UCLA
UCLA Previously Published Works

Title

12th ACM Conference on Web Science, WebSci '20

Permalink

Authors

Shahsavari, Shadi
Ebrahimzadeh, Ehsan
Shahbazi, Behnam
et al.

Publication Date

2020-07-01

DOI

10.1145/3394231

Peer reviewed

An Automated Pipeline for Character and Relationship Extraction from Readers' Literary Book Reviews on Goodreads.com

Shadi Shahsavari[1], Ehsan Ebrahimzadeh[1], Behnam Shahbazi[2], Misagh Falahi[1], Pavan Holur[1], Roja Bandari[1], Timothy R. Tangherlini[3], Vwani Roychowdhury[1]

1 Electrical and Computer Engineering, UCLA, Los Angeles, CA, USA
2 Computer Science, UCLA, Los Angeles, CA, USA
3 Scandinavian Section, UCLA, Los Angeles, CA, USA

ABSTRACT

Reader reviews of literary fiction on social media, especially those in persistent, dedicated forums, create and are in turn driven by underlying narrative frameworks. In their comments about a novel, readers generally include only a subset of characters and their relationships, thus offering a limited perspective on that work. Yet in aggregate, these reviews capture an underlying narrative framework comprised of different actants (people, places, things), their roles, and interactions that we label the "consensus narrative framework". We represent this framework in the form of an actant-relationship story graph. Extracting this graph is a challenging computational problem, which we pose as a latent graphical model estimation problem. Posts and reviews are viewed as samples of sub graphs/networks of the hidden narrative framework. Inspired by the qualitative narrative theory of Greimas, we formulate a graphical generative Machine Learning (ML) model where nodes represent actants, and multi-edges and self-loops among nodes capture context-specific relationships. We develop a pipeline of interlocking automated methods to extract key actants and their relationships, and apply it to thousands of reviews and comments posted on Goodreads.com. We manually derive the ground truth narrative framework from SparkNotes, and then use word embedding tools to compare relationships in ground truth networks with our extracted networks. We find that our automated methodology generates highly accurate consensus narrative frameworks: for our four target novels, with approximately 2900 reviews per novel, we report average coverage/recall of important relationships of >80% and an average edge detection rate of >89%. These extracted narrative frameworks can generate insight into how people (or classes of people) read and how they recount what they have read to others. [1]

KEYWORDS

graph theory, narrative theory, machine learning, knowledge base

----

[1] Our code will be posted at https://github.com/Roychowdhury-group/goodreads

----

1 INTRODUCTION

1.1 Motivation

Literary fiction attracts large reading audiences both in the United States and internationally. A National Endowment for the Arts survey reveals that, despite a long steady decline in literary reading in the United States, the number of American adults who read at least one work of fiction a year, even after excluding books read for school or work, still hovers around 43%. Social media, although often criticized for contributing to the decline in literary reading, has also offered opportunities for communities of readers to interact and engage in ongoing conversations, perhaps thereby reducing the otherwise negative impact of social media on reading. Book forums on social media provide readers an opportunity to share their experiences of reading and can, for some works of fiction, engender long running conversations about nuanced aspects of the work in question. These discussions range from explorations of twists and turns in the plot, to simple declarations of admiration for or familiarity with certain actants (characters, places, things). [26][12] *Taken individually*, book commentaries and reviews provide a highly individualized perspective on a work of fiction, focusing only on a few actants and their relevance to the narrative. *Taken together*, these comments provide insight into a broader reader consensus of a novel's overarching narrative framework, comprising a majority of the actants and their nuanced relationships.

1.2 Objectives and Challenges

In our work, we assume that we are given thousands of user reviews of a particular novel from a social cataloging/review website such as Goodreads.com. Given such a corpus, we ask the following questions: (i) Can one *automatically discover all the primary actants* as well as meta-actants (authors, actors and actresses from film adaptations, etc.) that are mentioned across all of the book reviews for a given novel? (ii) Can one also *discover and meaningfully cluster all the inter-actant relationships* that these reviews include? The results of goals (i) and (ii) provide, when properly thresholded and weighted, a representation of the consensus model of the novel

as perceived by those readers who review the book. Inspired by the actantial narrative model of Algirdas Greimas [8], we represent these results as an automatically generated narrative network, where nodes are actants and edges are directed multi-edges annotated with the extracted relationships. (iii) Finally, *given an expert generated ground truth narrative network*, can one *automatically compare that ground truth network with the auto-generated summary narrative framework network* and compute meaningful metrics such as recall and precision?

Solving the above problems is tantamount to developing a view of the reviewers' consensus about a target novel, as readers recollect and review the actual cast of actants and their inter-actant relationships.

The more often that an actant or relationship appears in the corpus, the more heavily it is weighted in the network graph. Importantly, the related methodologies presented here can be extended well beyond the realm of literary fiction to derive narrative frameworks undergirding nearly any collection of documents. We focus on literary fiction because of the unusual (for cultural datasets) presence of a ground truth against which to measure the accuracy of our results.

To construct the actant relationship narrative graph, we start with a dependency tree parsing of the sentences in each review and extract various syntactic structures, such as the Subject (captured as noun argument phrases), Object (also captured as noun argument phrases), actions connecting them (captured as verb phrases), as well as their alliances and social relationships (captured as explicitly connected adjective and appositive phrases; see Table 2; see the Methodology section for the tools used and relationship patterns extracted in this paper). *The task of aggregating these extracted phrases into a single narrative network poses unique computational challenges.*

First, as these extractions are both varied and extremely noisy, we need to reduce ambiguity across entity mentions. For example, in reviews of *The Hobbit*, Bilbo Baggins is referred to in numerous ways, including "Bilbo" (and its misspelling "Bilbos"), "The Hobbit", "Baggins" and "the Burgler" or "the Burglar". We refer to this disambiguation task as the *Entity Mention Grouping* (EMG) problem. Humans solve the EMG problem by using context: for the different mentions of a character to be the same, they must have the same relationships with other characters. The human ability to disambiguate in this manner has proven difficult to replicate with computational tools.

Second, the same challenge applies to inter-actant relationships. For example, the relationship "create" between Dr. Frankenstein and the monster in the novel *Frankenstein*, can be referred to by a cloud of different phrases, including "made", "assembled", and "constructed". To solve this ambiguity, one must computationally recognize that these words are contextually synonymous and identify the group as constituting a single relationship. To make matters more challenging, there are often numerous different relationships between the same actant pair. The decision tree parsing step produces an unordered list of phrases, which then has to be clustered into semantically similar groups, where each group captures one of the distinct relationships. For example, the extracted relationship phrases between Dr. Frankenstein and the monster include {*created, destroying, kill, regretting, constructed, denied, hates, disgusted, made,*

assemble, blaming, abandon, runs away}. These phrases, however, contain sample phrases from at least three distinct relationships: <u>Create</u>: [*created, constructed, made, assemble*], <u>Destroy</u>: [*destroying, kill*], and <u>Deny</u>: [*denied, hates, disgusted, blaming, abandon, runs away, regretting*]. We label this problem of reliably clustering relationships as the *Inter-actant Relationship Clustering* (IARC) problem.

