# UE and Nikon at the NTCIR-13 MedWeb Task

**Nga Tran Anh Hang**, Hiroko Kobayashi, Yu Sawai, Paulo Quaresma
University of Evora, Nikon Corp., Nikon Systems Inc.

# Outline

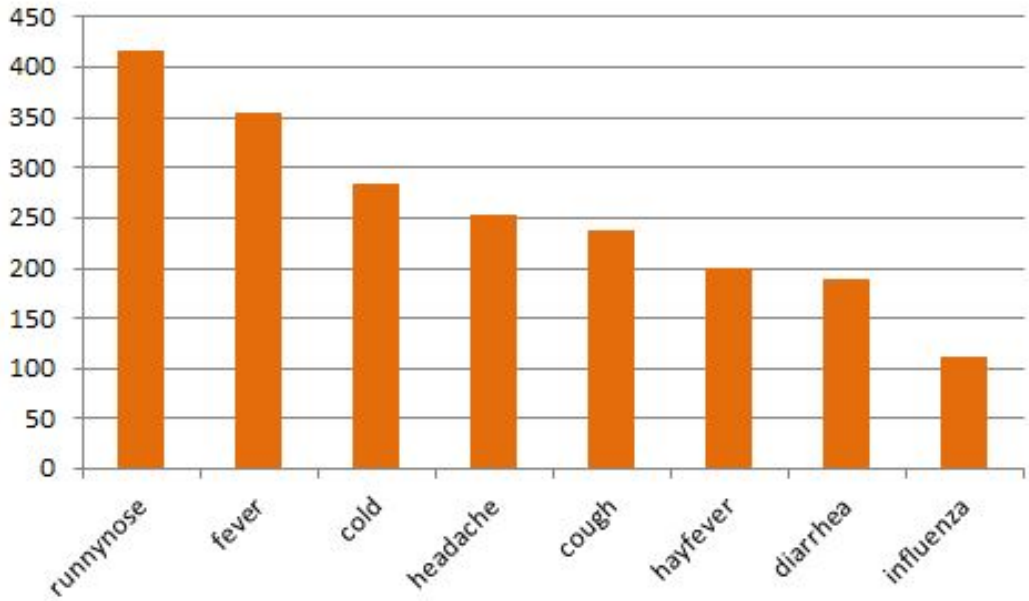
- **Introduction**
  - Task Motivation
- **Methodologies**
  - Rule-based Method (UE-ja-2)
  - Feature-engineering (UE-ja-1, UE-ja-3, UE-en-1)
  - Distributed Representations (UE-en-2, UE-en-3)
- **Results and Discussion**
- **Conclusion**

# Introduction

NLP research is focusing on rather **“clean”** language data.
In reality, there are many difficult cases to detect.

- 犬って鼻づまりとかするのかな？
  (I wonder if dogs get things like stuffy noses?)
- うちのテレビ熱だしすぎで大丈夫かな、これほんと。
  (My TV is giving off an awful lot of heat. Is it okay? Seriously.)

