10

Narrative Homogeneity and Heterogeneity in Document Categories

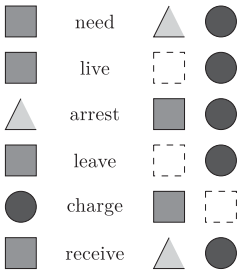
Dan Simonson and Anthony R. Davis

Abstract. In this chapter, we present techniques for examining the distributional properties of narrative schemas (Chambers and Jurafsky, 2009) in the news, particularly in a subset of the New York Times (NYT) Corpus (Sandhaus, 2008), to see how well they capture the events and stories presented there. In one technique, the narrative argument salience through entities annotated (NASTEA) task, we use the event participants indicated by narrative schemas to replicate salient entity annotations from the NYT Corpus. In another technique, we measure narrative schema stability by generating schemas with various permutations of input documents. Both of these techniques show differences between homogeneous and heterogeneous document categories, where homogeneous categories being those written from templates such as Weddings and Obituaries. Homogeneous categories tend to perform better on the NASTEA task using fewer schemas and exhibit more stability, whereas heterogeneous categories require more schemas applied on average to peak in performance at the NASTEA task and exhibit less stability. This suggests that narrative schemas succeed at detecting and modeling the repetitive nature of template-written text, whereas more sophisticated models are required to understand and interpret the complex novelty found in heterogeneous categories.

10.1 Introduction: Narrative Schemas and Their Evaluations

Two core components of the storyline of a narrative are the events of the story and the participants in those events. One technique that captures these two aspects of the storylines of a corpus is *narrative schemas* (Mooney and DeJong, 1985; Chambers and Jurafsky, 2009; Balasubramanian et al., 2013), generalizations over narratives that reflect common patterns of events and their participants. Narrative schemas complement other approaches to

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press



...A few minutes later, Mr. Solorzano said, he left the dealership to see his friend being questioned... Chief Russo said, "The injuries he received were from the street..." Mr. Solorzano was charged with assaulting a police officer...
He said he was still missing $160 that was in his pocket when he was arrested.
(NYT Corpus 831741 - 2/22/1996)

Figure 10.1 An example of a narrative schema with an associated text. The schema is represented by rows representing each event and its associated slots. Each column of symbols represents a particular slot, either SUBJ, OBJ, or PREP, which participated in that event. Each symbol represents a chain of mentions of a particular entity in different slots around those events. For example, the square here represents a person, student, or "self" type. Dashed squares in the schema indicate singleton chains not linked to any other slot in the schema. In the prose, underlines indicate events that occur in the schema; rectangles indicate the chain of mentions that correlated with the square in the schema.

the automated analysis of topical and narrative information in documents. Unlike template-filling techniques, they do not require a defined set of human-crafted templates; instead, template-like structures are induced. Unlike topic models, they generate representations in which event types and participant types are organized into relational structures, specifying shared participants between events. Unlike automatic summarization, they generalize over similar but distinct narratives to reveal their underlying common elements.

Figure 10.1 features an illustration of a narrative schema generated in this study, with an example of a document that contains events described by the schema. The schema correctly predicts that, in the given text, a particular individual should leave somewhere, be arrested, be charged with something, and receive something. However, it makes some inaccurate predictions as well – for example, that whatever arrested the individual should also be received by the individual. Because a schema is generated from many thousands of documents, it makes generalizations that are not guaranteed to be represented in every text.

Fundamentally, this chapter reviews efforts to evaluate narrative schemas. Determining whether a narrative schema is "correct" is not a well-defined task; yet, we do not want to abandon evaluation entirely. Rather, we chose two tasks to assess properties that reflect intuitions of what a good schema might be. The first is the *narrative argument salience through entities annotated* or *NASTEA* task (Section 10.4), where entities are retrived using narrative

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

schemas. The intuition behind this is that a "good" narrative schema should include elements that involve prominent participants in narrative. The second is a *stability procedure* (Section 10.5), which measures the stability of a set of schemas by ablating and cross-validating a set of documents to see how consistent the schemas themselves are. A "good" set of narrative schemas should be resilient to small perturbations in the source corpora. These properties of narrative schemas demonstrate the existence of two types of document categories: *homogeneous* and *heterogeneous*. *Homogeneous* categories are categories of documents with a consistent set of storylines with relatively fixed events and participant slots, albeit with new participant identities. They are often written from templates, such as *Obituaries* and *Weddings and Engagements*, whereas *heterogeneous* categories often describe new combinations of events or circumstances, or what is typically thought of as "news." The evidence for this distinction seems to be robust across both measurements of properties. Given their difference from one another in terms of what they measure and how they measure it, the two provide convergent evidence for such a distinction between document categories.

