University of Colorado Boulder

Who am I talking to? Topic-based Classification for Voice Applications

amazon alexa

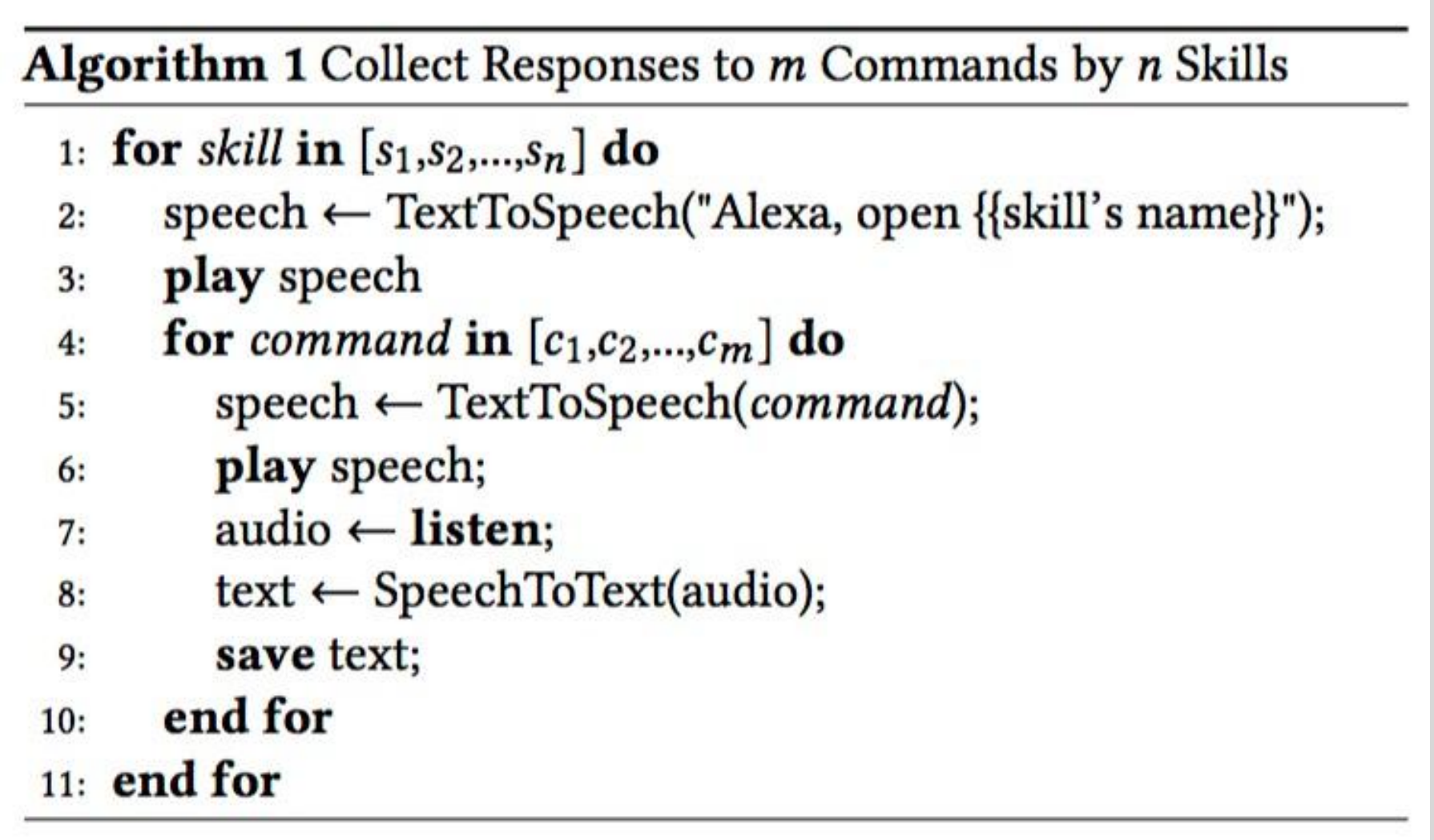
Xu Han
CU Boulder, Dept of Computer Science

