WTF-LOD – A New Resource for Large-Scale NER Evaluation

Lubomir Otrusina and Pavel Smrz
Brno University of Technology, Faculty of Information Technology, IT4Innovations Centre of Excellence
Bozetechova 2, 612 66 Brno, Czech Republic
{iotrusina,smrz}@fit.vutbr.cz

Abstract

This paper introduces the Web TextFull linkage to Linked Open Data (WTF-LOD) dataset intended for large-scale evaluation of named entity recognition (NER) systems. First, we present the process of collecting data from the largest publically-available textual corpora, including Wikipedia dumps, monthly runs of the CommonCrawl, and ClueWeb09/12. We discuss similarities and differences of related initiatives such as WikiLinks and WikiReverse. Our work primarily focuses on links from "textfull" documents (links surrounded by a text that provides a useful context for entity linking), de-duplication of the data and advanced cleaning procedures. Presented statistics demonstrate that the collected data forms one of the largest available resource of its kind. They also prove suitability of the result for complex NER evaluation campaigns, including an analysis of the most ambiguous name mentions appearing in the data.

Keywords: named entity evaluation, linked open data, CommonCrawl, ClueWeb, Wikipedia

1. Introduction

The need to bridge the semantic gap between the semi-structured "web of documents" and the structured "web of knowledge" (Buitelaar and Cimiano, 2008) has led to the development of various semantic enrichment systems in recent years. Named entity (NE) recognition and linking present a key component of semantic enrichment. Tools such as DBpedia Spotlight (Daiber et al., 2013), Illinois Wikifier (Ratinov et al., 2011), or AIDA (Hoffart et al., 2011) enable annotating mentions of NEs in a plain text and "anchoring" the annotations in linked open data resources (most frequently in DBpedia/Wikipedia).

Various evaluation campaigns have also appeared that compare quality of the NE annotation on collected datasets. Initiatives such as NIST TAC KBP[1] – Knowledge Base Population – Entity Discovery and Linking Track (Ji et al.,), NEEL[2] – Named Entity rEcognition and Linking Challenge on Microposts (Rizzo et al., 2015), or ERD[3] – Entity Recognition and Disambiguation Challenge (Carmel et al., 2014) rank participating systems based on their overall performance on collections of specific textual fragments (selected sentences, tweets...) that had been manually annotated. As the manual dataset preparation is tedious, the provided training and test data is limited to few thousands of entity mentions. Developers of NER tools can measure improvements in annotation quality w.r.t. a particular available dataset or they can use specific benchmarking frameworks such as NERD (Rizzo and Troncy, 2011) or Gerbil (Cornolti et al., 2013), embracing several datasets.

Since the manual creation of evaluation datasets is expensive and time-consuming, researchers started to think how to collect usable data without the burden of additional human labour. Internal links of Wikipedia can provide an apt resource and datasets such as TagMe[4] Wiki-Disamb30 and Wiki-Annot30 explore this potential. Although the data is rich (2 million mentions), it is suitable as a training dataset rather than a testing one.

Other works extract Wikipedia links from web crawled pages. In 2012, researchers from Google and the University of Massachusetts released Wikilinks (Singh et al., 2012) – a dataset, comprising 40 million mentions of over 3 million entities (see the discussion on the numbers below), based on finding hyperlinks to Wikipedia from a web crawl using anchor text as mentions. The dataset was further extended by including complete document contents (with cleaned DOM structure), extracting context for the mentions and aligning the mentions to Freebase/WikiData entities. Similarly, the WikiReverse[5] project contains 36 million links to English Wikipedia articles extracted from Common Crawl's July 2014 web crawl (3.6 billion web pages).

The resource introduced in this paper builds on the datasets mentioned above. We clean and deduplicate existing data first. A special attention is paid to contexts in which entity mentions appear. We disregard mentions that have no valuable context or that appear in repeating boilerplate-like contexts. The original content is extended by the data from other available web crawls. Various data filtering and enhancement steps are followed to gain the resulting dataset. We also extract a subset of ambiguous names (that can refer to more than one entity) and categorize them according to ambiguity types (e. g., person v. place) and an estimated complexity of the disambiguation task (by proportions of link frequencies). The whole corpus is regularly (monthly) expanded as entity linking information from newly added pages accumulates.