10.2 Background

Narrative schemas originate as an interpretation of Schank and Abelson's (1977) "scripts" – a conception of cognition and episodic memory where abstractions of repeated sequences of events are learned as abstractions of the events themselves. Attempts were made to learn these scripts during the prestatistical era of natural language processing (NLP), such as Mooney and DeJong's (1985) work generating schemas from individual documents.

Chambers and Jurafsky (2008, 2009) reintroduced the idea of scripts to the NLP literature. They used advances in parsing and coreference to aggregate statistics on the relationships between event verbs through shared, coreferring arguments, selected through their relationships to the verb, as either a `SUBJ`, `OBJ`, or `PREP`. Through these relations, they count pairs of event-dependency pairs. Though space does not allow a full overview, the formula for sim expresses a core intuition of their technique. $\langle e,d\rangle$ and $\langle e',d'\rangle$ are tuples of events and dependencies, respectively, and a is an argument type based on the most frequent noun phase type in each coreference chain:

$$sim(\langle e,d\rangle,\langle e',d'\rangle,a) = pmi(\langle e,d\rangle,\langle e',d'\rangle) + \lambda \log freq(\langle e,d\rangle,\langle e',d'\rangle,a) \tag{10.1}$$

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

pmi is the pointwise mutual information,[1] $freq$ is the frequency that both tuples and the argument type a appeared together, and λ is a weighting parameter, which balances the influence between the simple "generic" coreferent sharing expressed in the pmi with the more precise yet sparse typed coreferent sharing contained in the log $freq$ term.

Balasubramanian et al. (2013) followed up Chambers and Jurafsky (2009) with additional architectural improvements to schema generation. Additionally, they conducted the first manual evaluation of schemas, both of their own and of Chambers and Jurafsky's (2009) schemas, showing broadly that, to some extent under human evaluation, a portion of unsupervised schemas reflect some sort of reality for layman annotators.

For Chambers and Jurafsky's (2008) evaluation, they introduced the *cloze task* as a metric for understanding performance of their system. However, the cloze task does not measure schemas directly. A substantial body of work has been produced to further performance on the cloze task. Much of this work optimizes performance solely on the cloze task and does not generate schemas (Pichotta and Mooney, 2014, 2016). Some work critical of the cloze task has either presented new versions of the cloze task more focused on narrative (Mostafazadeh et al., 2016) or more fundamental tasks (Bisk et al., 2019) or looked at the problem of script induction as one of language modeling (Rudinger, Demberg et al., 2015; Rudinger, Rastogi et al., 2015).

10.3 Data and Schema Generation

The data for both of these experiments comes from the New York Times (NYT) Corpus (Sandhaus, 2008), a corpus containing 1.8 million articles from the (NYT) from January 1987 to June 2007. Each document in the corpus itself is tagged with document categories and entity annotations. The document categories were selected to represent a broad range of topics with similar frequencies (Table 10.1). The schemas used in this study are induced from this set of documents – minus a holdout set of 10% of the documents – and the document categories used in this study refer to this set.

Once the documents of these categories were extracted, they were preprocessed using Stanford CoreNLP (de Marneffe et al., 2006; Lee et al., 2013; Manning et al., 2014).[2] Dependency parsing and coreference resolution are

[1] For our purposes, $pmi(a,b) = p(a,b)/(p(a) \times p(b))$.
[2] Version 3.4.1.

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

Table 10.1. *Counts of document categories selected from the* `online producer` *tag for use in this study*

`online producer` category	Counts
Law and Legislation	52,110
Weddings and Engagements	51,195
Crime and Criminals	50,981
Education and Schools	50,818
United States Armament and Defense	50,642
Computers and the Internet	49,413
Labor	46,321
Obituaries	36,360

Note: Frequencies vary but were chosen to be around the same order of magnitude and to represent different sorts of topics.

effectively the first step of schema generation. Documents where parsing or coreference failed to complete[3] were removed from processing as well.

