Explaining Mixtures of Sources in News Articles

Anonymous ACL submission

Abstract

Human writers plan, *then* write (Yao et al., 2019). For large language models (LLMs) to play a role in longer-form article generation, we must understand the planning steps humans make before writing. We explore one kind of planning, source-selection in news, as a case-study for evaluating plans in long-form generation. We ask: why do *specific* stories call for *specific* kinds of sources? We imagine a process where sources are selected to fall into different categories. Learning the article's *plan* means predicting the categorization scheme chosen by the journalist. Inspired by latent-variable modeling, we first develop metrics to select the most likely plan underlying a story. Then, working with professional journalists, we adapt five existing approaches to planning and introduce three new ones. We find that two approaches, or schemas: *stance* (Hardalov et al., 2021) and *social affiliation* best explain source plans in most documents. However, other schemas like *textual entailment* explain source plans in factually rich topics like "Science". Finally, we find we can predict the most suitable schema given just the article's headline with reasonable accuracy. We see this as an important case-study for human planning, and provides a framework and approach for evaluating other kinds of plans, like discourse or plot-oriented plans. We release a corpora, *NewsSources*, with schema annotations for 4M articles, for further study.

1 Introduction

As language models (LMs) become more proficient at long-form text generation and incorporate resources (Lewis et al., 2020) and tools (Schick et al., 2023) to support their writing, recent work has shown that planning before writing is essential (LeCun, 2022; Spangher et al., 2023a; Park et al., 2023). However, supervised datasets to support learning and studying plans are few: they are difficult or expensive to collect, synthetic, or narrowly tailored to specific domains (Zhou et al., 2023).

Headline: NJ Schools Teach Climate Change at all Grade Levels
Michelle Liwacz asked her first graders: what can penguins do to adapt to a warming Earth? ← *potential labels:* Academic, Neutral
Gabi, 7, said a few could live inside her fridge. ← *potential labels:* Unaffiliated, Neutral
Tammy Murphy, wife Governor Murphy, said climate change education was vital to help students. ← *poten. labels:* Government, Agree
Critics said young kids shouldn't learn disputed science. ← *labels:* Unaffiliated, Refute
A **poll** found that 70 percent of state residents supported climate change being taught at schools. ← *potential labels:* Media, Agree

Table 1: Informational sources synthesized in a single news article. *How would we choose sources to tell this story?* We show two different explanations, given by two competing schema: affiliation and stance. Our central questions: (1) *Which schema best explains the sources used in this story?* (2) *Can we predict, given a topic sentence, which schema to use?*

One approach to collecting large amounts of diverse planning data is to observe natural scenarios in which planning has already occurred. In this work, we consider one such real-world scenario: source selection by human journalists. Consider the article shown in Table 1. The author shares her plan[1]:

> *NJ schools are teaching climate change in elementary school. We wanted to understand: how are **teachers** educating children? How do **parents** and **kids** feel? Is there **pushback**?*

As can be seen, the journalist planned, before writing, the different kinds of sources (e.g. teachers,

[1]Plan: `https://nyti.ms/3Tay92f` [paraphrased]. Final article: `https://nyti.ms/486I11u`, see Table 1.

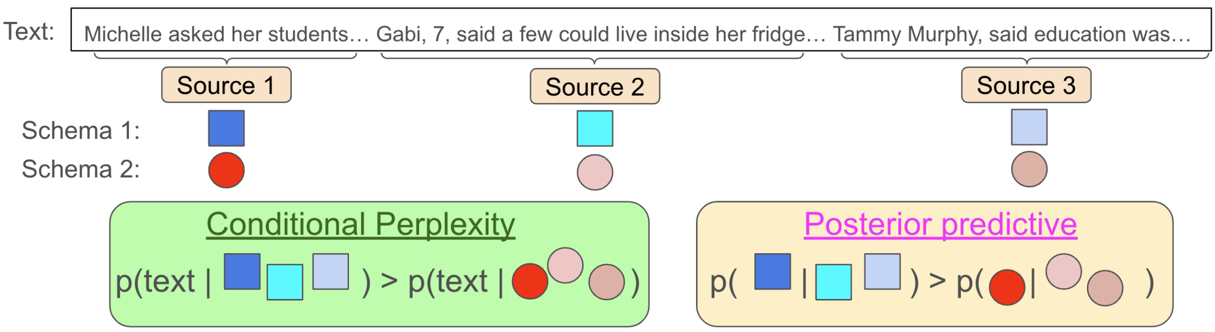


Figure 1: We seek to infer unobserved *plans* in natural data, focusing on one scenario: source-selection made by human journalists during news writing. Although the *reasons* why sources are chosen are unobservable, we show that one explanation (in the diagram, represented by *squares*: { , , }), is preferred over another (represented by *circles*: { , , }) if it better predicts the observed text (*conditional perplexity*) and the explanation is more internally consistent (*posterior predictive*).

kids) she wished to use. *Why did she choose these groups?* Was it: A. to include varied social groups? B. to capture different sides of an issue?

Answering this question, we argue, allows us to infer why she chose each source. If the answer is A, we can infer, then, that the writer probably chose her sources because each fell into a different social group. If the answer is B, the sources were more likely chosen because each agreed or disagreed with the main event. Table 1 shows this duality. Establishing $P(A) > P(B)$ means we can better infer why each source was used, allowing us to collect plans from natural text data.

Now, the core problem in this endeavor emerges: a document's plan is not typically observable. We directly address this and show that *we can differentiate between plans in naturally observed text*. Inspired by latent variable modeling approaches (Airoldi and Bischof, 2016), we uncover a document's most likely plan on the following basis: *a proposed plan better describes a document's actual plan if it gives more information about the completed document*. We introduce simple metrics for this goal: conditional perplexity and posterior predictive likelihood, in Figure 1 (Section 2.2).

Next, to create a straightforward setting to demonstrate the power of these metrics, we work with professional journalists from multiple major news organizations to identify planning approaches they regularly take. We operationalize these approaches as schema, or explanatory frameworks under which each source in the news article is assigned to a different discrete category (e.g. in the *affiliation* schema, for example, the source-categories would be *Government*, *Media*...). We adapt five schema from parallel tasks and introduce three novel schemas to better describe sourcing criterion. We implement our schemas by annotating over 600 news articles with 4,922 sources and training supervised classifiers. We validate our approach with these journalists: **they deem the plans we infer as correct with** $>$ **.74 F1 score.**

Finally, the choice of schema, we find, can be predicted with moderate accuracy using only the headline of the article (ROC=.67), opening the door to new computational journalism tooling.

In sum, our contributions are threefold:

- We frame *source-type planning* as a lens through which to study planning in writing.
- We collect 8 different plan descriptions, or *schemas* (5 existing and 3 we develop **with professional journalists**). We build a pipeline to extract sources from 4 million news articles and categorize them, building a large public dataset called *NewsSources*.
- We introduce two novel metrics: *conditional perplexity* and *posterior predictive* to compare plans. We find that different plans are optimal for different topics. Further, we show that the right plan can be predicted with .67 ROC given just the headline.

With this work, we hope to inspire further unsupervised inferences in document generation. Studying journalistic decision-making is important for understanding our information ecosystem (Winter and Krämer, 2014; Manninen, 2017; DeButts and Pan, 2024), can lead to important computational journalism tools (Quinonez and Meij, 2024) and presents a real-world case-study in planning.

2 Source Categorization

2.1 Problem Statement

Our central question is: why did the writer select sources $s_1, s_2, s_3...$ for document d? Intuitively, let's say we read an article on a controversial topic. Let's suppose we observe that it contains many opposing viewpoints: some sources in the article "agree" with the main topic and others "disagree". We can conclude that the writer probably chose sources on the basis of their *stance* (Hardalov et al., 2021) (or their opinion-based support) rather than another explanation, like their *discourse* role (which describes their narrative function).

More abstractly, we describe source-selection as a generative process: first, journalists plan *how* they will choose sources (i.e. the *set* of k categories sources will fall into), then they choose sources, each falling into 1-of-k categories. Different plans, or categorizations, are possible (e.g. see Figure 1): *the "right" plan is the one that best predicts the final document.*

Each plan, or categorizations, is specified by a *schema*. For the 8 schemas used in this work, see Figure 2. To apply a schema to a document, we frame an approach consisting of two components: (1) an attribution function, a:

$$a(s) = q \in Q_d \text{ for } s \in d \tag{1}$$

introduced in Spangher et al. (2023b), which maps each sentence s in document d to a source $Q_d = \{q_1^{(d)}, ...q_k^{(d)}\}$[2] and (2) a classifier, c:

$$c_Z(s_1^{(q)}, ...s_n^{(q)}) = z \in Z \tag{2}$$

which takes as input a sequence of sentences attributed to source $q^{(d)}$ and assigns a type $z \in Z$ for schema Z.

