NASTEA: Investigating Narrative Schemas through Annotated Entities

Dan Simonson
Department of Linguistics
Georgetown University
Washington, DC 20057
des62@georgetown.edu

Anthony Davis
Ashland, OR 97520
tonydavis0@gmail.com

Abstract

In this paper, we investigate the distribution of narrative schemas (Chambers and Jurafsky, 2009) throughout different document categories and how the structure of narrative schemas is conditioned by document category, the converse of the relationship explored in Simonson and Davis (2015). We evaluate cross-category narrative differences by assessing the predictability of verbs in each category and the salience of arguments to events that narrative schemas highlight. For the former, we use the *narrative cloze* task employed in previous work on schemas. For the latter, we introduce a task that employs narrative schemas called *narrative argument salience through entities annotated*, or *NASTEA*. We compare the schemas induced from the entire corpus to those from the subcorpora for each topic using these two types of evaluation. Results of each evaluation vary by each topical subcorpus, in some cases showing improvement, but the NASTEA task additionally reveals that some the documents within some topics are significantly more rigid in their narrative structure, instantiating a limited number of schemas in a highly predictable fashion.

1 Introduction

A number of approaches for detecting narrative structures in text have been devised in recent years. Drawing from the work of Schank and Abelson (1977) and subsequent efforts to automatically populate templates with specific events and particpants referred to in text, Chambers and Jurafsky (2008; 2009) created the first statistically induced versions of such models. Chambers and Jurafsky (2008) described a basic version of their approach, in which single narrative chains involving a participant are generated; Chambers and Jurafsky (2009) builds on that work to create entire *narrative schemas*—generalized story lines that contain events and chains of potential role fillers that span across events. Other models have been devised for analyzing narrative (Vossen et al., 2015; Miller et al., 2015), but we will employ a variant of Chambers and Jurafsky (2009)'s narrative model in this work.

Chambers and Jurafsky (2008; 2009) introduced the narrative cloze task for evaluating their results. This involves removing a single word from a narrative chain in a held out document; the language model must then predict the missing word. The model is scored by how highly it ranks the true hidden word compared to all other possible replacements. A number of generative models have been introduced to further improve performance on cloze, and have done so successfully (Jans et al. 2012; Cheung et al., 2013; Chambers, 2013; Pichotta and Mooney 2014; Nguyen et al. 2015). More recently, it has been shown that a LSTM recurrent neural network can improve performance as well (Pichotta and Mooney 2015). These models focus on improving performance on the narrative cloze. Typically, they use the ordering of words in a chain as a factor, endowing them with the ability to anticipate the linguistic structure of the documents they model, but less able to produce schemas that represent conventionalized narrative structures. For instance, a model of news text that guesses the widespread, nonspecific verb "say" may perform well on narrative cloze,

Proceedings of 2nd Workshop on Computing News Storylines, pages 57–66,
Austin, TX, November 5, 2016.

but such a model is unlikely to reveal the world knowledge forming part of a conventionalized sequence of events.

Thus in some ways, while recent work has succeeded in raising the bar for solving the cloze task, it has sidestepped the original goal, which was to act as "a comparative measure to evaluate narrative knowledge" (Chambers, 2011, 26 – 27). Conservative guesses on narrative cloze alone create strong linguistic templates but poor narratological ones. Two issues raise concerns about the value of cloze as an evaluation of the narrative aspect of schemas. First, the focus on statistical associations between verbs misses a key component of narrative, namely, the connections between participants common to the events within a narrative, which establish it as a coherent narrative in the first place. Second, these measures of statistical association will make it clear which types of actions tend to be mentioned in concert within a document, but they may be less successful in detecting associations between participants in those events, for at least two reasons: there are many more participants (e.g., named individuals) referred to in a corpus than there are verbs, and there are various ways of referring to the same participant within the course of a narrative (e.g., different name strings, descriptions, titles, and pronouns).

Additionally, little work has been done exploring the properties of Chambers' narrative schemas. Simonson and Davis (2015) attempt to determine whether the events in narrative schemas can be used as especially sensitive features for a naïve Bayes classifier. They demonstrate that schema events alone do not seem to predict document category (e.g. $schemas \not\rightarrow category$) However, they do not demonstrate the converse, whether constraining document category can produce better schemas (e.g. $category \rightarrow schemas$?), which we will attempt to show here.

