# On the Long-tail Entities in News

José Esquivel[1,2], Dyaa Albakour[2], Miguel Martinez[2], David Corney[2], and Samir Moussa[2]

[1] School of Computer Science and Electronic Engineering, University of Essex, UK
[2] Signal Media Ltd., 32-38 Leman Street, London, E1 8EW, UK
research@signalmedia.co

**Abstract.** Long-tail entities represent unique challenges for state-of-the-art entity linking systems since they are under-represented in general knowledge bases. This paper studies long-tail entities in news corpora. We conduct experiments on a large news collection of one million articles, where we devise an approach for measuring the volume of such entities in news and we uncover insights on the challenges associated with linking these entities to general knowledge bases.

## 1 Introduction

In the modern world of fast-flowing news delivery and consumption, searching and filtering documents for entities is becoming a more common information retrieval task. This has been echoed in a number of information retrieval evaluation initiatives such as the TREC KBA track [1] and the NewsIR workshop [2]. Filtering news documents using entities relies on effective Entity Linking (EL) approaches that are capable of identifying mentions of entities in the text and linking them to their entries in knowledge bases (KB)s [3].

State-of-the-art approaches for EL focus on popular entities and rely on general KBs, such as Wikipedia. The success of these approaches depends heavily on the availability of a sufficient quantity of relevant information about the entities in the KB. This includes the textual content of the pages representing the entities from which to learn an appropriate language model that describes them [4]. In addition, the links to the Wikipedia pages representing the entities provide a set of candidate mentions for each entity, as well as the semantic relations between entities in the KB as inferred from the graph of links [5]. In other words, state-of-the-art EL systems rely on general KBs covering popular entities with rich textual content and meta-data about them [6].

Entities which have a less complete profile cannot be easily linked by these approaches [6]. Many less popular or domain-specific entities are under-represented in general KBs such as Wikipedia [7]. We refer to these as *long-tail* entities, and examples of them include small-medium organizations, less popular individuals and rarely-mentioned geographical places. In the literature, long-tail entities have been defined as the large number of entities with relatively few mentions in text corpora [8]. They are characterized as those with limited or no KB profile and sparse or absent resources outside the KB [3]. In this paper, we study long-tail entities in news corpora.

ID: f7ca322d-c3e8-40d2-841f-9d7250ac72ca
Title: Worcester breakfast club for veterans gives hunger its marching orders
VETERANS saluted **Worcester**'s first ever breakfast club for ex-soldiers which won over hearts, minds and bellies. The Worcester Breakfast Club for HM Forces Veterans met at the Postal Order in **Foregate** Street at 10am on Saturday. ...

Fig. 1: An example from the Signal-1M dataset. **Bold** represents entities identified by the linker, while underlined are entities identified by the NER tagger.

A concrete example of popular and long-tail entities is given in the excerpt from a news article shown in Figure 1. This shows mentions of two classes of entities. The word "Worcester" is a reference to the town in Worcestershire, England. On the other hand, "Worcester's Breakfast Club for HM Forces and Veterans" is a mention of a specific organization, an entity which does *not* have an entry in Wikipedia and therefore cannot be linked by an off-the-shelf entity linker.

In this paper, we perform an analysis of a large collection of news articles, namely the Signal Media One Million News Articles (Signal-1M) dataset [9], to estimate the volume of long-tail entities which cannot be linked to general KBs. To do this, we compare the entity mentions identified by a Named Entity Recognizer (NER) and the entities linked to a general KB by a state-of-the-art entity linker. Our analysis shows that a large number of entities in news articles are difficult to link as they are either ambiguous or unpopular. Our assumption is that entities that cannot be easily linked are generally long-tail entity mentions, i.e. not well covered in general KBs. Furthermore, we show that even some common entities in the news are not well covered in general KBs.

To summarize, our main contributions are devising an approach for estimating the volume of long-tail entities in the news and uncovering insights into the volume and the types of entities that cannot be easily linked to general KBs.

