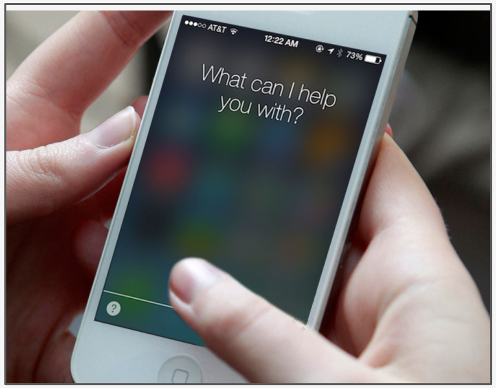
UT Dialogue System at NTCIR-12 STC

Shoetsu Sato[1], **Shonosuke Ishiwatari**[1], **Naoki Yoshinaga**[2], **Masashi Toyoda**[2], **Masaru Kitsuregawa**[2,3]

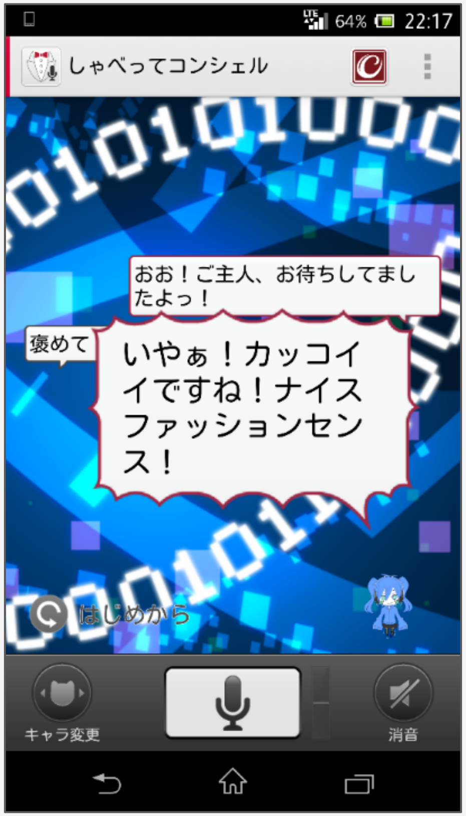
The University of Tokyo, [2]IIS, the University of Tokyo, [3]NII, Japan

A lot of dialogue systems that can chat have appeared



Siri (Apple)

Cortana (Microsoft)



しゃべってコンシェル (NTT Docomo)

http://www.idownloadblog.com/
http://techcrunch.com/2015/01/05/facebook-wit-ai/
http://ameblo.jp/cos-120/entry-11748747974.html

Recent approaches for chatting dialogue systems

- Data-driven approaches using dialogue data from **social media** are promising [Ritter+, '10]

U: Utterance
R: Response

U: あの人どう思う？
R: ああいう人間ほんと嫌い

U: また残業か・・・
R: 生き残ろうな・・・

U: 魚介嫌いでした？
R: そんなこと無いですよ。

Challenge we have tackled in STC task

- The diversity of **domains** (topics, speaking styles, etc…) makes it difficult to learn

U: Utterance
R: Response

U: あの**人**どう思う**？**
R: ああいう人間ほんと嫌い

U: また**残業**か・・・
R: 生き残ろうな・・・

U: **魚介**嫌い**でした？**
R: そんなこと無いですよ。

OVERVIEW

Goal: building a domain-aware dialogue model

Idea: Divide conversation data into **domain-consistent subsets** to train multiple specific LSTM-based dialogue models

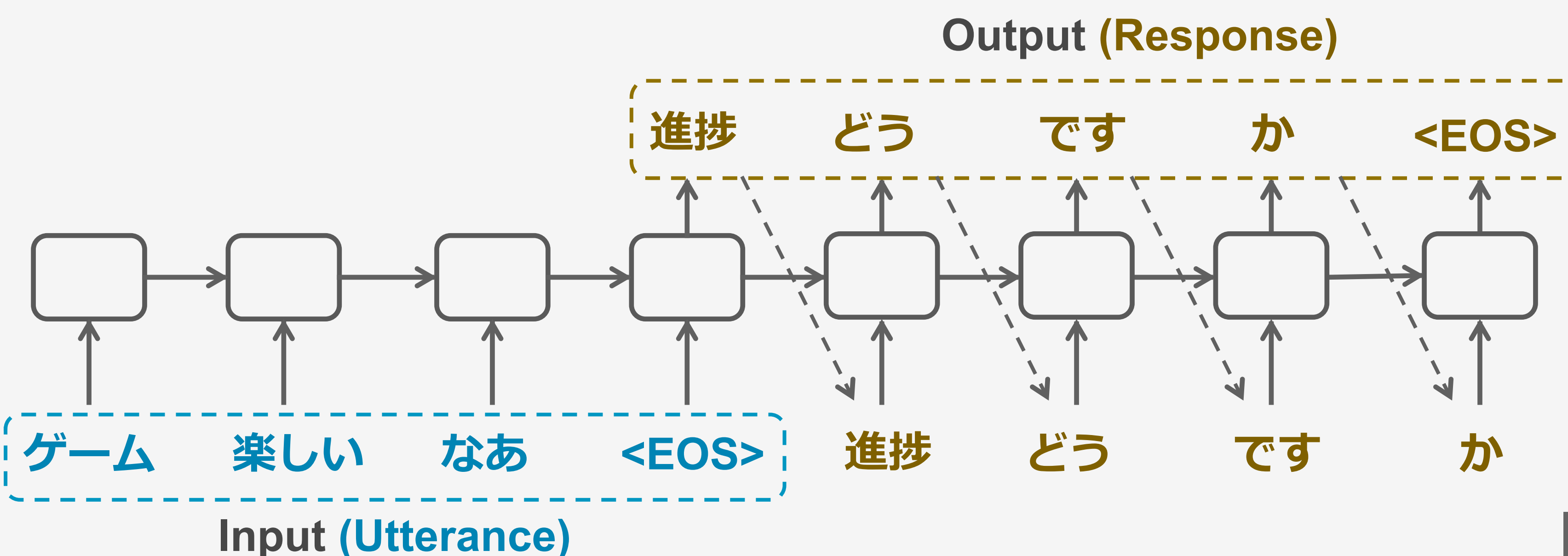
Evaluation: **response selection** from candidates

