A Cross-Topic Method for Supervised Relevance Classification

Jiawei Yong
kai.yuu@jp.ricoh.com
Ricoh Company, Ltd.

Abstract

In relevance classification, we hope to judge whether some utterances expressed on a topic are relevant or not. A usual method is to train a specific classifier respectively for each topic. However, in that way, it easily causes an underfitting problem in supervised learning model, since annotated data can be insufficient for every single topic. In this paper, we explore the common features beyond different topics and propose our cross-topic relevance embedding aggregation methodology (CREAM) that can expand the range of training data and apply what has been learned from source topics to a target topic. In our experiment, we show that our proposal could capture common features within a small amount of annotated data and improve the performance of relevance classification compared with other baselines.

1 Introduction

Relevance classification is a task of automatically distinguishing relevant information for a specific topic (Kimura et al., 2019). It can be regarded as a preprocessing task of stance detection, since potential stances should be refined into relevant ones to improve accuracy and efficiency. In Table 1, we show a simple example of relevance classification task in NTCIR-14.

Topic: we should move Tsukiji Market to Toyosu.
Utterance1: I do not agree to move Tsukiji Market because of its long history. **Relevance**: relevant
Utterance2: The number of foreign tourists to Japan has been on the rise. **Relevance**: irrelevant

Table 1: An example of relevance classification.

Here utterance1 is relevant to the topic not only for the contained topic words but also for its related semantics, and then we could leverage its features available for further stance detection. On the contrary, utterance2 is irrelevant to the topic, and its further calculation of stance detection is meaningless. Previously, the relevance task could be approached in an unsupervised way by calculating pairwise semantic distances between topic and utterance (Achananuparp et al., 2008; Kusner et al., 2015). However, in most instances, their performance is not as good as a supervised approach. As to the supervised method, traditionally, a specific topic-oriented classifier could be trained for prediction on a single topic (Hasan and Ng, 2013; Y Wang et al., 2017), but this method actually builds up an isolation among different topics and wastes existing annotated data for new predictions.

Cross-topic classification, which enables the classifier to adapt different topics even in different domains, is an alternative to a supervised approach (Augenstein et al., 2016; Xu et al., 2018). It allows the model to assimilate the common features from existing topics and make inferences for a new topic. For example, in the NTCIR-14 relevance classification task, we could start with an existing classifier containing a well-prepared set of ground-truth data from some other Tsukiji Market history or economic topics, to give a prediction about Tsukiji Market relocation topic.

In this paper, aiming to alleviate insufficient annotated data problems for a specific topic, we have concentrated on cross-topic relevance classification by our novel CREAM proposal. The basic idea of the CREAM method is to capture the common pairwise features between existing topic and utterance, and then apply them to relevance prediction on a target topic. By analyzing F1-scores in experiment results, we have known that CREAM has shown its better performance on a known topic's relevance classification compared with baselines. In addition, an associated value to

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 147–152
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

the unknown topic relevance has also been evaluated.

2 Related Work

To establish a cross-topic relevance classification model for supervised learning, here we regard it as a two-step procedure including pairwise text embedding and binary text classifier. Besides, the literatures around stance detection bright us inspiration as well.

2.1 Text Embedding

There are 3 well-known embedding methods named Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and fastText (Joulin et al., 2016) for word-level representation. Although GloVe and fastText show higher performance on some specific aspects, there's no escaping the fact that Word2Vec (CBOW, Skip-Gram) is most popular and widely used among different languages.

As to sentence-level embeddings, the Word2Vec inventor Mikolov proposed doc2vec (Quoc et al., 2014), as its name implies, to learn sentence or document embeddings. What's more, averaged word embeddings (Han and Baldwin, 2016) is also a common sentence-level embedding method.

2.2 Text Classifier

There are several classical ML/DL models utilized for text classification such as Support Vector Machine (SVM) (Vapnik, 1998; Vapnik, 2013), and an RNN variant Long Short-term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). It is noteworthy that SVM has an advantage in processing low-resource data.

Besides, nowadays we also could utilize a pre-trained model such as BERT (Devlin et al., 2018) or ELMO (Matthew et al., 2018) as a contextual text classifier. However, note that they are always pre-trained by a tremendous amount of open data (E.X. Wikipedia), we still need fine-tuning data on a large scale for root domain recognition.

