Early Discovery of Emerging Entities in Microblogs

Satoshi Akasaki,

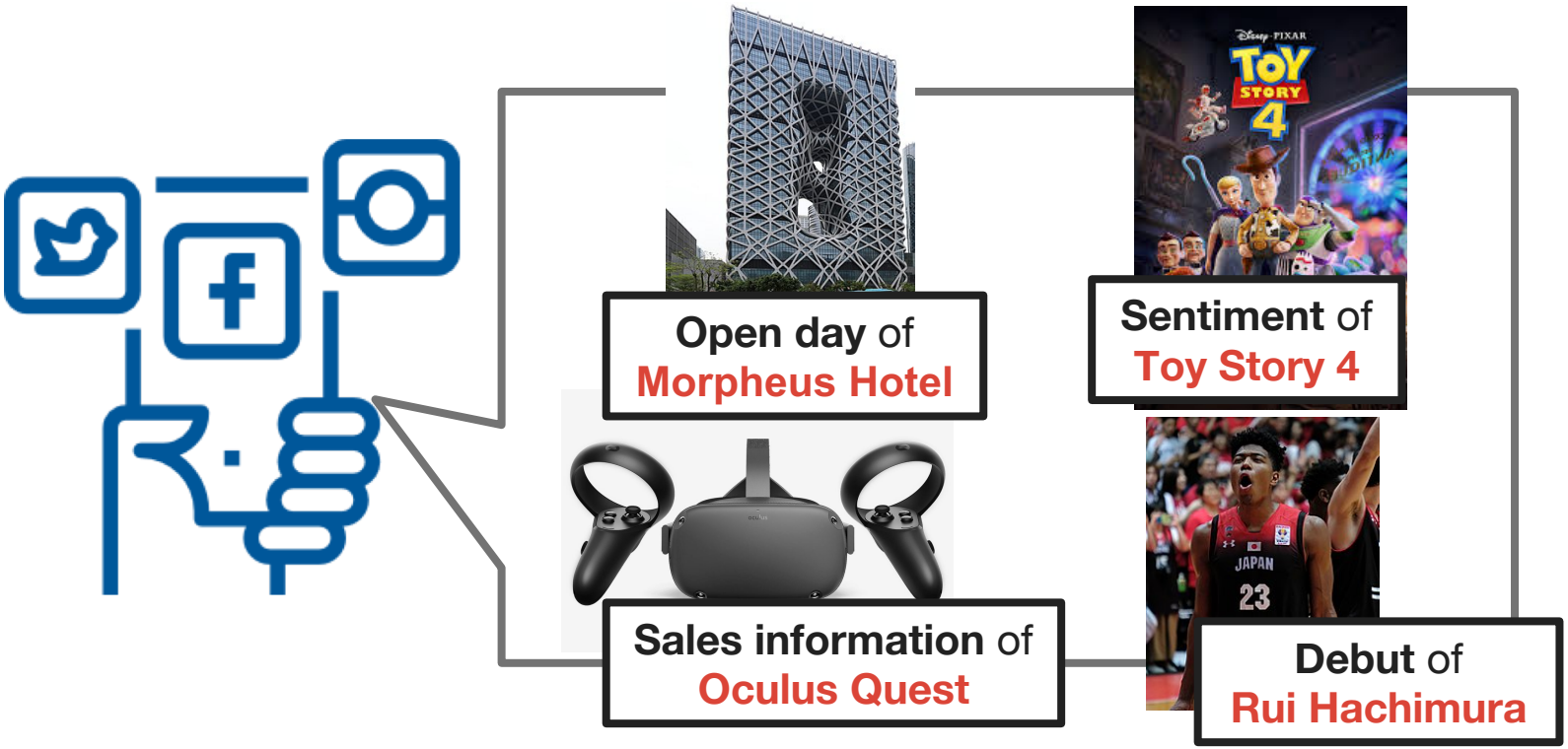
Naoki Yoshinaga,

Masashi Toyoda

THE UNIVERSITY OF TOKYO

Microblogs as sources of new information

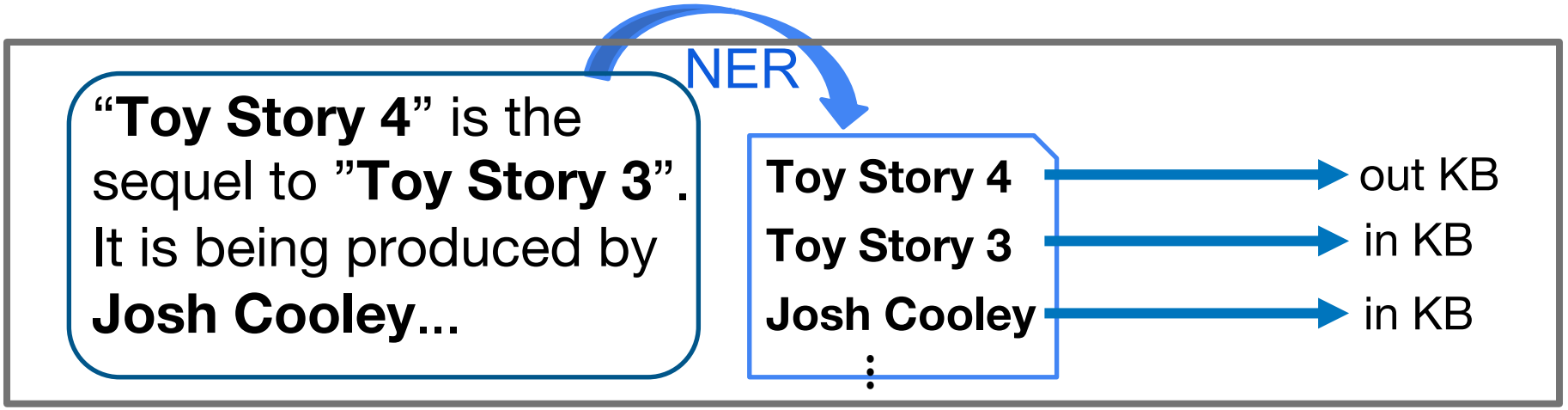
We can find impressions, opinions, and thoughts on **new artworks, products, persons** etc (**social listening**)



Social listening need to comprehensively collect such "**emerging entities (EEs)**"

