Early Discovery of Emerging Entities in Microblogs

Satoshi Akasaki[1] , **Naoki Yoshinaga**[2] and **Masashi Toyoda**[2]
[1]The University of Tokyo
[2]Institute of Industrial Science, the University of Tokyo
{akasaki, toyoda}@tkl.iis.u-tokyo.ac.jp, ynaga@iis.u-tokyo.ac.jp

Abstract

Keeping up to date on emerging entities that appear every day is indispensable for various applications, such as social-trend analysis and marketing research. Previous studies have attempted to detect unseen entities that are not registered in a particular knowledge base as emerging entities and consequently find non-emerging entities since the absence of entities in knowledge bases does not guarantee their emergence. We therefore introduce a novel task of discovering truly emerging entities when they have just been introduced to the public through microblogs and propose an effective method based on time-sensitive distant supervision, which exploits distinctive early-stage contexts of emerging entities. Experimental results with a large-scale Twitter archive show that the proposed method achieves 83.2% precision of the top 500 discovered emerging entities, which outperforms baselines based on unseen entity recognition with burst detection. Besides notable emerging entities, our method can discover massive long-tail and homographic emerging entities. An evaluation of relative recall shows that the method detects 80.4% emerging entities newly registered in Wikipedia; 92.8% of them are discovered earlier than their registration in Wikipedia, and the average lead-time is more than one year (578 days).

1 Introduction

Understanding the latest world events is an important objective for many applications such as social-trend analysis, marketing research, and reputation management. Such applications often require knowledge of *emerging entities*, such as new products, works, and individuals, which ceaselessly emerge one after another, for real-time monitoring of their activities. For example, social listening companies such as Salesforce and Oracle that monitor customer reputations need to track new product trends including those of customer's competitors. Although knowledge bases (KBs), such as Wikipedia, could be used as a reference list of entities, there is a certain delay until the emerging entities are registered in those KBs, and only notable entities are selected for registration. Therefore, instead of relying on KBs, we need to discover as many emerging entities as possible including *long-tail (less frequent but wide variety of) emerging entities* that are mostly overlooked in KBs before they become prevalent or their information appears frequently.

Nintendo *announces* **Super Smash Bros. Ultimate** - the game features every single character from past games. *Release data: December 7, 2018*
JohnnyDepp is set to play celebrated war photographer W. Eugene Smith in *upcoming drama* "**Minamata**" which HanWay Films *will launch at the upcoming* AFM. *Filming starts in Japan then Serbia in January 2019*
Can't wait for this one. **Big Dom's Bagel Shop** *will open* Aug. 25 in Cary. Here are all the details on Pizzeria Faulisi getting into the bagel business URL

Table 1: Example tweets on emerging entities (bold) with expressions suggesting their emergence (italic)

Previous studies [Nakashole *et al.*, 2013; Hoffart *et al.*, 2014; Wu *et al.*, 2016; Färber *et al.*, 2016] focus on detecting out-of-KB entities which are not registered in a particular KB, and consequently find massive non-emerging entities, since the absence of the entities in KBs does not guarantee their emergence (§ 3). Extracting emerging entities from the obtained out-of-KB entities is difficult since out-of-KB entities are mostly mere long-tail entities and we cannot expect many contexts to judge their emergence. The contexts can be even noisy when they are *homographic emerging entities* (*e.g.,* Go for new programming language and classical board game). Even worse, it is problematic to prepare training (and evaluation) datasets for out-of-KB detection since we need to manually annotate entities that are not registered (but notable) on the basis of the specific state of the given KB.

Considering these difficulties, we introduce a novel task of discovering emerging entities in a microblog when they have just introduced to the public through the microblog (§ 2). This task is more solid than the existing out-of-KB entity classification task since the task definition is independent of a particular KB. In our task, we use the fact that people write about emerging entities with *expressions suggesting their emergence* when those entities are not well known to the public (Table 1), considering that potential readers would be unfamiliar with them. By taking advantage of these contexts, we can effectively discriminate emerging entities from prevalent ones, even if they are long-tail or homographic emerging entities, and can find them in the early-stage of their appearance.

To obtain contexts of emerging entities, we propose a time-sensitive distant supervision method based on distant supervision [Mintz *et al.*, 2009]. Our method collects early-stage posts in a massive amount of time-series text where non-homographic entities registered in a KB first emerge. At this time, we also collect adequately-later posts after the first appearance as negative examples to robustly discriminate them from emerging contexts. We then train sequence-labeling models from those contexts to discover emerging entities.

We applied our method to our large-scale Twitter archive and compared the discovered emerging entities with those obtained with baselines, which regards entities that are unseen in a KB [Nakashole *et al.*, 2013] or our Twitter archive as emerging. Experimental results showed that the proposed method effectively detected emerging entities in terms of precision of the acquired entities including homographic and long-tail emerging entities. As the evaluation of relative recall and detection immediacy, using the entities newly registered in Wikipedia as a reference, our method detected most entities in the reference, and in most cases, these entities were discovered earlier than their registration in Wikipedia.

