Hannah Smith. 2023. Voice and Context: Building a Corpus of Events to Assess Potential Bias in Digital News Headlines. NASKO, Vol. 9. pp. 43-50.

Hannah Smith - University of Illinois Urbana-Champaign

Voice and Context: Building a Corpus of Events to Assess Potential Bias in Digital News Headlines

Abstract

Racism, sexism, or other forms of bias may be reinforced by the delivery of a news story. This delivery refers to the grammatical structure of the story including the order in which details of a story are delivered. One tool for organizing grammatical structure – voice, which describes whether an action is active (performed by a subject on an object) or passive (performed on a subject by an object) – is especially useful for embedding impressions of a story in news headlines due to headlines' brevity. Take for example the headline "Coroner: Man shot by police had BAC of 0.469". In this headline, active voice is used to describe the victim's blood alcohol content, emphasizing the victim's active choices which may incriminate them in this situation. The use of passive voice to describe the shooting de-emphasizes the police's active choice to shoot the victim. This demonstrates how voice may be used to influence a reader's perception of responsibility in an event by emphasizing the active choices of one involved party but not the other. Implication of responsibility can create a positive or negative image of an involved party depending on the sentiment of the action carried out by the involved party, so the distribution and context of these instances of voice may prove to be significant in understanding how they are used to create an impression of a news story for the reader – especially when analyzed in a specific social context where voice may be used to support existing bias.

Introduction

Language is used to express models of reality through its semantic elements and the relations between them (Hjørland, 2007, 2008). Bender and Koller (2020) identify three parts of the linguistic model: *form* or the semantic elements and their structure and relations, *conventional meaning* or the meaning of an expression that is constant across all instances of its use, and *communicative intent* or the purpose or motivation of a speaker or author behind some specific instance of linguistic communication. They refer to understanding of some expression as an ability to extract the communicative intent from that expression through an analysis of its form and of the conventional meaning of the expression and its parts. Additionally, they posit that communicative intent is about something outside of language. Whether the communicative intent of some expression is to convey some information, give an instruction, or socialize, the expression is referencing real-life objects or concepts. In this way, we can treat communicative intent as some knowledge of the real world that is being represented through language, where the organization of this knowledge is observed in the form of the expression and meanings of its parts.

Bender and Koller clarify that communicative intent is not the same as ground truth, as the speaker or author may be misinformed or intend to mislead. For news, the communicative intent is often to convey some information, although persuasive

arguments may also be embedded in the stories. Existing research has identified ways in which form is constructed (such as through inclusion or omission of certain details or strategic word choice) in news stories to express a certain communicative intent or, in other words, evoke a certain reading of a news story from the audience (Metila, 2013; Molek-Kozakowska, 2014; Reah, 2002). These constructed readings can reflect the political dispositions, ideologies, and biases of the author (Montejo & Adriano, 2018). In this way, racism, sexism, or other forms of bias may be reinforced by the form of a news story. The presence of this bias in the headlines of news stories may be particularly problematic as headlines strongly influence readers' perceptions of news stories due to their attention-grabbing visual style and placement at the beginning of a story (Barthelson, 2002; Metila, 2013). Additionally, the sheer number of articles that are available online means that a reader will read only read the headline of most of the articles they encounter (Holmqvist et al., 2003). Because of this, the quality of a headline can have a significant impact a viewer's understanding of the story being reported, regardless of the quality of the story itself. The goal of my research is to identify potential bias in news headlines through the development and exploratory analysis of a corpus of events and supplementary linguistic data extracted from news headlines.