This supervised framing is not typical in latent-variable settings; the choice of z and the *meaning* of Z are typically jointly learned without supervision. However, learned latent spaces often do not correspond well to theoretical schemas (Chang et al., 2009), and supervision has been shown to be helpful with planning (Wei et al., 2022). On the other hand, supervised models trained on different schema are challenging to compare, especially when different architectures are optimal for each schema. A latent-variable framework here is ideal: comparing different graphical models (Bamman et al., 2013; Bamman and Smith, 2014) *necessitates* comparing different schemas, as each run of a latent variable model produces a different schema.

2.2 Comparing *Plans*, or Schemas

We can compare plans in two ways: (1) how well do they explain each observed document? and (2) how structurally consistent are they?

Explainability A primary criterion for a *plan* is for it to explain the observed data well. To measure this, we use *conditional perplexity*[3]

$$p(x|z) \tag{3}$$

which measures the uncertainty of observed data, x, given a latent structure, z. Measuring $p(x|z)$ for different z (fixing x) allows us to compare z. Conditional perplexity is a novel metric we introduce, inspired by metrics to evaluate latent unsupervised models, like the "left-to-right" algorithm introduced by (Airoldi and Bischof, 2016). [4]

Structural Likelihood: A second basic criterion for a latent structure to be useful is for it be consistent, which is a predicate for learnability. We assess the consistency of a set of assignments, z, by calculating the *posterior predictive*:

$$p(z|z_{-}, x) \tag{4}$$

Deng et al. (2022) exploring using full joint distribution, $p(z)$, *latent perplexity*, to evaluate the structure text x produced by generative language models ("*model criticism*"). We simplify using the full distribution and instead evaluate the conditional predictive to study document structure. This, we find in early experiments, is easier to learn and thus helps us differentiate different Z better ("*schema criticism*").[5] Now, we describe our schemas.

For an illustration of each metric, please refer to Figure 1. The overall goal of the metrics is to determine *which schema, or labeling of sources, best explains the observed news article*. As the

[2] These sources are referenced in d. There is no consideration of document-independent sources.

[3] We abuse notation here, using p as both probability and perplexity: $p(x) = \exp\{-\mathbb{E} \log p(x_i|x_{<i})\}$.

[4] We note that the term, *conditional perplexity*, was originally introduced by Zhou and Lua (1998) to compare machine-translation pairs. In their case, both x and z are observable; as such, they do not evaluate latent structures, and their usage is not comparable to ours.

[5] Our work is inspired by Spangher et al. (2023b)'s work, where z was the choice of specific action, rather than a general action-type.

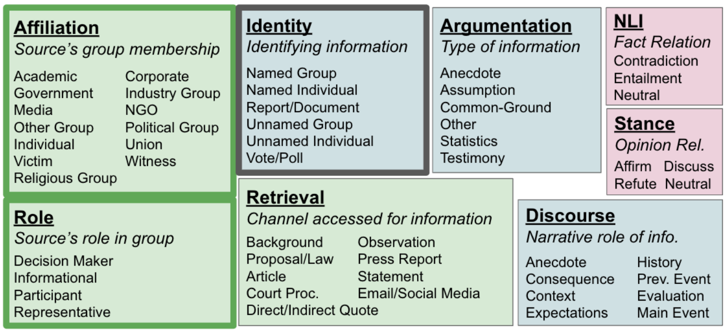


Figure 2: Label-sets for source-planning schemas. **Extrinsic Source Schemas** Affiliation, Role and Retrieval-method (Spangher et al., 2023b) capture characteristics of sources *extrinsic* to their usage in the document. **Functional Source Schemas:** Argumentation (Al Khatib et al., 2016), Discourse (Choubey et al., 2020) and Identity capture functional narrative role of sources. **Debate-Oriented Schemas**: Natural Language Inference (NLI) (Dagan et al., 2005) and Stance (Hardalov et al., 2021) capture the role of sources in encompassing multiple sides. The three novel schemas we introduce are shown with borders: Affil., Identity and Role. For definitions, see App. D.

figure shows, if schema A describes an article better than schema B, then labels assigned to each source under schema A (e.g. in Figure 1: squares, , ,) will outperform labels assigned under Schema B (e.g. circles, , ,).

2.3 Source Schemas

Our schemas, or descriptions of plans, are shown in Figure 2. In this work, we collect 8 schemas including three we introduce: *Identity, Affiliation and Role*. Each schema provides a set of categories describing the sources used in a news article. See Appendex D for more details and definitions for each schema.

We note that ***none*** of these schemas are complete and that real-world plans likely have elements outside of any one schema (or are combinations of multiple schema). However, this demonstration is important, we argue, to prove that we *can* differentiate between purely latent plans in long-form text. We now introduce each schema:

Debate-Oriented Schemas Both the *Stance* and *NLI* schemas are debate-orienced schemas. They each capture the relation between the information a source provides and the main idea of the article. *NLI* (Dagan et al., 2005) captures factual relations between text, while *Stance* (Hardalov et al., 2021) captures opinion-based relations . A text pair may be factually consistent and thus be classified as "Entailment" under a *NLI* schema, but express different opinions and be classified as "Refute" under *Stance*. These schemas say a writer uses sources for the purpose of expanding or rebutting information in the narrative.

Schema	Macro-F1	Schema	Macro-F1
Argumentation	68.3	Retrieval	61.3
NLI	55.2	Identity	67.2
Stance	57.1	Affiliation	53.3
Discourse	56.1	Role	58.1

Table 2: Classification f1 score, macro-averaged, for the 8 schemas. We achieve moderate classification scores for each of schema. In Section 2, when we compare schemas, we account for classification acc. differences by introducing noise to higher-performing classifiers.

Functional Source Schemas The following schemas: *Argumentation*, *Discourse* and *Identity* all capture the role a source plays in the overall narrative construction of the article. For instance, a source might provide a "Statistic" for a well-formed argument (*Argumentation* (Al Khatib et al., 2016)), or "Background" for a reader to help contextualize (*Discourse* (Choubey et al., 2020)). *Identity*, a novel schema, captures how the reader identifies the source. For example, a "Named Individual" is identifiable to a reader, whereas an "Unnamed Individual" is not. As identified in Sullivan (2016) and our journalist collaborators, this can be a strategic planning choice: some articles are about sensitive topics and need unnamed sources.

Extrinsic Source Schemas *Affiliation*, *Role* and *Retrieval* schemas serve to characterize attributes of sources external to the news article. Stories often implicate social groups (McLean et al., 2019), such as "academia" or "government." Those group identities are extrinsic to the story's architecture but important for the selection of sources. Sources may be selected because they represent a group (i.e. *Affiliation*) or because their group position is important within the story's narrative (e.g. "participants" in the events, i.e. *Role*). *Retrieval*, introduced by Spangher et al. (2023b), captures the channel through which the information was found. Although these schema are news-focused, we challenge the reader to imagine ones that might exist in other fields. For instance, a machine learning article might compare models selected via, say, a *Community* schema: each from *open-source*, *academic* and *industry research* communities.

3 Building a Silver-Standard Dataset of Different Possible Plans

The schemas described in the previous section give us theoretical frameworks for identifying writers' plans. To *compare* plans and *select the plan that best describes a document*, we must first create a dataset where informational sources are labeled *according to each schema*.

3.1 Dataset Construction and Annotation

We obtain the NewsEdits dataset (Spangher et al., 2022), which consists of 4 million news articles, and extract sources using a methodology developed by Spangher et al. (2023b), which authors established was state-of-the-art for this task. This dataset spans 12 different news sources (e.g. BBC, NYTimes, etc.) over a period of 15 years (2006-2021). For our experiments, we sample $90,000$ news articles that are long and contain more than 3 sources (on average, the articles contain ~ 7.5 sources).

We annotate sources under each of our new schemas. We recruited two annotators, one an undergraduate and the other a former journalist. The former journalist trained the undergraduate for 1 month to identify and label sources, then, they independently labeled 425 sources in 50 articles with each schema to calculate agreement, scoring $\kappa = .63, .76, .84$ on *Affiliation*, *Role* and *Identity* labels. They then labeled 4,922 sources in 600 articles with each schema, labeling roughly equal amounts. Finally, they jointly labeled 100 sources in 25 documents with the other schemas for evaluation data over 1 month, with $\kappa \geq .54$, *all in the range of moderate to substantial agreement* (Landis and Koch, 1977).

3.2 Training Classifiers to Label Sources

We train classifiers to label sources under each schema. Unless specified, we use a sequence classifier using RoBERTa-base with self-attention pooling, as in Spangher et al. (2021a). We deliberately chose smaller models to scale to large amounts of articles. We will open-source all of the classifiers trained in this paper.

Affiliation*, *Role*, *Identity We use our annotations to train classifiers which take as input all sentences attributable to source q and output a category in each schema, or $p(t|s_1^{(q)} \oplus ... \oplus s_n^{(q)})$.

