Fine-grained Structure-based News Genre Categorization

Zeyu Dai, Himanshu Taneja and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University
{zeyudai,himanshu27,huangrh}@tamu.edu

Abstract

Journalists usually organize and present the contents of a news article following a well-defined structure. In this work, we propose a new task to categorize news articles based on their content presentation structures, which is beneficial for various NLP applications. We first define a small set of news elements considering their functions (e.g., *introducing the main story or event, catching the reader's attention* and *providing details*) in a news story and their writing style (*narrative* or *expository*), and then formally define four commonly used news article structures based on their selections and organizations of news elements. We create an annotated dataset for structure-based news genre identification, and finally, we build a predictive model to assess the feasibility of this classification task using structure indicative features.

1 Introduction

There exist many guidelines for journalists in organizing and presenting contents in a news story. For example, when writing news briefs or breaking news, it is recommended to present the most newsworthy and key events first and then provide any additional details (e.g., sub-events of key events) (Po¨ ttker, 2003). While in other types of news, it is common to use a narrative hook (Myers and Wukasch, 2003) in the opening of a story that "hooks" the reader's attention so that the reader is willing to keep on reading the main story. Recognizing the overall structure of a news article can benefit many NLP tasks and applications, such as discourse parsing (Dijk, 1983), text segmentation, news summarization, information extraction and question answering system. Understanding the overall structure can also help reveal the events structure in the news. For example, the sequences of events in the news with *Narrative* structure usually follow the chronological order.

To categorize news articles based on their content organization and presentation differences, we first define a small set of news elements (section 3.1), and then formally define four commonly used news structures based on their different ways to select and organize news elements (section 3.2).

A news element is defined based on functions it plays in a news story as well as its writing style, and each news element is realized as a set of one or more consecutive paragraphs in a news article. The functions of a news element can be *introducing the main story and key events, catching the reader's attention* or *providing further details* etc.. We consider writing style in news stories as either *narrative* or *expository*. A narration section in a story usually describes surroundings, characters, and a sequence of events in a chronological order (Bal, 2009; Pentland, 1999; Smith, 2005), so that the reader can easily visualize the story with great details. An expository section is meant to provide information in a concise manner and usually answers the socalled *"5W1H"* questions: *what* are the events, *who* are involved, *where / when / why / how* did the events happen. Please see Table 1 for specific examples.

We then formally define four commonly used news structures (Wri, 2011; Jou, 2014; Po¨ ttker, 2003), *Inverted Pyramid*, *Kabob*, *Martini Glass* and *Narrative*, based on their selections and organizations of news elements. We then prepare annotation guidelines and create a dataset[1] containing around 900 news articles, where each article is annotated with its news structure and news elements. The annotated news

[1]The dataset will be made publicly available.

Functions: Introducing the Main Story and Key Event (1) *Title: Harsh Storm Batters Island off the Coast of Russia* Five days of blizzards and avalanches have paralyzed the Russian island of Sakhalin, cutting off air and sea links to the mainland, stranding dozens of motorists on highways, and burying a train, along with three railway workers, under snow drifts 10 feet deep.
Functions: Catching the Reader's Attention (2) *Title: Twitter becomes a player in customer service world* Have a problem with a business? Don't pick up the phone, or even log on to the company's Web site. Instead, Tweet it. Twitter, the 3-year-old social networking site, allows users to send 140-character text updates called "tweets" to groups of followers. "The modern-day consumer has gained considerable power and clout because of social media, and especially Twitter." says Larry Weintraub, CEO of Fanscape, a digital marketing agency. "Companies are on high alert, monitoring what people are saying about them in everyday conversations, or tweets."
Narrative writing style: (3) The accident occurred March 28 as workers digging tunnels broke through a wall into an old shaft filled with water, flooding their V-shaped shaft. Five of the workers' nine platforms were submerged. The exit out of the pit was blocked. Of the 261 miners underground that day, 108 made it to safety. The rest were trapped and feared dead.
Expository writing style: (4) Some 150 politicians, civil servants, tribal chiefs, police officers, Sunni clerics and members of Awakening Councils have been assassinated throughout Iraq since the election – bloodshed apparently aimed at heightening turmoil in the power vacuum created by more than three months without a national government.

Table 1: News Examples.

articles were sampled from four news domains, including *politics, crime, business* and *disaster* reports, for studying distributional differences of news structures across domains.

Finally, we design news structure indicative features and train a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) classifier to label each news article with one of the proposed news structures. Experimental results show that reasonable performance can be achieved for automatic structure-based news genre classification by using our structure indicative features, even though results on minority classes remain low.

