Going Beyond the "Clip Search": Relation Extraction From Web News Articles for for Journalistic Network Analysis

Arijit D. Sen,
University of California at Berkeley School of Journalism,
University of California at Berkeley School of Information,
The Dallas Morning News

ari_sen@berkeley.edu

Abstract

Journalists routinely review existing news reporting before undertaking a new project, in order to find new sources and "key players" on a given topic of interest. In industry parlance, this process is called a "clip search." But this process is often highly menial and time-consuming. In this paper, I explore a method of parsing natural language text from web news articles into a knowledge graph in order to speed up this process. The method may also give reporters new insights into potential connections between the players, which can help guide their reporting.

1 Introduction

Just as a researcher would conduct a literature review while writing and academic paper, it is common practice (at least for the author, who is a practicing investigative reporter at a major newspaper in the U.S.) at the beginning of an investigative project for journalists to conduct a "clip search," or systematic review of all existing reporting on the topic he or she is considering investigating. This process is often completed for two reasons:

1. The reporter doesn't want to repeat the work of others, or if they do, they have an ethical responsibility to cite previous work
2. The reporter wants to gain new insight on the topic they are researching, including the manual extraction of entities which could be potential sources or focuses of further investigative scrutiny.

However, this process is often highly time-consuming and repetitive, involving the journalist reading the article, summarizing it and extracting metadata (author, publication, publication date etc.), key passages and entities. Due to its repetitive nature, this process is ripe for automation with web scraping and machine learning methods.

Computer scientists, academic researchers and technology company employees have long taken advantage of machine learning methods to speed up and improve their work. By contrast, until very recently, the use of ML in newsrooms for journalistic projects has been largely non-existent.

I propose one way for newsrooms and independent journalists to achieve the same productivity gains that academicians and tech employees have — a first-of-its kind free tool to quickly and efficiently generate network diagrams between person and organization entities named in web news articles, without the need for any expensive computational resources or domain knowledge of programming, machine learning or data journalism.

The method also may give investigative reporters greater insight into the topic they are studying, allowing them to draw connections between previously unconnected "players," sources or topics.

2 Related Work

According to Zhang et. al., relation extraction as a task traces back to Message Understanding, a DARPA-funded conference which ran from 1987 to 1998 (Grisham and Sundheim). At the 1998 conference, researchers introduced the task of extracting location, employee and product relations from forms and evaluated participants using F1 scores.

Since then, the field has grown tremendously. Contemporary relation extraction models, according to Google researchers Sores et. al, can be broken into four categories. Supervised methods learn mappings between entities in limited schema from annotated training data. A related method, distantly supervised models, attempts to map semantic relationship between two entities annotated on the unstructured text with respect to an underlying knowledge graph (Ye and Ling). These models often rely upon an annotated text corpus of news articles, like the New York Times Corpus introduced by Reidel et. al. Significant progress has been made using these methods — earlier this year researchers introduced KGPool, a graph neural network model which achieved state of the art precision in extracting relations from the NYT corpus.

Open information extraction, introduced in 2007 by Banko et. al., extracts relational tuples from unstructured text without any human input. The universal schema approach, proposed by Reidel et. al. removes the limitation that relations be bounded to a fixed and finite target schema by combining all involved schemas. At the time, this method was able to out-perform all distantly-supervised models.

Researchers have created a number of benchmarks to evaluate the performance of relation extraction models. One of the most common and widely-used benchmarks is SemEval-2010 Task 8, introduced by Hendrickx et. al., which evaluates models on their ability to infer semantic relationships between pairs of nominals. The current state of the art on this task is from Soares et. al. which used the transformer-based model BERT to achieve and F1 of 89.2% on the test set.

Researchers have successfully applied relation extraction models across a number of domains to achieve real-world tasks, including mapping gene-disease relationships (Chun et. al.) and protein-protein interactions in the medical sciences (Huang et. al.).

