
Best paper award (2 out of 1592)

Links:
- Code/Data [https://github.com/miyyer/rmiy](https://github.com/miyyer/rmiy)

Downloaded from [http://umiacs.umd.edu/~jbg/docs/2016_naacl_relationships.pdf](http://umiacs.umd.edu/~jbg/docs/2016_naacl_relationships.pdf)

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.
Feuding Families and Former Friends: Unsupervised Learning for Dynamic Fictional Relationships

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, Hal Daumé III

1University of Maryland, Department of Computer Science and UMIACS
2University of Colorado, Department of Computer Science
{miyyer,aguha,snigdac,hal}@umiacs.umd.edu, Jordan.Boyd.Graber@colorado.edu

Abstract

Understanding how a fictional relationship between two characters changes over time (e.g., from best friends to sworn enemies) is a key challenge in digital humanities scholarship. We present a novel unsupervised neural network for this task that incorporates dictionary learning to generate interpretable, accurate relationship trajectories. While previous work on characterizing literary relationships relies on plot summaries annotated with predefined labels, our model jointly learns a set of global relationship descriptors as well as a trajectory over these descriptors for each relationship in a dataset of raw text from novels. We find that our model learns descriptors of events (e.g., marriage or murder) as well as interpersonal states (love, sadness). Our model outperforms topic model baselines on two crowdsourced tasks, and we also find interesting correlations to annotations in an existing dataset.

1 Describing Character Relationships

When two characters in a book break bread, is their meal just a result of biological needs or does it mean more? Cognard-Black et al. (2014) argue that this simple interaction reflects the diversity and background of the characters, while Foster (2009) suggests that the tone of a meal can portend either good or ill for the rest of the book. To support such theories, scholars use their literary expertise to draw connections between disparate books: Gabriel Conroy’s dissonance from his family at a sumptuous feast in Joyce’s The Dead, the frustration of Tyler’s mother in Dinner at the Homesick Restaurant, and the grudging respect for a blind man eating meatloaf in Carver’s Cathedral.

However, these insights do not come cheap. It takes years of careful reading and internalization to make connections across books, which means that relationship symmetries and archetypes are likely to remain hidden in the millions of books published every year unless literary scholars are actively searching for them.

Natural language processing techniques have been increasingly used to assist in these literary investigations by discovering patterns in texts (Jockers, 2013). In Section 6 we review existing techniques that classify or cluster relationships between characters in books using a fixed set of labels (e.g., friend or en-
3.1 Formalizing the Problem

Assume we have two characters $c_1$ and $c_2$ in book $b$. We define $S_{c_1,c_2}$ as a sequence of token spans where each span $s_t \in S_{c_1,c_2}$ is itself a set of tokens to character name clustering, which are further expanded upon in Vala et al. (2015), for future work.

1While this pipeline works reasonably well, it is unreliable for first-person narratives; we leave the necessary improvements

Mrs. Reilly looked at her son slyly and asked, “Ignatius, you sure you not a communiss?”

“Oh, my God!”

Our goal is to make \( r_t \) similar to \( v_s \). We use a contrastive max-margin objective function similar to [72x262]

As in other neural network models for natural language processing, we begin by associating each word type \( w \) in our vocabulary with a real-valued embedding \( v_w \in \mathbb{R}^d \). These embeddings are rows of a \( V \times d \) matrix \( \mathbf{L} \), where \( V \) is the vocabulary size.

Similarly, characters and books have their own embeddings in rows of matrices \( \mathbf{C} \) and \( \mathbf{B} \). We want \( \mathbf{B} \) to capture global context information (e.g., “Moby Dick” takes place at sea) and \( \mathbf{C} \) to capture immutable aspects of characters not related to their relationships (e.g., Javert is a police officer). Finally, the \( \text{RMN} \) learns embeddings for relationship descriptors, which requires a second matrix \( \mathbf{R} \) of size \( K \times d \) where \( K \) is the number of descriptors, analogous to the number of topics in topic models.