Related work: Find out-of-KB entities (1)

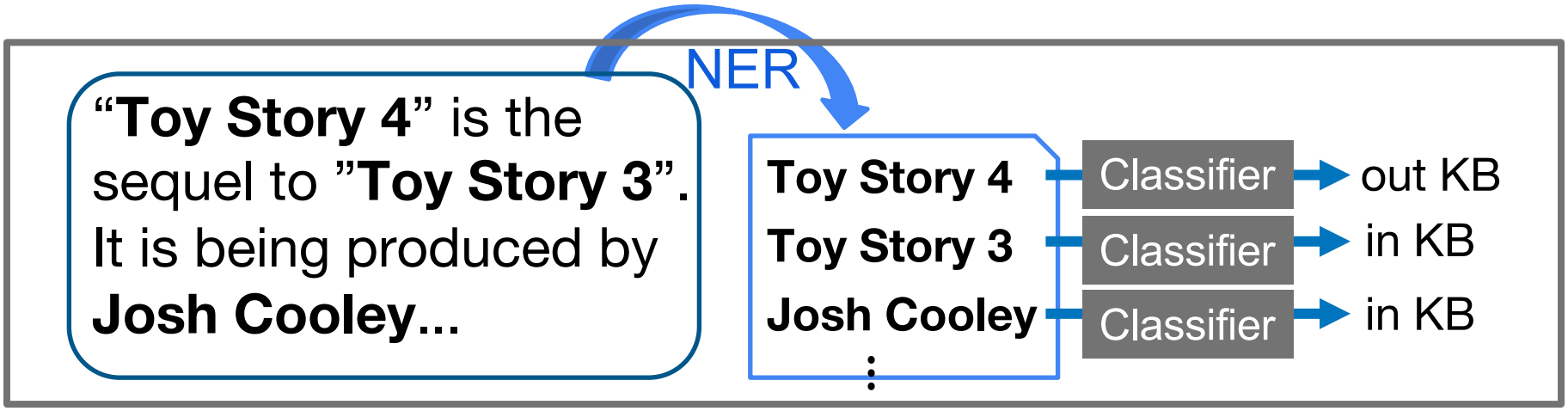
- Identify out-of-KB entities that are not registered in knowledge bases (KBs) [Nakashole+,'13]



- Regard extracted entities as out-of-KB If there are no entries in KBs with the same name
  - ✓ Overlook **homographic EEs** that need disambiguation
    e.g. **Go** (Programming language vs Table game)

Related work: Find out-of-KB entities (2)