Even more relevantly to this area, Yamamoto et. al., researchers at Toshiba, used relationship extraction on business news articles related to semi-conductors to extract competitive and cooperative relationships between companies for the purposes of making management decisions and then visualized them in force-directed graphs. However, the researchers only achieved a precision of 67% for cooperative relations, and a competitive relation precision of 87% and admitted their model suffered from overfitting.

In the journalistic domain, data-savvy reporters at large regional and national news outlets have just begun applying machine learning techniques to support investigative projects. Perhaps the earliest major example of this was a 2015 *Los Angeles Times* story in which reporters Ben Poston, Joel Rubin and Anthony Pesce used linear support vector machines (SVM) and maximum entropy classifiers to reveal that the Los Angeles Police Department had misclassified serious assaults for years. A year later, reporters at the *Atlanta Journal-Constitution* applied machine learning methods to analyze more than 100,000 physician disciplinary documents and assigned each a probability that it was related to sexual assaults perpetrated by doctors.

Perhaps the most prolific and notable reporting done using machine learning methods has come from the International Consortium of Investigative Journalists (ICIJ), a collective which brings together reporters from newsrooms around the world to work on stories, usually involving incredibly large document sets. For "The Implant Files," an investigation which revealed more than 1.7 million injuries and at least 83,000 deaths related to implanted medical devices, the group used Talend Real-time BigData Platform, Microsoft SQL Server 2017 and the programming language R to identify 2100 misclassified reports of device malfunctions or injuries, when in fact the patient had died.

In 2019, ICIJ joined forces with the Quartz AI Studio and began publishing stories in its "Luanda Leaks" investigation. The team used TensorFlow's Universal Sentence Encoder to identify semantically similar documents in a 356GB corpus, ultimately revealing that Africa's wealthiest woman, Isabel dos Santos, had funneled millions of taxpayer dollars out of one of the poorest nations in the world.

Though it has been a common technique in academic research and law enforcement for years, network analysis is equally new to the journalistic domain. A survey conducted by AI researcher and journalist Jonathan Stray in 2017, found only 34 stories which employed network analysis or graphs in some way. Stray suggests the most sophisticated of these is ICIJ's investigation into the so-called "Panama Papers."

In this series of stories, reporters examined a 2.6TB corpus of leaked financial records from Mossack Fonseca, a law firm which was ultimately revealed to have been implicated in a multi-billion-dollar money laundering scheme. For the investigation, developers used Apache Tika and Solr for document indexing and processing, Talend for ETL and the graph database Neo4j (with its associated query language Cypher) to store entities and their relations. This graph was then visualized using Linkurious.

But even in this project, the reporters and developers did not attempt to extract entities and their relations from unstructured text data; rather they relied exclusively on data already parsed into structured tabular form.

3 Data

In order to evaluate the performance of the model, the 30 top-ranked articles relating to Social Sentinel (a popular social media monitoring technology employed by schools) were scraped from Google News and parsed into a knowledge graph.

Potential entitles were preprocessed by remove trailing white spaces and newline characters. If, after this preprocessing step, a potential entity resulted in an empty string, it was removed from the evaluation. Stopword removal was not employed due to the importance of these tokens in the names of some entities (ie. "the Future of Privacy Forum.")

To evaluate the performance of the method, I reviewed all of the extracted nodes and relations manually and assessed whether they were correctly categorized as either a person or organization by spaCy.

If a potential entity was referred to in the possessive form (ie. "Social Sentinel's" instead of "Social Sentinel") it was evaluated as if it had no apostrophe. The same decision was made if the entity had additional non-essential tokens, so long as these tertiary tokens did not refer to another entity.

Ambiguous potential entities, such as vague coreferences (ie. "the university") or entities which refer to multiple things in the real world, such as "Black Lives Matter" or "Columbine" were generally considered to be entities.

All non-English tokens extracted from the text were not considered to be proper entities.

4 Method

Unlike most methods for relation extraction, which rely upon either large existing knowledge graphs, sophisticated neural networks or both, I propose a largely-unsupervised approach for parsing these news articles into knowledge graphs. This method uses two popular machine learning libraries: NLTK for sentence tokenization and text extraction and spaCy, a popular named entity recognition tool originally proposed by Honnibal and Montani in 2015.

