

Online News Event Extraction for Crime Analysis

Federica Rollo¹, Laura Po¹ and Giovanni Bonisoli¹

¹*"Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Via Vivarelli 10, Modena, Italy*

Abstract

Event Extraction is a complex and interesting topic in Information Extraction that includes methods for the identification of event's type, participants, location, and date from free text or web data. The result of event extraction systems can be used in several fields, such as online monitoring systems or decision support tools. In this paper, we introduce a framework that combines several techniques (lexical, semantic, machine learning, neural networks) to extract events from Italian news articles for crime analysis purposes. Furthermore, we concentrate to represent the extracted events in a Knowledge Graph. An evaluation on crimes in the province of Modena is reported.

Keywords

crime analysis, NLP, word embeddings, question answering, localization, deduplication

1. Introduction

In most recent years, the idea of Safe City is spreading. It is related to the strategies that aim to help the government in the development of a city security system for reducing the possibility of crime and providing an environment where people feel safe and comfortable [1]. A sound analysis of the crime events distribution, i.e., detecting where and when crimes occur and identifying the causes, is paramount to implement a Safe City. Crime analysis [2] is not merely crime events counting; it is an in-depth examination of the different criminogenic factors (e.g., time, place, socio-demographics) that helps to understand why the crime occurs. It consists of systematic, analytical processes for providing timely and pertinent information related to crime patterns and trend correlations to assist the police in crime reduction, prevention, and evaluation. Data-driven policing and associated crime analysis are still dawning. The use of Geographic Information System (GIS) techniques in crime mapping helps crime analysis and allows localizing crimes to identify the high-risk areas. Several countries provide statistics on crime, but they are often available with some delay. The delay between the occurrence of the event and the report publication can reach some days, months or even years. In most cases, they are provided as aggregated data, not as single crime events. Moreover, police reports are usually private documents, as in Italy. Therefore, although they are very useful documents, police reports cannot be considered a possible source for timely crime analysis for citizens. The reports of the Italian National Institute of Statistics (ISTAT) provide a clear picture of the types of crime happen in each province during the year, however, the information provided is aggregated by time and space and become available after (at least) one year from the crime event happening. In those cases, newspapers are a valuable source of authentic and timely information. Extracting crime events from news articles published on the web by local newspapers can help overcome


SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

✉ federica.rollo@unimore.it (F. Rollo); laura.po@unimore.it (L. Po); giovanni.bonisoli@unimore.it (G. Bonisoli)

🆔 0000-0002-3834-3629 (F. Rollo); 0000-0002-3345-176X (L. Po); 0000-0001-8538-8347 (G. Bonisoli)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

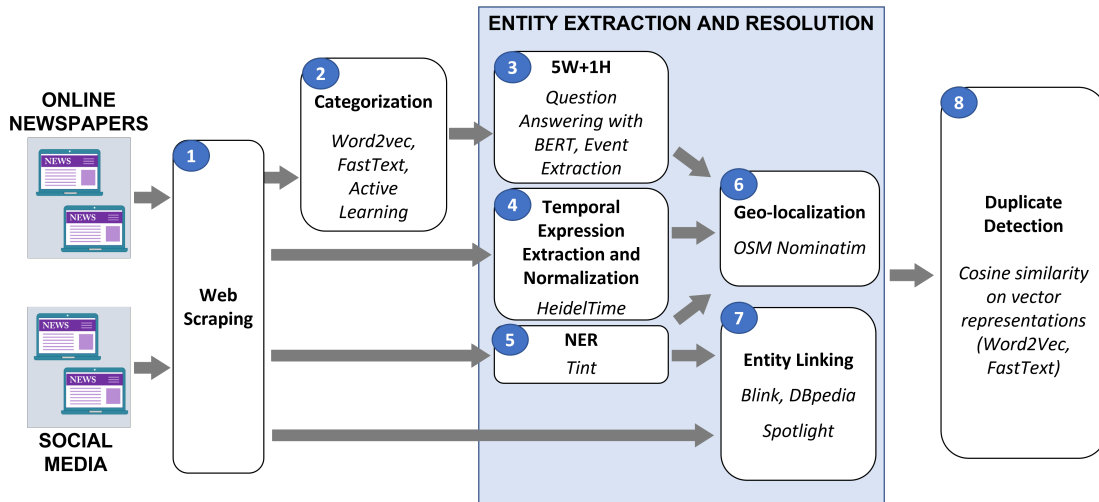


Figure 1: The pipeline of the Crime Analysis framework.

