Multi-label Topic Classification of Turkish Sentences Using Cascaded Approach for Dialog Management System

Gizem Soğancıoğlu, Bilge Köroğlu, and Onur Ağın

R&D and Special Projects Department
Yapı Kredi Technology, Istanbul, Turkey
{gizem.sogancioglu, bilge.koroglu, onur.agin}@yapikredi.com.tr

Abstract. In this paper, we propose a two-stage system which aims to classify utterances of customers into 10 categories including daily language and specific banking problems. Detecting the topic of customer question would enable the dialogue system to find the corresponding answer more effectively. In order to identify the topic of customer questions, we built a machine learning based model consisting of two stages which uses feature sets extracted from text and dialogue attributes. In the first stage, where utterances are categorized into two classes, namely *daily language* and *banking domain*, the binary classification problem was studied and different learning algorithms such as Naive Bayes, C4.5 have been evaluated with different feature sets. Utterances classified as *banking domain* by the first classifier, are classified by second stage classifier. In this stage, automatic detection of specialty of banking-related sentences is aimed. We approached this as a multi-label classification problem. Our proposed cascaded approach has shown a quite good performance with the score of 0.558 in terms of Micro-averaged F-score measure.

Key words: text classification, question answering, dialogue system, multi-label classification

1 Introduction

With the growing size of the knowledge in the web, the demand of reaching the needed answers for customers has significantly increased in last decades. However, searching for the answer via huge amounts of knowledge is time consuming. Consequently, customers usually prefer to use online customer services to ask solutions for their questions/problems instead of finding them from the web. On the other hand, the frequency of the usage of customer services for even general questions available in the web search results is painstaking work for the customer assistance. A chatbot which is a conversational agent that interacts with users via natural language is promising research field to overcome this problem. A dialog system provides person/month saving by decreasing the need of customer assistance and increases customer satisfaction by answering their questions fast.

Text based domain-independent dialog system consists of mainly 3 components[1]: Natural Language Understanding, Dialog Manager, Response Generator. Natural Language

A. Horák, K. Pala, P. Rychlý, A. Rambousek (Eds.): CBBLR 2016, pp. 77–85, 2016.

Understanding is one of the most challenging and crucial tasks for a good performing dialogue system[2]. The aim of this module is to automatically capture and tag the semantic properties from user sayings. One of the most common approaches as a solution of the task is detection of the user intent and question topic from given sentences.[3]. Topic classification task has been studied a lot so far as a crucial component of dialogue systems[4,5,6] as well as it has been used for various different purposes such as enhancing information retrieval performance[7]. In this paper, we investigated topic classification problem that would be utilized in a Turkish dialogue system for banking domain. We provided a cascaded approach based on machine learning model. In the first stage, customer sentences are classified into *daily language* or banking related sentences namely *banking domain*. This first stage task is binary classification problem which is studied a lot by researchers. In the second stage, the aim is to classify banking-domain specific sentences to 9 categories as money transfer, credit card, credit, flexible account, invoice payment, personal investment, web banking,personal account and other banking operations. Detailed definition of topic classes are defined in the Section 2.1. We approached this task as a multi-label classification problem where a sentence may be related to more than one class. For example; the customer may have a problem during invoice payment via web banking interface. This problem is related to both web/mobile banking and invoice payment classes. Therefore, multi-label classification method fits to our problem.

The language used in our data set, consisting of natural language conversations, differs from formally written text and people do not feel forced to write grammatically correct sentences, generally write like they talk or they type wrong characters by mistake. For improving the performance of the system, text normalization methods which normalize wrong and incomplete spelling are needed. In this study, we also applied a simple text normalization technique to overcome this problem.

The contribution of this paper is two-fold. First, we compared various multi-label and single-label classification algorithms in each stage. Second, we analyzed the significance of different features for topic classification problem. The rest of this paper is organized as follows: In Section 2, we define a taxonomy of the topic classes and give some brief explanation about the data set used in experiments. Moreover, we present our pre-processing algorithm used for text normalization and our cascaded machine learning approach in more detail. In Section 3, we describe the experimental setup and present the results for first-level and second-level models. In Section 4, we present summary of research findings and discuss future research pointers.

2 Methodology

2.1 Taxonomy

Detailed taxonomy definition is given in Table 1. *Daily language* and *banking domain* are classes aimed to classify in first stage of the system. Other defined 9 classes are classified by second stage classifier. Since this study was developed as a part of bigger question answering system, pre-defined classes were determined according to future system.

