Command based Task Oriented Hybrid System for the Music Field

Yuquan Le, Xian Li*, Suixue Wang, Peng Wang, Haiqian Lin, Guanyu Jiang
Guangzhou Shiyuan Electronic Technology co., LTD
{leyuquan@yeah.net}
{lixian, wangsuixue, wangpeng, linhaiqian, jiangguanyug}@cvte.com

ABSTRACT: *In the field of natural language applications, dialogue system plays a central role. In the music field, a set of challenges were fixed during the knowledge graph conclave. As the part of this exercise we considered the Intent Identification in music field and Slot Filling in field task. A Multifast text and conditional random field methods for the task is proposed in this work. While testing we found that the intent identification scores are significant and the accuracy rates are convincing. The overall results are more satisfactory which provide reliability of the proposed system.*

Keywords: Intent Identification, Slot Filling, Conditional Random Field, Multi-fastText

Received: 12 April 2020, Revised 28 July 2020, Accepted 9 August 2020

DOI: 10.6025/ed/2020/9/2/42-45

1. Introduction

In recent years, the dialogue system [1] plays an important role in natural language processing, and thus has received much attention. However, the current research is rarely involved in the music field. For this purpose, the 2018 China conference on knowledge graph and semantic computing (CCKS) challenge sets up a competition for the command understanding task oriented to the music field, which includes two parts (Music domain intent identification and slot filling). The goal of the music domain intent identification is to determine whether a certain utterance of the user expresses an intention, which belongs to music field. If the utterance expresses the related music field, the relevant parameters mentioned in the utterance need to be extracted. This task is called slot filling. In this paper, we take the text classification methods to deal with the first task and named entity recognition (NER) approaches to handle the second task. mentioned in the utterance need to be extracted. This task is called slot filling. In this paper, we take the text classification methods to deal with the first task and named entity recognition (NER) approaches to handle the second task.

Text classification is an important task in natural language processing with many applications [2]. The key problem in text classification is feature representation, which is commonly based on bag-of-words model. Several feature selection

approaches, which include frequency and term frequency-inverse document frequency (TF-IDF), are applied to select more features. Owing to the success of word embeddings [3], recent popular neural network methods [4] have applied on text classification, obtaining attractive performance. We propose Multi-fastText based on fastText [4] to handle music domain intent identification. The most existing approaches of NER are based on machine learning methods, which include Hidden Markov Model (HMM) [5], Support Vector Machine (SVM) and Conditional Random Field (CRF) [6]. In recent years, some neural network methods [7] have been successful in NER, and achieve competitive performance.

Although many methods have gained competitive performance in text classification and named entity recognition, respectively. There are lots of problems in applying the above methods to the current task. Since music domain intent identification and slot filling together require comprehensive factors, and at the same time, the spoken dialogue text in a specific scenario (music scenario in this paper) is irregular. In this study, we develop a hybrid system and participate in 2018 CCKS challenge. The proposed hybrid system is based on three mainly component methods (rule, CRF and Multi-fastText). To summarize, the main contributions of this paper are: (1) We develop a hybrid system, which attempts to comprehensively consider the performance of music domain intent identification and slot filling tasks. (2) For the spoken dialogue text generated by the particular music scene, we have explored some favorable rules (including some external dictionary resources). (3) We experiment on the CCKS-2018 task 2 datasets and the result proves the effectiveness of our system.

2. Background

2.1 FastText

FastText [4] is a library for the learning of word embedding and the text classification. The architecture is similar to the cbow model [8], where the middle word is replaced by a label. We use the softmax function to compute the probability distribution over the predefined classes.

2.2 CRF

Conditional Random Field (CRF) is a kind of discriminative undirected probabilistic graphical model. It is often used for labeling or parsing of sequential data. Particularly, it has been shown to be useful in POS tagging, shallow parsing [9] and named entity recognition [6].