the lack of crime up-to-date information [3, 4, 5, 6, 7]. Detailed information about the crime events can be extracted through Natural Language Processing (NLP) techniques applied to the news articles' text. Newspapers provide reliable, localized, and timely information (the time delay between the occurrence of the event and the publication of the news does not exceed 24/48 hours). The main drawback is that newspapers do not collect and publish all the facts related to crimes, but only the ones that arouse the readers' interest. Therefore, a percentage of police reports will not be turned into news articles and is lost.

The scope of this paper is to describe a framework to extract crime data from news articles, enrich them with semantic information and provide useful visualization. The strategy employs several techniques and extends a previous work [8]: crime categorization, named entity extraction, 5W+1H extraction, linked data mapping, geo-localization, time expression normalization, entity linking and duplicate detection. The novelty of such a framework is the integration of multiple techniques, previously used in different contexts, for solving various sub-problems into a common framework for crime analysis. Moreover, the framework transforms texts contained in news articles into a Crime knowledge graph that accurately describes and links the crimes. The framework has been tested successfully on news articles related to the city of Modena. However, it can be adapted to manage data of other cities or areas.

The outline of the paper is the following: in Section 2 the pipeline of the framework is presented, while Section 3 is devoted to the description of the use case in the province of Modena. Finally, Section 4 depicts some conclusions.

2. Crime Analysis framework

The pipeline of the framework to extract semantic information related to an event starting from news articles published on the web and alerts shared on social media consists of 8 phases. The phases should be executed mainly in sequence for each news article (except for some phases where the execution can be run in parallel); in any case, different news articles can be processed in parallel. The entire process is executed periodically to extract the latest published

news articles, analyze them and add information to the knowledge base (KB). The frequency of execution depends on the need of having a real-time up-to-date KB and how often the selected online newspapers publish news articles. Figure 1 illustrates the phases, in bold, and some techniques and tools, in italic:

1. Data extraction is performed by harvesting online newspapers and social media (*web scraping*). The content of each news article is labeled, structured and semantically annotated [9, 10]. Some web content may already expose a predefined structure, i.e., HTML pages encoded with the Document Object Model (DOM), and some libraries allow accessing the data encapsulated into HTML tags;¹ if this is not the case, other methods can be used, such as RSS Feed, API, and so on.
2. The categorization of the event is crucial to map a news article w.r.t. a type of event (business, sports, crime, politics, arts, culture, etc.). Given some pre-categorized news articles, i.e., annotated training data, machine learning algorithms can be applied to uncategorized news articles to assign them a type of event [11, 12]. Word embeddings can be exploited to extract the vector representations of the news articles, then, classifiers can take in input such representations to assign a category to each news article. Moreover, active learning can be used to enhance the quality of categorization retraining the classification model on the original dataset enriched with high-confidence categorized news articles. Other approaches can exploit topic detection algorithms [13, 14].
3. The identification and extraction of the 5W+1H (*What, When, Where, Who, Why, How*) might be performed by employing Event Extraction models or through the Question Answering task using BERT (Bidirectional Encoder Representations from Transformers) by adopting different questions according to the type of event [15, 16]. The 5W+1H are the questions that a reporter must answer through the reporting. Therefore, these are the essential elements of any news and also contribute to improve the value of news and newsworthiness in journalism.
4. By analyzing the news article's body, temporal expressions can be identified (for example, words like "two days ago", "this morning") and then normalized in date format. This operation allows identifying the exact date of the event, taking into account the date of publication.
5. The Named Entity Extraction (NER) is applied to the text of the news articles to identify the reference to persons, organizations, places, and temporal expressions and can be executed in parallel with the second phase. Its results can intersect the output of the 5W+1H phase.
6. With the Entity Linking, the entities identified in phase 5, such as persons, organizations, and locations are linked to resources (URI) available in Linked Datasets. For example, DBpedia Spotlight [17] can be used to link to resources of DBpedia and Linked Geo Data. Besides, an Italian version of Blink² [18] can be used to link entities to Wikipedia or to populate a new KB with entities not linked to external resources.
7. The geographical localization exploits the entities that have been identified as locations in phase 5 or as answers to "where" in phase 3 and processes them to be geo-referenced.

