Lifting News into a Journalistic Knowledge Platform

Tareq Al-Moslmi[a], Marc Gallofré Ocaña[a]

[a]University of Bergen, Fosswinckelsgt. 6, Postboks 7802, 5020 Bergen, Norway

Abstract

A massive amount of news is being shared online by individuals and news agencies, making it difficult to take advantage of these news and analyse them in traditional ways. In view of this, there is an urgent need to use recent technologies to analyse all news relevant information that is being shared in natural language and convert it into forms that can be more easily and precisely processed by computers. Knowledge Graphs (KGs) offer offer a good solution for such processing. Natural Language Processing (NLP) offers the possibility for mining and lifting natural language texts to knowledge graphs allowing to exploit its semantic capabilities, facilitating new possibilities for news analysis and understanding. However, the current available techniques are still away from perfect. Many approaches and frameworks have been proposed to track and analyse news in the last few years. The shortcomings of those systems are that they are static and not updateable, are not designed for large-scale data volumes, did not support real-time processing, dealt with limited data resources, used traditional lifting pipelines and supported limited tasks, or have neglected the use of knowledge graphs to represent news into a computer-processable form. Therefore, there is a need to better support lifting natural language into a KG. With the continuous development of NLP techniques, the design of new dynamic NLP lifters that can cope with all the previous shortcomings is required. This paper introduces a general NLP lifting architecture for automatically lifting and processing news reports in real-time based on the recent development of the NLP methods.

Keywords

Natural language processing (NLP), Journalistic knowledge platforms, Knowledge Graphs, Computational journalism, Stream data processing, Semantic technologies, Big data

1. Introduction

For several years we have seen how the traditional news press has moved to online content and new online press has appeared, publishing more online content than ever. Social networks enhanced that phenomenon facilitating real-time interactions and sharing, allowing pre-news to come to the surface, and bringing users with newer ways to digest news. Analysing news in real-time for supporting journalist work requires lifting those news to machine-understandable formats. Semantic representation of news using knowledge graphs is one of such formats that could be employed. Since news texts are expressed as natural language, there is a crucial need for processing and lifting these texts into a knowledge graph.

This paper presents an NLP lifting architecture component of the Journalistic Knowledge Platforms (JKP) for lifting natural language news text into knowledge graphs. JKP is a system intended for analysing, lifting, and representing news using knowledge graphs to support journalists exploiting knowledge from and about news being shared on the web and social media networks. JKPs have become crucial for press industry. Yet, many works have proposed to process the news texts in many different ways in order to apply different JKP processes.

Our group have been developing a series of JKP prototypes called News Hunter [1, 2, 3] in collaboration with a developer of newsroom tools for the international market. News Hunter moves forward the JKP to address the journalistic needs proposing a system to harvest real-time news stories from RSS feeds and social media, lifting news using SOTA approaches, and representing stories into knowledge graphs using Semantic Web standard technologies, Linked Open Data and NIF formats. News Hunter also explores detection and suggestion of news angles and exploitation of Semantic Web to support journalistic work [4, 5, 6, 7, 8].

Differently from previous works, our introduced NLP subsystem's architecture for News Hunter aims to lift all processed news into a semantic knowledge graph in real-time. Moreover, two Natural Language Processing (NLP) lifting tracks could be chosen: the traditional pipeline and the end-to-end which follows the state-of-the-art (SOTA) development of deep neural network. That would avoid some limitations reported in previous lifting tasks [9, 10].

The rest of the paper is organised as follows: Section 2 presents the background for our work. Section 3 introduced the general architecture of JKP. Section 4 constitutes the bulk of the paper and introduces the

Proceedings of the CIKM 2020 Workshops, October 19-20, Galway, Ireland.
EMAIL: Tareq.Al-Moslmi@uib.no (T. Al-Moslmi); Marc.Gallofre@uib.no (M. Gallofré Ocaña)
ORCID: 0000-0002-5296-2709 (T. Al-Moslmi); 0000-0001-7637-3303 (M. Gallofré Ocaña)

general NLP lifting process for real-time news lifting to a knowledge graph. Section 5 concludes the paper and outlines plans for future work.

