The American Local News Corpus

Ann Irvine[1] **Joshua Langfus**[1] **Chris Callison-Burch**[2]

[1] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD
[2] Computer and Information Science Department, University of Pennsylviania, Philadelphia, PA

Abstract

We present the American Local News Corpus (ALNC), containing over 4 billion words of text from $2,652$ online newspapers in the United States. Each article in the corpus is associated with a timestamp, state, and city. All 50 U.S. states and $1,924$ cities are represented. We detail our method for taking daily snapshots of thousands of local and national newspapers and present two example corpus analyses. The first explores how different sports are talked about over time and geography. The second compares per capita murder rates with news coverage of murders across the 50 states. The ALNC is about the same size as the Gigaword corpus and is growing continuously. Version 1.0 is available for research use.

Keywords: Corpus Creation, Newspapers, American English

1. Motivation

Gun violence has plagued the United States for decades. In 1996, the U.S. congress effectively eliminated the Center for Disease Control's (CDC) ability to gather data on instances of gun violence. Fortunately, the CDC's power to track gun violence was recently reinstated, and the National Academies released a report in 2013 outlining a public health agenda for reducing gun violence (Leshner et al., 2013). Among the recommendations in the report is a call for more useable data, "including community-level data and data concerning the circumstances of firearm deaths, types of weapons used, victim-offender relationships, role of substance use, and geographic location of injury" (Leshner et al., 2013, pp. 24).

We are interested in mining the text of news articles for data on when, where, and what type of gun violence occurs throughout the U.S. Using news articles to gather such data has the advantages of a very short time lag and, frequently, in-depth reporting. We have collected the ALNC for the purposes of this ongoing research project, but its potential applications extend far beyond researching gun violence.

2. Corpus Collection

The Newspaper Map website[1] contains a comprehensive database of nearly all of the newspapers in the world. The creators of the website provided us with a list of the newspapers which are published in the United States, along with their URLs and publication locations (city and state). We have compiled the ALNC by repeatedly crawling each news website.

We began our web crawls for each newspaper in August of 2013 and have collected articles from each every other day since. In this work, we present the ALNC Version 1.0, which is the result of six months of web crawls. As the crawls continue, we plan to release subsequent versions of the corpus. Although the corpus already contains an enormous amount of data, its size and value continues to increase.

Our raw crawled data consists of HTML pages. We have removed HTML tags and most other boilerplate text to the degree that it is possible to automatically identify given the wide variety of website structures. Because existing tools typically assume that webpages are well-formed and many in our corpus are not, we developed our own scripts (consisting primarily of a series of regular expressions) for cleaning the data, which yield consistently clean article content. Similarly, because article headers and metadata across newspaper websites is inconsistent, we have assigned an approximate timestamp to each article corresponding to the date of the web crawl. In some cases, an article's publication date may be earlier than the date that the crawler accessed the page. We treat multi-page articles as separate documents. However, it may be possible to automatically cluster these multi-document articles using their original URLs, which are included in the corpus. We use off-the-shelf tools[2] to sentence segment, tokenize, and lowercase each article.

Throughout crawling and data cleaning, we make several efforts to avoid duplicate data. First, in crawling, we only extract the content of articles with URLs which have not been extracted in the past. Second, we ignore articles with a large (>90% of lines) amount of overlap with previously extracted pages from the same website. This removes content duplication due to articles appearing under multiple URLs. Finally, within a single website, we eliminate individual lines of text that have already appeared twice in previous articles. This effort eliminates content that appears redundantly in many articles, such as suggestions to follow a writer's twitter feed or to share the story via email. One remaining source of content duplication is, for example, Associated Press (AP) stories that are published by multiple news agencies. We have deliberately chosen *not* to remove this type of duplication because information about which news agencies disseminate which news stories may be of great value to researchers using the corpus. After all preprocessing and deduplication, the ALNC Version 1.0 contains 4,015,186,410, or just over 4 billion, words of text.

Figure 1 shows the number of words of text in the ALNC from newspapers published in each of the 48 continental

[1] `http://newspapermap.com/`

[2] We use the preprocessing scripts available in the Moses toolkit, `http://www.statmt.org/moses/`.