For schema generation, we use Chambers and Jurafsky's (2009) original generation technique with some modifications. The model employed here is conditioned by document category; separate sets of schemas are trained on each document category instead of all documents. Furthermore, though Chambers and Jurafsky's (2009) schema germination technique has no intrinsic limit, we cut off generation for each category at 800 schemas. This was the limit it was practical to evaluate within the two proposed frameworks.

Additionally, 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 the various slots from an event newly added to a schema should be connected into forming chains within that schema. This occurs in a separate step – after it is decided that an event should be added to a schema, each individual slot from the candidate event is scored against the existing chains in the schema. The highest scoring chain for each slot has the slot added as a link in that chain; if the score is not high enough, the slot starts a new singleton chain in the schema. Also, an event may be added to multiple schemas if the score is high enough.

Lastly, we genericize some types – similar to Balasubramanian 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, first checking the Stanford

[3] 14,239 documents, or 0.7% of the 1.8 million total documents.

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

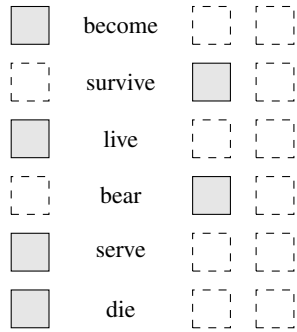


Figure 10.2 A relatively simple schema from the *Obituaries* document category. The squares indicate a chain that is strongly represented by the generic type PERSON but with many other lionizing human types: scholar, hero, advocate, philosopher, etc. The dashed squares represent slots attested in the data but not connected during schema generation.

Named Entity Recognition (NER) (Finkel et al., 2005) output for alternatives and, then guessing a type based on pronouns in a given chain. Finally, if nothing else is found, it aborts to a fallback type. Figures 10.1 and 10.2 depict schemas generated by this procedure.

10.4 Evidence through NASTEA Task

Ideally, a model of narrative should be able to extract the same entities that humans think are important with respect to a certain story. Because the NYT Corpus has entity annotations, the NASTEA task attempts to do exactly that: use narrative schemas to extract a set of annotated salient entities from a document. We use as a gold standard those contained in the NYT Corpus. Ideally, if a set of narrative schemas is a sufficient model of narrative, the participants a schema captures should match those marked as salient by NYT library scientists.

It is worth noting that our objective is not to achieve state-of-the-art performance on entity extraction. Rather, our goal is to use entity extraction as a proxy for schema quality. NASTEA provides a local measure of schema success, seeing whether in instances of particular documents it can successfully identify salient entities.

The most complicated component of this is the identifcation of the presence of a schema in a document, which is not trivial to determine. We explain our technique for doing so in Section 10.4.1, followed with a few of the particulars

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

about how NASTEA was done here (Section 10.4.2), and ending this section with the results of the task (Section 10.4.3).

10.4.1 The Presence of 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. The NASTEA task relies on some sort of notion of *presence* to determine what schemas should be applied to which documents.

In the following sections, 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 because coreference is error prone; though we trust it *en masse* for generalizing over many documents, we do not trust it on a document-to-document basis.

Measuring the presence $p_{S,D}$ of a schema S in a document D begins with $V_{S,D}$, the set of verbs in D that represent events in S:

$$V_{S,D} = \{v_i : v_i \in D \wedge v_i \in S_e\}, \tag{10.2}$$

where $v_i \in D$ is true when an instance of verb v_i is inside document D. S_e is the set of events in a schema, each represented by a verb. The same verb type can appear multiple times in the set, because each instance is uniquely indexed. As with the schemas, the set of verbs does not include nominalizations. 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 contribute to defining presence: *density* and *dispersion*, illustrated in Figure 10.3. Density ρ is defined as

$$\rho_{S,D} = \frac{|V_{S,D}|}{|D|}. \tag{10.3}$$

$\rho_{S,D}$ measures the fraction of sentences in document $|D|$ that contain 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 the schema S.