In this study, we intend to explore the properties of narrative schemas by investigating the influence of document category on schemas generated. Intuitively, detecting narrative sequences of events and their participants in text seems important, and the ability, e.g., to automatically generate as well as populate such schemas or templates is one clear application of this line of research. However, evaluation of schemas on this and similar tasks is not straightforward, as a gold standard is not clearly defined. We discuss and compare two techniques that are readily implemented and for which a gold standard is available: *narrative cloze* and *NASTEA*, an entity extraction task. NASTEA's reliance on schemas should add more transparency to the evaluation process, with schemas providing clear representations of patterns at the discourse level.

In Section (2), we will describe in detail the prior schema generation work we will modify for looking at topical conditioning. In Section (3), we describe our dataset. In Section (4), we describe our modifications of prior work for generating schemas. In Section (5), we describe in detail the NASTEA task we used to investigate schemas in this paper. In Section (6), we describe our results, followed by discussion (Section 7) and conclusions (Section 8).

2 Chambers and Jurafsky's Schema Model

In this paper, we work with Chambers and Jurafsky (2009)'s pmi-based narrative schemas, using a nearly identical score and generation procedure, though with a different data set and some extensions to explore the role of topic in a schema-learning procedure. These changes will be discussed in Section (4); here we will discuss the original model.

Fundamentally, Chambers and Jurafsky (2009) consider the problem of how well a new verb-dependency pair $\langle f, g\rangle$ fits into a chain of an existing schema, where f is some verb and g is a dependency. This relationship is defined in Equation (1) as $chainsim'$:

$$chainsim'(C, \langle f, g\rangle) = \max_a \left(score(C, a) + \sum_{i=1}^{n} sim(\langle e_i, d_i\rangle, \langle f, g\rangle, a) \right) \quad (1)$$

There are two main components of note here: $score(C, a)$, which assesses how well an argument type a fits in with chain C and $\sum_{i=1}^{n} sim(\langle e_i, d_i\rangle, \langle f, g\rangle, a)$, which determines how well the new pair $\langle f, g\rangle$ fits in with the rest of the existing chain, given argument type a.

$score$ is defined as:

$$score(C, a) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sim(\langle e_i, d_i \rangle, \langle e_j, d_j \rangle, a) \quad (2)$$

which checks, for every pair in C, the compatability of argument a. Both of these depend on sim, which is defined as:

$$sim(\langle e, d \rangle, \langle e', d' \rangle, a) = pmi(\langle e, d \rangle, \langle e', d' \rangle) + \lambda \log freq(\langle e, d \rangle, \langle e', d' \rangle, a) \quad (3)$$

sim establishes the relationship between two verb/dependency pairs $\langle e, d \rangle$ and $\langle e', d' \rangle$ on two different levels: the pmi establishes their general strength through coreference; if a verb/dependency pair shares a coreferring argument with another verb/dependency pair, this counts toward increasing the joint probability used in computing the pointwise mutual information between the two. $\lambda \log freq(\langle e, d \rangle, \langle e', d' \rangle, a)$ defines the strength of that connection with argument a in the mix, with the $freq$ being the counts of $\langle e, d \rangle$ and $\langle e', d' \rangle$ appearing together with a shared argument a.

3 Data

Our data comes from the New York Times corpus (Sandhaus 2008), a corpus containing 1.8 million articles from the New York Times from January 1987 to June 2007. Each article is annotated with metadata, including document categories—for our purposes, the `online producer` tag in the New York Times corpus—and salient entity annotations—people, organizations, and locations. Each article has been human-annotated with extensive metadata, including document categories—for our purposes, the `online producer` tag—and salient entity annotations—people, organizations, and locations.

To investigate our research question, we select a subset of document categories in the corpus that appear with a similar frequency and represent a broad range of topics (Table 1). The schemas used in this study are induced from this set of documents. In one procedure, the entire set of documents serves as the corpus for a single set of schemas. In a second, we create a topic-specific set of schemas, using the set of documents assigned to a given topic as the corpus for a set of schemas. One aim of this is to investigate the extent to which evaluation measures are affected by topic specificity. A second is to examine how the sets of topic-specific schemas might differ.