Our contributions are as follows:

- We introduce a novel task of discovering emerging entities in microblogs as early as possible.
- We propose a time-sensitive distant supervision method for easily and automatically constructing a large-scale training dataset from microblogs.
- Our method found emerging entities accurately (high precision), abundantly (high recall), and quickly (substantially earlier than their registration in Wikipedia).
- We will release all the datasets (tweet IDs)[1] used in experiments to promote the reproducibility.

2 Definition of Emerging Entity

In this section, we define what is meant by the term *emerging entity* in this study. Our definition of emerging entity is motivated from the report of [Graus *et al.*, 2018] and meets requirements for social-analysis applications.

Graus *et al.* analyzed how newly registered entities in Wikipedia have appeared in news and social media before they are registered as individual articles. They found that most of those entities shift from the state of “sporadically mentioned in news and social media” to that of “established as one article due to enhancement of references.”

Fortunately, when users submit posts about entities that appeared newly but are not famous yet to social media, they usually indicate the emergence of the entities, as in Table 1, despite their popularity. We thereby define emerging entities in terms of how they are described in contexts, in other words, how their state is perceived by people as follows:

Emerging contexts. *Contexts in which the writers assumed the readers do not know the existence of the entities.*

Emerging entities. *Entities in the state of being still observed in emerging contexts.*

We later confirm the solidness of these definitions by evaluating inter-rater agreement of emerging entities acquired from text (§ 5.4). We also define other terms on entities as follows:

Prevalent contexts. *Contexts in which the writers assumed the readers know the existence of the entities.*

Prevalent entities. *Entities in the state of being mainly observed in prevalent contexts.*

Long-tail entities. *Entities that are less frequent individually but have wide varieties.*

Homographic entities. *Entities that share the namings with other entities.*

3 Related Work

To the best of our knowledge, there has been no study attempting to find emerging entities in microblogs. We review the current tasks related to our task and clarify the term “emerging entities,” which has various meanings.

Emerging and Rare Entity Recognition

This is a task organized at the 2017 Workshop of Noisy User-generated Text (WNUT 2017) [Derczynski *et al.*, 2017] and focused on recognizing both “emerging and rare” entities from text. With this task, named entities (NEs) that appeared zero times in a specific (past) portions of datasets are regarded as emerging entities, and manually annotated NE tags to these entities as the target of detection regardless of the contexts in which they have appeared. The dataset used in this task includes the following example (the target entities are in bold):

> ... found photo storage tank that is 5x size of my **iPhone** with less capacity than **iPhone 4** ...

Consequently, this task is designed to detect (past) data-dependent emerging entities even after they become known to the public (*e.g.,* iPhone). The definition of emerging entities based on a specific data makes it difficult to distinguish emerging entities from prevalent entities. In fact, the state-of-the-art model achieved an F_1 of 49.59%, which is much lower than usual named-entity recognition (NER) on a dataset such as CoNLL-2003 (F_1 of 93.18%) [Akbik *et al.*, 2019].

Our task discovers emerging entities when they are introduced in microblogs. This enables us to take advantage of the fact that emerging entities tend to show their emergence at the early-stage of their appearance.

Out-of-KB Entity Identification on News Articles

This task has been studied to identify NEs that are not registered in a KB (referred to as “emerging” entities in the following studies but as out-of-KB entities here for clarity). Since this task is intended to detect entities absent in the KB, it does not distinguish emerging entities from mere long-tail entities.

[Nakashole *et al.*, 2013] proposed a method for extracting NEs using NER and regards all extracted NEs as out-of-KB if they are not registered in a KB. Since this method ignores

[1] http://www.tkl.iis.u-tokyo.ac.jp/~akasaki/ijcai-19/

contexts in which NEs appear, if the target NE has homographic entities in the KB, it is wrongly classified as an in-KB entity regardless of its emergence (false negatives). Similarly, if the target NE appears with unseen surface (mention), it is wrongly classified as an out-of-KB entity (false positives).

[Hoffart *et al.*, 2014], [Wu *et al.*, 2016], and [Färber *et al.*, 2016] proposed methods of classifying whether a given NE in a news article is out-of-KB. Their task is part of the task solved by [Nakashole *et al.*, 2013] since the target NEs are given (assumed to be recognized). Note that NEs are, however, not easily recognizable for languages in which NEs are not capitalized (*e.g.,* German, Chinese, and Japanese). In addition, their methods do not scale to ever-increasing emerging entities because the manual annotations of out-of-KB entities depend on the specific state of the KB and the approaches (and features of the classifier) are tailored for news text.