Though a high density value is a strong indicator of presence, some cases where the density is not as high may still be interesting. If a set of relevant verbs is close together, this indicates some expression of the schema, whereas

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

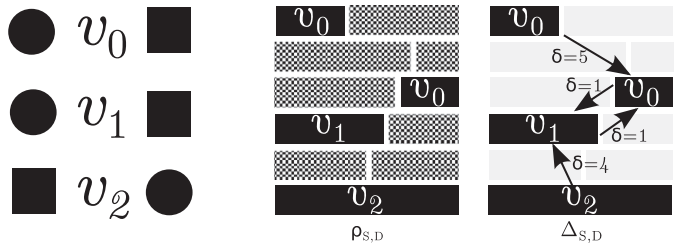


Figure 10.3 An illustration of how a document looks through the two components of schema presence. In other words, it is how the document D looks through density $\rho_{S,D}$ and dispersion $\Delta_{S,D}$ for a hypothetical schema S. In D, a rectangular block represents each sentence. v_i in that each rectangular block indicates an instance of the verb corresponding to the v_i event in S.

a disperse set of verbs is less likely to be an expression of the events listed in the schema. This we call Δ, defined as

$$\Delta_{S,D} = \frac{1}{|V_{S,D}|} \sum_{v_i \in V_{S,D}} \min_{v_j \in V_{S,D} - \{v_i\}} \delta(v_i, v_j), \tag{10.4}$$

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 $V_{S,D}$.

The presence measure should be higher for those documents in which the elements of a schema are both dense (throughout the document) and not disperse. We define *canonical presence* p as

$$p_{S,D} = \frac{\rho_{S,D}}{\Delta_{S,D}}. \tag{10.5}$$

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.

10.4.2 Evaluating Schemas at the Document Level with NASTEA

Once schemas have been ranked for presence, the best match must be applied to the matching 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

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

importance. From each pair, any NP governed through the indicated dependency is extracted in whole. Only NPs containing proper nouns (`/NNP.*/`) are retained, because common nouns are not indicated in the NYT metadata. Additionally, we exclude any schemas containing only one event from the NASTEA task.

The entities extracted are compared with the entities indicated in the NYT metadata. Each person, organization, or location from the metadata is tokenized with NLTK (Bird et al., 2009) and 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, because 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 NYT but differ in their metadata. A manual inspection of this low r value showed a meta-accuracy of around 98% (Simonson, 2017, p. 112).

The fraction of entities from the metadata captured represents the *recall* and the fraction of entities extracted that are actually found in the metadata indicates *precision*. NASTEA scores are reported as the F1 score of both of these values. In evaluation, only schemas generated with documents from a specific category were applied to that category. Documents that were members of multiple categories (about 9% of the held-out documents) were removed from the hold-out data to remove any possible penalties due to categorical overlap.

10.4.3 Results

Figure 10.4 illustrates results for the NASTEA task. 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 40%. These values remain more or less stable *ad infinitum* with a few minor variations in value as n continues to increase.

However, two categories are exceptions to this trend: *Weddings and Engagements* and *Obituaries*. These two categories, instead of producing concave down curves, produce curves that are concave up, indicating peak perfor-

