Towards a Cross-article Narrative Comparison of News

Martino Mensio
martino.mensio@open.ac.uk

Harith Alani
h.alani@open.ac.uk

Alistair Willis
alistair.willis@open.ac.uk

The Open University, United Kingdom

Abstract

In the world of public misinformation, there are many cases where the information is not false or fabricated, but rather has been manipulated using more subtle techniques such as word replacements, selection of details, omissions and argument distortion. These techniques can have the effect of influencing the reader's frame of mind towards the events reported. We currently lack the necessary tools to uncover such manipulations automatically. In this position paper, we propose an integrated analysis framework and pipeline to identify various narrative signals in news articles; such as structural roles, framing, and subjectivity. By comparing these at the document level and sentence level, it will be possible to highlight differences of narrative techniques used to report the same news events.

1 Introduction

Narrative analysis refers to the processing of a piece of text to understand and characterise its structure [Rie93]. Such an analysis could help to distinguish between event reports based on their narrative structure. These are usually reflected through linguistic signals that can be more or less explicit, such as emphasising certain aspects, changing the order in which certain information is presented, or using specific terminology to impose or stress a certain opinion.

In the specific case of news articles, their narrative structure usually follows a complex non-chronological sequence, which tends to differ from other kinds of narrative that proceed more linearly [ZZBBN19]. It is a choice that is made to "*get a good story*" [Bel05], and can be exploited to emphasise or introduce non-objective statements or causality relationships between events [Dah10].

To avoid being manipulated, one solution suggested in the literature is to gather information from multiple sources [ABS14, GAR97], and to cross-compare them in order to get a broader view of the event. The same information, for example, may be presented by some sources and omitted by others, or the sequence of events be presented differently to emphasise different aspects. Therefore, we believe that readers should be made more aware of the narrative and framing embedded in the piece of news they are consuming, and how they compare with those in other articles reporting the same event. Currently, there are hardly any automated tools that offer such functionality: the best readers can do is to use news aggregators that show articles grouped by events, but they have to do such comparison on their own.

In this position paper, we suggest a framework to automatically highlight the differences in how the same story is presented by different articles, by cross-comparing their narratives. To this end, the contributions of this paper are: *i)* integrating several signals characterising the narrative of news; *ii)* presenting a processing pipeline to link together similar articles at the document and sentence level, integrating the signals identified; and *iii)* introducing a set of cross-article signals that aim to highlight the difference of narrative techniques applied.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

2 Related work

In this section, we provide an overview of previous studies in two areas of research. First, the investigation on relationships between news articles which aims to find documents that cover the same information. Second, the detection of narrative linguistic signals, which investigates and characterises several aspects of structure, framing, and subjectivity. For both of them, we gather a set of techniques that enable our approach described in the next Section 3.

2.1 Relationships between news articles

There are different possible types of relationships between news articles, such as similarity (covering the same information), referencing (one is citing another one), and temporal proximity. They can be performed at the document level (e.g., the whole article is similar to another one) or at the sentence level (e.g., the same sentence is corroborated by a sentence in another article [BMTH18]) or even at the paragraph level. Since we are interested in finding articles discussing the same information, we focus on similarity relationships. Other relationships could add interesting features, such as the order of publication which would help to identify which of the articles might have taken inspiration from the other. For the time being, we focus on studying and understanding the role of similarity.

At the article-level, there is a wide variety of work that investigates article clustering, and the methods mostly used are Latent Dirichlet Allocation (LDA) or document embedding. LDA [BNJ03] is the most used technique for topic modelling, as it allows the discovery of topics and to group articles accordingly using word distributions. Another technique for grouping articles together is to compute a similarity measure (e.g., cosine similarity) between numeric representations of the documents (TF-IDF [Jon72] or Language Models [DCLT18, CYyK+18, YDY+19]). We plan to study these models in order to select the one that can efficiently discriminate articles that talk about the same events, even if they use different linguistics, from articles that may use the same subset of words but talk about different events.

Furthermore, there are works that not only link the articles at a document level, but also investigate in more detail the connections between sentences. In one recent work [BMTH18], groups of similar articles are found, then broken down to pieces of information and analysed to find if these details are *corroborated* (occurring in multiple documents) or *omitted* (occurring in other documents of the same group, but not the current one). We aim to use this idea of applying similarity to both article-level and sentence-level, extending it even to the word-level. By doing so, not only we might be able to recognise which sentences appear in multiple documents (with different degrees of similarity) but also we would be able to identify the specific words that have been changed.