In contrast to these studies, we focus on "truly" emerging entities that are defined independent of KBs and develop an early-detection method of them using a dataset constructed by time-sensitive distant supervision. We targeted microblogs, *i.e.,* timely social media, as sources for emerging entities since [Graus *et al.*, 2018] reported that emerging entities appear on social media more and earlier than in news articles.

Notable Account Prediction on Twitter

This task is to discover long-tail "rising" entities (*e.g.,* rising brands) that are expected to be notable in the future within Twitter [Brambilla *et al.*, 2017]. Although this task uses Twitter as the source of entities, the same as ours, it requires experts to provide example notable entities. Also, since the target entities are limited to only those with Twitter accounts, it cannot acquire various types of entities that are not linked to Twitter accounts. We also focus on Twitter but discover emerging entities (§ 2) without relying on domain experts and without restricting the types of entities to be discovered.

Overall, these related studies define labels (emergence or rare, out-of-KB, or notability) based on specific past data, KBs, or domain experts and annotated them manually. We compare our method with two baselines that detect unseen NEs in a KB [Nakashole *et al.*, 2013] or in the past Twitter (the same setting as WNUT17) as emerging. We chose these methods because they are only methods applicable to our task, which do not rely on manually annotated data.

4 Proposed Method

The proposed method discovers emerging entities in microblogs. We target a microblog (Twitter) since [Graus *et al.*, 2018] reported that compared to news articles, a more diverse range of emerging entities appear earlier on social media, and generally speaking, microblogs include the most timely posts among various types of social media. Note that we do not exploit Twitter-specific functions with our method; thus, it is also applicable to other microblogs such as Weibo.

To build a supervised model for discovering emerging entities, we exploit the fact that emerging entities are likely to appear in specific contexts (§ 2). By properly identifying such contexts, we can discover corresponding emerging entities effectively and instantaneously even if they are long-tail or homographic ones. The major challenge lies in how to collect such emerging contexts as the training data. To cover various emerging contexts for a diverse range of entity types, we develop a method that automatically collects such contexts and corresponding emerging entities.

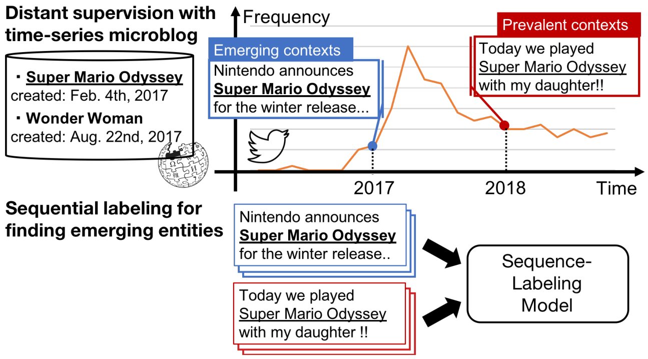
Distant supervision with time-series microblog
• Super Mario Odyssey created: Feb. 4th, 2017
• Wonder Woman created: Aug. 22nd, 2017
Frequency
Emerging contexts
Nintendo announces Super Mario Odyssey for the winter release...
Prevalent contexts
Today we played Super Mario Odyssey with my daughter!!
2017
2018
Time
Sequential labeling for finding emerging entities
Nintendo announces Super Mario Odyssey for the winter release..
Today we played Super Mario Odyssey with my daughter !!
Sequence-Labeling Model

Figure 1: Time-sensitive distant supervision: for the entities retrieved from a KB, emerging and prevalent contexts are collected from microblogs, and sequence labeling models are trained from the obtained emerging and prevalent contexts

4.1 Time-sensitive Distant-supervision

To meet the expected requirements on the training data for this task, we developed the proposed method (Figure 1) based on time-series text and the distant supervision [Mintz *et al.*, 2009], which automatically collects training data using an existing KB for a specific knowledge-acquisition task. Since our method does not incur any annotation cost, it is easy to prepare and construct the training data. The major difference from the original distant supervision is that labels are not defined only with the KB. We utilize the nature of time-series text to obtain labels for training a emerging entity recognizer.

The idea is to first extract non-homographic entities with unique namings from a KB that emerge when microblog posts are available and to collect their emerging contexts from the time-series microblog posts. The procedure is as follows:

Step 1 (Collecting candidates of emerging entities)

We start by collecting titles of articles in Wikipedia as existing entities and then associate them with the time-stamps of registration to collect emerging entities that newly appeared within the available period of the microblog (Twitter). We exclude entities that appeared on Twitter more than k times in the first one-year period where microblog posts are available. This is to exclude homographic entities that share the naming with prevalent entities since it is difficult to collect their emerging contexts only by searching the entities.