Argumentation*, *Retrieval*, *Discourse We use datasets, without modification, that were directly released by the authors. Each is labeled on a sentence-level, on news and opinion datasets. We train classifiers to label each sentence of the news article, s. Then, for each source q, we assign a single label, y, with the most mutual information[6] across sentences attributed to that source, $s_1^{(q)}, ...s_n^{(q)}$.

NLI, Stance We use an NLI classifier trained by Williams et al. (2022) to label each sentence attributed to source q as a separate hypothesis, and the article's headline as the premise. We use mutual information to assign a single label.

We create a stance training dataset by aggregating several news-focused stance datasets[7]. We then fine-tune GPT3.5-turbo[8] to label news data and label 60,000 news articles. We distill a $T5$ model with this data (Table 2 shows T5's performance).

3.3 Classification Results

As shown in Table 2, we model schemas within a range of f1-scores $\in (53.3, 67.2)$, showing moderate success in learning each schema[9]. These scores are middle-range and likely not useful on their own; we would certainly have achieved higher scores with more state-of-the-art methods. However, we note *these classifiers are being used for comparative, explanatory purposes, so their efficacy lies in how well they help us compare plans*, as we will explore in the next section.

[6] $\arg\max_y p(y|q)/p(y))$

[7] FNC-1 (Pomerleau and Rao, 2017), Perspectrum (Chen et al., 2019), ARC (Habernal et al., 2017), Emergent (Ferreira and Vlachos, 2016) and NewsClaims (Reddy et al., 2021). We filter these sets to include premises and hypothesis ≥ 10 words and ≤ 2 sentences.

[8] We use OpenAI's GPT3.5-turbo fine-tuning endpoint, as of November 16, 2023.

[9] When using these classifier outputs for evaluating plans, in the next section, we introduce noise (i.e. random label-swapping), so that all have the same accuracy.

Schema	n	Conditional Perplexity $p(x\|z)$ PPL	Δ base-k (↓)	Δ base-r (↓)	Posterior Predictive $p(\hat{z}\|z_{-}, x)$ F1	÷ base-k (↑)	÷ base-r (↑)
NLI	3	22.8	0.62	-0.08	58.0	1.02**	1.01 **
Stance	4	21.5	-1.71	-3.21**	39.1	0.88**	0.83 **
Role	4	22.3	-0.06	-0.33**	38.7	1.11**	1.10 **
Identity	6	21.8	-0.42	-0.94	25.0	1.00	1.15 **
Argumentation	6	21.7	-0.52	-1.04	30.7	1.10 **	1.12 **
Discourse	8	22.3	0.54	-0.75	19.2	1.06 **	1.08 **
Retrieval	10	23.7	1.47	0.36	15.8	1.10 **	1.12 **
Affiliation	14	20.5	-2.11**	-3.04**	10.5	1.26 **	1.16 **

Table 3: Comparing our schemas against each other. In the first set of experiments, we show *conditional perplexity* results, which tell us how well each schema explains the document text. Shown is PPL (the mean perplexity per schema), $\Delta kmeans$ (PPL - avg. perplexity of kmeans) and $\Delta random$ (PPL - avg. perplexity of the random trial). Statistical significance ($p < .05$) via a t-test calculated over perplexity values is shown via **.In the second set of experiments, we show *posterior predictive* results, measured via micro F1-score. We show F1 (f1-score per schema), ÷ kmeans (F1 / f1-score of kmeans), ÷ random (F1 / f1-score of random trial). Statistical significance ($p < .05$) via a t-test calculated over 500-sample bootstrapped f1-scores is shown via **.

4 Comparing Schemas

We are now ready to explore how well these schemas explain source selection in documents. We start by describing our experiments, then baselines, and finally results. All experiments in this section are based on the $90,000$ news articles filtered from NewsEdits, labeled as described in the previous section. We split $80,000/10,000$ train/eval.

4.1 Implementing Planning Metrics

We now describe how we implement the metrics introduced in Section 2.2: (1) *conditional perplexity* and (2) *posterior predictive*.

Conditional Perplexity To measure *conditional perplexity*, $p(x|z)$, we fine-tune GPT2-base models (Radford et al., 2019) to take in it's prompt a sequence of latent variables, each for a different source, and *then assess likelihood of the observed article text.*[10] This is similar to measuring *vanilla perplexity* on observed text, except: (1) we provide latent variables as conditioning (2) by fixing the model used and varying the labels, *we are measuring the signal given by each set of different labels*. Our template for GPT2 is:

$$\langle \mathrm{h} \rangle \; h \; \langle \mathrm{l} \rangle \; (1) \; l_1 \; (2) \; l_2 \ldots \langle \mathrm{t} \rangle$$

$$(1) \; s_1^{(q_1)} \ldots s_n^{(q_1)} \; (2) \ldots$$

Red is the prompt, or conditioning, and green is the text over which we calculate perplexity. <tokens> (e.g. "(1)", "⟨text⟩") are structural markers while variables l, h, s are article-specific. h is the headline, l_i is the label for source i and $s_1^{(q_1)} \ldots s_n^{(q_1)}$ are the sentences attributable to source i. *We do not use GPT2 for generation, but for comparative purposes, to compare the likelihood of observed article text under each schema.* We note that this implements Eq. 3 only if we assuming green preserves the meaning of x, the article text. Our data processing (Section 3.1), based on high-accuracy source-extraction models (Spangher et al., 2023b), gives us confidence in this.[11]

Posterior Predictive To learn the *posterior predictive* (Equation 4), we train a BERT-based classification model (Devlin et al., 2018) to take the article's headline and a sequence of source-types *with a one randomly held out*. We then seek to predict *that* source-type, and evaluate using F1-score. Additionally, we follow Spangher et al. (2023b)'s observation that some sources are *more important* (i.e. have more information attributed). We model the posterior predictive among the 4 sources per article with the most sentences attributed to them.

4.2 Baselines

Vanilla perplexity does not always provide accurate model comparisons (Meister and Cotterell, 2021; Oh et al., 2022) because it can be affected by irrele-

[10] We note that this formulation has overlaps with recent work seeking to learn latent plans (Deng et al., 2022; Wang et al., 2023; Wei et al., 2022).

[11] Initial experiments show that text markers are essential for the model to learn structural cues. However, they also provide their own signal (e.g. on the number of sources). To reduce the effects of these artifacts, we use a technique called *negative prompting* (Sanchez et al., 2023). Specifically, we calculate perplexity on the *altered* logits, $P_\gamma = \gamma \log p(x|z) - (1-\gamma) \log p(x|\hat{z})$, where $\hat{z}$ is a shuffled version of the latent variables. Since textual markers remain the same in the prompt for z and $\hat{z}$, this removes markers' predictive power.

vant factors, like tokenization scheme. We hypothesized that the dimensionality of each schema's latent space might also have an effect (Lu et al., 2017); larger latent spaces tend to assign lower probabilities to each point. Thus, we benchmark each schema against baselines with similar latent dimensions.

Base-r, or Random baseline . We generate k unique identifiers[12], and randomly assign one to each source in each document. k is set to match the number of labels in the schema being compared to.

Base-k, or Kmeans baseline . We first embed sources as paragraph-embeddings using Sentence BERT (Reimers and Gurevych, 2019) [13] Then, we cluster all sources across documents into k clusters using the kmeans algorithm (Likas et al., 2003), where k is set to match the number of labels in the schema being compared to. We assign each source it's cluster number.

4.3 Results and Discussion

As shown in Table 3, the supervised schemas mostly have have lower conditional perplexity than their random and unsupervised kmeans baselines. However, only the *Stance*, *Affiliation* and *Role* schemas improve significantly (at $p < .001$), and the *Role* schema's performance increase is minor. *Retrieval* has a statistically significant less explainability relative to it's baselines.

There is a simple reason for why some schemas have either the same or more conditional perplexity compared to their baselines: they lack explainability over the text of the document, but are not random and thus might lead to overfitting. We examine examples and find that *Retrieval* does not impact wording as expected: writers make efforts to convey information similarly whether it was obtained via a quote, document or a statement.

We face a dilemma: in generating these schemas, we chose *Retrieval* because we assumed it was an important planning criterion. However, our results indicate that it holds little explanatory power. *Is it possible that some plans do not get reflected in the text of the document?*

To address this question, we assign $\hat{Z} = \arg\min_Z p(x|z)$, the schema for each datapoint with the lowest perplexity, using scores calculated in the prior section[14], we calculate the lowest-perplexity schema. Table 5 shows the distribution of such articles. We then task 2 expert journalists with assigning their *own* guess about which schema best describes the planning for the particular article, for 120 articles. **We observe an F1-score of 74, indicating a high degree of agreement.**

Interestingly, we also observe statistically significant improvements of kmeans over random baselines in all cases (except $k = 3$). In general, our baselines have lower variance in perplexity values than experimental schemas. This is not unexpected: as we will explore in the next section, we expect that some schemas will best explain some articles, resulting in a greater range in performance. For more detailed comparisons, see Appendix B.