Table 1: Counts of document categories selected from the `online producer` tag for use in this study. Frequencies vary, but were chosen to be around the same order of magnitude and to represent different sorts of topics.

`online producer` category	counts
Law and Legislation	52110
Weddings and Engagements	51195
Crime and Criminals	50981
United States Armament and Defense	50642
Computers and the Internet	49413
Labor	46321
Top/News/Obituaries	36360

Once the documents of these categories were extracted, they were pre-processed using Stanford CoreNLP (Manning et al. 2014). Of particular importance are the Stanford Parser (de Marneffe et al. 2006) and `dcoref` (Lee et al. 2013), used for coreference resolution. These play a central role in the schema generation process described in the next section. Documents where parsing or coreference failed to complete were removed from processing as well.

4 Modifications to Schema Generation

We now briefly discuss our modifications to Chambers and Jurafsky (2009)'s schema generation technique, described in detail in Section (2). Our model varies fundamentally from Chambers and Jurafsky (2009)'s in that it is conditioned by document category, in this case selected from the `online producer` categories from the NYT corpus that we were interested in. Separate models are trained for each document category, only on documents contained in that category. The only exception to this is the baseline model, which is trained on all documents into one single model. We surmise that the resulting schemas should be "more topic-specific" than those generated by the baseline model, which lumps all topics together.

Conditioning schema generation by document category, as noted above, is one key difference. Ad-

ditionally, there are a few small changes at some of the post-score steps in the procedure. The score value from Chambers and Jurafsky (2009) does not explicitly describe how a newly added event's argument slots should be tied to the existing chains in the schema it is being added to. We handle this in a separate step—after it is decided that an event should be added to a schema, connections are made at that point where the threshold can be crossed. Also, we allow for an event to be added to multiple schemas if the score is high enough. In part, this is to allow for the words meaning to be captured across multiple contexts.

Lastly, we genericize some types—similar to Balsubramanian et al. (2013)—but not in all circumstances; instead, we do so only in the event that there is no common noun available to learn from. Our algorithm first checks the Stanford NER (Finkel et al. 2005) to see if there are any available types. Then it checks if there are any pronouns in the chain, and attempts to guess a type for the chain based on that. Finally, if there are no other types available, it aborts to a fallback type.

During the process of generation, a random selection of 10% of documents were held out for evaluation.

Figure (1) depicts a schema generated by our procedure.

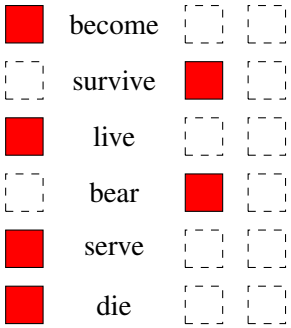

Figure 1: A relatively simple schema from the *Top/News/Obituaries* document category. The red squares indicate a chain that is strongly represented by the generic type PERSON, but with many other lionizing human types: scholar, hero, advocate, philosopher, etc. The dashed squares represent slots attested in the data but not connected during schema generation. In other words, this schema contains a single chain such that: PERSON/hero/advocate was born, lived, served, became, died, and was survived by...

5 Narrative Argument Salience Through Entities Annotated (NASTEA)

In this section, we will describe our technique for evaluating schemas using annotated entities.

Any evaluation applied to narrative schemas implicitly defines the notion of narrative. Since there are many aspects of the somewhat vague concept of narrative, and since, as noted above, there is no single obvious and clearly defined task and gold standard for evaluation of narrative schemas, a single type of evaluation is unlikely to gauge all of these aspects adequately. To address some of these shortcomings, we propose a task that is solvable, evaluates schemas directly, and concerns an aspect of narrative orthogonal to what the cloze task involves—the participants. Salient entity annotations in the New York Times corpus, performed by trained human indexers, appear well suited to this task. We investigate whether we can use narrative schemas to identify these salient entities, under the assumption that entities deemed important by the annotators indicate *Narrative Argument Salience Through Entities Annotated*, or *NASTEA*.

