selected relationships from Wikipedia article for '2003 Invasion of Iraq'

[Agarwal et al., 2013] Agarwal, A., Kotalwar, A., and Rambow, O. (2013). Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *the Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*.

[Bamman et al., 2013] Bamman, D., O'Connor, B., and Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 352–361.

[Bamman et al., 2014] Bamman, D., Underwood, T., and Smith, N. A. (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 370–379.

[Bramsen et al., 2011] Bramsen, P., Escobar-Molano, M., Patel, A., and Alonso, R. (2011). Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 773–782. Association for Computational Linguistics.

[Cartwright and Harary, 1956] Cartwright, D. and Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological review*, 63(5):277.

[Chambers, 2013] Chambers, N. (2013). Event schema induction with a probabilistic entity-driven model. In *EMNLP*, volume 13, pages 1797–1807.

[Chambers and Jurafsky, 2008] Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 789–797.

[Chaturvedi et al., 2017] Chaturvedi, S., Iyyer, M., and Daumé III, H. (2017). Unsupervised learning of evolving relationships between literary characters. In *Association for the Advancement of Artificial Intelligence*.

[Chaturvedi et al., 2016] Chaturvedi, S., Srivastava, S., Daume III, H., and Dyer, C. (2016). Modeling evolving relationships between characters in literary novels. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[Collins, 2002] Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

[Das et al., 2014] Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

[Elson et al., 2010] Elson, D. K., Dames, N., and McKeown, K. R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147. Association for Computational Linguistics.

[Evgeniou and Pontil, 2004] Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM.

[Feng et al., 2013] Feng, S., Kang, J. S., Kuznetsova, P., and Choi, Y. (2013). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1774–1784.

[Finlayson, 2012] Finlayson, M. A. (2012). *Learning narrative structure from annotated folktales*. PhD thesis, Massachusetts Institute of Technology.

[Hassan et al., 2012] Hassan, A., Abu-Jbara, A., and Radev, D. (2012). Extracting signed social networks from text. In *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, pages 6–14. Association for Computational Linguistics.

[Heider, 1946] Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112.

[Huang et al., 2012] Huang, L., Fayong, S., and Guo, Y. (2012). Structured perceptron with inexact search. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–151. Association for Computational Linguistics.

[Iyyer et al., 2016] Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., and Daumé III, H. (2016). Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *North American Association for Computational Linguistics*.

[Krishnan and Eisenstein, 2015] Krishnan, V. and Eisenstein, J. (2015). "You're Mr. Lebowski, I'm the dude": Inducing address term formality in signed social networks. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, (to appear in Proceedings)*.

[Leskovec et al., 2010] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370. ACM.

[Liu et al., 2005] Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 342–351.

[Mostafazadeh et al., 2016] Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. *Proceedings of NAACL HLT, San Diego, California, June. Association for Computational Linguistics*.

[Schank and Abelson, 1977] Schank, R. C. and Abelson, R. P. (1977). Scripts, plans, goals and understanding: an inquiry into human knowledge structures. hillsdale, nj: L. *N.J.: Erlbaum*.

[Srivastava et al., 2016] Srivastava, S., Chaturvedi, S., and Mitchell, T. (2016). Inferring interpersonal relations in narrative summaries. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[Stenetorp et al., 2012] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

[Valls-Vargas et al., 2014] Valls-Vargas, J., Zhu, J., and Ontañón, S. (2014). Toward automatic role identification in unannotated folk tales. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*.