Does domain consistence compensate for reduction of training data per a model?

RELATED WORK

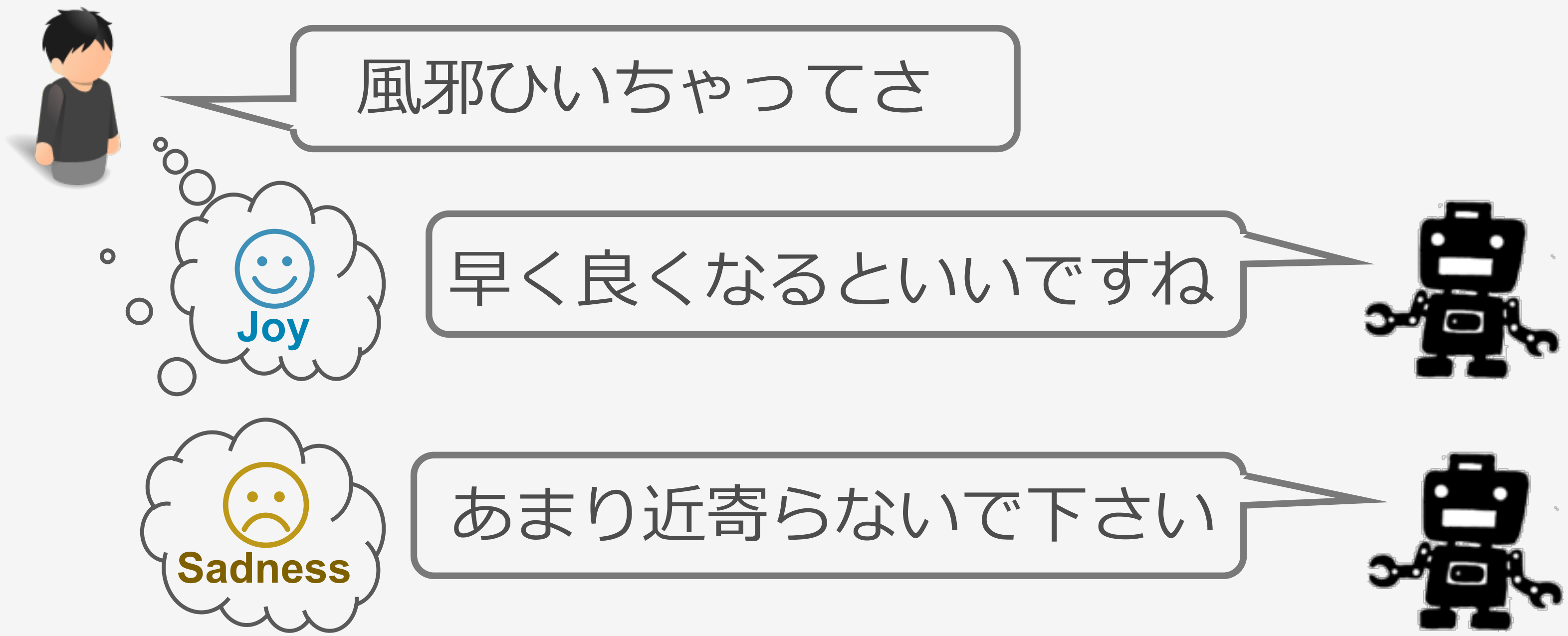
Recent promising approach to generate responses

- We employed recent promising Long-Short Term Memory based Recurrent Neural Network (LSTM-RNN) dialogue model [Vinyals+, '14]



Predicting and Eliciting Addressee's Emotion in Online Dialogue [Hasegawa+, '13]

- Generate a response that elicits a **specific emotion** in the addressee's mind