There are three steps of the NASTEA task that must be described in detail. First, in Section (5.1), we describe the notion of the *presence* of a schema in a document. Second, in Section (5.2), we describe how a present schema is used to extract salient entities from a text, and how those extractions are scored against the gold standard. Finally, in Section (5.3), we describe how this procedure is executed using an arbitrary number of schemas to produce curves indicating the performance of a group of schemas of the NASTEA task.

5.1 Identifying a Schema in a Document

Determining whether or not a word or n-gram appears in a document is a relatively simple task, but identifying whether a narrative schema is present or not is neither trivial nor categorical. In this study, we deploy a measure of *presence* that reflects the *canonicality* of a document—that is, how closely a document matches a schema. This measure uses the events of a schema as a proxy for its content—excluding the arguments from the measure. We explicitly exclude coreference information from the measure since coreference is error prone; while we

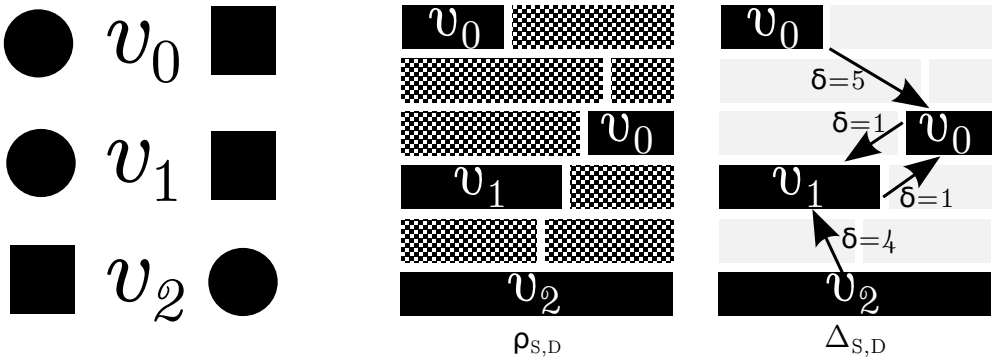

Figure 2: An illustration of how the presence $p_{S,D}$ (Formula 6) of a schema S (left) is measured with respect to a document D (both center and right). Each of the two illustrations of D shows how the document appears with respect to the respective measure: density $\rho_{S,D}$ (Formula 4) in the center and dispersion $\Delta_{S,D}$ (Formula 5) to the right. In this example, $\rho_{S,D}$ is $4/11$; $\Delta_{S,D}$ is $1/4 \times (5+1+1+4) = 11/4$; $p_{S,D}$ is $4^2/11^2$.

trust it en masse for generalizing over many documents, we are not so sure coreference can be trusted while considering one single document.

Measuring the presence $p_{S,D}$ of a schema S in a document D begins with $V_{S,D}$, the set of tokens from D that represent events in S. The same token type can appear multiple times in the set as long as multiple tokens of it appear throughout D. A sentence can have multiple verbs, and all relevant verbs are included in $V_{S,D}$.

There are two ways to consider the distribution of verbs within a document, both of which we want to contribute to defining presence: *density* and *dispersion*. Density ρ is defined as:

$$\rho_{S,D} = \frac{|V_{S,D}|}{|D|} \tag{4}$$

where $|D|$ is the number of sentences in the document, and $V_{S,D}$ is defined above. In other words, $\rho_{S,D}$ measures how much of the document D is composed of verbs $V_{S,D}$ representing the events in schema S. If this factor is high, then the document as a whole is very close to being only the series of events expressed in relevant schema. This is illustrated in the centered $\rho_{S,D}$ component of Figure (2)—the full black segments of ρ illustration represent members of $V_{S,D}$, the checker-patterned components represent sentences that do not contain any members of $V_{S,D}$.

While a high density value is a strong indicator of presence, some cases where the density is not as high may still be interesting. We hypothesize that verbs belonging to a schema appearing close together probably indicate an expression of that schema, while the same verbs more widely dispersed in the document are less likely to instantiate it. We therefore define the dispersion $\Delta_{S,D}$ with respect to a schema S and a document D as:

$$\Delta_{S,D} = \frac{1}{|V_{S,D}|} \sum_{v_i \in V_{S,D}} \min_{v_j \in V_{S,D} - \{v_i\}} \delta(v_i, v_j) \tag{5}$$

where $\delta(v_i, v_j)$ indicates the distance in sentences between two verbs v_i and v_j. The minimization seeks to find the nearest v_j to v_i in $V_{S,D}$, which is computed for every v_i contained in $V_{S,D}$. This is illustrated in Figure (2) as well, on the far right. Each arrow points from a specific v_i to the specific v_j where the distance is smallest.