¹An example is the Java HTML Parser named *jsoup*.

²<https://github.com/rpo19/BLINK>

In case a location is not specified in the news article, organizations (identified in phase 4) can also be exploited to geolocate the event.

8. The identification of duplicates or storylines aims to find the same event described in more news articles, this might occur also within the same newspaper where updates about one event are published over time. To avoid too many comparisons among news, it is possible to identify candidate duplicates and apply text similarity analysis methodology to these candidates. In the end, the information of duplicates can be merged.

The use of semantic technologies is a key point in the presented approach for detecting events from news articles and enriching them with information automatically extracted from the text.

3. Crime Knowledge Graph for Modena

The Crime Analysis framework has been applied to a collection of news articles related to the crimes that occurred in the province of Modena. We select two major newspapers that publish on average 850 news articles per year related to crimes in the Modena province and cover the 95% of the total news articles published in Modena newspapers. The framework collected 17,500 reports from June 2011 to December 2021 (approximately 10 years) and is currently running to analyze news articles published every day by two local newspapers. On 17,500 reports, the framework was able to geolocalize almost 100% of the crime events and normalize the time expressions on 83% of the news articles. The results produced allow performing crime mapping studies and the identification of crime hot spots in semi real-time: visualizations of these results are shown through the “Modena Crime” web application.³

Figure 2 shows an example of Knowledge Graph generation, geo-localization and duplicate detection of an Italian news article reporting a theft. The news article is derived from the translation from Italian to English of the news taken from the “Gazzetta di Modena” newspaper.⁴ The 5W+1H are extracted from the text and reported in the event-centric Knowledge Graph. The central node (the one colored in red in Figure 2) identifies the event, while the other nodes report information related to the 5W+1H. The time reference (“last Thursday”) and the publication date (“12 March 2022”) are used to identify the date of the event, while the entities categorized as locations are exploited by OSM Nominatim to find the GPS coordinates where the theft occurred. Then, the coordinates are used to represent the event on the map. Thanks to the duplicate detection algorithm, two news articles are identified as duplicates of the examined news article. They are follow-up news since they report updates on the theft. The publication date is used to build the storyline of the event.

We created a Crime Knowledge Graph for 1246 thefts that occurred in Modena in 2020 using the Neo4j tool. An analysis of the interconnections between the crimes has been conducted using centrality algorithms that determine, on the basis of the graph topology, the importance of the individual nodes, and community detection algorithms to distinguish groups of nodes within the overall graph. Each crime event can share some connected nodes with other events, such as the place where the theft happened or the stolen objects, etc. We added direct relationships among the crime nodes to represent that they share some connected nodes. Using Pagerank, we classified the crime nodes based on their importance in the graph. The higher the Pagerank

