Systematic Study of Long Tail Phenomena in Entity Linking

Filip Ilievski, Piek Vossen, Stefan Schlobach
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{f.ilievski, piek.vossen, k.s.schlobach}@vu.nl

Abstract

State-of-the-art entity linkers achieve high accuracy scores with probabilistic methods. However, these scores should be considered in relation to the properties of the datasets they are evaluated on. Until now, there has not been a systematic investigation of the properties of entity linking datasets and their impact on system performance. In this paper we report on a series of hypotheses regarding the long tail phenomena in entity linking datasets, their interaction, and their impact on system performance. Our systematic study of these hypotheses shows that evaluation datasets mainly capture head entities and only incidentally cover data from the tail, thus encouraging systems to overfit to popular/frequent and non-ambiguous cases. We find the most difficult cases of entity linking among the infrequent candidates of ambiguous forms. With our findings, we hope to inspire future designs of both entity linking systems and evaluation datasets. To support this goal, we provide a list of recommended actions for better inclusion of tail cases.

1 Introduction

The task of Entity Linking (EL) anchors recognized entity mentions in text to their semantic representation, thus establishing identity and facilitating the exploitation of background knowledge, easy integration, and comparison and reuse of systems. The past years featured a plethora of EL systems: DBpedia Spotlight (Daiber et al., 2013), WAT (Piccinno and Ferragina, 2014), AGDISTIS MAG (Moussallem et al., 2017), to name a few. These systems propose various probabilistic algorithms for graph optimization or machine learning, in order to perform disambiguation, i.e., to pick the correct entity candidate for a surface form in a given context. The reported accuracy scores are fairly high, which gives an impression that the task of EL is both well-understood and fairly solved by existing systems.

At the same time, several papers (Ilievski et al., 2016; Van Erp et al., 2016; Esquivel et al., 2017; Ilievski et al., 2017) have argued that state-of-the-art EL systems base their performance on frequent 'head' cases, while performance drops significantly when moving towards the rare 'long tail' entities. This statement seems intuitively obvious, but no previous work has quantified what the 'head' and 'tail' of EL entails. In fact, the lack of definition of head and tail in this task prevents the (in)validation of the hypothesis that interpreting some (classes of) cases is more challenging for systems than others. This, in turn, means that we are currently unable to identify the difficult cases of EL for which current systems need to be adapted, or new approaches need to be developed. Previous linguistic studies which analyze words distributions can not be applied for this purpose, because they do not study reference, nor the relation of the head-tail distribution to system performance.

Understanding the tail cases better and explicitly addressing them in systems design will be beneficial because: 1. a lot of textual data and requirements for processing it are made up of long tail cases, 2. unlike the head entities, the knowledge about tail entities is less accessible (in structured or unstructured form), not redundant, and hard to obtain. 3. to perform well on the tail, systems are required to interpret entity references without relying on statistical priors, but by focusing on high-precision reasoning.

Proceedings of the 27th International Conference on Computational Linguistics, pages 664–674
Santa Fe, New Mexico, USA, August 20-26, 2018.

This paper addresses the question: **which data properties capture the distinction between the head and the tail in entity linking, and to what extent?** Its main contributions are the following:

1. We define the long tail properties of entity linking **(Section 3)**.[1] This is the first work that looks systematically into the relation of surface forms in datasets and their instances in DBpedia, and provides predictions in the form of a series of hypotheses about long tail phenomena.

2. We analyze existing entity linking datasets with respect to these long tail properties, demonstrating that data properties have certain inter-correlations that follow our hypotheses **(Section 4)**.

3. We describe how the performance of systems correlates with head and long tail cases, proving that the long tail phenomena and their interactions influence the performance of systems **(Section 5)**.

4. We provide recommendations on how to address the tail in future EL research (**Section 7**).

2 Related work

Analyzing well-known datasets for semantic tasks, Ilievski et al. (2016) measured very low ambiguity and variation, and a notable bias towards dominance. Overall, tasks and datasets show strong **semantic overfitting** to the head of the distribution (the popular part) and are not representative for the diversity of the long tail. In the related task of Word Sense Disambiguation, Postma et al. (2016) analyzed the impact of frequency on system accuracy, showing that the accuracy on the most frequent words is close to human performance, while the least frequent words can be disambiguated correctly in at most 20% of cases. According to Van Erp et al. (2016), EL datasets contain very little referential ambiguity and evaluation is focused on well-known entities (i.e., with high PageRank (Page et al., 1999) values).

NIL entities are entities without a representation in a knowledge base (Ji and Grishman, 2011). These are typically considered to have low frequencies within a corpus and/or to be domain-specific. Esquivel et al. (2017) report that around 50% of the people mentioned in news articles are not present in Wikipedia. Considering that Wikipedia and its structured data derivatives are almost exclusively used as an anchor in EL, this means that for half of all people mentions, the EL task is nonsensical. While NILs are a challenge that concerns the long tail of EL, in this work we focus on those entities that have been linked to Wikipedia, but are still infrequent, since this provides an entry for extensive analysis of their properties.