*Table 1.*
*Counts of symptom labels in the training data (1920 pseudo-tweets)*



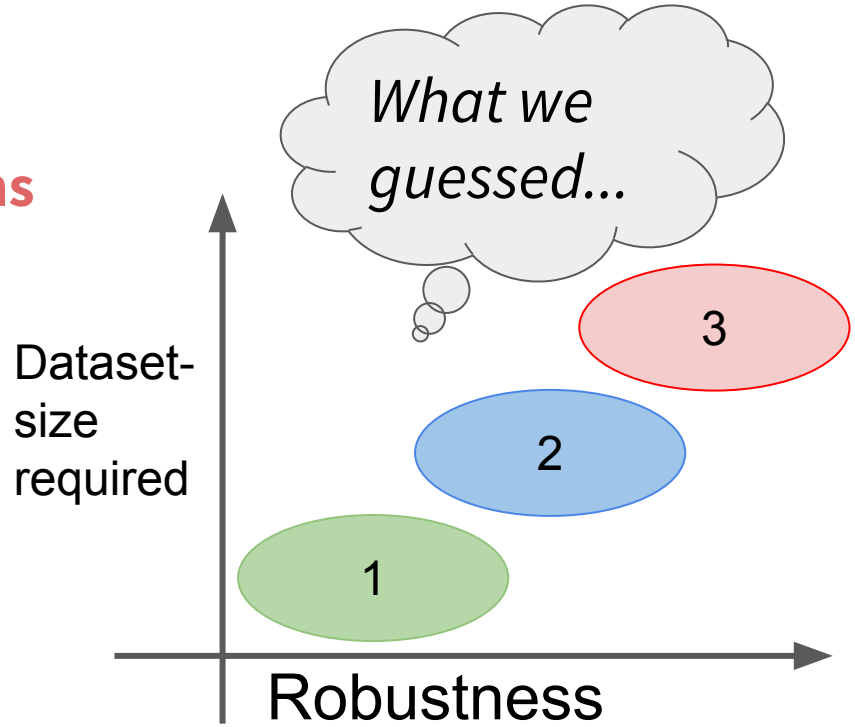
# Task Motivation

- We want to know strength and weakness of popular methods on **"real-world datasets"**.

**1. Rule based**

**2. Feature engineering**

**3. Distributed representations**



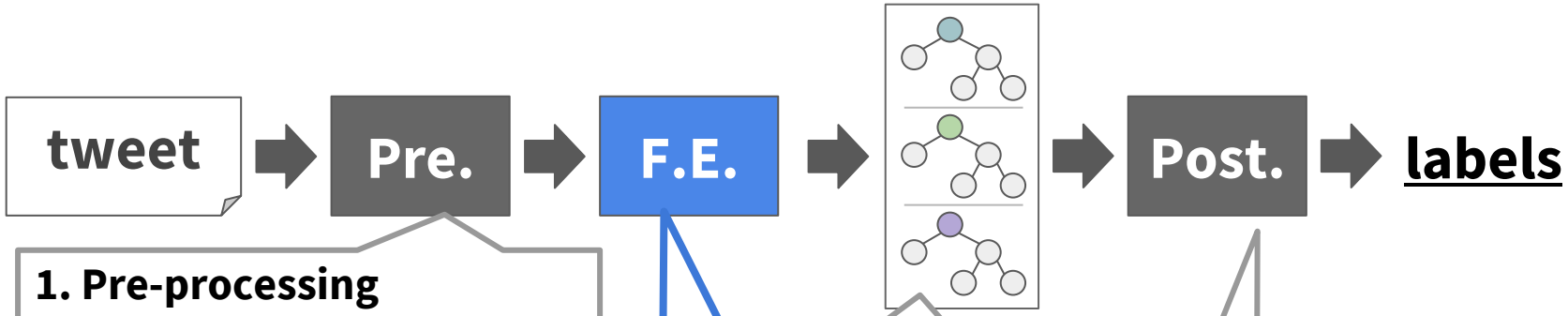
# Methodology: Rule-based Approach (UE-ja-2)


- Pre-processing
  - Extract nouns (Mecab, NEologd)
- Filtering
  - Use NEGATIVE (not symptoms) dictionary (e.g.” 鳥インフルエンザ(bird flu)”)
  - Use rule (except future phrase “明日(tomorrow)” )
- Detection of symptoms
  - Use symptoms dictionary

| | |
|---|---|
| **influenza** | インフル、インフルエンザ |
| **Diarrhea** | 下痢 |
| **..** | **..** |
| **Cold** | 風邪、鼻風邪 |

# Methodology: Feature-engineering Approach

(UE-ja-1, UE-ja-3, UE-en-1)




**1. Pre-processing**
to reduce sparseness and noise

- Normalization of characters, nouns
- For En., replace pronouns with special tokens.

**2. Feature Extraction**
surface features for robustness,
semantic features for long-distance relations

- Surface 1 to 2-grams
- Named-entity (for Ja.)
- SRL based features (subj. verb. pairs, for Ja.)