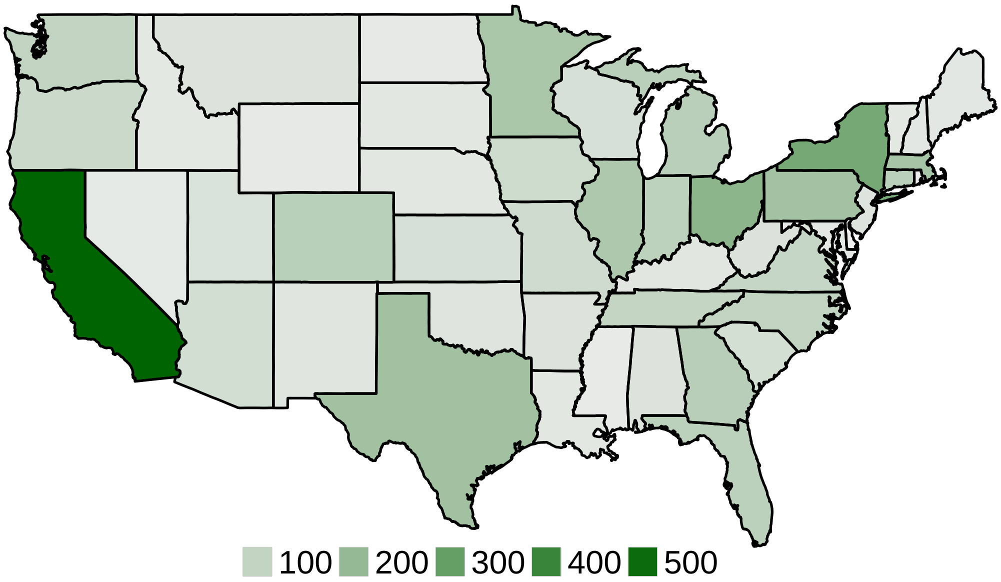

Figure 1: Millions of words of newspaper data contained in the ALNC, by state.

states (Alaska and Hawaii are not shown). Not surprisingly, the corpus contains the most data from newspapers published in populous states such as California, New York, Texas, and Ohio. It contains the smallest amount of data, about 1 million words, from Delaware.

3. Prior Work

Other corpora of English newswire text have been released in the past. The North American News Text Corpus[3] (NANTC), for example, contains the text of articles published between 1994 and 1998 in several major American newspapers: the LA Times, the Washington Post, the New York Times, Reuters News Service, the Wall Street Journal, and the Associated Press (Graff, 1995). This NANTC, including its supplement, contains about 600 million words. The fifth and latest version of the English Gigaword Corpus[4] is another English newswire corpus (Parker et al., 2011). It contains articles from the Associated Press and New York Times as well as English articles from the Agence France Press and the Xinhua News Agency. The Gigaword Corpus is about the same size of the ALNC, containing approximately 4 billion words.

The NANTC and the Gigaword Corpus only contain articles from *major news outlets*. In contrast, our ALNC corpus contains articles from a wide variety of news sources, including small weekly and monthly publications from rural regions of the United States. This distinction makes it a very different type of resource. For example, it contains many news articles covering major events, but articles have a variety of regional perspectives and political biases. Additionally, the ALNC contains coverage of local events that are not newsworthy enough for national publications, including, for example, local high school sporting events, town and county elections, and crime reports.

Version 1.0 of the ALNC is about five times the size of the NANTC and about the same size as the Gigaword corpus. The ALNC is continually growing at a rate of about 600 million words per month. We expect that a future version of the ALNC containing a year's worth of data will consist of over 7 billion words.

A variety of prior work in natural language processing has made extensive use of the NANTC and Gigaword Corpus. McClosky et al. (2006), for example, use the NANTC to adapt a parser to a new domain. Suzuki and Isozaki (2008) use the Gigaword Corpus to improve performance on a variety of basic NLP tasks, including part-of-speech tagging, chunking, and named entity recognition. Recently, Ganitkevitch et al. (2013) used the Gigaword Corpus to score a very large corpus of paraphrases for monolingual distributional similarity.

4. Example Corpus Analyses

In addition to releasing the ALNC, we have done some data analysis on the corpus to shed light on its coverage and the type of research that it enables.

First, we present an example about the relatively popularity of different sports in the United States, across locations and across time. Figure 2 shows the relative frequency of words related to different sports across time. We only consider two words for each sport: the name of the sport itself (e.g. *basketball*) and the acronym of the corresponding North American major league (e.g. *nba*). Figure 2 clearly shows the start and end of American football (henceforth *football*) season in the United States in early September and early February. The football trend line also clearly indicates days of the week when football games are approaching and happening (Thursday-Sunday) and when they are not (Monday-Wednesday). Football coverage decreases as many teams are eliminated from the playoffs in December and January, but it peaks again around the time of the Su-

