Report

Anna Liednikova

January 9, 2019

Contents

6 Conclusion   9

7 References   9

1 Introduction

This project was set within the framework of a collaboration with the ALIAE startup where the aim is to develop a chatbot to collect information from clinical patients. The main idea is to replace strictly defined surveys by a more natural conversation in order to let users express themselves freely so that more information could be collected. First, we should be able to analyse the user utterances (does the user speaks about pain, physical activity etc. ?) and detect at least one most probable intent of user message so that chatbot could choose the right strategy for fulfilling information.

A small manually annotated dataset (roughly 100 sentences for each category) was available. So the purpose of the project is to find word embeddings that will allow to expand it by external data using the existing dataset with labelled intents.

We have tried models for sentence representation by word and the whole context in order to use semi-supervised learning by gradually increasing train data by labelling external sources.

2 Literature review

[FLWH18] faced a similar problem with a small labelled dataset. In their article they describe the approach of automatically expanding dataset and improving the performance of baseline models, that was event trigger identification system in their case.

First, they trained a baseline classifier on available data. Then they identified clusters of additional data to obtain grouped paraphrases inspired by the NewsSpike idea introduced in [ZHDCZ15]. After they labelled clusters with baseline model trained fully-supervised. Combining new labelled data and original one they retrained the event extractor [FLWH18].

3 Methodology

sentence representation, clustering , classification

3.1 Sentence representation models

3.1.1 By Word representation

The most popular word-basis techniques to represent sentences are LDA (Latent Dirichlet Allocation) and Word2Vector model. We used their im-

plementation from gensim library in order to build models based on the initial labelled dataset.

3.1.2 By Sentence representation

In order to use context of words more PositionwiseFeedForward and BiLSTM neural networks were implement on Pytorch.

3.2 Clustering models

The main idea behind using clustering models for evaluating sentence representations is to detect whatever sentences with the same intent populate the same clusters.

For these purposes following clustering models were used:

- K-Means (KM)

- AgglomerativeClustering(ward) (AG)

- GaussianMixture (GM)

The number of clusters was set to the number of intents (20) that allows using not only purity and Silhouette coefficients to evaluate the clusters but also homogeneity and completeness to get the idea how resulting groups correlate with initial intents. Also, for each model, we plotted a confusion matrix with a background gradient to visualize how well clusters separate different intents.

3.3 Classification models

SVC on labelled data
    test on unlabelled data

4 Datasets

4.1 Dataset with intents

The initial dataset is created manually covering 20 main possible users intents. Each sentence represents one intent. For, example,

```
{'text': 'After sleep, for 2−3 hours,
I am better and then start feeling tired again',
 'intent': 'sleep',
}
```
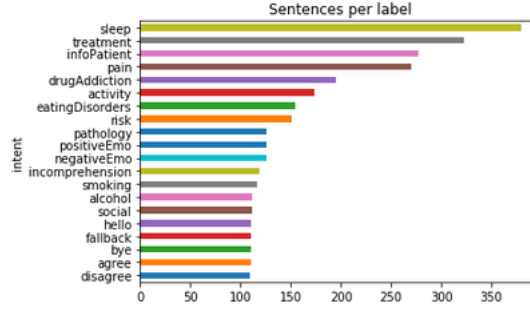
Distribution of labels is show on fig. 1

Figure 1: Label distribution in dataset
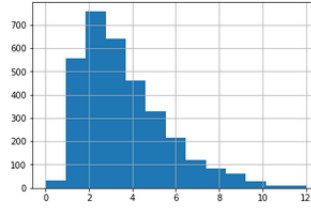
'min': 0, 'max': 12, 'mean': 3.438, 'std': 2.16367

Figure 2: Sentence dist

Total number of sentences is 3305. Found 1882 words after filtering through stopwords list.

4.2 Forum data

To improve initial classifier we needed to extend dataset by much more data and more naturally constracted one. No test set is available was available at the beginning of the project. To solve this problem, we opted to created our own test set following [ZHDCZ15] and [SGZH10] example.