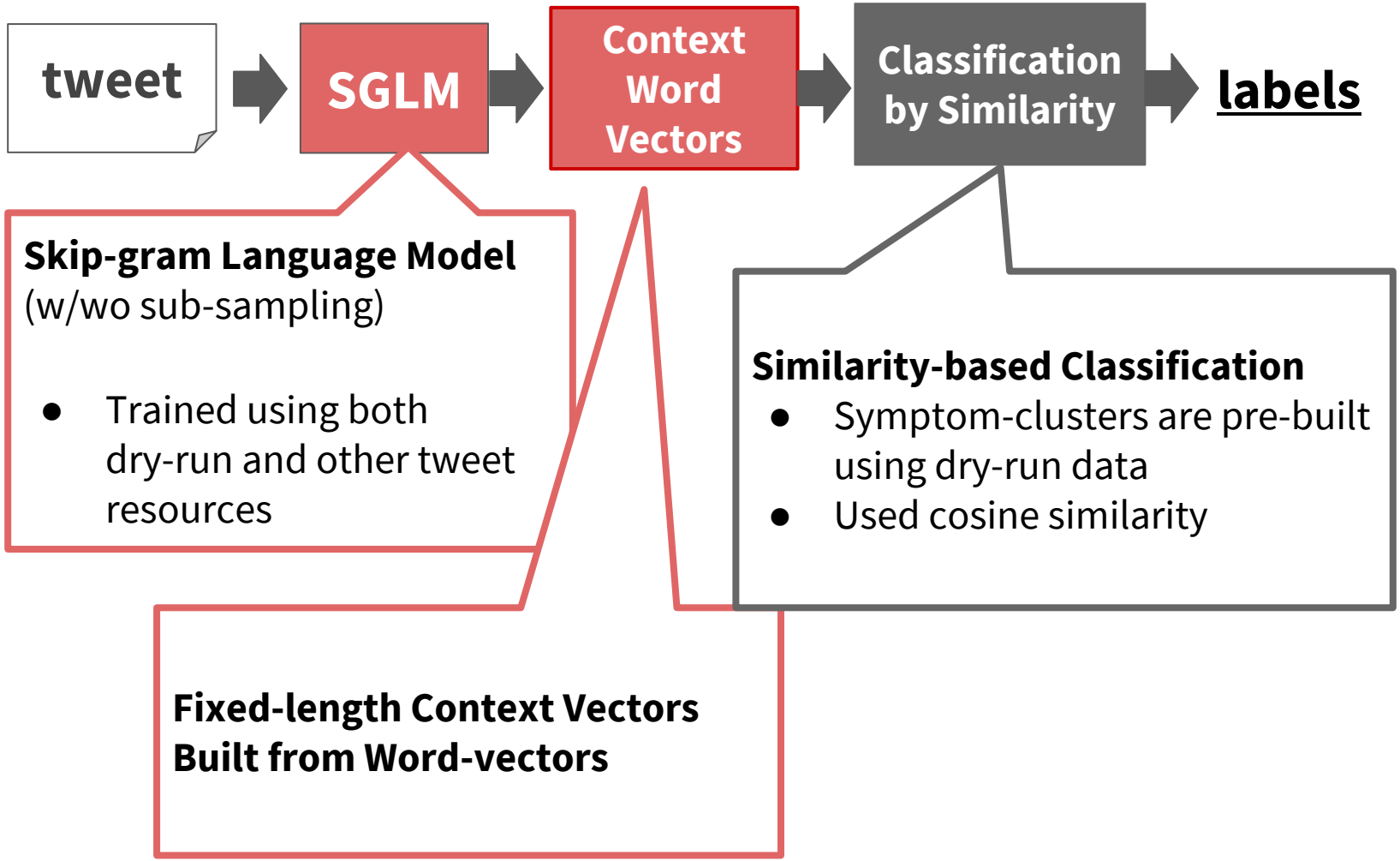
**3. Random Forests**

**4. Post-processing**

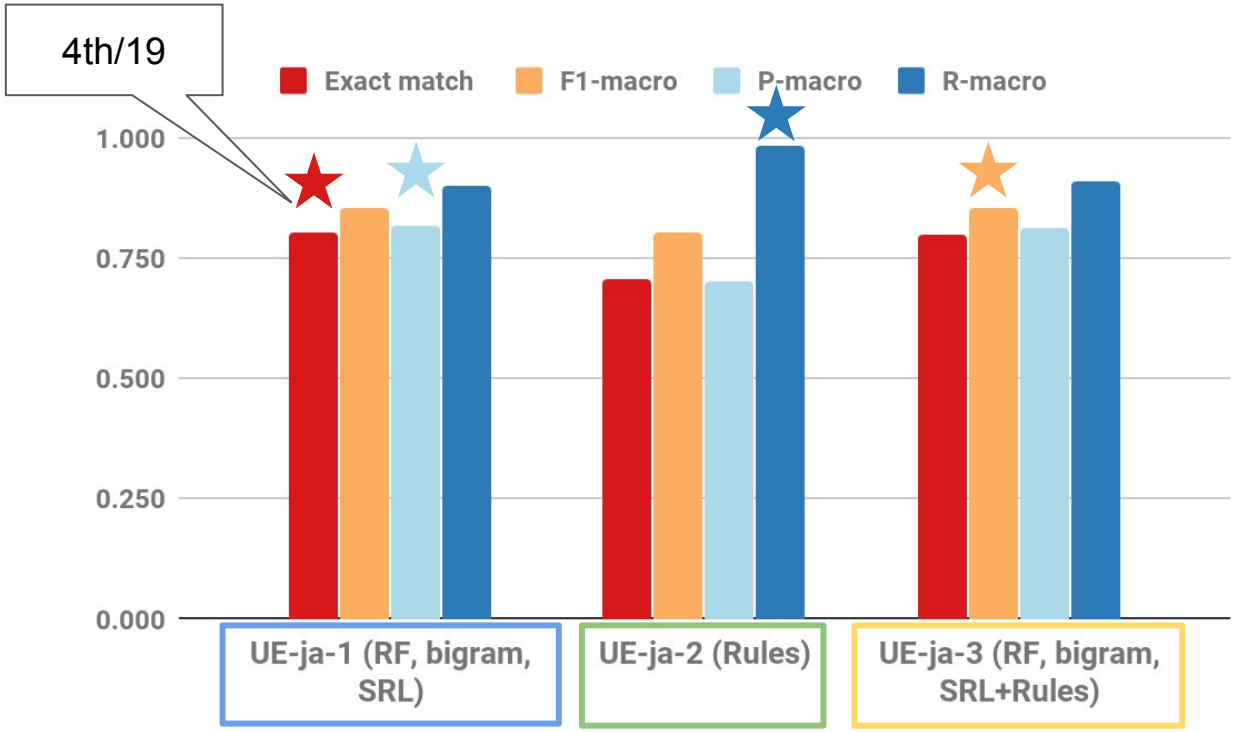
- Co-occurrence rules e.g. Influenza **+ Fever**