However, this set of approaches are limited to bringing to the attention of the reader the linked information pieces with a measure of similarity, without characterising the differences. The reader would then need to evaluate the differences in the role of the sentence, the framing that it implies and how it compares with other sentences in terms of subjectivity. Different documents may express the same set of details, but give them a different role (reporting an action, commenting, contextualising, doing a digression, identifying causes and consequences) and use different words that are semantically similar but may imply a different framing perspective. For this reason, the next subsection presents a set of narrative linguistic signals that could provide us with the missing features.

2.2 Narrative linguistic signals

There is much research on exposing the narrative using linguistic signals [ZZBBN19], with specific words that indicate the *structural role*, *framing* and *subjectivity* of the part of text they belong to. One limitation is that most of such works are applied to single articles, with little comparison between them.

On one hand, some research considers the *structural role* of a sentence in the document (e.g., is it providing some background, the main event, an evaluation). Different structural roles have been defined in the literature, such as news schema [Bel91], which identifies hierarchical categories (e.g., action, reaction, consequence, context, history), narrative structure [Bel05] (e.g., abstract, orientation, evaluation, complication, resolution), or linguistic signals [ZZBBN19, Mar00]. Such signals could be used to identify the differences between similar sentences with regards to their structural roles in the articles.

On the other hand, there is much literature on *framing*, defined as how a certain story is presented to shape mass opinion [Gof74], the addition to the underlying facts that reflects the sociocultural context and acts as an underlying force to persuade the reader. The work by [GM89] describes a set of *framing packages*, made

of *framing devices* (e.g., word choice, metaphors, catchphrases, use of contrast, quantification) and *reasoning devices* (e.g., problem definition, cause, consequence, solution, action). Additionally, the Frame Semantics Theory [Fil06] can be used to recognise lexical units of known frames. By extracting these linguistic signals, we could represent the framing behind a certain piece of text, and there exist different approaches to extract the listed features [MGB+17, GCCZ18, Asg16, STDS17].

In addition to these two characterisations, we can add other signals derived from studies on *subjectivity*. As found by recent research, in contemporary journalism the line between opinion and facts is blurring more and more [JWJ+19]. For this reason, having signals of subjectivity on the document and paragraph-level would be very useful [Liu10]. In this way, each article and each paragraph can be characterised with an indication of subjectivity.

All these features have been used in previous research, but as mentioned above, they are mainly applied to single-article analysis. Extending this kind of analysis by taking into consideration the relationships both at the article level and the sentence level would bring a big contribution by providing contrastive signals that would not come up otherwise.

3 Cross-article comparison framework

In this section, we propose a description of our comparison framework. We plan to use methods coming from both the research areas identified (document linking and linguistic signals) as a starting point. In order to do so, we propose the following processing pipeline:

- **preprocessing**: documents are retrieved, cleaned up and fragmented into paragraphs and sentences;
- **narrative features** are attached to each document, paragraph and sentence belonging to three main types: *structural role* using and highlighting the linguistic devices provided by [ZZBBN19]; *framing features* are extracted (framing and reasoning devices) finding some linguistic representatives from [GM89, Fil06]; *subjectivity* is computed, and strong word choices are highlighted [Liu10];
- **linking**: *similar articles* are found by using document-level similarity measures: in this way it would be possible to find groups of documents that describe the same events; *similar sentences and paragraphs* are found by sentence-level similarity measures, inside each group of documents: corroborated and omitted sentences are identified [BMTH18].

Figure 1 shows the result of such processing over two articles, where we have several features attached to the sentences, with similar paragraphs across the two articles linked together using a similarity measure [CYyK+18].