Finally, the task of *quantitative evaluation* – comparison of the extracted networks with ground truth networks – shares many of the same challenges as the previous two tasks. One has to *semantically align* any expert-created network with the automatically created one. For example, one should be able to match an expert annotated relationship of "X → Captured → Y," to an automatically aggregated relationship, such as "Y →{ escaped, rescued} from→ X."

1.3 Related Work

Numerous studies have explored book review collections while several other works have attempted to recreate story plots based on these reviews [26, 29, 30]. The sentence-level syntax relationship extraction task has been studied widely in work on Natural Language Processing and Open Information Extraction [1, 6, 7, 16, 21, 31] as well as in relation to the discovery of actant-relationship models for corpora as diverse as conspiracy theories and national security documents [15, 20]. There is considerable recent work on word and phrase embedding for encoding semantic similarity. While word embedding methods such as word2vec, fastText and GloVe [3, 14, 17] yield vectors that are context invariant, more recent models such as ELMo and BERT [5, 18] allow for polysemy (context-dependent embedding). This polysemic feature allows entire phrases to be encoded to both word-level and phrase-level embedding. We use BERT embedding in this paper.

While there is work, such as Clusty [19], which categorizes entities into different categories in a semi-supervised manner, the category examples are fixed. Similarly, works such as ConceptNet [23] use a fixed set of selected relations to generate their knowledge base. Other recent entity mention grouping work [9] seeks to map entity mentions via context vectors produced as an aggregated feature from high-level document metadata and proximal phrases to the mention within the text. Similar work in story graph applications [11] create co-scene presence character networks predicated on higher-level annotated knowledge, such as joint scene presence and/or duration of dialogue between a pair of characters. Moreover, these works assume perfect reliability in character mentions (thus obviating the need for the critical step of Entity Mention Grouping that is needed for social media reviews), an assumption we cannot make given our data or data from similarly informal domains.

A major challenge in work on reader reviews of novels is that predefined categories for novel characters and for the diverse inter-character relationships do not exist. In addition, document level features are missing while the proximal text is sparse due to the inherent size of a review (or tweet, comment, opinion, etc.). An unsupervised scheme such as ours for grouping entity mentions into characters and clustering of relationships into semantically distinct groups, as an approximate imitation of human processes, has not been addressed previously.

	# of posts	# of sentences
Frankenstein	2947	38432
The Hobbit	2897	37529
Of Mice and Men	2956	30205
To Kill a Mockingbird	2893	33000

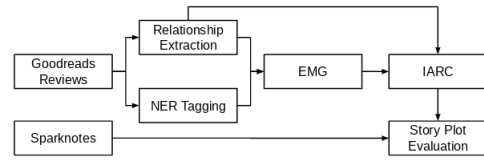**Table 1: Data description and size.**

Figure 1: Pipeline to extract actant-relationship graphs. Our contributions introduce the Entity Grouping and the Inter-actant Relationship Clustering blocks

1.4 Outline of the paper

In Section 2, we describe our data, our selection of the four novels for analysis, and our method for generating ground truth narrative frameworks. In Section 3 we describe our methodology and how we solve the EMG and IACR problems. In Section 4, we provide an overview of the limitations of our pipeline. In section 5, we present our results and evaluation, and in section 6, we discuss the findings. Lastly, in section 7, we suggest potential improvements that can be incorporated into the pipeline in future work.

2 RESOURCES

We use reader reviews of four works of fiction from the community forums on Goodreads: *Frankenstein* (1818); *Of Mice and Men* (1937); *The Hobbit* (1937); and *To Kill a Mockingbird* (1960) [10, 22, 24, 28]. The works were chosen from the list of the most frequently rated books on the Goodreads site (number of ratings $> 500,000$). For highly rated novels, the number of reviews is also quite high, although significantly lower than the number of ratings. For example, *The Hobbit* has been rated over 2.5 million times, but has 44, 831 reviews (at the time of our data collection). For each of the novels, we downloaded the maximum allowed three thousand reviews given the Goodreads API limits on review requests.

The reviews were harvested using a crawler specifically designed for this project. Not all reviews were useful since numerous posts were either spam, posts on different topics, or written in languages other than English. Other reviews were either too short to include meaningful content, or so garbled as to be unintelligible. After filtering the reviews, we were left with a corpus of 8693 usable reviews: *Frankenstein* (2947), *The Hobbit* (2897), *Of Mice and Men* (2956), and *To Kill a Mockingbird* (2893). We discovered two types of phrases in the reviews: (i) Opinion phrases that reflected the readersâĂŹ opinions about the book, the author, or the various characters and events. Relationships extracted from these phrases are the dominant ones when aggregated over all readersâĂŹ posts, which is not surprising given that these posts are intended to be reviews. (ii) Plot phrases that describe what happened to a subset of the actants, and how they interacted with each other. These phrases contain both the actants and their relationships, and are of primary interest to us.

Although our initial study corpus consisted of sixteen novels, we selected these four novels for detailed analysis on the basis of the broad disparity in their narrative structures, large variability in the number of characters, and a broad range of character relationships. For example, *The Hobbit* can be characterized as a multi-episodic, linear narrative that takes place across many different settings in an elaborate fantasy world, and includes a large cast of both human and non-human characters, instantiating an elaborate version of a standard hero's journey plot. *Of Mice and Men*, by way of contrast, is a short novella with a limited cast of characters that takes place in a highly localized, realistic setting, and represents a straightforward version of VonnegutâĂŹs âĂIJFrom bad to worseâĂİ plot. *Frankenstein*, although told partly in flashback, has a largely linear plot and a limited cast of characters, with a strong central figure and a relatively clear villain, although this is complicated by its use of nested narratives. Finally, *To Kill a Mockingbird* has an overlapping set of complex characters with multiple subplots.

For our ground truth narrative framework graphs, we relied on the online SparkNotes resource for each of the four chosen novels. SparkNotes is a corpus of freely available, professionally generated summaries of works of fiction, and provides us with a list of actants, as well as a chapter level plot summary. These fine-grained summaries allowed us to manually create an actant-relationship narrative framework graph for each novel. These ground truth graphs were coded independently by two experts in literature, and a third expert was used to adjudicate any inter-annotator disagreements.