Posterior predictive results generally show improvement across trials, with the *Affiliation* trial showing the highest improvement over both baselines. This indicates that most tagsets are, to some degree, internally consistent and predictable. *Stance* is the only exception, showing significantly lower f1 than even random baselines. This indicates that, although Stance is able to explain observed documents well (as observed by it's impact on conditional perplexity), it's not always predictable how it will applied. Perhaps this is indicative that writers do not know a-priori what sources will agree or disagree on any given topic before talking to them, and writers do not always actively seek out opposing sides.

Finally, as another baseline, we implemented latent variable model. In initial experiments, it does not perform well. We show in Appendex G that the latent space learned by the model is sensible. Bayesian models are attractive for their ability to encode prior belief, and ideally they would make good baselines for a task like this, which interrogates latent structure. However, more work is needed to better align them to modern deep-learning baselines.

5 Predicting Schemas

Taken together, our observations from (1) Section 3.3) indicate that schemas are largely unrelated and (2) Section 4.3 indicate that *Stance* and *Affiliation* both have similar explanatory power (although *Stance* is less predictable). We next ask: which kinds of articles are better explained by one schema, and which are better explained by the other?

[12]Using MD5 hashes, from python's `uuid` library.

[13]Specifically, `microsoft/mpnet-base`'s model https://www.sbert.net/docs/pretrained_models.html.

[14]across the dataset used for validation, or 5,000 articles

Stance	*Affiliation*
Bush, George W	Freedom of Speech
Swift, Taylor	2020 Pres. Election
Data-Mining	Jazz
Artificial Intelligence	Ships and Shipping
Rumors/Misinfo.	United States Military
Illegal Immigration	Culture (Arts)
Social Media	Mississippi

Table 4: Top keywords associated with articles favored by stance or affiliation. Keywords are manually assigned by news editors

In Table 4, we show topics that have low perplexity under the *Stance* schema, compared with the *Affiliation* schema (we calculate these by aggregating document-level perplexity across keywords assigned to each document in our dataset). As we can see, topics requiring greater degrees of debate, like "Artificial Intelligence", and "Taylor Swift" are favored under the *Stance* Topic, while broader topics requiring many different social perspectives, like "Culture" and "Freedom of Speech" are favored under *Affiliation*. We set up an experiment where we try to predict $\hat{Z} = \arg\min_Z p(x|z)$, the schema for each datapoint with the lowest perplexity. We downsample until assigned schemas, per articles, are balanced and train a simple linear classifier[15] to predict $\hat{Z}$. We get .67 ROC-AUC (or .23 f1-score). These results are tantalizing and offer the prospect of being able to *better plan source retrieval* in computational journalism tools, by helping decide an axis on which to seek different sources. More work is needed to validate these results.

6 Related Work

Latent Variable Persona Modeling Our work is inspired by earlier work in persona-type latent variable modeling (Bamman et al., 2013; Card et al., 2016; Spangher et al., 2021b). Authors model characters in text as mixtures of topics. We both seek to learn and reason about about latent character-types, but their line of work takes an unsupervised approach. We show that supervised schemas outperform unsupervised.

Multi-Document Retrieval In multiple settings – e.g. multi-document QA (Pereira et al., 2023), multi-document summarization (Shapira et al., 2021), retrieval-augmented generation (Lewis et al., 2020) – information *from a single source* is assumed to be insufficient to meet a user's needs. In

[15]Bag-of-words with logistic regression

Affiliation	41.7%	Argument.	1.2%
Identity	22.7%	Discourse	1.1%
Stance	17.7%	NLI	1.1%
Role	13.4%	Retrieval	1.1%

Table 5: Proportion of our validation dataset favored by one schema, i.e. $\hat{Z} = \arg\max_Z p(x|z)$

typical information retrieval settings, the goal is to retrieve a single document closest to the query (Page et al., 1998). In settings where *multiple sources are needed*, on the other hand, retrieval goals are not clearly understood[16]. Our work attempts to clarify this, and can be seen as a step towards better retrieval planning.

Planning in Language Models Along the line of the previous point, chain-of-thought reasoning (Wei et al., 2022) and few-shot prompting, summarized in (Sanchez et al., 2023), can be seen as latent-variable processes. Indeed, work in this vein is exploring latent-variable modeling for shot selection (). Our work, in particular the *conditional perplexity* formulation and it's implementation, can be seen as a way of comparing different chain-of-thought plans as they relate to document planning.

Computational Journalism seeks to apply computational techniques to assist journalists in reporting. Researchers have sought to improve detection of incongruent information (Chesney et al., 2017), detect misinformation (Pisarevskaya, 2017) and false claims made in news articles (Adair et al., 2017).

7 Conclusions

In conclusion, we explore ways of thinking about sourcing in human writing. We compare 8 schemas of source categorization, and adapt novel ways of comparing them. We find, overall, that *affiliation* and *stance* schemas help explain sourcing the best, and we can predict which is most useful with moderate accuracy. Our work lays the ground work for a larger discussion of discovering plans made by humans in naturally generated documents. It also takes us steps towards tools that might be useful to journalists. Naturally, our work is a simplification of the real human processes guiding source selection; these categories are non-exclusive and inexhaustive. We hope by framing these problems we can spur further research in this area.

[16]As Pereira et al. (2023) states, *"retrievers are the main bottleneck"* for well-performing multi-document systems.

8 Limitations

A central limitation to our work is that the datasets we used to train our models are all in English. As mentioned previously, we used English language sources from Spangher et al. (2022)'s *NewsEdits* dataset, which consists of sources such as nytimes.com, bbc.com, washingtonpost.com, etc. Thus, we must view our work with the important caveat that non-Western news outlets may not follow the same source-usage patterns and discourse structures in writing their news articles as outlets from other regions. We might face extraction and labeling biases if we were to attempt to do such work in other languages.

Another limitation of our work is that we only considered 8 supervised schemas. While we worked closely with journalists to develop these schemas and attempted to make them as comprehensive and useful as possible, it's entirely possible, in fact probable, that these 8 schemas do not describe sources that well. As mentioned in the main body, we fully anticipate that more work needs to be done to determine further, more optimal schemas. And it's likely, ultimately, that unsupervised approaches to developing more nuanced plans are desirable.

Furthermore, the metrics we evaluated are approximate and depend on schemas learned by ML models. Both of these facts could incentivize biased models. However, we attempted to mitigate this by conducting an experiment afterwards with journalists to blindly label articles.

Our annotation approach was done only two annotators, in a master-apprentice style and hence might be skewed in distribution. However, because the master was an experienced journalist with many years of newsroom experience at a major newsroom, we took their tagging to be gold-standard.

9 Ethics Statement

9.1 Risks

Since we constructed our datasets on well-trusted news outlets, we assumed that every informational sentence was factual, to the best of the journalist's ability, and honestly constructed. We have no guarantees that our classification systems would work in a setting where a journalist was acting adversarially.

There is a risk that, if planning works and natural language generation works advance, it could fuel actors that wish to use it to plan misinformation and propaganda. Any step towards making generated news article more human-like risks us being less able to detect and stop them. Misinformation is not new to our media ecosystem, (Boyd et al., 2018; Spangher et al., 2020). We have not experimented how our classifiers would function in such a domain. There is work using discourse-structure to identify misinformation (Abbas, 2022; **?**), and this could be useful in a source-attribution pipeline to mitigate such risks.

We used OpenAI Finetuning to train the GPT3 variants. We recognize that OpenAI is not transparent about its training process, and this might reduce the reproducibility of our process. We also recognize that OpenAI owns the models we fine-tuned, and thus we cannot release them publicly. Both of these thrusts are anti-science and anti-openness and we disagree with them on principle. We tried where possible to train open-sourced versions, as mentioned in the text.

9.2 Licensing

The dataset we used, *NewsEdits* (Spangher et al., 2022), is released academically. Authors claim that they received permission from the publishers to release their dataset, and it was published as a dataset resource in NAACL 2023. We have had lawyers at a major media company ascertain that this dataset was low risk for copyright infringement.

9.3 Computational Resources

The experiments in our paper required computational resources. We used 64 12GB NVIDIA 2080 GPUs. We designed all our models to run on 1 GPU, so they did not need to utilize model or data-parallelism. However, we still need to recognize that not all researchers have access to this type of equipment.

We used Huggingface models for our predictive tasks, and will release the code of all the custom architectures that we constructed. Our models do not exceed 300 million parameters.

9.4 Annotators

We recruited annotators from our educational institutions. They consented to the experiment in exchange for mentoring and acknowledgement in the final paper. One is an undergraduate student, and the other is a former journalist. Both annotators are male. Both identify as cis-gender. The annotation conducted for this work was deemed exempt from review by our Institutional Review Board.

References

Ali Haif Abbas. 2022. Politicizing the pandemic: A schemata analysis of covid-19 news in two selected newspapers. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 35(3):883–902.