³<https://dbgroup.ing.unimo.it/modenacrime>

⁴<https://shorturl.at/vKSX2>

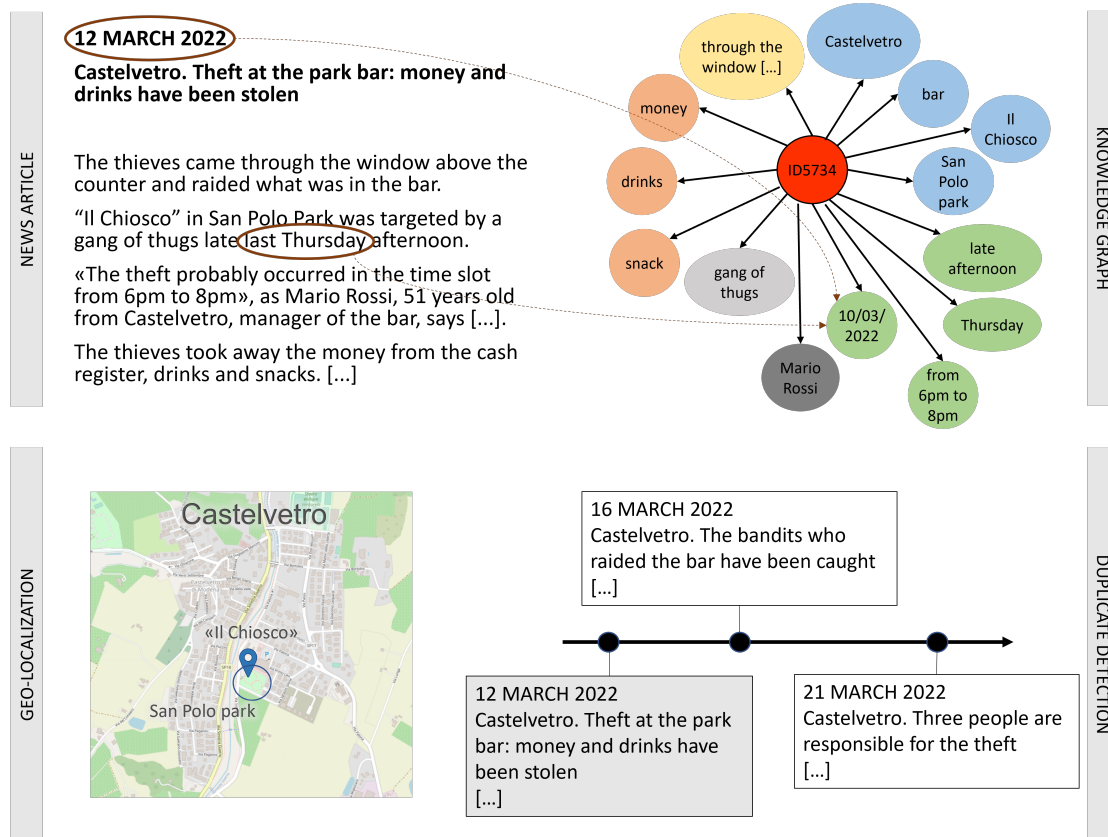


Figure 2: Example of news article with the corresponding knowledge graph, the geo-localization and the timeline with the duplicates.

value of a node, the greater the connections of the event node with the other event nodes. For example, thefts in which gold and valuables are stolen occur more frequently, and therefore events reporting such stolen items are strongly connected. To detect the communities, we used the label propagation algorithm on 5 different subgraphs obtained by examining the 5W+1H relationships separately. For the *Where* subgraph, the result highlighted communities of nodes sharing several locations, i.e., WHERE nodes. These first experiments provide some insights on how Crime Analysis can benefit from graph-based methods.

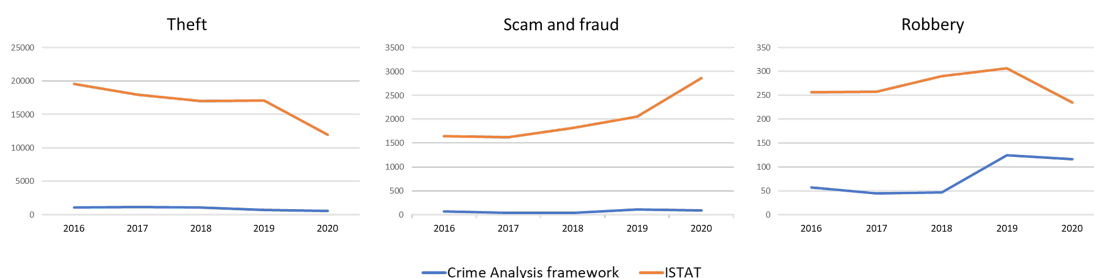


Figure 3: Distribution of crime reports from 2016 to 2020 in the province of Modena.