2. Related Work

Current JKPs [11, 12, 13, 14, 15, 16] deal with big data multilingual text and multimedia sources of news-related items from which they have implemented their different NLP pipelines. These JPKs implemented NLP pipelines for lifting news into knowledge graphs and detect events normally by using traditional Named Entity Recognition (NER) and Named Entity Linking (NEL) systems, and pre-processed news text using linguistic techniques such as Part-of-Speech tagging (PoS), tokenisation, lemmatization and translation. In addition, NEWS project [11] used pattern matching to detect events, implemented NEL using PageRank and classified items, concepts and events using IPTC codes. NewsReader [13] used DBpedia Spotlight for NEL and mined opinion, causal, factual, temporal and semantic role information from news. ASRAEL [16] used SpaCy for NER, ADEL for NEL and Wikidata for linking events. SUMMA [15] used support vector machines (SVM) for NEL and classified topics from news. And both EvenRegistry [12] and SUMMA [15] used clustering techniques to detect events.

The NEWS project [17, 11] aimed to provide fresh multilingual information to news agencies (Spanish EFE and Italian ANSA agencies) analysing both textual and multimedia items. NEWS uses Ontology Ltd. (currently part of EXFO Nova Context real-time active topology platform[1] to implement the NLP pipeline to provide item categorization, concept representation, abstract generation, event recognition and NER using the ITPC codes. The NLP pipeline combines both linguistic techniques (patterns and rules such as PoS tagging) and traditional NER and NEL techniques (statistical techniques and PageRank). For recognizing events, NEWS project used pattern recognition techniques to describe and find the desired events.

The process of recognizing events is a relevant feature of such systems, which is approached in many different ways. For example, Event Registry [12] uses clustering algorithms to detect and group similar articles which represent the same event. Following the central idea of events, NewsReader project [13, 18] proposed a method, tools and a system to automatically leverage and represent events from news.

The NewsReader NLP pipeline performs language specific NER and NEL, event and semantic role detection and temporal relation detection over four different languages dealing and millions of news articles. The NLP pipeline processes each item starting with linguistic techniques (tokenizer, PoS, multiwords tagger), traditional NER and NEL (based on DBpedia Spotlight), opinion miner, semantic role labeler, event resolution, temporal recognizer and causal and factuality relation extraction. To overcome the large amount of news articles, NewsReader implemented its NLP pipeline using Big Data oriented technologies (i.e., Hadoop and Storm) into an scalable and real-time system [14].

Big data, multimedia and multilingual sources together are encountered in SUMMA project [15] which is an open-source platform for automated, scalable and distributed monitoring of real-time media broadcasts to support news agencies work like BBC or Deutsche Welle. The platform is built using big data-oriented technologies and services running in Docker[2] containers. SUMMA converts multimedia sources into text which is translated into English when found in other languages. Then, the text is processed through a NLP pipeline which classify them by topic using a hierarchical attention model, cluster them into storylines using clustering algorithms, and represent them using traditional NER (dependency parsing) and NEL (SVM-Ranking) techniques.

Likewise, the previous works ASRAEL project [16] uses knowledge graphs to represent events in news articles for searching purposes. To do so, they map AFP articles to Wikidata using NER (based on spaCy) and the NEL system ADEL.

As observed in the previous works there is a need for big data, real-time and semantic technologies approaches to deal with high volumes of news items that comes from multilingual and multimedia sources, and a common interest for detecting events among journalists and the different projects. Moreover, the proposed NLP techniques follow traditional approaches and similar pipelines which may not be always suitable for big data and real-time or for providing the best results.

Many approaches for lifting natural language to knowledge graphs are based on previous-generation NER techniques, and new lifting approaches that add disambiguation and linking to recent best-of-breed NE recognisers are needed . There is also a lack of standards for comparing lifting approaches[10]. This can partly be attributed to a lack of commonly accepted benchmarks, but it also a consequence of the recognition-disambiguation-linking pipeline. For ex-

[1] https://www.exfo.com/en/ontology/

[2] www.docker.com

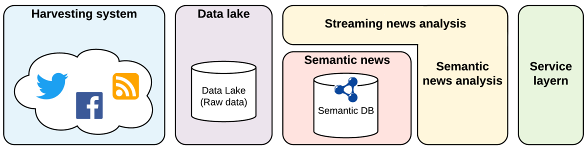

Figure 1: News Hunter architecture [2]

ample, it is hard to fairly compare pure NER with combined NER-NED-NEL techniques, when the latter is restricted to identifying named entities in the KB that is used for disambiguation and linking. Moreover, traditional sequential steps are now being integrated by joint learning or end-to-end processes. Consequently, mentions and entities that were previously analysed in isolation are now being lifted in each other's context. The current culmination of these trends are the deep-learning approaches that reported promising results recently. Most of those developments are not considered in previous works and this paper targets to cope with these gaps.