Reviewers who post to Goodreads have a variety of motivations for posting. The majority of reviewers use the site as part of a social network focused on reading, with the gender balance of active reviewers skewing slightly toward women [27]. There appear to be several categories of active reviewers on the Goodreads site, including students reviewing books as part of school assignments, members of book clubs, and people who aspire to become professional book reviewers. We make no discrimination as to classes of reviewers, but rather consider each review equally, as our goal is to understand the aggregate narrative model of a reviewed book. At the same time, we recognize that reviews of a book are often conditioned by the pre-existing reviews of that same book, including reviews such as those found in SparkNotes, Cliff Notes, and other similar resources. In certain cases, we recognize that these reviews may be influenced by the filmed adaptations of the target novels or professionally written summaries.

3 METHODOLOGY

Our methodology focuses on the underlying structure of the narrative framework that captures how a storytelling instance emerges via a collective negotiation process. Each post to a forum describes relationships among only a subset of actants (which are yet not known to our automated algorithms). To write a sentence, a reviewer first picks a context $C_i \in C$ and then samples an underlying context-dependent network $G_{C_i}(V_{C_i}, E_{C_i})$ (to be estimated by the algorithm) by drawing a pair of actants (A_k, A_j) according to a conditional actant recall distribution across all the actants, $p_{C_i}(A_j)$.

A context could represent a particular situation in the plot. For example, when someone wants to recount the scene in *Frankenstein* where Dr. Frankenstein creates the monster, then certain actants and relationships are described much more often than others.

Following this, the reviewer draws a relationship for the pair (A_k, A_j) from a distribution associated with the context-dependent edges: $D_{(E_{C_i}, (j,k))}(\mathcal{R})$. The writer then composes the review according to these outcomes by choosing the proper words and syntax. In particular, the reviewer chooses noun phrases (as mentions of the actants A_j and A_k) and the associated verb/relationship phrases (or other syntactical constructs) for the sampled relationship.

Recall that we have neither any knowledge of the underlying actants nor of the contexts that define different semantic relationships among them. After syntax-based relationship extractions from the reviews, we have multiple mentions/noun-phrases for the same actants, and multiple semantically equivalent relationship phrases to describe different contexts. In order to accurately estimate the different contexts C_i, actant frequency distributions $p_{C_i}(A_j)$, and the relationships $D_{(E_{C_i}, (j,k))}(\mathcal{R})$, we must aggregate the different mentions of the same actant into a single group. In order to do that, we need to consider relationships: two mentions refer to the same actant only if the key relationships with other actants are semantically identical. Thus, the estimations of entity mention groups and relationships need to be done jointly.

The following subsections describe our approach to the estimation of the aggregate narrative network in the three steps of our pipeline presented in figure 1: (i) Syntax-Based Relationship Extraction, (ii) Entity Mention Grouping (EMG), and (iii) Inter-actant Relationship Clustering (IARC). The resulting graph constitutes an end-state ranked consensus model of all actants and relationships. The evaluation of our results focuses on the similarity of the ground truth and learned narrative graph based on a matching of actants and their contextual relationships. The frequency distributions of the actants, p, and relationships, D, can be estimated based on the counts of the occurrences of the associated groups of phrases. Currently, we use a threshold to decide whether an actant or a relationship is included in the consensus narrative graph. We leave a more detailed study of these frequency distributions and their relationship to reader consensus to ongoing and future work. These probabilities encode the relative importance of the different actants and relationships in ways not captured by the thresholded network. For example, in The Hobbit, the actant node "Ring" has only a single relationship edge (i.e., "Bilbo" finds the "Ring") yet, due to the centrality of the "Ring" to the story, it has a frequency rank in the top ten among all noun phrases.

Syntax-Based Relationship Extraction: Each sentence in the text corpus is processed to extract specific patterns of syntax relationship tuples in the form of (arg_1, rel, arg_2) where arg1 and arg2 are noun phrases, and rel is a verb or other type of phrase. Our relation extraction combines dependency tree and Semantic Role Labeling (SRL) [7][13]. As opposed to limiting our extractions to agent-action-target triplets, we design a set of patterns (for example, Subject-Verb-Object (SVO) and Subject-Verb-Preposition (SVP)) to mine extractions from dependency trees using the NLTK package and various extensions. The patterns are based on extensions of Open Language Learning for Information Extraction (OLLIE) [21]

and ClauseIE [4]. Next, we form extractions from the SENNA Semantic Role Labeling (SRL) model. We combine dependency-based extraction techniques with SRL to increase the recall of our system. A list of all the syntax relationship patterns, their definitions, and related examples are provided in the GitHub link for our research.

Following these steps, we apply cleaning and de-duplication techniques to select unique and high precision extractions. Relationship tuples scraped from reviews only include those entity mentions that match or exceed a frequency lower bound (≥ 50).

Entity Mention Grouping (EMG): As a semantically identifiable character in a book is expressed in reviews as diverse entity mentions, it is necessary to group these mentions and label them with the same character label.

Let the frequently-occurring set of entity mentions be M and let R_{ik} be the relationships between entity mention m_i and m_k, where m_i is the Subject and m_k be the Object. The set R_{ki} then denotes the relationships when the roles are reserved. First, we note that if there is a relationship triplet (Subject $= m_i$, Verb, Object $= m_j$) then clearly m_i and m_j are mentions of different actants and are not to be grouped together. In order to avoid any noise-induced exclusion of such a pairing, we consider a pair m_i, m_j as incompatible if $|R_{ij}| + |Rji| \geq \gamma$. Based on our observation of the low frequency of noisy relationships, the hyperparameter γ is set to 3 in this paper. In the following we assume that for each mention m_i we have removed all incompatible nodes m_j.

Intuitively, two compatible mentions m_i and m_j correspond to the same actant if, for every other mention m_k, the relationships between the pair (m_i, m_k) are semantically the same as the relationships between the pair (m_j, m_k). In practice, different mentions of the same actant will share only a subset of the relationships when aggregated over all the extractions. In the following we provide an algorithm to quantify this intuitive idea that yields robust EMGs.