The presence measure should be higher for those documents in which the elements of a schema are both dense (throughout the document) and not dispersed, so we define *canonical presence* p as:

$$p_{S,D} = \frac{\rho_{S,D}}{\Delta_{S,D}} \tag{6}$$

This defines the extent to which a schema is present in a document—more specifically, the degree to which a document itself comes close to being an exemplar of the schema. The components of p are illustrated in Figure (2).

5.2 Extracting Salient Entities with a Schema

Once schemas have been ranked for presence, they must be applied to a document in some way. We use the verb/dependency pairs found in that document that are also present in a schema to extract entities of importance. From each pair, any NP governed through the indicated dependency is extracted in whole. Only NPs containing proper nouns (`/NNP.*/`) are retained, as common nouns are not indicated in the NYT Metadata.

One side effect of Chambers' algorithm is a large number of schemas containing only a single verb—having only weak connections with the events in any other schema. We excluded these schemas from the NASTEA task.

The entities extracted are compared with the entities indicated in the NYT Metadata, a union of the `person`, `organization`, and `location` tags for each document. Each person, organization, or location from the metadata is tokenized with NLTK's (Bird et al. 2009) `wordpuncttokenizer` and is normalized for capitalization. Punctuation tokens are removed. Each entity extracted from the data is considered equal to the metadata entity if a fraction of the tokens r are equal between the two. This r value is set at 0.2, which is quite low, but justifiable, as any overlap between the open-class proper noun components likely indicates a match expressed differently from the normalized representation in the metadata: for example, an extraction of "Mr. Clinton" should match "William Jefferson Clinton" in the metadata. A higher threshold would have excluded these sorts of matches, which are typical of the writing style of the New York Times but differ in their metadata.

The fraction of entities from the metadata captured represents the *recall* while the fraction of things extracted actually found in the metadata indicates *precision*. NASTEA scores are reported as the F1 score of both of these values.

5.3 NASTEA Curves and Their Interpretation

As much as we would wish for it to be the case, the most present schema does not always yield the correct entities. In many cases, adding additional schemas of high presence is required. We use a set of schemas for each document, increasing this quantity by groups of five, starting at one. This allows us to see how well the first schema applied performed, followed by the the top 6, followed by the top 11, etc. If only the highest presence schema is applied, then that is expressed as "N_1;" for the top 6, that is reported as "N_6," etc. Nevertheless, N_1 results are of particular interest to us—this is the "I'm feeling lucky" narrative schema, the one with the highest presence with respect to a document. The N_1 performance should be highest in documents where canonicality most strongly applies.

We split the data by document category, then generated schemas for each category. In evaluation, only schemas generated with documents from a specific category were applied to that specific category. Analogously, this was done for the narrative cloze task, but instead of schemas, each model—learned from the documents in that one single category—was applied to predict events for that specific category. In both experiments, documents that were members of multiple categories, about 9% of the held-out 27498 documents, were removed from the hold-out data to remove any possible penalties due to categorical overlap.

6 Results

Table 2: Average rank of answers in the narrative cloze.

Test Model	Avg. Rank
Baseline	1329
Topical	1273
Top/News/Obituaries	565
Weddings and Engagements	1058
Law and Legislation	1279
Labor	1297
Crime and Criminals	1268
Computers and the Internet	1346
United States Armament and Defense	1805

Of the narrative schemas generated,[1] around 13% were shared between document categories on average. Each categorical set of schemas shares around 26% of its schemas with the baseline set.

[1]The schemas are available for download at `http://schemas.thedansimonson.com`.