3. Journalistic Knowledge Platform architecture

In our previous work on News Hunter[2] we have proposed a general architecture for journalistic knowledge platforms (Figure 1) which is intended for big data real-time news lifting and processing. The still evolving architecture consists of 5 main parts: (1) The harvesting system which harvests the news from the web (e.g., RSS feeds, Facebook, Twitter) or daily produced in-house news (e.g., agency daily news activity) and its associated metadata (e.g., URL, source, author, ID, timestamp), and represents them using JSON in order to facilitate its parsing, transferring and simplifying it further processing. (2) The data lake or storage system for big data and real-time which is designed for sharing the news items across the different processes. (3) The semantic news component which contains the NLP lifter and the semantic DB (knowledge graph). (4) The semantic and streaming news analysis services, which due to the importance of social media can provide real-time analysis like trend monitoring, and event detection. (5) The service layer which allows users interact with the JKP.

News items can be collected from multiple sources: online news (e.g., RSS feeds), social media (e.g., Facebook, Twitter), archives or daily produced in-house news (e.g., agency daily news activity). The news crawler is oriented to harvest news from any source or multiple sources of interest. Due to the high amount of news items and their velocity of production, the harvested items are represented using standard lightweight formats like JSON, in order to facilitate its parsing, execution, transfer, sharing between components and temporal storage. News items are gathered together with its associated metadata (e.g., URL, source, author, ID, timestamp) which is included in the JSON files to benefit, speed and simplify its further processing and NLP tasks.

News items are processed according to their source: social media or news agencies . The news histories coming from news agencies (RSS feeds, news websties or archive) in JSON format are lifted into the knowledge graph as RDF triples using the NLP lifter, which can be adapted to the domain specific of the news history (e.g., economics, politics, sports). On the other hand, the news items coming from social media can be either pre-news (i.e., real-time information about events or something that is happening at the moment but not yet or incomplete as news histories) or small summaries/abstracts about news. Thus, identifying the topic they are related to and cluster them into groups of pre-news items that represent the same event and topic facilitates its processing. As these clusters of pre-news items represent a potential event with richer information that a single one item, they can be lifted using NLP techniques into the Knowledge Graphs.

Furthermore, as the social media items are potential real-time pre-news or events which can be breaking news, they are of highly importance for journalists. Yet, the clusters are analysed and monitored in order to find trends or breaking news events, that are reported in real-time to journalists.

In this paper, we are introducing the NLP lifting architecture that received the input from the harvester that have been explained previously[2]. The harvester is taking care of getting the data from different sources and standardise the data type into a unified format like JSON, XML, or NIF. The text can be stored in a big-data oriented databases such as Apache Cassandra [3] or HBase [4], which are oriented for distribution and large-scale processing pipelines. Moreover, the text can be distributed along the different NLP tasks using API or distribution framework like Kafka [5] or RabbitMQ [6]. The NLP liifter then has to deal with the data and lift it into a proper semantic format that will then be inserted to the KG.

[3]https://cassandra.apache.org
[4]https://hbase.apache.org
[5]https://kafka.apache.org
[6]https://www.rabbitmq.com