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward “the holy grail”: The continued quest to automate fact-checking. In *Computation+ Journalism Symposium, Evanston*.

Edoardo M Airoldi and Jonathan M Bischof. 2016. Improving and evaluating topic models and other models of text. *Journal of the American Statistical Association*, 111(516):1381–1403.

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

David Bamman, Brendan O’Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Ryan L Boyd, Alexander Spangher, Adam Fourney, Besmira Nushi, Gireeja Ranade, James Pennebaker, and Eric Horvitz. 2018. Characterizing the internet research agency’s social media operations during the 2016 us presidential election using linguistic analyses.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of NAACL-HLT*, pages 542–557.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Harald Cramér. 1999. *Mathematical methods of statistics*, volume 43. Princeton university press.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Matt DeButts and Jennifer Pan. 2024. Reporting after removal: the effects of journalist expulsion on foreign news coverage. *Journal of Communication*, page jqae015.

Yuntian Deng, Volodymyr Kuleshov, and Alexander M Rush. 2022. Model criticism for long-form text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11887–11912.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Kun Lu, Xin Cai, Isola Ajiferuke, and Dietmar Wolfram. 2017. Vocabulary size and its effect on topic representation. *Information Processing & Management*, 53(3):653–665.

Ville JE Manninen. 2017. Sourcing practices in online journalism: An ethnographic study of the formation of trust in and the use of journalistic sources. *Journal of Media Practice*, 18(2-3):212–228.

Kate C McLean, Moin Syed, Kristin Gudbjorg Haraldsson, and Alexandra Lowe. 2019. Narrative identity in the social world: The press for stability. *Handbook of Personality Psychology*.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The pagerank citation ranking: Bring order to the web. Technical report, Technical report, stanford University.

Kyeongman Park, Nakyeong Yang, and Kyomin Jung. 2023. Longstory: Coherent, complete and length controlled long story generation. *arXiv preprint arXiv:2311.15208*.

Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. 2023. Visconde: Multi-document qa with gpt-3 and neural reranking. In *European Conference on Information Retrieval*, pages 534–543. Springer.

Dina Pisarevskaya. 2017. Deception detection in news reports in the Russian language: Lexics and discourse. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 74–79, Copenhagen, Denmark. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (fnc-i): Stance detection. *Retrieved March*, 15:2023.

Claudia Quinonez and Edgar Meij. 2024. A new era of ai-assisted journalism at bloomberg. *AI Magazine*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677.

Alexander Spangher, Xinyu Hua, Yao Ming, and Nanyun Peng. 2023a. Sequentially controlled text generation. *arXiv preprint arXiv:2301.02299*.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021a. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2021b. " don't quote me on that": Finding mixtures of sources in news articles. *arXiv preprint arXiv:2104.09656*.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023b. Identifying informational sources in news articles. *arXiv preprint arXiv:2305.14904*.

Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. 2020. Characterizing search-engine traffic to internet research agency web properties. In *Proceedings of The Web Conference 2020*, pages 2253–2263.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Margaret Sullivan. 2016. Tightening the screws on anonymous sources. *New York Times*.

Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: a corpus of quotations from a decade of news. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 328–336.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. Anlizing the adversarial natural language inference dataset.

Stephan Winter and Nicole C Krämer. 2014. A question of credibility–effects of source cues and recommendations on information selection on news sites and blogs. *Communications*, 39(4):435–456.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

GuoDong Zhou and KimTeng Lua. 1998. Word association and MI-Trigger-based language modeling. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1465–1471, Montreal, Quebec, Canada. Association for Computational Linguistics.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Appendix

In Appendix A, we include more, precise detail about our experimental methods. Then, Appendix B, we present more exploratory analysis to support our experiments, including comparisons between schemas. In Appendix D, we give a more complete set of definitions for the labels in each schema. In Appendix G, we define the unsupervised latent variable models we use as baselines, including providing details on their implementation.

A Additional Methodological Details

A.1 Source Extraction

Before classifying sources, we first need to learn an attribution function (Equation 1) to identify the set of sources in news articles. Spangher et al. (2023b) introduced a large source attribution dataset, but their models are either closed (i.e. GPT-based) or underperforming. So, we train a high-performing open-source model using their dataset. We fine-tune GPT3.5-turbo [17], achieving a prediction accuracy of 74.5% on their test data[18]. Then, we label a large silver-standard dataset of 30,000 news articles and distill a BERT-base span-labeling model, described in (Vaucher et al., 2021), with an accuracy of 74.0%.[19] We use this model to score a large corpus of $90,000$ news articles from the NewsEdits corpus (Spangher et al., 2022). We find that 47% of sentences in our documents can be attributed to sources, and documents each contain an average of 7.5 +-/5 sources. These statistics are comparable to those reported by Spangher et al. (2023b).

B Exploratory Data Analysis

We explore more nuances of our schemas, including comparative analyses. We start by showing a view of $\hat{Z}$, the conditions under which a schema best explains the observed results. In Tables 6 and 7, we show an extension of Table 4 in the main body: we show favored keywords across all schemas. (Note that in contrast to Table 4, we restrict the keywords we consider to a tighter range). When topics require a mixture of different informa-

[17] As of November 30th, 2023.

[18] Lower than the reported 83.0% accuracy of their Curie model. We formulate a different, batched prompt aimed at retrieving more data.

[19] All models will be released.

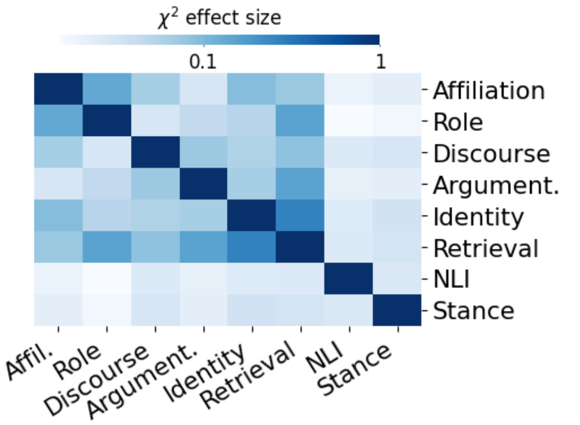


Figure 3: Correlation between 8 schemas, measured as Cramer's V (Cramér, 1999), or the effect-size measurement of the χ^2 test of independence.

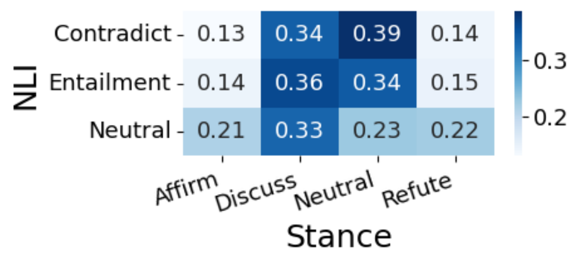


Figure 4: Stance and NLI schema definitions are not very aligned. We show conditional probability of labels in each category, $p(x|y)$ where x = Stance and y = NLI.

tion types, like statistics, testimony, etc. *Argumentation* is favored. When story-telling is on topics like "Travel", "Education", "Quarantine (Life and Culture)", where it incorporates background, history, analysis, expectation, *Discourse* is favored. In Table 9, we show the top *Affiliations* per section of the newspaper, based on the NYT LDC corpus (Sandhaus, 2008).

Next, we further explore the relation between different labelsets. In Figure 5, we show the same story as in Table 3 in the Main Body, except with a broader view of the distributional shifts. As can be seen, by comparing differents between the means in Table 3 and the medians in 5, we see that the effect of outliers is quite large, which reduces the significance we observe. In 7, we show the correlation between perplexities across labelsets. We observe clusters in our schemas of particularly high correlation. Interestingly, this stands in contrast to Figure 3, which showed almost no relation between the tagsets. We suspect that outlier effects on perplexity (e.g. misspelled words, strange punctuation)

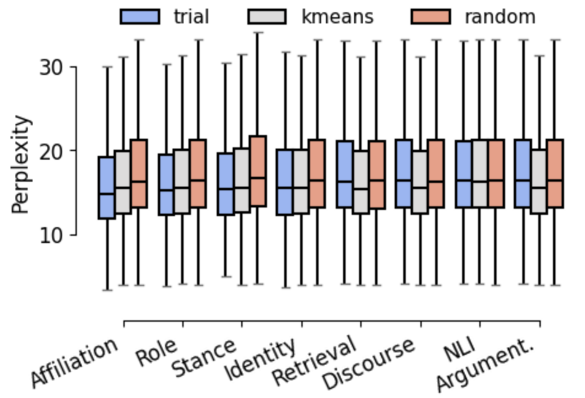
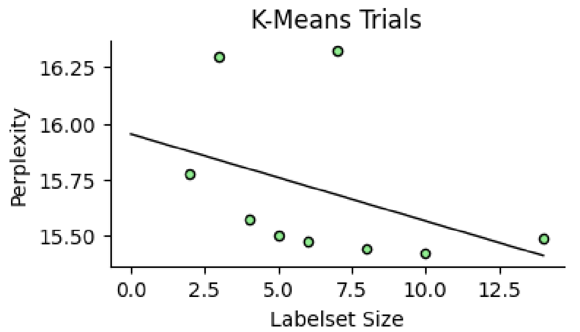
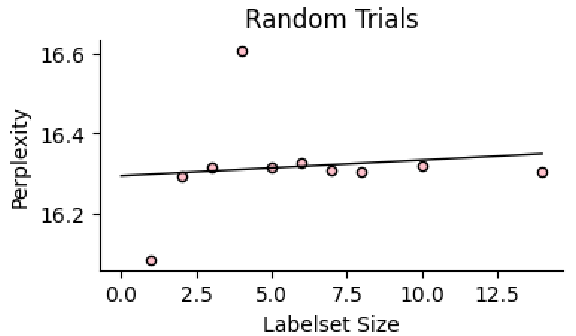