[3] LDC95T21,LDC98T30

[4] LDC2011T07

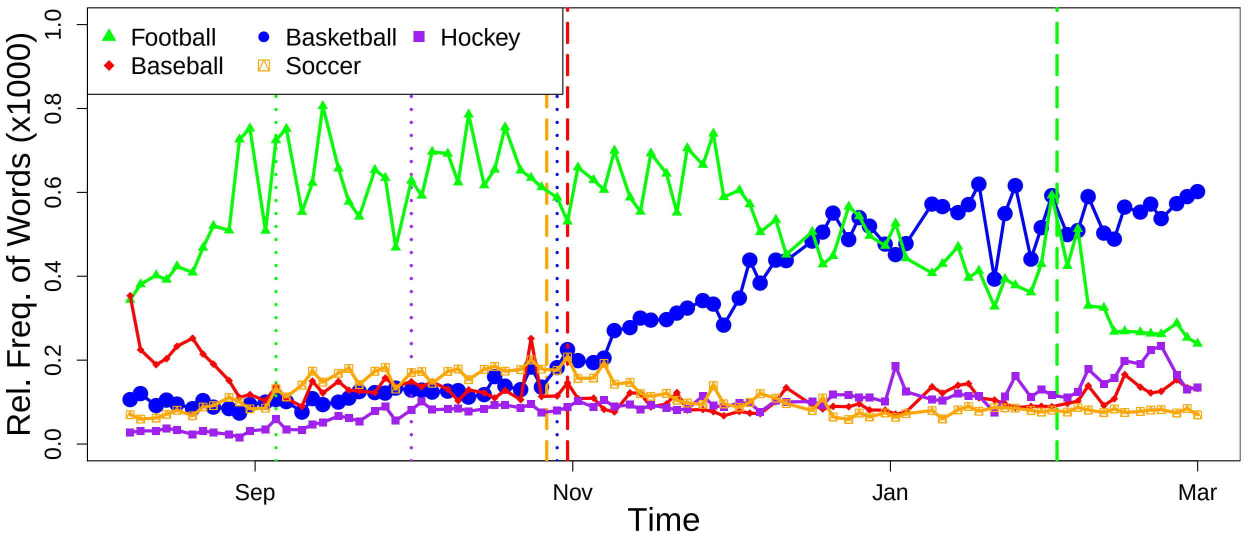

Figure 2: Relative frequency of words related to different sports across time, for all locations. The beginning of professional sports leagues' seasons are indicated with dotted vertical lines and the end of seasons with dashed vertical lines. For example, the National Football League's season began in early September and ended in early February.

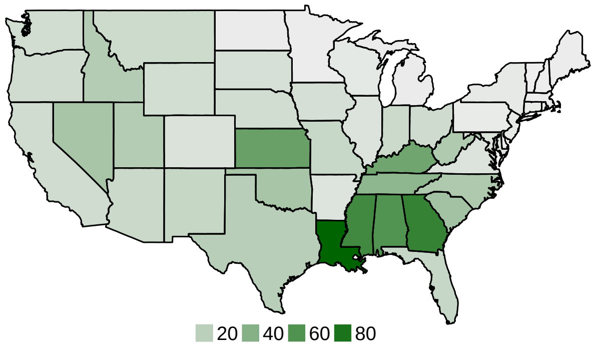

Figure 3: Ratio of football words to hockey words, by state. The darkest shade indicates that football words are mentioned 80 times as often as hockey words; the lightest that football words are mentioned only about 20 times as often.

perbowl[5] in early February before then dropping off dramatically. As football receives more press in August and September, baseball, although still in-season, tends to receive less. As the baseball playoffs begin and many teams' seasons end in late September, baseball receives even less coverage. However, during the world series in late October, news coverage of baseball peaks again. As basketball seasons at all levels of the game, little league through professional, begin in late October, its coverage increases dramatically.

Figure 3 shows the ratio between football words and hockey words for each state. Darker shades indicate that football is mentioned much more often; light shades indicate that the ratio is smaller. Football is mentioned more often in all states, but that is particularly true in Louisiana, where it is mentioned about 80 times as often as hockey. This is no surprise; Louisiana is famously fanatical about football. Newspapers in northeastern states, although they still mention football much more often than hockey, have a much less dramatic ratio of mentions.

Our second example showcases how we may use the ALNC to study gun violence in the United States. In our ongoing research, we are working to automatically classify whether articles report on an instance of gun violence or not. Here, we have identified instances of the word *murder* in articles. We have observed that articles containing the word *murder* are more often about incidents of gun violence than those containing the word *gun*, and we base our preliminary analysis on this set. Figure 4 shows the relative frequency of the word *murder* in articles published in each of the 48 continental states. The term has the highest relative frequency in Iowa and the second highest relative frequency in Louisiana, which has the highest per-capita murder rate among the 50 states.[6] Figure 5 plots the relative frequency of *murder* in the ALNC versus the per capita murder rate in each state. There is a weak positive correlation between the two ($r = 0.19$). The term *murder* is used relatively more frequently in Iowa than any other state, despite it having a very low murder rate. Similarly, newspapers originating in Hawaii and Indiana also use the term *murder* frequently despite those states having relatively few murders per capita. In contrast, news sources from Maryland and Delaware discuss murders much less frequently than we would expect given those states' murder rates.

