Narrative Schema Stability in News Text

Dan Simonson
Georgetown University
Washington, DC
des62@georgetown.edu

Anthony Davis
Ashland, Oregon
tonydavis0@gmail.com

Abstract

We investigate the stability of *narrative schemas* (Chambers & Jurafsky, 2009) automatically induced from a news corpus, representing recurring narratives in a corpus. If such techniques produce meaningful results, we should expect that small changes to the corpus will result in only small changes to the induced schemas. We describe experiments involving successive ablation of a corpus and cross-validation at each stage of ablation, on schemas generated by three different techniques over a general news corpus and topically-specific subcorpora. We also develop a method for evaluating the similarity between sets of narrative schemas, and thus the stability of the schema induction algorithms. This stability analysis affirms the heterogeneous/homogeneous document category hypothesis first presented in Simonson & Davis (2016), whose technique is problematically limited. Additionally, increased ablation leads to increasing stability, so the smaller the remaining corpus, the more stable schema generation appears to be. We surmise that as a corpus grows larger, novel and more varied narratives continue to appear and stability declines, though at some point this decline levels off as new additions to the corpus consist essentially of "more of the same."

1 Introduction

Narrative schemas complement other approaches to 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, with the goal of revealing their underlying common elements.

Scripts (or schemas or frames) have long been touted as a way to provide artificial intelligence systems with world knowledge and understanding (Schank & Abelson, 1977). Building these scripts by hand is labor intensive, so automatically learning script knowledge has attracted attention for decades (Balasubramanian et al., 2013; Chambers & Jurafsky, 2009; Mooney & DeJong, 1985; Pichotta & Mooney, 2014, 2015). However, we are also interested in what such techniques reveal about broad, quantitative properties of discourse in general.

Like many unsupervised tasks, the evaluation of schemas is still a matter of debate and depends on their intended use; there is no one broadly accepted method, and nothing that closely models human intuitions about narrative. Chambers & Jurafsky (2008, 2009) proposed and used the *cloze task* to evaluate their procedure, widely adopted in subsequent work (Cheng & Erk, 2018; Jans et al., 2012; Pichotta & Mooney, 2014, 2015). In the cloze task, an event from a sequence in holdout data—either a coreference or sentence-to-sentence sequence—is removed and must be guessed by a model of the events in text.

However, it is questionable whether the cloze task actually requires script knowledge to perform well (Mostafazadeh et al., 2016; Rudinger et al., 2015; Simonson, 2018), and thus whether it is an *effective* measure of schema quality. Cloze does not evaluate schemas directly—only indirectly through the score

used to generate schemas—and is fundamentally unsolvable, even by humans, so it is an open question whether the information needed to perform well on a cloze task corresponds closely to the information within narrative schemas. Others have evaluated schemas directly. Balasubramanian et al. (2013) evaluated schemas manually using mechanical turkers, but this is labor intensive and pre-supposed specific properties about schemas—for example, that "a child exploded a blast" cannot be part of a valid schema, despite that we live in a world where such incidents occur on a regular basis. Simonson & Davis (2016) introduced the NASTEA task for schema evaluation where schemas are used to identify salient entities in a document, but this is dependent on a set of "salient entity annotations" from the New York Times corpus (Sandhaus, 2008). We noted that some document categories are *homogeneous*—that is, requiring only one schema to obtain optimal performance on their task—and others are *heterogeneous*—requiring multiple schemas for optimal performance. It is not clear whether their findings are robust or a mere artifact of the idiosyncrasies of NYT annotations.

In this paper, we turn instead to the stability of narrative schemas, as a method of gauging schema quality complementary to cloze and the NASTEA task. We assume that high-quality schemas are robust; that is, the addition, deletion, or modification of a few narratives in the corpus should not drastically affect the schemas generated. We conducted experiments using both corpus ablation and cross-validation at each stage of ablation. This kind of evaluation is relatively easy to carry out in many sorts of corpora; it does not require any annotation effort and is conceptually straightforward. Two challenges, however, are the computationally intensive nature of these experiments, and the similarity measure employed in comparing sets of narrative schemas.

In Section (2), we describe the data set used in this paper. In Section (3), we describe the schema germinators used in this study, including one novel technique intended to complement the others from prior work. In Section (4), we explain the ablation and cross-validation procedure used to perform a schema stability analysis. in Section (5), we propose the "Fuzzy Jaccard" coefficient used for comparing sets of schemas. In Section (6), we present the results of this study. In Section (7), we discuss the results. In Section (8), we conclude this paper.