Each input to the \( \text{RMN} \) is a tuple that contains identifiers for a book and two characters, as well as the spans corresponding to their relationship: \((b, c_1, c_2, S_{c_1,c_2})\). Given one such input, our objective is to reconstruct \( S_{c_1,c_2} \) using a linear combination of relationship descriptors from \( \mathbf{R} \) as shown in Figure 2; we now describe this process formally.

### 3.2.1 Modeling Spans with Vector Averages

We compute a vector representation for each span \( s_t \) in \( S_{c_1,c_2} \) by averaging the embeddings of the words in that span,

\[
v_{s_t} = \frac{1}{l} \sum_{w \in s_t} v_w.
\]

Then, we concatenate \( v_{s_t} \) with the character embeddings \( v_{c_1} \) and \( v_{c_2} \) as well as the book embedding \( v_b \) and feed the resulting vector into a standard feed-forward layer to obtain a hidden state \( h_t \).

\[
h_t = f(W_h \cdot [v_{s_t}; v_{c_1}; v_{c_2}; v_b]).
\]

In all experiments, the transformation matrix \( W_h \) is \( d \times 4d \), and we set \( f \) to the ReLu function, \( \text{ReLU}(x) = \max(0,x) \).

### 3.2.2 Approximating Spans with Relationship Descriptors

Now that we can obtain representations of spans, we move on to learning descriptors using a variant of dictionary learning (Olshausen and Field, 1997; Elad and Aharon, 2006), where our descriptor matrix \( \mathbf{R} \) is the dictionary and we are trying to approximate input spans as a linear combination of items from this dictionary.

Suppose we compute a hidden state for every span \( s_t \) in \( S_{c_1,c_2} \) (Equation 2). Now, given an \( h_t \), we compute a weight vector \( \mathbf{d}_t \) over \( K \) relationship descriptors with some composition function \( g \), which is fully specified in the next section. Conceptually, each \( \mathbf{d}_t \) is a relationship state, and a relationship trajectory is a sequence of chronologically-ordered relationship states as shown in Figure 1. After computing \( \mathbf{d}_t \), we use it to compute a reconstruction vector \( r_t \) by taking a weighted average over relationship descriptors,

\[
r_t = \mathbf{R}^T \mathbf{d}_t.
\]

Our goal is to make \( r_t \) similar to \( v_{s_t} \). We use a contrastive max-margin objective function similar to
previous work (Weston et al., 2011; Socher et al., 2014). We randomly sample spans from our dataset and compute the vector average \( v_{s_n} \) for each sampled span as in Equation 1. This subset of span vectors is \( N \). The unregularized objective \( J \) is a hinge loss that minimizes the inner product between \( r_t \) and the negative samples while simultaneously maximizing the inner product between \( r_t \) and \( v_{s_t} \),

\[
J(\theta) = \sum_{t=0}^{\left| S_{t_1,t_2} \right|} \sum_{n \in N} \max(0, 1 - r_t v_{s_t} + r_t v_{s_n}), \tag{4}
\]

where \( \theta \) represents the model parameters.

### 3.2.3 Computing Weights over Descriptors

What function should we choose for our composition function \( g \) to represent a relationship state at a given time step? On the face of it, this seems trivial; we can project \( h_t \) to \( K \) dimensions and then apply a softmax or some other nonlinearity that yields non-negative weights.\(^3\) However, this method ignores the relationship states at previous time steps. To model the temporal aspect of relationships, we can add a recurrent connection,

\[
d_t = \text{softmax}(W_d \cdot [h_t; d_{t-1}]) \tag{5}
\]

where \( W_d \) is of size \( K \times (d + K) \) and \( \text{softmax}(q) = \exp q / \sum_{l=1}^{d} \exp q_l \).

Our hope is that this recurrent connection will carry some of the previous relationship state over to the current time step. It should be unlikely for two characters in love at time \( t \) to fall out of love at time \( t + 1 \) even if \( s_{t+1} \) does not include any love-related words. However, because the objective function in Equation 4 maximizes similarity with the current time step’s input, the model is not forced to learn a smooth interpolation between the previous state and the current one. A natural remedy is to have the model predict the next time step’s input instead, but this proves hard to optimize.