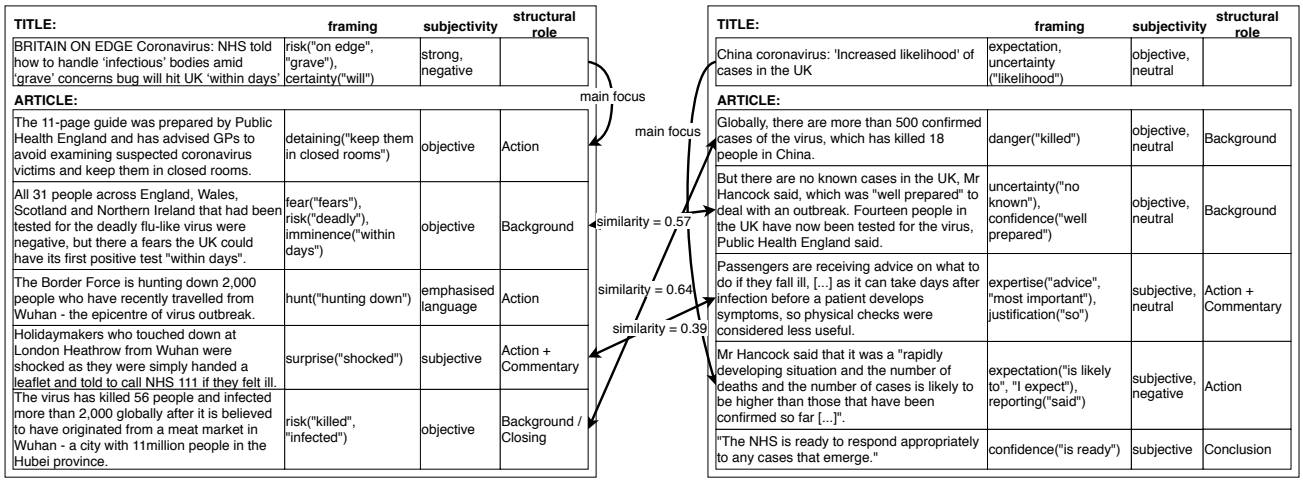

Figure 1: An example of analysis between two news articles that both talk about the risk of coronavirus spread in the UK. The first one (from The Sun) emphasises the risks from the virus, while the second article (from BBC) is more focused on presenting the UK as ready to face the problem. Each paragraph is characterised with framing, subjectivity and structural signals, and the links between the articles represent the most similar pairs of sentences.

This is is the starting point to identify the differences, with a contrastive analysis. We propose here a set of *cross-article comparative signals* that can bring the narrative analysis a step further:

- The **main focus** of the compared articles is on a different part or detail of the story: this means that while they are both describing the same broad event, they are trying to emphasise or prioritise two different aspects. This signal can be computed by looking at the most similar sentence to the article title (proxy of the emphasis), and seeing how it is represented in other documents.
- **Ordering**: the compared articles present the same details, but in a different order. Re-ordering events tends to be an efficient way of creating implicit cause-effect relationships. To do this comparison, it is sufficient to find the crossovers in the sentence-level connections.
- **Selection of details**: One article is *omitting* certain details that have been reported by other articles, or is describing events that are *corroborated* by other sources, or has *unique parts* that do not occur in other articles [BMTH18]. In addition to seeing which parts are selected or omitted, the narrative analysis can help us to find some insights about them (e.g., the article is omitting subjective statements reported by others, or is describing a background event that others did not include).
- The articles are **framing** the narrative in different ways from each other. This manifests through comparing linked sentences to observe the differences in terms of framing features: the considered articles are describing the same events but with different framing and reasoning. One concrete example is the usage of *causality*: one article may contain causality signposting between a pair of sentences that is absent elsewhere. Or as another example, the usage of *specific words* can reveal a specific framing: talking about the same detail or entity, the usage of verbs or adjectives may change. For detecting such peculiarities, features as Named Entities and subjectivity may be combined.
- The comparison can be also done on the **subjectivity** of the article: both at the document level (saying that this is an opinion piece, while a similar one is more factual) or at the sentence level, by interweaving this signal with the ones proposed before.

From the signals in Figure 1, we can see that the first article pushes the narrative towards `risk` and other negative frames, to sustain the idea presented in the title "Britain on Edge". The second article, even though it has a lot of information in common with the first one, is more confident on the preparedness of the National Health Service to face the virus (e.g., `confidence`, `expertise`). The extraction of these cross-article signals is the first step to finding possible cases of manipulation.