- Identify out-of-KB entities including homographic ones using binary classification [Hoffart+,'14],[Farber+,'16],[Wu+,'16]



- Definition of entities depend on KBs of a specific time

✓ It is unrealistic to recreate the dataset when the KBs are updated

Related work: Find out-of-KB entities (2)

- Identify out-of-KB entities including homographic ones using binary classification [Farber+,'16], [Wu+,'16]

Out-of-KB entities **do not guarantee their emergence** because there are many **mere long-tail entities**

We focus on **emerging entities** that are defined without depending on KBs

Task setting of this research

Discovering diverse **emerging entities (EEs)** as early as possible from microblogs

Challenges:

- How to define EEs without depending on resources such as KBs?
- How to collect diverse EEs including long-tail and homographic ones?
- How to discover EEs early while their frequencies are low?

Definition of emerging entity (EE)

[Graus+,'18] analyzed how emerging entities in Wikipedia behave, and found that there are two states:

1. **Initially mentioned in medias** such as news and microblogs
2. **Established as articles** by the enrichment of references

1.

2.

Since we want to find EEs early without depending on KBs, we define emerging entities based on 1.

Definition of emerging entity (EE) (cont.)

We define EEs only in terms of how people describe their contexts, without relying on **any tentative resources**

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

Definition of emerging entity (EE) (cont.)

We define EEs only in terms of how people describe their contexts, without relying on **any tentative resources**

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

We can discover diverse **EEs** early since:

- ✓ ECs are observed even if EEs are long-tail or homographic
- ✓ ECs appear in the early-stage of their appearance

Approaches for discovering EEs

Automatically construct training data that includes ECs and non-ECs of the same entity for discrimination

1. Collect **ECs** and non-ECs with **EEs** using distant supervision on time-series posts **(Timely Distant Supervision (TDS))**
2. Train an NER model that detects **EEs** using the collected data

Proposed timely distant supervision (collecting positive examples)

Collect ECs assuming that **initial posts of entities usually include ECs**

1. Extract titles of Wikipedia articles as entities
2. Collect **first N reposts** containing the extracted entities

Proposed timely distant supervision (collecting negative examples)

Collect non-ECs assuming that **prevalent posts of entities usually include non-ECs**

1. Collect **last N posts** one year after the time of collecting positive examples for each entity as negative examples

Settings: Training data

Timely distant supervision successfully collected **222,092 Japanese tweets** that include **19,604 entities**

registered in Wikipedia from Mar. 11th, 2012 to Dec. 31st, 2015

Statistics of positive examples (emerging contexts)

Type*	# ent.	# ex.	Examples of entities
PERSON	4,932	23,939	Sora Amamiya (Voice Actor), Naomi Osaka (Athlete)
CREATIVE WORK	6,460	47,267	Kakuyomu (Web site), Shin Godzilla (Movie)
LOCATION	371	1,554	Odaei Station (Station), Ogijima Library (Building)
GROUP	366	2,173	Cocoro SB (Company), Suigetsu kai (Political Faction)
OTHER	130	561	Sado Flog (Species), 2014AA (Celestial Body)
UNMAPPED	7345	35,552	Miracast (Technology), Apple A9 (SoC)
TOTAL	**19,604**	**111,046**	

