An Evaluation Framework for Mapping News Headlines to Event Classes in a Knowledge Graph

Steve Fonin Mbouadeu
St. John's University
steve.mbouadeu19@stjohns.edu

Martin Lorenzo
IBM Research
mlorenzo@ibm.com

Ken Barker
IBM Research
kjbarker@us.ibm.com

Oktie Hassanzadeh
IBM Research
hassanzadeh@us.ibm.com

Abstract

Mapping ongoing news headlines to event-related classes in a rich knowledge base can be an important component in a knowledge-based event analysis and forecasting solution. In this paper, we present a methodology for creating a benchmark dataset of news headlines mapped to event classes in Wikidata, and resources for the evaluation of methods that perform the mapping. We use the dataset to study two classes of unsupervised methods for this task: 1) adaptations of classic entity linking methods, and 2) methods that treat the problem as a zero-shot text classification problem. For the first approach, we evaluate off-the-shelf entity linking systems. For the second approach, we explore a) pre-trained natural language inference (NLI) models, and b) pre-trained large generative language models. We present the results of our evaluation, lessons learned, and directions for future work. The dataset and scripts for evaluation are made publicly available.

1 Introduction

Businesses and organizations can benefit from seeking knowledge of new events that may have an impact on their business. To assist in this task, there are several media monitoring solutions with features that can provide alerts and real-time analysis for ongoing events. The majority of existing solutions are centered around entities and/or topics. For example, they identify mentions of key companies or people, group texts by topics, and analyze contents for sentiment. On the other hand, there is great value in an event-centric solution that identifies ongoing events and analyzes the characteristics of the identified events to enable event-based reasoning. In particular, such a solution would enable causal reasoning to determine the causes and consequences of ongoing events and identify potential risks and opportunities (Hassanzadeh et al., 2022).

To enable a knowledge-driven event-centric news analysis and monitoring solution, a key requirement is the ability to accurately map ongoing news to event-related classes in a knowledge base. One way to perform this mapping is to treat event-related classes as a set of categories (or topics) and classify news headlines into these categories. Prior work has studied classification methods for news headlines (e.g., see Awasthy et al. (2021); Rana et al. (2014) and references therein). The majority of existing methods rely on supervised learning and therefore require a training corpus. For a generic solution that can adapt to changing event classes or one that can be tuned easily for different domains, it is not feasible to rely on the availability of training corpora large enough for accurate classification.

In the absence of training data, the alternative solution is to apply unsupervised or weakly supervised classification methods that rely on little or no training data. Such methods often rely on rules and pre-trained generic models. More recently, pre-trained language models, and in particular large language models, have shown superior performance in such settings. As a result, we have seen a surge in the number of available models, each using different architectures, parameters, pre-training corpora, and fine-tuning strategies. Choosing the right model for a given task requires an evaluation framework to measure the accuracy of the models on the end task.

In this paper, we present an evaluation framework for unsupervised mapping of news headlines to event classes in a knowledge graph. To the best of our knowledge, this is the first benchmark dataset and evaluation framework for this task. In what follows, we first present the task definition and use cases we envision for the task. We then describe our methodology for creating the benchmark dataset. Next, we present the results of our evaluation of a number of methods belonging to two different kinds of unsupervised techniques. We discuss key lessons learned and a number of avenues for future work. The datasets used in our

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 44–52, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_006

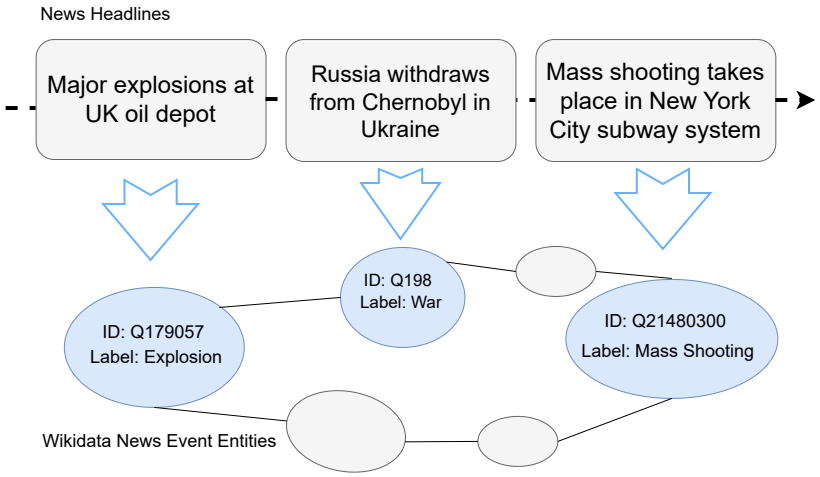

Figure 1: Example of News Headlines and Event Classes in Our Benchmark

experiments as well as the evaluation framework are publicly available (Mbouadeu et al., 2023).