Let $T_{ik} = H(R_{ik})$ describe the set of headwords in R_{ik}. Also let G be the directed bipartite graph from the entity mentions M to M (see Fig. 2) with the edges representing the relationships between the entity mentions. We would like to find an Entity Mention Grouping (EMG) function $g : M \rightarrow [1, ..., N]$, $N \leq |M|$, where (i) if $g(m_i) = g(m_j) = k$ then entity mentions (m_i, m_j) are grouped together to form the kth actant. Moreover, (ii) we want the groups to be complete: that is, for two groups $g^{-1}(k_1)$ and $g^{-1}(k_2)$ (with $k_1 \neq k_2$ and $k_1, k_2 \in [1, ..., N]$), the entity mentions are semantically similar within each set and are semantically differentiated across the sets. To measure semantic similarity between m_i and m_j, we consider the following measure involving another mention m_k:

$$s_{(ij)k} = \Pr(T_{ik}|T_{jk}) + \Pr(T_{jk}|T_{ik}),$$
$$\Pr(T_{ik}|T_{jk}) = \frac{|H(R_{ik}) \cap H(R_{jk})|}{|H(R_{jk})|}. \tag{1}$$

To understand why $s_{(ij)k}$ is an effective similarity measure, consider the following cases: (i) If $H(R_{ik}) = H(R_{jk})$, implying that m_i and m_j share the exact relationships with m_k and hence should be grouped together, then $s_{(ij)k}$ achieves the maximum value of 2, (ii) the m_j mention of an actant occurs less frequently then m_i and is reflected by $H(R_{ik}) \subset H(R_{jk})$, then $s_{(ij)k} \geq 1$. This captures the case where m_j shares all its relationships with m_i but not vice versa, (iii) m_i and m_j are indeed mentions of different actants, in

which case $|H(R_{ik}) \cap H(R_{jk})|$ is expected to be a lot smaller than both $|H(R_{ik})|$ and $|H(R_{jk})|$ and $s_{(ij)k} \ll 1$.

To ensure that we compute similarity when m_k is the Subject, we define an analogous similarity score:

$$s_{k(ij)} = \Pr(T_{ki}|T_{kj}) + \Pr(T_{kj}|T_{ki}),$$

$$\Pr(T_{ki}|T_{kj}) = \frac{|H(R_{ki}) \cap H(R_{kj})|}{|H(R_{kj})|}. \tag{2}$$

Finally, the score matrix S is computed where the score S_{ij} between m_i and m_j aggregates the measure on all feasible $m_k \in M - \{m_i, m_j\}$ and provides a metric for similarity across all entity mentions:

$$S_{ij} = \sum_{m_k \in M - \{m_i, m_j\}} s_{(ij)k} + s_{k(ij)}. \tag{3}$$

The grouping function g is now constructed as follows: For every entity mention m_i, the scores in the vector S_i are ranked in descending order. We next introduce two hyperparameters for each novel, $\alpha, \beta \geq 0$, such that an entity mention m_i is grouped with m_j only if the score S_{ij} satisfies: $S_{ij} \geq \alpha$ and $\frac{S_{i(j-1)}}{S_{ij}} \geq \beta$ (for $j \geq 2$).

We compute α from novel-specific distribution statistics. In particular, we compute the histogram of all non-zero S_{ij} and compute α as the 75th percentile (i.e. 25% of S_{ij}'s are $\geq \alpha$). For all considered books (except *To Kill a Mockingbird* where $\alpha = 2.6$), $\alpha = 2.0$. The hyperparameter β is set to 2.

The parameters α and β are similar to those in works such as the Elbow K-Means method [2], in which β correlates to inertia if the scores S_i correlate to the distortion, and α provides a means of resolution if the elbow is unreliable (common in our model for rarer entity mentions).

The entity mention groups, once found, are labeled with the most frequent mention in the respective groups. Empirically, these automatically computed labels match the ground truth entities as derived from SparkNotes.

Inter-actant Relationship Clustering (IARC): The aggregated entity mentions captured in g are fed back into the standard relationship extraction task. Then, the relationships aggregated between any pair of actants, represented by their respective entity mention groups (e.g.: $A_1 = g^{-1}(k_1)$ and $A_2 = g^{-1}(k_2)$) is computed as:

$$R_{A_1 A_2} = \bigcup_{p \in A_1, \; q \in A_2} R_{pq}. \tag{4}$$

$R_{A_1 A_2}$ is a richer and potentially multi-modal set of relationships. This process enables a form of transfer learning, aiding relationship extractors in identifying connections at a higher semantic level of characters and not merely at the level of entity mentions. The associated relationship clusters are found using the cosine similarity measure in the BERT embedding space (Algorithm 1).

Algorithm 1: Inter-actant Relationship Clustering

Result: $C_{A_1 A_2}$
$\hat{R}_{A_1 A_2}, C_{A_1 A_2} = \{\}$;
for $r \in R_{A_1 A_2}$ **do**
   append BERT(r) to $\hat{R}_{A_1 A_2}$
end
$C_{A_1 A_2}$ = Elbow K-Means Method on $\hat{R}_{A_1 A_2}$

$C_{A_1 A_2}$ is the set of clusters of relationships that describe the multi-modality in $R_{A_1 A_2}$. For each cluster C we compute its dispersion (using the cosine similarity measure), β_C. We retain only those clusters with β_C greater than a threshold (here, we set it to 0.8) as a valid semantic relationship group.

Evaluation: We compare these relationship clusters to the ground truth relationships between characters (e.g.: $J_{A_1 A_2}$). We aim to find a mapping $h_{A_1 A_2} : J_{A_1 A_2} \rightarrow C_{A_1 A_2}$. This process is described in Algorithm 2, where $f_{cos}(a, b)$ is the function to compute the cosine similarity between a, b, and β_C is the dispersion of a cluster C using the cosine similarity measure. Thus, a ground truth relationship phrase is mapped to an automatically clustered semantic group only if its embedding is close enough to the centroid of the cluster.

Algorithm 2: Evaluation: Mapping Relationship Clusters to Ground Truth

Result: $h_{A_1 A_2}$
for $C \in C_{A_1 A_2}$ **do**
   if $\beta_C \geq 0.8$ **then**
      if $\max\limits_{r \in C, \; j \in J_{A_1 A_2}} f_{cos}(r, \mathsf{BERT}(j)) \geq 0.8$ **then**
         $h_{A_1 A_2}(j) = C$
      end
   end
end

Similar to the EMG task, the clusters are well differentiated, resulting in high-fidelity labels. Furthermore, Algorithm 2 seeks to approximate a maximum likelihood estimation problem, where $\mathcal{L}$ represents the cosine similarity f_{cos} implemented with thresholds:

$$h_{A_1 A_2}(j) = \underset{C \in C_{A_1 A_2}}{\operatorname{argmax}} \; \mathcal{L}(C, j), \; \forall j \in J_{A_1 A_2}. \tag{5}$$

4 LIMITATIONS

Data can be noisy, particularly when social media posts, which are informal by nature, are the primary source. This informality creates noise in the relationship extraction phase. A missing punctuation mark, for example, can significantly change the dependency tree structure and lead to erroneous extractions of both the arguments and the relationship phrases.