2.3 Stance Detection

Stance detection, which is the task of classifying the attitude expressed in text towards a target, also provides us with valuable inspiration on text classification. For example, Augenstein (Augenstein et al., 2016) tried to utilize conditional LSTM encoding to build a representation for stance and target independently, and an end-to-end memory network (Mohtarami et al., 2018), which integrates CNN with LSTM, has also been presented to solve this classification task. What's more, a simple but tough-to-beat baseline (Riedel et al., 2017) shows the potential of TF-IDF and cosine similarity on this pairwise classification task. Note that relevance classification can be regarded as a preprocessing of stance detection, since irrelevant stances should be excluded before being classified into support, against or even a neutral stance.

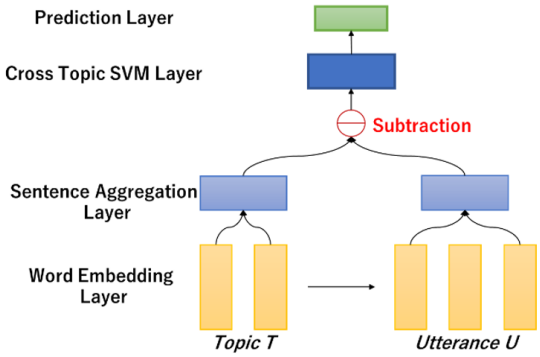

Figure 1: The overall architecture of CREAM.

3 Methodology

In this section, we would like to give a comprehensive introduction about our proposed cross-topic method CREAM, for supervised relevance classification. The overall architecture of CREAM is depicted in Figure 1. As described in the previous section, we briefly divide the whole model into 2 parts including text embedding and text classifier. In the text embedding part, we have implemented Word Embedding Layer and Sentence Aggregation Layer, and as to the text classifier, the SVM Layer and Prediction Layer would achieve their functions. The expected input includes a pair of topic text and topic-oriented utterance in the same domain, and the output would be predicted binary relevance label. In the following, we would illustrate the implementation details of each layer in CREAM.

3.1 Word Embedding Layer

Here we adopt pre-trained Word2Vec embeddings to represent each word of two inputs (a topic text T containing n words and a topic-oriented utterance U, e.g., topic and utterance1 in Table 1). Note that utterance could be much longer than topic text, so here we select the same number of words as topic T from utterance U. For each selected word of T,

we select one word with the highest Word2Vec similarity from U. The outputs of this layer are two sequences of word vectors with the same length $T = \{\vec{t}_1, ..., \vec{t}_n\}$ and $U = \{\vec{u}_1, ..., \vec{u}_n\}$.

3.2 Sentence Aggregation Layer

The sentence aggregation layer is the key to our cross-topic method CREAM. Here we manage to aggregate topic and utterance vectors by two steps to represent common features.

Separated Aggregation: In this step, we aim to provide a sentence-level embedding for T and U respectively. Here we separately aggregate $|T|$ word vectors for topic and utterance by averaged word embeddings:

$$\vec{T} = \frac{\sum_i^n \vec{t}_i}{n} \quad \vec{U} = \frac{\sum_i^n \vec{u}_i}{n} \tag{1}$$

Topic-Utterance Aggregation: Here we further concentrate on applying an aggregation between topic and utterance to represent the common features of relevance. As we have known there exists a classical conclusion from Word2Vec: $\vec{king} - \vec{man} + \vec{woman} = \vec{queen}$, we could get an inference that there exist some common features between word pairs (king, man) and (queen, woman) since $\vec{king} - \vec{man} = \vec{queen} - \vec{woman}$ is still workable.

As to sentence-level relevance classification, here we also conduct a vector subtraction between topic $\vec{T}$ and utterance $\vec{U}$ to represent relevance vector $\vec{R}$ as below.

$$\vec{R} = \frac{\vec{T} - \vec{U}}{|T|} \tag{2}$$

It is noteworthy that here we normalize each dimension value of relevance vector $\vec{R}$ by dividing $|T|$ to limit the subtraction result in the same range. Therefore, assuming that we have a relevance vector $\vec{R}_1$ (topic1) and $\vec{R}_2$ (topic2), they would be treated equally for the same cross-topic training if they all denote the same relevant relationship (e.g., $\vec{R}_1$ represents a utterance is relevant to topic1, $\vec{R}_2$ represents another utterance is relevant to topic2).