3.1. Impact and scalability

To evaluate the impact of the proposed framework, the number of crimes collected by the framework and the number of crimes published in the official report of ISTAT (i.e., the crimes reported to the police) have been compared. The report related to the period from 2016 to 2020⁵ has been taken into account. The information is only quantitative, the types of crime are reported per province and no information about where and when, during the year, the crime happened is provided. For providing a comparison between the two datasets, only the crime categories in common have been taken into account. Unfortunately, a location-based comparison is not possible because ISTAT provides a unique report for the entire province. With the total number of crimes in the city of Modena of 9590 from 2016 to 2020, the built KB covers around 10% of the crimes reported by ISTAT. A hypothesis on this low coverage can be attributed to the fact that not all the criminal events recorded by ISTAT, and therefore in the police reports, are of high impact and public interest. Therefore, not all of them are reported in local news articles. The most frequent crimes in both datasets are thefts each year from 2016 to 2020. Figure 3 shows the total number of the top three types of crimes recorded in the report of ISTAT and compared to the number collected by the framework. As can be seen, the lower coverage is reported in scams and frauds (the percentage is between 2% and 5% each year), while the higher one is in robbery (up to 50% in 2020).

Even if the approach has been applied in a medium area, it highlights its potentiality. In Italy, it is not possible to collect real-time crime information from official sources, since official criminal statistics are reported annually with a delay of 6 months. The proposed approach can be applied everywhere, also in small or medium cities/areas, since there will be always one or more newspapers that report the main crimes to happen in that place.

A first scalability test has been executed to ingest all the news articles related to crimes that happened in the entire Emilia-Romagna region. Other 9 newspapers which publish news related to the 9 provinces of the Emilia-Romagna region were selected. All the available news articles, from 2011 till now, which refer to 11 crime types have been collected. The total number of news articles is 35,000 (on average 3,900 news articles for each province). The crime ingestion can be run in parallel for different newspapers and different crime types. Therefore, 99 ingestion processes have been executed in parallel to extract, analyze and store data of the region. The total loading time, which depends on the loading time of the province with the higher number of news articles from 2011, is 3 hours for 35,000 news articles.⁶

4. Conclusion and Future Work

The framework presented in this paper is able to extract crime data from news articles, enrich them with semantic information and provide a Knowledge Graph that can be exploited for further analysis. It suggests multiple techniques for solving various sub-problems: extracting crime events from news articles, geo-locating them, linking entities to Linked Data resources, and detecting duplicates. The framework has been successfully employed in the province of Modena and has allowed collecting a consistent dataset of more than 17,500 news articles about

⁵<http://dati.istat.it/Index.aspx?QueryId=25097&lang=en>

⁶The test has been performed on a Microsoft Windows 10 Pro, 16GB RAM, Processor Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz, 2208 Mhz, 6 Cores, 12 Logical Processors.

13 types of crimes. A comparison with the official crime reports provided by ISTAT unveil that this approach has allowed collecting about 10% of the crime events. This can be considered a satisfactory result since we are aware that news articles do not cover all the crimes that happen in a city. The approach is domain-independent; it can be applied to any kind of news article, not only crime news, and can also be adapted to other languages.

In future work, we will work on the definition of a crime ontology to describe the crime events. In addition, Neo4j will be deepened to better analyze the Crime Knowledge Graph.

Acknowledgments

This work is partially supported by the project “Deep Learning for Urban Event Extraction from News and Social media streams” founded by the Engineering Department “Enzo Ferrari” of the University of Modena and Reggio Emilia.