* We use types of DBpedia and aggregate them into six types

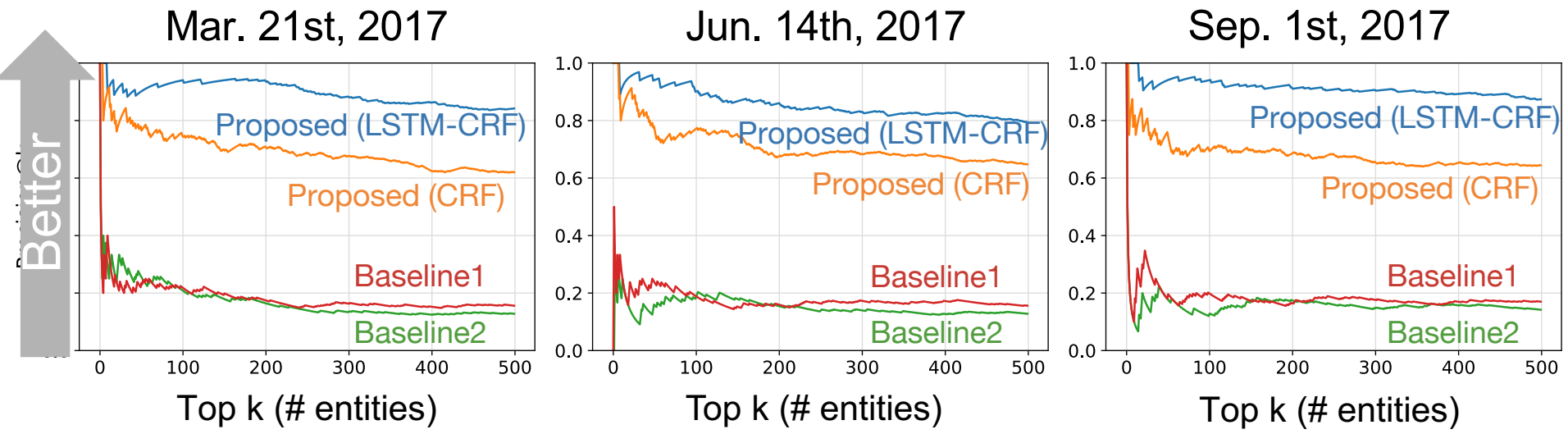
Settings: Compared methods

To validate the usefulness of the auto-constructed training data, we prepare baselines without using the training data

- **Proposed method:**
  - ✓ Train **LSTM-CRF** [Lample+,'16] and **CRF** using training data
- **Baselines:**
  - ✓ Train **LSTM-CRF based generic NER model** using noisy web text and **output recognized NEs that <u>do not</u>**
    - ➢ exist in a dictionary or in Wikipedia (**Baseline1**)
    - ➢ appear as NEs in the past Twitter (**Baseline2**)

Results: Precision

Evaluate top-500 discovered entities from daily retweets that are sorted using confidence scores of each method



Proposed method **outperformed performances of the baselines** (>80%)

Results: Precision (classification of EEs)

Classify **discovered EEs** into three types:

- **Head:** appeared in Twitter more than 100 times
- **Long-tail:** appeared in Twitter less than 100 times
- **Homograph:** whose surfaces are already registered in Wikipedia

Daily tweets	Head (n > 100)	Long-tail (n <= 100)	Homograph	Total
Mar. 21st, 2017	227	110	84	422
Jun. 14th, 2017	214	106	77	397
Sep. 1st, 2017	261	110	66	437

Our method **discovered long-tail** and **homographic EEs** by capturing ECs

Results: Precision (examples of output)*

Entity	Example contexts (input tweet)
DRAGON BALL FighterZ	New fighting game **DRAGON BALL FighterZ** is announced! Trailer is become public. Expected to appear early 2018.
Godzilla: Planet of the Monsters	Theater release date announced! The movie "**Godzilla: Planet of the Monsters**" will be released nationwide on Nov. 17th (Friday)..
Tokyo Ghoul	[Theater] 7/29 National movie "**Tokyo Ghoul**" preview video lifted !! Nobuyuki Suzuki (Gekidan EXILE) appeared in the video...
LOVE and LIES	The new visual of the single "**LOVE and LIES**" released on April 19th has been unveiled! Please have a look!
Kayakku LIVING	Kayakku established WEB company by business acquisition, which is related to housing named "**Kayakku LIVING**"
Rio Nakano	The new member who was announced today is **Rio Nakano** ! 💗💗💗 It will be appeared from 4/2 (Sat) 😊💕
Next Schubert	Started a new group activity as a classical rock girls unit. We will have a live on Feb. 26th under the name "**Next Schubert**" !!

* **Head EE** **Homographic EE** **long-tail EE** EC

Results: (Relative) Recall

Evaluate how many target entities registered in Wikipedia from Jan. 2017 to Jun. 2018 could be discovered from all the retweets

Type	# entities	# found (%)	Lead-days against Wikipedia (mean/median)
PERSON	3851	3211 (83.38%)	665 / 556
CREATIVEWORK	4122	3683 (89.35%)	379 / 179
LOCATION	223	179 (80.27%)	660 / 394
GROUP	240	148 (61.67%)	559 / 412
OTHER	59	18 (30.51%)	758 / 977
UNMAPPED	4891	3523 (72.03%)	697 / 622
TOTAL	**13386**	**10762 (80.40%)**	**579 / 417**

- Discovered **80% of the entities** on average
- **Found every types of entities more than one year earlier** than the date of registration in Wikipedia