## 2 Identifying long-tail entities

To empirically estimate the volume of long-tail entities in a corpus of documents, first we run each document through a NER tagger and an EL tagger separately. The NER tagger identifies mentions of entities in the document along with their types (the NER tag set), while the EL tagger identifies and links entity mentions to their entries in a general KB (the EL tag set). Then, we compute the overlap between these tag sets. We consider this overlap a reasonable proxy for estimating the volume of long-tail entities. In particular, long-tail entities will be typically identified by the NER tagger but not linked by the EL tagger due to their low coverage in the KB. A high overlap indicates a smaller volume of long-tail entities, while a low overlap indicates the opposite. In our approach, we consider two tags as overlapping if either of their start or end offsets is within the other tag's offsets. For example, Figure 1 shows two cases of overlapping tags. In the first case, *Worcester* is identified by both taggers. In the second case, *Foregate Street* identified by the NER tagger, whereas the EL tagger marked only *Foregate*.

One limitation of our approach is that it relies on the correctness of the NER tagger. However, we think the resulting NER tag set is an unbiased approximation of the complete set of entities in the corpus. Also, we understand that this is only one possible way of estimating the long-tail entity set. Other approaches, such as getting the least frequent entities in a KB, or the out-of-database entities in the same, should also be explored.

## 3 Estimating the long-tail of entities in news

### 3.1 Experimental Setup

To estimate the long-tail of entities in news articles, we applied the procedure described in Section 2 on the one million articles in the Signal 1M dataset, originally sourced from tens of thousands of news and blog sources in September 2015. For NER, we used the Stanford tagger [11], and we used DBPedia Spotlight for EL [3], which uses Wikipedia as a KB. When measuring the overlaps of the tagger outputs, we aggregate the results by entity type and by unique entity mentions. The latter is done after normalizing each of the entity mentions by removing any white-space and non-ASCII characters from it, and converting them to their lower case representation. We do this to get a better estimate of the amount of unique entity mentions in the corpus identified by Stanford NER tagger as we decrease the number of duplicate mentions, which only differ in formatting.

Moreover, to further examine the effectiveness of Spotlight in linking long-tail entities, we ran the same procedure described in Section 2, but with different subsets of the unique entity mentions identified by the Stanford NER tagger in the corpus. We achieve this by specifying a cut-off point $x$, at which we consider only the top $x\%$ of normalized unique entity mentions ranked by their frequency in the corpus.

We configured both taggers (Stanford NER Tagger and DBPedia Spotlight) with the recommended parameters according to their documentation. For the Stanford NER Tagger, we used the default English 3-class model trained on news articles without part-of-speech tagging [11]. For DBPedia Spotlight, we used the 'annotate' end-point of the API adjusting the confidence and the support input parameters to 0.4 and 5 respectively as recommended by the API documentation. [4] The API was deployed locally with a Wikipedia dump from July 2013 (two years prior to the dates of the news articles in the dataset). We believe that with this configuration, we may capture newly emerging entities which typically appear in Wikipedia after some lag [10].

Finally, we aggregate entity types into the Stanford types: *PERSON*, *LOCATION*, and *ORGANIZATION*. To do this we map all DBPedia Spotlight types falling under the *Person*, *Place*, and *Organization* hierarchy to their corresponding Stanford types. We also introduced two other types: (i) any DBPedia type that does not fall under any of these top-level hierarchies is mapped to *MISC*; (ii) the DBPedia's default top-level *Thing* type, is mapped to another custom type *None*. [5]

[3] `https://github.com/dbpedia-spotlight/`
[4] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service`
[5] `http://mappings.dbpedia.org/server/ontology/classes/`

Table 1: Overlaps ratio with different types of linked entities by Spotlight. Each row represents all entity mentions for a certain Stanford type; each column corresponds to one Spotlight type.

