On Emerging Entity Detection

Michael Färber, Achim Rettinger*, and Boulos El Asmar

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{michael.faerber,rettinger}@kit.edu
boulos.el-asmar@bmw.de

Abstract. While large Knowledge Graphs (KGs) already cover a broad range of domains to an extent sufficient for general use, they typically lack emerging entities that are just starting to attract the public interest. This disqualifies such KGs for tasks like entity-based media monitoring, since a large portion of news inherently covers entities that have not been noted by the public before. Such entities are unlinkable, which ultimately means, they cannot be monitored in media streams. This is the first paper that thoroughly investigates all types of challenges that arise from out-of-KG entities for entity linking tasks. By large-scale analytics of news streams we quantify the importance of each challenge for real-world applications. We then propose a machine learning approach which tackles the most frequent but least investigated challenge, i.e., when entities are missing in the KG and cannot be considered by entity linking systems. We construct a publicly available benchmark data set based on English news articles and editing behavior on Wikipedia. Our experiments show that predicting whether an entity will be added to Wikipedia is challenging. However, we can reliably identify emerging entities that could be added to the KG according to Wikipedia's own notability criteria.

Keywords: Emerging Information Discovery, Evolving Knowledge, Novelty Detection, Entity Linking, Text Annotation.

1 Introduction

Although existing knowledge graphs (KGs) such as DBpedia, Wikidata, and YAGO are already quite powerful in terms of their size, they are inherently incomplete, since they contain concepts and facts of an ever-changing world: constantly, new knowledge needs to be added. Considering Wikipedia, each day over 700 new articles are added to the English Wikipedia which stay permanently.[1] We do regard those articles as *novel entities* w.r.t. Wikipedia, as each article describes one entity which has not been part of Wikipedia before.

* The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

[1] This fact results from our empirical analysis, see Section 2.2 for more details.

In this work we attempt to automatically identify such *out-of-KG entities* that are of great importance to numerous time-sensitive tasks which require up-to-date KGs, like semantic media monitoring or automatic speech recognition for TV news. Clearly, not all entities that potentially will get added to Wikipedia should be reported to someone who is interested in breaking news, since entities like *Antonio Ferramolino*, a 16th century Italian architect, were added recently to Wikipedia but not because of a current newsworthy event.

To take this into account we identify a crucial condition that makes an out-of-KG entity a potential candidate for news monitoring tasks: it needs to be *trending* and become *notable* for *the first time*. Those out-of-KG entities which show a notable increase in public interest for the first time are thereby referred to as *emerging*[2]. Hoffart et al. [10] use the same term, however, they do not require out-of-KG entities to be *notable*. *Notability* is an officially specified requirement for novel articles on Wikipedia. In order to determine whether an entity is notable, Wikipedia provides *notability guidelines*[3] to editors. An entity is thereby regarded as notable if it is traceable by reliable secondary literature.

Firstly, this paper presents the first full picture of how missing surface forms (i.e., phrases by which entities are referred to in text) and missing KG-entities impact the task of entity linking. Our empirical analysis of news streams in combination with the editing behavior on Wikipedia reveals that the need for incorporating emerging out-of-KG entities for entity linking occurs frequently, but is least investigated in existing research. Secondly, we create a publicly available benchmark data set and present a machine learning approach to automatically predict emerging out-of-Wikipedia entities. Our assumption is that emergence can be measured by analyzing media streams such as online news. The results show that making predictions about which entities will actually be added to Wikipedia is tricky. However, we are able to identify (actual) emerging entities with high confidence that could be added to the KG according to Wikipedia's own notability criteria. Those entities can be suggested to Wikipedia editors for inserting them into Wikipedia which helps to keep Wikipedia up-to-date with current events.

In summary, we make the following contributions in this paper:

- We describe and formalize the different entity linking challenges arising from out-of-KG entities and surface forms.
- We examine the occurrence and importance of the entity linking challenges regarding emerging entities "in the wild," i.e., based on Wikipedia as KG and annotated English news articles.
- We provide the first public benchmark data set for the *Emerging Entity Detection* challenge.[4]
- We present the first approach for predicting emerging entities based on the history of Wikipedia edits and noun phrases extracted as potential entity mentions from news streams.

[2] *Emerging* relates to *trending*: Entities can *emerge* only once. Once they have become *notable*, any (repeated) increase in public interest is just a *trend*.

[3] See `https://en.wikipedia.org/wiki/Wikipedia:Notability`.

[4] See `http://people.aifb.kit.edu/mfa/emerging-entity-detection/`.

The remainder of this paper is organized as follows: In Section 2, we analyze the conceptually different challenges for entity linking resulting from out-of-KG entities and out-of-KG surface forms In Section 3, the previous work is reported in respect to each challenge. Finally, we introduce our approach for emerging entity prediction in Section 4 and conclude in Section 5.