3.3 Cross-Topic SVM Layer

In this layer, we decide to adopt a supervised learning model SVM for cross-topic binary classification. The reason is because of low-resource data we have stated in chapter 2.2. In our case, SVM can efficiently perform a non-linear classification using kernel function (Mark et al., 1964) to fit the maximum-margin hyperplane in a transformed feature space. Here the following sigmoid kernel function for relevance vectors $\vec{R}_1$ and $\vec{R}_2$ makes SVM acted as multi-layer neural networks even they are different topics.

$$K(\vec{R}_1, \vec{R}_2) = \tanh(a\, \vec{R}_1^{\mathrm{T}}\, \vec{R}_2 - b) \tag{3}$$

After applying the kernel function, the target function of maximum-margin hyperplane could be written in:

$$y = sign(\sum_{i \in S} \alpha_i^* t_i K(\vec{R}_i, x) - h^*) \tag{4}$$

Here h^*, α^* are optimal parameters to distinguish binary hyperplane, and t is the correct class label for training.

3.4 Prediction Layer

We predict the relevance label of each topic-utterance pair via sigmoid-fitting method:

$$p_i = \frac{1}{1 + \exp(Ay_i + B)} \tag{5}$$

Where we apply the sigmoid operation to get the predicted probability for relevant and irrelevant classes with parameters A and B.

4 Experiments

In this section, we would introduce the evaluation results of our proposed methodology. We have evaluated our CREAM on the NTCIR-14 QALab dataset (Kimura et al., 2019). Note that NTCIR-14 QALab dataset maybe is the first dataset focusing on relevance classification besides fact-check and stance detection. Besides our own method, we have also taken three baseline approaches to cross-topic relevance classification.

Word Mover's Distance (WMD): this classical unsupervised learning method is often utilized to calculate a word travel cost between two documents. Here we predict the relevance label based on switch cost boundary from utterance to its topic.

Bidirectional LSTM (BiLSTM): this approach receives encoded-word sequences (topic and utterance) and makes a concatenation to merge them into one sequence. Finally, the concatenated vector would be fed into its prediction layer to give a relevance label prediction.

BERT: There is no doubt that BERT is the state-of-the-art model to solve NLP issues such as text classification. It is well-known that BERT could receive pairwise texts as inputs and output the label between them. Therefore, BERT is also applicable to this relevance classification theoretically. Here

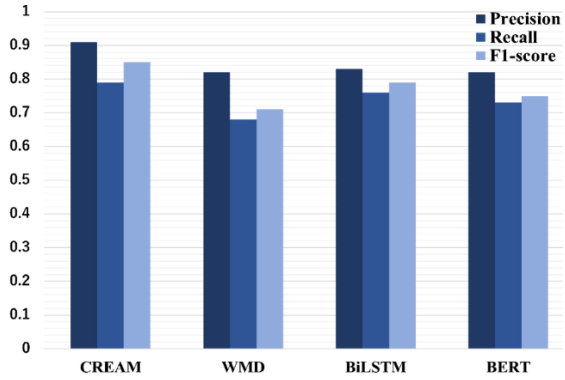

Figure 2: The averaged precision recall and F1-score of CREAM and baselines in experiment 1.

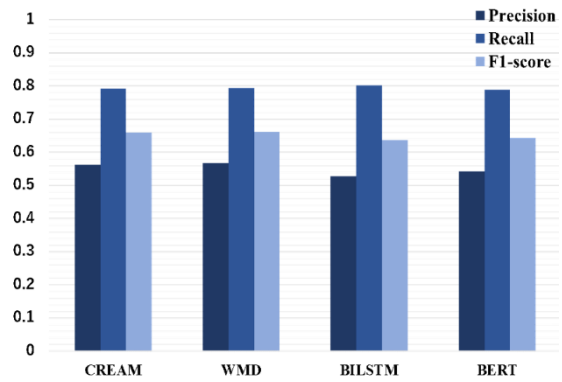

Figure 3: The averaged precision recall and F1-score of CREAM and baselines in experiment 2.

we beforehand input labelled topic and utterance separately into pre-trained BERT for fine-tuning.

4.1 Experiment dataset

NTCIR-14 QALab: This dataset is a Japanese collection for the relevance classification task, which contains around 10000 topic-oriented utterances on 14 different topics. Although task organizers do manual labeling by crowdsourcing, it is still difficult to provide an even larger amount of labeled dataset for each topic. Therefore, the traditional method with low-resource data would easily cause an underfitting problem.