Finally, 272552 unique posts

4.3 Preprocessing routine

Some proprocessing routine should be covered. This routine includes sentence tokenization, word tokenization, stop words removing and lemmatization.

Most popular and common tools for nlp are NLTK and SpaCy but they differ in behaviour.

For example, in word tokenization they give different results that can influence not only simple statistics but meaning too. Some example of different tokenization can be seen in table 1. Though concating words to ones like 'flulike' or '35mg' or 'longterm' sometimes gives more robust and concrete meaning in case of big dataset, in our case it seems better to stay with spacy way of tokenization in order to have smaller and more simple vocabualary.

For stopwords removing there were three options: nltk, spacy and the longest one. The last option was rejected due to containing words like 'want', 'stop', 'successfully' etc. that can be useful for detecting basic intents like positive or negative emotion, social. Finally nltk one was selecting because of containing shorts like 'm' from 'am', 've' from 'have'. Final dictionary contained 1882 words. Also all numbers were changed to num. Chart (3) looks fine.
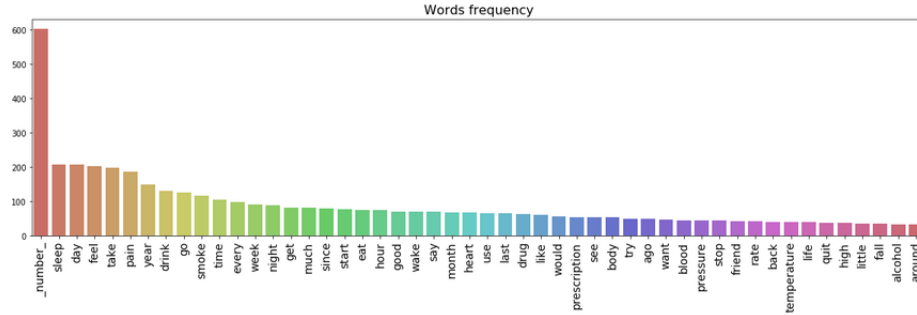

Figure 3: Words frequency

Length ¡ 30.

Next problem is empty sentences. But they don't change the dataset much.

intent fallback 12 disagree 11 hello 4 agree 3 incomprehension 2 positiveEmo 1

array(['what?', 'same that again', 'I am up', 'same', 'can you', 'do that', 'do this', 'Will do', 'no', 'no i will not', "No i don't", 'no', 'no i will not', "No i don't", "What's up?", "What's up?", 'he', 'a', 'd', 'i', 'm', 'o', 's', 't', 'y', 'I will', 'Will do', 'No', 'no', 'No it is not', "No I don't", 'No', "I'm here"], dtype=object)

NLTK	SpaCy
['i', 'wouldnt', 'go', 'to', 'sleep', 'until', 'like', '5', '6', 'or', '8am']	['i', 'would', 'nt', 'go', 'to', 'sleep', 'until', 'like', '5', '6', 'or', '8', 'am']
['that', 'is', 'totally', 'wrongheaded']	['that', 'is', 'totally', 'wrong', 'headed']
['i', 'am', 'in', 'the', 'process', 'of', 'tapering', 'from', 'suboxone', 'longterm', 'use']	['i', 'am', 'in', 'the', 'process', 'of', 'tapering', 'from', 'suboxone', 'long', 'term', 'use']
['i', 'had', 'an', 'onandoff', 'opiateopioid', 'habit', 'from', 'about', '2010']	['i', 'had', 'an', 'on', 'and', 'off', 'opiate', 'opioid', 'habit', 'from', 'about', '2010']
['i', 'have', 'flulike', 'pathologysymptom']	['i', 'have', 'flu', 'like', 'pathologysymptom']
['i', 'have', 'exerciseinduced', 'insomnia']	['i', 'have', 'exercise', 'induced', 'insomnia']
['i', 'm', 'supposed', 'to', 'take', '6', '35mg', 'tablets', 'a', 'day', 'but', 'i', 'have', 'taken', '20', 'today']	['i', 'm', 'supposed', 'to', 'take', '6', '35', 'mg', 'tablets', 'a', 'day', 'but', 'i', 'have', 'taken', '20', 'today']

Table 1: Tokenization comparision

HealthBoards is a medical forum web portal that allows patients to discuss their ailments. We scraped 272553 unique posts contained in each category. Finally, the corpus consists of N sentences. Table 4 shows the dataset statistics.