2 Data

We performed our stability analysis of narrative schemas on a subset of the *New York Times* Annotated Corpus (Sandhaus, 2008), the same subset as Simonson & Davis (2016). containing over 1.8 million articles from 1987 to 2007.

We selected the same document set from Simonson & Davis (2016) for direct comparison with our previous work, to affirm or disprove our homogeneity-heterogeneity hypothesis. Note that in these earlier results, included an "Education and Schools" category is mentioned (Simonson & Davis (2016), Figure 3), but is unfortunately omitted in Table (1) in that paper. This has been included in our Table (1) here.

Table 1: Counts of document categories selected from the `online_producer` tag for use in this study after pre-processing. Simonson & Davis (2016) chose categories to contain roughly the same number of articles and to represent different sorts of topics.

`online producer` category	counts	`online producer` category	counts
Law and Legislation	52110	United States Armament and Defense	50642
Weddings and Engagements	51195	Computers and the Internet	49413
Crime and Criminals	50981	Labor	46321
Education and Schools	50818	Top/News/Obituaries	36360

For NLP preprocessing, the Stanford CoreNLP suite of tools (Manning et al., 2014),[1] was chosen for comparability to Simonson & Davis (2016)'s original work, itself chosen because it has both parsing and coreference capabilities (de Marneffe et al., 2006; Lee et al., 2013), which are essential for generating schemas. Using the same version of tools as Simonson & Davis (2016), we were able to replicate their

[1] Stanford CoreNLP, Version 3.4.1 (2014-08-27), (via Simonson (2018))

category counts after excluding a number of articles that produced no coreference chains or failed to survive pre-processing.

3 Germinators for Schema Generation

The heart of Chambers & Jurafsky (2009)'s schema induction technique, and our variants derived from it, is a similarity measure between *narrative chains* in the context of a schema and a candidate event verb. A narrative chain is a set of verb-dependency pairs—i.e. slots—that share a common participant filling all the slots. These are constructed from verbs that are statistically associated using a PMI-based measure plus a preference for verb-argument slot pairs in which the argument slot is filled by participants of a consistent type. This similarity measure determines the best candidate verb-argument slot pair to add to an existing chain; a score termed $chainsim'$ is computed for each candidate pair, summing this similarity measure for each candidate and the existing elements of the chain. Lastly, narrative schemas, consisting of merged sets of chains sharing events, are produced.

Here, we focus primarily in generating schemas. The cloze task does not directly evaluate schemas. Rather, it evaluates a component of the model used to generate schemas. For example, in the case of Chambers & Jurafsky (2008, 2009), $chainsim$ and $chainsim'$ are used to rank event verbs in the cloze task. The schemas therein generated are not directly evaluated. Thus, we draw a line after $chainsim'$ and other techniques used on the cloze task, referring to them as the *candidate score*.

However, a candidate score alone is not enough to generate schemas. We must decide on a technique for traversing candidates and interpreting their scores as prospective additions to a schemas under construction. We will refer to these techniques collectively as *germinators*.

We employ two previously devised germinators for schema generation (Section 3.1): counter-training (Simonson & Davis, 2015) and Chambers' original technique or schema germination (Chambers & Jurafsky, 2009), though here referred to as "linear induction." Both of these techniques are relatively deterministic, so for comparison, we present a novel stochastic technique for schema generation called the "random walker" germinator (Section 3.2). We expect this stochastic technique to be less stable than its deterministic counter-parts.

All of the implementations for narrative schema germinators can be found in `durruti`.[2]

3.1 Prior Techniques

Linear induction is what we call Chambers & Jurafsky (2009)'s technique for inducing schemas. In this technique, the event verbs are considered in order for adding to narrative schemas. Each event verb is compared against an ever-growing set of schemas. New schemas are created when the event verb does not pass a threshold parameter β. In Chambers & Jurafsky (2009)'s original implementation, if a candidate event verb's $chainsim'$ score is greater than β, it is added to the schema that best fit the event verb. We differ in this regard, adding the candidate event verb to every schema for which its $chainsim'$ score crossed the threshold.

Alternatively, *counter-training* (Simonson & Davis, 2015)—inspired by Yangarber (2003)—considers a fixed set of schemas simultaneously, which start as a set of seeds. Candidate events are scored against all schemas, then those scores are penalized based on how many different schemas each candidate fit into. After penalties have been applied, the best candidate event for each schema is added to each schema.