In its whole, WTF-LOD represents one of the largest available and unified datasets for training and evaluation of NER tools. The subset of the most ambiguous names then allows in-depth investigation of advanced named entity recognition and disambiguation strategies which would not be possible with currently available resources. The large size of the collected dataset enables exploring interdependencies

[1] http://nlp.cs.rpi.edu/kbp/

[2] http://www.lancaster.ac.uk/Microposts2015

[3] http://web-ngram.research.microsoft.com/ERD2014/

[4] http://acube.di.unipi.it/tagme-dataset/

[5] https://wikireverse.org/

between two key performance characteristics – annotation accuracy and the time to process a text of a given size.

2. Data preparation and processing

Although WikiLinks and WikiReverse projects promise together more than 70 million Wikipedia links, a significant portion of these datasets is formed by incorrect and non-existent links. For example, if Wikilinks' 40,323,863 mentions of 2,933,659 entities are filtered out to contain only links to real articles of English Wikipedia, only 38,130,711 mentions of 1,474,554 distinct entities remain. Moreover, the WikiReverse resource does not distinguish original texts from word-to-word copies of Wikipedia texts on crawled web pages. For example, the site with the highest number of linked Wikipedia articles in WikiReverse is http://edwardbetts.com/ – a personal page of a software developer which enables metasearch in the Wikipedia content (see, e. g., `http://edwardbetts.com/find_link/Smetana`).

To overcome the identified issues, we carefully choose texts (crawled web pages) to be included in the WTF-LOD dataset, check whether they are not exact- or near duplicates of the Wikipedia content and that they do not contain contexts that have been already included in the resource. We also validate the links and extend them by additional information from DBpedia, WikiData/Freebase, GeoNames, Linked Movie DB, etc.

Wikipedia dumps form a basis of the WTF-LOD. We consider various language versions of Wikipedia. Most of the further processing steps are language-independent so that the effort can be easily replicated for additional languages. Yet, as there are not easily available collections resulting from regular crawlings comparable to CommonCrawl for other languages, we report statistics for English Wikipedia only in this paper.

As Wikipedia content includes invalid internal links to non-existent articles (due to errors as well as articles that existed earlier but have been removed), we validate all links identified in the text. We also resolve page redirections in order to unify all potential forms of URL to the same Wikipedia article. Only whole pages are considered as anchors for entity definitions. We disregard links to particular sections of wikipages, for example, [[Aachen#Main_sights]]. Name ambiguity is also occasionally recognized only after an article dealing with a particular entity referred to the ambiguous name is entered to Wikipedia. Unfortunately, resulting disambiguation pages sometimes use the original page title rather than indicate their special purpose by adding (disambiguation) to the original one (for example, Aaron_Johnson_(disambiguation) as opposed to Aaron_Bailey – both being disambiguation pages). To cope with the dynamicity of page titles, WTF-LOD keeps out links to disambiguation pages too. HTML as well as plain text versions of pages from the Wikipedia dump are derived from the cleared data and stored in a unified form. In total, there are 161,649,242 links to 4,970,399 articles resulting from the March 2016 dump of English Wikipedia.

The Wikipedia content naturally contains a large number of interlinks. As Wikipedia articles and DBpedia resources derived from them form primary means to define anchors for most of existing NER tools, the text is often used to train disambiguation modules of these systems. It is then not possible (not fair) to use the same dataset for evaluation of disambiguation results. That is why WTF-LOD specifically marks linking contexts appearing in Wikipedia and enables users to easily identify additional content from the real "wild" web (outside of Wikipedia and copies of its pages).

Existing large web corpora available to general public have been used to extend the primary base of our dataset. The instances of CommonCrawl[6] were processed first. Initial filtering steps involved language checks (only English pages kept), boilerplate removal (by means of BoilerPipe[7] (Kohlschütter, 2011)), and text cleaning (to deal with incorrect encoding of characters). As some crawled Wikipedia pages are included in the corpora and there is also a partial overlap between the CommonCrawl corpora, we employed another computationally intensive step – deduplication. Hashes of individual paragraphs contained in the data are computed and used to remove exact duplicates of paragraphs inserted earlier. We then detect near duplicates by hashing all n-grams of words (n=5) in the remaining paragraphs and evaluating overlaps with previously stored hashes of the n-grams (Pomikálek, 2011). Finally, we extract contexts of the Wikipedia links ($\pm$10 words) and unique them across the whole dataset.

The current version of CommonCrawl covers monthly runs of the large-scale crawling from October 2014 to November 2015. New as well as old crawlings will be added soon. Although there is a significant overlap between the data collected in each two consecutive months (35 – 46%), the enormous size of the crawls guarantees that WTF-LOD increases about 1 – 2 million wikilinks every month. This is demonstrated by the graph in Figure 1 showing the increasing size of the WTF-LOD in time (the CommonCrawl IDs are formed by a year and a week number within the year). Table 1 then details proportions of duplicate wikilinks in individual CommonCrawls.