2 Entity Linking Challenges Arising from Missing Entities and Missing Surface Forms

We first clarify some terminology:

- An *entity* is a thing which can be uniquely identified via a URI $u \in U$.
- A *Knowledge Graph* is an RDF graph, which consists of a set of RDF triples where each RDF triple (s, p, o) is an ordered set of the following RDF terms: a subject $s \in U \cup B$, a predicate $p \in U$, and an object $U \cup B \cup L$. An RDF term is either a URI $u \in U$, a blank node $b \in B$, or a literal $l \in L$. U, B, and L are pairwise disjoint. In this paper, we do not consider blank nodes.
- A *surface form* is a textual phrase referring to one or several specific entities (e.g., the title of a Wikipedia article). Each entity has none, one or several surface forms attached.
- If a surface form is mentioned in a text, we speak of a *mention* of an entity. The task of linking a mention to a KG entity is referred to as *entity linking*.
- The tuple of a mention and a corresponding entity in a KG is designated by us as *annotation*.
- Entities and surface forms can be present in the KG or not. In the latter case we call them *unknown*, *missing*, or *out-of-KG*. If an out-of-KG entity is trending and notable for the first time, we call it *emerging*. In the moment an entity is inserted into the KG, it is regarded as *novel*.

2.1 Overview of Challenges

Let the following be given:

- KG g at the time t_0 (e.g., Wikipedia at 2015-04-04) with the set of all entities E_{t_0} and for each entity $e \in E_{t_0}$ the set of associated surface forms $S^e_{t_0}$.
- KG g' at time t_1 (e.g., Wikipedia at 2015-05-15) with the set of – in comparison to g – newly added entities $E_{\Delta t}$ since t_0 (i.e., $\Delta t = t_1 - t_0$) and for each entity $e \in (E_{t_0} \cup E_{\Delta t})$ the set of surface forms of e, $S^e_{\Delta t}$, added within Δt. S^e_{t1} is then the set of surface forms for entity e at time t_1.[5]
- Mention $m \in M$ in a text document given in the time range Δt.
- Function $f : M \to (E_{t_0} \cup E_{\Delta t})$, indicating the correct entity linking of $m \in M$ to an entity $e \in (E_{t_0} \cup E_{\Delta t})$.

[5] As we are interested in novel/emerging entities, we do not consider deletions of entities or surface forms within Δt.

We can then differentiate between the following disjoint challenges for entity linking w.r.t. missing entities and entity surface forms in a KG (in the following called *Challenges*; see also Figure 1). We thereby first describe each Challenge, before we define it formally for a given mention m.

Challenge 1: Known surface form, known entity. This is the regular task of entity linking, i.e. without any aspect of missing entries. For a given mention in the text, one or several entities exist in the KG whose surface forms match the mention. With the help of a word-sense-disambiguation method, the appropriate entity in the KG is selected for the annotation of the mention. In our example in Figure 1, "Snowden", the person, is chosen and not the band.

$$\exists e \in E_{t_0} : (m \in S^e_{t_0} \wedge f(m) = e) \wedge \nexists e' \in E_{\Delta t} : m \in S^{e'}_{t_1}$$

Challenge 2: Unknown surface form, known entity. Given the mention in the text, no surface form can be found in the KG that matches the mention and, hence, the mention cannot be linked. However, the entity which should be linked to, already exists in the KG. Missing surface forms regarding this situation can be either small word variations (different (mis)spellings, abbreviations, or substrings; e.g., "Arslonbob" for `:Arslanbob`) or completely new word creations which emerge (e.g., "Kybella" for `:Deoxycholic_acid`).

$$\nexists e \in E_{t_0} : m \in S^e_{t_0} \wedge \exists e' \in E_{t_0} : (m \in S^{e'}_{\Delta t} \wedge f(m) = e')$$

Challenge 3: Known surface form, unknown entity. Here, when using a regular entity linking tool, the given mention in the text might be falsely linked to an existing entity in the KG, since this entity has a surface form which is matching (e.g., "Alphabet"). However, the mention actually refers to an entity which does not exist in the KG yet (e.g., the company `:Alphabet_Inc.`).

$$\exists e \in E_{t_0} : m \in S^e_{t_0} \wedge \exists e' \in E_{\Delta t} : (m \in S^{e'}_{t_1} \wedge f(m) = e')$$

Challenge 4: Unknown surface form, unknown entity. Given the mention in the text, none of the known surface forms of the entities in the KG can be matched and, hence, the mention cannot be linked. Also the entity to be linked to is unknown. Examples are `:Antonio_Ferramolino` and `:41st_G7_summit`.

$$\nexists e \in E_{t_0} : m \in S^e_{t_0} \wedge \exists e' \in E_{\Delta t} : (m \in S^{e'}_{\Delta t} \wedge f(m) = e')$$

2.2 Challenges in the Wild

We can now analyze the above mentioned EL Challenges by monitoring entity mentions in news streams and relating them to the editing behavior in Wikipedia. Due to page limit constraints, we thereby focus primarily on the task of Emerging Entity Detection (i.e., Challenge 4 with emerging entities).