In this work, we distinguish dark, emerging, and domain entities. **Dark entities** are those for which no relevant information is present in a given knowledge base (Van Erp et al., 2015). Dark entities thus expand the notion of NIL entities to cases where an entity representation exists, but it is insufficient to reason over. **Emerging entities** are time-bound entities, recently unknown but potentially becoming popular in news in a short time (Hoffart et al., 2014). A body of work has dealt with **domain entities**, whose relevance is restricted within a topic, e.g. biomedical (Zheng et al., 2015), or historical domain (Heino et al., 2017). Dark, emerging, and domain entities mostly make up the tail in EL. Their definition is, however, orthogonal to our work: we strive to provide an umbrella theory of the tail in EL based on linguistic properties and avoid a discussion on defining the distinction between head or tail in a categorical way.

Finally, studying distributional properties of entity expressions and their linking, as we do in this study, is different from the classical linguistic studies on the distribution of words (Zipf, 1935; Corral et al., 2015; Kanwal et al., 2017). Linked entity data provides information on the surface forms, the meaning, and the referent of the surface form, whereas distributional studies on words only provide knowledge on the surface forms and to a limited extent on their sense, but never on their reference.

3 Approach

To address our research goal of **quantifying the long tail of EL**, we first explain the notions of ambiguity, variance, frequency, and popularity. Next, we formulate a set of hypotheses regarding their interaction and our expected influence on system performance. We also describe our choice of data collections and

[1] We consider the following properties: ambiguity of surface forms, variance of instances, frequency of surface forms, frequency of instances, and popularity of instances.

EL systems to analyze. The code of this analysis is available on Github at `https://github.com/cltl/EL-long-tail-phenomena`.

3.1 The long tail phenomena of the entity linking task

Each entity exists only once in the physical world. However, this is different in our communication where: 1. certain surface forms are very prominent and others occur only rarely; 2. certain instances are very prominent and others are mentioned incidentally. The task of EL covers a many-to-many relation between surface forms observed in text and their instances potentially found in a knowledge base. Surface forms and instances both have their own **frequency distribution**, pointing to the same underlying Grice (1975) mechanisms, governed by an envisioned trade-off between efficiency and effectiveness.

Surface forms have various frequency of occurrence in textual documents. Frequent surface forms include "U.S." and "Germany", but also "John Smith". The frequency of a surface form can be explained by its relation to one (or a few) very popular instances (`United States`), but it can also be a result of high **ambiguity** ("John Smith" is a common name, so it simply refers to many possible instances).[2]

Similarly, some **instances** are more popular and therefore more frequently mentioned than others. Note that *frequency* here refers to the number of occurrences in a corpus, while popularity refers to the frequency as a topic and can, for example, be measured by the volume of knowledge about an instance captured with its **PageRank** in a knowledge base. Frequent and popular instances are intuitively quite prominent and relevant, very often across many different circumstances, and are typically referred to by a relatively wide set of surface forms, resulting in a high **variance**. In addition, frequent/popular entities tend to participate in metonymy relations with other entities topically related to them. For instance, `United States` as a country relates to `United States Army` and to `United States Government` - two other entities of a different type, but possibly referenced by the same surface forms.

3.2 Hypotheses on the long tail phenomena of the entity linking task

We look systematically at the relation of surface forms in datasets and their instances in DBpedia, and provide a series of hypotheses regarding the long tail phenomena and their relation to system performance (Table 1). Some of these hypotheses, e.g., D1 and D2, are widely accepted as common knowledge but have rarely been investigated in EL datasets. Others, such as S4 and S5, are entirely new.

ID	Hypothesis	Sec
D1	Only a few forms and a few instances are very frequent in corpora, while most appear only incidentally.	4.1
D2	A few instances in corpora are much more popular (have much higher PageRank) compared to most other.	4.2
D3	Only a small portion of all forms in corpora are ambiguous.	4.3
D4	Only a small portion of all instances in corpora are referred to with multiple forms.	4.4
D5	There is a positive correlation between ambiguity of forms and their frequency.	4.5
D6	There is a positive correlation between variance of instances and their frequency.	4.5
D7	There is a positive correlation between variance of instances and their popularity.	4.5
D8	There is a positive correlation between popularity of instances and their frequency.	4.5
D9	The frequency distribution within all forms that refer to an instance is Zipfian.	4.6
D10	The frequency and the popularity distribution within all instances that refer to a form is Zipfian.	4.6
S1	Systems perform worse on forms that are ambiguous than overall.	5.1
S2	There is a positive correlation between system performance and frequency/popularity.	5.2
S3	Systems perform best on frequent, non-ambiguous forms, and worst on infrequent, highly ambiguous forms.	5.3
S4	Systems perform better on ambiguous forms with imbalanced, compared to balanced, instance distribution.	5.4
S5	Systems perform better on frequent instances of ambiguous forms, compared to their infrequent instances.	5.4
S6	Systems perform better on popular instances of ambiguous forms, compared to their unpopular instances.	5.4

Table 1: Hypotheses on the data properties (D*) and on their relation to system performance (S*)

[2]The notion of ambiguity captures the amount of instances to which a surface form has been observed to refer. Non-ambiguous forms are those that refer to a single instance, whereas ambiguous forms refer to at least two instances. Highly ambiguous forms refer to a wide array of instances.

