Domain-independent Classification of Automatic Speech Recognition Texts

Lyubov Nesterenko
Evgenia Mescheryakova
1 June 2017

«Обращаем ваше внимание на то, что в целях улучшения качества обслуживания все разговоры записываются»

Intro

Some contact centers deal with about 10^5 calls daily.

Intro

Some contact centers deal with about 10^5 calls daily.

Monitoring all the calls would be time-consuming and tedious for supervisors

Intro

Some contact centers deal with about 10^5 calls daily.

Monitoring all the calls would be time-consuming and tedious for supervisors.

There is a need to make various analytical reports automatically.

Intro

Some contact centers deal with about 10^5 calls daily.

Monitoring all the calls would be time-consuming and tedious for supervisors.

There is a need to make various analytical reports automatically.

This includes classification/clustering by topic (= reason for a call).

Intro

We propose a simple though effective approach to domain independent automatic speech recognition (ASR) texts classification.

The use of clustering for semi--automatic training set annotation seems to be a solution to domain -independent classification.

Related work

[Agarwal et. al. 2007]: how ASR mistakes affect the supervised classification results

[Popova et al. 2014]: Russian ASR and manually transcribed texts clustering

[Wang, Wu, Shao 2014]: hierarchical clustering in a sliding window + clusters merging

- noise obviously affects ASR texts clustering/classification,
- training set annotation is costly,
- clustering requires knowing the number of clusters or makes us rely on optimization procedure results

Data

1370 ASR (badly transcribed) texts

topic	**documents**
luggage	653
booking	288
ticket return	257
flight status	74
flight info	98
Total	**1370**

Noisy texts

trancription	**corrected**
спасибо за ногти коня	спасибо за звонок всего доброго до свидания
давайте кот бронирования вам назову	давайте код бронирования вам назову
дрова зажигания	спасибо за ожидание
брони юношеского труда скажет	брони ? ? скажет
мне бы на попозже рябина	мне бы на попозже ?

Implementation: lite preprocessing

- lemmatization

Implementation: lite preprocessing

- lemmatization
- stop words removal

Implementation: lite preprocessing

- lemmatization
- stop words removal

domain-independent stop words list: standard Russian list + words typical for call-centers

- *спасибо, всего доброго, до свидания, говорите, скажите*

no customized stop words lists

Implementation: vectorization choice

Classifier, vectorization	F1-score
RFC, **tf*idf**	0.85
Logistic Regression, **tf*idf**	0.86
SVM, **tf*idf**	0.84

RFC with doc2vec ≈ 0.65
SVM, Logistic with doc2vec < 0.45

Implementation: clustering and clusters merging

- Fixed number of clusters: deliberately larger than expected

Implementation: clustering and clusters merging

- Fixed number of clusters: deliberately larger than expected
  K-means (k-means++ initialization, 15 clusters)
  Average cluster homogeneity = 0.66

Implementation: clustering and clusters merging

- Fixed number of clusters: deliberately larger than expected
  K-means (k-means++ initialization, 15 clusters)
  Average cluster homogeneity = 0.66
- Clusters are merged via their lexical similarity → nice-looking topic names

Implementation: clustering and clusters merging

- Fixed number of clusters: deliberately larger than expected
  K-means (k-means++ initialization, 15 clusters)
  Average cluster homogeneity = 0.66
- Clusters are merged via their lexical similarity → nice-looking topic names
- Clustering results {text: ClusterId} are a training set for a classifier

Implementation: clustering and clusters merging

- Fixed number of clusters: deliberately larger than expected
  K-means (k-means++ initialization, 15 clusters)
  Average cluster homogeneity = 0.66
- Clusters are merged via their lexical similarity → nice-looking topic names
- Clustering results {text: ClusterId} are a training set for a classifier
- New texts classification: preprocessing → vectorization → cluster id → topic name

Results: Logistic regression trained on clustering results

Тематика	Precision	Recall	F1-score	Support
luggage	0.96	0.90	0.93	125
booking	0.83	0.37	0.51	65
ticket return	0.48	0.90	0.63	52
flight status	0.58	0.69	0.63	16
flight information	0.73	0.50	0.59	16

Weighted Precision	Weighted Recall	Weighted F1
0.80	0.74	0.74

Conclusion

We observed the problem of domain-independent classification of automatic speech recognition texts and proposed a solution that allows to avoid fully manual annotation of the documents collection.

Our results show that using clustering techniques as an automatic training set annotation tool does not worsen the classification results greatly.

We regard the described pipeline as an acceptable solution for the case when one cannot afford manual annotation of a large training set.

Thank you!