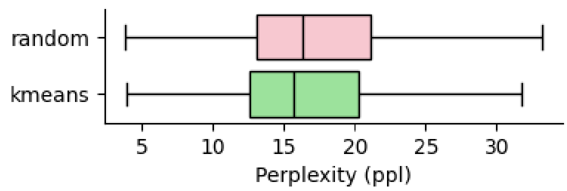
Figure 5: Distribution of conditional perplexity measurements across different experimental groups.



(a) Relationship between the size of the labelset and perplexity for kmeans trials



(b) Relational between the size of the labelset and perplexity for random trials.



(c) Distribution over perplexity scores for all random trials and kmeans trials, compared.

Figure 6: To explore the effects of labelset size, and confirm that conditional perplexity does align with basic intuitions, we compare Random trials and Kmeans trials across all of our labelset sizes.

Affiliation	Argumentation	Discourse	NLI
Inflation (Economics)	Race and Ethnicity	Travel and Vacations	Deaths (Fatalities)
Writing and Writers	Books and Literature	Quarantine (Life and Culture)	Murders, Homicides
United States Economy	Demonstrations, Protests and Riots	Education (K-12)	Law and Legislation
Race and Ethnicity	Travel and Vacations	Fashion and Apparel	States (US)
Disease Rates	Suits and Litigation	Murders, Homicides	Science
Real Estate and Housing (Residential)	Senate	Great Britain	Politics and Government
China	United States International Relations	Deaths (Fatalities)	Personal Profile
Supreme Court (US)	Deaths (Fatalities)	Pop and Rock Music	Children/ Childhood
Ukraine	Labor and Jobs	Demonstrations, Protests and Riots	China

Table 6: Keyword topics that are best explained (i.e. have the lowest conditional perplexity) by the following schemas: Affiliation, Discourse, NLI. Broader topics, like "Inflation" which require sources from different backgrounds, favor Affiliation-based source selection, while topics integrating many different, possibly conflicting, facts, favor NLI-based selection.

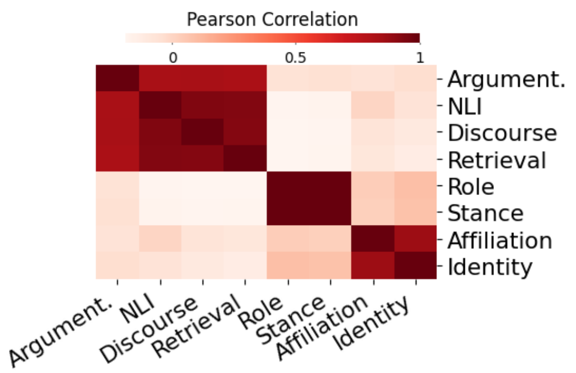


Figure 7: Pearson Correlation between conditional perplexity per document under different schemas.

has a high effect on relating different conditional perplexities, swamping the effects of the schema. This points to the caution in using perplexity as a metric; it must be well explored and appropriately baselined.

In Figure 4, we explore more why NLI and Stance are not very related. It turns out that many of the factual categories can fall in any one of the opinion-based categories. A lot of "Entailing" facts under NLI, for example, might be the the basis of "Discussion" under Stance. This points to the need to be cautious when using NLI as a stand-in for Stance, as in (Reddy et al., 2021).

In Figures 6, we compare random and kmeans perplexities across the latent dimension size. Our experiments show that indeed, we are learning important cues about perplexity. As expected, "Random" assignments have almost no affect on the perplexity of the document, while "kmeans" assignments do. Increasing the dimensionality space of Kmeans, interestingly, *decreases* the median perplexity, perhaps because the Kmeans algorithm is allowed to capture more and more meaningful semantic differences between sources.

Finally, we discuss label imbalances in our classification sets. We do not observe a strong correlation between the number of labels in the schema and the classification accuracy ($\rho = -.16$). As seen in Table 8, many schema are highly skewed, with, for example, the minority class in Argumentation ("common ground") being present in less than .22% of sources. Using our classifiers to label the news articles compiled in Section A.1, we find that the schemas all offer different information. Figure 3 shows the effect size of the χ^2 independence test, a test ranging from $(0, 1)$ which measures the relatedness of two sets of categorical variables (Cramér, 1999). The schemas are largely uncorrelated, with the highest correspondence be-

Retrieval	Role	Identity	Stance
Actors and Actresses	Inflation (Economics)	United States Economy	Midterm Elections (2022)
Fashion and Apparel	House of Representatives	Disease Rates	Presidential Election of 2020
Pop and Rock Music	Presidential Election of 2020	Real Estate and Housing (Residential)	California
Elections	United States Economy	Movies	Storming of the US Capitol (Jan, 2021)
Personal Profile	Trump, Donald J	Education (K-12)	Vaccination and Immunization
Deaths (Fatalities)	Education (K-12)	Race and Ethnicity	News and News Media
Primaries and Caucuses	Elections, House of Representatives	Ukraine	United States Economy
Politics and Government	Supreme Court (US)	Trump, Donald J	Defense and Military Forces
Regulation and Deregulation of Industry	Computers and the Internet	Presidential Election of 2020	Television

Table 7: Keyword topics that are best explained (i.e. have the lowest conditional perplexity) by the following schemas: Retrieval, Role, Identity, Stance. Political topics, like “House of Representatives” which often have a mixture of different roles, favor Role-based source selection, while polarizing topics like “Storming of the US Capitol” favor Stance.

Schema	n	H	% Maj.	% Min.
Affiliation	14	2.2	32.9	0.46
Role	4	1.0	53.3	4.61
Identity	6	1.3	52.2	0.69
Argument.	6	1.1	62.9	0.22
NLI	3	1.1	40.4	22.6
Stance	4	1.3	34.8	15.5
Discourse	8	1.9	30.0	1.09
Retrieval	10	2.0	21.4	0.05

Table 8: Description of the size of each schema (n) and the class imbalance inherent in it, shown by: Entropy (H), % Representation of the Majority class (% Maj.) and % Representation of the Minority class (% Min.).

ing $\nu = .34$ between “Identity” and “Retrieval”. We were surprised that NLI and Stance were not very related, as they have similar labelsets and have been used interchangeably (Reddy et al., 2021). This indicates that significant semantic differences exist between fact-relations and opinion-relations, resulting in different application of tags. We explore this in Appendix B.

C Article Example

Here is an article example, annotated with different schema definitions, along with a description by the journalist of why they pursued the sources they did.

> *We mined state and federal court paperwork. We went looking for [previous] stories. We called police and fire communications people to determine [events]. We found families for interviews about [the subjects'] lives.*[20]

D Further Schema Definitions

Here we provide a deeper overview of each of the schemas that we used in our work, as well as definitions that we presented to the annotators during annotation.

- **Affiliation:** Which group the source belongs to.