With respect to both techniques, once the decision has been made to add an event to a schema, we insert new events into each schema the same way in both germinators, following Chambers & Jurafsky (2009)'s technique for doing so. Additionally, for both prior germinators and the random walker—described in Section 3.2—$chainsim'$ is used as the underlying score.

3.2 Random Walker

Both prior techniques are deterministic. Adding a "random walker" stochastic germinator potentially provides a nice contrast to these deterministic techniques. For a particular schema, all possible insertions are treated as weighted random choices based on the scores between chains already in a schema and

[2] `https://github.com/thedansimonson/durruti`

candidate events. However, because the scores are pmi-based, the values, if turned into a probability distribution, are nearly uniformly distributed. This causes events selected for schemas to be effectively unrelated to one another. Instead, we use a weighting function to weight the score more appropriately for random selection, in this case $weight(C, vd) = 2^{chainsim'(C,vd)}$, which effectively undoes the $\log$ contained in the score function.

This algorithm departs from a simple random walk, because the weights on the graph change at each step; the current score value for each schema depends not only on the last event added, but on all of the events previously added. The "graph" is thus recomputed based on the current state of the schema being grown.

4 Stability Ablation and Cross-Validation

The stability evaluation procedure alternates two stages: an ablation step and a cross-validation step. At each ablation step, 10% of the total set of documents are removed (not 10% of the previous ablation). Then, using the corpus at each ablation stage, ten-fold cross-validation partitions the set of documents and 9/10ths of the available documents are used in each fold to generate schemas. These splits are not preserved across ablations. While 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).

We generate schemas using each of the three schema germinators described above. Separate sets of schemas are generated from the documents in each category in Table (1), using separate PMI models produced from documents in each category, at each stage of ablation, and for each fold of cross-validation. These sets of schemas are what we compare to one another to gauge their stability in different experimental configurations.

Since the total number of schemas generated is about 4 million, we did not attempt to optimize various parameter values pertinent to each germinator, instead using a single set of approximately optimized values from an earlier study of schemas in a salient named entity detection task.

5 Fuzzy Jaccard Coefficient and the Jaccard Reciprocal Fraction

The result from each run (using a given schema germinator, on documents from a given category at a specified stage of ablation and fold of cross-validation) is a set of narrative schemas. How can we then measure the similarity between different sets of schemas? A schema in one set may be highly similar but not identical to a schema in the other set; see, for example, any two schemas in Figure (3). We wish to take their similarity into account in our overall measure of the similarity between the two sets. It is simple to measure the similiarity J_e between two individual schemas σ and τ, using the Jaccard coefficient, where σ_e and τ_e are the sets of events contained in each schema:

$$J_e(\sigma, \tau) = \frac{|\sigma_e \cap \tau_e|}{|\sigma_e \cup \tau_e|} \tag{1}$$

But evaluating the similarity between two sets of schemas is not so straightforward, particularly when we need to compare thousands of pairs of sets of schemas. Essentially, we would like to determine, for each schema in one set, how similar its best match is in the other. We therefore define a fuzzy Jaccard measure over two sets (of schemas, in this case), by redefining the intersection cardinality in a "fuzzy" way, as:

$$J_{J_e}(S, T) = \frac{|S \cap_{J_e} T|}{|S \cup_{J_e} T|} \tag{2}$$

where S and T are sets and J_e is a symmetric and well-defined comparison between elements of S and T, with values between 0 and 1. We invoke the identity ($|S \cup T| = |S| + |T| - |S \cap T|$) to derive a fuzzy counpart for the cardinality of the union. The fuzzy similarity measure then becomes:

$$J_{J_e}(S, T) = \frac{|S \cap_{J_e} T|}{|S| + |T| - |S \cap_{J_e} T|} \tag{3}$$

This allows us to have to only define $|S \cap_{J_e} T|$, which we define as:

$$|S \cap_{J_e} T| = \sum_{\tau \in T} \max_{\sigma \in S} J_e(\sigma, \tau) \tag{4}$$

The values rendered by this approximation are somewhat misleading, however. It tends to skew low, especially when considering schemas. Two schemas with six events each, sharing five events, have a Jaccard score of only 71%, since the schemas are "punished" for having two events that do not match. The Fuzzy Jaccard measure J_{J_e} further exaggerates this discrepancy. Assume that all of the schemas in two sets of schemas have a J_e value as described above, at 71%. In other words, these are fundamentally two very similar sets of schemas—each schema in one set has a counterpart in the other set sharing five out of six events. Assume also that both sets of schemas are the same size. This gives us:

$$J_{J_e}(S, T) = \frac{|T| \times 0.71}{2|T| - |T| \times 0.71} = 55\% \tag{5}$$

This value misleadingly implies that the sets of schemas are only 55% similar despite each schema in one set having a close match in the other.