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

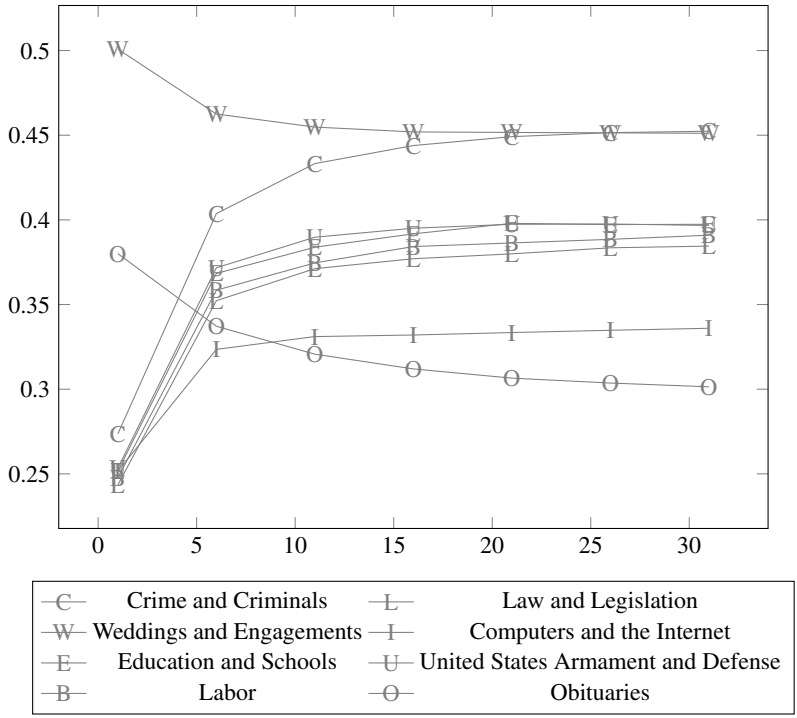


Figure 10.4 Plot of test-by-test performance on the NASTEA task for each topic. The x-axis indicates number of top-n present schemas applied. The y-axis indicates F1 score (i.e., N_n).

mance when only one schema is applied (at N_1) and decreasing performance as more schemas are applied.

This exceptional N_1 performance necessitates closer inspection. Because 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 10.5 takes a subset of 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 homogenous at N_1, choosing a single schema as most present more often than their more heterogeneous counterparts.

In the next section, this distinction arises from a very different sort of experiment, one that does not use annotated entities at all.

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

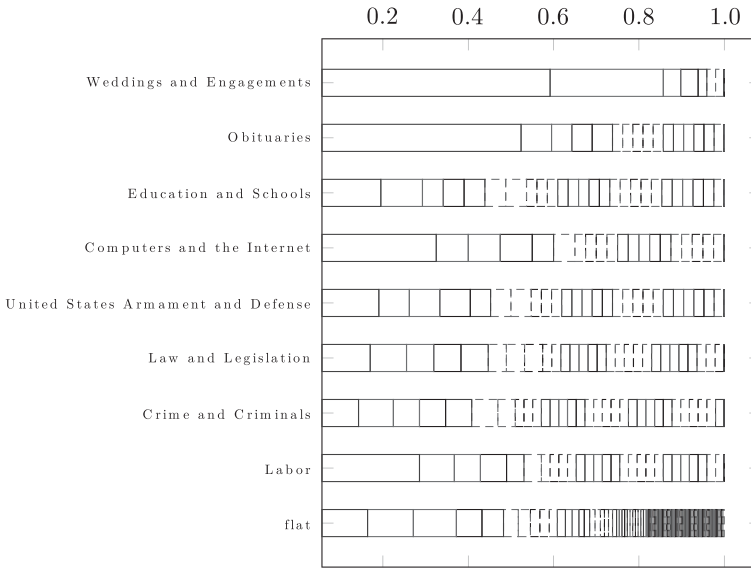


Figure 10.5 Plot of N_1 document categorical narrative homogeneity. A representation of the variety of schemas with the highest presence in documents in each category ($n = 1$ for the NASTEA task) in a subset of 324 of the holdout documents; "flat" represents a set of schemas generated without categorical distinctions and applied to all documents in the corpus. Fewer slices represent a smaller fraction of schemas being most present. A larger slice indicates that the single schema it represents had the highest presence for more documents.

10.5 Evidence through Schema Stability

The NASTEA task provides one angle to examine narrative schemas, through the correlation between what human annotators thought were central to a narrative and what the extracted schemas presented as central. This has limitations. It requires human annotations to evaluate schemas, as well as recorrelating output of the system with documents to examine them.

An alternative is to examine schemas against one another. Ideally, a set of schemas should be consistent, even given perturbations of the input data; in other words, a few missing documents should not significantly alter the resulting schemas. These result in a more global measure than NASTEA provided: the inputs as a whole are modified, and the outputs as a whole are scored collectively for their intrinsic stability.

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

Though there are things to be learned from different schema germination techniques, we will not be examining those differences here.[4] Instead, we will be focused on how the stability further reifies the homogeneous/heterogeneous distinction exhibited by the NASTEA task.