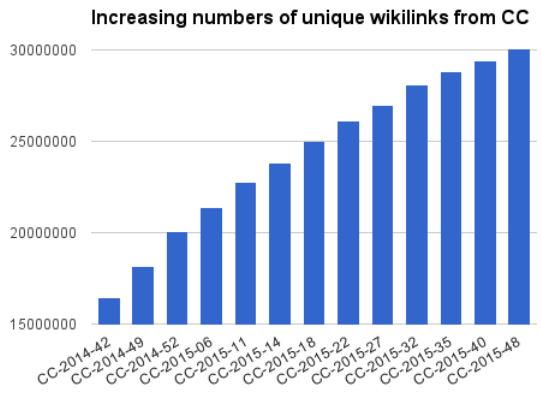

Figure 1: Increasing number of unique wikilinks from CC

[6] `https://commoncrawl.org/`

[7] `https://github.com/kohlschutter/boilerpipe`

	no deduplication	after deduplication	after near-deduplication	after incremental deduplication	after incremental near-deduplication
CC-2014-42	79,904,310	16,455,939	12,418,046	16,455,939	12,418,046
CC-2014-49	42,325,274	13,055,878	10,279,016	1,706,815	784,513
CC-2014-52	38,018,686	15,106,106	11,745,023	1,931,388	795,824
CC-2015-06	34,062,611	13,465,921	10,504,876	1,255,081	314,279
CC-2015-11	32,232,586	13,294,477	10,339,573	1,406,866	531,464
CC-2015-14	29,210,218	12,471,128	9,754,312	1,036,227	313,846
CC-2015-18	35,673,742	14,151,249	10,897,237	1,183,061	314,126
CC-2015-22	34,158,765	13,923,581	10,793,573	1,133,671	287,955
CC-2015-27	28,645,461	12,201,711	9,557,059	869,764	176,945
CC-2015-32	41,582,430	15,831,706	12,112,394	1,087,893	203,758
CC-2015-35	33,515,074	13,139,806	10,237,094	728,166	159,878
CC-2015-40	23,524,714	10,249,359	8,101,424	630,847	139,209
CC-2015-48	33,886,734	12,891,876	10,130,819	857,967	224,074

Table 1: Unique wikilinks in monthly CommonCrawls

The same procedure is applied to ClueWeb09[8] containing about 489 million web pages and ClueWeb12[9] with about 733 million pages. We extracted additional 5,534,518 and 5,663,221 links from ClueWeb09 and ClueWeb12, respectively.

3. Entity disambiguation subset and overall statistics

Figure 2 shows a histogram of the numbers of distinct links per anchors (in log_{10} scale) in WTF-LOD. As one can see, unambiguous anchors form a significant part of the resource. Yet, there are still many cases that need to be disambiguated. To have a representative dataset for evaluating entity disambiguation algorithms, we extracted a subset of WTF-LOD consisting only of anchor texts linked to ambiguous mentions referring to people and places (including imaginary ones – those appearing in books/movies/songs). We took advantage of the DBpedia ontology[10] and some additional processing of the Wikipedia categorization to identify the types.

It is crucial to distinguish the level on which the name ambiguity demonstrates. There are full names such as Michael Jordan that correspond to several people in Wikipedia or other authoritative resources (as well as to many people in the world that have no such record). Moreover, there are songs, movies, video games, statues, ships, places... that bear the same name. Natural languages and socio-cultural settings apply various systems to prevent confusion in such cases (middle names, titles, the Elder/the Younger qualifiers, etc.). Potential ambiguity increases if one considers references made only by a surname or by a given name. Indeed, partial name references increase the pool of potentially referred entities significantly. Some authors and some challenges focus on the task of entity disambiguation and linking in the specific context of short sentences with several partial name references (for example, "Thomas and Mario are strikers playing in Munich" in (Navigli and Moro, 2014)).