If we make a few assumptions about J_{J_e}, we can find a better interpretation for the values it computes. Essentially, what we want to know is the typical value of $|\sigma_e \cap \tau_e|$ implied by a given J_{J_e} value. We define $|\sigma_e \cap \tau_e| = x$ since that is what we want to solve for. If there is such a value, we will assume it is fixed under the maximization in the defined cardinality of the intersection. This also allows us to reduce the summation over T to the cardinality of T. We also assume that $|\sigma_e| = |\tau_e| = \sigma'$ since our schema germinator ceases at a maximum of six events for all germinators:

$$|S \cap_{J_e} T| = \sum_{\tau \in T} \max_{\sigma \in S} \frac{|\sigma_e \cap \tau_e|}{|\sigma_e| + |\tau_e| - |\sigma_e \cap \tau_e|} = |T| \frac{x}{2\sigma' - x} \tag{6}$$

Let us allow for one more approximation, that $|S| = |T|$. This is absolutely true for the counter-training, random walker, and linear induction truncated germinators, since they generate a fixed number of schemas. It is approximately true for the linear induction germinator. Substituting Formula (6) into Formula (3) and making the given approximation yields:

$$J_{J_e}(S, T) = \frac{|S \cap_{J_e} T|}{|S| + |T| - |S \cap_{J_e} T|} = \frac{|T| \frac{x}{2\sigma' - x}}{2|T| - |T| \frac{x}{2\sigma' - x}} = \frac{\frac{x}{2\sigma' - x}}{2 - \frac{x}{2\sigma' - x}} = \frac{x}{2(2\sigma' - x) - x} = \frac{x}{4\sigma' - 3x} \tag{7}$$

Solving for x, we get the Jaccard Reciprocal Fraction, or JRF:

$$x = \frac{4}{J_{J_e}^{-1}(S, T) + 3} = JRF \tag{8}$$

This gives us 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. For the example shown above, $JRF = 4/(1/0.55 + 3) = 0.83 = 5/6$. This is more intuitive while simultaneously preserving the underlying set theoretic machinery and justification for the comparisons performed.[3]

It has been suggested that we include a more sophisticated lexical similarity metric—for example, those presented by Corley & Mihalcea (2005) or Kusner et al. (2015). However, fundamentally, these would obscure the answers to the questions we are asking here, which is, through the lens of coreference chains, dependencies, event verbs, and their pointwise mutual information, how stable are the schemas produced to perturbations in the data used? Were the models used to generate the schemas more dependent on a lexical similarity measure, such would be essential in answering that question—and while a

[3]Preliminary results indicate that the JRF values are quite similar to the raw numeric values computed.

great potential next step in improving schema induction, the problem of lexical similarity has largely been neglected in narrative schemas and the narrative cloze task. We suspect adherence to editorial guidelines within a single publication will likely normalize the most interchangeable terms—for example, selectively choosing "pursue" over "chase." In a multi-newspaper corpus, we suspect lexical distance would be much more important.

6 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. Presenting these values here is too cumbersome, however. As an overview, average JRF values across germinators and document categories are shown in Figures (1) and (2).