The stability evaluation procedure alternates two stages: an ablation step and a cross-validation step. At each ablation step, 10% of the starting set of documents is removed – not 10% of the previous ablation – on a category-by-category basis. Then 10-fold cross-validation partitions the set of documents; the 9/10ths of documents in each fold are used to generate 10 sets of schemas for each category at that ablation. These splits are not preserved across ablations. Though these procedures involve removing portions of the original corpus, the most intuitive way to interpret the intent is in reverse – that is, to think of some sort of search and retrieval procedure yielding slightly different results (cross-validation step) at each step in a larger data collection effort (ablation step). This results in 100 sets of schemas generated for each document category.

10.5.1 Fuzzy Jaccard Coefficient and the Jaccard Reciprocal Fraction

The described stability ablation procedure still needs a technique for comparing the hundreds of thousands of schemas across sets of them. Evaluating the similarity between two sets of schemas is not so straightforward, particularly when a measure that awards partial credit for partial matches would return the most intuitive results. Essentially, we would like to determine, for each schema in one set, how similar its best match is in the other.

To give an intuitive but also set-theoretically informed measure of the similarity between sets of schemas, we report values for the schema stability in terms of the *Jaccard Reciprocal Fraction* or JRF:

$$JRF(S,T) = \frac{4}{J_{J_e}^{-1}(S,T) + 3}, \tag{10.6}$$

where S and T are sets of schemas, and $J_{J_e}^{-1}$ is the reciprocal of the *Fuzzy Jaccard measure* (J_{J_e}):

[4] For a comparison between different germinator types, see Simonson and Davis (2018).

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

$$J_{J_e}(S,T) = \frac{|S \cap_{J_e} T|}{|S| + |T| - |S \cap_{J_e} T|}, \tag{10.7}$$

where S and T are sets of schemas and $|S \cap_{J_e} T|$ is a fuzzy measure of the cardinality of the intersection of two sets, where

$$|S \cap_{J_e} T| = \sum_{\tau \in T} \max_{\sigma \in S} J_e(\sigma, \tau), \tag{10.8}$$

where σ and τ are sets of events contained in a single schema in S and T and $J_e(\sigma, \tau)$ is the Jaccard coefficient between the two sets. The full derivation for these is detailed in Simonson and Davis (2018). Most important, the JRF gives approximately the typical fraction of shared events between schemas in two sets of schemas, regardless of the size of schemas in each set. As the fuzzy Jaccard value approaches 1, so does the JRF; as the fuzzy Jaccard value approaches 0, the denominator approaches infinity, and thus the JRF approaches 0.

10.5.2 Results

For each individual pair of sets of schemas within an ablation, we compute fuzzy Jaccard scores, their means, and their standard deviations, transformed into JRF form.[5] Average values are shown in Figure 10.6.

Note that increasing ablation number refers to a decreasing number of documents; in other words, ablation 8 refers to 8/10ths of the documents having been *removed*. In total, the experiments generated 2,698,865 schemas, cut down to 640,000: 800 per category, across 8 categories, 10 cross-validations, and 10 ablations. These are not unique because the goal was to generate schemas as similar to one another as possible.

In Figure 10.6, the document categories found to be homogeneous – Weddings and Obituaries – are notably more stable than the categories shown to rely on fewer schemas to identify participants in Section 10.4. The difference is larger for Weddings and Engagements than for Obituaries; the gap between Obituaries and the other categories is small at ablation 0 but increases as fewer documents are used.