Using these tools, I first extracted the full text of articles related to a query, split the articles into sentences and tokens and indexed both. Entities were also extracted from these articles, with their sentence and token indices noted. Intra-article co-reference resolution was performed using the neuralcoref package for spaCy. The co-referents, along with their token indices, were stored as a Pandas DataFrame.

Nodes were stored as a list of dictionary objects, with each node containing the name of the entity, its type, the sentence index, its token indices and the hyperlink of the article that it came from.

A relation was created if two potential entities co-occurred in the same sentence or if a target potential entity was in a sentence with a source potential entity's co-referent. Additionally, no relations between two organization nodes were created, as (in the experience of the author) these are unlikely to be useful in a journalistic context. Limited cross-article co-reference resolution was employed — if an entity was extracted in one article with the same string as an extracted entity in another article, these were represented as a single node, rather than as two.

When parsed, these nodes and relations were stored as a JSON file for later use in the popular JavaScript visualization library D3.

While this method may seem crude and ineffectual, there is reason to suggest it may perform well, given common journalistic syntactic structure.

It is standard practice in journalism to introduce a human subject using their first and last names, as well as their affiliation, on first reference. For example, a sentence in an article on Facebook may read "Facebook CEO Mark Zuckerberg…" or "Sheryl Sandberg, the chief operating officer at Facebook,…." In this form, later sentences typically refer to the person by either a pronoun or their last name ie. "Zukerberg changed the name of the company to Meta last month…" or "He changed the name to Meta last month."

In the case of these sentences, a correct parsing creates nodes for "Mark Zuckerberg" and "Sheryl Sandberg" with the person type and connect them to "Facebook" which would be an organization type node. An additional edge would be drawn between "Mark Zuckerberg" and "Meta" since "Mark Zuckerberg" is co-referred to as either "he" or "Zuckerberg" in the sentence with "Meta," (an organization-type node).

5 Analysis

In order to evaluate the performance of my model, I compared the potential entities that were extracted to my assessment of their entity type validity.

The sample of 30 articles generated 1442 entities and 707 distinct string representations. Based on my assessment, 1048 of these 1442 entities were properly categorized as either a person or an organization, for an estimated precision of roughly 0.72 and an estimated f1 of 0.84.

Slightly higher performance was achieved on the potential entities classified as organizations. 892 of the 1442 extracted potential entities were classified in the organization type. Of these, 702 were classified correctly, with an estimated precision of 0.79 and F1 of 0.88.

Potential entities labeled as people achieved an estimated precision of 0.63 and an F1 of 0.77.

I also evaluated the potential entities which were linked together against my human assessment of whether both entities were classified as the correct type. Of the 346 extracted relationship pairs, 211 had both entities categorized correctly, according to my human labels, achieving an estimated accuracy of 61%.

To assess the semantic validity of the connections between linked potential entities, I reviewed each article to determine whether there was a direct link between the pair. I define a direct link as either working with someone (in the case of person to person links) or working for a company (in the case of person to organization links). I deemed 108 of the 212 (~0.501) properly type-classified relations to be direct links.

However, even those linked nodes without direct relationships may encode useful information for the journalist. For example, the relation identified between "Lam Thuy Vo" and "Social Sentinel" identifies one of the reporters who investigated the company for Buzzfeed News in 2019. And, as Stray identifies in his 2017 presentation, journalists would far prefer to have false positives (entities extracted and linked that turn out to not be related) than false negatives (entities not extracted or not linked when they are related).

6 Discussion

Since my model is meant to be applied across a number of topic phrases, an estimation of its true performance is near-impossible. But an assessment of these estimated metrics and underlying data does shed useful insight into its successes and shortcomings.

Upon examination of the underlying data, it appears that many of these underlying relation errors come from two-word proper noun phrases, which spaCy often confuses with names of people. More puzzlingly, spaCy's NER system classified entities with the same string representation as different entity types depending on the article, or even sentence, it was in. While we wait for the developers of spaCy to resolve these errors, further improvements to this method may choose to take the majority class for these disagreements.