Step 2 (Collecting contexts of emerging entities)

For each entity obtained in Step 1, we then retrieve first n early-stage microblog posts posted before the time-stamps of registration as emerging contexts. Although contexts of long-tail emerging entities are not covered in the obtained training data, similar emerging contexts can be shared by other entities in the KB. This is because if the coarse type of entities are the same, their emerging contexts tend to be common regardless of their popularity (*e.g.,* product types tend to be introduced with the term *released*).

There are two issues to be addressed: 1) how to filter noisy examples of emerging contexts and 2) how to prevent overfitting that detects only the entities used for collecting training data. We explain how we address these issues.

Filtering Noisy Emerging Contexts

Although distant supervision can generate abundant training data, incorrectly labeled data can also be included. We therefore collect only reposts (retweets) from the day when the included entities first appeared in retweets more than k' times. This is inspired from the report of [Graus *et al.*, 2018] that emerging contexts are likely to be shared by many users since they include information novel to the public.

Collecting Prevalent Contexts as Negative Examples

When a model is trained only with the collected emerging contexts, it will be overfitted to detect only mentions of the emerging entities used to collect the training data. To avoid this, in Step 2, we collect prevalent contexts for the same entities collected in Step 1 as negative examples (Figure 1). Specifically, as the prevalent contexts for each entity, we collect the same number of microblog posts one year after the time of collecting emerging contexts. This enables the model to discriminate between emerging contexts and prevalent contexts and reduces the effect of noisy (prevalent) contexts incorrectly included in the positive examples.

We finally label only the acquired entities in the emerging contexts as emerging entities and combine them with their prevalent contexts to form the training data for sequence labeling described below. We tried several values for the three hyperparameters of our method, k, n, and k', and confirmed that the accuracy of the models trained from the resulting training data did not markedly change. We therefore empirically set the parameters to $k = 5$, $n = 100$ and $k' = 10$.

4.2 Sequence Labeling for Finding Emerging Entities

We next train a sequence-labeling model for finding emerging entities from the collected training data. We adopted and compared classical conditional random field (CRF) [Lafferty *et al.*, 2001] and modern long short-term memory (LSTM) with CRF output layer (LSTM-CRF) [Lample *et al.*, 2016] as the sequence-labeling models. We adopted BIOES as the tagging scheme, which was reported to be better than other schemes [Ratinov and Roth, 2009]. We tagged emerging entities in positive examples with BIES and the others with O.

Because our task includes detecting NEs, we referred to features for NER to solve our task. As the CRF features, we use part-of-speech tags, character types, and the results of NER[2] for the posts, and cluster IDs [Miller *et al.*, 2004] obtained from Brown clustering [Brown *et al.*, 1992] for each token in input and the two tokens before and after that. LSTM-CRF inputs a word embedding and character embeddings encoded by character LSTM of each token into bidirectional LSTM, which are followed by the CRF layer.

5 Experiments

We applied the proposed method to actual Twitter archive and performed our task of discovering emerging entities.

5.1 Data

We constructed the training dataset by using the time-sensitive distant supervision detailed in § 4.1 from our Twitter archive, which we have been compiling since March 11th, 2011 (more than 50 billion tweets have been accumulated).

In Step 1 of § 4.1, we collected titles of articles that were registered in the Japanese version of Wikipedia from March 11th, 2012 to December 31st, 2015 using the Wikipedia dump on June 20th, 2018. We then excluded redirects and disambiguation pages from the titles, and then ran Step 2. We obtained a total of 222,092 tweets including the same number of emerging and prevalent contexts for 19,604 entities as the training data. For model selection, we used 10% of the training data as the development data. We tokenized each example by using MeCab (ver. 0.996)[3] with ipadic dictionary (ver. 2.7.0) and then removed URLs, usernames, and hashtags.

We then analyzed the obtained emerging contexts by mapping the included emerging entities to their corresponding types assigned in the DBpedia ontology; for example, the entity "Spider-Man: Homecoming" is mapped to the type "Film." Out of the 19,604 emerging entities in our dataset, we have 12,259 type-mappings (51 types). As shown in Table 2, the entity types that are manually categorized into PERSON and CREATIVE WORK account for a large proportion. This is because these entities tend to generate a great deal of attention at the time of their appearance than other entities. The unmapped entities included artifacts (*e.g.*, devices, products), Web services, and other terminology because there are no mappings for them in the DBpedia ontology. We also see that emerging contexts could be diverse according to the type of entity they include. We thus have to capture those contexts properly to discover various types of emerging entities.

5.2 Models

The following models were implemented for comparison:

Proposed (CRF). We used the implementation using MALLET (ver. 2.0.6) [McCallum, 2002] with L-BFGS as an optimizer. The hyperparameter C was tuned to 0.125 using the development data. To obtain Brown clusters, we applied Brown clustering with 1024 clusters to 200 million Japanese tweets sampled from March 11th, 2011 to March 11th, 2012.