Overview of the related work and target of our method

Point

– General LSTM based methods employed a single model trained from all data

– It is impossible to enumerate all domains in human dialogues

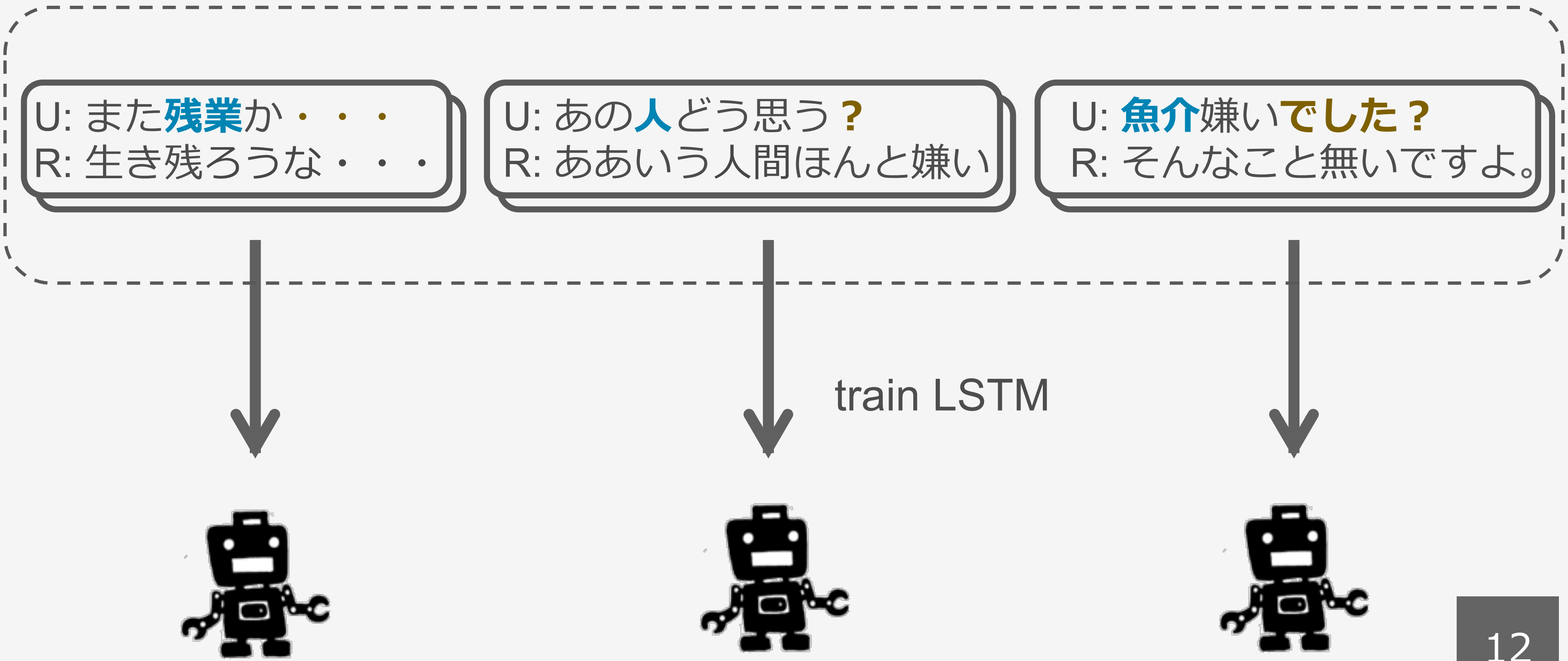
Purpose

– Capture the difference of domains **automatically** as clusters and train **multiple** models

METHOD

Our approach: K-cluster model (1/2)

- **Cluster** the dialogues for each of the **unlabeled domain**, and train multiple models



Our approach: K-cluster model (2/2)