[20] https://www.nytimes.com/2017/01/23/insider/on-the-murder-beat-times-reporters-in-new-yorks-40th-precinct.html

Newspaper Sections	Proportion of Sources in each Category		
Arts	Individual: 0.29	Media: 0.19	Witness: 0.17
Automobiles	Corporate: 0.41	Witness: 0.17	Media: 0.11
Books	Individual: 0.26	Media: 0.19	Witness: 0.18
Business	Corporate: 0.51	Government: 0.2	Industry Group: 0.06
Dining and Wine	Witness: 0.28	Individual: 0.18	Media: 0.17
Education	Government: 0.36	Academic: 0.19	Witness: 0.1
Front Page	Government: 0.5	Political Group: 0.09	Corporate: 0.08
Health	Government: 0.33	Academic: 0.19	Corporate: 0.12
Home and Garden	Individual: 0.21	Witness: 0.19	Corporate: 0.17
Job Market	Corporate: 0.26	Individual: 0.15	Witness: 0.14
Magazine	Witness: 0.23	Media: 0.2	Individual: 0.18
Movies	Individual: 0.28	Media: 0.18	Witness: 0.18
New York and Region	Government: 0.36	Witness: 0.13	Individual: 0.12
Obituaries	Government: 0.18	Individual: 0.18	Media: 0.16
Opinion	Government: 0.43	Media: 0.14	Witness: 0.12
Real Estate	Corporate: 0.33	Government: 0.21	Individual: 0.12
Science	Academic: 0.4	Government: 0.19	Corporate: 0.1
Sports	Other Group: 0.38	Individual: 0.15	Witness: 0.14
Style	Individual: 0.23	Witness: 0.2	Corporate: 0.17
Technology	Corporate: 0.41	Government: 0.17	Academic: 0.09
The Public Editor	Media: 0.44	Individual: 0.16	Government: 0.16
Theater	Individual: 0.34	Witness: 0.18	Media: 0.14
Travel	Witness: 0.25	Corporate: 0.21	Government: 0.15
U.S.	Government: 0.44	Political Group: 0.12	Academic: 0.08
Washington	Government: 0.6	Political Group: 0.1	Media: 0.08
Week in Review	Government: 0.37	Academic: 0.11	Media: 0.1
World	Government: 0.54	Media: 0.09	Witness: 0.09

Table 9: Distribution over source-types with different *Affiliation* tags, by newspaper section.

- **Institutional:** The source belongs to a larger institution.
  1. **Government:** Any source who executes the functions of or represents a government entity. *(E.g. a politician, regulator, judge, political spokesman etc.)*
  2. **Corporate:** Any source who belongs to an organization in the private sector. *(E.g. a corporate executive, worker, etc.)*
  3. **Non-Governmental Organization (NGO):** If the source belongs to a nonprofit organization that operates independently of a government. *(E.g. a charity, think tank, non-academic research group.)*
  4. **Academic:** If the source belongs to an academic institution. Typically, these are professors or students and they serve an informational role, but they can be university administrators, provosts etc. if the story is specifically about academia.
  5. **Other Group:** If the source belongs or is acting on behalf of some group not captured by the above categories (please specify the group).
- **Individual:** The source does **NOT** belong to a larger institution.
  1. **Actor:** If the source is an individual acting on their own. *(E.g. an entrepreneur, main character, solo-acting terrorist.)*
  2. **Witness:** A source that is ancillary to events, but bears witness in either an active *(e.g. protester, voter)* or inactive *(i.e. bystander)* way.

Headline: Services failed to prevent crime
__'s voice became a preoccupation of __, who told the police that he heard her calling his name at night. ← Government, Neutral
"Psychotic Disorder," detectives wrote in their report. ← *labels:* Government, Refute
"She had a strong voice," said Carmen Martinez, 85, a neighbor. ← Witness, Neutral
Records show a string of government encounters failed to help __ as his mental health deteriorated. ← *labels:* Government, Agree
"This could have been able to be avoided," said __'s lawyer. ← *labels:* Actor, Agree

Table 10: Informational sources synthesized in a single news article[21]. Source categorizations under two different schema: affiliation and stance. Our central question: *which schema best characterizes the kinds of sources needed to tell this story?*

3. **Victim:** A source that is affected by events in the story, typically negatively.
4. **Other:** Some other individual (please specify).

- **Role:**
  1. **Participant:** A source who is either directly making decisions on behalf of the entity they are affiliated with, or taking an active role somehow in the decision-making process.
  2. **Representative:** A source who is speaking on behalf of a *Participant*.
  3. **Informational:** A source who is giving information on ongoing decisions or events in the world, but is not directly involved in them.
  4. **Other:** Some other role that we have not captured (please specify).
- **Role Status:**
  1. **Current:** A source who is currently occupying the role and affiliation.
  2. **Former:** A source who *used* to occupy the role and affiliation.
  3. **Other:** Some other status that we have not captured (please specify).

We note that **Rote Status** was a schema that we collected, but ultimately did not end up modeling.

E Example GPT Prompts

We give more examples for prompts.

E.1 Source Attribution Prompts

In Section A.1, we discuss training a GPT3.5-Turbo model with Spangher et al. (2023b)'s source attribution dataset to create more labeled datapoints, which we then distil into a BERT model. We train a batched model to save on costs. The prompt takes the following form:

```
Input:
1. <sent 1>
2. <sent 2>
3. <sent 3>...
Response:
1. <attribution 1>
2. <attribution 2>
...
```

Here is an example:

```
System Prompt:
You are a journalist's
fact-checker who identifies
sources providing information
for each sentence. The user
will show you a sentences in
an article and you'll respond
with the source of the sentences.
Consider the whole article and be
sure to answer every question.
Answer either by directly
copying text in the article
OR with "passive-voice" when
a canonical source is clearly
consulted OR "journalist" when
a direct observation is made OR
"No source" when no source is
referenced, the information is
vague, or the source is unclear.
Do not make up names, or say
anything that is not in the
article besides those phrases
above.
User Input: 1: BANGKOK | A
plane carrying key senior Laotian
government officials crashed
Saturday morning, leaving at
least four people dead, Laotian
diplomats said Saturday.
2: Killed in the crash were
two top figures in the security
apparatus of the authoritarian
```

```
Lao government: the deputy prime
minister, Douangchay Phichit,
and Thongbane Sengaphone, the
minister of public security,
according to two Lao diplomats.
3: For a Communist party that
relies on force and intimidation
to stay in power, the loss of
what were arguably the two most
powerful people in the security
apparatus was a significant blow.
4: The governor of Vientiane
province was also killed in the
crash.
5: In addition to his post
as deputy prime minister Mr.
Douangchay was defense minister
and a member of the Politburo
, the highest decision - making
body of the Communist party.
6: Mr. Thongbane, the public
security head, was feared in the
country and was said to be one of
the officials leading a crackdown
against dissent over the past
year and half.
Response:
1: Laotian diplomats
2: No Source
3: No Source
4: No Source
5: No Source
6: No Source
```

E.2 Stance-Based Prompts

In Section 3.2 we discuss the prompts we formulated to do appropriate transfer learning from the stance datasets others have annotated to our news setting. Because in Stance detection, there are usually many claims made for each hypothesis, we used batched prompts to save costs, in the following form:

```
Premise: <premise>
Claim:
1. <claim 1>
2. <claim 2>
Response: 1. <label 1>
2. <label 2>
...
```

Here is an example:

```
System Prompt: You are a
journalist's assistant who spots
opposing claims. The user will
give you a premise and 5 claims.
Respond to each one, in numbered
order from 1 to 5, with a choice
from: ['Neutral', 'Affirm',
'Discuss', 'Refute'].
Don't say anything else, and be
sure to answer each one.
User Prompt
Premise: 3-D printing will
change the world.
Claims:
1: I can see 3D printing for
prototypes, and some custom work.
However manufacturing industries
use thousands of plastics and
thousands of metal alloys...
2: Flash backwards to 1972,
Colorado, where the newly
enfranchised...
3: This is precisely the way I
feel about 3D printers...another
way to fill the world with
plastic junk that will end up
in landfills, beaches, and yes,
mountains and oceans. ...
4: I am totally terrified with
the thought of 3-D printed,
non-traceable, guns and bullets
in every thugs hands. May that
never happen. But then Hiroshima
did (bad thing)...
5: Hate to point out an obvious
solution is to tie the tax rate
to unemployment....
Response:
1: Refute
2: Neutral
3: Refute
4: Affirm
5: Neutral
```

E.3 GPT-2 Conditional Perplexity Prompts

In Section 4.1, we discuss crafting prompts for GPT2-base models in order to calculate conditional perplexity. We give the outline of our prompt. Here is an example:

```
Revelations from the artist's
autobiography threaten to cloud
```

```
her new show at the San Francisco
Museum of Modern Art.
<labels>
(1): NGO,
(2): Media,
(3): Media,
(4): Media,
(5): Corporate
<text>
(1): In a telephone interview
on Tuesday, the museumś current
director, Christopher Bedford ,
said he welcomed the opportunity
to "be very outspoken about
the museumś relationship to
antiracism" and ...
(2): Last week a Chronicle
critic denounced the museumś
decision to proceed with the
show.
(3): Its longest-serving
curator, Gary Garrels, resigned
in 2020 soon after a post quoted
him saying, "Dont́ worry, we will
definitely continue to collect
white artists."
(4): The website Hyperallergic
surfaced those comments in June .
(5): And its previous director,
Neal Benezra, apologized to
employees after removing critical
comments from an Instagram post
following the murder of George
Floyd.
(6): And the San Francisco
Museum of Modern Art has been
forced to reckon with what
employees have called structural
inequities around race.
(7): The popular Japanese artist
Yayoi Kusama, whose " Infinity
Mirror Rooms " have brought
lines around the block for one
blockbuster exhibition after
another, has...'
```

F Combining Different Schema

We show how two schema, *Role* and *Affiliation* may be naturally combined. One function of journalism is to interrogate the organizations powering our society. Thus, many sources are from

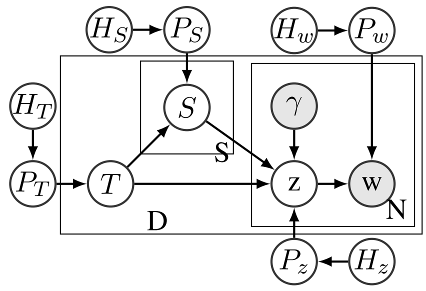


Figure 8: Plate diagram for Source Topic Model

Affiliations: *Government*, *Corporations*, *Universities*, *Non-Governmental Organizations* (NGOs). And, they have different *Roles* in these places. Journalists first seek to quote *decision-makers* or *participants*: presidents, CEOs, or senators. Sometimes decision-makers only comment though *Representatives*: advisors, lawyers or spokespeople. These sources all typically provide knowledge of the inner-workings of an organization. Broader views are often sought from *Informational* sources: experts in government or analysts in corporations; scholars in academia or researchers in NGOs. These sources usually provide broader perspectives on topics. Table 11 shows the intersection of these two schema.