Introduction

- Voice user interface (VUI) platforms allow the third party developers to freely build their own voice applications. However, how to evaluate these voice applications still needs exploring.
- Topic Analysis for these voice applications is crucial in evaluating them. [1]
- This project deploys deep learning methods (e.g. DAN, text CNN, skip-gram) to conduct topic analysis based on my self-collected Alexa skill dataset.
- This project uses a machine learning algorithm (LR) to generate baseline. Both DAN & text CNN have better performance (with fastText embedding) , as well as DAN with self-generated skip-gram embedding matrix.

Data Collection

•An voice application crawler [2] is used to collect responses data of 45,708 Alexa skills. The following algorithm clearly demonstrate how the crawler works.



```
Algorithm 1 Collect Responses to $m$ Commands by $n$ Skills
 1: for skill in [$s_1$,$s_2$,...,$s_n$] do
 2:     speech ← TextToSpeech("Alexa, open {{skill's name}}");
 3:     play speech
 4:     for command in [$c_1$,$c_2$,...,$c_m$] do
 5:         speech ← TextToSpeech(command);
 6:         play speech;
 7:         audio ← listen;
 8:         text ← SpeechToText(audio);
 9:         save text;
10:     end for
11: end for
```

Alg 1. Working Process of the Voice Crawler

•In order to generate the commands list (line 4 in Alg. 1), the package Stanford CoreNLP is used to parse the sentence of the instruction from each Alexa skill, stimulated by "Alexa, help" command, and do text to annotation analysis.

Dataset

- From the voice application crawler, I retrieved the Alexa Skills Responses Dataset. The details are displayed in the following table and Fig 1..

Number of Alexa skills	45708 *(15 topics labeled)*
Dialogue turns on average	7.3
Percentage of skills which cannot conduct multiple turns of dialogue	13.3%
Responses length on average	32.1 words (*with 1 word at least and 106 words at most*)