2 Related Work

The previous works on automated text categorization have considered various dimensions for categorization, such as *topic* (Kazawa et al., 2005; Zhou et al., 2009), *style* (Argamon-Engelson et al., 1998) and *author* (Stamatatos et al., 2000). Although news structures have been extensively studied in linguistics and journalism (Schokkenbroek, 1999; Van Dijk, 1985; Ytreberg, 2001), there are few studies trying to categorize a news article based on its content organization structure and there is no published dataset for developing such data-driven methods. To the best of our knowledge, we are the first to consider categorizing news articles according to news structures. Our main contributions include defining news elements and news structures, creating the first dataset for news structure identification as well as identifying news structure indicative features and conducting the first computational study for structure-based news genre categorization.

The well-studied text segmentation task (Ponte and Croft, 1997; Mulbregt et al., 1998; Dharanipragada et al., 1999) has focused on segmenting a document based on topics and identifying topic transition boundaries. Labov and Waletzky (2003) conducted an in-depth analysis of 14 narrative news stories and decomposed each story into six elements[2]. In contrast, we define a small set of news elements and determine the overall structure of each news based on the selection and organization of these elements.

3 Defining and Annotating News Structures

3.1 Five News Elements

We define each news element based on its functions in a news story and its writing style[3]. Based on their characteristics, we define five types of elements below:

Standard Lede: Located at the beginning of a news article; used to introduce the main story and key events to the reader in a very concise manner; written in the expository style; e.g., the first paragraph of example (1) in Table 1.

Image Lede: Located at the beginning of a news article; unlike Standard Lede, it does not directly discuss key events of the news, instead it catches the reader's attention by providing an anecdote related

[2]The six elements are abstract, orientation, complicating action, evaluation, resolution and coda.

[3]The two characteristics are often correlated.

to key events, quoting a catchy phrase or comment, or reporting an impressive fact or statistics (Jou, 2014); written in either narrative or expository style; e.g., the first paragraph of example (2) in Table 1.

Synopsis: Preceded by an Image Lede, the main purpose of Synopsis is to summarize the main story, inform the reader about key events and acts as a bridge between the Image Lede and the rest of the story; written in expository style; e.g., the second and third paragraphs of example (2) in Table 1.

Narration: Provides great details and often indicates the presence of a set of chronologically ordered events (Bal, 2009; Mani, 2012); written in narrative style; e.g., the example (3) in Table 1.

Body Section: Provides additional details and supplementary information about key events; written in expository style. Essentially, an element that does not belong to any of the above categories is annotated as a Body Section.

3.2 Four News Structures

News Structure	Inverted Pyramid	Martini Glass	Kabob	Narrative
First Element	Standard Lede	Standard Lede	Image Lede	Image Lede*
Second Element	Body Section	Body Section*	Synopsis	Narration
Third Element		Narration	Body Section	

Table 2: Element arrangement of each news article structure. * means this element is optional.

We distinguish four news structures based on their selections and organizations of news elements. Table 2 summarizes organization patterns of news elements for each news structure.

Inverted Pyramid (IP): *Inverted Pyramid* (Po¨ ttker, 2003) as a news article structure has been widely used by newspapers since the beginning of the 20th century. In this news structure, contents are presented in the descending order of importance and relevance (Scanlan, 2003). It means that key events will be placed first, and additional details related to key events will be discussed later. Naturally, this structure can be represented as a *Standard Lede* followed by a *Body Section* as shown in Table 2.

Martini Glass (MG): Relied on a specific narrative chronology, *Martini Glass* (Wri, 2011; Jou, 2014) begins by presenting a summary of a story following the *Inverted Pyramid* structure, and then transitions into a detailed chronological elaboration of the story. This structure is better suited for stories that rely on a specific narrative chronology. Therefore, different from the *Inverted Pyramid* structure, a *Narration* element is included in the *Martini Glass* structure as well.

Kabob (Kab): In the *Kabob* (Wri, 2011; Jou, 2014) structure, a news story usually begins with an anecdote to catch the reader's attention, then introduces the main story and key events, and finally broadens into a general discussion with more details. Therefore, the *Kabob* structure starts with an *Image Lede*, and then uses a *Synopsis* as a transition followed by a *Body Section*.

Narrative (Nar): A narrative news story captivates the reader by presenting a chronologically ordered sequence of events with a greater amount of details than usual news. We label an article as *Narrative* if the majority paragraphs form a single *Narration* element with an optional preceding *Image Lede*.

Based on above definition, we can see that only the *Inverted Pyramid* and *Martini Glass* structures place key events of a news story at the beginning paragraphs; and only the *Martini Glass* and *Narrative* structures contain an Narration element written in narrative style. These commonalities and differences provide insights when designing features for categorizing news based on their structures.