2 Task Definition and Use Cases

Our target task in this paper is as follows: Given a news headline and a set of event classes from a knowledge graph, find the most relevant event class to the news headline. The news headline is a short text (typically a sentence) that indicates the content of a news article by providing a concise summary of the article's contents. The knowledge graph contains event-related classes. Each class comes with one or more labels, a description of the class, and possibly a class hierarchy and other attributes. Figure 1 shows examples of news headlines, event classes, and their mappings. We refer to this task as *News Headline Event Mapping*. Note that this task is different from the event linking task defined by Yu et al. (2023) which takes an event mention (a phrase) and a context as input, and finds a specific Wikipedia article as output. Nevertheless, as described in Section 4.1, such methods can be used for our task.

Figure 2 shows example use cases for news headline event mapping in the context of a knowledge-based news event analysis solution (Hassanzadeh et al., 2022). In this context, news headlines from a variety of sources or a news content aggregation service (e.g., EventRegistry (Leban et al., 2014)) are monitored in order to identify major news that could have an impact on a users' organization, on a certain region, or more generally on society. This domain of interest is defined through a knowledge graph of events that contains a rich source of knowledge about past events and event classes. Such a source of knowledge can be gathered through automated knowledge extraction methods (Hassanzadeh et al., 2020; Heindorf et al., 2020) or be derived from domain-specific or general-domain knowledge sources such as Wikidata (Vrandečić and Krötzsch, 2014). The knowledge graph provides event classes along with labels and descriptions to be used for news headline event mapping. The output of headline event mapping is then used for an analysis of the potential causes and effects of the identified event. The outcome can be used as a part of a news monitoring solution to create alerts for the identified event or its consequences so that it can assist with managing a potential risk or opportunity. It can also provide the required knowledge for an analyst looking at the implications of ongoing news for a business or organization. Finally, it can be used as an input for scenario planning (Sohrabi et al., 2019) or event forecasting (Muthiah et al., 2016; Radinsky et al., 2012).

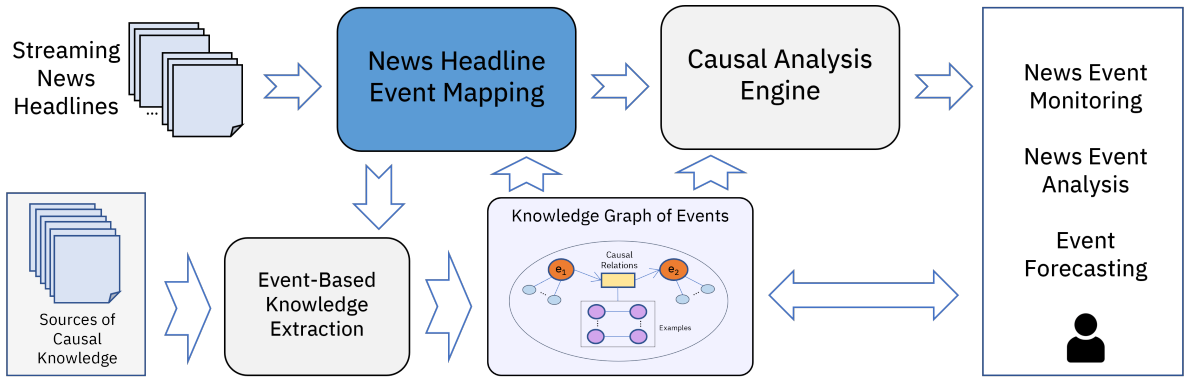

Figure 2: News Headline Event Mapping in a Knowledge-Based Event Analysis Solution (Hassanzadeh et al., 2022)