We instead force the model to use the previous relationship state by modifying Equation 5 to include a linear interpolation between \( d_t \) and \( d_{t-1} \),

\[
d_t = \alpha \cdot \text{softmax}(W_d \cdot [h_t; d_{t-1}]) + (1 - \alpha) \cdot d_{t-1}. \tag{6}
\]

Here, \( \alpha \) is a scalar between 0 and 1. We experiment with setting \( \alpha \) to a fixed value of 0.5 as well as allowing the model to learn \( \alpha \) as in

\[
\alpha = \sigma(v_{\alpha}^T \cdot [h_t; d_{t-1}; v_{s_t}]), \tag{7}
\]

where \( \sigma \) is the sigmoid function and \( v_{\alpha} \) is a vector of dimensionality \( 2d + K \). Fixing \( \alpha = 0.5 \) initially and then tuning it after other parameters have converged improves training stability; for the specific hyperparameters we use see Section 4.\(^3\)

### 3.2.4 Interpreting Descriptors and Enforcing Uniqueness

Recall that each descriptor is a \( d \)-dimensional row of \( \text{R} \). Because our objective function \( J \) forces these descriptors to be in the same vector space as that of the word embeddings \( \text{L} \), we can interpret them by looking at nearest neighbors in \( \text{L} \) using cosine distance as the similarity metric.

To discourage learning descriptors that are too similar to each other, we add another penalty term \( X \) to our objective function,

\[
X(\theta) = \| \text{RR}^T - \text{I} \|, \tag{8}
\]

where \( \text{I} \) is the identity matrix. This term comes from the component orthogonality constraint in independent component analysis (Hyvärinen and Oja, 2000).

We add \( J \) and \( X \) together to obtain our final training objective \( L \),

\[
L(\theta) = J(\theta) + \lambda X(\theta), \tag{9}
\]

where \( \lambda \) is a hyperparameter that controls the magnitude of the uniqueness penalty.

### 4 Evaluating Descriptors and Trajectories

Because no previous work explores the interpretability of unsupervised relationship modeling over time, evaluating the \text{RMN} is tricky. Further compounding the problem is the subjective nature of the task; for example, is a trajectory that ignores a key event better than one that hallucinates episodes absent from source text?\(^3\)

---

\(^3\)We experiment with a variety of nonlinearities but find that the softmax yields the most interpretable results due to its predisposition to select a single descriptor.
With these issues in mind, we conduct three evaluations to show that our output is reasonable. First, we conduct a crowdsourced interpretability experiment that shows RMNs produce significantly more coherent descriptors than three topic model baselines. A second crowdsourced task indicates that our model produces trajectories that match plot summaries more accurately than topic models. Finally, we qualitatively compare the RMN’s output to existing static annotations of literary relationships and find both expected and surprising results.

4.1 Topic Model Baselines

Before moving onto the evaluations, we briefly describe three baseline models, all of which are Bayesian generative models. Latent Dirichlet allocation (Blei et al., 2003, LDA) learns a single document-topic distribution per document; we can apply LDA to our dataset by concatenating all spans from a relationship into a single document. Similarly, NUBBI (Chang et al., 2009a) learns separate sets of topics for relationships and individual characters.\footnote{NUBBI requires additional spans that mention only a single character to differentiate character topics from relationship topics. None of the other models receives these extra data.}

LDA and NUBBI are incapable of taking into account the chronological ordering of the spans because they view all relationships tokens as exchangeable. While we can compare the descriptors learned by these models to those of the RMN, we cannot evaluate their trajectories. We turn instead to the hidden topic Markov model (Gruber et al., 2007, HTMM), which foregoes the bag-of-words assumption of LDA and NUBBI in favor of modeling topic segments within a document as a Markov chain. This model outputs a smooth sequence of topic assignments over a document, so we can compare the trajectories it learns on our dataset to those of the RMN.

4.2 Experimental Settings

In our descriptor interpretability experiments, we vary the number of descriptors (topics) for all models ($K = 10, 30, 50$). We train LDA and NUBBI for 100 iterations with a collapsed Gibbs sampler, and the HTMM uses the default setting of 100 EM iterations.