3.3 Datasets and systems

We focus on well-known EL datasets with news documents, preferring larger sets with open licenses. Many customary EL datasets are however quite small ($< 1,000$ mentions). We opted to perform our analysis on the following two data collections, with five corpora in total:

AIDA-YAGO2 (Hoffart et al., 2011) - we consider its train, test A, and test B sets, summing up to 34,929 entity forms in 1,393 news documents, published by Reuters from August 1996 to August 1997.

N3 (Röder et al., 2014) is a collection of three corpora released under a free license. We consider the two N3 corpora in English: RSS-500 and Reuters-128. Reuters-128 contains economic news published by Reuters, while RSS-500 contains data from RSS feeds, covering various topics such as business, science, and world news. These two corpora consist of 628 documents with 1,880 entity forms in total.

We analyzed the EL performance of recent public and open-sourced entity linkers, as the state-of-the-art: 1. **AGDISTIS MAG** (Moussallem et al., 2017)[3] combines graph algorithms with context-based retrieval over knowledge bases. 2. **DBpedia Spotlight** (Daiber et al., 2013)[4] is based upon cosine similarities and a modification of TF-IDF weights. 3. **WAT** (Piccinno and Ferragina, 2014)[5] combines a set of algorithms, including graph- and vote-based ones.[6]

3.4 Evaluation

We apply the customary metrics of precision, recall, and F1-score to measure system performance in Section 5. In Table 2, we briefly describe the computation of true positives (TPs), false positives (FPs), and false negatives (FNs) per class. For example, if the gold instance belongs to C_1, then we count a TP when the system instance is also C_1. In case the system instance belongs to another class C_i, $i \neq 1$, this leads to a FN for C_1 and a FP for C_N. A special class are the NILs: predicting a NIL case incorrectly by the system results in a FN for the correct class; inversely, if the system was supposed to predict a NIL and it did not, then we count a FP. See the details in later sections for what constitutes a class. In our analysis we exclude the cases referring to NILs.

$G \backslash S$	C_1	...	C_N	NILL
C_1	TP(1)		FP(N), FN(1)	FN(1)
...				
C_N	FP(1), FN(N)		TP(N)	FN(N)
NILL	FP(1)		FP(N)	-

Table 2: Computation of TPs, FPs, and FNs per class C_1, ..., C_N. 'G'=gold instance, 'S'=system instance.

4 Analysis of data properties

4.1 Frequency distribution of forms and instances in datasets

We hypothesize that only a few forms and a few instances are very frequent in corpora, while most appear only incidentally (*D1*). This represents a variation of Zipf's law (Zipf, 1935) for the EL task.

The log-log frequency distributions of forms and instances (Figure 1a and 1b) show an almost ideal Zipfian distribution in the case of AIDA (with a slope coefficient of -0.9085 for forms and -0.9657 for instances) and to a lesser extent for N3 (a slope of -0.4291 for forms and -0.5419 for instances). The less Zipfian curves of N3 are probably because this data collection is significantly smaller than AIDA.

The similar shape of the form and the instance distribution per dataset can be explained by the dependency between the two aspects. Namely, the form 'U.S.' denoting the instance `United_States` 462 times is reflected in both the form and the instance distributions. However, these two distributions are only identical if the ambiguity and variance are both 1. In practice, the mapping between forms and instances is M-to-N, i.e., other forms also denote `United_States` (such as 'America') and there are other instances referred to by a form 'US' (such as `United_States_dollar`).

[3]AGDISTIS API: `http://akswnc9.informatik.uni-leipzig.de:8113/AGDISTIS`, used on 24/05/2018.

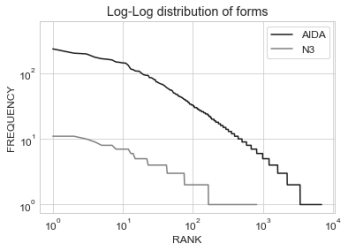
[4]Spotlight API: `http://spotlight.fii800.lod.labs.vu.nl/rest/disambiguate`, used on 24/05/2018.

[5]WAT API: `https://wat.d4science.org/wat/tag/json`, used on 24/05/2018.

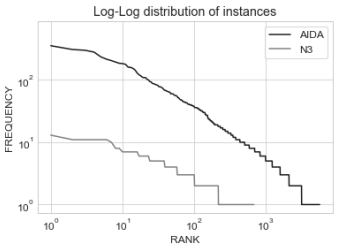
[6]All three APIs link to the Wikipedia dump from April 2016. Since the official DBpedia Spotlight endpoint at `http://model.dbpedia-spotlight.org/en/disambiguate` links to a newer Wikipedia version (February 2018 at the moment of writing of this paper), we set up our own endpoint that performs linking to the model 2016-04, to enable fair comparison with the other two systems. We reached similar conclusions with both versions of DBpedia Spotlight.

