Missing Mr. Brown and buying an Abraham Lincoln – Dark Entities and DBpedia

Marieke van Erp,[1] Filip Ilievski,[1] Marco Rospocher[2] and Piek Vossen[1]

[1] VU University Amsterdam
{marieke.van.erp,f.ilievski,piek.vossen}@vu.nl
[2] Fondazione Bruno Kessler
rospocher@fbk.eu

Abstract. We argue for the need for the community to address the issue of "dark entities", those domain entities for which a knowledge base has no information in the context of the entity linking task for building Event-Centric Knowledge Graphs. Through an analysis of a large (1,2 million article) automotive newswire corpus against DBpedia, we identify six classes of errors that lead to dark entities. Finally, we outline further steps that can be taken for tackling this issue.

Keywords: natural language processing, domain-specific entity linking

1 Introduction

Knowledge graphs are becoming ever more important for companies to organise and access their data [13]. A key starting point for most knowledge graphs is data extracted from Wikipedia. Hence, DBpedia [4] and DBpedia Spotlight [6] often serve as basis for knowledge graph construction and entity grounding efforts [14]. However, due to the breadth and the general nature of Wikipedia, it is often insufficient for domain-specific use cases such as biomedical entity linking [18]. Further, there are bias and coverage issues with Wikipedia and its derivatives as evidenced by the work of Kittur [10] and Tiddi [15] as well as lack of location information shown in Wikidata World Maps[3].

Another dimension that is not sufficiently covered by DBpedia is events. As we aim to build Event-Centric Knowledge Graphs (ECKGs), we need to know more about an entity than its type. If it is for example a person, we need to know things such as what his/her role is, when (s)he was appointed, what (s)he did before, etc. For organisations, we need to know what deals they made with whom, what products they offer etc.

Given the importance of DBpedia for entity grounding efforts and its apparent skew, we assert that more attention needs to be paid to the issue of *dark entities* in our natural language processing pipelines. Dark entities are those entities for which no relevant information was identified in a given knowledge base / entity repository. In many cases, this means that there is no resource present in the knowledge base. There are also cases where a resource is present, but it contains very little or no relevant information to further reason about the entity with. For example, `http://dbpedia.org/page/Maurice_Lippens`, contains no other information than redirect links.

[3] `https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en`

To obtain insight in the scale of the problem, we performed an analysis of the dark entities phenomenon with respect to a newswire corpus and DBpedia. For news, the performance of NLP pipelines are typically good and the coverage of DBpedia should be satisfactory. We find that for over 40,000 entities, there was no additional information available in an external resource. Based on this analysis, we catalog six classes of errors that lead to dark entities. We also discuss *true dark entities*, those in which the failure is due to the fact that no resource within DBpedia exists. True dark entities are those domain entities that, while not "famous" enough to deserve a page in Wikipedia (and hence an entry in DBpedia), play some role in the given domain. We note that the community has been concerned with NIL entities [9] (entities that may not have high rates of occurrence within a corpus or are domain-specific). Dark entities subsume this problem and expand it to consider problems in the entity grounding itself as well as deriving ECKGs involving these entities. For example, the event "crash" can mean different things. Knowing its participants are entities is not enough. We need to know what they do and how they are related. If the subject and object are the CEO and CFO of the same company this is valuable information to judge the implication of them crashing, whilst a company that crashes has an entirely different meaning. Dark entities are also related to the notion of *emerging entities* [8], but this is focused on identifying potentially ambiguous entities becoming highly popular in news in a short time.

This paper is structured as follows. In Section 2, we describe the dataset used for our analysis, followed by a discussion of the entity grounding pipeline used (Section 3). In Section 4, we present our analyses. We conclude with thoughts on ways forward.

2 The Global Automotive Industry Dataset

The Global Automotive Industry Dataset was compiled in the context of the NewsReader EU project,[4] to investigate the impact of the 2008-2009 financial crisis on the automotive industry.