A common stylistic feature in journalism, referring to the author by the name of the organization, was also the source of some of these issues. For example, in the 2019 Buzzfeed News article "Your Dumb Tweets Are Getting Flagged To People Trying To Stop School Shootings" "Margolis" is linked to "Buzzfeed News" because of the sentence '"Our children are sharing all kinds of things digitally, and included in that are potential acts of harm,' Margolis told BuzzFeed News." A similar error occurs with "Breault" and "Buzzfeed News" later on in the same story, in the sentence "Breault told BuzzFeed News she wasn't surprised that an algorithm didn't get her peer group's dark sense of humor." In future iterations of this method, this stylistic feature may be able to be accounted for, albeit with a small sacrifice in computational efficiency.

When comparing the models estimated performance to the performance of the state of the art, as expected, it does not perform as well. However, these systems are being tested on different corpora thus comparing them may not be valid. Also, this model does not require any existing relational pairs for training — the only training data provided is spaCy's CPU-optimized small English pipeline. This means the model can be run on lower-powered machines without access to GPU computing resources. However, as the number of articles increases, the speed of this model decreases, meaning that existing knowledge graph models may be a better choice for more well-reported on topics (ie. Facebook).

The performance of the system is also highly reliant upon the underlying systems it uses, mainly spaCy and Google search. At present this means that it will be unable to perform better than the named entity recognition built into spaCy, which is itself not the state of the art. Search results also limit the the model's performance, as the system can only be as good as the articles it parses. Further refinements to this method may choose to filter which articles are parsed based on variables like text length and source.

As referenced by Stray in his 2017 talk, more insights could be drawn by using centrality algorithms from graph theory, allowing investigative reporters to identify the key players in a story even faster. And with advancements in optical character recognition technologies this method could be applied to other corpora reviewed later in the reporting process, such as documents received via Freedom of Information Act requests, though this would likely come with a significant decrease in performance.

7 Conclusion

The tedious clip search process journalists routinely undertake is an essential part of producing strong investigative work. For years, this process has taken journalists hours of precious time which could be spent broadening their understanding or developing deeper relationships with sources.

In this paper I have outlined a computationally-inexpensive method to extract relation pairs from unstructured web news text and parse that text into knowledge graphs. Despite its shortcomings, this proposed method, and its future iterations, have the potential to automate much of this process even for the least tech-savvy reporters, allowing them to produce better, more impactful stories in a fraction of the time.

8 References

Q. Zhang, M. Chen and L. Liu (2017). “A Review on Entity Relation Extraction.”

R. Grisham and B. Sundheim (1996). “Message Understanding Conference – 6: A Brief History,” *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics,* Volume 1, pp. 466 – 471.

S. Reidel, L. Yao and A. McCallum (2010). “Modeling Relations and Their Mentions without Labeled Text.” *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III.* pp. 148-163.

L. Soares, N. FitzGerald, J. Ling, T. Kwiatkowski. “Matching the Blanks: Distributional Similarity for Relation Learning.” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2895–2905.

Z. Ye and Z. Ling (2019). “Distant supervision relation extraction with intra-bag and inter-bag attentions.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, Volume 1, pp. 2810–2819.

A. Nadgeri A. Bastos, K. Singh, I. Onando Mulang J. Hoffart, S, Shekarpour and V. Saraswat. “KGPool: Dynamic Knowledge Graph Context Selection for Relation Extraction” *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021).*

M. Banko, M. Cafarella, S. Soderland, M. Broadhead and O. Etzioni (2007). “Open Information Extraction from the Web.” *The Association for the Advancement of Artificial Intelligence Conference 2007.* Pp. 2670-2676.

S. Reidel, L. Yao, A. McCallum and B. Marlin (2013). “Relation Extraction with Matrix Factorization and Universal Schemas.” *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics 2013*, pages 74–84.