- Combined with rule-based model

# Methodology: Distributed-representations Approach

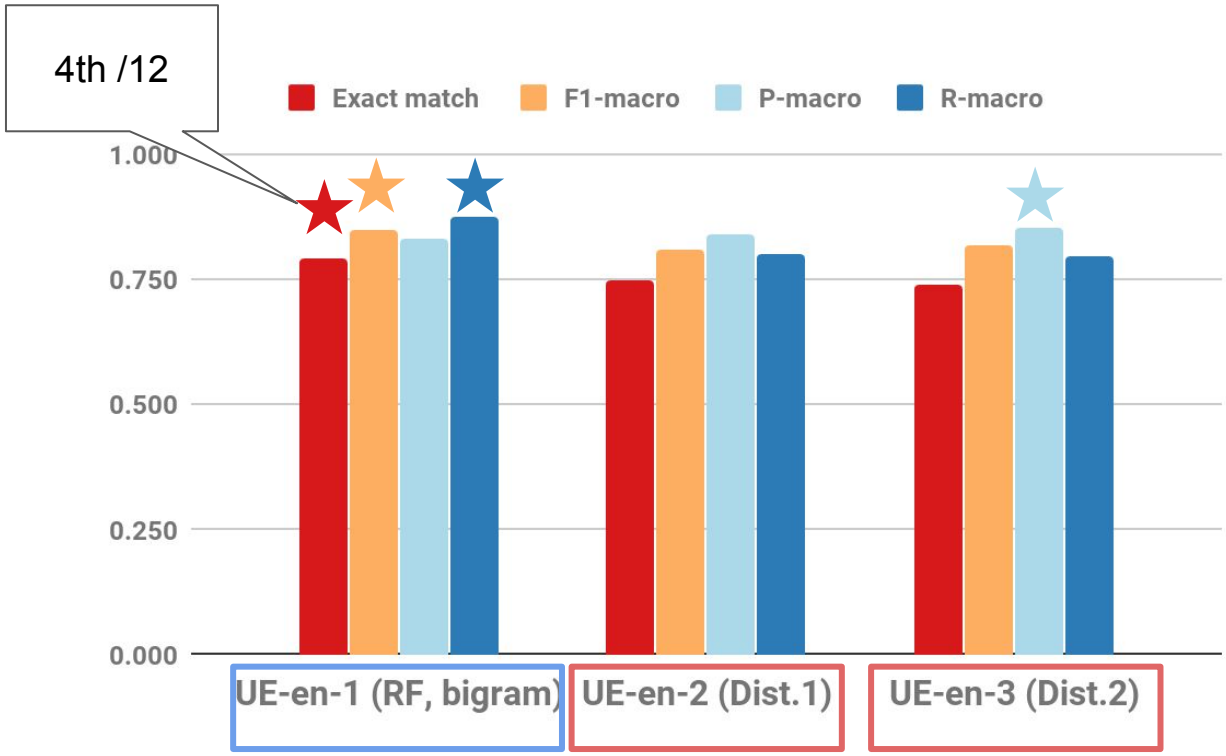
(UE-en-2, UE-en-3)