Utterance

就職したくない・・・

Find the nearest domain by human utterance and utterances in training subsets

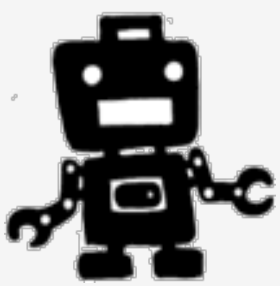
U: また残業か・・・

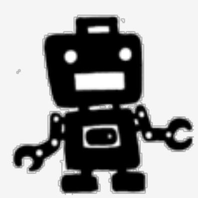
U: あの人どう思う？

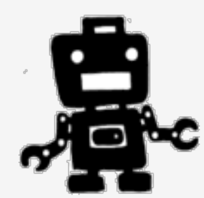
U: 魚介嫌いでした？

Response

市民、労働は義務です。

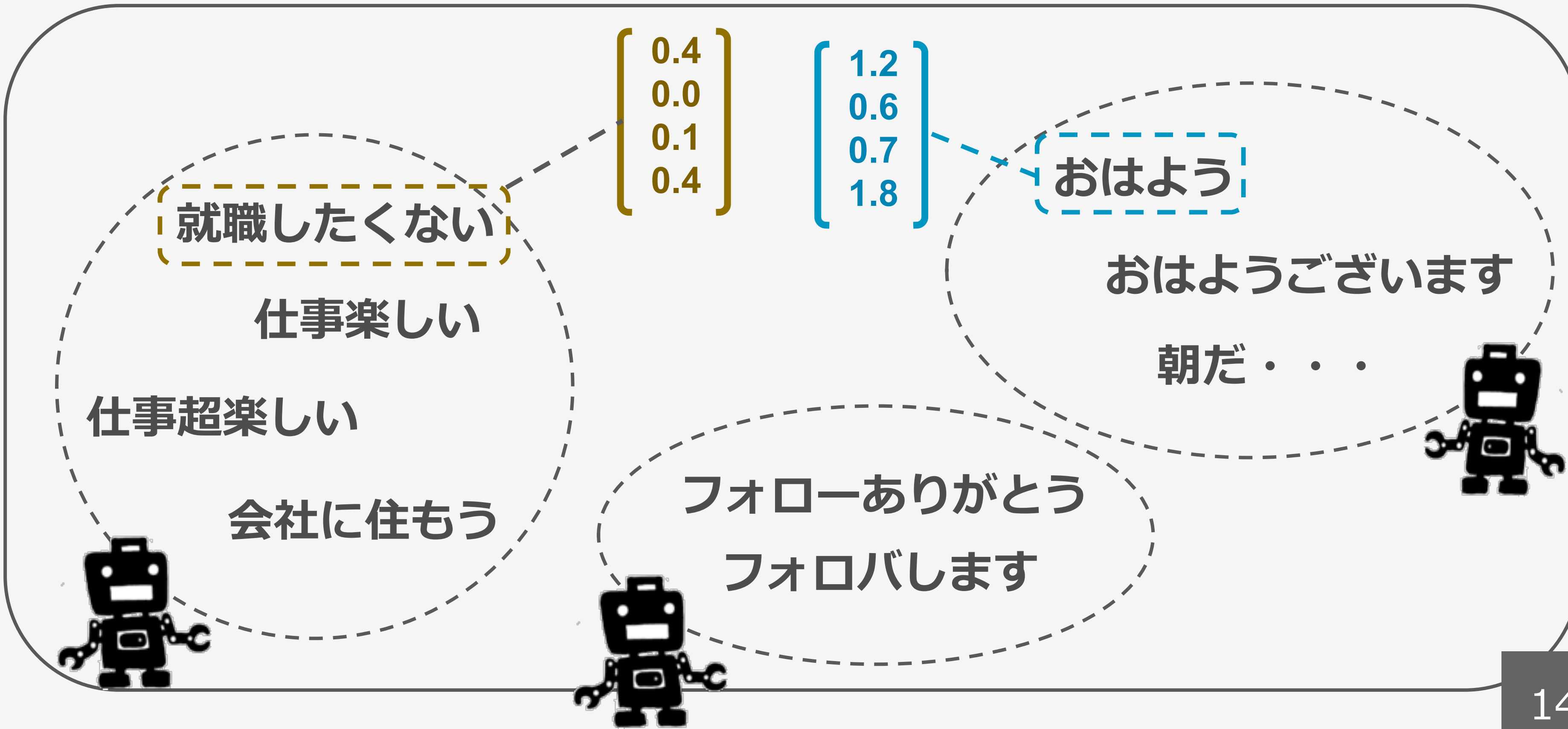






How to automatically handle the domains in each utterance (1/2)

- Apply **k-means clustering** to the utterance vectors and regard **clusters as subsets** of the training data



How to automatically handle the domains in each utterance (2/2)

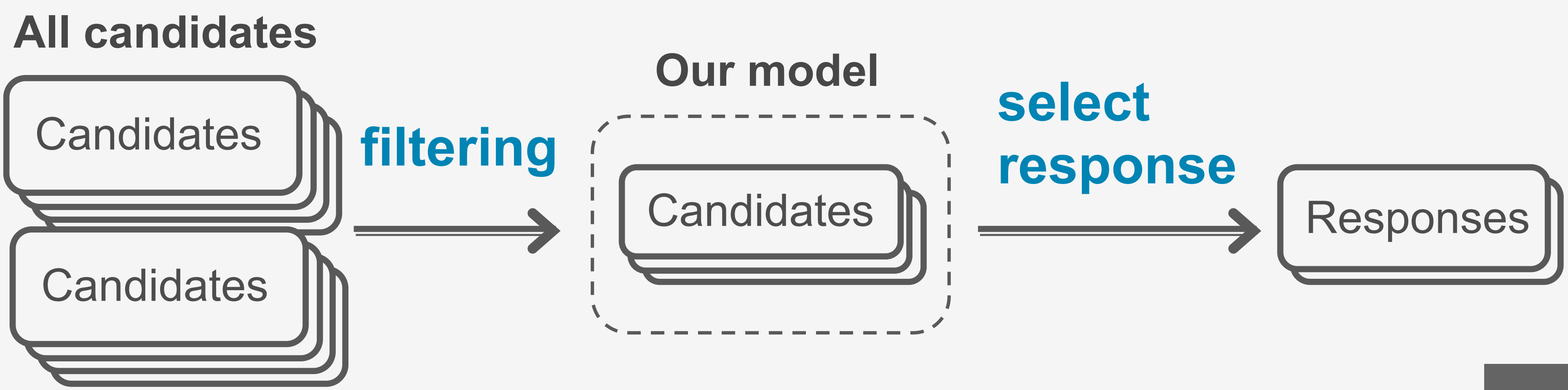
- Represent each utterance as a vector built from **word embeddings** [Mikolov+, ‘13]
- The density of word embeddings would solve sparseness problems in short texts compared with Bag-of-Words