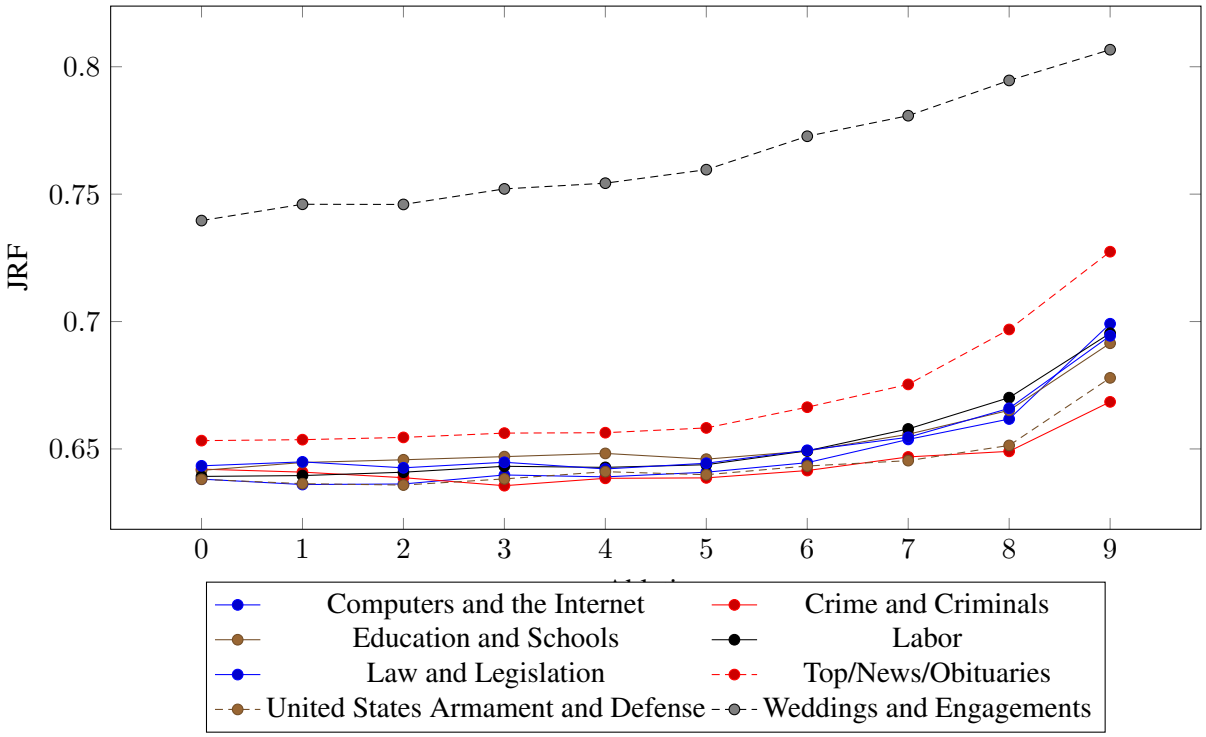

Figure 1: Stability averaged across document categories. Ablation is on the x-axis; Jaccard Reciprocal Fraction (e.g. events typically shared) is on the y-axis.

Note that increasing ablation number refers to a decreasing number of documents; in other words, ablation 8 refers to $8/10^{ths}$ of the documents having been *removed*.

In total, the series of experiments generated a total of 3,978,865 schemas. These are not "unique," as the whole point was to generate schemas that are hopefully as similar to one another as possible. Linear induction produced 2,698,865 schemas, in part a product of its open-ended generation process. Both counter-training and random walker generated 640,000 schemas, as the number of schemas generated within each category was fixed at 800 for practical computational reasons. [4]

Figures (1) and (2) contain the stability values as averaged across different dimensions. Figure (1) averages all stability values for a given ablation across all algorithms, leaving separate values for each category. Conversely, Figure (2) averages the stability values across document categories, leaving each algorithm individually expressed.

In both figures, stability generally increased as the number of total documents decreased. The one exception to this was the linear induction schemas, as shown in Figure (2). The causes and consequences

[4]Since the linear induction truncated algorithm is simply linear induction but with a few schemas clipped off the end, it is not counted as a separate instance of generation.