3 Benchmark Dataset

To the best of our knowledge, there is no benchmark dataset for the task of news headline event mapping. There are benchmarks on related tasks such as entity linking (van Erp et al., 2016) that include news headlines. However, none of these benchmarks provide ground truth event class annotations. We have, therefore, curated a new dataset designed for the news headline event mapping task using Wikidata and Wikinews. First, we leveraged the links to Wikinews articles in Wikidata to gather a collection of event-related instances. To focus on event classes, we then filtered out instances that are not subclasses of `occurrence (Q1190554)` as well as classes with very short labels, as some non-event related entities are also linked to Wikinews. Finally, these articles were reviewed manually to check whether they related to news headlines and news events. This yielded 105 Wikinews headlines mapped to Wikidata event classes. We manually added five headlines from other sources for a final dataset of 110 mappings of headlines to Wikidata event classes. The labels of all of the Wikidata classes included in our benchmark are shown in Figure 3. The examples of Figure 1 were taken from the dataset.

There are a number of other benchmark datasets in the literature for related tasks. Closest to our task is that of zero-shot sentence classification. Yin et al. (2019) present an excellent review of benchmarks for this task. Many benchmarks for news headline classification and for zero-shot sentence classification target binary classification (e.g., for emotions or sentiment or clickbait detection), or a small number of topics. The closest to our benchmark is the Yahoo! dataset (Zhang et al., 2015), which consists of 10 topics. To our knowledge, there is no benchmark that targets the task of assigning news headlines to event classes in a knowledge graph or a large number of well-defined topics.

4 Evaluation

We use our benchmark and evaluation framework to evaluate the effectiveness of a number of different kinds of methods in news headline event mapping. We first describe the approaches and implementation details for each method. We then present the results of the evaluation and a detailed discussion on key lessons learned and directions for future work.

4.1 Methods

We experiment with news headline event mapping methods ranging from a simple similarity-based baseline to adaptations of classic entity linking tools and large generative language models.

4.1.1 Zero-Shot Classifiers

We evaluate two zero-shot text classification methods. One is a simple baseline based on textual similarity, while another uses state-of-the-art pre-trained language models for classification.

Similarity-Based Baseline (`Fuzzy`) This method identifies a substring of a headline that is suggestive of an event occurrence, and finds the most similar event class label in the knowledge graph to that substring. The substring is found through a sliding window of bigrams and trigrams of word tokens in the input, and matching them using Levenshtein distance (Chandrasekar et al., 2017) to our target event class labels. The event class with the lowest distance is returned as the most similar class to the headline.

aircraft_crash attack aviation_accident bomb_attack climate_change
coup_d'état crime cyberattack death disease_outbreak earthquake espionage
explosion fire flood impeachment_in_the_United_States infectious_disease
killing mass_murder mass_shooting massacre murder natural_disaster
procession scandal school_shooting social_issue stabbing_attack
terrorist_attack transport_accident volcanic_eruption war work_accident

Figure 3: News Event Classes in Our Benchmark

Zero-Shot Text Classifier using Natrual Language Inference (ZSTC) This classifier is a standard "MNLI" model: a pre-trained language model fine-tuned on the multi-genre textual entailment corpus for the Natural Language Inference (NLI) task from the RepEval workshop (Williams et al., 2018). Instances in the MNLI corpus are pairs of sentences with a label indicating whether the first sentence Entails the second sentence, is Contradicted by it, or is independent of it ("Neutral"). MNLI models can be used for zero-shot text classification by supplying the text to be classified as sentence1, and a textual representation of a target class as sentence2. For our experiments, the textual representation of a class (sentence2) is simply the English label of the class in Wikidata. Sentence1 is the headline to be classified. The target class whose textual representation (label) has the highest Entailment score is the predicted class for the text. Our zero-shot text classifier uses the RoBERTa-large (Liu et al., 2019) language model fine-tuned on MNLI.

4.1.2 Classic Entity Linkers & Adaptations

We considered a number of state-of-the-art open-source entity linking (EL) systems to adapt to include in our experiments. Most entity linking solutions are trained to work only on named entities (e.g., people, locations, organizations) and fail when it comes to events. We considered EL systems that are more easily adaptable for mapping to event classes. The systems we considered include BLINK (Wu et al., 2020), OpenTapioca (Delpeuch, 2019), Falcon (Sakor et al., 2020), and Wikifier (Brank et al., 2017). Out of these, our adaptation of BLINK failed to perform well, and OpenTapioca required a training corpus. Although training OpenTapioca using our dataset provided promising results (another potential use case for our dataset), we excluded the results in this paper to focus on fully unsupervised (zero-shot) methods.