$$\overrightarrow{\text{就職}} \quad \overrightarrow{\text{し}} \quad \overrightarrow{\text{たく}} \quad \overrightarrow{\text{ない}} \qquad\qquad \overrightarrow{\text{就職したくない}}$$

$$\begin{bmatrix} 0.3 \\ 0.2 \\ -0.1 \\ 0.6 \end{bmatrix} \begin{bmatrix} 0.1 \\ -0.2 \\ 0.3 \\ 0.5 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0.1 \\ -0.1 \\ 0.3 \end{bmatrix} \begin{bmatrix} 1.0 \\ -0.1 \\ 0.3 \\ 0.2 \end{bmatrix} \xRightarrow{\textbf{average}} \begin{bmatrix} \mathbf{0.4} \\ \mathbf{0.0} \\ \mathbf{0.1} \\ \mathbf{0.4} \end{bmatrix}$$

Response candidate filtering

- In response selection task from many candidates, our model's **high computational cost** causes a problem
- To reduce the number of candidates into **500**, we employed a fast SVM classifier [Yoshinaga+, '10]



EXPERIMENTS

3 experiments we did

- **Experiment 1: Small response selection task**
  - Evaluate how our method effects in response selection
  - Select response from 20 candidates without filtering
- **Experiment 2: Filtering performance**
  - Evaluate to what extent our filter can select proper candidates
- **Experiment 3: NTCIR-12 formal run**
  - Evaluate whole performance of our system (clustered-LSTM, and filter)

Experiment 1 : small response selection task

- **Dataset: Twitter**

  Utterance-response (tweet-reply) pairs crawled from Twitter: **100K** for training, **1K** for test

  – Provided for *NTCIR-12 Short Text Conversation Japanese Task* [Shang+, ‘16]

- **Evaluation: Response selection**

  The proportion of test tweets where we succeeded to select the correct (actually replied) response from randomly chosen **20 candidates**