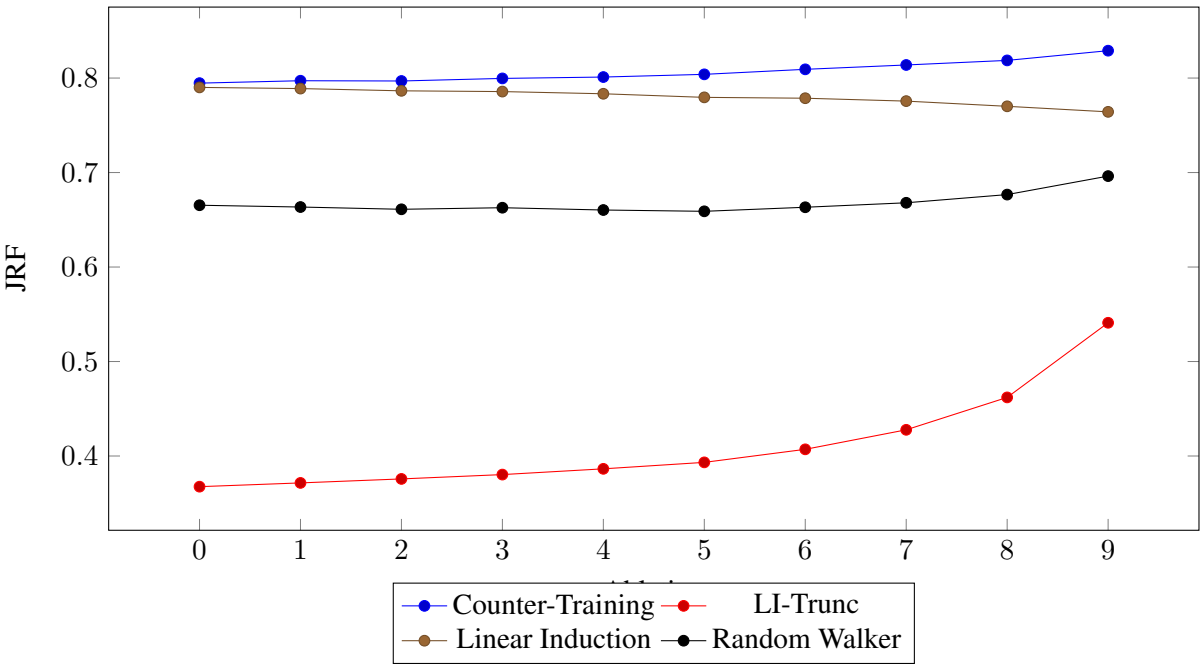

Figure 2: Performance averaged across algorithms. Ablation is on the x-axis; Jaccard Reciprocal Fraction (e.g. events typically shared) is on the y-axis.

of this will be discussed in the next section.

In Figure (1), the document categories Simonson & Davis (2016) found to be homogeneous are notably more stable than the categories we found to be heterogeneous. The difference between similarity scores in each cross-validation was significant ($p < 0.001$) in 140/140 comparisons (t-tests) between the similarity scores of "Weddings and Engagements" and the other document categories for both counter-training and the random walker germinators; the difference was significant ($p < 0.001$) in 137/140 comparisons between "Top/News/Obituaries" and the other document categories for both as well. Two of the insignificant differences occurred during the 9^{th} ablation against "Computers and the Internet" and "Labor," one occurred during the 2^{nd} ablation against "Education and Schools." This does not itself have to do with the content of the schemas, but the differences in similarities internal to the document category itself.

Some examples of schemas generated in this process—six from the counter-training germinator in two different cross-validations of the 0^{th} ablation—are contained in Figure (3). Schemas containing these sorts of events are typical of the Obituaries section. Note that "bear" is simply the lemmatized form of "be born". In each schema, participants of the same chain of argument slots of the verbs are indicated by the same color and shape, but this does not extend across schemas. Each column of shapes is, from left to right, in each schema, `SUBJ`, `OBJ`, and `PREP`, respectively. Dotted boxes indicate slots attested in the data but not linked to a chain; completely blank slots were never attested in the data. Each chain shares a number of types, too many to explicitly enumerate here. The object slot of "survive," for example from the idiom "person was survived by," was typically a generic person type or type that would qualify as a subset of person: "woman," "man," "boy," "microbiologist," "lawyer," etc.

7 Discussion

The first striking aspect of these results is the affirmation that the Weddings and the Obituaries categories, the two determined to be *homogeneous* by Simonson & Davis (2016), were consistently more stable at all ablations than the rest of the document categories. The differences were significantly different the vast majority of the time (277/280), and in cases where this was not the case, only a handful of the stability scores of the heterogeneous categories had increased. This affirms our earlier findings from another angle

