Analysis of Utterance Embeddings and Clustering Methods Related to Intent Induction for Task-Oriented Dialogue

Jeiyoon Park[1], Yoonna Jang[1], Chanhee Lee[2] Heuiseok Lim[1*]

[1]Department of Computer Science and Engineering, Korea University,

[2]Naver Corporation

KOREA UNIVERSITY

NAVER

Overview

- **Why Intent Induction?**
  - With the skyrocketing demand for conversational AI, the more user utterances a dialogue system encounters, the more unknown intents it does.
  - Predefining user intent is expensive and it is impossible to annotate all the user intents.

- **Our Approach**
  - Previous studies did not conduct an in-depth analysis of the application of existing models to intent induction that might cause performance degradation problems
  - We postulate there are two salient factors for automatic induction of intents: (1) clustering algorithm for intent labeling and (2) user utterance embedding space
  - We analyze how the two key factors affect to user intent clustering and intent induction.

- **Contribution**
  - We demonstrate that the combined selection of utterance embedding and clustering method in the intent induction task should be carefully considered.
  - We present that pretrained MiniLM with Agglomerative clustering shows significant improvement in NMI, ARI, F1, accuracy and example coverage in intent induction tasks.
  - We find that there is a trade-off relationship between NMI, ARI and Example coverage.

Task 1: Intent Clustering

- **Result Analysis (Embedding)**
  - m denotes multilingual model, u stands for unsupervised model, and + means large model.

Clustering Method	K-means Clustering							
Metric	**NMI**	**ARI**	**ACC**	**Precision**	**Recall**	**F1**	**Example Coverage**	**#K**
$\text{EASE}^{m}_{ROBERTA}$	33.4	14.2	28.0	28.0	**69.0**	39.9	43.9	5
EASE_{BERT}	36.1	18.1	35.3	36.8	58.9	45.3	53.5	8
EASE^{m}_{BERT}	38.4	10.9	23.6	42.7	24.2	30.9	87.6	44
$\text{EASE}_{ROBERTA}$	45.8	25.4	40.0	43.7	60.7	50.9	65.6	12
$\text{DiffCSE}^{trans}_{BERT}$	43.8	15.5	29.4	49.5	30.6	37.8	90.1	40
$\text{DiffCSE}^{sts}_{BERT}$	46.6	16.9	29.9	50.4	30.8	38.2	89.9	43
$\text{DiffCSE}^{sts}_{ROBERTA}$	53.9	29.0	45.1	57.2	46.8	51.5	91.1	30
$\text{DiffCSE}^{trans}_{ROBERTA}$	55.0	23.1	35.7	60.6	35.7	44.9	96.7	50
SimCSE^{u}_{BERT}	31.7	13.3	27.9	27.9	<u>65.0</u>	39.0	43.9	5
$\text{SimCSE}^{u}_{BERT+}$	47.0	25.7	38.4	46.9	46.9	46.9	76.8	19
$\text{SimCSE}^{u}_{ROBERTA}$	51.2	29.0	44.4	49.7	61.6	55.0	70.5	14
SimCSE_{BERT+}	53.0	27.8	39.8	55.0	42.2	47.8	91.0	31
SimCSE_{BERT}	53.1	24.4	39.0	60.5	39.5	47.8	96.3	44
$\text{SimCSE}^{u}_{ROBERTA+}$	53.2	25.9	42.2	56.8	43.8	49.5	91.7	32
$\text{SimCSE}_{ROBERTA}$	56.6	28.8	42.7	60.8	43.3	50.6	91.3	36
$\text{SimCSE}_{ROBERTA+}$	56.8	28.9	41.3	62.5	41.6	49.9	98.4	42
Glove^{avg}	30.5	7.0	20.6	34.6	22.2	27.0	92.2	50
MPNet	59.3	32.3	46.1	66.0	47.1	54.9	96.5	42
MiniLM_{L6}	59.3	35.7	52.6	62.2	54.9	58.4	92.4	28
$\text{MiniLM}_{MULTIQA}$	<u>61.7</u>	<u>38.2</u>	**55.1**	<u>66.6</u>	55.4	<u>60.5</u>	<u>98.8</u>	30
MiniLM_{L12}	**63.1**	**38.9**	<u>54.9</u>	**68.0**	54.9	**60.8**	**100.0**	31