For the RMN, we initialize the word embedding matrix $L$ with 300-dimensional GloVe embeddings trained on the Common Crawl (Pennington et al., 2014). The character and book embeddings ($C$ and $B$) are initialized randomly. We fix $\alpha$ to 0.5 for the first 15 epochs of training; after the descriptor matrix $R$ has converged, we fix $R$ and tune $\alpha$ using Equation 6 for 15 more epochs.\footnote{Preliminary experiments show that learning $\alpha$ and $R$ simultaneously results in less interpretable descriptors.} Since the topic model baselines do not have access to character and book metadata, for fair comparison we also train a “generic” version of the RMN (GRMN) where the metadata embeddings are removed from Equation 2. The uniqueness penalty $\lambda$ is set to $10^{-4}$.

All of the RMN model parameters except $L$ are optimized using Adam (Kingma and Ba, 2014) with a learning rate of 0.001 for 30 epochs; the word embeddings are not fine-tuned during training.\footnote{Tuning $L$ ruins descriptor interpretability; pretrained embeddings are likely already a good solution for our problem.} We also apply word dropout (Iyyer et al., 2015) to the input spans, removing words from the vector average computation in Equation 1 with probability 0.5.

4.3 Do Descriptors Make Sense?

The goal of our first experiment is to compare the descriptors $R$ learned by the RMN to the topics learned by the topic model baselines. We conduct a word intrusion experiment (Chang et al., 2009b): workers identify an “intruder” word from a set of words that—other than the intruder—come from the same topic. For the topic models, the five most probable words are joined by a highly-probable word from a different topic as the intruder. We use the same procedure for the RMN and GRMN, except that cosine similarity to
Table 1: Three high-precision (top) and three low-precision (bottom) descriptors for the RMN and HTMM, along with labels from an external evaluator and model precision (MP) computed via word intrusion experiments. The RMN is able to learn a variety of interpersonal states (e.g., love, sadness), while the HTMM’s most coherent topics are about concrete objects or events.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>RMN Nearest Neighbors</th>
<th>HTMM Most Probable Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>sadness</td>
<td>regretful rueful pity</td>
<td>violence sword shot</td>
</tr>
<tr>
<td>love</td>
<td>love delightful happiness</td>
<td>boats ship captain</td>
</tr>
<tr>
<td>murder</td>
<td>autopsy arrested homicide</td>
<td>food kitchen mouth</td>
</tr>
<tr>
<td>worship</td>
<td>toil pray devote yourselves</td>
<td>sci-fi suppose earth</td>
</tr>
<tr>
<td>moodiness</td>
<td>glumly snickered quizzically guiltily</td>
<td>fantasy agreed magician</td>
</tr>
<tr>
<td>informal</td>
<td>kinda damn heck guess shitty</td>
<td>military ship captain</td>
</tr>
</tbody>
</table>

4.4 Do Trajectories Make Sense?

While the previous evaluation focused only on descriptor quality, our next experiment compares the trajectories learned by the best RMN model from the intrusion experiment (measured by highest mean model precision) to those learned by the best HTMM model, which is the only baseline capable of learning relationship trajectories. Workers read a plot summary and choose which model’s trajectory best represents the relationship in question. We use the $K = 30$ setting because it provides the best balance between descriptor variety and trajectory interpretability.

For this evaluation, we crawl Wikipedia, Goodreads, and SparkNotes for plot summaries associated with our 1,383 books. We then remove all relationships where each involved character is not mentioned at least five times in the summary, which results in a final evaluation set of 125 relationships. We present workers with two characters, a plot summary, and a visualization of trajectories learned by the RMN and the HTMM (Figure 4). The workers then select the trajectory that best matches the relationship described by the summary.

To generate the visualizations, we first have an external annotator label each descriptor from both models with a single word as in Table 1. For fairness, the annotator is unaware of the underlying models. For the RMN, we visualize trajectories by displaying the label of the argmax over descriptor weights $d_t$ at each time step $t$. Similarly, for the HTMM, we display the most probable topic at each time step.