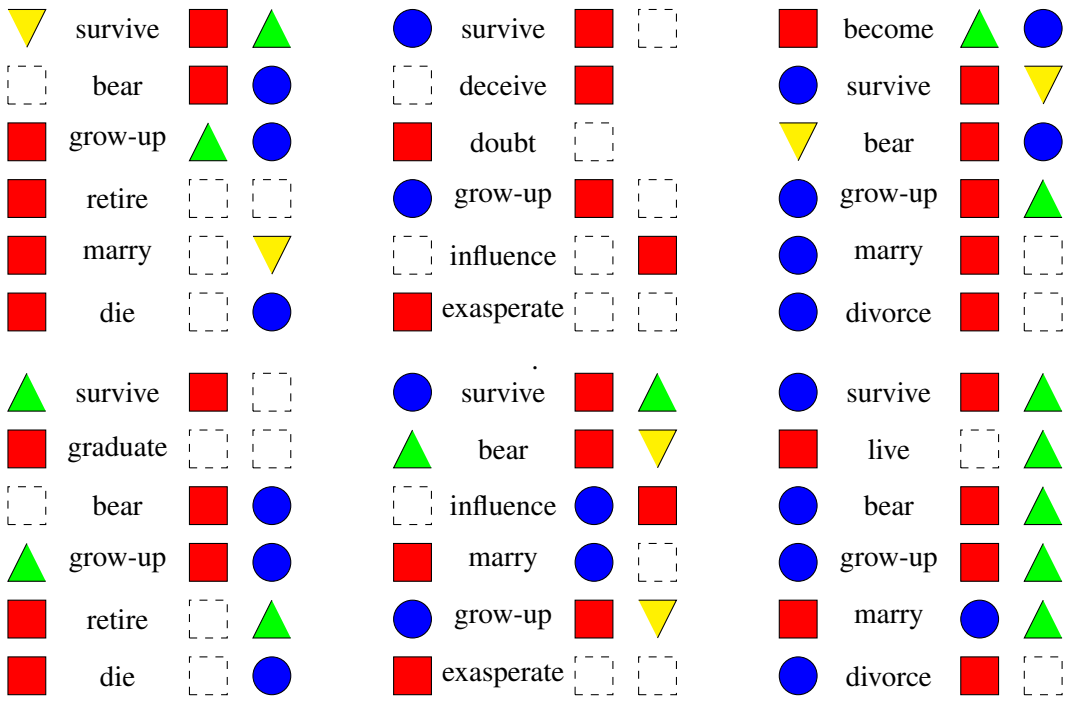

Figure 3: Six schemas generated by the counter-training germinator in two cross-validations, showing varying degrees of similarity between schemas. The two rows show schemas generated from two different cross-validation of the data in the same ablation.

unrelated to the NYT corpus' salient entity annotations, as we managed to show a similar distinction between homogeneous and heterogeneous document categories without salient entity annotations. The document categories are written from templates, so in some respects, it's not surprising that a template extractor should exhibit different properties on the categories written from templates. Simonson & Davis (2016) showed that this can be ascertained from labelled evaluation data; the stability procedure used here shows the same distinction, significantly without labelled data.

Unexpectedly, however, stability increased as the number of documents decreased.

Our initial expectations were that stability would increase with the number of documents. The rationale, as with many statistical learning algorithms, is that the more documents in the training data, the better the algorithm performs, as it has more examples to leverage and overall improves performance. This presupposes that better stability mirrors performance improvements, and that with more data, a germinator can better converge on the ground truth reflecting underlying knowledge about narratives in the news.

In most circumstances, contrary to expectations, as the number of documents decreases, stability increases, with linear induction constituting the sole exception. There are two possible explanations for this: (1) the number of documents withheld in cross-validation decreases with the number of documents, and stability depends more on the number of documents that differ between cross-validations than on the fraction of documents that differ, or (2) a fixed number of schemas can better capture the information contained in a smaller set of documents because new documents add more novel and unique narratives, thereby making the content of the narratives more difficult to capture in a finite set. In other words, provided with more and more documents, the germinators do not converge to some finite set of schemas, but instead are presented with an increasingly difficult problem to solve. These two explanations are not completely orthogonal; however, the first is a far more mechanical explanation. It is simply that more word types are contained in a larger set of documents, and therefore will be subject to a larger Zipfian tail, which is more difficult for a finite set of recorded event verbs to cover. The second presupposes that the system has some semblance of understanding the language data and narrative, and the challenge comes from the increased diversity of narratives contained therein. The second explanation could cause the first,

but the second claim requires a much greater burden of evidence because of its deeper implications.

Unlike the other germinators, linear induction has no limit to the number of schemas it can produce, albeit with a great number of schemas containing single events. This means there are two possible explanations for its behavior. The first is the explanation originally hoped for—that as the algorithm is given more documents, it begins to converge on a stable core of knowledge derived from the source data that in one form or another represents a consistent understanding of the data. The second explanation is more mundane—that what's actually stable is the schemas containing single events, schemas which then bias the apparent stability upward as more events that do not cross the β threshold are observed in a greater number of documents.

The second explanation is more likely here. The linear induction schemas here generated 1,544,879 single event containing schemas, roughly 57.2% of the total schemas generated by linear induction. Additionally, the LI-Truncated stability provides some insight here. If there is a stable content core that linear induction is approaching, then the first 800 schemas generated should reflect that. What we see instead is the lowest stability scores across the board. However, note that as ablation 9 is approached, the stability of the LI-Truncated schemas increases, likely because the number of schemas actually produced by linear induction is approaching 800, so the effect of the truncation is vanishing.