[5] The full table of values is available at https://schemas.thedansimonson.com/

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

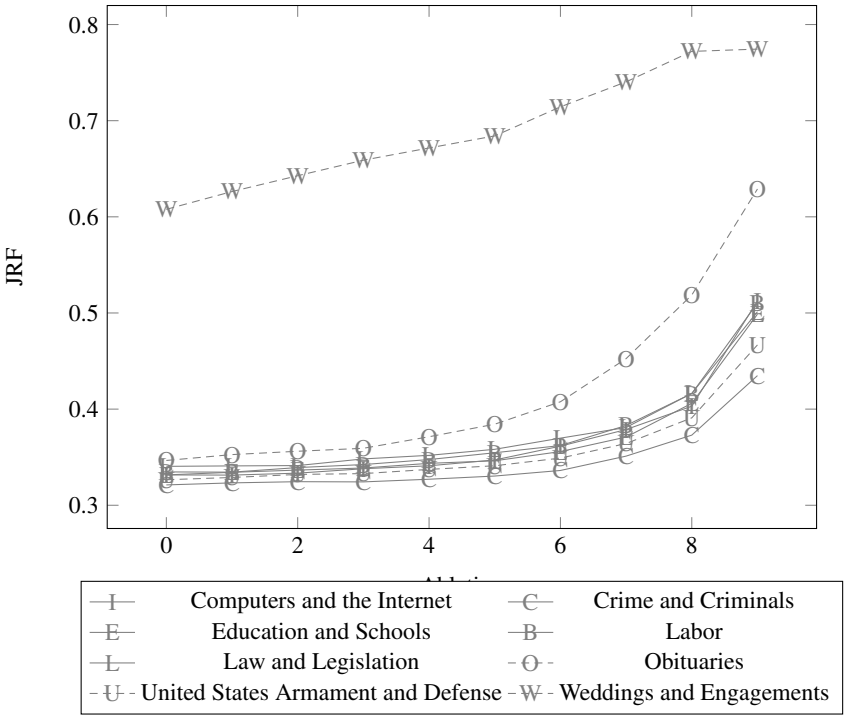


Figure 10.6 Stability in each ablation in each document category. Ablation is on the x-axis; Jaccard Reciprocal Fraction (e.g., events typically shared) is on the y-axis.

10.6 Discussion

A homogeneous category is one with a consistent set of storylines – the identities of the participants may change but the events and the roles stay the same. We can see clear evidence for homogeneity in the *Weddings and Engagements* and *Obituaries* categories of the (NYT), distinct from the other, more heterogeneous categories examined. The NASTEA task shows that, for the homogeneous categories, the participants in those narratives can be identified with a handful of schemas; for heterogeneous narratives, it requires far more schemas to identify the participants. The stability procedure shows that the schemas derived from a homogeneous document category remain more consistent when derived from a different subset of documents. Both of these results are clear evidence of consistent storylines – of homogeneity – among the Weddings and Obituaries categories.

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

The strength of this result is reinforced by the very different nature of the tasks used here, examining properties of narrative schemas from very different angles. Whereas the NASTEA task looks at participants in a narrative, the stability procedure compares events across schemas. The NASTEA task uses human annotations to accomplish its objective, whereas the stability procedure requires none. The NASTEA task is very localized, gathering participants from a specific narrative to score; the stability procedure is global, examining properties of schemas as a whole set of sets. Nevertheless, in both cases, we see evidence of the same phenomena: heterogeneous and homogeneous document categories. The stability procedure does this most intuitively. Schemas derived from a homogenous document category should be more stable under perturbations of its data, and this is what is seen quite consistently across the board. The NASTEA task requires more interpretation but also shows a clear distinction between the two types of categories. In the categories that are more homogeneous, a single schema does most of the work of the entity retrieval, as shown in Figure 10.5.

The homogeneous document categories are written from templates, so in some respects it is not surprising that a template extractor should exhibit different properties on the categories written from templates. However, a quantitative procedure for identifying this is valuable, especially when trying to leverage or understand properties of new kinds of data, such as the variety of genres found in the GUM corpus (Zeldes, 2017). The NASTEA task showed that this can be ascertained from broadly labeled data about participants in a narrative; the stability procedure used here shows the same distinction without any labeled data.

10.7 Conclusions

We showed two experimental results that both confirm that some document categories are homogeneous, whereas others are heterogeneous, based on how well narrative schemas can be used to extract entities in the NASTEA task and how consistent the schemas are that they produce under perturbations of the data. Both of these tasks shared a corpus and a schema generation technique in common but showed consistently the separation between the types of document categories despite coming at the problem from very different angles.

This distinction has implications not just for analyzing storylines but for problems beyond them as well. For example, when working on the problem of event extraction, the variety of events extracted is contingent on the type of data under analysis. A homogeneous category will have a predictable and tightly

https://doi.org/10.1017/9781108854221.013 Published online by Cambridge University Press

constrained set of events, whereas the events extracted from a heterogeneous document category are far more variable. Similarly, if a system is reasoning about storylines, such as Qin et al.'s (2019) work on counterfactuals, homogenous categories' storylines should have a more constrained search space than hetergeneous categories.