Table 1. Alexa Skills Responses Dataset

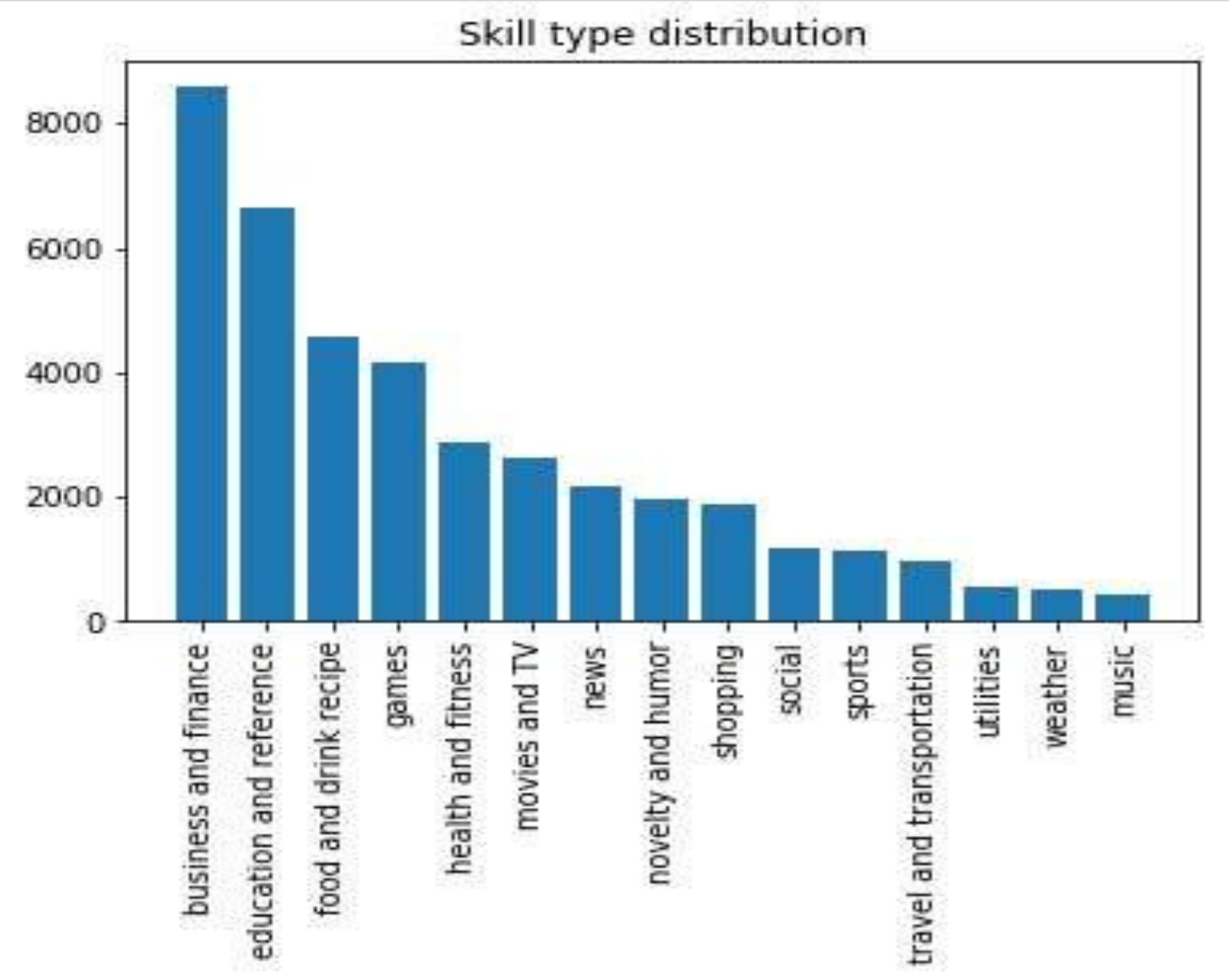


Fig 1. Alexa Skills Topic Distribution

Baseline Generation

- Used Logistic Regression (LR) to conduct topic classification (15 topics)
    - Used grid search to find the best parameters
- 6 feature sets
    - Bag of words
    - Tags frequency
    - Name-entity frequency
    - Sentence length
    - Sentence entropy
    - LDA

Deep Learning Methods

- Text CNN with fastText embedding
- Deep Average Network (DAN) with fastText embedding (Fig 2)