Figure 1: Log-log distribution of:

(a) form frequency



(b) instance frequency



(c) PageRank

4.2 PageRank distribution of instances in datasets

Similar to the instance frequency, we expect that a few instances in the corpora have an extremely high PageRank compared to most others (*D2*). Figure 1c shows the PageRank distribution of our both datasets. We observe that most entity mentions in text refer to instances with a low PageRank value, while only a few cases have a high PageRank value. Not surprisingly, the instance with the highest PageRank value (`United_States`) is at the same time the instance with the highest corpus frequency.

We inspect the effect of frequency and PageRank on system performance in Section 5.2.

4.3 Ambiguity distribution of forms

We hypothesize that only a small portion of all forms in a corpus are ambiguous (*D3*). As shown in Table 3, when both datasets are merged and NIL entities excluded, only 508 surface forms (6.73%) are ambiguous, as opposed to 7,037 monosemous forms (93.27%). This extremely high percentage validates our hypothesis. Moreover, in Sections 5.1, 5.3, and 5.4, we show that it also has a strong effect on systems performance.

	1	2	3	4	5	6	..	12
AIDA	6,400	359	78	29	7	3		1
N3	794	18	2	1	0	0		0
BOTH	7,037	381	84	29	10	3		1

Table 3: Ambiguity distribution per dataset, after NILs are excluded. Columns represent degrees of ambiguity.

4.4 Variance distribution of instances

We expect that only a small portion of all instances in a corpus are referred to with multiple forms (*D4*). The results of our variance analysis are given in Table 4. Over both datasets, 1,568 instances (25.61%) are referred to by multiple forms, as opposed to 4,555 instances (74.39%) which are always referred to by the same form. While the distribution of variance is much more even compared to that of ambiguity, we observe that most of the instances have a variance of 1.[7]

	1	2	3	4	5	6	7	8	9	10	11
AIDA	4,156	1,118	230	56	19	10	6	0	1	1	1
N3	550	106	15	7	1	0	0	0	0	0	0
BOTH	4,555	1,206	247	74	22	10	6	0	0	2	1

Table 4: Variance of instances with respect to the number of surface forms that reference them. Columns represent degrees of variance.

4.5 Interaction between frequency, PageRank, and ambiguity/variance

In the previous four Sections we analyzed the frequency distribution of individual data properties. Here we move forward to analyze their interaction. Figure 2 shows these results with mean as an estimator.[8]

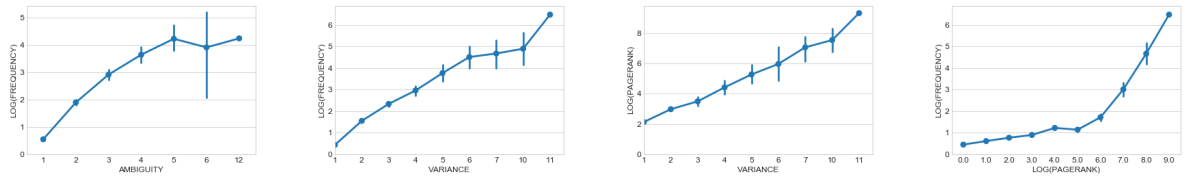
Firstly, we predict a positive dependency between ambiguity of forms and their frequency (*D5*). We expect that frequently used forms tend to receive new meanings, often as a result of metonymy or meronymy

[7] We expect that this skewness will also dominate system performance. For space reasons, we have not included this analysis.

[8] We observed comparable results with median as an estimator.

Figure 2: Correlations between long tail phenomena (estimator=mean):

(a) Ambiguity and frequency. (b) Variance and frequency. (c) Variance and PageRank. (d) PageRank and frequency.



of their dominant meaning. Figure 2a confirms this tendency, the Spearman correlation being 0.3772.

Secondly, we expect a positive correlation between variance of instances and their frequency (*D6*) or popularity (*D7*). Frequently mentioned and popular instances tend to be associated with more forms. Indeed, we observe that instances with higher frequency (Figure 2b) or PageRank (Figure 2c) typically have higher variance. The Spearman correlations are 0.6348 and 0.2542, respectively.

Thirdly, we compare the frequency of instances to their popularity measured with PageRank, predicting a positive correlation (*D8*). On average, this dependency holds (Figure 2d), though there are many frequent instances with low PageRank, or vice versa, leading to a Spearman correlation of 0.3281. The former are instances whose prominence coincides with the creation time of the corpus, but are not very well-known (e.g., the now-retired football player `Predrag_Mijatovic`). The latter are generally popular entities which were not captured sufficiently by the corpus, because their topical domain is marginal to this corpus (e.g., scientists), or they became relevant after the corpus release (emerging entities).

Hence, besides the high corpora skewness in terms of frequency, popularity, ambiguity, and variance, these factors also have positive interdependencies. Section 5 shows their effect on system performance.