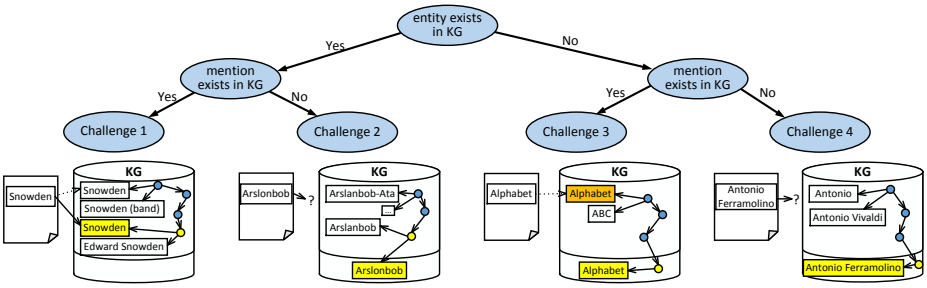

Fig. 1: Four different challenges arising in entity linking tasks when novel entities and novel surface forms start to appear.

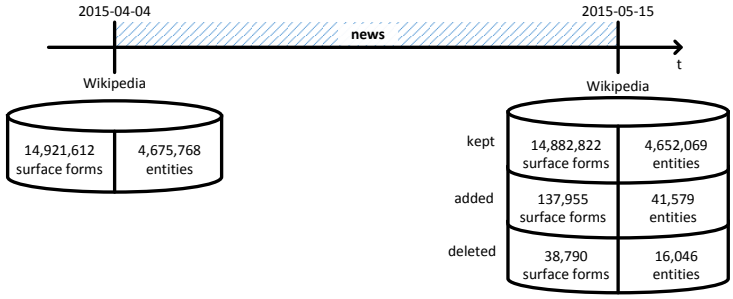

Fig. 2: Timeline with the Wikipedia versions and news used.

Experimental Setup

Wikipedia: Given the Wikipedia states from two different points in time, we first form the set of distinct entities and the set of distinct surface forms for both Wikipedia versions (using the xLiD framework [21] where the surface forms are derived by the title of Wikipedia pages, the redirect pages, the disambiguation pages, and the anchor texts of hyperlinks in Wikipedia). We can now calculate the difference between these two Wikipedia versions which identifies the novel entities and surface forms. The result is depicted in Figure 2. We can see that 41,579 entities are in the version of May 2015, but not in the earlier version of April 2015. Also, 137,955 surface forms have been added to Wikipedia in this time range. While the major part of these new surface forms belongs to "old" entities, there still are many that correspond to "novel" entities which were added in the time range.

News: To assess the importance of each entity linking challenge (as defined in Section 2.1), we need to tap into another data source where traditional entity linking suffers from the mentioned challenges. We choose news articles to investigate how often each entity linking challenge occurs in a real-world news stream. We gather all English news articles from the IJS newsfeed service [17], covering more than 30,000 English news sources within this time range. This results in 1,966,540 English news articles in total. We annotate the news

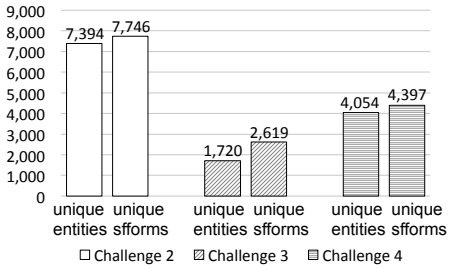

Fig. 3: Number of unique entities and number of unique surface forms used in the annotation. Challenge 1 is not displayed as it neither deals with novel entities nor with new surface forms.

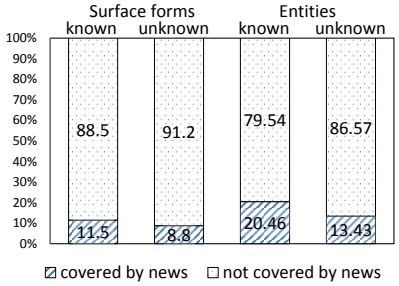

Fig. 4: Proportions to which known and novel entities/surface forms of the KG were also detected in the news, considering the annotations of all Challenges.

articles with Wikipedia entities using the x-LiSA tool [22], a state-of-the-art entity linking system, given the Wikipedia state of May 15, 2015. For this setting, we gain 205,225,526 annotations in total. Given the *diff* of entities and surface forms between the mentioned Wikipedia states, we can now calculate the distribution among the Challenge 1, 2, 3 and 4.