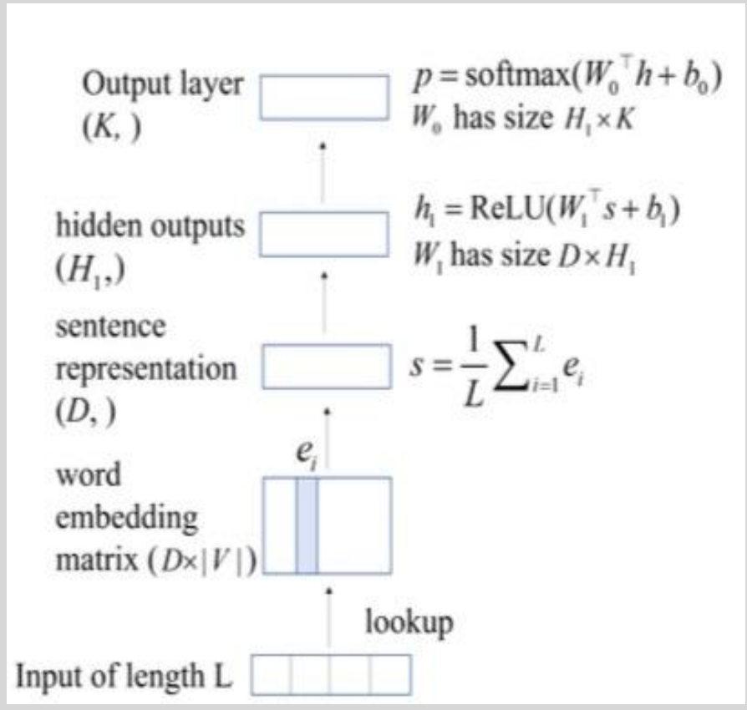

Output layer
(K,)
$p = \text{softmax}(W_0^\top h + b_0)$
W_0 has size $H_1 \times K$
hidden outputs
(H_1,)
$h_1 = \text{ReLU}(W_1^\top s + b_1)$
W_1 has size $D \times H_1$
sentence representation
(D,)
$s = \frac{1}{L}\sum_{i=1}^{L} e_i$
e_i
word embedding matrix ($D\times|V|$)
lookup
Input of length L

Fig 2. DAN Structure

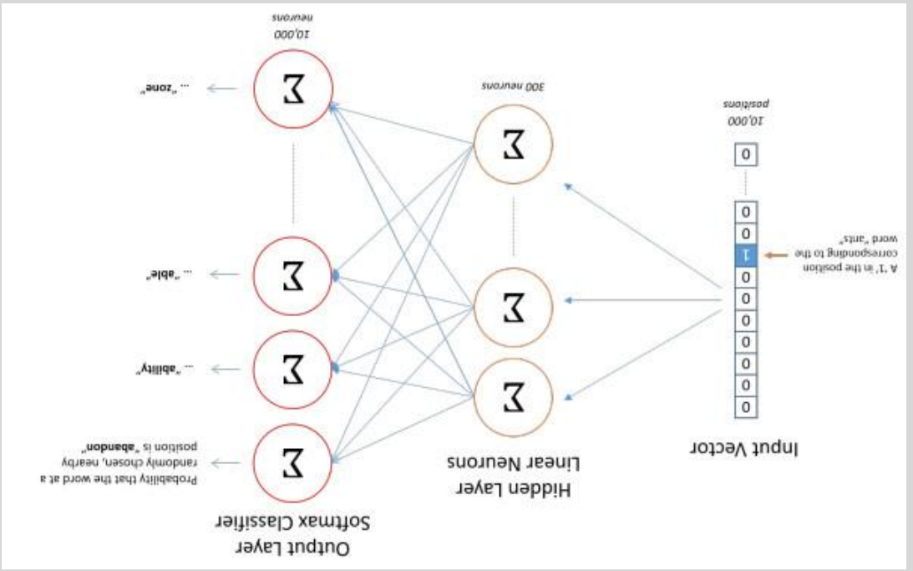


Fig 3. Skip-Gram Structure

- Skip-Gram to train our own embedding (Fig 3)

Results

- DAN training process (Fig 4)