Toward a broader understanding of storylines and narrative, Caselli and Vossen (2016) critiqued narrative schemas as a model of narrative for the lack of causality in them. Causality is a core part of narrative understanding, yet schemas do not point toward causality in any particular direction between events. When we set out on this work, we hoped that narrative schemas could be used to analyze narrative structure more broadly, possibly employed as a sort of "tokens of narrative" model. Though schemas might provide a primitive example of a structured model of narrative knowledge, they remain incomplete. We further augment Caselli and Vossen's (2016) critique: many types of narrative are flexible, but discrete schemas are rigid. That said, the existence of some stable schemas, even in heterogeneous categories, may help highlight the components of text that exist in more calcified forms. Further, in certain genres, the determination of homogeneity maybe prove helpful in their analysis. The results here indicate that a complete model of storylines will require both.

References

Balasubramanian, Niranjan, Solderland, Stephen, Mausam, and Etzioni, Oren. 2013. Generating Coherent Event Schemas at Scale. Pages 1721–1731 of: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Bird, S., Loper, E., and Klein, E. 2009. *Natural Language Processing with Python*. Cambridge: O'Reilly Media.

Bisk, Yonatan, Buys, Jan, Pichotta, Karl, and Choi, Yejin. 2019. Benchmarking Hierarchical Script Knowledge. Pages 4077–4085 of: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics.

Caselli, Tommaso, and Vossen, Piek. 2016. The Storyline Annotation and Representation Scheme (StaR): A Proposal. Pages 67–72 of: *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*. Austin, TX: Association for Computational Linguistics.

Chambers, Nathanael, and Jurafsky, Dan. 2008. Unsupervised Learning of Narrative Event Chains. Pages 789–797 of: *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Chambers, Nathanael, and Jurafsky, Dan. 2009. Unsupervised Learning of Narrative Schemas and Their Participants. Pages 602–610 of: *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*. Association for Computational Linguistics.

de Marneffe, M., MacCartney, B., and Manning, C. D. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In: *Proceedings of the Language Resources and Evaluation Conference of the European Language Resources Association (LREC)*. Association for Computational Linguistics.

Finkel, J. R., Grenager, T., and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Pages 363–370 of: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Lee, H., Chang, A., Peirsman, Y., et al. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-Ranked Rules. *Computational Linguistics*, **39**(4), 885–916.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, et al. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. Pages 55–60 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mooney, Raymond, and DeJong, Gerald. 1985. Learning Schemata for Natural Language Processing. Pages 681–687 of: *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI)*.

Mostafazadeh, Nasrin, Grealish, Alyson, Chambers, Nathanael, Allen, James, and Vanderwende, Lucy. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. Pages 51–61 of: *Proceedings of the Fourth Workshop on Events*. San Diego: Association for Computational Linguistics.

Pichotta, Karl, and Mooney, Raymond. 2014. Statistical Script Learning with Multi-argument Events. Pages 220–229 of: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Pichotta, Karl, and Mooney, Raymond J. 2016. Using Sentence-Level LSTM Language Models for Script Inference. Pages 279–289 of: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1. Long Papers*. Berlin: Association for Computational Linguistics.

Qin, Lianhui, Bosselut, Antoine, Holtzman, Ari, et al. 2019. Counterfactual Story Reasoning and Generation. Pages 5043–5053 of: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics.

Rudinger, Rachel, Demberg, Vera, Modi, Ashutosh, Van Durme, Benjamin, and Pinkal, Manfred. 2015. Learning to Predict Script Events from Domain-Specific Text. Page 205 of: *Lexical and Computational Semantics (* SEM 2015)*.

Rudinger, Rachel, Rastogi, Pushpendre, Ferraro, Francis, and Van Durme, Benjamin. 2015. Script Induction as Language Modeling. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-15)*.

Sandhaus, Evan. 2008. *The New York Times Annotated Corpus LDC2008T19*.
Schank, Roger, and Abelson, Robert. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures.* NJ: London: Lawrence Erlbaum.
Simonson, Dan. 2017. *Investigations of the Properties of Narrative Schemas*. Ph.D. Thesis, Georgetown University.
Simonson, Dan, and Davis, Anthony. 2018. Narrative Schema Stability in News Text. In: *Proceedings of COLING 2018*. Santa Fe, NM: Association for Computational Linguistics.
Zeldes, Amir. 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, **51**(3), 581–612.