4.6 Frequency distribution within a synset

We observed that the distribution of form frequency, instance frequency, and PageRank all have a Zipfian shape. But do we also see this behavior on a single form or instance level?

Supposedly, the frequency distribution within all forms that refer to an instance is Zipfian (*D9*). We test *D9* on the instance with highest variance and frequency, `United_States` (Figure 3). As expected, the vast majority of forms are used only seldom to refer to this instance, while in most cases a dominant short form "U.S." is used to make reference to this entity.

Figure 3: Form frequencies for the instance `United_States`

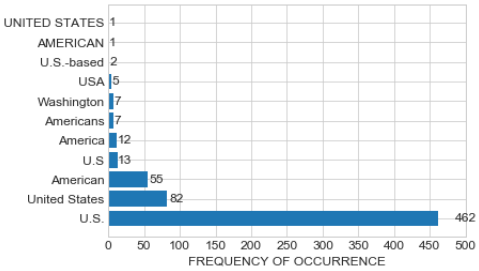


Figure 4a presents the frequency distribution of all instances that are referred to by the most ambiguous form in our data, "World cup". Figure 4b shows their PageRank distribution. In both cases, we observe a long-tail distribution among these instances (*D10*). Comparing them, we observe a clear match between frequency and PageRank, deviating only for instances that were prominent during the corpus creation, like `1998_FIFA_World_Cup`.

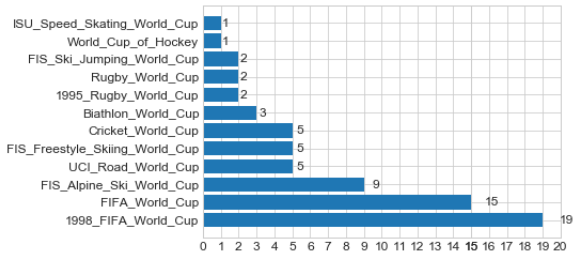
For analysis on the effect of frequency and popularity on system performance, please see Section 5.2.

5 Analysis of system performance and data properties

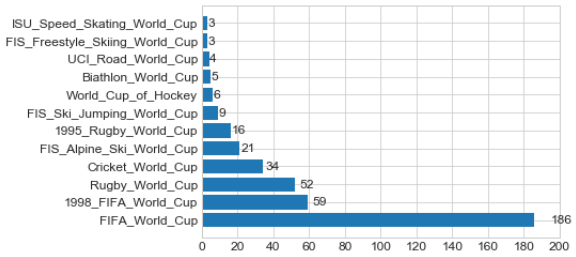
Next, we analyze systems performance in relation to the data properties: ambiguity (Section 5.1), form frequency, instance frequency, and PageRank (Section 5.2), as well as their combinations (5.3 and 5.4).

Figure 4: Distributions of the instances denoted by the most ambiguous form ("World Cup")

(a) Frequency distribution.



(b) PageRank distribution.

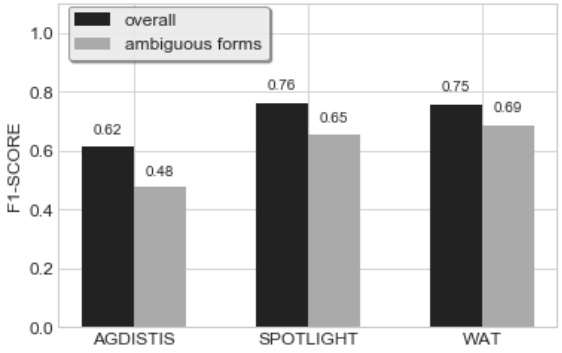


5.1 Correlating system performance with form ambiguity

Figure 5 displays the micro F1-scores of AGDISTIS MAG, Spotlight, and WAT on the two data collections jointly. For each system, we show its overall F1-score and F1-score on ambiguous forms only.

Section 4.3 showed that most of the forms in our corpora are not ambiguous. We expect that these forms lift the performance of systems, i.e., that they can resolve non-ambiguous forms easier than ambiguous ones (*S1*). Figure 5 confirms this for all systems: the F1-score on ambiguous forms is between 6 and 14 absolute points lower than the overall F1-score.

Figure 5: Micro F1-scores of systems: overall and on ambiguous subsets



When computing macro- instead of micro-F1 scores, we observe similar findings for *S1*. Interestingly, the macro-F1 scores are consistently lower than the micro-F1 scores, especially in case of the ambiguous subsets evaluation. Namely, the overall macro-F1 scores are between 0.44 and 0.52, and between 0.14 and 0.34 on the ambiguous forms. This suggests that frequent forms boost system performance, especially on ambiguous surface forms. We investigate this further in the next Sections.