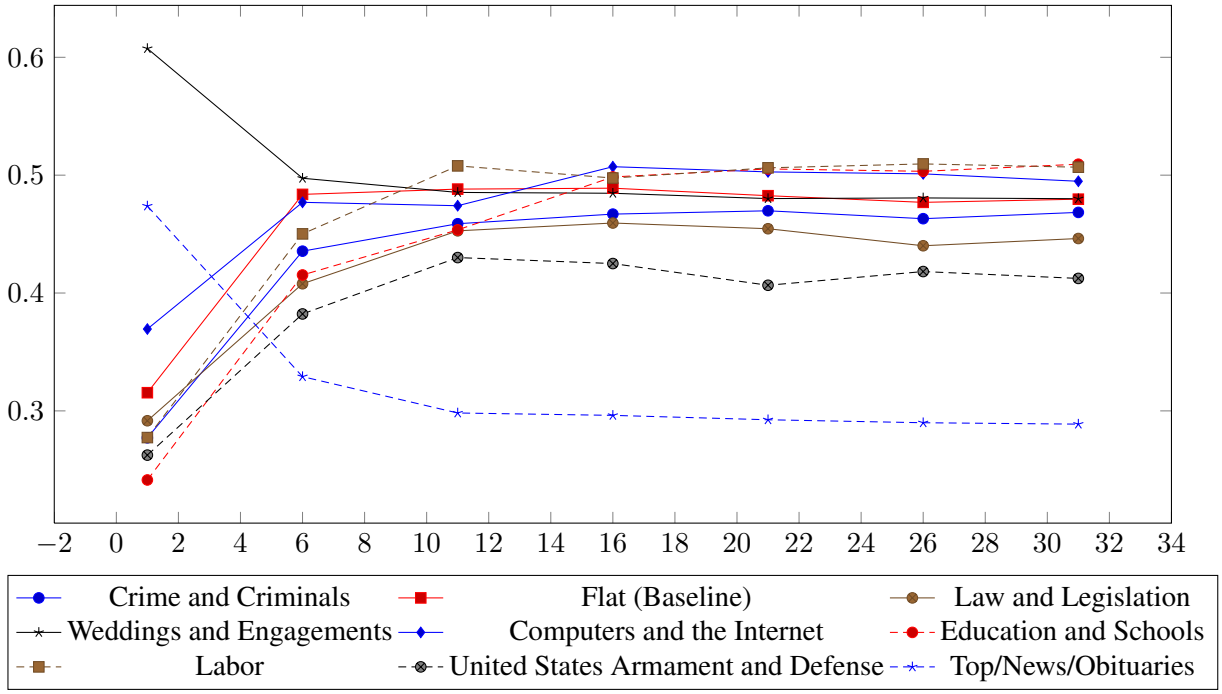

Figure 3: Plot of test-by-test performance on the NASTEA task for each topic. The x-axis indicates number of top-n present schemas applied. The y-axis indicates F1 score (i.e. N_n) on the number of entities retrieved by the set of top-n schemas.

Table (2) contains the cloze task results. Figure (3) illustrates results for the NASTEA task, broken down by document category. Most categories follow a general trend of performing poorly with the highest-presence guess alone. As more schemas are applied, the system is better able to retrieve annotated entities on most categories, with F1-scores leveling off around 45%. These values remain more or less stable *ad infinitum* with a few minor variations in value as n continues to increase. The "flat" baseline model follows this trend adequately as well.

However, two categories are exceptions to this trend: *Weddings and Engagements* and *Top/News/Obituaries*. Their N_1 performances are significantly[2] higher than their counterparts' scores, and their curves are concave up. This difference is supported by the results of the cloze task as well.[3]

This exceptional N_1 performance invites closer inspection, which can be seen in Figure (4). Since NASTEA is applying schemas to documents, those schemas can be retained and counted allowing for illustration of the variety of different schemas that seem to best fit a particular document, what we will refer to as *narrative homogeneity*. Figure (4) takes the N_1 results and illustrates the totals of counts for schemas that were applied in each N_1 case. Categories that performed well on N_1 were also more homogeneous at N_1, choosing a single schema as most present more often than their more heterogeneous counterparts.