We assume that the random variables sequence X and Y are of the same length, and use $x = x_1, x_2, ...x_n$ and $y = y_1, y_2, ...y_n$ for the generic input sequence and label sequence, respectively. A CRF on (X, Y) is specified by a vector f of local features and a corresponding weight vector λ. The CRF global feature vector is given by $F(y, x) = \Sigma_i f(y, x, i)$, where x is the input sequence, y is the label sequence and i ranges over the input positions. The conditional probability distribution defined by the CRF is then $p_\lambda(Y|X) = \frac{exp\ \lambda.F(Y, X)}{\Sigma_y exp\ \lambda.F(y, x)}$. For training example $\{(x_t, y_t)\}_{t=1}^N$, the goal is to maximize the log-likelihood $L_\lambda = \Sigma_t log p_\lambda(y_t|x_t)$.

3. Model

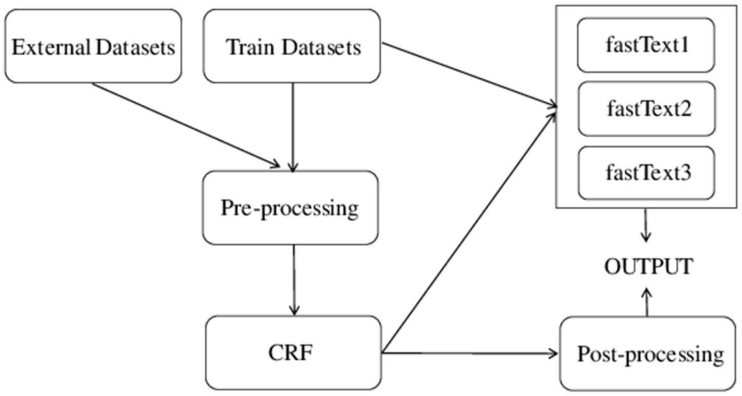

Figure 1. The architecture of system

Figure 1 shows the overview architecture of our system for command understanding task oriented to the music field. In order to increase the diversity of data for named entity recognition, we used some external data (NLPCC 2018 task 2 datasets). Pre-processing includes character segmentation and part of speech tagging (we treat a character as an independent individual for POS). Since the datasets come from the real discourse of the user in the dialogue system. Each sample is a segment of the user's utterance, including three sentences. Our goal is to determine if the last sentence is a musical intent and and slot filling. Through statistics, we found that there are a large number of datasets in which the last sentence is a short one. Therefore, we must make reasonable use of the previous information. We cut the datasets into three parts: each contains the first, the second, and the third sentence of each sample. We have designed a Multi-fastText method in order to better mine the user's multi-round dialogue information. Multi-fastText method works as follows: (1) If fastText3 determines that the third sentence is a musical intent, then the third sentence is classified as musical intent; (2) Otherwise, when fastText1, fastText2 judges that the first and second sentences are all musical intents, the third sentence is also classified as musical intent; (3) In other cases, the third sentence is no musical intent. Post-processing includes some rules, the details of rules are as follows:

Rule 1: Full Sentence Matching Rule (FSMR): (a) The entire sentence contains only one artist name, which is labeled as the artist label. Specifically, we crawled 28712 artist names from the Internet as an external artist dictionary resources. (b) The entire sentence contains only one song name and the song name is not in the ambiguous song dictionary, it is marked as a song label. Specifically, we crawled 172511 song names from the Internet as an external song dictionary resources. However, there are some ambiguous song names in the song dictionary, such as "点歌", "一首歌", etc. Therefore, we also maintain an external ambiguous song dictionary.

Rule 2: Entity Re-identify Rule (ERIR): Specifically, we consider two entities: "artist", "song". By using the song dictionary, if the model-based result is the substring of certain entities in the dictionary (and the entity is the substring of sentences), then correct the result to the one with the shortest length that meets the requirements. For example, assuming the model-based result is "路口", while the song dictionary include "下一个路口" and the utterance includes "下一个路口". The final result can be revised as "下一个路口" by entity re-identify rules. If some of the entities in the dictionary (entity is the substring of the sentence) are also substrings of the model-based result, the result is corrected to the one with the longest length that meets the requirements. Just as one example, assuming the model-based result is "刘德华冰雨", while the song dictionary includes "冰雨". The final result can be revised as "冰雨" by entity re-identify rules. Artist external dictionary resources also perform similar operations.