Proposed (LSTM-CRF). We used the implementation using Theano (ver. 0.9.0) provided by [Lample *et al.*, 2016].[4] We set hyperparameters as suggested in [Yang *et al.*, 2018], who explored the practical settings of neural-sequence-labeling. We optimized the model using stochastic gradient descent and chose the model at the epoch with the highest F_1 on the development data. To initialize the embedding layers, we trained 200-dimensional word embeddings using GloVe [Pennington *et al.*, 2014] from 800 million Japanese tweets posted from March 11th, 2011 to March 11th, 2012.

[2] We used CaboCha (https://taku910.github.io/cabocha/).

[3] https://taku910.github.io/mecab/

[4] https://github.com/glample/tagger/

TYPE	**# entities**	**# posts**	**examples of emerging context (translated and truncated)**
DBpedia types			
PERSON	**4932**	**23939**	
Actor	885	2863	... who is the partner of me has changed her name from Mika Tadokoro to **Momo Nonomiya**
MusicalArtist	731	4616	An idol unit who can do fishing, "**TSURIxBIT**" debuts on May 22th!
SoccerPlayer	531	2327	We are pleased to announce that **Kiyotaka Miyoshi** has joined Shimizu S-Pulse.
VoiceActor	484	1596	Expected new voice actor "**Sora Amamiya**" appeared for the first time on live broadcasting! ...
BaseballPlayer	419	4390	... announced that they have agreed to sign a player contract with **Spencer Patton**.
AdultActor	299	1731	... **Iori Furukawa**, who debuted last month is also cute! SOD's newcomer is always amazing!!
Model	281	1343	Nice to meet you, I am **Tamotsu Kansyuji** from the Juno-Super-Boy Contest. From now on ...
Politician	260	1164	In Kawasaki Mayor's election, Mr. **Norihiko Fukuda** defeated other candidates and ...
Person	177	729	Former CIA official **Edward Snowden** has revealed the US intelligence gathering
Writer	152	542	Ms. **Daruma Matsuura**, who won the newcomer award seems to start serializing from ...
Others (19 types)	713	2638	You'd better follow the youngster, **Atsugiri Jason**, who has been working for 2 months ...
CREATIVE WORK	**6460**	**47267**	
MusicSingle	1321	11685	The title of Nana Fujita's single has been decided as "**Right Foot Evidence**"!
TelevisionShow	1153	8478	[Kayoko Okubo] TBS's new program "**o-ku bon bon**" start from today! 24:50-25:20 on air
MusicAlbum	970	6092	Kis-My-Ft2 3rd album "**Kis-My-Journey**" (provisional) will be released on July 2 this summer!
Film	917	6307	... announced that the title of the rebooted version Spider-Man"**Spider-Man: Homecoming**".
VideoGame	652	6355	Latest videos and key art of "**DARK SOULS III**" released! [E3 2015]
Manga	623	2561	It was announced on Shonen Jump released today, new series "**My Hero Accademia**" starts ...
Anime	323	2983	Kyoto Animation's TV anime "**Tamako Market**" started broadcasting in January 2013!
RadioProgram	266	1010	... as we will record the first broadcast **Sasara Night of Fujioka Minami** on the STV.
Book	146	983	Congratulations on **Re: Zero-Starting Life in Another World** for upcoming publication!
Others (5 types)	89	813	KADOKAWA and Hatena's novel posting site **Kakuyomu** will open on February 29, 2016. ...
LOCATION	**371**	**1554**	
Building	121	756	A new gourmet building "**Ueno no Mori Cherry Terrace**" is born in Ueno. Both lunch and ...
Museum	42	184	**Kumagai Morikazu Tsukechi Museum of Art** opening ceremony commenced. Director ...
Station	34	115	... the name of the station to be built at the JR Nambu branch line is decided as **Odaei Station**!
Settlement	28	47	We started construction of a new town in **Slavticci**, 50 km west of the Chernobyl power plant.
School	24	34	I attended the **Kaishi International High School** Opening Ceremony.
City	17	46	Currently, I am in the core city of **La Hadadatu**, about 130 kilometers away from the war area ...
University	14	47	I was scheduled to attend the symposium on the establishment of **Akita University of Art** ...
Others (18 types)	91	325	Although it is late at night, we have released the **Ogijima Library** website!
GROUP	**366**	**2173**	
Company	259	1441	With the entry of robot business, Softbank established the new company named **Cocoro** ...
SoccerClub	55	304	Via Tin Kuwana changed team name to "**Vir Tin Mie**" Press release is available!
Organization	28	179	Mr. Ishiba decided on Friday to set up a political faction and name it "**Suigetsukai**."
PoliticalParty	24	249	... considers a new political party = Breaking up from nuclear power "**Japan's Future Party**"
OTHER	**130**	**561**	
Species	77	337	New species called **Sado flog**, found on Sado Island, features yellow feet and yellow belly.
CelestialBody	17	75	[With image] A new Earth-like planet "**Grisee 832c**" is discovered
Others (8 types)	36	149	Hitachi unveils new "**Class 800 Series**" for high-speed railways in the UK.
UNMAPPED	7345	35552	JMA named the recent heavy rain as "**Heavy Rainfall in / Kanto Tohoku H27.9**". DoCoMo's summer new model "AQUOS PHONE ZETA **SH-09D**" quick photo review ...
Total	19604	111046	