While the arbitrary clipping of the linear induction schemas greatly decreases stability, this may point to the unexpected decrease in stability in the other germinators as the number of documents increases. Given their hard limit on the number of schemas used, they are in some sense performing a similar "clipping" of the content contained in the documents. However, their increased stability, while still conducting a form of clipping, could be attributed to performing a more informed clipping than the LI-Truncated schemas.

8 Conclusions

We have explored the stability of narrative schemas. We used both an ablation and cross-validation of the data to produce different sets of schemas and compared them using the Fuzzy Jaccard coefficient and the Jaccard Reciprocal Fraction, which produces a transformation of the Fuzzy Jaccard coefficient that is easier to interpret.

Our results affirmed the homogeneous-heterogeneous distinction found in Simonson & Davis (2016). The homogeneous categories produced more stable batches of schemas than the heterogeneous ones. The counter-training and linear induction germinators produced more stable results, but the linear induction truncated results indicate that much of the apparent stability of the linear induction germinator is contained in its long tail of schemas.

Additionally, contradicting expectations, the schemas produced were more stable when given fewer documents. It is difficult to say whether this is because fewer documents were removed at every step of the cross-validation or because there is simply fewer narratives to capture in the same number of schemas. These explanations do not necessarily contradict one another, and the effect witnessed may be a mixture of both.

Understanding the stability of schemas provides us with a window into the quantitative properties of news narratives at large. Additionally, stability provides another technique for evaluating what makes a "good set of schemas:" as objects that are consistent regardless of what subset of a corpus they are derived from. Not for all purposes is this property necessarily "good," but the stability procedure provides a quantitative metric for this property if it is desired.

For sake of comparison, further analysis, and potential use in other projects, we have also made our schemas publicly available for potential use in the future.[5]

Acknowledgements

We must thank the Georgetown University Department of Linguistics for its continued support, as well as Amir Zeldes and Nate Chambers for feedback on this work. We also would like to thank the reviewers, whose reviews exhibited a great deal of care and attention to detail.

[5] https://schemas.thedansimonson.com/

References

Balasubramanian, N., Solderland, S., Mausam, & Etzioni, O. (2013). Generating coherent event schemas at scale. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1721–1731). Association for Computational Linguistics.

Chambers, N., & Jurafsky, D. (2008). Unsupervised learning of narrative event chains. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)* (pp. 789–797). Association for Computational Linguistics.

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

Cheng, P., & Erk, K. (2018). Implicit argument prediction with event knowledge.. Retrieved from `arXiv:1802.07226`

Corley, C., & Mihalcea, R. (2005). Measuring the semantic similarity of texts. In *Proceedings of the acl workshop on empirical modeling of semantic equivalence and entailment* (pp. 13–18).

de Marneffe, M., MacCartney, B., & Manning, C. (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.

Jans, B., Bethard, S., Vulić, I., & Moens, M. (2012). Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 336–344).

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning* (pp. 957–966).

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, *39*(4), 885–916.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Mooney, R., & DeJong, G. (1985). Learning schemata for natural language processing. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 681–687).

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., ... Allen, J. (2016). A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Pichotta, K., & Mooney, R. J. (2014). Statistical script learning with multi-argument events. In *EACL* (Vol. 14, 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.*

Rudinger, R., Rastogi, P., Ferraro, F., & Van Durme, B. (2015). Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-15).*

Sandhaus, E. (2008). The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*, *6*(12), e26752.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* New Jersey: Lawrence Erlbaum.

Simonson, D. (2018). *Investigations of the properties of narrative schemas* (Unpublished doctoral dissertation). Georgetown University.

Simonson, D., & Davis, A. (2015). Interactions between narrative schemas and document categories. In *Proceedings of the First Computing News Storylines Workshop (CnewS) at ACL-IJCNLP 2015* (p. 1). Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Simonson, D., & Davis, A. (2016). NASTEA: Investigating narrative schemas through annotated entities. In *Proceedings of the Second Workshop on Computing News Storylines (CNS 2016) at EMNLP 2016* (pp. 57–66). Austin, Texas: Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W16-5707

Yangarber, R. (2003). Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1* (pp. 343–350). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/1075096.1075140 doi: 10.3115/1075096.1075140