4. NLP lifter

This section describes the NLP lifter for news natural language texts to knowledge graphs. The NLP lifter which is a component of the JKP architecture consists of the main NLP lifting tasks as well as some additional related tasks. Differently from others proposed systems, our proposed NLP lifter is docker-based and contains the most possible tasks (traditional and recently developed ones) as shown in Figure 2. This allow the development of the platform and ensure using the most recent technology all the time. There will be two main NLP tracks: the traditional pipeline that is updated by recent technologies and the end-to-end track which is the SOTA in many tasks. In addition, there is the ensemble service that could combine more than one lifter to produce better results. The purpose and advantage of this is that the user can choose to use the most suitable track for his case and data as well as the most recent techniques. In the traditional pipeline the tasks like NER, NED, and NEL are implemented separately and mostly using the off-the-shelf software. The off-the-shelf systems are usually based on old approaches and their performance is not the SOTA. Moreover, traditional lifting methods neglect the relations between entity types and entity context. However, there will be a possibility in our introduced architecture to ensure the using of the most updated ones or using newest systems by just replacing or adding their dockers to the related component. The news item annotation ontology that has already been designed by[7] defines how the semantic annotations of news items should be represented in the knowledge graph. Each harvested news item is associated with one or more annotations, which may be, for example, named entities, concepts, topics, times or geolocations or relations between annotations. The ontology also describes how the sources of news items and annotations are represented in the knowledge graph to maintain provenance[7]. We describe the general NLP lifter components as the following:

4.1. Pre-processing

The quality of the data plays a key role in determining the suitable pre-processing techniques. Since we are dealing with the real-time streaming, the cleaning and normalization are required to remove unnecessary or noisy terms (like ASCII codes, currency symbols, hashtags, and so forth). The most frequently used pre-processing techniques are tokenization and POS tagging [19, 20]. Other common steps are sentence splitting, lemmatisation, chunking and dependency parsing, and structural parsing. Recent works indicate that robust lifting systems require accurate tuning of several steps, especially tokenization and semantic similarity [21]. Recently, deep neural networks, especially end-to-end methods, have reduced the need for pre-processing steps. Moreover, using deep neural networks for pre-processing tasks such as tokenization has recently produced promising results [22]. The proposed NLP lifter could include as many pre-processing steps as possible, which will be in separate dockers, so the user can choose all suitable ones for the target data.

4.2. Named entity recognition

Named entity recognition is the task that identifies the named entities contained in the text like persons, locations, organizations, time, date, money, etc. NER approaches could be categorised into three main groups: knowledge-based approaches, learning-based methods, and feature-inferring neural network methods. Despite the existence of recent SOTA NER results (especially recent deep NN approaches) such as [23, 24, 25, 26], these approaches have not been utilized and exploited in the process of lifting natural language to knowledge graphs as mentioned earlier. This paper aims to implement those SOTA NER methods in docker-based components to tackle this shortcoming.

4.3. Named entity linking

NEL annotates each mention in a text with the identifier of its corresponding entity that is described in a KB in the LOD cloud. Our paper has defined NEL as a wider task that includes NED as one of its processes. Many NEL approaches are utilizing off-the-shelf systems for NER task. It is, however, a challenging task to choose which particular model to use for those systems. That is because it requires to estimate the similarity level between the system's training datasets and the dataset that needs to be processed in which we strive to accurately recognize entities, according to [27]. Most recent SOTA systems on AIDA-CoNLL dataset includes [28, 29, 30, 31]. There is no perfect NEL model for all datasets and one model might be the best on one dataset but perform poorly on others. Accordingly, having the top N best SOTA implemented in dockers will allow the user to pick the most suitable model for his data and/or replace or update them at any time when needed.