Observations and Discussions

Frequency of Unique Entities and Unique Surface Forms per Challenge: Fig. 3 shows (i) to how many unique novel entities the detected entity mentions link and (ii) how many unique new surface forms were found as mentions. We can observe that Challenge 2 covers more distinct entities than Challenge 4 and that then Challenge 3 follows (always with considerable differences). Apparently, apart from the annotations of Challenge 1, most frequently only new surface forms of already existing KG entities are used in the annotations. This is reasonable since our KG Wikipedia already covers millions of entities and it is likely that a part of these entities get new surface forms which occur as mentions in the news.

Considering Challenge 3 and 4, entities of Challenge 4 are more often linked than entities of Challenge 3. This is comprehensible: Novel entities are more likely to have a surface form which is not existing in the KG so far than having a surface form which is already known. In the latter case, additional entities for existing surface forms are added to the KG, intensifying the ambiguity problem.

Proportion of Named Entities Among Novel Entities: To get a better characterization of the novel entities, we approach the question: "How many novel entities (which had been inserted within the time range) are named entities?" Bunescu and Pasca [2] have developed heuristics for determining whether a Wikipedia entity is a named entity. In order to answer our question, we implemented these heuristics and gained 33,052 named entities and 8,523 non-named entities.[6] Our evaluation on this classification (given a sample of

[6] The remaining few entities are not parseable by the Stanford parser.

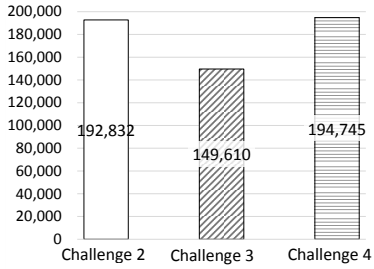

Fig. 5: Number of all annotations w.r.t. the different Entity Linking Challenges for the given time range.

300 manually classified, randomly chosen novel entities) revealed an accurracy of 85.67% for the chosen NER classification method. Note that in the manual evaluation we tended to classify more Wikipedia entities as named entities than the heuristics. For instance, we also considered events which can be given unique names as named entities.[7] In total we can state that focussing on *named* novel entities might be sufficient, especially for emerging entity detection tasks.

Proportion of Emerging Entities and Emerging Surface Forms in the KG: Fig. 4 shows on the right side the ratio of the number of novel entities and surface forms, respectively, detected in the news and being in the KG at time t_1 (and not yet in the KG at time t_0) to the total number of novel entities and surface forms, respectively, being in the KG at time t_1. 13.43% of the novel entities are found as annotations in the news and 8.80% of the new surface forms appear as mentions in the news. We can assume that these 13.43% of the novel entities are the ones which are of highly public interest (i.e., *emerging*), since they occur in the news. On average, each novel entity appears 45.55 times. Mostly, those emerging entities were categorized in Wikipedia at time t_1 under the non-disjoint categories of living people, dead people (especially dying in the given time range Δt), and politicians.

Frequency of Annotations per Entity Linking Challenge: Fig. 5 shows the number of all annotations per Challenge for the given time range. Note, that the reported occurrence numbers deliberately include repeated mentions of the same entity or surface form, since here we want to assess the total number of successfully linked mentions in the news. First of all, all Challenges occur considerably often. Challenge 1, as expected, happens most frequently (204,675,773 occurrences; not depicted due to the high number), since Wikipedia even at the early point in time already covers millions of entities and grows only by 41,579 novel entities in the given time range.[8] Then, Challenge 4 follows in the

[7] Given the set of 300 novel entities manually tagged as named entities, 95 of them got classified as of type PERSON, 51 of type LOCATION, 27 of type ORGANIZATION, and 24 of type EVENT (as subtype of MISC).

[8] For 11,639 of those 41,579 novel entities, however, only the Wikipedia title or redirects changed (due to typo correction or outsourcing of parts of a page). I.e.,

occurrence ranking. In contrast to the distribution of unique novel entities and new surface forms (see Fig. 3), annotations of Challenge 4 appear more frequent than annotations of Challenge 2. This shows that our news stream captures novel entities (detected in Challenge 3/4) well and over time (on average 45.55 times per novel entity, see above).

Persistence of Detected Emerging Entities: Considering the annotations where novel entities were found (i.e. Challenges 3 and 4), almost all of those entities still exist in the current Wikipedia version (as of July 2016; above 99% regarding all Challenges). Those entities seem to be permanently relevant for the KG, which is a strong indicator for the importance of emerging entities and of their detection (see Section 4).

