"Don't quote me on that": Finding Mixtures of Sources in News Articles

ALEXANDER SPANGHER, NANYUN PENG, JONATHAN MAY, and EMILIO FERRARA, Information Sciences Institue, University of Southern California

ACM Reference Format:
Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2020. "Don't quote me on that": Finding Mixtures of Sources in News Articles. 1, 1 (September 2020), 6 pages. https://doi.org/10.1145/nnnnnn.nnnnnnn

1 INTRODUCTION

A dominant form of information published in news articles is derived from people, called *sources*. Through direct conversation, statements or written correspondence, journalists interact with sources to obtain quotations that inform news consumers' understanding of current events, facilitate the voting decisions we make in our democracy and hold powerful individuals accountable.

Computational journalism is an emerging discipline that seeks to apply computational techniques to enhance journalists' ability to seek new information [4]. Researchers in this field attempt to build models for machine-in-the-loop systems to aid journalistic inquiry and produce more robust news coverage. Here we introduce a taxonomy and a model for one of many generative processes in newsmaking: the inclusion of *named sources*, or named-entities associated with quotations, in news articles. Our motivation is a first-step towards tools that can help journalists identify gaps in pieces, find sources more quickly and produce more robust coverage.

1.1 Contributions of this work

Our research advances three distinct directions:

(1) We propose a problem definition for the analysis of named sources, as well as an ontology of named sources that categorizes sources into different *source-types* by their *affiliation* and *role* (*cf.,* Section 2).
(2) We implement a probabilistic graphical model that captures the mixture of source-types in each news article as a function of news-article type and the words that are associated with each source (*cf.,* Section 4). We evaluate our model with expert annotators and show a predictive accuracy of 80%, well above existing baselines (*cf.,* Section 5).
(3) We present analytical insights that (1) lay the groundwork for future studies aimed at helping journalists find sources more quickly; (2) show how our model can be used to analyze trends in news. For instance, we find that between 1999-2002 in *New York Times* front page articles,

Authors' address: Alexander Spangher, spangher@isi.edu; Nanyun Peng, npeng@isi.edu; Jonathan May, jonmay@isi.edu; Emilio Ferrara, ferrarae@isi.edu Information Sciences Institue, University of Southern California.

XXXX-XXXX/2020/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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

high-level government officials were quoted less frequently while academic experts were quoted more.

2 PROBLEM STATEMENT

We seek to model news stories as mixtures of sources, where each source is labeled by a *source-type*. The *source-type* is defined as a concatenation of a source's identified *affiliation* and *role*. A source's *affiliation* refers to the kind of organization a source belongs to while *role* represents their role *in that organization*.[1] Each news article is defined by a *document-type*, which influences the mixture of *source-types* present in the article. We next present the source ontology, shown in Table 1, based around the notion of *affiliation* and *role*. We leave to future work a similar explication of news-article types – in this work, we model them as latent variables to be inferred (*cf*. Section 4).

2.1 Source Ontology

One function of journalism is to interrogate the organizations powering our society. Thus, many sources are from Institutions: *Government*, *Corporations*, *Universities*, *Non-Governmental Organizations* (NGOs). Journalists first seek to quote *decision-makers*: 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.

A different category of sources do not belong to formal organizations. They are Individuals: *Actors*, *Victims* and *Witnesses*. These sources differ based on how active a role they take in the events around them: actors affect events around them, while witnesses and victims are neutral or affected by the events around them. Often, these sources cannot be directly reached and journalists seek proxies: family members, lawyers, doctors or spokespeople.

2.2 Source Identification and Representation

We define *sources*, formally, as PERSON named-entities that are quoted.We represent documents as combinations of *source-words* as well as *background-words*. *Source-words* are all words in the first sentence that mentions a source, as these usually contain identifying information (e.g.: "Mick

[1]We emphasize that the focus of the *role* category is on the source's role in the organization, not the story itself.

Mulvaney, the president's chief of staff.") as well as all sentences that contain a quote by that source (e.g.: "'Get over it', said Mulvaney."). Background words are all other words.

3 RELATED WORK

This work focuses on people quoted in news articles and is part of a broader field of character-based analysis in text.

Persona Modeling Our work builds off [1] – which was extended by [3]. Authors model characters in text as mixtures of topics, which are themselves influenced by latent "personas." Both their work and ours seek to learn latent character-types. There are key differences between our work and theirs: [1] view their characters as *doers*. Their characters are villains or heroes who have substantive roles in a plotline. As such, the text associated with characters is *verb* focused. Our work, in contrast, views characters as information providers, not necessarily active participants in the story.[2] Thus, we build a different set of rules for associating text with characters. Additionally, there are differences in model structure which we will discuss in Section 4.

Computational Journalism This work also falls into the field of *Computational Journalism*, which seeks to apply computational techniques to enhance the news environment. Within this broad field, our work aims at aiding journalists by leading towards machine-in-the-loop systems. Overview, for instance, is a tool that helps investigative journalists comb through large corpora [2]. Work by [6] aims to surface social media posts that are *unique* and *relevant*. Our work is especially relevant in this vein. We envision characterizations of source types being combined with knowledge graphs to lead to similar tools for finding relevant sources, and suggesting sources to add to a story.

4 MODEL

Our model observes a switching variable, γ and the words, w, in each document.[3] The model then infers source-type, S, document type T, and word-topic z.