Table 2: Statistics of the emerging entities and their contexts obtained from our Twitter archive by our time-sensitive distant supervision

Baselines. Since our methods use automatically constructed training data, we prepared two baselines that do not utilize such data. Baseline1 regards NEs obtained by NER as emerging if they are not detected as NE on Twitter from one year to one week before the posting time of the input tweets. We set the period up to one week before to find NEs that emerge near the target day. Baseline2 regards NEs obtained by NER as emerging if they do not exist in a KB [Nakashole *et al.*, 2013]. We regard the obtained NEs as emerging when they are not registered in Wikipedia as of the month before the posting time of the input tweets because there is a time lag to use the latest Wikipedia dump in actual settings. To make NER robust, we use LSTM-CRF trained with a dataset combining KWDLC[5] and KNBC,[6] both of which are corpora in which NE tags are attached to noisy Web text.

[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KWDLC

[6] http://nlp.ist.i.kyoto-u.ac.jp/kuntt/#ga739fe2

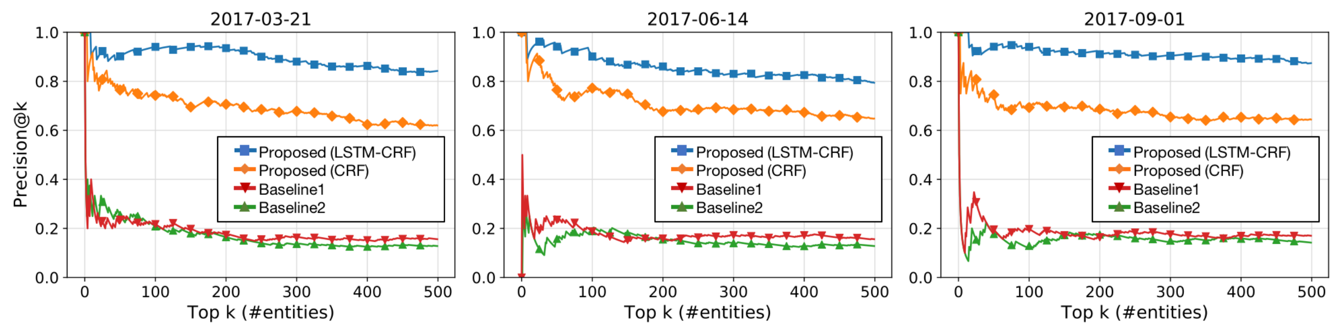

Figure 2: Precision@k for the top-500 emerging entities obtained from Twitter streams by each model.

5.3 Evaluation Procedures

To evaluate the proposed method, we designed two evaluation procedures for emerging entities discovered from Twitter.

Precision

To evaluate the precision of the obtained emerging entities, we applied each model to daily tweets, ranked the discovered entities using their confidence scores, and finally computed the accumulative precision for the top 500 entities. As the test sets, we randomly picked three sets of daily Japanese retweets, on March 21st, 2017 (1,695,423 tweets), June 14th, 2017 (2,041,833 tweets), and September 1st, 2017 (1,901,305 tweets) so that the seasons do not overlap. As the confidence score of Proposed (CRF) and Proposed (LSTM-CRF), we used the marginal probability obtained using the constrained forward-backward algorithm [Culotta and McCallum, 2004]. We adopted the maximum scores for the extractions when several mentions of the same entity were recognized. Since baselines do not provide any scores regarding the emergence of entities, we used the number of extractions of each entity normalized with the extraction number of the previous day as the confidence score. This captures the bursty feature that takes into account the appearance ratio of the previous day.

We asked three annotators, including the first author and two student volunteers, to decide whether the outputs were accompanied by emerging contexts defined in (§ 2) by referring to the input tweets and then adopt the majority labels to mediate the conflicts. We obtained an inter-rater agreement of 0.798 by Fleiss's Kappa [Fleiss and Cohen, 1973], which indicates substantial agreement. This high agreement justifies the solidness of the task setting.

Relative Recall and Detection Immediacy

To evaluate the recall and detection immediacy of the obtained emerging entities, we ideally want to refer to the complete list of entities that have emerged in certain periods. However, it is unrealistic to have such a list for a diverse range of entities including long-tail emerging entities. We instead evaluated the relative recall and immediacy against a KB, by determining how many entities registered in Wikipedia could be found from the tweets and how early they were detected against their registration date in Wikipedia.