Other parts of our pipeline are equally sensitive to noise, including pronoun resolution and BERT embeddings. While pronoun resolution is needed to improve coverage (that is, to capture relationships amongst entity mention references when they are expressed in terms of pronouns), the process adds additional noise by occasionally resolving pronouns to the wrong entity mentions. Error from pronoun resolution is more noticeable in relation to rare words. For example, in the sentence, "The example their single father Atticus sets for *them* is one all parents wish we could parallel.", *them* is mapped to the single character *Dill. Dill* is among the characters mentioned least frequently in reviews of *To Kill a Mockingbird*. In such a scenario, the extracted relationships have a low fidelity because of the sparse sample space. In addition, while the BERT embeddings that we use for this paper provide useful vectors in cosine-measured k-means clustering, the approach also suffers from sensitivity to noise.

Entity	Descriptors
	The Hobbit
Bilbo	(a, the, simple, clean) hobbit, a burglar, baggins, hero, protagonist
Smaug	(a, the, horrible, vicious) dragon
Gandalf	(a, the, wise) wizard
	Frankenstein
Frankenstein	(a, the, fantasy) book, (the, a) creator, (a, the) doctor
Monster	(his, a, the) creation
	Of Mice and Men
George	a small (man,-, in height), Lennie's (caretaker, best friend, father figure, protector)
Lennie	(the, pitiful, unique, favorite) character, George's (foil, best friend)
	To Kill a Mockingbird
Jem	(big, the older, strong) brother
Atticus	(the, loving, ordinary, her) father
Scout	(a, hotheaded, young, an interesting) Tomboy

Table 2: Examples for Appos and SVcop candidate descriptors for entity mentions across the four novels.

Entity Mention (m_i)	Ranked Similarity Scores for other Mentions (m_j) $(S_{ij}$'s, see Eq. 3)
Bilbo	baggins,42.14 hobbit,14.47 burglar,3.80
Burglar	bilbo,3.80 dwarves,2.79
Wizard	gandalf,22.49 gandolf,7.00 grey,5.34 thorin,3.32
Hobbit	bilbo,14.47 baggins,6.06

Table 3: Given two entity mentions (m_i, m_j), the similarity score S_{ij} (see Eq. 3) measures the semantic "fungibility" of the mentions (i.e., whether both mentions are used interchangeably to refer to the same actant). The table shows several popular entity mentions (m_i's) and the similarity scores of other candidate mentions, m_j's, in *The Hobbit*. Clearly, the mentions [Bilbo, baggins, Hobbit, Burglar] form a clique representing the same actant, *Bilbo Baggins*. One can also see the emergence of another EMG [Wizard, Gandalf, Gandolf, Grey] for the actant *The wizard*.
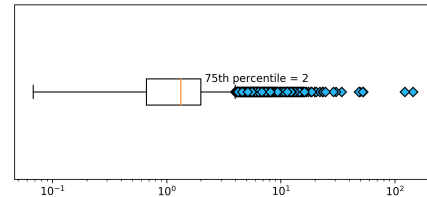
Using SparkNotes as a ground truth also raises some issues, as the summaries in these reader guides are less detailed than the novels that they summarize. Consequently, comparing our extractions to the limited relationships described in SparkNotes means that some of our discovered relationships, which may be in the novel but not in the SparkNotes summary, are improperly evaluated (i.e. the relationship exists in both the target novel and our extractions but is missing in SparkNotes). For example, while our extractions reveal that George cares for or loves Lennie in *Of Mice and Men*, this relationship is missing from the SparkNotes summary. Similarly, certain actants or relationships that exist in the ground truth summaries may simply be absent from the reader review corpus, as is the case for certain Frankenstein actants such as M. Krempe. Our methods are not able to discover actants or relationships that do not appear in reader reviews–this elision of characters and relationships, however, may be indicative of interesting aspects of reader review practice.

5 RESULTS

We first examine the syntactic method of establishing actant-actant relationships for clustering. In Table 2, the Appos and SVCop relationships suggest not only limiting sentence-level associations, but also semantically invariant associations mentioned explicitly in the reviews. While this syntactic approach may work in many situations, book reviewers often *assume* a basic shared knowledge of the plot of a novel. This assumption dissuades reviewers from explicitly writing out the relationships between actants. In addition, book reviews are not very descriptive in general, focusing more on specific plot points or a character's trajectory. This tendency in book reviews further weakens direct Appos and SVCop actant-relationship extraction.

Figure 4: A Box plot of the similarity scores, S_{ij}'s (see Eq. 3), for all entity mention pairs (m_i, m_j) in *The Hobbit*. For any entity mention, m_i, its Entity Mention group (EMG) is first pruned to contain m_j's with scores, $S_{ij} \geq \alpha$, where α is the 75th percentile of the score distribution. From the plot we find $\alpha = 2$. This EMG is further pruned by first sorting the list by their scores, and then ensuring that the ratio of any two successive scores is bounded below, i.e., $\frac{S_{i(j-1)}}{S_{ij}} \geq \beta$ (for $j \geq 2$). We found that $\beta = 2$ provided a good cutoff.