Table 1: Taxonomy of Topics

Intent Class	Definition
Daily Language	Domain-independent sentences that do not require informative answer. e.g.: 'Hello, Ok, Thanks, Are you there?'
Banking Related Topic	Contains all banking domain specific questions and sentences
Money Transfer	The act of transferring money from one account to another that contains various transferring types such as electronic funds transfer
Personal Investment	Financial investment by a person or business such as investment on government bonds
Credit Card Operations	Convenient substitute for cash e.g. what happens if i do not pay debt of my credit card on time
Credit Operations	An agreement between a buyer and a seller in which the buyer receives the good or service in advance and makes payment later
Personal Account	A bank account one uses for purposes other than business. One uses a personal account for regular expenses, such as rent.
Flexible Account	An account type that allows customers spend more money than they have
Invoice Payment	Invoice payments such as water bill, electric bill
Web/mobile Banking	A web interface of the bank to enable customers do some operations e.g. i could not pay invoices via mobile banking,it is not working
Other Banking Operations	Other banking operations that do not belong to any of the defined classes above

2.2 Data Set

Since there is no available public data set consisting of customer questions related to banking domain, we used web chat conversations consisting of dialogues between customer and customer assistance. It contains around a hundred twenty thousand dialogues and 1 million utterances only for two years, between 2013 and 2015. This data set was annotated by 5 different banking experts. 4500 utterances in total were manually annotated and only 4160 utterances which have more than 60% inter-annotator agreement were considered. 2/3 of these data is allocated for training, and remaining 1/3 used for testing.

Table 2: Data Set for First Stage Classification

Class	Training Set	Test Set
Daily Language	1650	550
Banking Domain	1470	490

Table 3: Training Data Set for Second Stage Classification

Money Transfer	231
Personal Investment	143
Credit Card	660
Credit	164
Personal Account	189
Flexible Account	117
Invoice Payment	141
Web Banking	183
Other Banking Operations	132

2.3 Preprocessing

Our empirical analysis showed that questions are needed to be normalized due to spelling errors. We have applied simple text normalization technique to convert incorrectly written words to well-formed. We manually constructed Turkish abbreviations dictionary consisting of hundred words. Moreover, we extracted lexicon consisting of unique words from the well-formed banking news. Stages applied for normalizing text are expressed as follows:

1. Non-alphanumeric characters are removed from the text.
2. If a text contains any abbreviation, algorithm converts it into well-formed version by using the Turkish abbreviations dictionary.
3. If a word in text does not occur in our lexicon, algorithm computes edit distance of this word with each of the word in lexicon and replaces the wrong written word with the most similar one in lexicon. Edit distance [8] is computed by counting minimum number of steps required to convert $word_1$ to $word_2$.

2.4 Topic Classification

Topic classification system consists of two stages as shown in Figure 1. Firstly, text is normalized using the algorithm mentioned in Section 2.3. Then, the normalized text is given to the feature extractor module of first-stage classifier. Features, which are bag-of-words (BOWs), length and history, are extracted in this module. By using the extracted features, topic class of a text is obtained. If the class of the text is detected as *banking domain* by first-stage, then the second stage is proceeded. In feature extractor module of the second stage classifier, Lexicon-based Classifier (LBC) is applied and the result of the system is used as feature as well as BOWs model. First and second stage classification methods are introduced in the Sections 2.4 and 2.4 in detail.

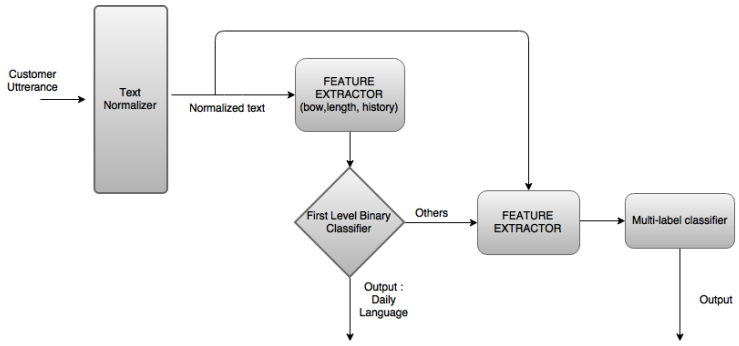

Fig. 1: System Design

First Stage Binary Classification

- **Baseline Model**: For baseline model of the first-stage classification, each utterance in test set was considered as the most frequent class in training set. We compared our method with this model.
- **Binary Classifier Using Dialogue and Utterance Attributes**:
  In the first stage, binary classification has been studied to detect *daily language* or *banking domain* topic class from a given sentence. C4.5, Naive Bayes, Support Vector Machine (SVM), Multilayer Perceptron (MLP) and Random Forest implemented in Weka[1] were evaluated.
  As features of the first stage binary classifier, BOWs, length and history were tried out respectively.
  `BOWs:` Text is represented as a set of its words in simple vector space.
  `Length:` As a length feature, count of the words in a sentence was considered.
  `History:` Topic of the previous customer utterance in dialog was used as history feature. If the corresponding utterance is the first sentence written by customer in dialogue, history feature is considered as EMPTY.