| | PERSON | LOCATION | ORG. | MISC | None | No Overlap |
|---|---|---|---|---|---|---|
| PERSON (total=7.71M) | 26.59% | 4.71% | 2.52% | 1.29% | 10.51% | **54.38%** |
| LOCATION (total=5.52M) | 0.65% | **64.55%** | 6.43% | 1.62% | 19.42% | 7.33% |
| ORG. (total=5.37M) | 1.49% | 11.91% | **39.44%** | 4.68% | 27.94% | 14.54% |

### 3.2 Results

Table 1 shows the overlap between the Stanford entities and the Spotlight entities grouped by type. Overall, we observe that the same-type overlap is relatively poor across the different types of entities considered. In particular, the same-type overlap is worst for people (26.59%) and best for locations (64.55%). The last column "No Overlap" in the table shows the percentage of misses; from this we can see that the Spotlight linker is not able to provide a link for almost half of the "people mentions" (more than 4.1 million people mentions in the Signal 1M Dataset). This indicates that there is a large number of people mentioned in news articles that are hard to link to general KBs. Organizations have a lower rate of misses, but there is still a large percentage of organizations that are in the long-tail and hard to link to general KBs. It should be noted that there are significant cases where Spotlight was able to link the entity but where the linked entity did not have an identifiable type. This is because there are a large number of entities in Wikipedia which do not have an explicit type, especially in the case of organizations, where 27.94% of entity mentions are linked to KB entities that have no type. This data illustrates that a large number of entities in news articles are hard to link to general KBs, which is an indication that they are either not covered in the KB at all or that they are very ambiguous.

We conducted another analysis where we looked at how the overlap changes for more popular mentions of entities in the corpus. Figure 2 plots the overlap between Stanford entities and Spotlight entities for different cut-off points of Stanford entities ranked by their frequency (see Section 3.1 for the definition of cut-off points). Likewise, we plot the misses rate (No Overlaps) in Figure 3. We observe that at higher cut-offs, the average same-type overlap increases for all entity types, with the largest increase being for people names. Similarly, the Spotlight linker is more successful in finding a link for these mentions, but again the decrease in the misses rate is only marginal. Therefore, even for the very commonly-mentioned entities, the Spotlight linker is still not capable of finding them in Wikipedia.

To examine whether this is due to coverage in the KB or entity ambiguity, we aggregate the overlap per Stanford entity mention at the various cut-off points. The intuition is that understanding the distribution of overlap across entity mentions for different cut-off points (degree of mention popularity) would give more explanation on the effectiveness of Spotlight. For each cut-off point, we present the distribution of overlap percentages per entity mention as a box plot in Figure 4. For very popular mentions (cut-off point 0.1%) the average overlap is high and the variance is small meaning that the majority of mentions can be

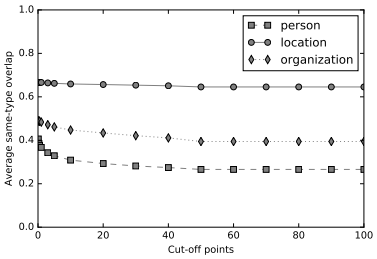


Fig. 2: Same-type overlap between Stanford and Spotlight entities for different cut-off points of Stanford entities ranked by their frequency

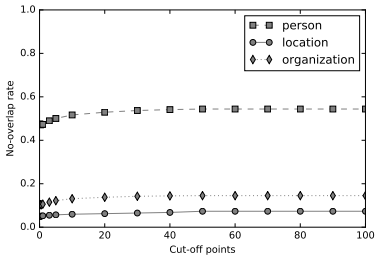


Fig. 3: No-overlap rate with Spotlight entities for different cut-off points of Stanford entities ranked by their frequency

linked, in most cases, but there are still hard ones which are never linked to a KB. At higher cut-off points (5%, and 10%), we observe that the average overlap decreases and the variance is very high.

The average lower overlap is expected since less popular entities are less likely to be represented in Wikipedia. However, the high variance indicates that Spotlight is generally either very successful in linking the entity for most of its occurrences or not successful at all. This indicates that the linking is mainly suffering because of the lack of coverage of these entities in Wikipedia.