Evaluation detail

Utterance

発表つらいんだけど

Ranked 20 Response Candidates

① わかる

② 自分の研究を知ってもらう良い機会だと思うよ

③ 今完全に鬱だよ

④ その店美味いよね

⋮

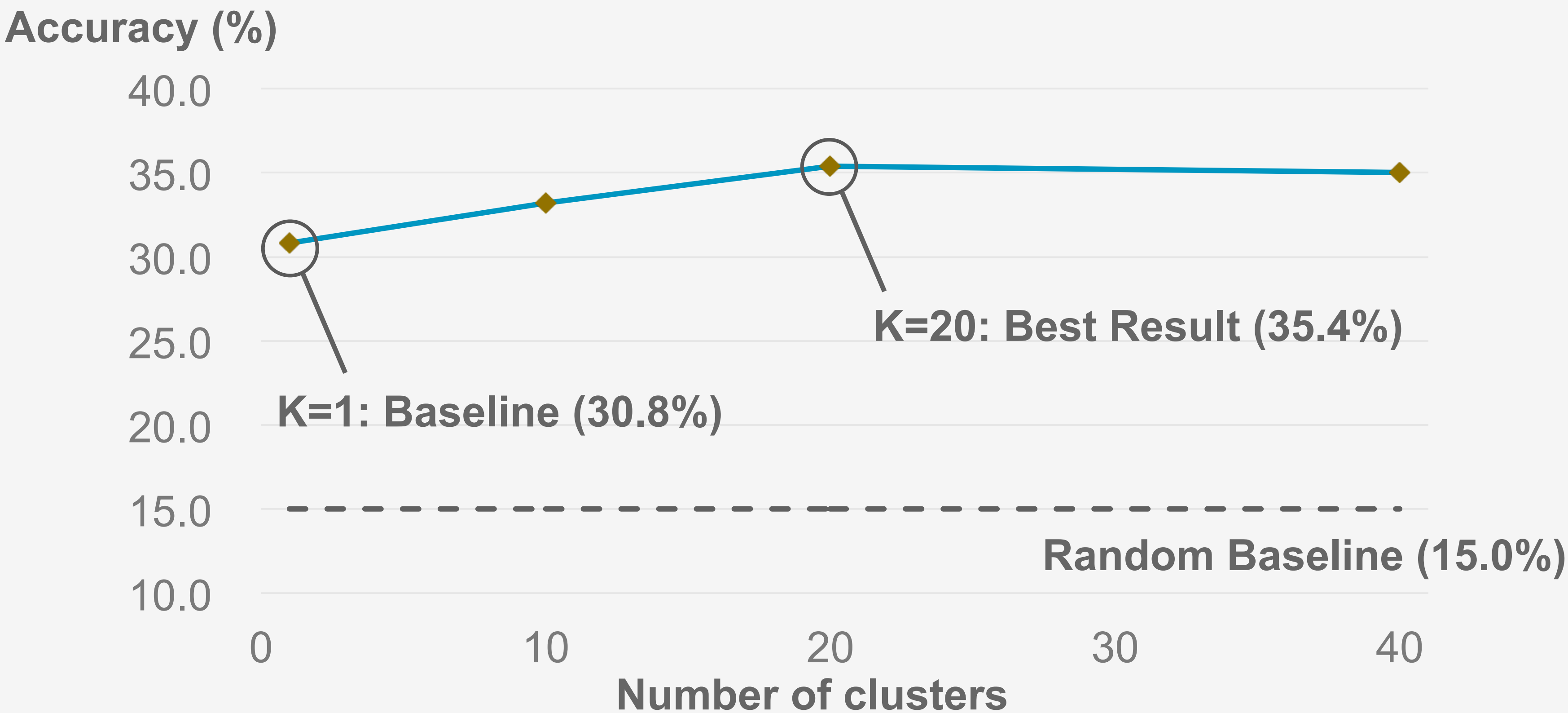
⑳ くぁwせｄｒｆｔｇ

Correct Response

We defined it as success if the **top-3 responses** include the correct response

Results of K-cluster model

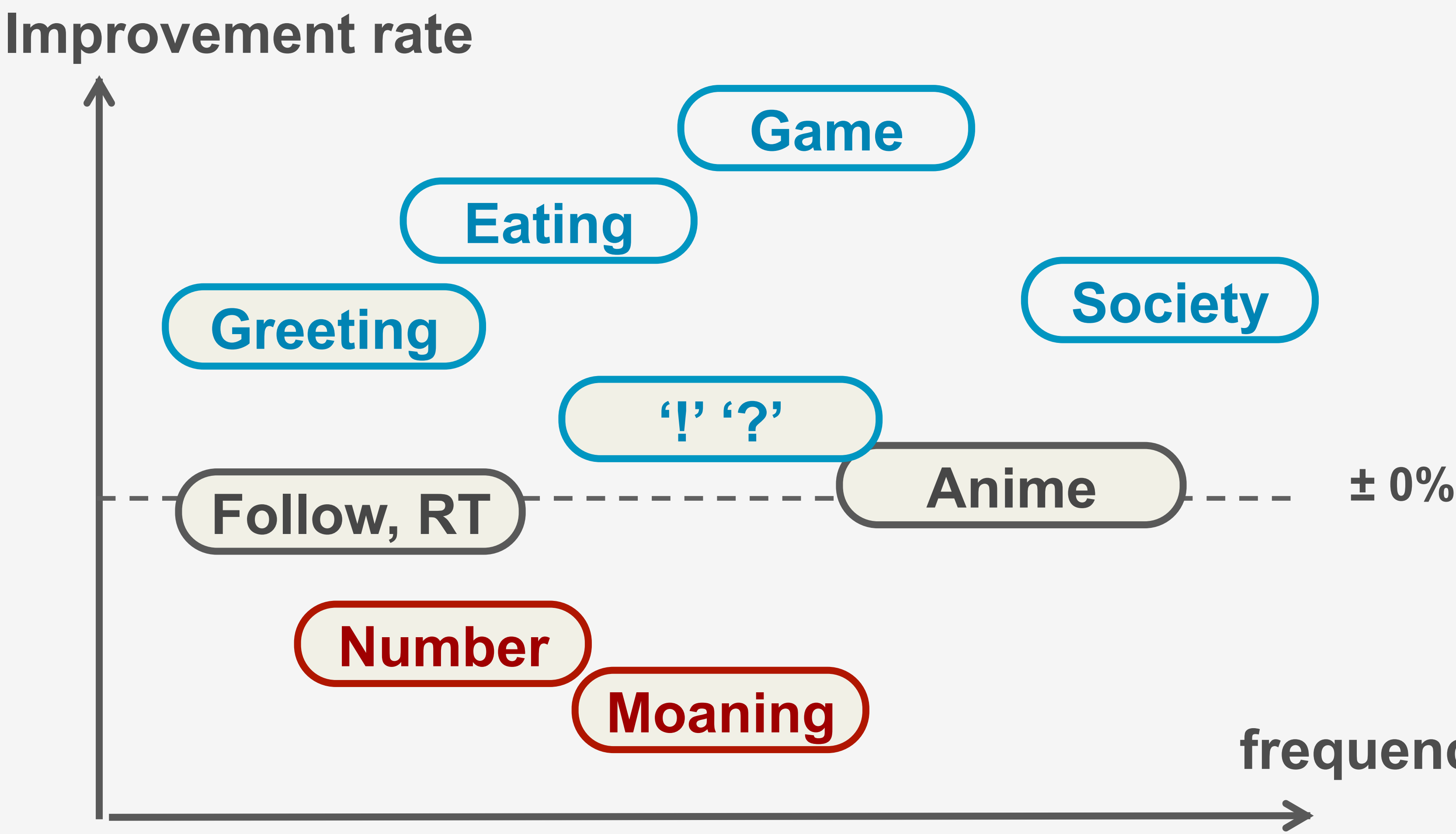
- We compared 1, 10, 20, and 40 cluster models increasing number of clusters until the accuracy was saturated



Our best 20-cluster model outperformed baseline at **4.6%**

Examples of clusters we labeled by hand from their frequent utterances

- 13 models out of 20 cluster got improved



Examples of the selected responses

(1)