- The number of parameters. Both DiffCSE and SimCSE denote $\text{RoBERTa}_{\text{base}}$ model

Method	DiffCSE	SimCSE	MPNet	MiniLM
# Param	250M	125M	110M	21.3M

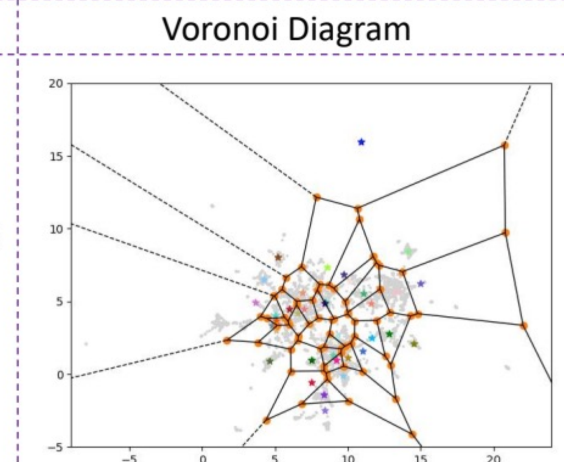
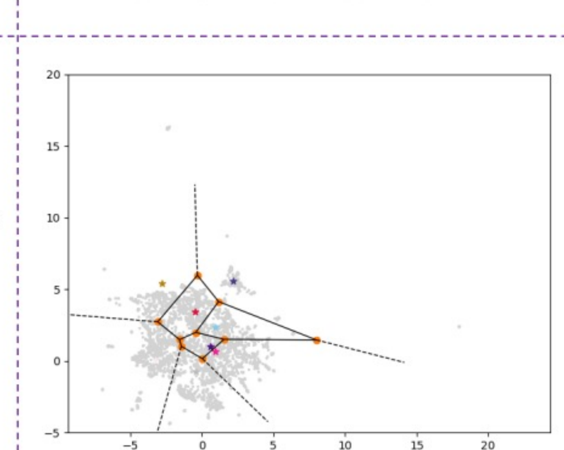
- These results demonstrate that (i) performance increases as the number of parameter decreases, which means excessively large embedding model leads to performance degradation, and (ii) the use of a student network which is trained by the teacher's self-attention distributions and guiding layer improves intent clustering performance

