Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
May 2, 2024
A Regional Conference of the Society for Industrial and Systems Engineering

Designing an Intelligent System to Map Global Connections

Elijah Bellamy, Kseniya Farrell, Aiden Hopping, James Pinter, Michael Saju, and David Beskow

Department of Systems Engineering,
United States Military Academy,
West Point, NY 10996

Corresponding author's Email: jwpinter2@gmail.com

Author Note: The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Extended Abstract: This study develops a knowledge graph from the Common Crawl News Dataset to provide situational awareness and answer information requirements for national security leaders. We map dynamic global interactions between entities using news data collected over a defined period. We developed a data pipeline to extract semantic content from the Common Crawl News feed and filter it to content related to national security. We build a knowledge graph using national security-related content using Named Entity Recognition, Named Entity Disambiguation, and Named Entity Linking. After developing this intelligent system, we demonstrate its value in a case study on a Russian language knowledge graph focused on the Russian invasion of Ukraine. The knowledge graphs developed in this study provide decision-makers an understanding of the entities and links in our highly interconnected world.

The internet offers vast amounts of data for decision-making but poses challenges in sifting through the information. This paper proposes using knowledge graphs to organize internet news data for better decision-making. Knowledge graphs integrate information into structured human knowledge, aiding information retrieval and analysis. Named Entity Recognition (NER), Named Entity Disambiguation (NED), and Named Entity Linking (NEL) are crucial steps in building knowledge graphs from large online content archives.

Researchers from the University of Groningen and the University of Twente (in the Netherlands) introduced the term knowledge graph in the 1980's as a knowledge-based system (Ehrlinger & Wöß, 2016). A knowledge graph fundamentally "*...integrates information into an ontology and applies a reasoner*" (Ehrlinger & Wöß, 2016). Knowledge graphs have been used for information retrieval, content analysis, question answering, and knowledge base population (Shen, Wang, & Han, 2014).

A knowledge graph is a graph representation of a knowledge base. A knowledge graph can be represented by *factual triple* in the form of (subject, predicate, object), for example (Michael Jordan, member of, Chicago Bulls) (Ji, Pan, Cambria, Marttinen, & Philip, 2021). Building a knowledge graph begins with extracting entities from semantic content, known as Named Entity Recognition (NER), first introduced by Ralph Grishman at the 6th Message Understanding Conference in 1996 (Grishman & Sundheim, 1996). NER extracts and labels distinct entities within unstructured semantic content (Yadav & Bethard, 2018). Entity types vary but typically include persons, organizations, locations, products, and events. Named Entity Disambiguation (NED) attempts to normalize the specific name used to refer to each entity. For example, NED identifies that "President Reagan" and "Ronald Reagan" are the same entity. It would also differentiate between "Jordan" the NBA basketball player, and "Jordan" the country. One of the primary methods for NED uses max prior probability to determine the appropriate link to a knowledge base (select the most frequently used candidate). A more advanced method measures the similarity between entity context and a knowledge base (like Wikipedia) description either with a bag of words (BOW) or embedding representation (Al-Moslmi, Ocaña, Opdahl, & Veres, 2020).

Common Crawl provides a rich source of internet data. The paper focuses on six months of English and Russian news data from Common Crawl News Stream (60 million articles). A supervised machine learning filter is used to identify articles relevant to national security (removing news categories such as lifestyle, sports, and technology).

The methodology parses common crawl data from raw HTML to structured data and conducts language-agnostic BERT embeddings before extracting all named entities. Named entities are filtered only to include people, locations, geo-political entities, products, organizations, events, and facilities. These named entities are then disambiguated by linking them to Wiki Data with max prior linking (we select the most frequently used candidate terms). The Wiki Data IDs are used to disambiguate entities in the data. These entities are associated by co-mentioning in an appropriate context window. We tested relating the entity's sentence, paragraph, and article context windows and evaluating the resulting graph density, as seen in Figure 1. The paragraph context window proved the most valuable. The Wiki Data was also used to geocode geo-

political entities and locations, allowing us to visualize news events and news density across the globe, as is seen in Figure 2.