2.3 Conclusions

The results of our analysis provide interesting insights into novel Wikipedia entities and surface forms and how they appear in the news. Below is a selection of key findings that we believe are most revealing:

1. Challenge 4 covers most of the novel entities inserted into the KG in Δt. In addition, Challenge 4 occurs – besides Challenge 1, which does not cover any novelty – most frequent regarding the set of all annotations. Thus, when dealing with novel entities, Challenge 4 is the most pressing issue to address.
2. About 13.4% of the novel Wikipedia-entities are also mentioned in the news. Since those entities start to be mentioned in the news with increasing frequency at a certain point in time (occurring on average 45.55 times), we assume they are *emerging* entities, i.e., of increasing public interest. This clearly shows that *emerging* entity detection is different from *novel* entity detection and should not be treated equally as done by previous work [10].
3. Furthermore, we found out that almost all emerging entities remain in Wikipedia constantly. Together with the item 2 above, i.e. the frequent occurrence of emerging entities in the news, it indicates the great importance of emerging entities for being in the KG and for being detected as early as possible (see Section 4).
4. About 75% of the novel entities are named entities. This indicates that focusing on named entities might be sufficient for many real-world novel entity detection applications. Emerging entities are most frequently living people which are of public interest (e.g., politicians) or people who recently died.

On our website, we present further results of our analysis, such as the string similarity between the mention and already given surface forms of the target entities.

on average over 700 entities are inserted into Wikipedia each day which are "really" novel. For the task of Emerging Entity Detection (see Section 4), we only consider real novel entities which emerge (i.e., recently gained public interest for the first time).

3 Related Work

In the following, we describe, how the different entity linking challenges w.r.t. novelty have been pursued by the research community. Due to the focus of this paper on Emerging Entity Detection in Section 4, we elaborate related work regarding Challenge 4.

3.1 Challenge 1: Linking to in-KG Entities via Known Surface Forms

There is an extensive amount of published work on entity linking (i.e., linking mentions to entries in a KG) and text annotation (entity linking including mention detection for unstructured text). The first approaches on entity linking to Wikipedia have been proposed by Bunescu et al. [2] and Cucerzan [5]. In 2008, Milne and Witten [13] built a system including a more sophisticated mention detection step. The annotation of the news texts used in our evaluations is provided by x-LiSA [22].

3.2 Challenge 2: Linking to in-KG Entities via Unknown Surface Forms

Dredze et al. [6] design a system for entity disambiguation taking into account the challenges of *name variations*, *entity ambiguity*, and *absence* of entities in the KG. The authors hence approach the Challenges 1, 2, and 3. They use different features for name variant detection and calculate a similarity score between the entity mention and the KG entity. SVM ranking is used to get the best candidate for each mention. In order to face Challenge 3, they introduce *NIL* as out-of-KG entity to which mentions can always be linked to.

Gottipati and Jiang [9] cover the Entity Linking Challenge 2 besides the traditional entity linking scenario. For that, their system considers not only the entity name for finding the in-KG entity, but also alternative name strings; these strings are gathered (i) from the document containing the mention using a NER tool and (ii) from Wikipedia exploiting page redirects.

3.3 Challenge 3: Linking to Out-of-KG Entities via Known Surface Forms

AIDA, a system for disambiguating named entities, was extended in 2014 [10] so that it can link to out-of-KG entities which share their entity names with in-KG entities. For each mention, besides the in KG-entity candidates, an additional out-of KG entity candidate is introduced which is initially represented by the mention string and later enriched by characteristic keyphrases. Wang et al. [18] also focus on the disambiguation of named entities. They detect so-called *target entities* in the text. These are entities (i) which are not necessarily contained in a KG, but whose names are known and where text documents containing

them are available, and (ii) which all come from a so called *target domain* such as "IT companies". They leverage these two aspects for a graph-based model that disambiguates all mentions across all documents collectively. Wu et al. [19] want to classify whether a given mention belongs to an existing KG entity or not, thereby targeting Challenge 3 and 4. The authors use five different spaces to model entities (a contextual, neural embedding, topical, query, and lexical space), but they do not consider the evolution of KGs.

3.4 Challenge 4: Linking to Out-of-KG Entities via Unkown Surface Forms

In this context, it is noteworthy to mention both schema-independent and schema-dependent novel entity detection approaches. All approaches only cover the prediction of whether given mentions (in unstructured text or already extracted) are KG entity candidates and which semantic types these entity candidates can be assigned to. However, they do not focus on *emerging* entities (as entities being of increasing public interest) and also do not correlate their predictions with the actual entity evolutions in a KG (such as the editing behavior in Wikipedia). Thus, to the best of our knowledge, this paper is the first one to define and propose an approach to solve the task of *emerging* entity detection.