We applied our EMG algorithm to obtain the actants as documented in Table 4. Table 3 and Fig. 4 provide example statistics obtained during the execution of the EMG algorithm. Each actant, once formed, aggregates relationships that the individual entity mentions imply. The clustering of relationships aggregated under the now-formed entity mention groups yield higher granularity and confidence in the IARC task, as semantic connections between entity mentions reinforce the relationships from one actant to another. This effect is observed across the four books as shown in Fig. 3. The relative size of words in the figure correlate to their frequency in the aggregated relationships between the entity mention groups.
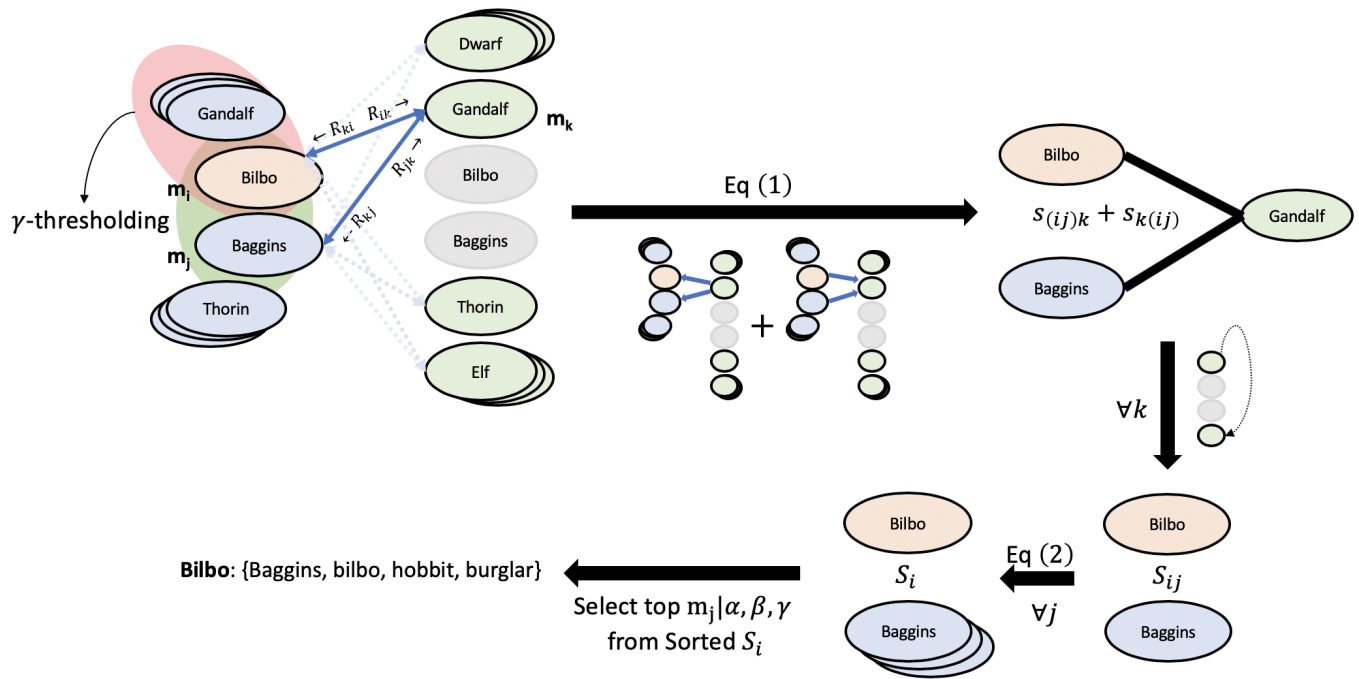
Figure 2: The pipeline of the EMG task shows the formation of the bipartite graph G with the computation of the Score Matrix S, along with hyperparameters α, β, γ
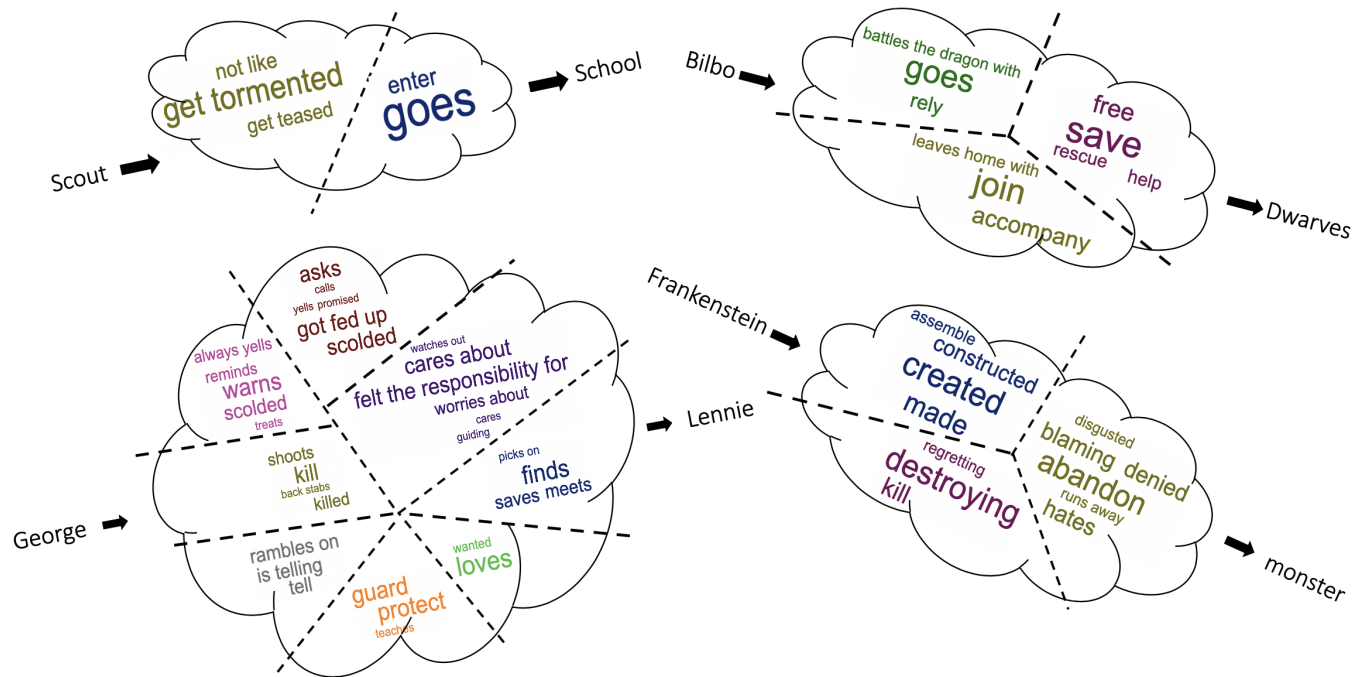

Figure 3: Directed and clustered relationships emergent after IARC between 2 actants per novel. In clockwise direction from top left: from Scout to School in *To Kill a Mockingbird*, from Bilbo to Dwarves in *The Hobbit*, from Frankenstein to Monster in *Frankenstein* and from George to Lennie in *Of Mice and Men*.

Figure 5: Evaluation phase: matching 2 clusters of relationships in Of Mice and Men, from George to Lennie, to ground truth labels, in accordance to Algorithm 2. β_c determines the set of edges.

The task of mapping relationship clusters to particular ground truth labels is shown for the "converse" and "warn" clusters from George to Lennie in *Of Mice and Men* (Figure 5). The rich clusters, in comparison to the ground truth labels from SparkNotes suggests recall as a good measure of performance for our pipeline. A summary of our results for all four books including recall is presented in Table 5.