Falcon 2.0 (Falcon EM) Falcon 2.0 (Sakor et al., 2020) leverages NLP techniques to achieve state-of-the-art entity linking performance on a number of EL datasets, notably on question-structured prompts (Sakor et al., 2020). Given a prompt, it generates a list of entity surface forms, similar to event mentions. After generating these surface forms or tokens, it selects candidate entities for each of them by searching them in an information retrieval (IR) index (powered by Elasticsearch) of a Wikidata data dump. We only included Wikidata concepts that were recursively instances or subclasses of event classes in the dump to tailor it to our task. In our evaluation, we used Falcon to match headlines to Wikidata concept labels. If Falcon did not generate at least one candidate concept, we successively stripped tokens from the right of the headline, approximating more general phrases. we repeated the process until either a candidate concept was found or the phrase became empty. The resulting candidate concepts were then ranked using SPARQL ASK queries, measuring the taxonomic distance between the candidate concepts and our chosen news event classes. The class from our set of target event classes that was the shortest distance from a Falcon-generated candidate concept was chosen as the predicted class for the headline.

Wikifier (`Wikifier`) Wikifier (Brank et al., 2017) is a service for the task of "wikification" – taking an input text and annotating phrases in the text with Wikipedia URLs. Wikifier employs surface forms of hyperlinks in Wikipedia to perform linking to Wikipedia entities. For example, the Wikipedia page for earthquakes contains a link to the tsunami page. This suggests that earthquake is related to tsunami. For any surface form throughout Wikipedia that is present in the given text, Wikifier makes a candidate entity of the underlying entity. A directed mention-concept graph is created, linking surface forms to these candidate entities. Wikifier performs a global disambiguation based on the distance between entities. Distance represents the number of hyperlink hops required to get from one page to another. The smaller the distance, the more related the entities are considered. The relatedness metrics are used to score the candidate entities. Wikifier returns these candidate entities as predictions along with their scores. We converted the Wikipedia hyperlinks to Wikidata concepts with a simple lookup query. For our evaluation, we picked the top prediction that was among one of our target event classes.

4.1.3 Large Generative Language Models

Another way to perform zero-shot classification is through the use of generative large language models (LLMs) and prompts. There are a number of LLMs available with different architectures, parameter sizes, and resource requirements. For the results in this paper, we decided to pick just one of the popular LLMs with reasonable resource requirements, namely GPT-J 6B (Wang and Komatsuzaki, 2021), so that our experiments are reproducible without requiring access to commercial APIs or expensive GPUs. We include two different prompting strategies for the results in this paper. Experiments with a wider variety of LLMs and more extensive prompt engineering are a subject for future work.

GPT-J Event Mapping (`GPT-J EM`) Our goal here is to form a prompt that yields the generation of the relevant event class by the LLM. One way to create a prompt is to provide a few examples (a "few-shot" strategy) of headline + delimiter + known event class label, followed by the headline to be classified and the same delimiter, and ask the model to generate completion text. Having experimented with a number of prompting strategies, we decided to use a co-training approach (Lang et al., 2022).

Co-training works similarly to cross-validation, where each individual headline is mapped with zero shots using GPT-J and then the best-performing headlines are used to generate a few-shot prompt. The output of this method is an event label that we then mapped to Wikidata.

GPT-J Event Mapping with Types (`GPT-J EMT`) We continued our experiments with GPT-J by including all the event classes in the prompt along with the pre-training. The set of labels from our news event classes were listed separately and prefixed with "types:". We then added this list to the beginning of the prompt to signal the categories to be picked from. We also prefixed each annotation in the pre-training examples with "type:" to establish that association. Additionally, we implemented a catch-all for non-event classifications. If a prediction didn't match an event class label, we performed textual similarity matching with our target event labels to find the most similar event class to return as output.