5.2 Correlating system performance with form frequency, instance frequency, and PageRank

Next, we consider frequency of forms and instances, as well as PageRank values of instances in relation to system performance. For each of these, we expect a positive correlation with system performance (*S2*), suggesting that systems perform better on frequent and popular cases, compared to non-popular and infrequent ones.

	all forms			ambiguous forms only		
	FF-F1	**FI-F1**	**PR-F1**	**FF-F1**	**FI-F1**	**PR-F1**
AGDISTIS	0.2739	0.3812	0.1465	0.3550	0.4073	0.3969
Spotlight	0.1321	0.1847	0.1357	0.3986	0.4196	0.3108
WAT	0.4663	0.5050	0.3164	0.5831	0.5319	0.4214

Table 5: Correlation between F1-score and: frequency of forms (FF-F1), frequency of instances (FI-F1), and PageRank (PR-F1). Left: on all forms, right: only on ambiguous forms.

The Spearman correlation for each of the systems and properties, over all forms, are shown in Table 5 (left half). While most of the correlation for frequency and popularity is positive, the values are in general relatively low (WAT being an exception). This shows that frequency/popularity by itself contributes, but is not sufficient to explain system performance. The right half of the Table shows the same metrics when applied to the ambiguous forms. We observe an increase in all values, which means that frequency and popularity are most relevant when multiple instances 'compete' sharing a form. These findings are in line with those in Section 5.1.

5.3 Correlating system performance with ambiguity and frequency of forms jointly

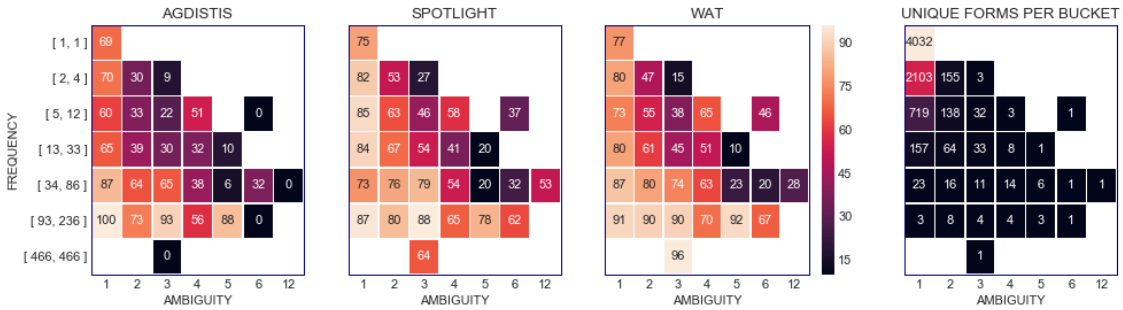


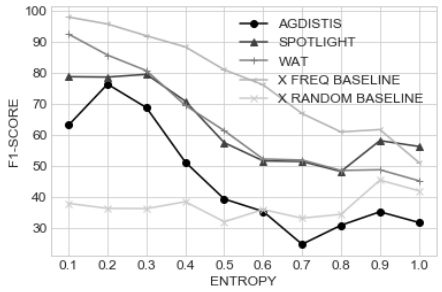
Figure 6: F1 per (ambiguity, frequency) pair. The first three heat maps show the F1-scores of systems per ambiguity degree and frequency range. The last plot shows the amount of unique forms in each cell.

We have shown that system performance is lower on ambiguous than unambiguous forms. We also observed a tendency of systems to perform better on frequent forms and instances as compared to infrequent ones. But how does performance differ across different levels of ambiguity? How do ambiguity and form frequency interact as a joint predictor of performance? The heat maps in Figure 6 show the interplay between ambiguity, frequency, and micro F1-score for each of the systems. Generalizing over the discussion in Sections 5.1 and 5.2, we observe the best scores on frequent, non-ambiguous forms (bottom-left), and worst F1-scores on non-frequent, highly ambiguous forms (top-right) (*S3*).

In Section 5.4, we investigate if some instances within ambiguous forms are more difficult than others.

5.4 Correlating system performance with frequency of instances for ambiguous forms

Figure 7: Micro F1-score per entropy bucket.



To measure the instance distribution within individual forms, we employ the notion of *normalized entropy*, similarly as Ilievski et al. (2016). The entropy of a form i with n_i instances and N_i occurrences is: $H_i = (-\sum_{j=1}^{n_i} p_{i,j} log p_{i,j})/log_2 N_i$, where $p_{i,j}$ is the probability that the instance j is denoted by the form i. For non-ambiguous forms $H_i = 0$, while forms with uniform frequency distribution of instances have a maximum $H_i = 1$. We predict an inverse correlation between system performance and entropy (*S4*). The results in Figure 7 show a dramatic drop in micro F1-score for uniformly distributed cases (high entropy) compared to skewed ones (low entropy). We compare these shapes to two baselines: frequency baseline, that picks the most frequent instance for a form on the gold data, and a random baseline, choosing one of the gold instances for a form at random. All three systems have a similar curve shape to the frequency baseline, whereas out of the three systems Spotlight's curve comes closest to that of the random baseline.

We also compute the macro F1-score per entropy bucket to help us understand whether the drop in performance in Figure 7 is due to: 1. a qualitative difference between low and high entropy forms, or 2. an overfitting of systems to the frequent interpretations of ambiguous forms. The macro F1-score reduces the effect of frequency on performance, by evaluating each form-instance pair once. We observe that the macro F1-scores are much more balanced across the entropy buckets compared to the micro F1-scores, and especially lower on the buckets with higher skewness (low entropy). This suggests that the high micro F1-score for low entropies is heavily based on frequent instances.