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 $z_{d,w} \sim Cat(P_z^{(S_s)})$
  (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: [1] 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 [3] extends [1]'s model to ameliorate this, 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.

[2]Our characters are primarily associated with a small set of relatively uninteresting speaking verbs: "say," "explain," "according to"

[3]The switching variable, γ is observed according to rules defined in Section 2.2 and takes one of two values: "source word" for words that are associated with a source "background", for words that are not.

5 DATA AND EXPERIMENTS

We use the *New York Times* Annotated Corpus[4], which contains 1.8 million articles published during 1987–2007, the date of publication and newspaper page of the article. We take all articles that appeared on the front-page (A1) of the *New York Times* on Monday-Friday, with at least one source. This results in approx. $25,000$ articles.

We run our topic model over a range of latent topics, K. We display results for $K = 25$. We specify a set of 26 *source-types* defined by our source-ontology. Our subject-matter experts manually tag $1,000$ source-types as training data (out of $125,000$ source-types total), which we use to train our topic model in a semi-supervised setting. To validate, we examine the latent source-types assigned to each source and our subject-matter experts manually check the labels assigned to $1,000$ of these sources as validation data.

We have an overall accuracy-rate of 79%, with an inter-annotator agreement $>$ 80% by two annotators. We compare our model against 4 baseline models, shown in Figure 2. The models are: **SM+L** is our semi-supervised source topic-model. **SM-L** is our source topic model run without labels. **PM** is [1]'s Persona topic model run on news corpora with our text-processing rules (described in Section 2)[5]. **VPM** is a vanilla version of the Persona topic model run on our news corpora with [1]'s text-processing rules. Finally, **BC** is a Spectral co-clustering approach [5].[6]

Model	VPM	BC	PM	SM-L	SM+L
Acc.	.01	.02	.08	.13	.80

Table 2. Overall accuracy on ground-truth labeled set across source-types. Our semi-supervised *Source Topic Model* (SM+L) outperforms all other models by a wide margin.

The overall accuracy of both **SM-L** and **SM+L**, as shown in Figure 2 beats the other baselines, indicating that our modeling choices provide necessary signal.

6 ANALYTICAL INSIGHTS

We show two analyses from our **SM+L** model: (1) the description of source-types, and (2) the breakdown of source-types by document-type. In the following section, source-type labels are assigned based on the source-type indices fixed to the gold labels in our training set.

Description of Source-Types We examine the breakdown of source-type over time. Figure 1 shows the count of a selected group of source-types during 1987–2008 in the *New York Times*. One startling shift is the sharp drop in *Government Decision-Makers* relative to other source-types shown. In 1999–2002, *Government Decision-Makers* went from having one of the largest presences in the press to having one of the smallest. This indicates a sharp change in the accountability of government. Finally, we examine the top three topics associated with a selection of source-types, shown in Table 3. We envision an additional computational journalism application for this work in being able to compile and categorize source-types from external knowledge bases for journalists to use.

Source-Types by Document Type Finally, we can interrogate the relationship between different document types and the source-types used in them. This direction is an active area of ongoing work: presently, we lack a collaborative understanding of the generative news-article types that newsrooms produce. Table 4 shows several interesting combinations of source-types and document-types learned by our model.

[4]`https://catalog.ldc.upenn.edu/LDC2008T19`

[5]For this run, we treat all words our rules associate with sources as *Agent* words in [1]'s schema

[6]We use scikit-learn's implementation.

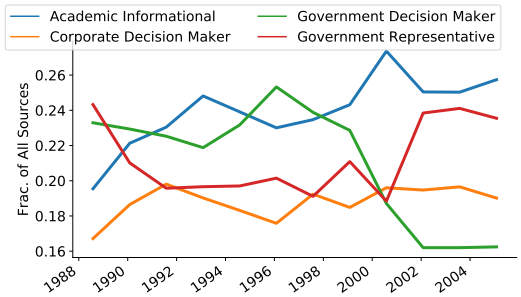

Fig. 1. Counts of Source Types used over time (18 month buckets), normalized by all sources.

Source-Type	Top Topics	Source-Type	Top Topics
academic-expert	research, child, student; like, hospital, study; care, come, time	**actor-individual**	year, include, agree; time, issue, party; make, woman, family
corporate-decision-maker	work, think, add; official, program, come; make, woman, family	**government-decision-maker**	interview, committee, member; make, election, lead; force, come, statement

Table 3. Top topics associated with selected source types. Top three topics are weighted by PMI.

Doc-Type	Top Source-Types		
1	witness-casual	academic-expert	actor-individual
3	government-decision-maker	victim-lawyer	corporate-victim
16	corporate-analyst	government-expert	academic-expert

Table 4. Topic Source-types per document-type, by PMI. Select combinations displayed.

7 CONCLUSIONS

In conclusion, we have shown a more nuanced way of thinking about the voices used in journalism. Future work holds promise both for (1) improving our categorization schemes, (2) improving our modeling approach and (3) finding downstream applications both in news production and news analysis for such an approach. Overall, we intend this work to serve as a demonstration of how the types of generative processes behind news can be quantified, and the results of such an effort.

REFERENCES

[1] 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.

[2] Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. 2014. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE transactions on visualization and computer graphics*, 20(12):2271–2280.

[3] 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.

[4] Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Communications of the ACM*, 54(10):66–71.

[5] Inderjit S Dhillon. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM.

[6] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. 2010. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *2010 IEEE Symposium on Visual Analytics Science and Technology*, pages 115–122. IEEE.