Voice as Form

Voice is a property of a verb which describes the relationship between the action (represented by the verb) and its participants through the order of presentation of detail. Take, for example, the headline "Coroner: Man shot by police had BAC of 0.496". Here, 3 details of an event are included: (1) a coroner made a report (2) a man was shot by a police officer, (3) the man who was shot had a high BAC (blood alcohol content). In the phrase "man shot by police", "man" is presented first before the action word "shot", and "police" is presented after the action word. The word presented before the action word – in this case, "man" – is the subject of the instance of voice. The word presented after the action word – in this case, "police" – is the object of the instance of voice. Note that subject/object status relates solely to the order in which the details are presented and not to the actions of the participants; both the actor in an event and the target of the action are capable of being the subject or object, and which is which determines the type of voice being used. In the example headline, the expression "Man shot by police" demonstrates passive voice because the actor ("police") is the object. If the order of details in the expression were changed so that the actor was the subject – "Police shot man" – the expression would demonstrate active voice.

Turner & Rommetveit (1968) found that subjects of an instance of voice were recalled faster than objects, and Bohner (2001) found that when presented with an event described using passive voice, readers tended to express beliefs that indicated that the subject was in some part responsible for the action occurring despite the subject being the target of the action and not the actor. Readers expressed these beliefs less frequently when the same event was described using active voice (Bohner, 2001). These findings indicate that participants are more memorable and are perceived as more responsible for the occurrence of an event when presented as subjects instead of objects in instances of voice,

regardless of whether the participants are the actors or targets in an event. The two versions of our example instance of voice – "Man shot by police" and "Police shot man" - can be compared to find two distinct readings of a story: one where the actions of the man are emphasized and one where the actions of the police are emphasized. The details of the event are not changed, so the conventional meanings of the two expressions should also be unchanged. Because of this, we can deduce that voice is functioning as a manifestation of form in this expression as it impacts the readings (or communicative intent) of the expression without changing the conventional meaning of the expression. Additionally, this method of constructing readings of stories may be particularly useful for news headlines as it does not require the addition of detail, only the arrangement of detail. This suits headlines as they must convey the most information in the least amount of text (Reah, 2002), so any opportunity to convey additional information without adding more details is beneficial to headline authors.

Identity as Meaning

Bender and Koller (2020) describe the conventional meaning as representing the communicative potential of some expression. The conventional meaning provides an understanding of what an expression represents outside of the specific context in which it is used – in other words, it provides an understanding of the objects or concepts represented by an expression as they exist in the real world and outside of language, thus grounding expressions in reality and connecting them with the real-world objects they reference. Competing theories surrounding the nature of conventional meaning exist in the field of linguistics; for this analysis I am focusing on Furner's (2009) descriptions of personal and social identity as a type of conventional meaning. Furner describes personal identity as a set of properties that individuates a person, and social identity similarly as a set of properties that distinguishes one group from others. These sets of properties include different facets of identity (such as age, class, nationality, etc.) which contribute to the representation of a person or group. Furner also clarifies that the identification and understanding of these facets are influenced by human intentionality and subjectivity, therefore they do not represent ground truths about these identities but rather they represent the reader's understanding of these identities, which may be based on biases, assumptions, and projections.

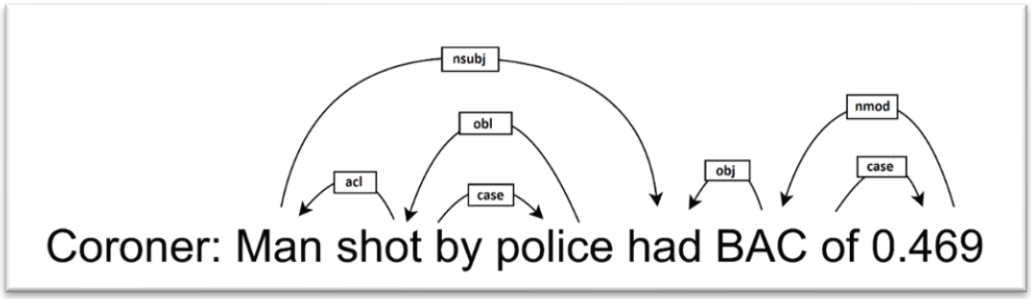
Metila (2013) identified a similar phenomenon in news headlines where reader perception of a story was influenced by extra-textual information such as cultural attitudes regarding the objects and concepts referenced in the headline. This influence can be observed in the example headline, where roles (coroner and police, which evoke authority) and details (drunkenness, which carries a negative sentiment) contribute to the perceived identity of the participants in these instances of voice (Card et al., 2016; Wallace et al., 2021). These identities may be further emphasized or downplayed using voice. For example, it may be that that presenting the man's actions before the police's actions implies a chronology which gives the impression of the police as reactionary (as opposed to instigative) which aligns with a wider public perception of the role of police. This, when combined with negative facets of the man's identity that the reader might