Conclusion

Discover emerging entities (EEs) from microblogs

- Define EEs and proposed the task of discovering EEs
- Proposed timely distant supervision that collect ECs
- Proposed method achieved **83.2% precision on ave.** and **80.4% recall on ave.** while discovered **diverse EEs early** (578 days earlier than the registration in Wikipedia)

Future works

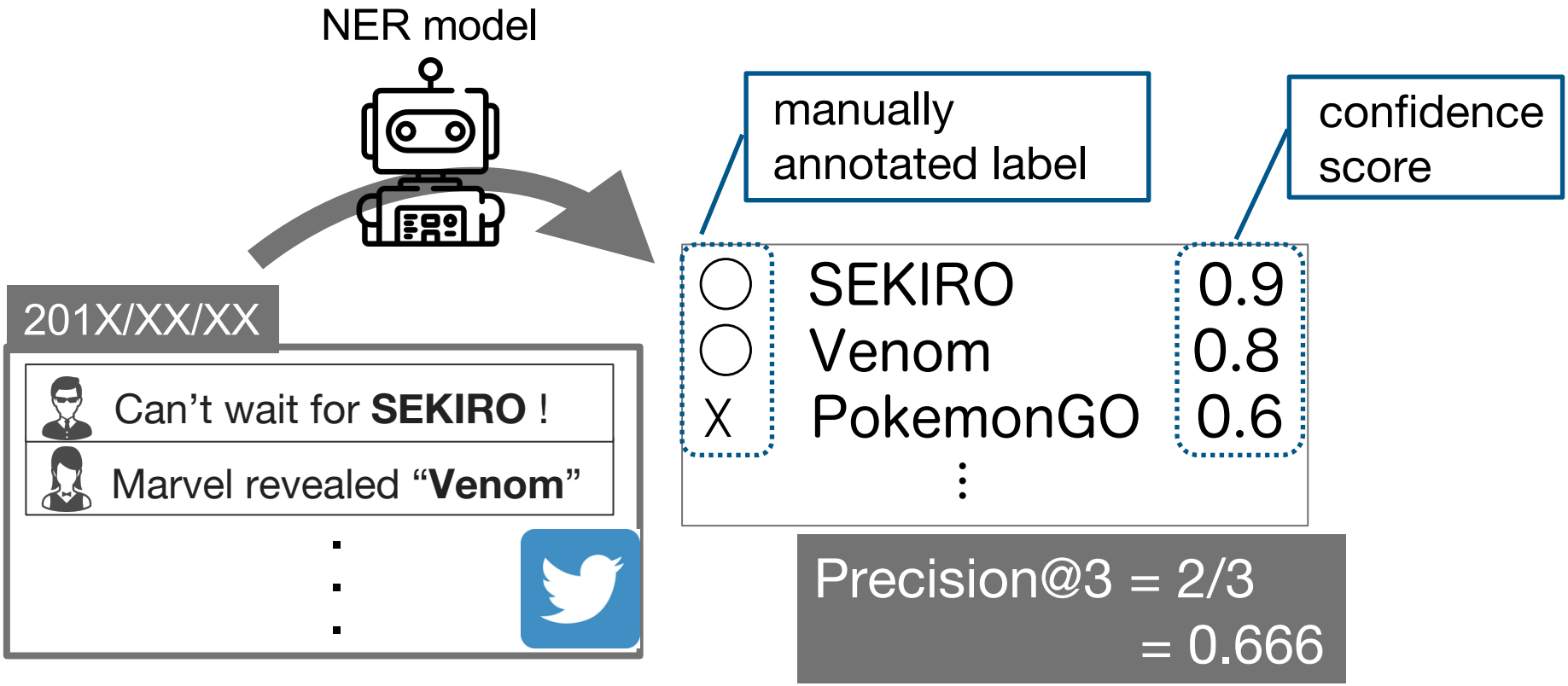
- Emerging entity typing
- Refine distant supervision to remove noises

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

Settings: evaluation method (Precision)

more than 1.5m tweets for each day

Apply each method to retweets of Mar. 21th, Jun. 14th and Sep. 1st, 2017 and then calculate Precision@K for each day



Settings: evaluation method (Recall)

Since we can't obtain a list of all the EEs for evaluating recall, we regard **titles of Wikipedia as a pseudo EE list**

- **To remove noises**, we targeted only entities registered from January 2017 to June 2018 and retweeted over 100 times
- Applied proposed methods to all the retweets since March 12th, 2012 that contain the pseudo EEs (about 9M tweets)