Early Discovery of Emerging Entities in Microblogs

Satoshi Akasaki ♣
akasaki@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga ♣
ynaga@iis.u-tokyo.ac.jp

Masashi Toyoda ♣
toyoda@tkl.iis.u-tokyo.ac.jp

♣ The University of Tokyo

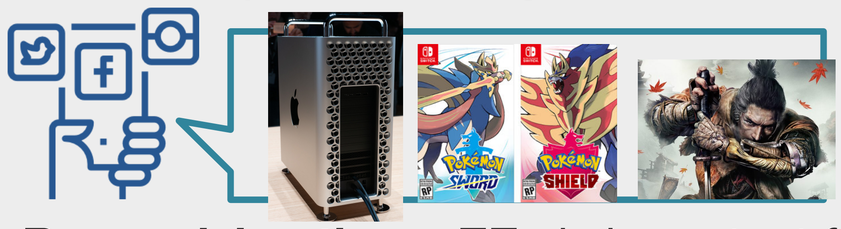
Dataset: http://www.tkl.iis.u-tokyo.ac.jp/~akasaki/ijcai19/

Contribution

- Introduce a novel task of discovering **emerging entities (EEs)** in microblogs
- Revisit the definition of EEs so that it does not depend on temporal resource (KB etc.)
- Propose a method that can discover EEs accurately, abundantly and quickly

Introduction

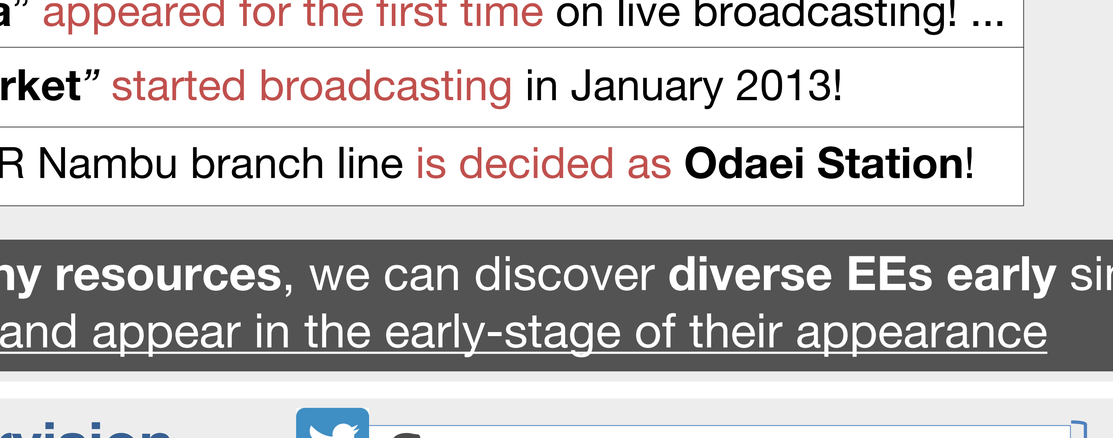
- **Emerging entities (EEs)** are appearing ceaselessly in microblogs real-time



- **Recognizing those EEs** is important for applications such as social listening

Task definition

Find EEs as **many and early** as possible from microblog posts



Challenges:

- ✓ How to solidly define EEs without relying on temporal resources?
- ✓ How to **discover diverse EEs** including long-tail ones?
- ✓ How to **discover EEs early** while their frequencies are low?

Resource-independent definition of EEs

Define **EEs** only in terms of how people describe their contexts (**emerging contexts (EC)**):

- ✓ **Emerging Contexts**: *Contexts in which the writer assumed the reader does not know the existence of the entities*
- ✓ **Emerging Entities**: *Entities in the state of being still observed in EC*

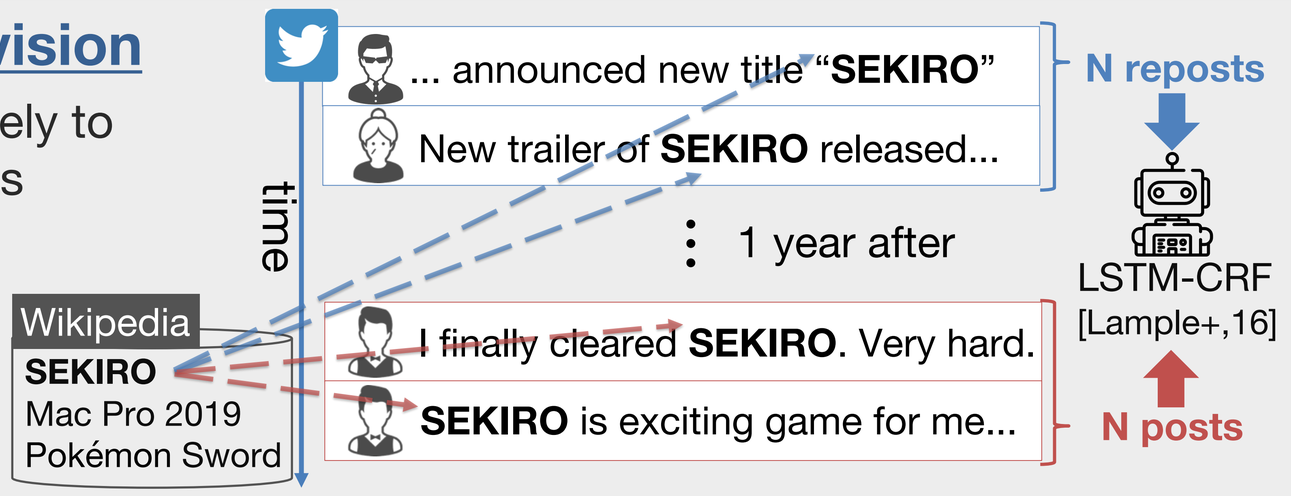
Expected new voice actor "**Sora Amamiya**" appeared for the first time on live broadcasting! ...
Kyoto Animation's TV anime "**Tamako Market**" started broadcasting in January 2013!
The name of the station to be built at the JR Nambu branch line is decided as **Odaei Station**!