Since entities newly registered in Wikipedia include emerging and prevalent entities, we obtained the reference list of emerging entities as follows. We collected entities that appeared more than 100 times on our Twitter archive from January 1st, 2017 to June 20th, 2018, and then extracted retweets containing each entity since the first appearance. To exclude prevalent entities as much as possible, we ignored entities that appeared more than five times on our Twitter archive from March 11th, 2011 to March 11th, 2012. We obtained 13,386 entities with 9,080,178 tweets (678 tweets per entity on average) since March 12th, 2012, and then applied our method to these tweets and calculated the recall and detection immediacy of the obtained entities.

daily tweets	HEAD $(n > 100)$	LONG-TAIL $(n \leq 100)$	HOMOGRAPH	total
Mar. 21st, 2017	227	110	84	422
Jun. 14th, 2017	214	106	77	397
Sep. 1st, 2017	261	110	66	437

Table 3: Details of the emerging entities discovered from the daily tweets with Proposed (LSTM-CRF)

5.4 Results and Analysis

Figure 2 depicts the cumulative precision (precision@k) for the top 500 entities discovered with each model. Proposed (LSTM-CRF) is superior to the others and mostly maintained a precision above 80% (on average 83.2% for top-500 entities for the three sets of daily retweets), while two Baselines remained mostly under 20%. Proposed (LSTM-CRF) is superior to Proposed (CRF) because it models the longer contexts (entire posts) with LSTM, and can properly capture the emerging contexts detectable by seeing the entire posts.

Table 3 lists the detected emerging entities falling under three categories to confirm whether our method could discover various types of emerging entities defined in § 2. HEAD represents entities whose surfaces appeared over 100 times in our Twitter archive from the detection date to one year later, and LONG-TAIL is less than that. HOMOGRAPH represents homographic entities whose namings are already registered in Wikipedia before the detection date. As a result, Proposed (LSTM-CRF) could discover not only entities that would be added to Wikipedia but also find many long-tail emerging entities (*e.g.*, good and evil (play), Photo X Art Field (exhibition)). Their frequency is low but they are use-

TYPE	**# entities**	**# found (%)**	**lead-days**
DBpedia types			**mean (median)**
PERSON	**3851**	**3211 (83.38%)**	**665 (556)**
Actor	651	506 (77.73%)	753 (750)
MusicalArtist	624	457 (73.24%)	749 (659)
SoccerPlayer	477	427 (89.52%)	775 (762)
VoiceActor	345	317 (91.88%)	508 (376)
AdultActor	306	300 (98.04%)	376 (297)
BaseballPlayer	238	213 (89.50%)	591 (405)
Model	225	210 (93.33%)	764 (654)
Politician	136	112 (82.35%)	665 (538)
Wrestler	116	97 (83.62%)	360 (165)
Others (16 types)	560	436 (77.85%)	705 (654)
CREATIVE WORK	**4122**	**3683 (89.35%)**	**379 (179)**
TelevisionShow	699	640 (91.56%)	228 (55)
MusicSingle	653	593 (90.81%)	305 (85)
Film	641	552 (86.12%)	391 (214)
MusicAlbum	550	498 (90.55%)	356 (161)
Manga	523	481 (91.97%)	678 (600)
VideoGame	440	391 (88.86%)	407 (249)
RadioProgram	228	203 (89.04%)	261 (38)
Anime	216	184 (85.19%)	261 (91)
Book	101	94 (93.07%)	627 (463)
Others (4 types)	71	47 (66.19%)	700 (515)
LOCATION	**223**	**179 (80.27%)**	**600 (394)**
Building	89	73 (82.02%)	505 (291)
Museum	33	32 (96.97%)	685 (447)
Station	25	21 (84.00%)	264 (154)
School	18	9 (50.00%)	554 (74)
Library	13	13 (100.00%)	995 (1328)
Park	11	9 (81.82%)	639 (193)
University	7	6 (85.71%)	904 (996)
Others (10 types)	27	16 (59.25%)	882 (956)
GROUP	**240**	**148 (61.67%)**	**559 (412)**
Company	188	113 (60.11%)	509 (361)
SoccerClub	26	13 (50.00%)	780 (741)
Organisation	16	14 (87.50%)	725 (511)
PoliticalParty	10	8 (80.00%)	622 (527)
OTHER	**59**	**18 (30.51%)**	**758 (977)**
Species	53	14 (26.42%)	825 (1008)
CelestialBody	3	1 (33.33%)	2 (2)
Train	2	2 (100.00%)	241 (241)
Aircraft	1	1 (100.00%)	1613 (1613)
UNMAPPED	4891	3523 (72.03%)	697 (622)
Total	13386	10762 (80.40%)	578 (417)

Table 4: Relative recall and time advantage over entity types of emerging entities detected with Proposed (LSTM-CRF)

ful for companies performing social listening and local users trying to find something interest. It also found homographic entities (*e.g.*, NEVER LAND (music album), Summer of Love (musical movie)), which were not found with **Baselines**. Although it has been reported that these homographic emerging entities are difficult to find [Hoffart *et al.*, 2014; Färber *et al.*, 2016], our method successfully discovered these entities by obtaining the emerging contexts of the entities.