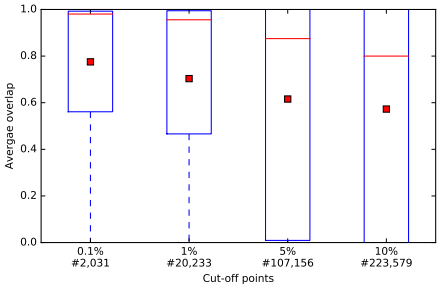


Fig. 4: Average overlap per entity mention at different cut-off points. The number of unique entity mentions is shown below the cut-off point

To further investigate the problem, we manually checked the entity mentions with high mean overlap and with very low mean overlap at the different cut-off points (examples shown in Table 2). As expected, entity mentions with high average overlap are usually referring to popular entities and are not ambiguous, which makes them easy cases for Spotlight. The examples in the table include popular people (Donald Tusk), organizations (CFA institute) and locations (Balkans). On the other hand, very popular mentions with low overlap of linked entities (second column of Table 2) are ambiguous mentions of people or organizations (e.g. Total and Andy) or emerging entities (e.g. Daesh and Diego Costa) that were not well covered in Wikipedia in 2013, the snapshot used in the experiment. Finally for common but less popular men-

Table 2: Examples of high-overlap entity mentions and low-overlap entity mentions at different cut-off points.

| **cut-off 0.1%** | **cut-off 0.1%** | **cut-off 10%** |
|---|---|---|
| **High Overlap** | **Low Overlap** | **Low Overlap** |
| cfa institute | andy | mark gleeson |
| rbc capital | nomura | mique juarez |
| donald tusk | total | pryce |
| balkans | daesh | amanda sue watson |
| barclays premier league | diego costa | asigra |

tions in the corpus (cut-off 10%), mentions with low overlap mainly represent people or organizations which are not represented in Wikipedia.

## 4 Conclusions and Future work

We have analyzed the overlap between state-of-the-art NER and EL systems and the results show that not only is their overlap relatively poor, but also EL systems clearly under-perform when linking long-tail entities (up to 50% missing rate for people), even for those which are very common in the news. This directly impacts the end-to-end quality of entity linking systems, and it could be especially relevant for scenarios where long-tail entities are common (e.g., niche areas such as law or medicine). Future work will consider other datasets from those areas. Also, we will consider experiments using more recent Wikipedia dumps with Spotlight to estimate the volume of emerging entities. Unsurprisingly, our experiments suggest that person names are the hardest to link by the Spotlight linker, as compared to organizations or locations. Our analysis also highlights some of the challenges of EL in news, such as emerging entities being problematic for EL and that ambiguous mentions of entities are never linked.

## References


1. J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, E. Voorhees, and I. Soboroff. TREC KBA Overview. In *Proceedigns of TREC 2014*.
2. Miguel Martinez, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David Corney, Ricardo Campos, and Dyaa Albakour. Report on the 1st International Workshop on Recent Trends in News Information Retrieval (NewsIR16). In *SIGIR Forum* **50**(1), pp 58–67.
3. Ridho Reinanda, Edgar Meij, and Maarten de Rijke. Document filtering for long-tail entities. In *Proceedings of CIKM 2016*.
4. Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.
5. Paolo Ferragina and Ugo Scaiella. Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of CIKM2010*.
6. Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of ELRA*, 2016.
7. Thomas Lin and Oren Etzioni. No noun phrase left behind: detecting and typing unlinkable entities. In *Proceedings EMNLP 2012*.
8. Mina H. Farid, Ihab F. Ilyas, Steven Euijong Whang, and Cong Yu. LONLIES: estimating property values for long tail entities. In *Proceedings of SIGIR 2016*, pages 1125–1128, 2016.
9. David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. What do a million news articles look like? In *Proceedings of ECIR NewsIR 2016 workshop*.
10. Fetahu, Besnik and Anand, Abhijit and Anand, Avishek. How much is Wikipedia Lagging Behind News? In *Proceedings of the ACM Web Science Conference 2015*.
11. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of ACL 2015*.