infer, could be interpreted as a justification of the police's actions in this event. Whether this is designed purposefully to reinforce this perception or is a subconscious manifestation of this perception, it creates a reading of the event that influences the readers' attitudes towards the event and its participants.

Methods

To extract instances of voice, I pass each headline through a natural language processing algorithm called a dependency parser to identify dependencies between words in the headline and tag them with labels. These dependency labels identify grammatical relationships between words i.e., "man" is the subject of "had" in our example headline (see Figure 1). I use Stanford's CoreNLP Stanza dependency parser (Qi et al., 2020) along with a set of pre-parsed headlines (Benton et al., 2022) to train the dependency parser. I then implement an algorithm that uses these dependency labels to extract the subject, verb, and object sets, referred to as SVO triples, and the type of voice of each SVO triple (Smith, 2023).

Figure 1: Example dependency-parsed headline



Identities in the headlines are extracted using a two-pass clustering approach where the first pass organizes nouns into categories using Stanford's Named Entity Recognizer (NER) (Finkel et al., 2005), then each category is passed through a k-means clustering algorithm to identify granular "roles". Each noun is mapped to a role and each verb is mapped to a sentiment found using Python's Natural Language Toolkit (Bird et al., 2009). I store these mappings as standoff annotation in a separate document.

Data

Table 1: Number of headlines by topic

Topic	Number of Headlines
Police shootings	1165
COVID-19	6130
Opioid crisis	973
Random	5001

Events are extracted from one of four sets of news headlines divided by topic. Table 1 shows these topics and the proportion of each topic in the dataset. The data itself consists of news headline text. The topic provides a context that is necessary for this analysis as it is within these contexts that social roles and their associated biases can be qualified and observed due to some roles manifesting differently in certain contexts (ex. "authority" manifesting as "officer" vs "doctor") or appearing with varying frequency in different contexts. My first three data sets are sourced from the University of Illinois's Cline Center Global News Index (https://clinecenter.illinois.edu/project/datascience/global-news-index) and contain headlines related to COVID-19, the opioid crisis, and police shootings. These specific topics were chosen due to their controversial nature – there may be a variety of readings that arise from sources on different sides of the controversy. Additionally, these topics all involve distinct power dynamics such as police/civilian or doctor/patient. The fourth set is sourced from the GoodNewsEveryone corpus, which is a data set of randomly sampled headlines from Reddit (Bostan et al., 2020) which acts as a control set with no specific topic.

Results

Evaluation of the SVO triple extraction algorithm produced a precision score of 0.276, a recall score of 0.364, and a false negative rate of 0.159. Common error cases for this algorithm include incorrect voice type identification and extracting partially correct triples (Smith, 2023).

Preliminary results include the word groups extracted using the two-pass clustering approach. The first pass identified ten NER groups: CAUSE_OF_DEATH, COUNTRY, CRIMINAL_CHARGE, DATE, DURATION, NATIONALITY, O, ORGANIZATION, PERSON, SET (a temporal class), STATE_OR_PROVINCE, TIME, and TITLE. Of the 511 unique nouns extracted from the headlines, 440 were placed in the O group, which seemed to be the designation for terms that did not fall into any other group. Implementing k-means clustering (MacQueen, 1967) with 10 clusters in the O group identified subgroups of interest including some which appear to be words that were not correctly identified by the NER pass including a cluster consisting mainly of names, a cluster of locations, a cluster of titles (specifically plural

titles, which did not seem to be placed in the TITLE group), and a cluster of organizations. This may suggest that layered approaches to clustering can generate coherent clusters even if some terms were not correctly identified on a first pass.