5. Potential Applications

Our example corpus analyses provide just a glimpse into how the ALNC may shed light on trends and events in the United States. Recent work in the growing field of com-

[5] Season-final championship game of the National Football League

[6] Murder rate data from FBI 2010 Uniform Crime Reports, http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/

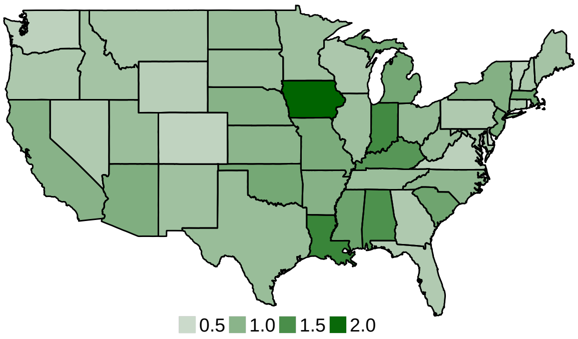

Figure 4: Relative frequency (x10k) of 'murder,' by state. The lightest shade indicates that 'murder' appears once in every 20, 000 words; the darkest, about four times as often.

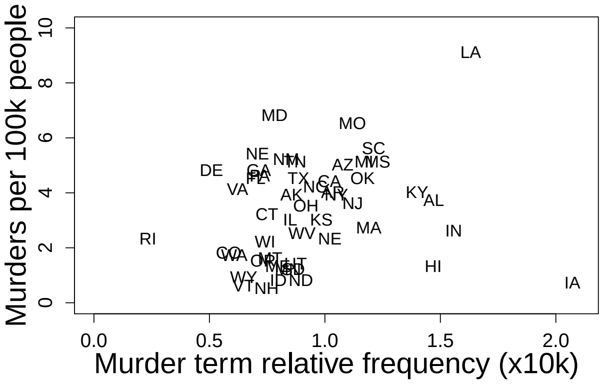

Figure 5: Relative frequency (x10k) of 'murder' vs. the number of murders per 100, 000 people.

putational social science has made use of other large corpora to study language variation by gender, age, personality, and geography (Eisenstein et al., 2010; Schwartz et al., 2013). Other work has shown that it is possible to track public health using language on social media sites (Paul and Dredze, 2011). One interesting direction for future work is to link events discussed in social media with those reported in local or national news stories in the ALNC.

Additionally, the ALNC has the potential to improve performance on core NLP research tasks. Like the NANTC and Gigaword Corpus, the ALNC contributes to the vast quantity of English text available for unsupervised and semi-supervised learning. Because it includes text from a large number of sources, it may also allow for domain adaptation at a geographic or temporal level. For example, the ALNC may be useful for language modeling, where it is known that training models on corpora that were created at a time chronologically similar to the test data yields performance improvements (Rosenfeld, 1995; Levenberg and Osborne, 2009).

6. Conclusion

We have presented and released the ALNC version 1.0. In conjunction with the corpus itself, we presented several interesting examples showcasing how it may be used to answer some simple social science research questions. With its rich location and temporal annotations, the corpus is also applicable to advanced research in public health, political science, sociology, and journalism.

7. Acknowledgements

We would like to thank the creators of the Newspaper Map website for providing us with their database of U.S. newspapers. This material is based on research sponsored by DARPA under contract HR0011-09-1-0044 and by the Johns Hopkins University Human Language Technology Center of Excellence. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

8. References

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

David Graff. 1995. North American News Text Corpus.

Alan I. Leshner, Bruce M. Altevogt, Arlene F. Lee, Margaret A. McCoy, and Institute of Medicine; Co Patrick W. Kelley, Editors; Committee on Priorities for a Public Health Research Agenda to Reduce the Threat of Firearm-Related Violence; Executive Office. 2013. *Priorities for Research to Reduce the Threat of Firearm-Related Violence*. The National Academies Press.

Abby Levenberg and Miles Osborne. 2009. Stream-based randomised language models for SMT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword, 5th Edition.

Michael Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *International Conference on Weblogs and Social Media (ICWSM)*.

Ronald Rosenfeld. 1995. Optimizing lexical and n-gram coverage via judicious use of linguistic data. In *Proceedings of the European Conference on Speech Technology*.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9), 09.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.