As the evaluation of relative recall, we focused on the best-performing method, *i.e.,* Proposed (LSTM-CRF) and computed its relative recall over the reference list of 13,386 emerging entities. We detected 10,762 emerging entities (80.4%). This is reasonably high considering that there was noise in the reference list such as a concept name that was defined after it became prevalent (*e.g.*, Virtual Youtuber) and periodic entities (*e.g.*, Tokyo prefectural election, 2017).

Table 4 shows the distribution of the types of the 13,386 entities obtained by the DBpedia mappings, detection ratio, and lead-time against the Wikipedia registration time for each type. For PERSON, CREATIVE WORK, and LOCATION types, our model found on average more than 80% of entities. On the other hand, for GROUP and OTHER types, detection rates dropped remarkably. We found that some of those entities do not appear in emerging contexts at all within our Twitter archive. Since our method utilizes such emergence signals as the clue, it is difficult to discover entities appearing without emerging contexts. This is the current limitation of our method. Note that the 13,386 entities used in this evaluation included some prevalent entities (*e.g.,* local company) that might also affect the performances.

We next evaluated detection immediacy. We found that 92.8% of the discovered entities (9,979 out of 10,762) were detected earlier than their registration in Wikipedia. We then investigated the remaining 783 (7.2%) entities and found that they were mostly periodic events such as Olympics and election, or incorrectly included prevalent entities. The mean (and median) lead days of the first day when Proposed (LSTM-CRF) detected each entity against their registration date were 578 (and 417) days, which supports the detection immediacy of our method. Compared to CREATIVE WORK types of entities, our method detected PERSON and LOCATION types of entities earlier than their registration in Wikipedia, which means those entity types take longer to be notable enough to be registered in Wikipedia [Graus *et al.*, 2018].

Overall, these results reconfirm that microblogs are useful sources for finding emerging entities and our method can detect such entities at the early-stage of their appearance. It also implies that relying on Wikipedia for source of entities misses valuable information on emerging entities.

6 Conclusions

We introduced a novel task of discovering emerging entities in microblogs (§ 1, 2). We pointed out the problems of related tasks (§ 3) and proposed an effective method for discovering emerging entities in microblogs by exploiting the contexts of those entities using time-sensitive distant supervision (§ 4). Experimental results demonstrated that our method performed accurately and showed that emerging entities, including homographic and long-tail ones, can be effectively and instantly discovered by obtaining emerging contexts (§ 5).

We plan to carry out semantic typing of emerging entities by exploiting emerging contexts, which are likely to include enough information for the public to understand the entities.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 18J22830. The authors thank the volunteers and the anonymous reviewers for their hard work.

References

[Akbik *et al.*, 2019] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 724–728, 2019.

[Brambilla *et al.*, 2017] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. Extracting emerging knowledge from social media. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pages 795–804, 2017.

[Brown *et al.*, 1992] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[Culotta and McCallum, 2004] Aron Culotta and Andrew McCallum. Confidence estimation for information extraction. In *Proceedings of the 5th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 109–112, 2004.

[Derczynski *et al.*, 2017] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (WNUT)*, pages 140–147, 2017.

[Färber *et al.*, 2016] Michael Färber, Achim Rettinger, and Boulos Asmar. On emerging entity detection. In *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 223–238, 2016.

[Fleiss and Cohen, 1973] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[Graus *et al.*, 2018] David Graus, Daan Odijk, and Maarten de Rijke. The birth of collective memories: Analyzing emerging entities in text streams. *Journal of the Association for Information Science and Technology*, 69(6):773–786, 2018.

[Hoffart *et al.*, 2014] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 385–396, 2014.

[Lafferty *et al.*, 2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 260–270, 2016.

[McCallum, 2002] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[Miller *et al.*, 2004] Scott Miller, Jethran Guinness, and Alex Zamanian. Name tagging with word clusters and discriminative training. In *Proceedings of the 5th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 337–342, 2004.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1003–1011, 2009.

[Nakashole *et al.*, 2013] Ndapandula Nakashole, Tomasz Tylenda, and Gerhard Weikum. Fine-grained semantic typing of emerging entities. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1488–1497, 2013.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Ratinov and Roth, 2009] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, pages 147–155, 2009.

[Wu *et al.*, 2016] Zhaohui Wu, Yang Song, and C Lee Giles. Exploring multiple feature spaces for novel entity discovery. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3073–3079, 2016.

[Yang *et al.*, 2018] Jie Yang, Shuailong Liang, and Yue Zhang. Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3879–3889, 2018.