Future Work

For my next steps for this project, I plan to research computation methods to improve both the extraction of SVO triples as well as the identification of word groups via clustering. I would also like to expand my datasets both within-topic and into new topics. Because this study ultimately concerns how identity, power dynamics, and public opinion impact the telling of news stories, I would like to analyze the use of voice in headlines related to other topics that exhibit these power dynamics and levels of controversy such as news stories covering transgender healthcare, development of labor unions, and protests.

Acknowledgments

I thank Jodi Schneider for supervising this study, Halil Kilicoglu and Roxana Girju for their assistance in researching and developing this project, Sara Perez for her assistance in developing the gold set, the Information Quality Lab for feedback, and the University of Illinois's Cline Center for providing access to their data repositories.

References

Ann Turner, E., & Rommetveit, R. (1968). Focus of attention in recall of active and passive sentences. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 543–548. https://doi.org/10.1016/S0022-5371(68)80047-7

Barthelson, M. (2002). *Behaviour in Online News Reading* [Master's, Lund University]. http://lup.lub.lu.se/student-papers/record/1329001

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Benton, A., Shi, T., İrsoy, O., & Malioutov, I. (2022). Weakly supervised headline dependency parsing. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6520–6535. https://aclanthology.org/2022.findings-emnlp.487

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media. http://doi.org/10.3115/1118108.1118117

Bohner, G. (2001). Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, *40*(4), 515–529.

https://doi.org/10.1348/014466601164957

Bostan, L. A. M., Kim, E., & Klinger, R. (2020). GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1554–1566. https://aclanthology.org/2020.lrec-1.194

Card, D., Gross, J., Boydstun, A., & Smith, N. A. (2016). Analyzing framing through the casts of characters in the news. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1410–1420. https://doi.org/10.18653/v1/D16-1148

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 363–370. https://doi.org/10.3115/1219840.1219885

Furner, J. (2009). Interrogating "Identity": A Philosophical Approach to an Enduring Issue in Knowledge Organization. *Knowledge Organization*, *36*(1), 3–16. https://doi.org/10.5771/0943-7444-2009-1-3

Hjørland, B. (2007). Semantics and knowledge organization. *Annual Review of Information Science and Technology*, *41*(1), 367–405. https://doi.org/10.1002/aris.2007.1440410115

Hjørland, B. (2008). What is Knowledge Organization (KO)? *Knowledge Organization*, *35*(2–3), 86–101. https://doi.org/10.5771/0943-7444-2008-2-386

Holmqvist, K., Holsanova, J., Barthelson, M., & Lundqvist, D. (2003). Chapter 30 - Reading or scanning? A study of newspaper and net paper reading. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The Mind's Eye* (pp. 657–670). North-Holland. https://doi.org/10.1016/B978-044451020-4/50035-9

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967*, 281–297.

Metila, R. A. (2013). A discourse analysis of news headlines: Diverse framings for a hostage taking event. *Asian Journal of Social Sciences & Humanities*, *2*(2), 71–78.

Molek-Kozakowska, K. (2014). *Coercive metaphors in news headlines: A cognitivepragmatic approach*. *40*(1), [149]-173. https://doi.org/10.5817/BSE2014-1-8

Montejo, G. M., & Adriano, T. Q. (2018). A critical discourse analysis of headlines in online news portals. *Journal of Advances in Humanities and Social Sciences*, *4*(2), 70–83. https://doi.org/10.20474/jahss-4.2.2

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acldemos.14

Reah, D. (2002). *The Language of Newspapers*. Psychology Press.

Smith, H. (2023). extracting voice from headlines: towards assessing potential bias in digital news. Under review

Wallace, R., Lawlor, A., & Tolley, E. (2021). Out of an abundance of caution: COVID19 and health risk frames in Canadian news media. *Canadian Journal of Political Science*, *54*(2), 449–462. https://doi.org/10.1017/S0008423921000214