In general, the relationships between actants reveal a high degree of consistency with the ground truth graph. The largest divergences consist of missed relationships rather than the identification of non-existent relationships, although these occur occasionally. This latter group of relationships is often the attribution of a relationship, such as the killing of Smaug (the dragon in *The Hobbit*), to an important character such as Bilbo Baggins. In other words, many readers *incorrectly believe* that Bilbo killed Smaug. Another small set of spurious relationships, including one that suggests that Jem killed Bob Ewell in *To Kill a Mockingbird*, are caused by reader confusion, âĂIJwhat-ifâĂİ scenarios or, more commonly, incorrect pronoun resolution and aggregation. Apart from the relatively infrequent misattribution of relationships, the reduction in relationships aligns with the corresponding reduction in the number of actants connected to the central component of the story graph.

Figure 6 depicts the narrative framework graph for *The Hobbit* with blue nodes representing ground truth actants or meta-actants. We also show four examples of resolved actants or meta-actants (colored green) not found in the ground truth: **Tolkien**:[tolkein, author], **novel**:[book, fantasy, story, novel], **Fili**:[fili] and **Film**:[film, movie, scene]. Blue edges represent relationships in the ground truth found by using our methods (frequency threshold ≥ 5), while red edges represent undetected ground truth relationships. Green edges connecting to green nodes (frequency threshold ≥ 10) are edges that cannot be verified; we include them to indicate the richness of the extracted graph as opposed to the ground truth. Figure 7 shows a graph similar to Figure 6 after the deletion of low frequency edges (≤ 5), and represents the core structure of the narrative covered in the reviews conditioned on the SparkNotes ground truth.
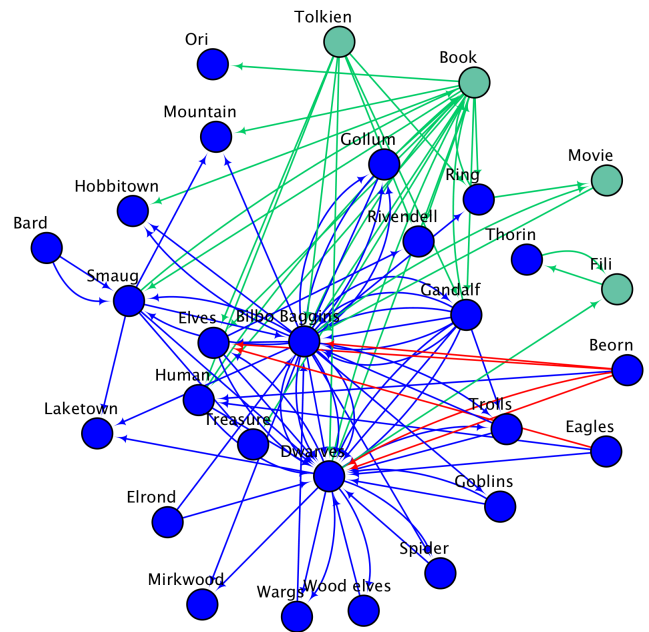

Figure 6: Narrative Framework graph of *The Hobbit*. Green nodes are extracted entities not part of the ground truth, red edges are ground truth edges which were not detected by the algorithm, blue edges are detected ground truth edges.
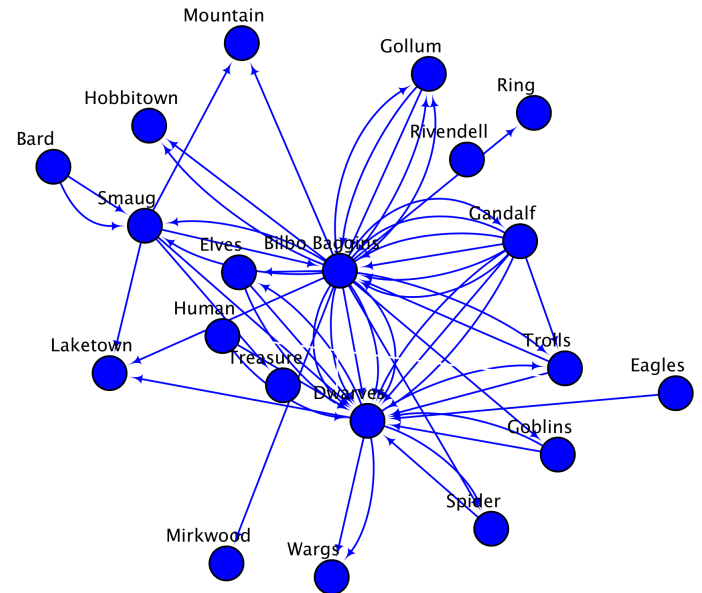

Figure 7: Narrative Framework graph of *The Hobbit after* thresholding on the frequency of relationship. Blue edges have at least 5 relationship instances.

There are shared structural properties (disregarding the specific relationships they encode) that can be used to automatically distinguish between actual characters in the novels and the various meta-actants. For example, the meta-actant **Tolkien** (the green node at the top center of Figure 6) has only outgoing edges, indicating that Tolkien appears only as the subject in any inferred relationship triplet. This lack of incoming edges is a significant feature of meta-actants: An important character in a novel usually has bi-directional relationships with other characters. An author of the novel, on the other hand, usually "acts" on the characters; hence the corresponding node is directionally isolated. The incoming edges for the meta-actant "Book" are all attributable to phrases such as " character XNZ is portrayed *in the* book/novel". A simple filtering of these preposition-induced relationships directionally isolates the meta-actant "Book." Further structural explorations of the derived networks, such as measures of centrality and importance of different characters, are part of our ongoing work.

6   DISCUSSION

The results support the idea that readers, when summarizing a novel, tend to reduce the scope of the story and to focus on the most memorable aspects of the plot, here modeled as inter-actant relationships. In the reviews we studied, people converge on a set of main actants and relationships that map well to a core set of actants and relationships in the ground truth summaries, suggesting that people are relatively adept at summarizing even complex novels. As part of their summaries, however, people tend to simplify. This simplification may be related to cognitive limits on the number of real-world relationships that a person can keep in mind.

Since reviews tend to be short, when compared to the length of the work summarized, it is not surprising that people reduce both the number of actants, particularly in works with very large casts of characters such as *The Hobbit*, and the relationships between those actants. The inter-actant relationships are also simplified in the reader reviews. Readers can simplify complex plots, such as that in *To Kill a Mockingbird*, into relatively straight forward stories of conflict, strategies to address that conflict, and the result of the use of those strategies. The reduction of plot complexity may also be influenced by the abstraction of the novel in other media. For certain books, such as *The Hobbit*, recent films have been highly successful, and it is quite possible that movie watching has had some impact on reader reviews. The same may apply to the other books in this study given, for example, the numerous references to the actor Gregory Peck in the reviews of *To Kill a Mockingbird*. Although we have not done so here, it may be interesting to compare reader reviews of filmatized novels to the summary story graphs for those films.