Second Stage Multi-label Classification In the second stage, the aim is to classify banking related sentences into 9 specific banking categories such as web/mobile banking, credit card. We approached this as a multi-label classification problem, where an instance may belong to more than one class. In multi-label classification, the examples (utterances) are associated with a set of labels.

Approaches tackling the multi-label classification can be grouped into two categories[9]: *problem transformation* and *algorithm adaptation*. Problem transformation methods see the problem as several single label classification problems, while algorithm adaptation methods adjust single label classifier to handle multi-label data. Since algorithm adaptation approaches mostly depend on single classifier, they lack generality.

Mulan[10], a java library for multi-label learning, implements various algorithms based on problem transformation and algorithm adaptation approaches. We have used Mulan library to evaluate learning models in our data set. As multi-label classifiers, RAndom k-labELsets(RAKEL) [11], MLkNN(multi-label KNN), Clustering-based [12], Hierarchical Multilabel classifier (HMC) [13] algorithms are evaluated on our data set. The following briefly describes each multi-label classification algorithm used in our experiments.

1. `Rakel`, one of the strongest approaches to multi-label classification problem, is based on problem transformation approach. It constructs an ensemble of Label Powerset (LP) classifiers. LP is a simple problem transformation method, which creates one binary classifier for every label combination. LP classifers are trained using a different random subset of the set of labels.
2. `MLkNN`, a multi-label lazy learning approach derived from traditional KNN(K Nearest Neighbor) algorithm, is based on *algorithm adaptation* approach. MLkNN

[1] https://sourceforge.net/projects/weka/

uses a Bayesian approach and utilizes the maximum a posterior principle to determine the label set of the test instance. Approach is based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors.

3. `Clustering-based` approach comprises a clustering algorithm and a multi-label classification algorithm. It consists of two steps: Firstly, it groups the training data into a user-specified number of clusters, k, using the clustering algorithm. As second step, it uses the multi-label algorithm on the data of each cluster and produces k multi-label classification models. For the classification of a new test instance, it first finds the closest cluster of this instance, and then uses the corresponding multi-label model to classify it.
4. `HMC`: HMC takes as parameter any kind of multi-label classifier and builds a hierarchy. Any node of hierarchy is a classifier and is trained separately. The root classifier is trained on all data and as getting down the hierarchy tree the data is adjusted properly to each node.

As features of the classifiers, the result of LBC and BOWs are used. 10-fold cross validation is applied and results are reported in Table 5.

– **Lexicon based classifier:**LBC performs multi-class classification which assigns one topic to the each utterance. For each topic classes, most frequent set of words are used as lexicon. As shown in Figure 2, n-grams (unigrams, bigrams,trigrams, four-grams and five-grams) are extracted from the given sentence. Then, the edit distances between n-grams and each of the lexicon belonging to different classes are computed. Result of the system was considered as a class having the minimum edit distance score with the given utterance. Pseudocode of the LBC is given in Algorithm 1.

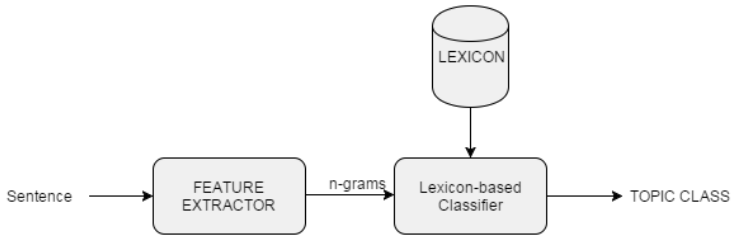

Fig. 2: Lexicon-based Classifier

3 Results

Experimental results of the each stage has been reported in terms of F-score. Since F-score is the harmonic mean of precision and recall, the following formulas express how

Algorithm 1 Pseudocode of Lexicon-based Classifier

```
distance ← 0
topic ← OTHER
distance_min ← 0
lexiconsList ← loadLexicons()
ngramsList ← featureExtractor(sentence)
while ngram in ngramsList do
   while lexicon in lexiconsList do
      distance ← editDistance(lexicon, ngram)
      if distance < distance_min then
         distance_min ← distance
         topic ← returnCorrespondingClass(lexicon)
      end if
   end while
end while
return topic
```

precision and recall values are computed for multi-label classification problem. (In the equations 1 and 2, P refers to predicted values while T denotes true labels.)