The dataset comprises 1,259,748 English news articles covering the period 2003 - 2013. It was extracted from the LexisNexis database[5] through a series of complex queries containing major car makers. The dataset has been used as a test bed to scale NLP and Semantic Web technologies to extract entities, events and information about them from natural language text, resulting in Event-Centric Knowledge Graphs that formalize long-term developments and changes. The system is described in [1] and its modules are openly available separately or as a virtual machine from `http://www.newsreader-project.eu/results/software/`. The 1.2M articles were first processed using NewsReader's natural language processing pipeline that performs information extraction at the document level. Then, a cross-document coreference module aggregated the different event and entity mentions in the text to unique instances, with links to the locations in the text documents where the events and entities were mentioned.[6] This resulted in 25,156,574 events and 1,967,150 entities. But for only less

[4] `http://www.newsreader-project.eu`

[5] `http://www.lexisnexis.com`

[6] The generated RDF TRiG files are available from `http://www.newsreader-project.eu/results/data/`. The sources cannot be made available due to copyright issues but can be accessed to LexisNexis subscribers via the document IDs in the RDF.

than 15% of the entity instances (277,425 out of 1,967,150) that the pipeline detected, the system was able to link the entity to a DBpedia resource.

3 Entity Grounding in Detail

To recognise entities in text, ixa-pipe-nerc [2][7] is employed, a perceptron based NERC system that is trained on the AIDA/CoNLL 2003 dataset,[8] enriched with unsupervised clusters. This NERC system outperforms state-of-the-art systems on both the AIDA/CoNLL 2003 benchmark and a set of 120 WikiNews corpus[9] articles annotated for this project ([7], tables 6 and 7).

The results from the NERC module in the pipeline, were used as input for DBpedia Spotlight statistical models [6] to ground the entity in the DBpedia knowledge base. On the benchmark datasets TAC 2011 [9] and AIDA/CoNLL 2003[10] scores of 79.77 precision and 60.68 recall and 79.67 precision and 75.94 recall respectively were obtained. We chose not to rely on DBpedia Spotlight's internal entity spotter as it links far more entities than the named entities we are currently investigating. Naturally, an analysis on the concepts spotted and linked by Spotlight lies in the extension of this work.

4 Dark Entities: An Analysis

An analysis of dark entity candidates shows that there are four main reasons for errors that are related to recognising the correct candidates to link:

Real world data is dirty Character encodings are still a hard problem, resulting in recognised entities such as JonBenâ©t and Bill %26 Melinda Gates Foundation. This is despite the fact that all data comes from a standardised digital repository.

Incorrect Named Entity Recognition Boundaries Named entity recognisers have difficulties with longer entities where parts are left out ('Ferdinand' instead of 'Rio Ferdinand'), or erroneously included ('Copyright Targeted News Services' instead of 'Targeted News Services').

Conjunctions The named entity recogniser may get the boundaries of the entity as it is referred to correct as it also relies on syntactic constituent information, but this chunk may be made up of two or more entities, such as 'Peugeot and Citroen'. Both are valid entities and have resources in DBpedia, but the conjunction proves difficult to analyse for the named entity disambiguator.

Coreference The state-of-the-art in-document coreference resolution hovers around F_1 score of 60 [5]. This means that the system might be able to link the first mention of 'Gordon Brown' in the text, but fail to connect this to a subsequent mention of 'Mr. Brown'.

[7] https://github.com/ixa-ehu/ixa-pipe-nerc

[8] http://www.cnts.ua.ac.be/conll2003/ner/

[9] http://en.wikinews.org/wiki/Main_Page

[10] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

These errors may not result in true dark entities, but they mostly make it difficult for the named entity linker to identify the correct context to link the entity candidate to. Data cleaning can alleviate some of the instances coming from the first two causes but is not trivial. Another possible remedy for the entity boundaries problem is to perform the tasks of entity recognition and disambiguation simultaneously [12]. As for conjunctions, at first sight, one could think of splitting entities that have a mention of 'and' or '&' in them, but this would incorrectly split 'Standard & Poor's Financial Services LLC'. And what does one do with names that may have commas as part of their name?