As a final analysis, we seek to understand whether frequent/popular instances of a form are indeed resolved easier than less frequent/popular instances of the same form. For that purpose, we pick the set of all ambiguous forms, and we rank their instances by relative frequency/PageRank value.

Considering the most ambiguous form "World Cup" as an example and ranking by frequency, its $r1$ (rank 1) instance is `1998_FIFA_World_Cup`, $r2$ is `FIFA_World_Cup`, ..., and $r12$ is `ISU_Speed_Skating_World_Cup`. We expect systems to perform much better on frequent instances of ambiguous forms, compared to infrequent instances of ambiguous forms, i.e., we expect F1-scores to decrease when moving from higher to lower ranks (*S5*). Figure 8 shows that our hypothesis holds to a large extent for precision, recall, and F1-scores, except for the occasional peaks at $r6$ and $r9$.

Figure 8: Precision, recall, and micro F1-score per instance frequency rank, averaged over the systems.

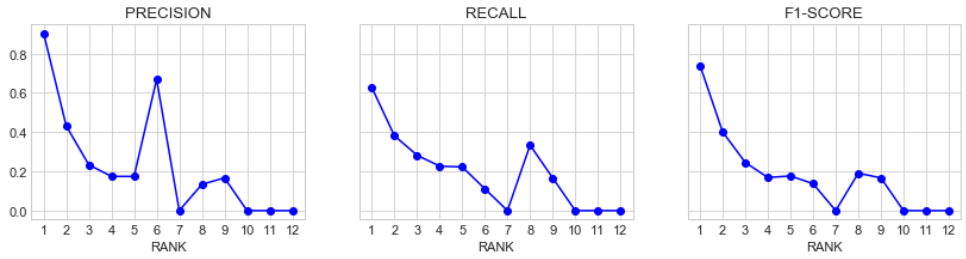
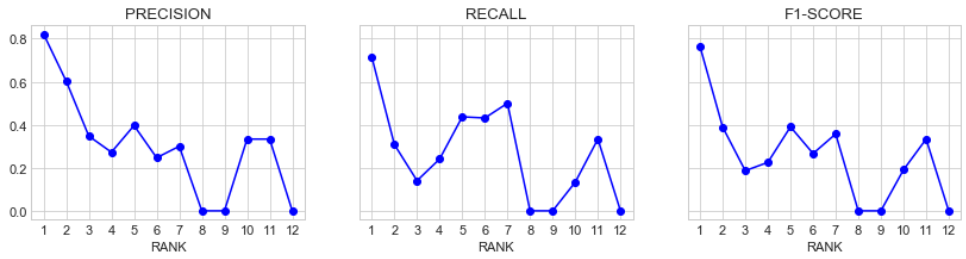


Figure 9: Precision, recall, and micro F1-score per PageRank-based rank, averaged over the systems.



Similarly, we order the instances denoting a form based on their relative PageRank value. We hypothesize that systems perform better on popular instances of ambiguous forms, compared to their unpopular instances (*S6*). Although less monotonic than the frequency ones in Figure 8, the resulting shapes of this analysis in Figure 9 suggest that popularity can also be applied to estimate system performance.

6 Summary of findings

We noted a positive correlation between ambiguity and frequency of forms, as well as between variance and frequency of instances. We noticed that the distribution of instances overall, but also per form, has a Zipfian shape. Similarly, the distribution of forms, both on individual and on aggregated level, is Zipfian. While some of these distributions are well-known in the community for words, this is the first time they have been systematically analyzed for surface forms of entities, their meaning and reference, and empirically connected with the performance of systems.

We observed that ambiguity of forms leads to a notable decline in system performance. Coupling it with frequency, we measured that low-frequent, highly ambiguous forms yield the lowest performance, while very frequent, non-ambiguous forms yield the highest performance. The entropy of forms, capturing the frequency distribution of their denoted instances, revealed that balanced distributions tend to be harder for systems, with the micro F1-value dropping with 20-40 absolute points between the highest and lowest entropy. Finally, the higher performance on skewed cases was shown to be a result of overfitting to the most frequent/popular instances.

Based on these outcomes, we can conclude that the intersection of ambiguity and frequency/popularity is a good estimator of the complexity of the EL task. The hard cases of EL should be sought among the low-frequent and unpopular candidates of highly ambiguous forms.

7 Recommended actions

We have shown that there are systematic differences between the head and the tail of the EL task, and that these reflect on how systems perform. Provided that systems show a weakness on tail cases, and that this weakness is simultaneously hidden by averaged evaluation numbers, how can we overcome this obstacle in practice? Here we list three recommendations:

1. When **creating a dataset**, we propose authors to include statistics on the head and the tail properties (ambiguity, variance, frequency, and popularity) of the data, together with a most-frequent-value baseline. By doing so, the community would be informed about the hard cases in that dataset, as well as about the portion of the dataset that can be resolved by following simple statistical strategies.