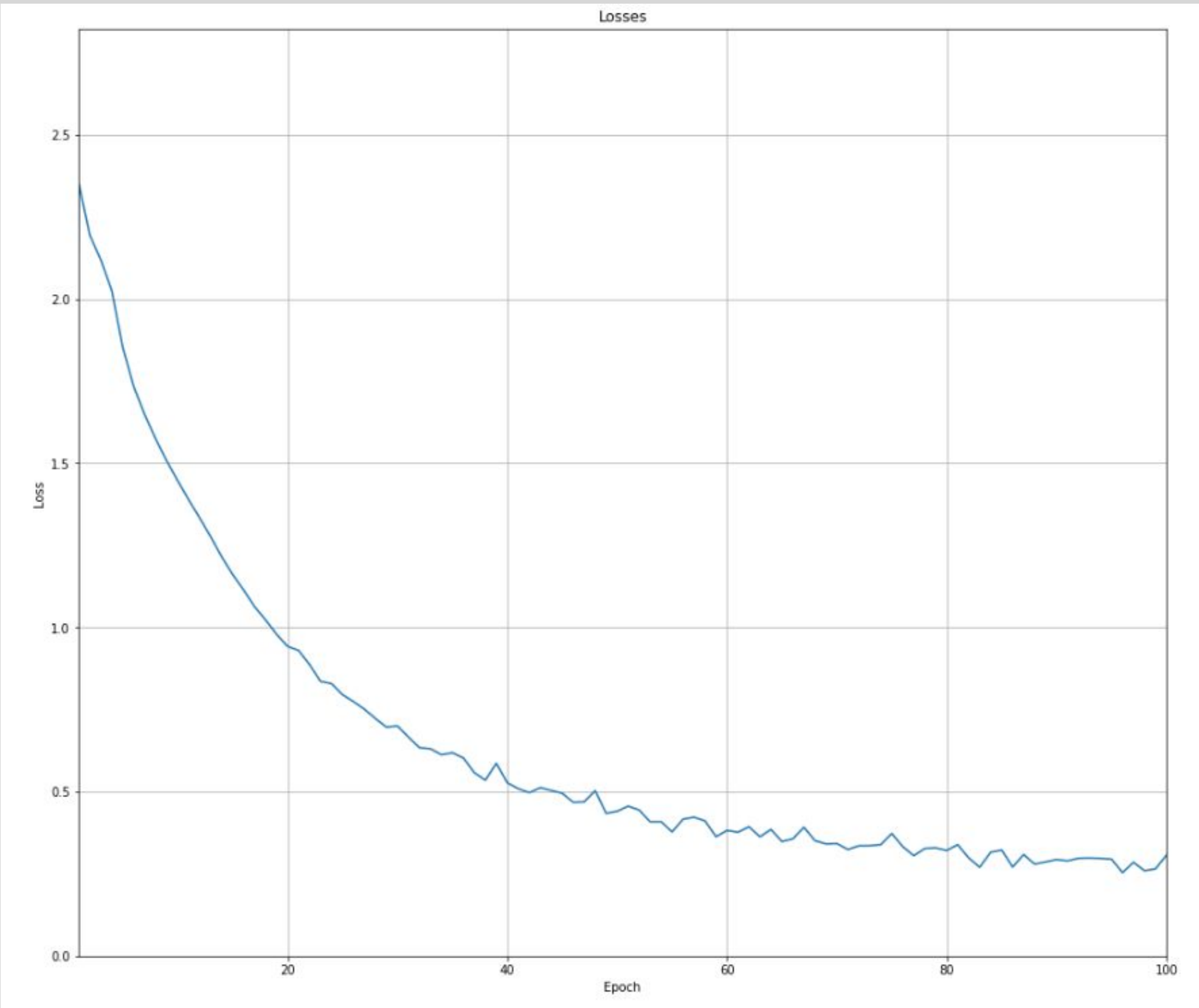


Fig 4. DAN Training loss

- Results of experiments (Table 2)

	Training time	Accuracy
LR	10 min	0.632
Text CNN with fastText embedding	20 min (50 epochs)	0.704
DAN with fastText embedding	30 min (100 epochs)	0.753
DAN with self-trained embedding	30 min (100 epochs)	0.691

Table 2. Training time and accuracy of the models

Analysis and Conclusion

- Deep Learning methods outperform the LR baseline in terms of classification accuracy.
- DAN with fastText embedding has the best performance. One biggest characteristic ofmy dataset is that there is too much noise (including the limitation of Alexa's error handling ability). In this case, DAN's robustness to noise can result in better performance, as well as its ability in memorizing keywords.
- Our DAN with self-trained skip-gram embedding matrix (Fig 5) also outperforms the baseline. This result indicates the potential of our datasets developing to a robust pre-trained word vectors.