7   CONCLUSION

The approach we describe here is widely applicable to other crowd-sourced review sites such as Rotten Tomatoes and Metacritic (for films) and LibraryThing and Love Reading (for literature) that, much like Goodreads, allow viewers or readers to present their own reviews of fiction, be it literature or film. An intriguing aspect of many

of these sites is the propensity of reviewers to provide "plot summaries" as opposed to critical engagements with more sophisticated thematic analysis. While this plot-based approach to reviewing works of fiction may drive literary scholars to the brink of insanity, it does allow us to consider questions regarding the popular engagement with literature and other forms of artistic production. In future work, we expect to include actant-relationship sequencing so that we derive automatically a reader consensus model of plot, represented as a dynamic narrative framework graph. Given the responses that people post, we can use the scale of these sites to derive insight into how people (or groups of people) not only read but also remember. Turning the process around, it may be possible to develop a dynamically updated crowd-sourced summary of a novel or film–as more people write reviews, the consensus summary would update, capturing the emphasis on actants, relationships, and events that commentators add. Such a system could act as a cultural response barometer since what people remember, and what they forget (or choose to leave out), can be telling indicators of popular engagement with art.

REFERENCES

[1] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics, 1998.

[2] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X.

[4] Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, 2013.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[6] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[7] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[8] AJ Greimas. Les actants, les acteurs et les figures in sémiotique narrative et textuelle coll. *L. paris*, 1973.

[9] Nevena Lazic, Amarnag Subramanya, Michael Ringgaard, and Fernando Pereira. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515, 2015. doi: 10.1162/tacl_a_00154. URL https://doi.org/10.1162/tacl_a_00154.

[10] Harper Lee. To kill a mockingbird. philadelphia & new york, 1960.

[11] O-Joun Lee and Jason Jung. Explainable movie recommendation systems by using story-based similarity. 03 2018.

[12] Wendy G Lehnert. Narrative text summarization. In *AAAI*, pages 337–339, 1980.

[13] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[15] John W Mohr, Robin Wagner-Pacifici, Ronald L Breiger, and Petko Bogdanov. Graphing the grammar of motives in national security strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics*, 41(6): 670–700, 2013.

[16] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations.

Book Name	Entity Mention Groups
Of Mice and Men	**Lennie**:[Lennie, lenny], **George**:[george, milton], **Curley's Wife**:[curley's wife, tart, wife], **Aunt Clara**:[aunt clara, aunt, clara], **men**:[workers, men], **ranch**:[ranch, farm], **soft things**:[soft things, soft, things], **mental disability**:[mental disability, mental, disability]
The Hobbit	**Bilbo**:[bilbo, baggins, burglar, hobbit], **Rivendell**:[rivendell, middleearth], **Gandalf**:[gandalf, wizard, gandolf, grey], **dwarf**: [dwarf, dwarves], **Thorin**: [thorin, company], **trolls**:[trolls, orcs], **elf**:[elf, elves], **Hobbitown**:[hobbitown, shire, hobbiton], **man**: [human, man, lakemen], **dragon**:[dragon, smaug]
Frankenstein	**monster**:[monster, creature, adam], **Frankenstein**:[frankenstein, victor, doctor, creator], **Mary Shelley**: [mary, shelley, author, mary shelley], **Elizabeth**:[elizabeth, wife], **Walton**:[walton, robert], **Henry**:[henry, clerval], **Justine**:[justine, moritz], **Caroline**:[caroline, beaufort]
To Kill a Mockingbird	**Scout**:[scout, sister], **Atticus**:[atticus, dad, father, finch], **Jem**:[jem, brother], **Harper Lee**: [lee, harper lee, author, harper], **Tom**: [tom, robinson, negro, mockingbird, africanamerican], **Bob**:[bob, ewell], **Boo**: [boo, arthur, arthur radley, boo radley], **Mayella**: [mayella, daughter], **aunt**: [aunt, alexandra], **Maycomb**: [maycomb, alabama, town], **Heck**:[heck, tate], **Cunningham**:[cunningham, walter]

Table 4: Final actants after EMG per book. Each actant group is labeled with the most frequent mention in the group. Empirically, these automatically computed labels match the ground truth entities as derived from SparkNotes.

	Of Mice and Men	The Hobbit	Frankenstein	To Kill a Mockingbird
Recall (%)	**88.33** (83.33)	**82.61** (59.42)	**69.04** (66.66)	**90.16** (68.85)
Edge detection rate (%)	**98.33** (96.66)	**92.75** (69.56)	**73.80** (73.80)	**93.44** (77.04)
Average Number of Relationships	**246.55** (209.15)	**139.34** (14.03)	**20.33** (13.38)	**72.09** (27.34)
Median Number of Relationships	**54** (48)	**43** (3)	**7** (7)	**36** (6)

Table 5: Performance on character relationship extraction with IARC after (<u>in bold</u>) and before (within parentheses) EMG. In the "before", scenario an actant group consisted of only the mention used in the ground truth. Thus for actant "Bilbo" only the mention "Bilbo" was used to compute its relationship. Post EMG, the mentions in the group Bilbo:[bilbo, baggins,burglar,hobbit] were aggregated to compute the actant Bilbo's relationships.

In *Proc. of NAACL*, 2018.

[19] Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 995–1004, 2015.

[20] Mattia Samory and Tanushree Mitra. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[21] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 523–534. Association for Computational Linguistics, 2012.

[22] Mary Shelley. Frankenstein. london: Lackington, hughes, harding, mavor, and jones, 1818. ed. stuart curran. *Romantic Circles Electronic Editions*, 16, 2015.

[23] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2016.

[24] J Steinbeck. Of mice and men. new york: Covici & friede, 1937.

[25] Timothy R Tangherlini, Vwani Roychowdhury, Beth Glenn, Catherine M Crespi, Roja Bandari, Akshay Wadia, Misagh Falahi, Ehsan Ebrahimzadeh, and Roshan Bastani. âĂIJmommy blogsâĂİ and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR public health and surveillance*, 2(2):e166, 2016.

[26] Mike Thelwall and Karen Bourrier. The reading background of goodreads book club members: a female fiction canon? *Journal of Documentation*, 2019.

[27] Mike Thelwall and Kayvan Kousha. Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4):

972–983, 2017.

[28] John Ronald Reuel Tolkien. *The hobbit*. Houghton Mifflin Harcourt, 2012.

[29] Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

[30] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1248. URL https://doi.org/10.18653/v1/p19-1248.

[31] Fei Wu and Daniel S Weld. Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 118–127. Association for Computational Linguistics, 2010.