# Results of Japanese Subtask



# Results of English Subtask



# Results and Discussion: Error Analysis

- More knowledge is needed, such as ontology
  - Non-human case : 「犬って鼻づまりとかするのかな？」
    *( I wonder if dogs get things like stuffy noses?)*
- Discourse level knowledge is needed (Jp corpus)
  - 「インフルかと思って病院に行ったけど、検査したら違ったよ。」
    *(I thought I had the flu so I went to the doctor, but I got tested and I was wrong.)*
- Other things to be mentioned
  - Dealing with dialects:「あかん」
  - New-born expressions (newborn words/phrases on the Internet)

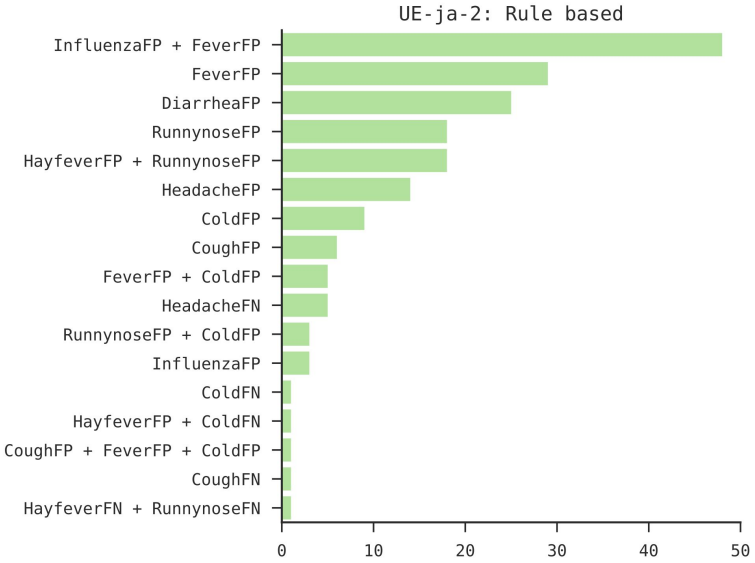