3.3 Dataset Creation

To understand distributional differences of news structures across domains, we randomly sampled 250 documents for each of the four news domains, including *politics, crimes, business* and *disasters*, from the New York Times section of the Gigaword corpus (Robert Parker and Maeda, 2011) by matching news documents with pre-defined domain keywords[4]. Due to ambiguities of domain keywords, not every document is relevant to its deemed domain. Therefore, we manually checked the title of each document and cleaned the dataset by removing unrelated documents from each domain, in total, 147 documents

[4]We will list keywords in appendix.

were removed. In addition, we shifted 96 of the remaining news articles across domains. After cleaning, the dataset contains 853 news articles in total that span over four news domains.

We trained two annotators to annotate the dataset. For each document, annotators were asked to read the whole document and determine if it has one of the four news structures we defined, and then divide the document into segments corresponding to news elements. First, the two annotators annotated the same 170 documents for measuring annotation inter-agreements. Then, each annotator was asked to annotate half of the remaining documents. The two annotators achieved a Cohen's κ inter-agreement score of 67% in identifying the news structure type of each document and agreed on news element segmentations[5] for 61% of times.

3.4 Dataset Statistics

News Domains	Inverted Pyramid	Martini Glass	Kabob	Narrative	Other
Politics	154	17	53	10	6
Crime	113	12	61	32	8
Business	121	3	81	16	7
Disaster	94	5	42	18	0
Total	482	37	237	76	21

Table 3: News article structures distribution.

Table 3 shows the distribution of news structures in each domains and the overall distribution of news structures in our dataset. We can see that most of the annotated articles manifest one of the four news article structures we defined and the distribution of news structures is heavily imbalanced. As expected, the *Inverted Pyramid* is the dominant news article structure across the four domains, while there are the least number of news articles in the *Martini Glass* structure, mostly in the domains of *politics* and *crime*. Furthermore, depending on news domains, certain types of news article structures are more common. For example, there are more *crime* reports written in the *Narrative* structure compared with other news domains, while there are more *business* news articles in the structure of *Kabob*.

4 Automatic Structure-based News Genre Classification

We randomly selected 53 documents as the development set and trained a multi-class classifier using the remaining 800 documents with 10-fold cross-validation for predicting the news structure type of each news article. We use the implementation of the SVM model in LIBSVM (Chang and Lin, 2011) library with default settings and tuned hyper-parameters using the development set.

4.1 The Feature Set

N-gram Features: As basic features, we consider both unigrams and bigrams (Brown et al., 1992). Both were widely used in text classification tasks.

Writing Style Features: As we discussed in section 3.2, only the *Martini Glass* and *Narrative* structures include an Narration element, we therefore create two sets of features for recognizing narrative writing style. First, we create features for grammar production rules and we use the frequency of each syntactic production rule[6] (e.g., S $\rightarrow$ NP VP) extracted from constituency-based parse trees[7] as a feature. Second, we create a feature for each semantic category in LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2015) dictionary and the feature value is the occurrences of all words in that category. These LIWC features capture presences of certain types of words, such as words denoting

[5]We only count news elements that were annotated with exactly the same paragraph boundaries and the same news element type.

[6]Note that the bottom level syntactic production rules have the form of POS tag $\rightarrow$ WORD and contain a lexical word, which made these rules dependent on specific contexts. Therefore, we exclude these bottom level production rules to obtain more general features.

[7]We used Stanford CoreNLP (Manning et al., 2014) to generate constituency-based parse trees for each sentence.

relativity (e.g., motion, time, space), which were reported effective for detecting narrative stories (Yao and Huang, 2018).

Key Event Placement (KEP) Features: Note that only the *Inverted Pyramid* and *Martini Glass* structures start with a Standard Lede, which introduces key events directly and may repeat key events and associated event attributes (e.g., character, time and location) that were mentioned in the title as well. Therefore, we design a simple feature representing the number of words in overlap[8] between the first paragraph and news title.

4.2 Experimental Results

Feature Sets	IP	MG	Kab	Nar	Macro	Micro
Unigrams	71.6/85.1/77.8	0/0/0	51.3/43.4/47.1	53.2/35.2/42.4	44.0/40.9/41.8	65/65/65
Bigrams	71.6/87.1/78.6	0/0/0	53.6/44.3/48.5	52.5/29.6/37.8	44.4/40.3/41.2	66/66/66
Unigrams + Bigrams	72.5/87.5/79.3	0/0/0	56.5/45.2/50.3	58.0/40.8/47.9	46.7/43.4/44.4	67/67/67
+ Writing Style	73.4/85.8/79.1	37.5/9.1/14.6	56.3/48.4/52.1	62.0/43.7/**51.2**	57.3/46.7/49.3	68/68/68
+ KEP Features	74.7/88.4/81.0	0/0/0	56.4/48.0/51.8	55.8/40.8/47.2	46.7/44.3/45.0	69/69/69
+ Both	76.0/88.2/**81.7**	44.4/12.2/**19.1**	60.1/52.5/**56.0**	60.0/42.3/49.6	60.2/48.8/**51.6**	71/71/**71**

Table 4: 10-fold cross-validation classification results. Each cell shows Precision/Recall/F1 score.