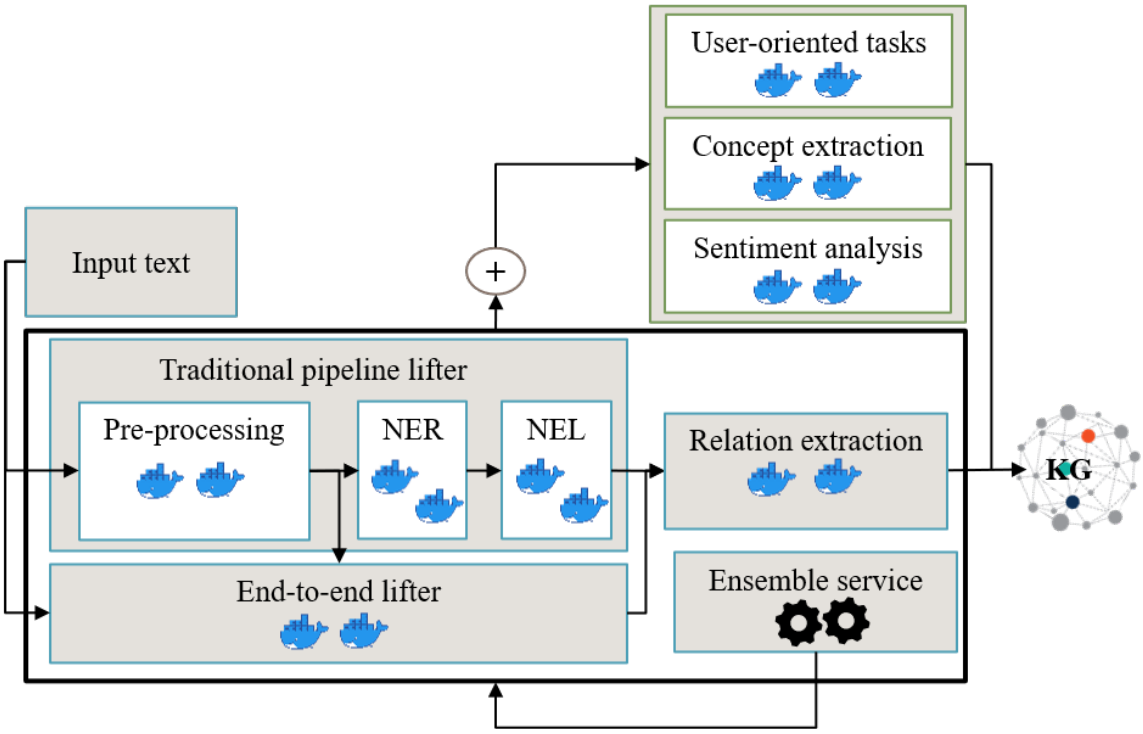

Figure 2: General NLP lifting architecture

4.4. End-to-end track

The majority of previous studies were mostly assuming the availability of mentions and entities and focused on the disambiguation process only. However, leveraging mutual dependency between mentions and their entities is neglected. Moreover, it is not a practical idea in a real-world application. Different from that and to overcome those shortcomings, end-to-end deals with row text and aims to extract all mentions and link them to their entities in the knowledge base. End-to-end entity linking has been recently proposed and is receiving increasing attention. Few studies have been published which claiming the application of the end-to-end approach [32, 33, 34, 35]. The most interesting ones are the most recent neural-based end-to-end linking models [36, 37, 38, 39]. One of the most recent SOTA is [38] followed by [36]. Our NLP lifter aims at including such techniques as an alternative recent track for lifting news texts into a semantic knowledge graph.

4.5. Relation and concept extraction

Our NLP lifter aims at covering lifting of general concepts and of relations between entities. Many recent approaches also lift relations jointly with entities (both named entities and concepts) and reported the SOTA results. Similar to previous components, the proposed lifter will implement those methods and include them as optional tasks as many others for the user.

4.6. User-oriented tasks

User-oriented tasks include those tasks specific and personalised for the project where the NLP lifting architecture is implemented. Apart from including SOTA NLP tasks like the previously described, the NLP lifting architecture takes into account purpose specific tasks such as news angles detection, event detection, IPTC media codes annotation, rumours detection and text completion.

4.7. Knowledge graph

In a knowledge graph, the nodes represent either concrete objects, concepts, information resources, or data about them, and the edges represent semantic relations between the nodes [40]. Knowledge graphs thus offer a widely used format for representing information in computer-processable form. They build on, and are heavily inspired by, Tim Berners-Lee's vision of the semantic web, a machine-processable web of data that augments the original web of human-readable documents [41]. Knowledge graphs can therefore leverage existing standards such as RDF, RDFS, and OWL. Moreover, the constructed knowledge graph could be used to implement more operations like question answering, knowledge graph-based sentence auto-completion, storytelling, fact-checking and so forth using semantic news analysis.

5. Conclusion

Lifting high-volume streams of news texts involves representing their content in machine-understandable formats. KGs is one such formats that has received much attention recently. NLP lifters are an important prerequisite for making the abundance of natural language news on the internet available as computer-processable knowledge graphs. Thus, the presented NLP lifting pipeline provides with an structured and formalised process for transforming natural language text into computer-processable knowledge graphs. The presented pipeline can incorporate any NLP technique like traditional or end-to-end approaches and combining its results or expand them with specific-purpose NLP method like sentiment analysis. Moreover, the introduced NLP lifter is designed to simplify its components replaceability by making use of docker technology, facilitating e.g., the update of all tasks and methods to SOTA approaches. Although the proposed JKP is designed mainly to help journalists, it could be used and customized for the public. The presented NLP lifting architecture aims to be used as reference for developers and researchers of JKP interested in real-time NEL. News organisations may need to adapt their systems, replace components, add new SOTA technologies, or integrate it with other JKP, thus having such NLP lifting pipeline as reference facilitate its management and understanding. Furthermore, it is not restricted to news text and could be used to lift other types of texts.