7 Discussion

The NASTEA task shows a clear, discrete distinction between two types of document categories: those that seem to be narratologically homogeneous and others that seem to be narratologically heterogeneous within the scope of this model of narrative. In the homogeneous case, the assertion that $category \rightarrow schema$ seems to be valid, while in more heterogeneous circumstances, this is much less the case. This affirms Miller et. al. (2015)'s observation that their own corpus is characterized by a "heterogeneity of the articles' foci," with their corpus likely fitting into the *United States Armament and Defense* category—a notably heteroge-

[2] $p < 0.001$ including the baseline with the heterogeneous categories; $p < 0.005$ excluding the baseline from the analysis.

[3] $p < 0.05$

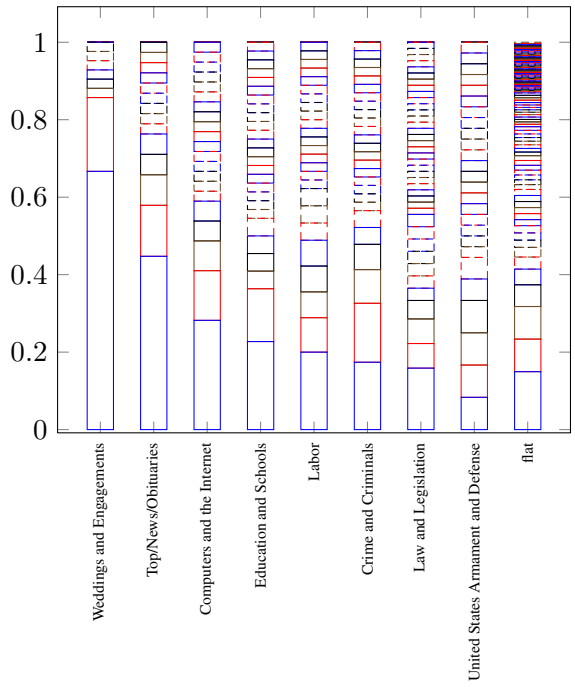

Figure 4: Plot of N_1 Document Categorical Narrative Homogeneity: A representation of the fractional distribution of schemas with the highest presence across all documents in a category ($n = 1$ for the NASTEA task). The y-axis Each slice of the whole indicates the fraction of a single schema having the highest presence for a document. A larger slice indicates that the single schema it represents had the highest presence for more documents in that topic than a smaller slice.

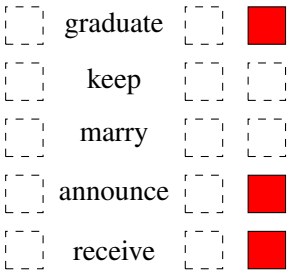

Figure 5: Schema generated in the *Weddings* document category. The dashed squares represent slots attested in the data but not connected during schema generation. The chain of red squares indicates a generic organization type. The other slots remain largely unlinked because they are frequently found as conjunctive arguments of reciprocal verbs the are not handled well by the existing narrative models.

neous one—were it derived from the NYT Corpus.

Of those we used in this study, the *Weddings and Engagements* and *Top/News/Obituaries* (referred to hereafter as *Weddings* and *Obituaries*, respectively) are distinctly homogeneous. This distinction is reaffirmed through the cloze task as well, where each of their respective rank averages are hundreds of ranks higher. This indicates that they are more rigid in their choice of wording and the events they describe, and those events point more strictly toward the entities the NYT library scientists annotated. It is not too surprising that these particular categories are different. Impressionistically, the writing styles of such documents are more rigid than their more news-typical counterparts. However, the objective measurability of this impression via two distinct forms of evaluation is a first.

There are two possible interpretations of this result. One is that the homogeneous categories are truly something different from the heterogeneous ones, and that this is a fact about news narratives and document categories at large. This is very much plausible, as *Weddings* and *Obituaries* are categories defined by the events contained within them: marriage and death, and the events that lead up to those. Events in *United States Armament and Defense* can vary dramatically: from roadside bombings to budget overruns. The other interpretation is that the homogeneous categories are ones that are better encapsulated by our model of narratives and that the heterogeneous ones are not captured properly. This makes the NASTEA task something to optimize performance on, making it a quantitative metric for evaluating improvements in narrative schemas. These are not necessarily contradictory interpretations if one accepts both of them as independently representing different aspects of the notion of narrative.