4 Evaluation

The evaluation of this framework needs to be performed at different levels. Firstly, we need to find a similarity model that performs well both at the article and sentence levels, going beyond the linguistic surface and being able to relate pieces of text that may use different terms for describing the same events. The evaluation of the similarity measure will be done at the article level using data coming from tools that aggregate articles talking about the same events, such as Google News Headlines[1] and AllSides[2] as well as research datasets such as NewsAggregator.[3] Instead, for the sentence-level similarity, user feedback will be needed to understand when and why a sentence is considered to describe the same detail while we are dealing with manipulations that can be significant.

Following that, we would also need to evaluate the whole framework with user studies to understand the relevance, quality and usefulness of the indicators proposed. Currently, and to the best of our knowledge, there are no similar approaches to the task we are addressing in this paper, and hence we are unable to establish comparisons with other baseline approaches from the literature.

5 Discussion

Much research exists that address the problem of misinformation. However, the vast majority of such research focuses on distinguishing what is true from what is false, and hence mainly applies to a small subsection of the

[1] `https://news.google.com/`
[2] `https://www.allsides.com/story/admin`
[3] `http://archive.ics.uci.edu/ml/datasets/News+Aggregator`

misinformation ecosystem.[4] There is a lack of research on identifying *misleading content*, *false connection* and *false context*. To this end, there is an immediate need for technological solutions to address such cases, where the information is manipulated in a subtle fashion, and thus cannot be easily dismissed as false. We want to reveal the differences in reporting, without declaring that one article contains true or false information, but rather to provide a tool that exposes such diversities.

In this paper we proposed a comparative approach that aims at bringing into light the differences in the narratives of news articles, using a set of cross-article narrative signals. These signals only exist when multiple documents are compared, in contrast to single-article ones that already exist. With this method, we aim to reveal the framing intentions of the writers, and making them more evident and comparable.

This analysis may be useful for empowering users to form a critical view of pieces of news they are consuming, to find missing pieces that have been omitted and to see the same information presented with a different framing by different articles and sources.

Acknowledgements

This work is partially supported by EU H2020 Project Co-Inform (grant no. 770302).

References

[ABS14] Øistein Anmarkrud, Ivar Bråten, and Helge I Strømsø. Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, 30:64–76, 2014.

[Asg16] Nabiha Asghar. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*, 2016.

[Bel91] Allan Bell. *The language of news media*. Blackwell Oxford, 1991.

[Bel05] Allan Bell. News stories as narratives. *The language of time: a reader*, page 397, 2005.

[BMTH18] Dimitrios Bountouridis, Mónica Marrero, Nava Tintarev, and Claudia Hauff. Explaining credibility in news articles using cross-referencing. In *SIGIR workshop on ExplainAble Recommendation and Search (EARS)*, 2018.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[CYyK+18] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[Dah10] Michael F Dahlstrom. The role of causality in information acceptance in narratives: An example from science communication. *Communication Research*, 37(6):857–875, 2010.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fil06] Charles J Fillmore. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400, 2006.

[GAR97] John T Guthrie, Solomon Alao, and Jennifer M Rinehart. Literacy issues in focus: Engagement in reading for young adolescents. *Journal of Adolescent & Adult Literacy*, 40(6):438–446, 1997.

[GCCZ18] Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*, 2018.

[GM89] William A Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology*, 95(1):1–37, 1989.

[4] https://firstdraftnews.org/latest/fake-news-complicated/

[Gof74] Erving Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

[Jon72] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[JWJ+19] Kavanagh Jennifer, Marcellino William, Blake Jonathan, Smith Shawn, Davenport Steven, and Tebeka Mahlet Gizaw. *News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms*. Rand Corporation, 2019.

[Liu10] Bing Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*, pages 627–666. CRC Press Book, 2010.

[Mar00] Daniel Marcu. *The theory and practice of discourse parsing and summarization*. MIT press, 2000.

[MGB+17] Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. Overview of the fire 2017 irled track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68, 2017.

[Rie93] Catherine Kohler Riessman. *Narrative analysis*, volume 30. Sage, 1993.

[STDS17] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*, 2017.

[YDY+19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

[ZZBBN19] Iqra Zahid, Hao Zhang, Frank Boons, and Riza Batista-Navarro. Towards the automatic analysis of the structure of news stories. In *Text2Story@ ECIR*, pages 71–79, 2019.