Utterance	あ、見るの忘れてた。おめでとう!
Proposed	ありー! 見直してくれてありがとう!
Baseline	今年は1年ありがとうございました

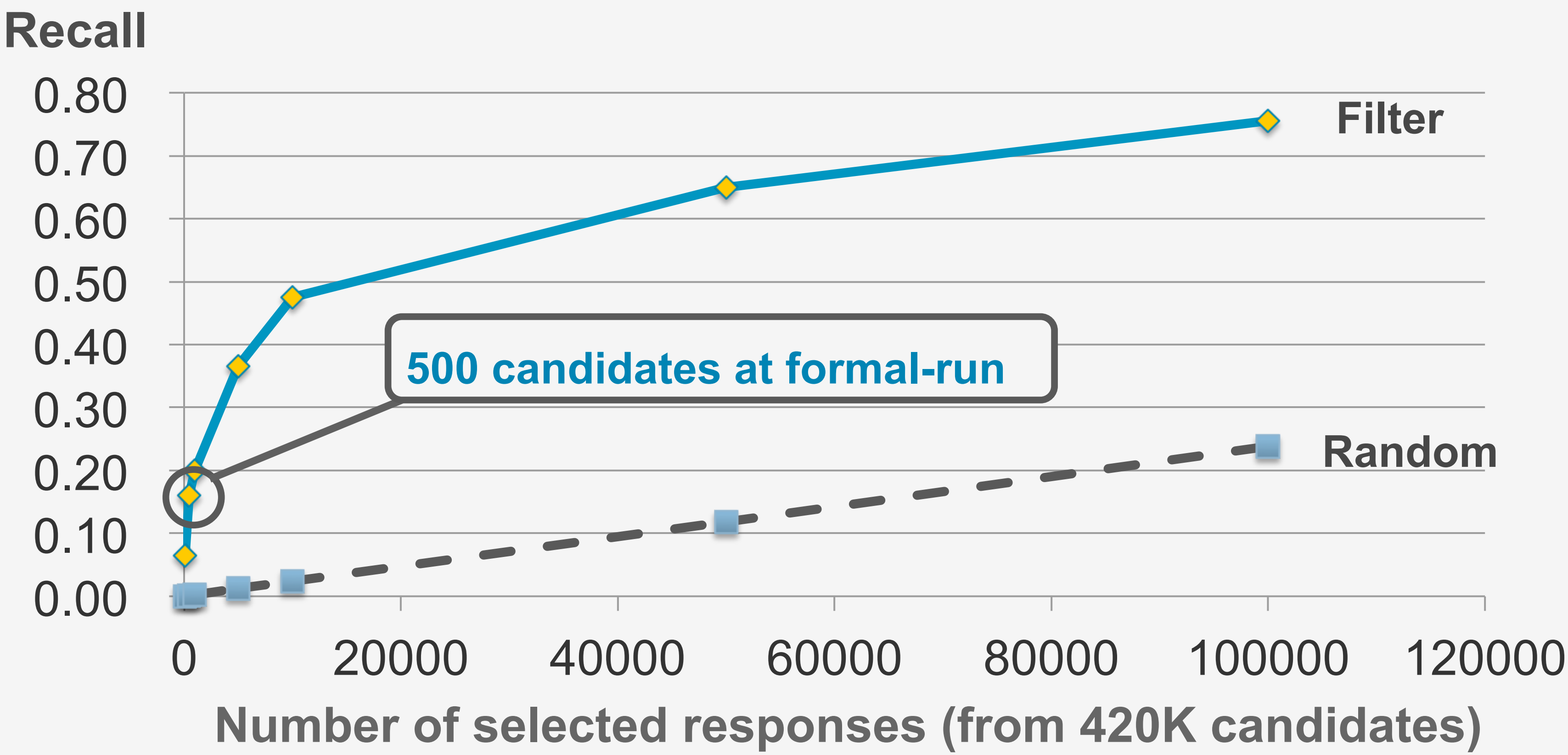
(2)

Utterance	カントリーマァムのドリンクのやつが 見つかりません。
Proposed	ローソン限定じゃなかったっけ?
Baseline	先輩、おはよーございます♪

Our proposed model tends to stop selecting typical responses

Experiment 2: Filtering performance

- Evaluate the filter by recall, whether top-N filtered candidates include the **correct response**