# Conclusions

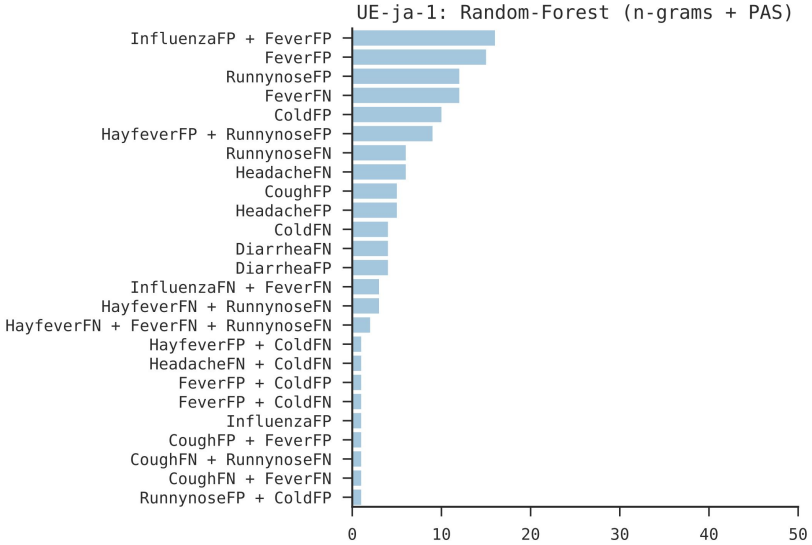
- **Simple methods can achieve good performance!**
  - We focused on **practical** application
  - Applied Rule-based, Feature-engineering based, Distributed-representation based systems
- **There are still many things to be improved**
  - Handle explicit knowledge of symptoms.
  - Discourse, and causal structure
  - Neologisms, slang, dialects (for Japanese corpus)
  - Jokes, time and space detection

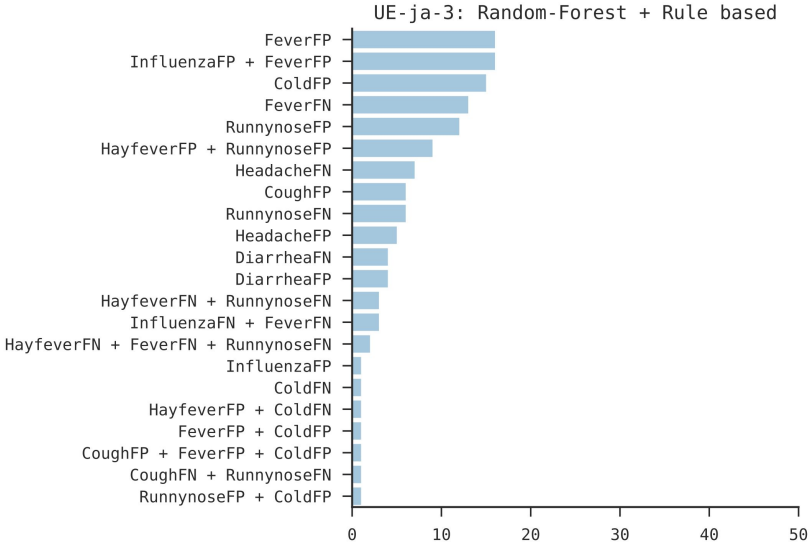
Thank you!

# Appendix

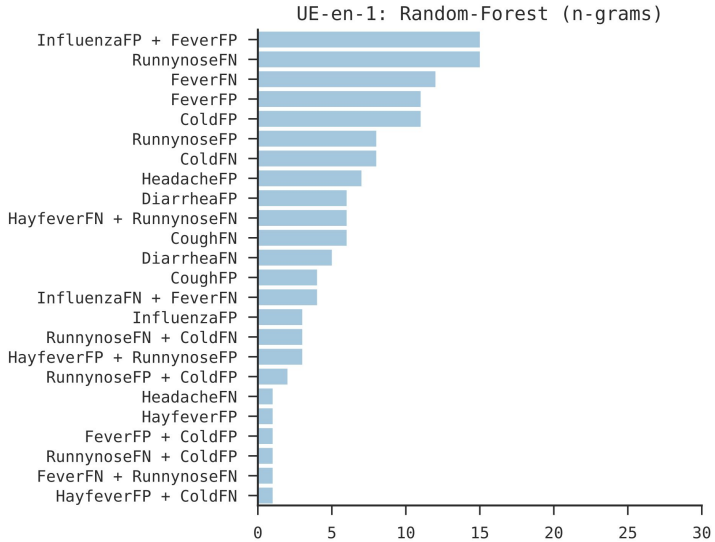
# Error Statistics (Ja. subtask)