Schema-free Novel Entity Extraction: Firstly it is noteworthy to mention that there are the Open Information Extraction approaches such as ReVerb [8] and NELL [3] which provide textual triples and, hence, entity mentions. Those mentions can be used to find out-of-KG entities (targeting Challenge 3 and 4) and to populate the KG [7]. Furthermore, NERC tools and general noun phrase extraction techniques can be used to gain novel entity candidates.

Within the Text Analysis Conference (TAC) tracks and the TREC tracks, the following tracks are related, but are too distant for a comparison with our approach and do not provide a suitable data set: 1. In the *TAC-KBP2015 Tri-lingual Entity Discovery and Linking (EDL) track* [11], besides the ordinary entity linking, non-linkable mentions should be clustered across languages. However, any non-linkable mention is considered as novel. 2. In the *TREC Novelty Detection tracks* [16], the topics (which are events and opinions) are very broad so that they cannot be used as entities. 3. In the *TREC KBA tracks*,[9] the systems had to fill slots on profiles. Like in case of the other mentioned tracks, the task is not to detect *emerging* out-of-KG entities.

The problem of novel entity detection also appears in the area of speech recognition, where it is often refered to as out-of-vocabulary (OOV) problem. Recent systems increase the set of known words by leveraging large external corpora such as the Web [15]. All OOV systems for speech recognition have in common that the OOV words are not assessed w.r.t. relevance, as any utterance needs to be matched, not only emerging entity mentions.

[9] See `http://trec-kba.org/`, requested June 26, 2016.

Schema-dependent Novel Entity Detection: Lin et al. [12] introduce the *unlinkable noun phrase problem*: Given a noun phrase that is not linked to Wikipedia as KG, determine whether it is an entity,[10] and if it is, determine its fine-grained entity types. In contrast to us, noun phrases are already given, so that no mention detection step is necessary. Furthermore, Lin et al. do neither consider *emerging* entities nor the evolution of a KG in general. Other works on predicting entity types for out-of-KG entities include HYENA proposed by Yosef et al. [20]. HYENA as first system assigns multiple types to an entity in a hierarchical order by applying a multi-label classifier. Nakashole et al. [14] propose PEARL, a system which assigns entity types to mentions of out-of-Freebase entities with the help of relational patterns.

4 Emerging Entity Detection

In this section, we present the first approach to the task of *emerging entity detection* on the basis of Wikipedia: We propose to train a machine learning model to detect out-of-Wikipedia entities which are emerging, i.e., which are for the first time reaching considerable public interest and are therefore conforming to Wikipedia's own notability requirements. Those entities can then be used for recommending the creation of new Wikipedia articles to Wikipedia editors or for enhancing entity linking in media monitoring systems.

4.1 Used Data Sets

For our analysis shown in Sec. 2.2, we annotated English news articles with the help of x-LiSA [22][11] given a "future" Wikipedia version (2015-05-15), so that we know which mentions refer to novel entities and which mentions just use new surface forms for existing entities. We now use the same annotated corpus for training and testing our machine learning model for *emerging entity detection*.

As emerging entity candidates we use all noun phrases (NPs) which were (i) extracted from the news articles by an implemented noun phrase extraction module (using a slightly extended rule set of [23] on the Part-of-Speech tags gained by the Stanford parser) and which were (ii) not linkable to in-KG entities by the entity linking system x-LiSA [22] given the Wikipedia state of 2015-04-04 (see Fig. 2). By means of the latter, we exclude all noun phrases for which KG entities already exist (i.e., filtering out annotations of Challenge 1 and 3). All noun phrases with annotations of Challenge 4 (i.e., mentions linking to emerging entities) are then labeled as "true", all noun phrases with annotations of Challenge 2 (i.e., new mentions linking to "old" in-KB entities) or without any Wikipedia annotation as "false". In total, after some initial filtering (e.g., considering only noun phrases with at least three alphanumeric characters) we

[10] An entity is here understood as "noun phrase that could have a Wikipedia-style article if there were no notability or newness considerations, and which would have semantic types." [12]

[11] Note that any text annotation method for Wikipedia could have been applied here.

Table 1: Number of *true* and *false* target labels for the data sets with different NP series lengths. Only the target labels of the respective last NP per NP series was considered.

NP series length	# *true*	# *false*
2	1754	100071
3	1246	48954
5	831	22066
10	474	8141
20	271	3076
30	169	1751
40	119	1142
50	86	806

Table 2: Examples of true positives and false positives for the data set with NP series length 20.