G Latent Variable Models

As shown in Figure 8, our model observes a switching variable, γ and the words, w, in each document. The switching variable, γ is inferred and takes one of two values: "source word" for words that are associated with a source "background", for words that are not.

The model then infers source-type, S, document type T, and word-topic z. These variables are all categorical. All of the variables labeled $P_{\cdot}$ in the diagram represent Dirichlet *P*riors, while all of the variables labeled $H_{\cdot}$ in the diagram represent Dirichlet *H*yperpriors.

Our generative story is as follows:

For each document $d = 1, ..., D$:

1. Sample a document type $T_d \sim Cat(P_T)$
2. For each source $s = 1, ..., S_{(d,n)}$ in document:
   (a) Sample source-type $S_s \sim Cat(P_S^{(T_d)})$
3. For each word $w = 1, ...N_w$ in document:
   (a) If $\gamma_{d,w}$ = "source word", sample word-topic $z_{d,w} \sim Cat(P_z^{(S_s)})$

			Role		
			Decision Maker	*Representative*	*Informational*
Affiliation	*Institutional*	*Government*	President, Senator...	Appointee, Advisor...	Expert, Whistle-Blower...
		Corporate	CEO, President...	Spokesman, Lawyer...	Analyst, Researcher...
		NGO	Director, Actor...	Spokesman, Lawyer...	Expert, Researcher...
		Academic	President, Actor...	Trustee, Lawyer...	Expert, Scientist...
		Group	Leader, Founder...	Member, Militia...	Casual, Bystander...
	Individ.	*Actor*	Individual...	Doctor, Lawyer...	Family, Friends...
		Witness	Voter, Protestor...	Spokesman, Poll...	Bystander...
		Victim	Individual...	Lawyer, Advocate...	Family, Friends...

Table 11: Our source ontology: describes the affiliation and roles that each source can take. A *source-type* is the concatenation of *affiliation* and *role*.

(b) If $\gamma_{d,w}$ = "background", sample word-topic $z_{d,w} \sim Cat(P_z^{(T_d)})$

(c) Sample word $w \sim Cat(z_{d,n})$

The key variables in our model, which we wish to infer, are the document type (T_d) for each document, and the source-type ($S_{(d,n)}$) for each source. It is worth noting a key difference in our model architecture: Bamman et al. (2013) assume that there is an unbounded set of mixtures over person-types. In other words, in step 2, S_s is drawn from a document-specific Dirichlet distribution, $P_S^{(d)}$. While followup work by Card et al. (2016) extends Bamman et al. (2013)'s model to ameliorate this, Card et al. (2016) do not place prior knowledge on the number of document types, and rather draw from a Chinese Restaurant Process.[22] We constraint the number of *document-types*, anticipating in later work that we will bound news-article types into a set of common archetypes, much like we did for *source-types*.

Additionally, both previous models represent documents solely as mixtures of characters. Ours, on the other hand, allows the type of a news article, T, to be determined both by the mixture of sources present in that article, and the other words in that article. For example, a *crime* article might have sources like a government official, a witness, and a victim's family member, but it might also include words like "gun", "night" and "arrest" that are not included in any of the source words.

G.1 Inference

We construct the joint probability and collapse out the Dirichlet variables: P_w, P_z, P_S, P_T to solve a Gibbs sampler. Next, we discuss the document-type, source-type, and word-topic inferences.

G.1.1 Document-Type inference

First, we sample a document-type $T_d \in 1, ..., T$ for each document:

$$p(T_d|T_{-d}, s, z, \gamma, H_T, H_S, H_Z) \propto \\ (H_{TT_d} + c_{T_d,*}^{(-d)}) \times \prod_{s=1}^{S_d} \frac{(H_{Ss} + c_{T_d,s,*,*})}{(c_{T_d,*,*,*} + SH_S)} \\ \times \prod_{j=1}^{N_T} \frac{(H_{zk} + c_{k,*,T_d,*})}{(c_{*,*,T_d,*} + KH_z)} \quad (5)$$

where the first term in the product is the probability attributed to document-type: $c_{T_d,*}^{(-d)}$ is the count of all documents with type T_d, not considering the current document d's assignment. The second term is the probability attributed to source-type in a document: the product is over all sources in document d. Whereas $c_{T_d,s,*,*}$ is the count of all sources of type s in documents of type T_d, and $c_{T_d,*,*,*}$ is the count of all sources of any time in documents of type T_d. The third term is the probability attributed to word-topics associated with the background word: the product is over all background words in document d. Here, $c_{k,*,T_d,*}$ is the count of all words with topic k in document type T_d, and $c_{*,*,T_d,*}$ is the count of all words in documents of type T_d.

G.1.2 Source-Type Inference

Next, having assigned each document a type, T_d, we sample a source-type $S_{(d,n)} \in 1, ..., S$ for each source.

$$p(S_{(d,n)}|S_{-(d,n)}, T, z, H_T, H_s, H_z) \propto \\ (H_{SS_d} + c_{T_d,S_{(d,n)},*,*}^{-(d,n)}) \\ \times \prod_{j=1}^{N_{S_{d,n}}} \frac{(H_z + c_{z_j,*,S_{(d,n)},*,*})}{(c_{*,*,S_{(d,n)},*,*} + KH_z)} \quad (6)$$

[22] Card et al. (2016) do not make their code available for comparison.

The first term in the product is the probability attributed to the source-type: $c^{-(d,n)}_{T_d,S_{(d,n)},*,*}$ is the count of all sources of type $S_{(d,n)}$ in documents of type T_d, not considering the current source's source-type assignment. The second term in the product is the probability attributed to word-topics of words assigned to the source: the product is over all words associated with source n in document d. Here, $c_{z_j,*,S_{(d,n)},*,*}$ is the count of all words with topic z_j and source-type $S_{(d,n)}$, and $c_{*,*,S_{(d,n)},*,*}$ is the count of all words associated with source-type $S_{(d,n)}$.

G.1.3 Word-topic Inference

Finally, having assigned each document a document-type and source a source-type, we sample word-topics. For word i, j, if it is associated with sources ($\gamma_{i,j} =$ Source Word), we sample:

$$p(z_{(i,j)}|z^{-(i,j)}, S, T, w, \gamma, H_w, H_S, H_T, H_z) \propto \\ (c^{-(i,j)}_{z_{i,j},*,S_d,*,*} + H_{zz_{i,j}}) \times \frac{c^{-(i,j)}_{z_{i,j},*,w_{i,j},*} + H_w}{c^{-(i,j)}_{z_{i,j},*,*,*} + VH_w} \tag{7}$$

The first term in the product is the word-topic probability: $c^{-(i,j)}_{z_{i,j},*,S_d,*,*}$ is the count of word-topics associated with source-type S_d, not considering the current word. The second term is the word probability: $c^{-(i,j)}_{z_{i,j},*,w_{i,j},*}$ is the count of words of type $w_{i,j}$ associated with word-topic $z_{i,j}$, and $c^{-(i,j)}_{z_{i,j},*,*,*}$ is the count of all words associated with word-topic $z_{i,j}$.

For word i, j, if it is associated with background word-topic ($\gamma_{i,j}$ = Background), we sample:

$$p(z_{(i,j)}|z^{-(i,j)}, S, T, w, \gamma, H_w, H_S, H_T, H_z) \propto \\ (c^{-(i,j)}_{z_{i,j},*,T_d,*} + H_{zz_{i,j}}) \times \frac{c^{-(i,j)}_{z_{i,j},*,w_{i,j},*} + H_w}{c^{-(i,j)}_{z_{i,j},*,*,*} + VH_w} \tag{8}$$

Equation 8 is nearly identical to 7, with the exception of the first term, the word-topic probability term, where $c^{-(i,j)}_{z_{i,j},*,T_d,*}$ refers to the count of words associated with word-topic $z_{i,j}$ in document-type T_d, not considering the current word. The second term, the word probability term, is identical.