2. When **evaluating** a system, we suggest splitting of all cases into head and tail ones. Afterwards, head and tail cases can be evaluated separately, as well as together. This provides a direct insight into the differences in scoring of the tail cases compared to the head cases, potentially signaling aspects of the EL tail that are challenging for the given system. In addition, the frequency skewness of head cases can be largely decreased by employing a macro instead of micro F1-score, as shown in this paper.

3. In addition to the suggestion in 2., when **developing or training** a system, it should be made explicit which heuristics target which cases, and to what extent resources and training data optimize for the target dataset in relation to the head and tail distributions.

8 Conclusions

Although past research has argued that the performance of EL systems declines when moving from the head to the tail of the entity distribution, the long tail has not been quantified so far, preventing one to distinguish head and tail cases in the EL task. Previous linguistic studies on words distributions can also not be applied for this purpose since they do not study reference. This paper is the first one that systematically looks into the relation of surface forms in EL corpora and instances in DBpedia, and provides a series of hypotheses on what long tail phenomena are. We analyzed existing EL datasets with respect to these long tail properties, demonstrating that data properties have certain inter-correlations that follow our hypotheses. Next, we investigated their effect on the performance of three state-of-the-art systems, proving that the long tail phenomena and their interaction consistently predict system performance. Namely, we noted a positive dependency of system performance on frequency and popularity of instances, and a negative one with ambiguity of forms. Our findings in this paper are meant to influence future designs of both EL systems and evaluation datasets. To support this goal, we listed three recommended actions to be considered when creating a dataset, evaluating a system, or developing a system in the future.

We see two directions for future improvement of our analysis: 1. To obtain a corpus-independent inventory of forms and their candidate instances, both with their corresponding frequencies, is a challenge in the case of EL and no existing resource can be assumed to be satisfactory in this regard (for Word Sense Disambiguation, this is usually WordNet). We approximated these based on the corpora we analyzed, but considering the fairly small size of most EL datasets, this poses a limitation to our current analysis. 2. Some of our current numbers are computed only on a handful of cases. This leads to unexpected disturbances in our results, like the occasional peaks for high ranks in Figure 8. We expect the outcome of this analysis to gain significance when more large EL datasets become available in the future.

Acknowledgements

We would like to thank Eduard Hovy, Frank van Harmelen, and Marieke van Erp for their valuable suggestions. In addition, we thank the anonymous reviewers for their feedback. The research for this paper was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza fund.

References

Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. 2015. Zipfs law for word frequencies: Word forms versus lemmas in long texts. *PloS one*, 10(7):e0129031.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM.

José Esquivel, Dyaa Albakour, Miguel Martinez, David Corney, and Samir Moussa. 2017. On the Long-Tail Entities in News. In *European Conference on Information Retrieval*, pages 691–697.

H Paul Grice. 1975. Logic and conversation. *1975*, pages 41–58.

Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. 2017. Named entity linking in a complex domain: Case second world war history. In *International Conference on Language, Data and Knowledge*, pages 120–133. Springer.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd international conference on World wide web*, pages 385–396. ACM.

Filip Ilievski, Marten Postma, and Piek Vossen. 2016. Semantic overfitting: what world do we consider when evaluating disambiguation of text? In *proceedings of COLING*.

Filip Ilievski, Piek Vossen, and Marieke Van Erp. 2017. Hunger for Contextual Knowledge and a Road Map to Intelligent Entity Linking. In *International Conference on Language, Data and Knowledge*, pages 143–149. Springer.

Heng Ji and Ralph Grishman. 2011. Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1148–1158. Association for Computational Linguistics.

Jasmeen Kanwal, Kenny Smith, Jennifer Culbertson, and Simon Kirby. 2017. Zipfs Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.

Diego Moussallem, Ricardo Usbeck, Michael Röeder, and Axel-Cyrille Ngonga Ngomo. 2017. MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach. In *Proceedings of the Knowledge Capture Conference*, page 9. ACM.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Francesco Piccinno and Paolo Ferragina. 2014. From TagME to WAT: a new entity annotator. In *ERD@SIGIR*.

Marten Postma, Ruben Izquierdo, Eneko Agirre, German Rigau, and Piek Vossen. 2016. Addressing the MFS Bias in WSD systems. In *Proceedings of LREC 2016*, Paris, France. ELRA.

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3-A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format. In *LREC*, pages 3529–3533.

Marieke Van Erp, Filip Ilievski, Marco Rospocher, and Piek Vossen. 2015. Missing Mr. Brown and Buying an Abraham Lincoln-Dark Entities and DBpedia. In *NLP-DBPEDIA@ ISWC*, pages 81–86.

Marieke Van Erp, P Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jörg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC. ELRA*.

Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC medical informatics and decision making*, 15(1):S4.

George Zipf. 1935. *The Psychobiology of Language: An Introduction to Dynamic Philology*. M.I.T. Press, Cambridge, Mass.