True positive	False positive
Michael Slager	Elton Simpson
Eric Courtney Harris	Garissa University
Dave Goldberg	Ananta Bijoy Das
Adult Beginners	Joan Kagezi
Dan Fredinburg	Gaioz Nigalidze
LG G4	Russell Begaye
Operation Maitri	Mitchell Santner
Struggle Street	Jose Urena
Oleg Kalashnikov	Severino Gonzalez
Masaan	Operation Fiela

came up with 15.6M extracted NPs (2.6M unique NPs), extracted from 1.8M English news articles. Note that this data set is highly unbalanced regarding the target labels (ratio true:false is 1:164). In order to reduce the unequal distribution between the classes and the overall data set size, we applied further filtering techniques. For instance, we considered only named entities (using the Stanford NER tagger; see our findings in Sec. 2.2). The reduced data set contained 840,101 NP occurrences and 404,263 unique NPs.

As the overall task is to predict emerging entities as soon as they reach public attention for the first time, we focused on the very first appearances of the NPs in the news stream. We therefore built series of the first n occurrences of each unique NP (with $n = 2, 3, 5, 10, 20, 30, 40, 50$; see Table 1). Based on those NP series, we calculated the features: For each NP occurrence, we extract a number of local noun phrase features (19; e.g., POS tag and suffix of the noun phrase), article features (17; e.g., source of the article), corpus features (7; e.g., slope reg. occurrences of the noun phrase over the last 24h), and global features (12; e.g., Wikipedia PageView slope over the last 24h). A detailed list of all extracted features is available at our website.[12] As most of the features (such as many slope values) are capturing the history of a whole NP group, we only used the the last NP occurrence in each NP series for training and testing.[13]

4.2 Feature Selection and Model Training

To alleviate the imbalance problem, we applied feature selection on the binary and numerical features. Our dimensionality reduction approach for binary features relies on a variance threshold per target class. For numerical features, the features whose values keep the same range and distribution in the positive class and in the negative class subsets were removed. The most important remaining features after the feature selection process include:

[12] See `http://people.aifb.kit.edu/mfa/emerging-entity-detection`.

[13] We also experimented with aggregating *all* features for each NP series, but did not yield better evaluation results.

Table 3: Evaluation results of Emerging Entity Detection.

NP series length	2	3	5	10	20	30	40	50
F1 score (in %)	4.6	6.7	10.7	12.2	23.1	24.7	25.0	19.0
accuracy (in %)	81.7	81.5	84.2	72.9	82.0	79.4	79.4	71.4

- *PosTag*: Part-of-Speech tag of the NP;
- *hostName*: host name of the article source (e.g., "www.n24.de");
- *npDiffArtsOccurrenceSlope24hSlope*: slope (using linear regression), based on the occurrence number of the NP over the last 24 hours;
- *pageViewSlope24h*: slope of the Wikipedia page view values[14] (requests for existing and non-existing Wikipedia pages) reg. the NP over the last 24h;
- *npAsTitleInDEWP*: true if the NP appears as title in the German Wikipedia.

We randomly split our data set and used 75% for training and 25% for testing. For training, the distribution among the two classes was equalized by removing instances. The training set was used to fit a Linear SVC model [4,1].[15]

4.3 Evaluation Results

The achieved F_1 scores and accuracy scores for the different data sets corresponding to the different NP series lengths are presented in Table 3. We can recognize an increase regarding the F_1 scores with increasing NP series length (with a maximum of 25.0 at NP series length 40), while the accuracy values roughly remain the same. Table 2 shows some examples of true positive and false positive predictions. Note that the displayed false positive examples were eventually all inserted into Wikipedia, just after the considered time range. This clearly shows that our approach is able to suggest emerging entities to Wikipedia editors before they noticed them.

To investigate this further, the top 100 false positive instances (with smallest distance to the hyperplane) were assessed by two independent assessors in order to find out to which extent the recommended NPs are valid Wikipedia emerging entities and thus could be added to Wikipedia. The assessors followed the *notability criteria* given by Wikipedia.[16] In case of varying judgements, the assessors agreed on a common judgment in a second round. Out of 100 assessed NPs which were classified as emerging entities (using the data with NP series length 20), (i) 37 had not been added to Wikipedia yet[17] but were judged manually as notable,[18] (ii) 20 had not beed added to Wikipedia before and judged manually as not notable, and (iii) 43 had already been added to

[14] See `http://dumps.wikimedia.org/other/pagecounts-raw/`.

[15] We also evaluated machine learning algorithms specialized on imbalanced and time-series data, such as cost-sensitive AdaBoost, cost-sensitive one class classifier and recurrent neural networks. However, this did not yield better results.

[16] See more information on our website.

[17] Given Wikipedia status of 2015-04-04 as the reference KG.

[18] Some of those entities were inserted later.

Wikipedia before, but were not detected as such.[19] If we disregard the mistakes introduced by the entity linking step, we can state a "false false positive" rate of about 65 %. In other words, 65 % of the out-of-KG instances predicted by our approach as emerging entities, actually were feasible emerging entities but not recognized by Wikipedia editors (yet).