By capturing those **ECs** that are **independent of any resources**, we can discover **diverse EEs early** since ECs are common whether they are long-tail or not, and appear in the early-stage of their appearance

(LSTM-)CRF based on timely distant supervision

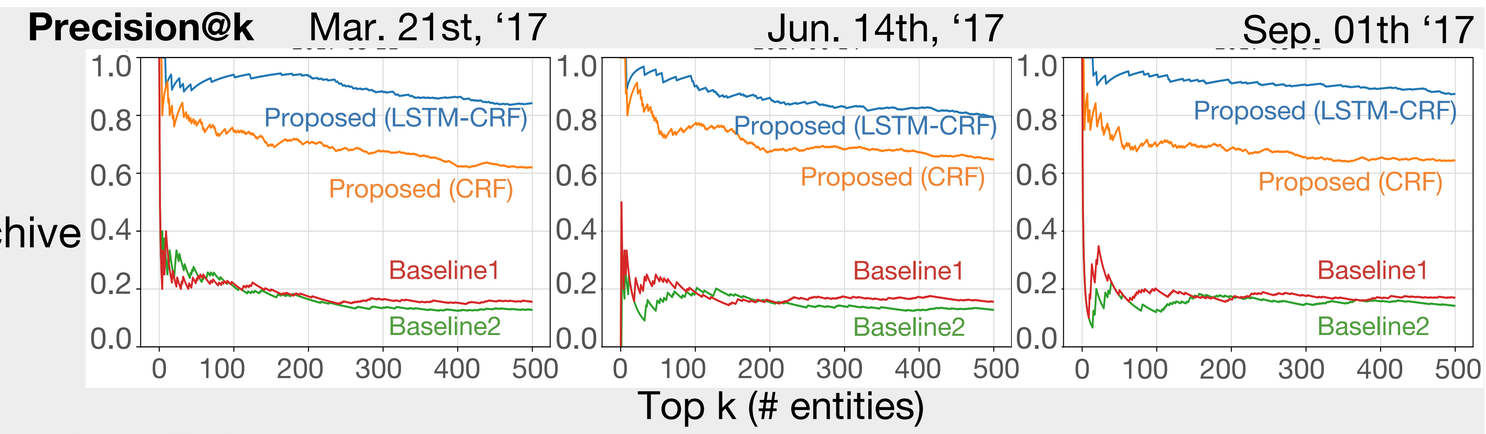
Collect **early-stage microblog posts** where EEs are likely to appear by using distant supervision on time-series posts

1. Retrieve **first N reposts** **where the titles of Wikipedia articles appear** as ECs with EEs (Pos. ex.)
2. Retrieve **last N posts** **one year after the time of collecting ECs** with collected EEs (Neg. ex.)



Experiments

- **Training data:**
  - ✓ 222,092 posts collected using proposed method from the period of 2012 to 2015 in our Twitter archive
- **Methods:**
  - ✓ **Proposed:** (LSTM-)CRF using the training data
  - ✓ **Baseline:** Output NEs found by NER that are absent in a dictionary or Wikipedia (Baseline1), or in the past tweets (Baseline2)
- **Evaluation:**
  - ✓ **Precision:** Evaluate top-500 discovered entities from daily retweets, which are sorted using confidence scores of each method
  - ✓ **(Relative) Recall:** Evaluate how many target entities registered in Wikipedia from Jan. '17 to Jun. '18 could be discovered from all the retweets that include those entity surfaces



Examples of discovered EEs and ECs

Discovered EE	Example contexts of entities (input tweet)
DRAGON BALL FighterZ	New fighting game **DRAGON BALL FighterZ** is announced! Trailer is become public. Expected to appear early 2018.
Kayakku LIVING	Kayakku established WEB company by business acquisition, which is related to housing named "**Kayakku LIVING**"

Discovered EEs from daily tweets

Day	# Head (freq. > 100)	# Long-tail (freq. < 100)	# Ambiguous (exist in Wikipedia)
Mar. 21st	227	110	84
Jun. 14th	214	106	77
Sep. 01st	261	110	66

Recall and lead-days over entity types of EEs

Type	# Entities	# Found (%)	Lead-days (mean/med.)
PERSON	3851	3238 (84.08%)	660 / 550
ARTWORK	4122	3703 (89.84%)	377 / 176
LOCATION	223	179 (80.27%)	597 / 385
GROUP	240	152 (63.33%)	545 / 396
OTHER	59	18 (30.51%)	758 / 977
UNMAPPED	4891	3523 (72.03%)	690 / 615
TOTAL	**13406**	**10852 (80.95%)**	**571 / 406**