In our future work, we plan to validate the results of our proposed NLP lifter by using both a manually collected and annotated corpus of news and gold-standards, and compare the results of our proposed lifter with current NEL systems such as ADEL, SpaCy lifter, NewsReader, Stanford CoreNLP or DBpedia Spotlight. Besides, we want to explore the possibilities that validations tools like GERBIL [42] can provide when applied inside our NLP lifter. We believe that validation tools can provide insights about the evolution and performance of the applied NLP processes which can be incorporated to reinforce, improve and keep updated the NLP models.

Acknowledgments

Supported by the News Angler project funded by the Norwegian Research Council's IKTPLUSS programme as project 275872.

References

[1] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, News hunter: building and mining knowledge graphs for newsroom systems, in: NOKOBIT, volume 26, 2018.

[2] M. Gallofré Ocaña, L. Nyre, A. L. Opdahl, B. Tessem, C. Trattner, C. Veres, Towards a big data platform for news angles, in: 4th Norwegian Big Data Symposium (NOBIDS) 2018, 2018.

[3] A. Berven, O. Christensen, S. Moldeklev, A. Opdahl, K. Villanger, A knowledge graph platform for newsrooms, Computers in Industry (2020). To appear.

[4] B. Tessem, A. L. Opdahl, Supporting journalistic news angles with models and analogies, in: 2019 13th RCIS, IEEE, 2019, pp. 1–7.

[5] A. L. Opdahl, B. Tessem, Towards ontological support for journalistic angles, in: Enterprise, Business-Process and Information Systems Modeling, Springer International Publishing, 2019, pp. 279–294.

[6] B. Tessem, Analogical news angles from text similarity, in: Artificial Intelligence XXXVI, Springer International Publishing, 2019, pp. 449–455.

[7] A. L. Opdahl, B. Tessem, Ontologies for finding journalistic angles, Software and Systems Modeling (2020) 1–17.

[8] E. Motta, E. Daga, A. L. Opdahl, B. Tessem, Analysis and design of computational news angles, IEEE Access (2020).

[9] M. Albared, M. Gallofré Ocaña, A. Ghareb, T. Al-Moslmi, Recent progress of named entity recognition over the most popular datasets, in: 2019 First International Conference of Intelligent Computing and Engineering (ICOICE), 2019, pp. 1–9.

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

[11] N. Fernández, D. Fuentes, L. Sánchez, J. A. Fisteus, The news ontology: Design and appli-

cations, Expert Systems with Applications 37 (2010) 8694 – 8704.

[12] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: learning about world events from news, in: Proceedings of the 23rd WWW, ACM, 2014, pp. 107–110.

[13] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, M. Rospocher, R. Segers, Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news, Knowledge-Based Systems 110 (2016) 60 – 85.

[14] M. Kattenberg, Z. Beloki, A. Soroa, X. Artola, A. Fokkens, P. Huygen, K. Verstoep, Two architectures for parallel processing of huge amounts of text, in: Proceedings of the Tenth LREC'16), European Language Resources Association (ELRA), 2016, pp. 4513–4519.

[15] U. Germann, P. v. d. Kreeft, G. Barzdins, A. Birch, The summa platform: Scalable understanding of multilingual media, in: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 2018.

[16] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, X. Tannier, Searching news articles using an event knowledge graph leveraged by Wikidata, in: 30th WWW Conference, 13-17 May 2019, 2019.

[17] N. Fernández, J. M. Blázquez, J. A. Fisteus, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, Z. Ben-Asher, News: Bringing semantic web technologies into news agencies, in: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. M. Aroyo (Eds.), The Semantic Web - ISWC 2006, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 778–791.

[18] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, Journal of Web Semantics 37-38 (2016) 132 – 151.

[19] G. Zhu, C. A. Iglesias, Exploiting semantic similarity for named entity disambiguation in knowledge graphs, Expert Systems with Applications 101 (2018) 8 – 24.