References

- [1] J. Ristvej, M. Lacinák, R. Ondrejka, On smart city and safe city concepts, *Mob. Networks Appl.* 25 (2020) 836–845. doi:10.1007/s11036-020-01524-4.
- [2] G. Oatley, B. Ewart, Data mining and crime analysis, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 1 (2011) 147–153. doi:10.1002/widm.6.
- [3] S. K. P. S. Thilagam, Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers, *Information Processing and Management* 56 (2019). doi:10.1016/j.ipm.2019.102059.
- [4] L. Po, F. Rollo, Building an urban theft map by analyzing newspaper crime reports, in: 13th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP Zaragoza, Spain, 2018, pp. 13–18. doi:10.1109/SMAP.2018.8501866.
- [5] T. Dasgupta, A. Naskar, R. Saha, L. Dey, Crime profiler: Crime information extraction and visualization from news media, in: A. P. Sheth, A. Ngonga, Y. Wang, E. Chang, D. Slezak, B. Franczyk, R. Alt, X. Tao, R. Unland (Eds.), *Proceedings of the International Conference on Web Intelligence, WI '17*, Association for Computing Machinery, New York, NY, USA, 2017, p. 541–549. doi:10.1145/3106426.3106476.
- [6] P. R. Boppuru, R. Kenchappa, Geo-spatial crime analysis using newsfeed data in indian context, *Int. J. Web Based Learn. Teach. Technol.* 14 (2019) 49–64. doi:10.4018/IJWLTT.2019100103.
- [7] V. Sharma, R. Kulshreshtha, P. Singh, N. Agrawal, A. Kumar, Analyzing newspaper crime reports for identification of safe transit paths, in: R. Mihalcea, J. Y. Chai, A. Sarkar (Eds.), *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, USA, May 31 - June 5, 2015, The Association for Computational Linguistics, 2015, pp. 17–24. doi:10.3115/v1/n15-2003.
- [8] F. Rollo, L. Po, Crime event localization and deduplication, in: J. Z. Pan, V. A. M. Tamma, C. d’Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference*, Athens, Greece, November 2-6, 2020, *Proceedings, Part II*, volume 12507 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 361–377. doi:10.1007/978-3-030-62466-8_23.

- [9] S. Bergamaschi, D. Beneventano, L. Po, S. Sorrentino, Automatic normalization and annotation for discovering semantic mappings, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6585 LNCS (2011) 85 – 100. doi:10.1007/978-3-642-19668-3_8.
- [10] R. Trillo, L. Po, S. Ilarri, S. Bergamaschi, E. Mena, Using semantic techniques to access web data, *Information Systems* 36 (2011) 117 – 133. doi:10.1016/j.is.2010.06.008.
- [11] G. Bonisoli, F. Rollo, L. Po, Using word embeddings for italian crime news categorization, in: M. Ganzha, L. A. Maciaszek, M. Paprzycki, D. Slezak (Eds.), *Proceedings of the 16th Conference on Computer Science and Intelligence Systems*, Online, September 2-5, 2021, 2021, pp. 461–470. doi:10.15439/2021F118.
- [12] F. Rollo, G. Bonisoli, L. Po, Supervised and unsupervised categorization of an imbalanced italian crime news dataset, in: E. Ziemba, W. Chmielarz (Eds.), *Information Technology for Management: Business and Social Issues*, Springer International Publishing, Cham, 2022, pp. 117–139.
- [13] F. Rollo, A key-entity graph for clustering multichannel news: student research abstract, in: *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, 2017*, pp. 699–700. doi:10.1145/3019612.3019930.
- [14] L. Po, F. Rollo, R. T. Lado, Topic detection in multichannel italian newspapers, in: A. Cali, D. Gorgan, M. Ugarte (Eds.), *Semantic Keyword-Based Search on Structured Data Sources - COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8-9, 2016, Revised Selected Papers, Lecture Notes in Computer Science, 2016*, pp. 62–75. doi:10.1007/978-3-319-53640-8_6.
- [15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.
- [16] C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, M. Iyyer, BERT with history answer embedding for conversational question answering, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, ACM, 2019, pp. 1133–1136. doi:10.1145/3331184.3331341.
- [17] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: shedding light on the web of documents, in: C. Ghidini, A. N. Ngomo, S. N. Lindstaedt, T. Pellegrini (Eds.), *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011, ACM International Conference Proceeding Series, ACM, 2011*, pp. 1–8. doi:10.1145/2063518.2063519.
- [18] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020*, pp. 6397–6407. doi:10.18653/v1/2020.emnlp-main.519.