The results of this task with seven workers per comparison favor the RMN: for 87 out of the 125 evaluated relationships (69.6%), the workers choose the RMN’s trajectory over the HTMM’s. We compute the Fleiss $\kappa$ value (Fleiss, 1971) to correct our inter-annotator agreement for chance and find that

---

9http://www.crowdflower.com

10To reduce visual clutter, we ignore descriptors that persist for only a single time step.
Summary: Govinda is Siddhartha’s best friend and sometimes his follower. Like Siddhartha, Govinda devotes his life to the quest for understanding and enlightenment. He leaves his village with Siddhartha to join the Samanas, then leaves the Samanas to follow Gotama. He searches for enlightenment independently of Siddhartha but persists in looking for teachers who can show him the way. In the end, he is able to achieve enlightenment only because of Siddhartha’s love for him.

Figure 4: An example from the Crowdflower summary matching task; workers are asked to choose the trajectory (here, “A” is generated by the RMN and “B” by the HTMM) that best matches a provided summary that describes the relationship between Siddartha and Govinda (from Siddartha by Hesse).

\( \kappa = 0.32 \), indicating fair agreement among the workers. Furthermore, thirty-four relationships had unanimous agreement among the seven workers; of these, twenty-six were unanimous in favor of the RMN compared to only eight for the HTMM.

4.5 What Makes a Relationship Positive?

While the previous two experiments show that the RMN is more interpretable and accurate than baseline models, we have not yet shown that its insights can aid in drawing connections across various books and genres. As a first step in this direction, we investigate what makes a relationship positive or negative by comparing trajectories from the RMN and HTMM to static affinity annotations from a recently-released dataset (Massey et al., 2015) of fictional relationships. Expected correlations (e.g., murder and sadness are strongly negative descriptors) emerge alongside surprising ones (politics is negative, religion is positive).

The affinity labeling task of Massey et al. (2015) requires workers to describe a given relationship as positive, negative, or neutral. We consider only non-neutral relationships for which two annotators agree on the affinity label and remove all books not present in our own dataset. This filtering step results in 120 relationships, 78% of which are positive and the remaining 22% negative.

Since the annotations are static, we first aggregate our trajectories across all time steps. We compute \( K \)-dimensional “average positive” and “average negative” weight vectors \( a_p \) and \( a_n \) by averaging the relationship states \( d_t \) for the RMN and the document-topic distributions for the HTMM across all time steps for relationships labeled with a particular affinity. Then, we compute the vector difference \( a_p - a_n \) and sort it to produce a ranked list of descriptors, where descriptors with positive differences occur more frequently in positive relationships. Table 2 shows the most positive and most negative descriptors; of particular interest is the large negative weight associated with political relationships from both models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMN</td>
<td>education, love, religion, sex</td>
<td>politics, murder, sadness, royalty</td>
</tr>
<tr>
<td>HTMM</td>
<td>love, parental, business, outdoors</td>
<td>love, violence, politics</td>
</tr>
</tbody>
</table>

Table 2: Descriptors most characteristic of positive and negative relationships, computed using existing annotations. Compared to the RMN, the HTMM struggles to coherently characterize negative relationships. Interestingly, both models show negative predispositions for political relationships, perhaps due to genre bias or class differences.

5 Qualitative Analysis

Our experiments show the superiority of the RMN over various topic model baselines in both descriptor interpretability and trajectory accuracy, but what causes the improved performance? In this section, we analyze similarities between the RMN and HTMM and look at qualitative examples where the RMN succeeds and fails. We also connect the findings of our affinity experiment to existing literary scholarship.

Both models are equally proficient at learning and assigning event-based descriptors (e.g., crime, violence, food). More specifically, the RMN and HTMM agree on environmental descriptions (e.g., boats, outdoors) and graphic sexual scenes (Figure 5, middle).

However, the RMN is more sophisticated with in-
terpersonal relationships. None of the topic model baselines learns negative emotional descriptors such as sadness or suffering, which explains the inaccurate HTMM trajectory of Arthur and Lucy in the left-most panel of Figure 5. All of the topic model baselines learn duplicate topics; in Table 2, one love descriptor is highly positive while a duplicate is strongly negative. The RMN circumvents this problem with its uniqueness penalty (Equation 8).