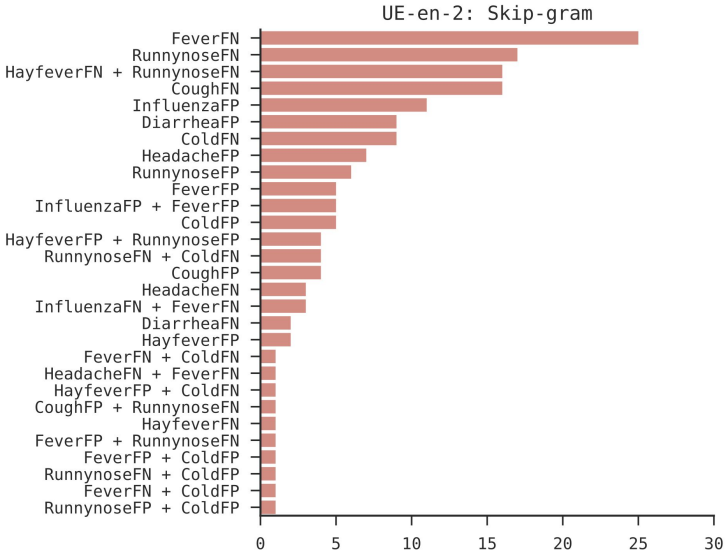


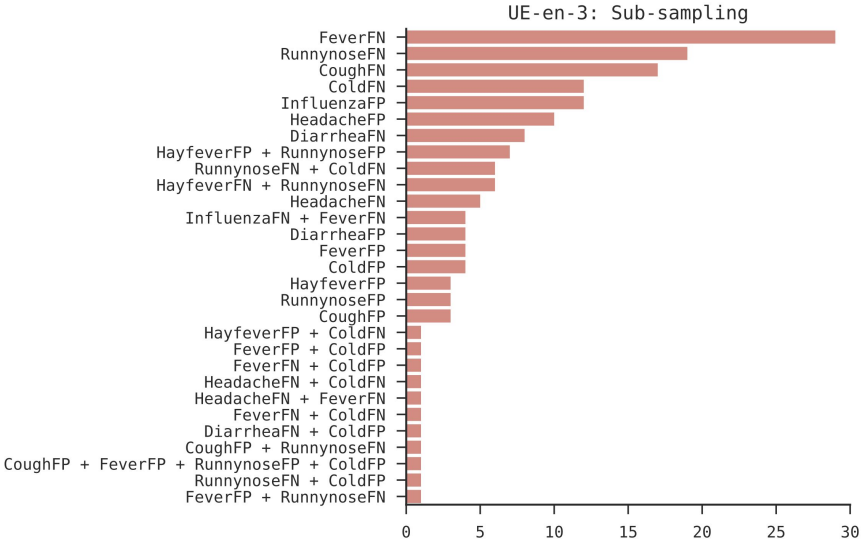



# Error Statistics (En. subtask)



# Details of Pre-processing & Custom Dictionary

(UE-Ja-1&3)

- Preprocessing
  - Applied normalization used in https://github.com/neologd/mecab-ipadic-neologd/wiki/Regexp
- Custom dictionary
  - Contains nouns which are not chunked properly by MeCab-IPADic-NEologd
  - Also used for normalizing by dictionary-form（原形）entries:
    e.g. {*鼻ずまり, 鼻づまり, 鼻詰まり -> 鼻づまり}
    *A word or phrase with *asterisk is marked as spelling or grammatical error.*
  - Some metaphorical usages found in dry-run data are also normalized:
    e.g. {頭痛の種, 頭痛のもと -> 面倒事}

# Methodology: Distributed-representations Approach

- Sub-sampling of frequent words

**SOURCE TEXT**

| | | | | | |
|---|---|---|---|---|---|
| **I** | have | a | headache, so I've decided to go home. | | |
| I | **have** | a | headache | so I've decided to go home. | |
| I | have | **a** | headache | so | I've decided to go home. |
| I | have | a | headache | **so** | I've decided to go home. |

**TRAINING SAMPLE**

(I, have)
(I, a)

(have, I)
(have, a)
(have, headache)

(a, I)
(a, have)
(a, headache)
(a, so)

(so, I)
(so, have)
(so, headache)
(so. I've)