5 Conclusions

In this paper, we presented a systematic overview of the different *entity linking* challenges arising from *out-of-KG entities* and *out-of-KG surface forms*. We provided an empirical analysis based on Wikipedia and on annotated English news articles regarding the importance of each of those challenges.

Based on that, we identified *emerging entity detection*, i.e. *trending* entities becoming *notable* for the first time, as the key task to facilitate semantic media monitoring. To address the task, we presented the first trained model for detecting emerging entities. The measured F_1 score lead to the conclusion that a robust prediction of emerging Wikipedia entities is tricky, due to the extreme imbalance in the data. However, this is to a large extend due to Wikipedia missing articles about valid emerging entities. Our approach is verifiably capable of identifying feasible candidate entities which could be added to Wikipedia according to Wikipedia's own notability guidelines. This would improve the up-to-dateness of Wikipedia and semantic media monitoring systems.

References

1. A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *The Journal of Machine Learning Research*, 2:125–137, 2002.
2. R. Bunescu and M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceesings of the 11th Conf. of the European Chapter of the Association for Comput. Linguistics (EACL-06)*, pages 9–16, Trento, Italy, 2006.
3. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka, and T. Mitchell. Toward an Architecture for Never-Ending Language Learning, 2010.
4. C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
5. S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on EMNLP-CoNLL*, pages 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
6. M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. Entity Disambiguation for Knowledge Base Population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 277–285, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[19] Investigations revealed that the already existing Wikipedia entities were not annotated by x-LiSA because no suitable surface form were available for those entities. In most of those cases, the entity was a person and in the news article only the family name was mentioned and extracted. However, in the set of known surface forms from Wikipedia only the full name of the entity was contained. Resolving those issues are left to future work.

7. A. Dutta, C. Meilicke, and H. Stuckenschmidt. Enriching structured knowledge with open information. In *Proceedings of the 24th Int. Conf. on World Wide Web*, WWW '15, pages 267–277, Republic and Canton of Geneva, Switzerland, 2015.
8. A. Fader, S. Soderland, and O. Etzioni. Identifying Relations for Open Information Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
9. S. Gottipati and J. Jiang. Linking Entities to a Knowledge Base with Query Expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 804–813, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
10. J. Hoffart, Y. Altun, and G. Weikum. Discovering Emerging Entities with Ambiguous Names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 385–396, New York, NY, USA, 2014. ACM.
11. H. Ji, J. Nothman, B. Hachey, and R. Florian. Overview of tac-kbp2015 tri-lingual entity discovery and linking. 2015.
12. T. Lin, Mausam, and O. Etzioni. No Noun Phrase Left Behind: Detecting and Typing Unlinkable Entities. In *Proc. of the 2012 Joint Conf. on EMNLP and CoNLL*, EMNLP-CoNLL '12, pages 893–903, Stroudsburg, PA, USA, 2012. ACL.
13. D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.
14. N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained Semantic Typing of Emerging Entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1488–1497, 2013.
15. C. Parada, A. Sethy, M. Dredze, and F. Jelinek. A spoken term detection framework for recovering out-of-vocabulary words using the web. *Paragraph*, 10(71.24):323K, 2010.
16. I. Soboroff and D. Harman. Novelty detection: the TREC experience. HLT '05, pages 105–112, Stroudsburg, PA, USA, 2005. ACL.
17. M. Trampuš and B. Novak. Internals of an aggregated web news feed. In *Proceedings of the Fifteenth International Information Science Conference IS SiKDD 2012*, pages 431–434, 2012.
18. C. Wang, K. Chakrabarti, T. Cheng, and S. Chaudhuri. Targeted Disambiguation of Ad-hoc, Homogeneous Sets of Named Entities. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 719–728, New York, NY, USA, 2012. ACM.
19. Z. Wu, Y. Song, and C. L. Giles. Exploring multiple feature spaces for novel entity discovery. In *AAAI 2016*. AAAI - Association for the Advancement of Artificial Intelligence, February 2016.
20. M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. HYENA: Hierarchical Type Classification for Entity Names. In *COLING 2012*, pages 1361–1370, 2012.
21. L. Zhang, M. Färber, and A. Rettinger. xLiD-Lexica: Cross-lingual Linked Data Lexica. In *Proceedings of the Ninth Int. Conf. on Language Resources and Evaluation (LREC'14)*, pages 2101–2105. ELRA, 2014.
22. L. Zhang and A. Rettinger. X-LiSA: Cross-lingual Semantic Annotation. *PVLDB*, 7(13):1693–1696, 2014.
23. S. Zhao, C. Li, S. Ma, T. Ma, and D. Ma. Combining POS Tagging, Lucene Search and Similarity Metrics for Entity Linking. In X. Lin et al., editors, *Web Inf. Systems Eng. – WISE 2013*, pages 503–509. Springer Berlin Heidelberg, 2013.