The resulting knowledge graph for one month contains over 500,000 nodes and 4.1 million links. It visually represents connections between entities, aiding in analysis and decision-making. A case study analyzing the Russian-Ukrainian conflict with a knowledge graph reveals extensive connections between individuals and organizations. We demonstrate how to identify and explore an entity of interest, namely Yevgeny Prigozhin. Knowledge graphs offer a valuable approach to organizing internet data for national security decision-making. They facilitate a deeper understanding of relationships between entities, enhancing situational awareness and enabling proactive responses to challenges.

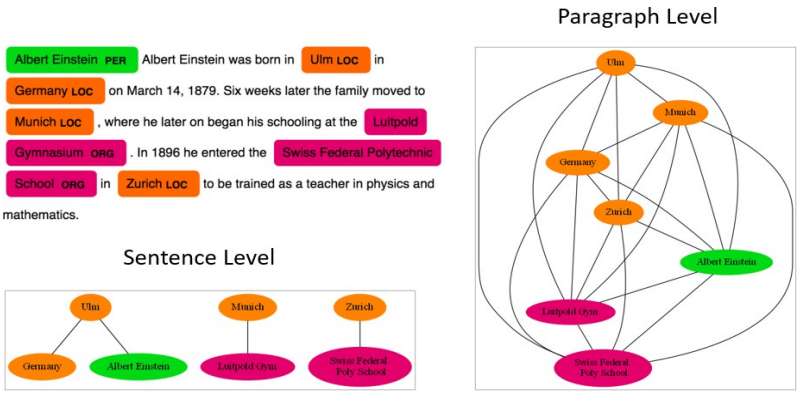

Figure 1: Demonstrating NER and Co-mentioning at the Sentence and Paragraph context windows

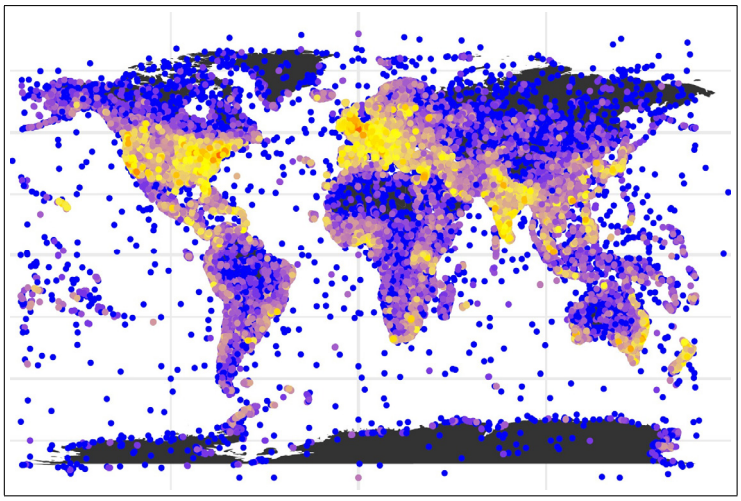

Figure 2: All location and geo-political entities plotted on map (coloring is the log of DBSCAN frequency).

Keywords: Knowledge Graph, Named Entity Recognition, Named Entity Linking

1. References

Al-Moslmi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, *8*, 32862–32881.

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. In *International conference on semantic systems*.
Retrieved from https://api.semanticscholar.org/CorpusID:8536105

Grishman, R., & Sundheim, B. M. (1996). Message understanding conference- 6: A brief history. In *International conference on computational linguistics*. Retrieved from https://api.semanticscholar.org/CorpusID:11986411

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, *33*(2), 494–514.

Shen, W., Wang, J., & Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transac- tions on Knowledge and Data Engineering*, *27*(2), 443–460.

Yadav, V., & Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. *ArXiv*, *abs/1910.11470*. Retrieved from https://api.semanticscholar.org/CorpusID:49587276