Method	$F1_E$	$F1_I$	*Acc*	*Score*
CRF+Multi-fastText	0.753	0.854	0.970	1.287
CRF+Multi-fastText+FSMR	0.771	0.863	0.970	1.304
CRF+Multi-fastText+FSMR+ERIR	**0.780**	**0.867**	**0.977**	**1.312**

Table 1. The experimental results for the test datasets

3. Experiments

3.1 Datasets and Implementation

We trained the 300 size pre-trained character embedding (as a pre-training embedding for the Multi-fastText) using the word2vec tool [1], and the training corpus used the entire wikipedia 2018 [2]. We migrated the NLPCC 2018 task 4 datasets [3] along with the CCKS datasets as a training set for the CRF (using crfsuite tools [4] in this paper) model, using the [5] tool for

[1] https://code.google.com/archive/p/word2vec/

[2] https://dumps.wikimedia.org/zhwiki/

[3] http://tcci.ccf.org.cn/conference/2018/taskdata.php#

[4] http://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html

[5] https://pypi.org/project/PyNLPIR/0.4.1/

part-of-speech tagging.

3.2 Experiment Result and Analysis

From the table 1, we can find that our system has achieved competitive performance. The best result is that $F1_E$ is 0.780, $F1_I$ is 0.867, *Acc* is 0.977, and the overall score is 1.312. Specifically, we find that the rules are effective. Without using rules, the bottleneck of $F1_E$ is 0.753, while CRF + rules achieves 0.780 $F1E$ score. However, adding ERIR based on FSMR has a very weak performance improvement (only 0.09).We speculate that there are several reasons for this : (1)

The resources used as external dictionaries are obtained through web crawlers, and some data may not be cleaned clearly, thus introducing new noise; (2) We only select two entities ("artist" and "song"), whose the total amount is huge, to join the rule action, but in fact, the task contains many types of entities.

4. Conclusion

In this paper, we proposed a hybrid system, which is named after CVTE SLU, for command understanding task oriented to the music field. Experiments on 2018 CCKS corpus prove the effectiveness of our system. In order to make the results better for future work, we will start to work on two aspects. First, maintaining external dictionary resources to make their quality more reliable. Second, applying the rule method proposed in the article to all entity categories.

References

[1] Levin, E., Pieraccini, R. (1997). *User modeling for spoken dialogue system evaluation*.

[2] Le, Y., Wang, Z.-J., Quan, Z., He, J., Yao, B. (2018). Acv-tree: A new method for sentence similarity modeling. *In*: *IJCAI*, p. 4137–4143.

[3] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C. (2003). A neural probabilistic language model, *Journal of machine learning research*, vol. 3, (February), p. 1137–1155.

[4] Joulin, A., Grave, E., Mikolov, P. B. T. (2017). Bag of tricks for efficient text classification, *EACL 2017*, p. 427.

[5] Morwal, S., Jahan, N., Chopra, D. (2012). Named entity recognition using hidden markov model (hmm), *IJNLC*, 1 (4) 15–23.

[6] Zhou, J., Dai, X., Yin, C., Chen, J.-J. (2006). Automatic recognition of chinese organization name based on cascaded conditional random fields, *Acta Electronica Sinica*, 34 (5), p. 804.

[7] Ma, X., Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf, *arXiv preprint arXiv:1603.01354*.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality, in *Advances in neural information processing systems*, p. 3111–3119.

[9] Sha, F., Pereira, F. (2003). Shallow parsing with conditional random fields, in *NAACL*, p. 134–141.