Table 1: Accuracy Results

	Fuzzy	ZSTC	Falcon EM	Wikifier	GPT-J EM	GPT-J EMT
Correct @1	22	23	33	49	65	74
Accuracy	0.2	0.209	0.3	0.445	0.591	0.673

4.2 Results

For our evaluation we ran each system from Section 4.1 on the headlines from our news event corpus to generate the systems' best predicted event classes. We calculated accuracy of each system as the percentage of top-ranked predictions matching the gold event class.

The results are shown in Table 1. In addition to the benchmark datasets, all of our outputs as well as our evaluation script are available on our GitHub repository (Mbouadeu et al., 2023).

The zero-shot classifier methods (`Fuzzy` and `ZSTC`) performed comparably. They both did well on headlines that have linguistic overlap with a target class label. `Fuzzy` works when there is surface/lexical overlap, whereas `ZSTC` takes advantage of semantic overlap. Examples of headlines having linguistic overlap with target classes are: "*Major explosions at UK oil depot*", "*Mass shooting takes place in New York City subway system*", and "*Myanmar military vows to abide by constitution amid coup fears*". The first two, for example, have *explosion* and *mass shooting* target event classes, and labels for those classes appear verbatim in the headlines.

Linguistic overlap can result in frequent false positives, particularly for very general target classes. For example, for the headline "*More than 80 people killed in Nice, France attack on Bastille Day*", both methods associated "killed" with the `killing` event class and "attack" with the `attack` class. Ideally, both classes would be included among the gold classes and a ranking metric used to give credit to multiple (ranked) system predictions. For simplicity, and for even comparison to systems without ranked/scored output, we only report accuracy (correct @1).

Among the classic entity linking methods, `Wikifier` performed better than `Falcon EM`. In general, it was able to map more challenging headlines having no obvious linguistic overlap with class labels. For example, it was able to map the headline "*Russia withdraws from Chernobyl in Ukraine*" to the `war` event class.

The LLM-based methods also showed the ability to map news headlines to event classes whose labels do not appear in the headline. Examples of such headlines are: "*Nine firefighters killed in South Carolina blaze*" (event class `fire`), and "*Attack at Texas elementary school kills at least 19, including 18 children*". (event class `school shooting`). The second example is particularly interesting because the LLM-based methods preferred the more specific `school shooting` event class in spite of the headline's overlap with the label of the `killing` class. The LLM-based methods (`GPT-J EM` and `GPT-J EMT`) also showed a more consistent ability to map news headlines to events with labels that are generalizations of text appearing in headlines, such as `violence` and `natural disasters`.

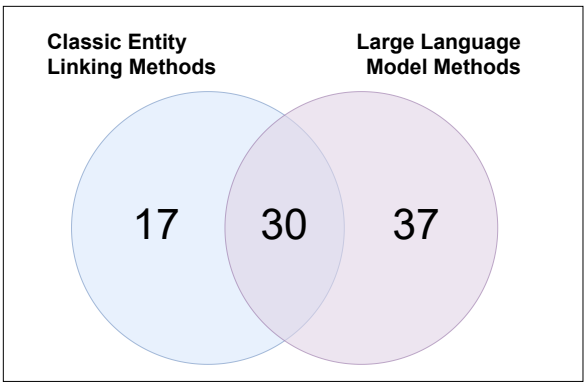

Figure 4: Overlap of Accurate Prediction Coverage of Entity Linking Adaptations and LLM-based Methods

4.3 Lessons Learned and Future Work

An Ensemble Approach Although the classifier and entity linking based methods did not perform as well as the LLM-based methods, they complement each other. Combining their coverage of successfully mapping headlines in the dataset yields 95% accuracy. When comparing their coverage of the dataset, they generally succeed and fail on different types of headlines. The classic entity linker adaptations do well with headlines with single-worded event mentions that match directly to event classes. LLM-based mappers do well with multi-worded event mentions that are not necessarily substrings of the event class labels and those without any clear event mentions as well. They are still able to make the association between these more ambiguous mentions and the event entities, presumably from their learning from large amounts of text. This is further supported by the fact that of the 33% of headlines that the LLM-based mappers failed to correctly map, 87% have a clear event mention that closely matches the labels of their event classes. Nevertheless, there is still a noticeable amount of overlap between the two types of methods, as shown in Figure 4. However, these results do suggest that an ensemble approach that combines techniques used in classic entity linking and leverages large language models, intentionally deciding how and when to apply them, would improve performance on this task.