[20] M. Fossati, E. Dorigatti, C. Giuliano, N-ary relation extraction for simultaneous t-box and a-box knowledge base augmentation, Semantic Web 9 (2018) 413–439.

[21] M. Conover, M. Hayes, S. Blackburn, P. Skomoroch, S. Shah, Pangloss: Fast entity linking in noisy text environments, in: Proceedings of the 24th ACM SIGKDD, KDD '18, ACM, 2018, p. 168–176.

[22] T. Boros, S. D. Dumitrescu, R. Burtica, NLP-cube: End-to-end raw text processing with neural networks, in: Proceedings of the CoNLL 2018, ACL, 2018, pp. 171–179.

[23] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli, Cloze-driven pretraining of self-attention networks, in: Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP, ACL, 2019, pp. 5359–5368.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[25] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the NAACL, ACL, 2018, pp. 2227–2237.

[26] L. Liu, X. Ren, J. Shang, X. Gu, J. Peng, J. Han, Efficient contextualized representation: Language model pruning for sequence labeling, in: Proceedings of the 2018 Conference on EMNLP, ACL, 2018, pp. 1215–1225.

[27] J. Plu, G. Rizzo, R. Troncy, Enhancing entity linking by combining ner models, in: H. Sack, S. Dietze, A. Tordai, C. Lange (Eds.), Semantic Web Challenges, Springer International Publishing, Cham, 2016, pp. 17–32.

[28] J. Raiman, O. Raiman, Deeptype: Multilingual entity linking by neural type system evolution, 2018. arXiv:1802.01021.

[29] I. Yamada, H. Shindo, Pre-training of deep contextualized embeddings of words and entities for named entity disambiguation, 2019. arXiv:1909.00426.

[30] Z. Fang, Y. Cao, Q. Li, D. Zhang, Z. Zhang, Y. Liu, Joint entity linking with deep reinforcement learning, in: The WWW Conference, WWW '19, ACM, 2019, p. 438–447.

[31] A. Luo, S. Gao, Y. Xu, Deep semantic match model for entity linking using knowledge graph and text, Procedia Computer Science 129 (2018) 110 – 114. 2017 International Conference on Identification, Information and Knowledge in the Internet of Things.

[32] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Transactions of the ACL 2 (2014) 231–244. arXiv:10.1162/tacl_a_00179.

[33] O.-E. Ganea, M. Ganea, A. Lucchi, C. Eickhoff, T. Hofmann, Probabilistic bag-of-hyperlinks model for entity linking, in: Proceedings of the 25th WWW, WWW '16, WWW Conference, 2016, p. 927–938.

[34] D. B. Nguyen, M. Theobald, G. Weikum, J-nerd: Joint named entity recognition and disambiguation with rich linguistic features, Transactions of the ACL 4 (2016) 215–229. arXiv:10.1162/tacl_a_00094.

[35] O.-E. Ganea, T. Hofmann, Deep joint entity disambiguation with local neural attention, in: Proceedings of the 2017 Conference on EMNLP, ACL, 2017, pp. 2619–2629.

[36] N. Kolitsas, O.-E. Ganea, T. Hofmann, End-to-end neural entity linking, in: Proceedings of the 22nd Conference on Computational Natural Language Learning, ACL, 2018, pp. 519–529.

[37] Y. Cao, L. Hou, J. Li, Z. Liu, Neural collective entity linking, 2018. arXiv:1811.08603.

[38] P. Le, I. Titov, Improving entity linking by modeling latent relations between mentions, in: Proceedings of the 56th ACL, ACL, 2018, pp. 1595–1604.

[39] P. H. Martins, Z. Marinho, A. F. T. Martins, Joint learning of named entity recognition and entity linking, in: Proceedings of the 57th ACL, ACL, 2019, pp. 190–196.

[40] D. Allemang, J. Hendler, Semantic Web for the Working Ontologist, second edition ed., Morgan Kaufmann, 2011.

[41] T. Berners-Lee, J. Hendler, O. Lassila, et al., The semantic web, Scientific american 284 (2001) 28–37.

[42] M. Röder, R. Usbeck, A. N. Ngomo, GERBIL - benchmarking named entity recognition and linking consistently, Semantic Web 9 (2018) 605–625.