Resolving errors from the entity recogniser, still leaves two more causes of error, related to dark entities:

Subdivisions The granularity level of an entity in text may not match its granularity in a knowledge base. For instance, only the 'bigger' organisation will have an entry in the knowledge base, making it difficult to link more specific entity mentions such as 'Volkswagen Middle East'. Linking to the generic 'Volkswagen' resource may not be optimal, for example in a potential case where Volkswagen closes the Volkswagen Middle East branch. When reasoning over such a statement, one would falsely state that dbpedia:Volkswagen[11] closes dbpedia:Volkswagen.

Domain Mismatch / Ambiguity In Figure 1, the distribution of selected DBpedia classes is charted against the class distribution of the entities recognised in the Global Automotive Dataset v2 by DBpedia Spotlight and the class distribution of the entities in 120 manually annotated Wikinews articles for this domain. The figure shows that organisations are much more important in the NewsReader domain than in DBpedia. It should be noted that there is a mismatch between the DBpedia class 'Work' and the 'Work' class in the Gold Standard dataset, as in the DBpedia classification 'Work' creative works such as artwork and software, whereas the NewsReader annotation also includes electronics and car models. In the DBpedia classification, these are covered by different classes such as 'Device' and 'MeanOfTransportation'. The DBpedia ontology also shows that there is much attention for sports and celebrities, as is demonstrated by the detailed subclass hierarchy under the 'Person' class. In the NewsReader domain, most persons are businesspeople, of different levels of fame (CEOs, politicians and journalists). The domain mismatches between DBpedia and the automotive dataset are often cause for linking errors, most apparent when an entity is consistently linked to the most popular DBpedia candidate, despite its low domain relevance or context fit. In the NewsReader data this for example happens with 'Lincoln' which should refer to the carmaker (or in some instances to 'Lincoln Matresses') but is consistently linked to the entry on 'Abraham Lincoln' (or 'Lincoln City F.C.', 'Lincoln Park', 'City of Lincoln') instead.

5 Conclusion

Our analyses showed that although newswire is considered a standard domain in NLP evaluations, with many tools tuned to it, recognising and linking entities mentioned is nontrivial. Similarly as the phenomenon of emerging entities, the case of dark entities has been treated as a side effect in contemporary linking systems. While the analysis is

[11] `PREFIX dbpedia: <http://dbpedia.org/resource/>`

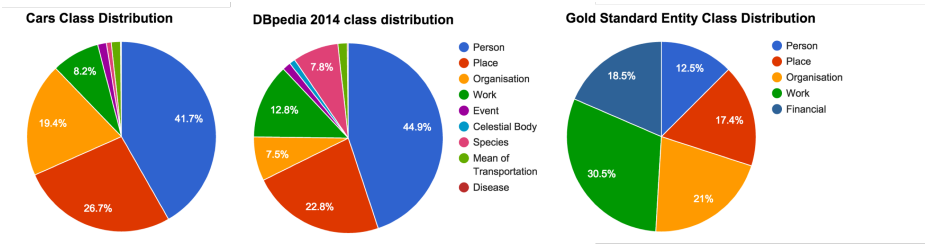

Fig. 1. Class distribution of DBpedia 2014, entities recognised and linked to DBpedia in the Cars dataset and Gold Standard annotation

preliminary, combined with anecdotal evidence from discussions with other community members emphasizes the need to address the problem of dark entities. What are some avenues for addressing our six classes of errors? Possibilities include:

- **Dynamic set of knowledge bases.** The LOD cloud contains vast amount of sources, diverse with respect to their topical domain and level of detail. Linking entities to domain-relevant knowledge bases should increase the recall of the dark entities and reduce the mismatch in granularity. Unfortunately, few practicalities need to be resolved. Firstly, currently it is hard to include additional sources into DBpedia Spotlight. Secondly, it may be nontrivial to decide which sources to consider.
- **Expand knowledge bases.** LOD resources are used as input for NLP applications. Recent work [17, 8] have exploited the possibility for the reverse direction: feeding the NLP results back into the LOD cloud. Although these are not perfect yet, they may be useful in extending the coverage of LOD. Besides, the LOD cloud resources themselves are also not error-free [16, 3]. New information in the cloud could in turn lead to different outcome for the NLP application, making their relationship somewhat circular. We propose to examine the applicability of this circularity for the dark entities use case, through a three-step strategy: 1) Identification of unresolved entities; 2) Analysis of the domain in order to normalise these entities and determine their semantic type and relations to other entities. Results are then fed into a domain knowledge base, which is an add-on to DBpedia; 3) Re-processing of the domain texts to obtain better interpretations of the entities and their activities.
- **Leverage latent semantics.** Techniques such as word2vec [11] that mine latent semantics may be effective for providing guidance on how to propagate information from one known entity to another. Such techniques are a promising mechanism to enrich the representation of an entity, however, the errors caused by the chosen latent semantics algorithm should be taken into account.

Going forward, we hope that by tackling dark entities we can make for natural language processing systems which are more accurate and have higher coverage.

Acknowledgments

The research for this paper was supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

References

1. Agerri, R., Aldabe, I., Beloki, Z., Laparra, E., de Lacalle, M.L., Rigau, G., Soroa, A., Fokkens, A., Izquierdo, R., van Erp, M., Vossen, P., Girardi, C., Minard, A.L.: Event detection, version 2 deliverable 4.2.2. Deliverable, NewsReader Project (2014)
2. Agerri, R., Bermudez, J., Rigau, G.: IXA Pipeline: Efficient and ready to use multilingual NLP tools. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland (2014)
3. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: Lod laundromat: a uniform way of publishing other people's dirty data. In: The Semantic Web–ISWC 2014, pp. 213–228. Springer (2014)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Journal of Web Semantics: : Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)
5. Clark, K., Manning, C.D.: Entity-centric coreference resolution with model stacking. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 1405–1415. Bejing, China (26 – 31 July 2015)
6. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 121–124. ACM (2013)
7. van Erp, M., Vossen, P., Agerri, R., Minard, A.L., Speranza, M., Urizar, R., Laparra, E., Aldabe, I., Rigau, G.: Annotated data, version 2. deliverable d3.3.2. Tech. rep., NewsReader Project (2015)
8. Hoffart, J., Altun, Y., Weikum, G.: Discovering emerging entities with ambiguous names. In: Proceedings of the 23rd international conference on World wide web. pp. 385–396 (2014)
9. Ji, H., Grishman, R.: Knowledge base population: Successful approaches and challenges. In: Proceedings of ACL 2011 (2011)
10. Kittur, A., Chi, E.H., Suh, B.: What's in wikipedia? : mapping topics and conflict using socially annotated category structure. In: CHI'09. pp. 1509–1512 (2009)
11. Mikolov, T., tau Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL HLT (2013)
12. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics 2 (2014)
13. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs - from multi-relational link prediction to automated knowledge graph construction. `http://arxiv.org/pdf/1503.00759.pdf` (Mar 2015)
14. Rizzo, G., Troncy, R., Hellmann, S., Bruemmer, M.: Nerd meets nif: Lifting nlp extraction results to the linked data cloud. In: LDOW'12 (2012)
15. Tiddi, I., d'Aquin, M., Motta, E.: Quantifying the bias in data links. In: Knowledge Engineering and Knowledge Management, pp. 531–546. Springer (2014)
16. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in dbpedia. In: The Semantic Web: Trends and Challenges, pp. 504–518. Springer (2014)
17. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: A probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 481–492. ACM (2012)
18. Zheng, J.G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., Ji, H.: Entity linking for biomedical literature. extraction 24, 19 (2014)