While cloze and NASTEA overall agreed on the exceptionality of *Weddings* and *Obituaries*, there remain some discrepancies between the two. *Obituaries* performs much better on cloze relative to *Weddings*, while on NASTEA, the reverse happens, and *Weddings* outperforms *Obituaries*. Within the rest of the categories, rankings shuffle around between the two. For example, *Computers and the Internet* performed well below average on cloze, but ranked third highest on N_1, with the homogeneity to match.

Narrative cloze's opacity makes these discrepancies difficult to understand without trolling through thousands of rankings. NASTEA has the transparency to show what is going on under the hood: clear differences in narrative homogeneity.

8 Conclusion

We have shown that constraining document category can influence a model's performance on the cloze task. NASTEA, the new technique we have introduced to evaluate the properties of narrative schemas, paints a more complex picture: that some document categories—*Weddings* and *Obituaries*—are more homogeneous in the narratives they express than other sorts of categories. In other words, at the narratological level, not all categories are the same—some are measurably different from others. In the process, we have also defined the first ever measure for the presence of a schema in a document, opening up the possibility for techniques that use schemas to perform quantitative analysis of documents at the narratalogical level.

Acknowledgments

We would like to thank Amir Zeldes, Nate Chambers, and the reviewers of this paper at multiple stages, whose feedback helped refine this work into its current form, as well as the Georgetown University Department of Linguistics and Graduate School for their continued support.

References

Balasubramanian, N., Soderland, S., Mausam, & Etzioni, O. 2013. Generating Coherent Event Schemas at Scale. In *EMNLP* (pp. 1721-1731).

Bird, S., Loper, E., and Klein E. 2009. Natural Language Processing with Python. OReilly Media Inc.

Chambers, N., & Jurafsky, D. 2008. Unsupervised Learning of Narrative Event Chains. In *ACL* (pp. 789-797).

Chambers, N., & Jurafsky, D. 2009. Unsupervised learning of narrative schemas and their participants. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 602-610). Association for Computational Linguistics. Chicago

Chambers, N. W. 2011. Inducing Event Schemas and their Participants from Unlabeled Text. *Stanford University*.

Chambers, N. 2013. Event Schema Induction with a Probabilistic Entity-Driven Model. In *EMNLP* (pp. 1797-1807).

Cheung, J. C. K., Poon, H., & Vanderwende, L. 2013. Probabilistic frame induction. In *NAACL-HLT 2013* Association for Computational Linguistics.

de Marneffe, M., MacCartney, B., and Manning, C.D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

Finkel, J.R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *ACL 2005* (pp. 363-370).

Jans, B., Bethard, S., Vuli, I., & Moens, M. F. 2012. Skip n-grams and ranking functions for predicting script events. In *EACL* (pp. 336-344). Association for Computational Linguistics.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics* 39(4).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL* (System Demonstrations, pp. 55-60).

Miller, B., Olive, J., Gopavaram, S., & Shrestha, A. 2015. Cross-Document Non-Fiction Narrative Alignment. In *Proceedings of the First Workshop on Computing News Storylines* (pp. 56-61). Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Nguyen, K.H., Tannier, X., Ferret, O., & Besancon, R. 2015. Generative Event Schema Induction with Entity Disambiguation. In *ACL* (pp. 188 - 197)

Pichotta, K., & Mooney, R. J. 2014. Statistical Script Learning with Multi-Argument Events. In *EACL* (pp. 220-229).

Pichotta, K., & Mooney, R. J. 2015. Learning statistical scripts with LSTM recurrent neural networks. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Sandhaus, E. 2008. *The New York Times Annotated Corpus*. Linguistic Data Consortium, Philadelphia.

Schank, R.C. & Abelson, R.P. 1977. Scripts, plans, goals and understanding. Lawrence Erlbaum.

Simonson, D. & Davis, A. 2015. Interactions between Narrative Schemas and Document Categories. In *Proceedings of the First Workshop on Computing News Storylines* (pp. 1-10). Association for Computational

Linguistics and The Asian Federation of Natural Language Processing.

Vossen, P., Caselli, T., & Kontzopoulou, Y. 2015. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines* (pp. 40-49). Association for Computational Linguistics and The Asian Federation of Natural Language Processing.