Table 4 shows the experimental results using different groups of features. Using N-gram features only achieves good performance for recognizing the *Inverted Pyramid* structure. Added the writing style features on top of N-gram features significantly improves the classification performance on the *Martini Glass* and *Narrative* structures which contain a Narration element. Adding the KEP features further helps to identify three news article structures except the *Narrative* category. Note that the classification performance on the *Martini Glass* structure is poor, mainly because it is a minority class and not sufficiently represented in our dataset. We conclude that SVM model using both lexical features and our designed structure indicative features can achieve reasonable performance for predicting news article structure type.

5 Conclusion

We conducted the first study on fine-grained structure-based news genre categorization by defining a small set of general news elements and formally defining four commonly used news article structures. We created the first dataset of news articles annotated with both news structures and news elements. Finally, we conducted the initial experiments and showed the feasibility of automatic news genre categorization. Future work may include investigating the structure of event story within different news structure type.

Appendix

Here is the full list of domain keywords we used to sample news documents in Section 3.3:

Politics: [government, president, congress, white house, senate, Republican, GOP, Democratic, Tea Party, foreign minister, cabinet ministers];

Crime: [assasinate, arrest, bomb, murder, kidnap, robbery, manhunt, al qaeda, charged with assault, charged with battery];

Business: [merger, investor, stock, market, shareholders, hedge fund, banker, bankruptcy];

Disaster: [disaster management, weather warn, severe weather, x.x magnitude, wind speed, rescue team, volcano erupt, earthquake, oil spill].

References

Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4.

Mieke Bal. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

[8]Stop words were removed.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

S Dharanipragada, M Franz, Jeffrey S McCarley, S Roukos, and Todd Ward. 1999. Story segmentation and topic detection in the broadcast news domain. In *Proceedings of the DARPA Broadcast News Workshop*, pages 65–68. Herndon: National Institute of Standar and Technology.

Teun A Dijk. 1983. Discourse analysis: Its development and application to the structure of news. *Journal of communication*, 33(2):20–43.

2014. Journalism story structure. `http://journalism-education.cubreporters.org/2010/08/journalism-story-structure.html`.

Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda. 2005. Maximal margin labeling for multi-topic text categorization. In *Advances in neural information processing systems*, pages 649–656.

William Labov and Joshua Waletzky. 2003. *Narrative analysis: Oral versions of personal experience*. University of Washington Press.

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, 5(3):1–142.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Paul van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. 1998. Text segmentation and topic tracking on broadcast news via a hidden markov model approach. In *Fifth International Conference on Spoken Language Processing*.

Jack Myers and Don C Wukasch. 2003. *Dictionary of poetic terms*. University of North Texas Press.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Brian T Pentland. 1999. Building process theory with narrative: From description to explanation. *Academy of management Review*, 24(4):711–724.

Horst Po¨ ttker. 2003. News and its communicative quality: The inverted pyramidwhen and why did it appear? *Journalism Studies*, 4(4):501–511.

Jay M Ponte and W Bruce Croft. 1997. Text segmentation by topic. In *International Conference on Theory and Practice of Digital Libraries*, pages 113–125. Springer.

Junbo Kong Ke Chen Robert Parker, David Graff and Kazuaki Maeda. 2011. English gigaword fifth edition.

Chip Scanlan. 2003. Writing from the top down: Pros and cons of the inverted pyramid. `https://www.poynter.org/news/writing-top-down-pros-and-cons-inverted-pyramid`.

Christina Schokkenbroek. 1999. News stories: structure, time and evaluation. *Time & Society*, 8(1):59–98.

Carlota S Smith. 2005. Aspectual entities and tense in discourse. In *Aspectual inquiries*, pages 223–237. Springer.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495.

Teun A Van Dijk. 1985. Structures of news in the press. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, 10:69.

2011. Writing the article: Leads, quotes, and organization. `https://springfieldnews.wikispaces.com/Writing+the+Article+--+Leads,+Quotes,+and+Organization`.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

Espen Ytreberg. 2001. Moving out of the inverted pyramid: narratives and descriptions in television news. *Journalism Studies*, 2(3):357–371.

Shibin Zhou, Kan Li, and Yushu Liu. 2009. Text categorization based on topic model. *International Journal of Computational Intelligence Systems*, 2(4):398–409.