$$Precision = \frac{T \cap P}{P} \tag{1}$$

$$Recall = \frac{T \cap P}{T} \tag{2}$$

Table 4 shows the experimental results for the first stage of topic classification system. Using only the simple BOWs model as a feature of classifier performed fairly well in terms of the F-score measure. Moreover, experimental results showed that length is an indicative feature for detecting sentences from daily language. Adding history feature, the topic class of the previous sentence in dialogue, has also increased the overall performance. Among different classification algorithms, C4.5 has obtained the best performance. Baseline model mentioned in Section 2.4 has been computed as 0.55 F-score. Our best performing system has improved significantly with the two-tailed P value that is less than 0.0001. This difference is considered to be extremely statistically significant. Table 5 shows results for second-stage classification method. Using LBC

Table 4: Experimental results of first stage classifier with respect to feature sets

Features	**Random Forest**	**MultiLayer Percep-tron**	**Naive Bayes**	**SVM**	**C4.5**
BOWs	0.651	0.722	.0.624	0.662	0.737
+Length	0.782	0.824	0.75	0.771	0.842
+History	0.801	0.83	0.774	0.81	**0.865**

feature along with BOWs has increased the classification performance. Among different multi-label algorithms, Rakel has obtained the best performance with 0.558 Micro-averaged F-score and 0.354 Macro-averaged F-score. Although the wrong classification by first-stage classifier directly effects the performance of the second-stage classifier, it has performed quite well.

Table 5: Results for second stage classification

Learner	**Features**	**Micro-averaged F-score**	**Macro-averaged F-score**
ML-kNN	BOWs	0.211	0.104
ML-kNN	BOWs, LBC	0.319	0.176
HMC	BOWs	0.344	0.186
HMC	BOWs, LBC	0.372	0.192
Clustering Based	BOWs	0.514	0.382
Clustering Based	BOWs, LBC	0.513	0.382
RAKEL	BOWs	0.533	0.341
RAKEL	BOWs, LBC	**0.558**	**0.354**

4 Conclusion

In this paper, we have proposed a two-stage system for topic classification of customer questions in Turkish language. Since there is no suitable data set publicly available, we crafted manually our own data set consisting of around 4160 sentences in banking domain. Moreover, there was no similar previous study on utterance classification in banking domain. So, we could not provide a comparison between our system and previous systems. On the other hand, we compared different machine learning algorithms for both first stage and second stage classifiers and analyzed the effect of using each feature in this paper.

As first stage classifier, a number of machine learning algorithms including Random Forest, SVM, Naive Bayes, C4.5, Multilayer Perceptron have been evaluated. It is observed that length of the sentence is an important and distinctive feature to discriminate *daily language* from *banking domain* sentences. C4.5 has shown the best performance among other classifiers with the accuracy of %86.5. As second stage classifier, MLkNN, Rakel, HMC and clustering based algorithms have been evaluated. The experimental results showed that RAKEL performed better than others and LBC feature has increased the performance of the overall system. Although there is still open room to improve the classification performance, our cascaded approach is promising for topic classification problem.

This study is conducted as a part of the Turkish banking-domain dialogue system. In future work, we aim to develop a chat bot which answers questions only related to this

pre-defined classes and evaluate our classification approach for this question answering system.

References

1. Arora, S., Batra, K., Singh, S.: Dialogue system: A brief review. arXiv preprint arXiv:1306.4134 (2013)
2. Chen, L., Zhang, D., Mark, L.: Understanding user intent in community question answering. In: Proceedings of the 21st international conference companion on World Wide Web, ACM (2012) 823–828
3. Kiyota, Y., Kurohashi, S., Kido, F.: Dialog navigator: A question answering system based on large text knowledge base. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics (2002) 1–7
4. Lee, G.H., Lee, K.J.: A topic classification system based on clue expressions for person-related questions and passages. KIPS Transactions on Software and Data Engineering **4**(12) (2015) 577–584
5. Li, X., Roth, D.: Learning question classifiers. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics (2002) 1–7
6. Aytekin, Ç., Say, A., Akçok, E.: Eliza speaks turkish: a conversation program for an agglutinative language. In: Third Turkish Symp. Artificial Intelligence and Neural Networks, Ankara. (1994) 435
7. Chan, W., Yang, W., Tang, J., Du, J., Zhou, X., Wang, W.: Community question topic categorization via hierarchical kernelized classification. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, ACM (2013) 959–968
8. Klabunde, R.: Daniel jurafsky/james h. martin, speech and language processing. Zeitschrift für Sprachwissenschaft **21**(1) (2002) 134–135
9. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006)
10. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. The Journal of Machine Learning Research **12** (2011) 2411–2414
11. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. In: European Conference on Machine Learning, Springer (2007) 406–417
12. Nasierding, G., Tsoumakas, G., Kouzani, A.Z.: Clustering based multi-label classification for image annotation and retrieval. In: Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, IEEE (2009) 4514–4519
13. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. In: Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08). (2008)