I. Hendrickx , S. Kim , Z. Kozareva, P. Nakov, D. Se´aghdha, S. Pado , M. Pennacchiotti, L. Romano and S. Szpakowicz (2010). “SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals.” *Proceedings of the 5th International Workshop on Semantic Evaluation,* ACL 2010, pp. 33–38.

H. Chun; Y. Tsuruoka; J. Kim; Rie Shiba; Naoki Nagata; Teruyoshi Hishiki; Jun-ichi Tsujii (2006). "Extraction of Gene-Disease Relations from Medline Using Domain Dictionaries and Machine

Learning". *Pacific Symposium on Biocomputing*.

M. Huang, X. Zhu, Y. Hao, D. Payan, K. Qu and M. Li (2004). "Discovering patterns to extract protein-protein interactions from full texts". *Bioinformatics*. Volume 20, pp. 3604–3612.

Hao Zhang, Frank Boons, Riza Batista-Navarro, "Whose story is it anyway? Automatic extraction of accounts from news articles," Information Processing & Management, Volume 56, Issue 5, 2019, Pages 1837-1848, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.02.012.

M. Kanakaraj and S. S. Kamath, "NLP based intelligent news search engine using information extraction from e-newspapers," *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1-5

Yamamoto, Y. Miyamura, K. Nakata and M. Okamoto, "Company Relation Extraction from Web News Articles for Analyzing Industry Structure," *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, 2017, pp. 89-92.

Hao Zhang, Frank Boons, Riza Batista-Navarro, "Whose story is it anyway? Automatic extraction of accounts from news articles," *Information Processing & Management*, Volume 56, Issue 5, 2019, Pages 1837-1848.

M. Kanakaraj and S. Kamath, "NLP based intelligent news search engine using information extraction from e-newspapers," *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1-5, doi: 10.1109/ICCIC.2014.7238500.

M. Honnibal and I. Montani, "An Improved Non-monotonic Transition System for Dependency Parsing," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1373 – 1378, https://aclweb.org/anthology/D/D15/D15-1162.

E. Diaz-Struck and R. Caravajal, "Algorithms, Analysis And Adverse Events: How ICIJ Used Machine Learning To Help Find Medical Device Issues," International Consortium of Investigative Journalists, 2018, https://www.icij.org/investigations/implant-files/algorithms-analysis-and-adverse-events-how-icij-used-machine-learning-to-help-find-medical-device-issues/.

J. Stray, "Network Analysis in Journalism: Practices and Possibilities," *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Halifax 2017,* 2017.

M. Guevara, "How Artificial Intelligence Can Help Us Crack More Panama Papers Stories," International Consortium of Investigative Journalists, 2019, https://www.icij.org/inside-icij/2019/03/how-artificial-intelligence-can-help-us-crack-more-panama-papers-stories/.

M. Cabra and E. Kissane, "The People and Tech Behind the Panama Papers," Source, 2016, https://source.opennews.org/articles/people-and-tech-behind-panama-papers/.

M. Cabra, "How the ICIJ Used Neo4j to Unravel the Panama Papers," Neo4j, 2016, https://neo4j.com/blog/icij-neo4j-unravel-panama-papers/.

B. Poston, J. Rubin and A. Pesce, "LAPD underreported serious assaults, skewing crime stats for 8 years," *Los Angeles Times*, 2015, https://www.latimes.com/local/cityhall/la-me-crime-stats-20151015-story.html.

D. Robbins, C. Teegardin, A. Hart, J. Ernsthausen, R. Horne, R. Watkins and L. Norder, "How the Doctors & Sex Abuse project came about," *Atlanta Journal-*

Constitution, 2016, https://doctors.ajc.com/

J. Merrill, “How Quartz used AI to sort through the Luanda Leaks,” Quartz, 2020, https://qz.com/1786896/ai-for-investigations-sorting-through-the-luanda-leaks/.

L. Vo and P. Aldhous, “Your Dumb Tweets Are Getting Flagged To People Trying To Stop School Shootings”, Buzzfeed News, 2019, https://www.buzzfeednews.com/article/lamvo/social-sentinel-school-officials-shootings-flag-social-media.