Various disambiguation approaches identifying the most probable combination of candidate references can be devised. On the other hand, the situations in which such sentences need to be disambiguated without any further context from the source document are rather rare. It is often the case that simple co-reference resolution helps in real disambiguation cases of this type. That is why WTF-LOD includes the whole documents in which name references appear. Disambiguation tools can then apply any suitable method to link correct entities. Non-trivial co-reference resolution is also needed in the case of common-noun references and other expressions not including a part of entity names (for example, the web search giant). This kind of reference are not frequent in the collected data (less than 1%) and it is not included in the "complex disambiguation"

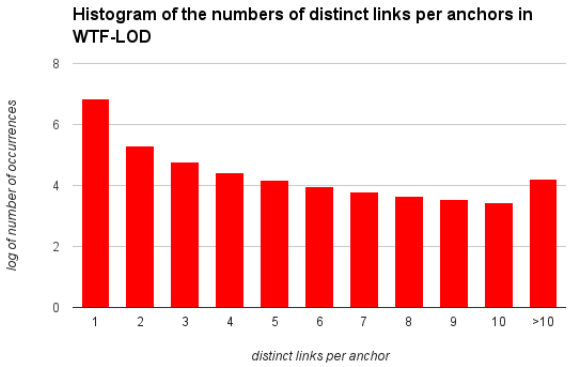

Figure 2: Histogram of the numbers of distinct links per anchors in WTF-LOD

[8] http://lemurproject.org/clueweb09/

[9] http://lemurproject.org/clueweb12/

[10] http://wiki.dbpedia.org/services-resources/ontology

subset.

The subset is further subdivided according to the above-mentioned semantic types of entities that can be referred to by each name. For example, there are 32,018 ambiguous people names (corresponding to 51,753 Wikipedia articles) in the data. Considering only those entities that are referred to more than 50 times in the web-crawled part of the WTF-LOD dataset, there are 4,565 ambiguous entity names with 1,277,974 unique contexts.

Table 2 summarizes overall statistics of WTF-LOD as a whole and the disambiguation subset. The data will be freely available on torrent. If it shows to be feasible, we will make it accessible also as a part of the WikiReverse site.

	Total dataset	**Disambiguation subset**
Number of mentions	203,130,666	13,145,189
Referred entities	5,036,070	241,203
Size in GB	68	4.2

Table 2: Overall statistics of WTF-LOD as a whole and the disambiguation subset

4. Conclusions and Future Work

The WTF-LOD dataset introduced in this paper lays the foundations for large-scale evaluation of named entity recognizers. As opposed to datasets employed in current NER evaluations (for example, the ESWC-16 Open Knowledge Extraction Challenge[11]), it enables exploring scalability of NER systems and their accuracy in real conditions. The subset of ambiguous names can then be used for comparison of disambiguation strategies and their applicability in realistic web deployments.

Our future work will concentrate on methods identifying sites with frequent Wikipedia links. This information will form a basis of focused crawling that could bring a significant number of additional wikilinks. The methods will be language independent which will help us to prepare similar datasets for other languages. The dataset will be further extended by links to other LOD sources from web pages (which are, unfortunately, rare). We will also produce PoS-tagged and parsed versions of link contexts.

5. Acknowledgements

This work was supported by the European Union Horizon 2020 project MixedEmotions, grant agreement No. 644632, and by The Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II), project IT4IXS – IT4Innovations Excellence in Science – LQ1602.

[11] http://2016.eswc-conferences.org/eswc-16-open-knowledge-extraction-oke-challenge

6. Bibliographical References

Buitelaar, P. and Cimiano, P. (2008). *Ontology learning and population: bridging the gap between text and knowledge*, volume 167. Ios Press.

Carmel, D., Chang, M.-W., Gabrilovich, E., Hsu, B.-J. P., and Wang, K. (2014). Erd'14: Entity recognition and disambiguation challenge. In *ACM SIGIR Forum*, volume 48, pages 63–77. ACM.

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013). A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 249–260, New York, NY, USA. ACM.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pages 782–792.

Ji, H., Nothman, J., Hachey, B., and Florian, R.). Overview of tac-kbp2015 tri-lingual entity discovery and linking.

Kohlschütter, C. (2011). *Exploiting links and text structure on the Web: a quantitative approach to improving search quality.* Ph.D. thesis, University of Hanover.

Navigli, R. and Moro, A. (2014). Multilingual word sense disambiguation and entity linking. In *COLING (Tutorials)*, pages 5–7.

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Faculty of Informatics, Masaryk University.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics.

Rizzo, G. and Troncy, R. (2011). Nerd: evaluating named entity recognition tools in the web of data.

Rizzo, G., Cano, A. E., Pereira, B., and Varga, A. (2015). Making sense of microposts (#microposts2015) named entity recognition & linking challenge. In *5th International Workshop on Making Sense of Microposts (#Microposts2015)*.

Singh, S., Subramanya, A., Pereira, F., and McCallum, A. (2012). Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. Technical Report UM-CS-2012-015.