Task 2: Intent Induction

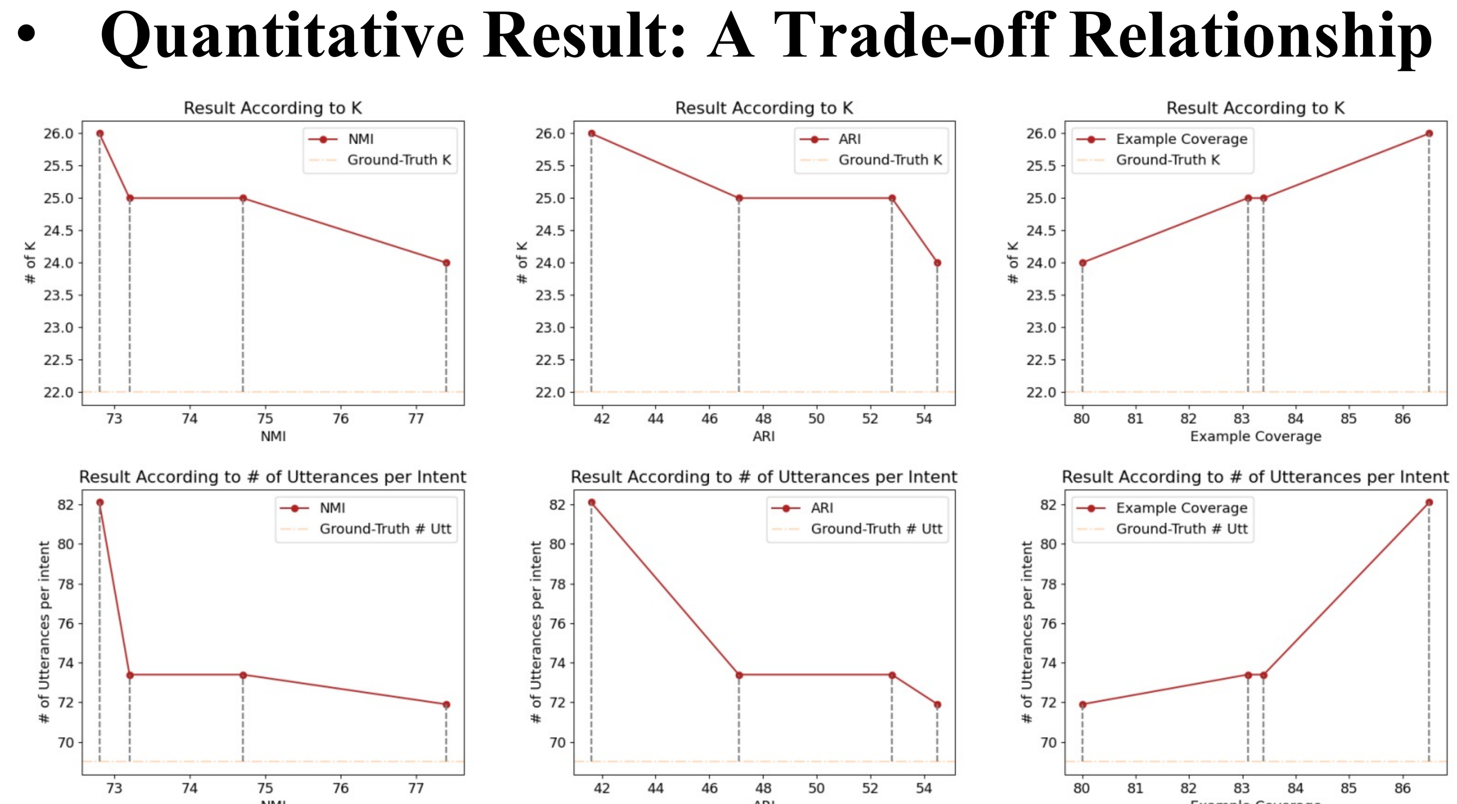
- **Result Analysis (Clustering Algorithm)**

Utterance Embedding	$\text{MiniLM}_{\text{MULTIQA}}$							
Metric	**NMI**	**ARI**	**ACC**	**Precision**	**Recall**	**F1**	**Example Coverage**	**#K**
Bisect K-means	72.3	55.7	57.4	63.0	<u>81.2</u>	70.9	76.8	27
VOT	75.3	50.6	60.0	71.9	72.3	72.1	**93.4**	33
Spectral	75.0	51.7	59.0	67.7	75.9	71.6	86.3	23
K-means	77.4	54.5	63.2	70.5	79.1	74.6	80.0	24
BIRCH	<u>79.5</u>	<u>62.8</u>	**68.1**	<u>71.9</u>	**84.6**	<u>77.7</u>	<u>89.9</u>	24
Agglomerative	**81.0**	**64.2**	<u>66.5</u>	**75.5**	80.6	**78.0**	86.2	25

- **Qualitative Result**
  - Agglomerative clustering on DiffCSE utterance embedding is not able to discern user intent well and amalgamate completely different user utterances into a single intent cluster.

Method	Voronoi Diagram	Examples	Label
Agglomerative + MiniLM		*Would it be on like any of my other bills?* *Oh okay. Where at on the bill. I'm looking now.* *Okay. When will I see that in my bill?* *Hi I am calling because I received a wrong bill.* *Thank you. How much will this make my bill go up?* *Yeah, so exactly how does this work? Do I just send a bill or what?* *crap! Well what can I do now? Can I make the payment and have the same policy that I had before?*	34
Agglomerative + DiffCSE		*Would it be on like any of my other bills?*	5
		Oh okay. Where at on the bill. I'm looking now.	2
		Okay. When will I see that in my bill?	7
		Hi I am calling because I received a wrong bill.	6
		Thank you. How much will this make my bill go up?	0
		Yeah, so exactly how does this work? Do I just send a bill or what?	1
		crap! Well what can I do now? Can I make the payment and have the same policy that I had before?	4

- **Quantitative Result: A Trade-off Relationship**



SIGdial & INLG 2023 workshop: The 11th Dialog System Technology Challenge (DSTC11)