Filtering effectively reduced the number of candidates

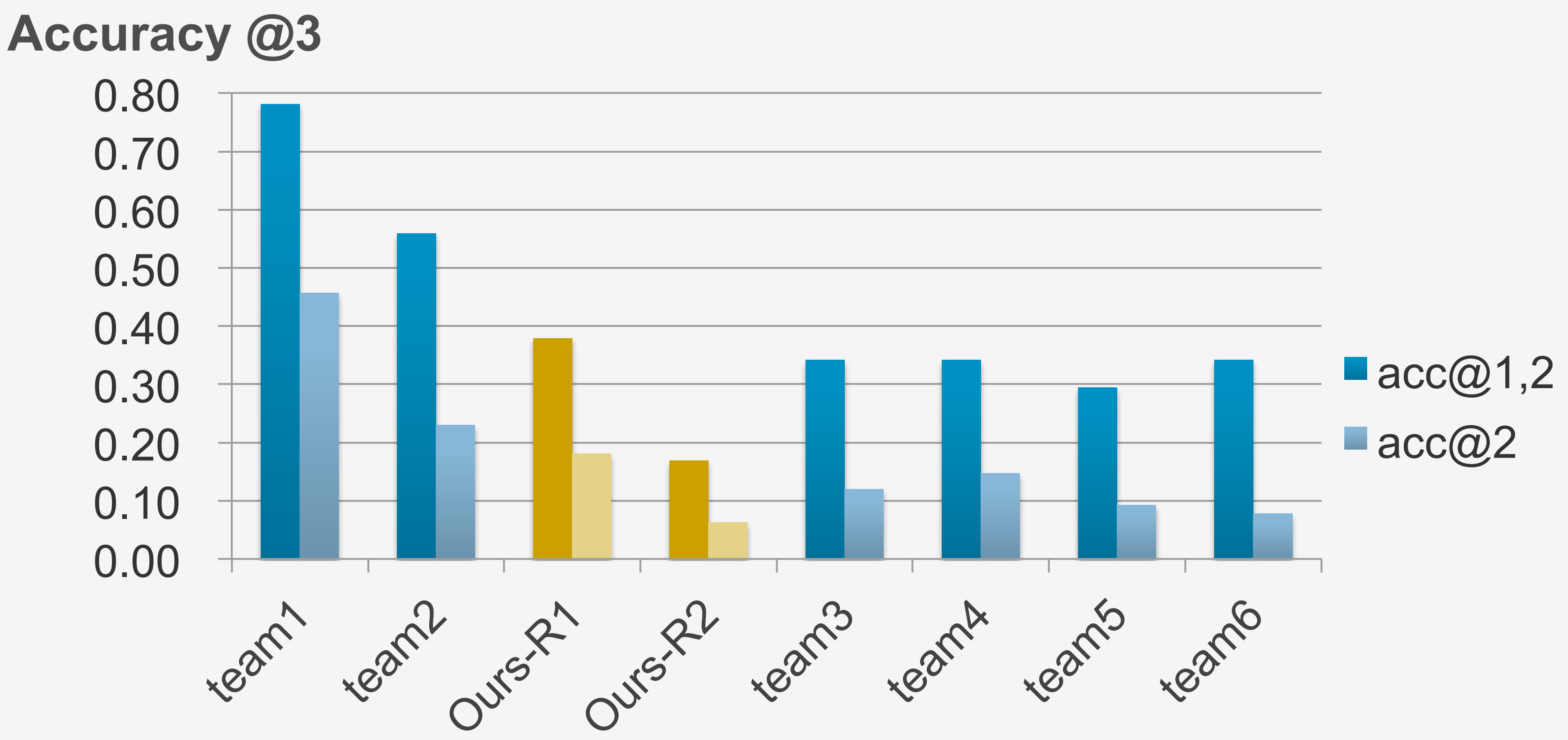
Experiment 3 : NTCIR-12 STC Japanese Task

- **Model: 20 cluster model**
  The best one evaluated at experiment 1, 20 cluster model trained from **100k** utterance-response pairs

- **Evaluation:**
  – For the **204** provided test utterances, select responses from **500k** candidates
  – responses are assigned scores of **0 (inappropriate)**, **1 (appropriate in some context)**, **2 (appropriate)** by human annotators

Accuracies of selected top-1 responses at NTCIR-12 STC Japanese Task

Ours-R1: Filtering + 20-cluster LSTM model
Ours-R2: Filtering only



Our system selected better responses from filtered candidates

Summary

By response selection test we confirmed the effect of cluster-based **domain-aware** dialogue model

- Domain-consistent training subsets made better results in spite of reduction of training data

- By filtering candidates, our system could effectively select responses

RESULTS FOR EACH CLUSTER

Results in each cluster (20-cluster model)

domain (topics, wording, writing style)	#elems train	#elems test	#corr ours	#corr baseline	improvement $\frac{\Delta \#corr}{\#elems\ (test)}$
-	11801	108	**38**	27	10.19%
-	11524	124	**37**	32	4.03%
politics, economics, social matters	10294	130	**48**	38	7.69%
-	9743	94	**32**	23	9.57%
animation, comics	6747	56	**11**	10	1.79%
-	6552	66	**24**	23	1.52%
game	5677	50	**13**	5	16.00%
-	5627	45	**14**	13	2.22%
end with '?' r '!'	5190	63	**17**	15	3.17%
moaning (esp., sleepy, weary)	5064	52	17	**21**	-7.69%
-	4908	50	22	**24**	-4.00%
numbers	3803	31	5	**7**	-6.45%
eating	2630	16	**6**	4	12.50%
frank acknowledgment (follow, RT)	2252	33	29	**30**	-3.03%
end with '!!!'	1869	17	**8**	**8**	0.00%
polite acknowledgement (follow, RT)	1553	13	**12**	**12**	0.00%
greetings	1537	21	**7**	6	4.76%
end with '. . .'	1326	12	**3**	2	8.33%
polite morning greetings	1174	13	**9**	6	23.08%
shouting with word lengthing or repetition	729	6	**2**	**2**	0.00%
	100000	1000	354	308	4.60%