'min': 0, 'max': 2027, 'mean': 10.932588521491454, 'std': 8.964716092053479
'min': 3.2580388329566468, 'max': 35.23600634007455, 'mean': 11.224462266261876, 'std': 6.963225985041148

Data should be divided in subsets by increasing sentence length because of the difference in mean values for both datasets.

5   Experiment

5.1   Clustering

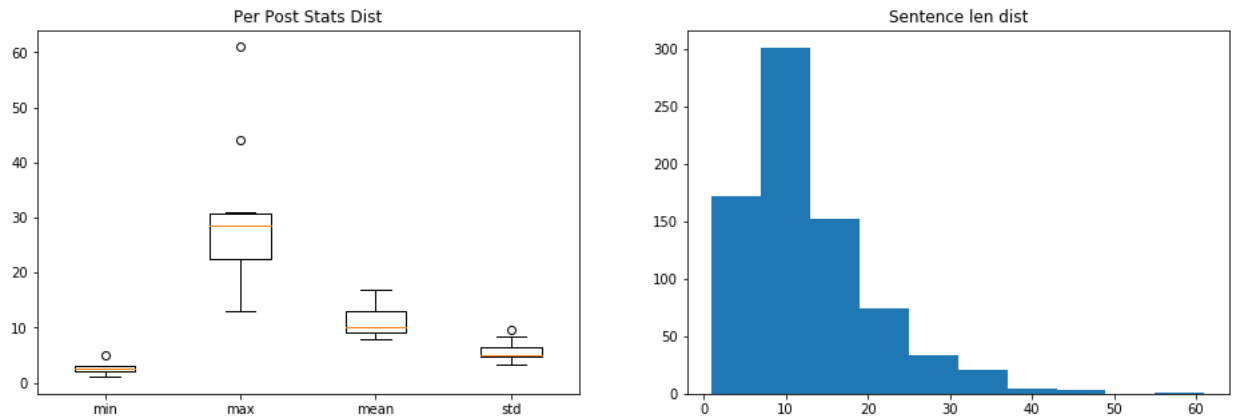### 5.1.1   Evoluation metrics

5.1.2   W2V model

Comparision table

Figure 4: Text stat

WE	Cluster	purity	Silhouette	homogeneity	complete	Clf CV score
Defenders	KM AC GM					
M	KM AC GM					
Forward	FW					
S	KM AC GM					

5.1.3 LDA model

5.1.4 GloVe pretrained

5.1.5 Google News W2V pretrained

5.1.6 CNN by word

From simple encoder, w2v, lda, w2v + lda

5.1.7 BiLSTM by word

From simple encoder, w2v, lda, w2v + lda

5.1.8 Overall comparision

5.2 Classifier

5.2.1 Evoluation metrics

5.2.2 Results for each model

6 Conclusion

7 References

References

[FLWH18]   James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh
Hajishirzi. Semi-supervised event extraction with paraphrase
clusters. In *Proceedings of the 2018 Conference of the North
American Chapter of the Association for Computational Lin-
guistics: Human Language Technologies, Volume 2 (Short Pa-
pers)*, pages 359–364. Association for Computational Linguis-
tics, 2018.

[SGZH10]   Parikshit Sondhi, Manish Gupta, ChengXiang Zhai, and Julia
Hockenmaier. Shallow information extraction from medical fo-
rum data. In *Coling 2010: Posters*, pages 1158–1166, Beijing,
China, August 2010. Coling 2010 Organizing Committee.

[ZHDCZ15] Thomas Zhang, Jason H D Cho, and Chengxiang Zhai. Under-
standing user intents in online health forums. *IEEE journal of
biomedical and health informatics*, 19, 03 2015.