A Larger Dataset Despite the relatively small size of the current version of our dataset, we believe our results are informative, and highlight the strengths and weaknesses of different classes of methods. We also believe the small size of the data reflects well the real-world use case of building a generic and adaptable event monitoring solution, where gathering ground truth data for supervised solutions could be prohibitively expensive. Still, the methodology we outlined in Section 3 can be extended to gather a larger and more diverse collection of news headlines mapped to event classes. At the time of writing this manuscript, we are applying a similar strategy to news headlines that are referenced from within Wikipedia-related event articles to curate a second, much larger version of our dataset.

More Experiments on LLMs With the ever-growing number of publicly-available LLMs as well as commercial APIs enabling access to such models and allowing a more extensive prompt engineering effort, our dataset and its larger extensions can be used for a study on various LLM-based news headline event mapping methods.

5 Conclusion

In this paper, we defined the task of news headline event mapping and outlined a few use cases for the task in event monitoring, analysis, and forecasting solutions. We presented an approach for creating a benchmark dataset, and used it to create the first benchmark dataset for the evaluation of news headline event mapping methods. We used the benchmark to evaluate different classes of mapping methods, including a) zero-short classification based methods, b) adaptations of classic entity linking methods, and c) methods based on large generative language models. Our results provide interesting insights on the strengths and weaknesses of each of the methods. We outlined several avenues for future work, including our plan to extend the dataset, work on an ensemble method, and further experiments on LLM-based methods. Our dataset, as well as our evaluation script and outputs of the models, are publicly available on our GitHub repository.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146, Online. Association for Computational Linguistics.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, 472.

B. Chandrasekar, Bharath Ramesh, Vishalakshi Prabhu, S. Sajeev, Pratik K. Mohanty, and G. Shobha. 2017. Development of intelligent digital certificate fuzzer tool. In *Proceedings of the 2017 International Conference on Cryptography, Security and Privacy*, ICCSP '17, page 126–130, New York, NY, USA. Association for Computing Machinery.

Antonin Delpeuch. 2019. OpenTapioca: Lightweight entity linking for Wikidata. *CoRR*, abs/1904.09131.

Oktie Hassanzadeh, Parul Awasthy, Ken Barker, Onkar Bhardwaj, Debarun Bhattacharjya, Mark Feblowitz, Lee Martie, Jian Ni, Kavitha Srinivas, and Lucy Yip. 2022. Knowledge-based news event analysis and forecasting toolkit. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5904–5907. ijcai.org.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2020. Causal knowledge extraction through large-scale text mining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13610–13611. AAAI Press.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3023–3030, Virtual Event Ireland. ACM.

Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. 2022. Co-training improves prompt-based learning for large language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11985–12003. PMLR.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 107–110, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Steve Mbouadeu, Ken Barker, and Oktie Hassanzadeh. 2023. An Evaluation Framework for Mapping News Headlines to Event Classes in a Knowledge Graph. `https://github.com/mbouadeus/news-headline-event-linking`.

Sathappan Muthiah, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, Jose Cadena, Chang-Tien Lu, Anil Vullikanti, Achla Marathe, Kristen Summers, Graham Katz, Andy Doyle, Jaime Arredondo, Dipak K. Gupta, David Mares, and Naren Ramakrishnan. 2016. EMBERS at 4 years: Experiences operating an open source indicators forecasting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 205–214, New York, NY, USA. Association for Computing Machinery.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 909–918, New York, NY, USA. Association for Computing Machinery.

Mazhar Iqbal Rana, Shehzad Khalid, and Muhammad Usman Akbar. 2014. News classification based on their headlines: A review. In *17th IEEE International Multi Topic Conference 2014*, pages 211–216.

Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3141–3148, New York, NY, USA. Association for Computing Machinery.

S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, M. D. Feblowitz, and A. Riabov. 2019. IBM scenario planning advisor: Plan recognition as AI planning in practice. *AI Commun.*, 32(1):1–13.

Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4373–4379, Portorož, Slovenia. European Language Resources Association (ELRA).

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, Nitish Gupta, and Dan Roth. 2023. Event linking: Grounding event mentions to Wikipedia. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2679–2688, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.