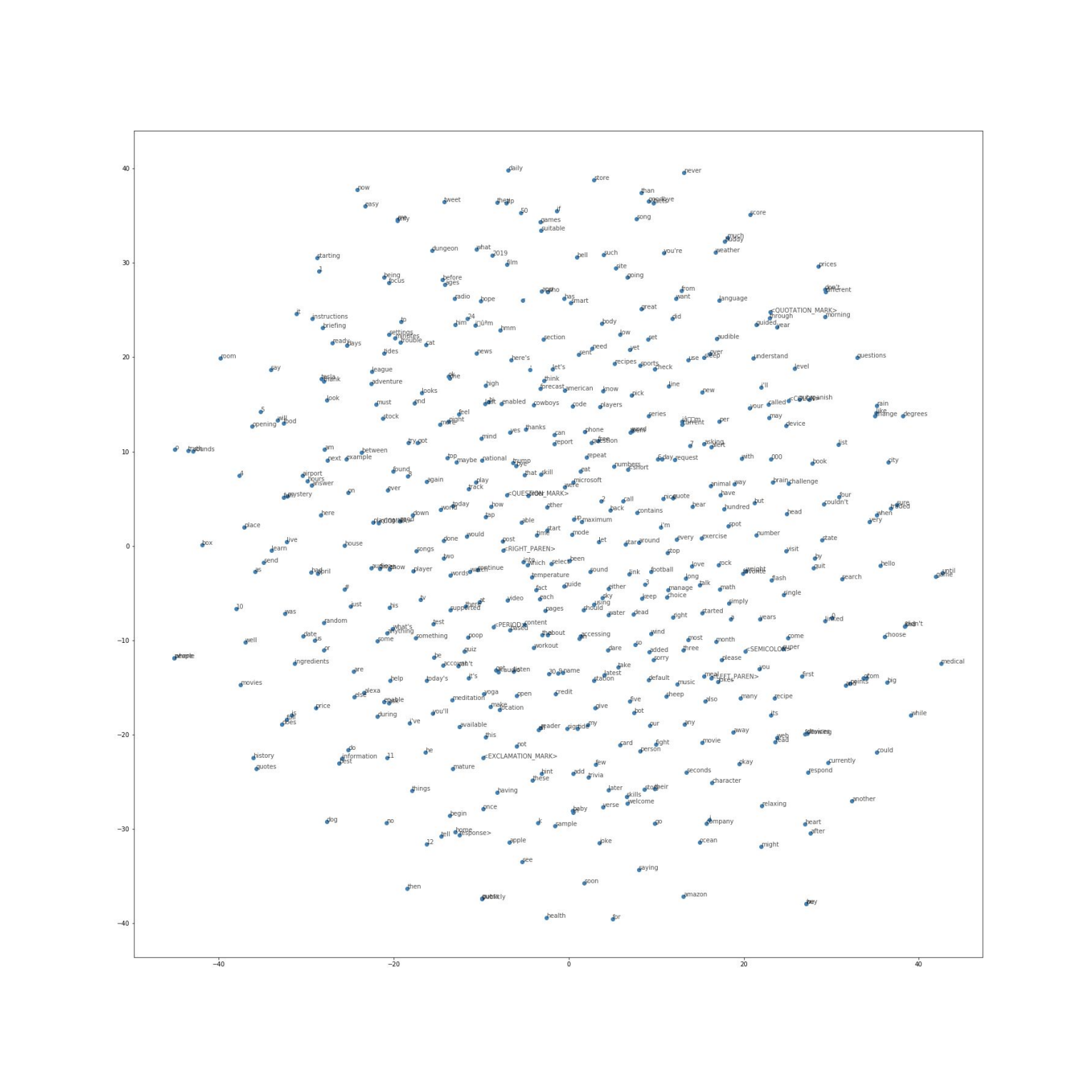


Fig 5. Clustering of self-trained word embeddings

Future Work

- The analysis above proves the feasibility of deploying deep learning methods to conduct voice application evaluation. Based on this, In the future work, a chatbot trained by deep learning methods using the Alexa responses dataset can be deployed to do evaluation.

References

[1] Guo, Fenfei, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. 2018. "Topic-Based Evaluation for Conversational Bots." *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1801.03622.
[2]Han, Xu, and Tom Yeh. 2019. "Evaluating Voice Applications by User-Aware Design Guidelines Using an Automatic Voice Crawler." http://ceur-ws.org/Vol-2327/IUI19WS-USER2AGENT-6.pdf.