While the increased descriptor variety is a positive, sometimes it leads the RMN astray. The model largely ignores the love between Charles Darnay and Lucie Manette in Dickens’ A Tale of Two Cities due to book’s sad tone; meanwhile, the HTMM’s trajectory, while vastly simplified, does pick up on the romance (Figure 5, right). While the RMN’s learnable book and character embeddings help, the signal in a span cannot lead to the “proper” descriptor.

Both the RMN and HTMM learn that politics is strongly negative (Table 2). Existing scholarship supports this finding: Victorian-era authors, for example, are “obsessed with otherness . . . of antiquated social and legal institutions, and of autocratic and/or dictatorial abusive government” (Zarifopol-Johnston, 1995), while in science fiction, “dystopia—–precisely because it is so much more common (than utopia)—–bears the aspect of lived experience” (Gordin et al., 2010). Our affinity data comes primarily from Victorian novels (e.g., by Dickens and George Eliot), leading us to believe that that the models are behaving reasonably. Finally, returning to the “extra” meaning of meals discussed in Section 1, food occurs slightly more frequently in positive relationships.

6 Related Work

There are two major areas upon which our work builds: computational literary analysis and deep neural networks for natural language processing.

Most previous work in computational literary analysis has focused either on characters or events. In the former category, graphical models and classifiers have been proposed for learning character personas from novels (Bamman et al., 2014; Flekova and Gurevych, 2015) and film summaries (Bamman et al., 2013). The NUBBI model of Chang et al. (2009a) learns topics that statically describe characters and their relationships. Because these models lack temporal components (the focus of our task), we compare instead against the HTMM of Gruber et al. (2007).

Closest to our own work is the supervised structured prediction problem of Chaturvedi et al. (2016), in which features are designed to predict dynamic sequences of positive and negative interactions between two characters in plot summaries. Other research in this area includes social network construction from novels (Elson et al., 2010; Srivastava et al., 2016) and film (Krishnan and Eisenstein, 2015), as well as attempts to summarize and generate stories (Elsner, 2012).

While some of the relationship descriptors learned by our model are character-centric, others are more events-based, depicting actions rather than feelings;
such descriptors have been the focus of much previous work (Schank and Abelson, 1977; Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Orr et al., 2014). Our model is more closely related to the plot units framework (Lehnert, 1981; Goyal et al., 2013), which annotates events with emotional states.

The RMN builds on deep recurrent autoencoders such as the hierarchical LSTM autoencoder of Li et al. (2015); however, it is more efficient because of the span-level vector averaging. It is also similar to recent neural topic model architectures (Cao et al., 2015; Das et al., 2015), although these models are limited to static document representations. We hope to apply the RMN to nonfictional datasets as well; in this vein, Iyyer et al. (2014) apply a neural network to sentences from nonfiction political books for ideology prediction.

More generally, topic models and related generative models are a central tool for understanding large corpora from science (Talley et al., 2011) to politics (Nguyen et al., 2014). We show representation learning models like RMN can be just as interpretable as LDA-based models. Other applications for which researchers have prioritized interpretable vector representations include text-to-vision mappings (Lazari-dou et al., 2014) and word embeddings (Fyshe et al., 2015; Faruqui et al., 2015).

7 Conclusion

We formalize the task of unsupervised relationship modeling, which involves learning a set of relationship descriptors as well as a trajectory over these descriptors for each relationship in an input dataset. We present the RMN, a novel neural network architecture for this task that generates more interpretable descriptors and trajectories than topic model baselines. Finally, we show that the output of our model can lead to interesting insights when combined with annotations in an existing dataset.

Acknowledgments

We thank Jonathan Chang and Amit Gruber for providing baseline code, Thang Nguyen for helpful discussions about our model, and the anonymous reviewers for their insightful comments. This work was supported by NSF grant IIS-1320538. Boyd-Graber is also partially supported by NSF grants CCF-1409287 and NCSE-1422492. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.
References


Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In Association for the Advancement of Artificial Intelligence.


Vinodh Krishnan and Jacob Eisenstein. 2015. “You’re Mr. Lebowski, I’m The Dude”: Inducing address term formality in signed social networks. In Conference