4.2 Experiment Setup

Our initial word embeddings are obtained from the pre-trained Wikipedia word vectors (Suzuki et al., 2016).

In experiment 1, we divide our dataset into training data (1620) and test data (180) with the proportion 9:1. Note that there is no new topic in test data of experiment 1 since all topics have been included for training in the learning phase.

In experiment 2, we hope to verify the performance of our method compared with others on unknown topic relevance prediction. Therefore, we extract 13 topics' data for training to predict the last one topic in cross-validation.

4.3 Experiment Results

We mainly use F1-score to evaluate classification performance. Figure 2 illustrates the F1-score and averaged precision/recall as well among four methods in experiment 1, and the averaged evaluation results of cross-validation in experiment 2 have been summarized in Figure 3.

Furthermore, the relationship between the threshold of word mover's distance and F1 score is shown as an example in Figure 4. We just go through all the potential thresholds to find out the optimal one on the peak point to give a prediction for test data.

5 Discussion

As shown in Figure 2, we have known our CREAM has improved performance of relevance classification through experiment 1 since its F1-score is higher than others. The potential reasons of improvement are listed in the below.

- The sentence aggregation layer could extract common features between topic-utterance pairs and demonstrate the pairwise relevance degree by sentence aggregation processing.
- The cross-topic SVM layer shows high performance especially in low-resource data even compared with BiLSTM and BERT model. The BERT model pre-trained with open data perhaps is limited by the fine-tuning need for larger-scale data.

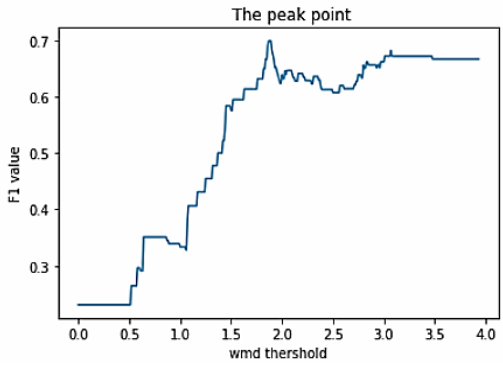

Figure 4: The relationship between the threshold of word mover's distance and F1 score.

As to the unknown topic's relevance prediction in experiment 2, the result of our method is close to the unsupervised WMD method which shows a

powerful predictive power to new data. We believe our CREAM method has an associated value on relevance prediction for unknown topics since the impact of a specific topic has been deducted by topic-utterance aggregation across different topics.

6 Conclusion and Future Work

In this paper, we have proposed a novel cross-topic aggregation model named CREAM to generalize the common features for solving low-resource data problems in relevance classification. Experiment results show its excellent performance on a known topic's relevance classification by F1-score over baselines. Meanwhile, we have also known that CREAM has an associated value to the unknown topic relevance prediction.

In the future, CREAM for relevance classification deserves more experiments with different datasets. For example, we could evaluate our methodology on multilingual datasets, in order to make it more impressive. Moreover, we could also input extern synonyms from the domain-based thesaurus to expand topic texts. Finally, self-attention mechanisms can be a promising improvement for imbalance length problems between topic and utterance instead of Word2Vec-style extraction.

Acknowledgments

Thanks to all relevant members of Ricoh System research center and NLU research group for their advices. Also, we would like to thank the anonymous reviewers and chairs for valuable feedback and discussions.

References

Isabelle Augenstein, Tim Rocktaschel, Andreas Vla- ¨ chos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 876–885.

Xu Chang et al. *Cross-Target Stance Classification with Self-Attention Networks*. arXiv preprint arXiv: 1805.06593 (2018).

Sepp Hochreiter and J¨urgen Schmidhuber. 1997. *Long Short-term Memory*. Neural computation, 9(8): pages 1735–1780.

Devlin Jacob et al. *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805 (2018).

Lau Jey Han and Baldwin Timothy. 2016. *An empirical evaluation of doc2vec with practical insights into document embedding generation*. Computing Research Repository, arXiv preprint arXiv:1607.05368.

Armand Joulin, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. *Bag of Tricks for Efficient Text Classification*. Computing Research Repository, arXiv:1607.01759.

Aizerman Mark A, Braverman, Emmanuel M, and Rozonoer, Lev I. 1964. *Theoretical foundations of the potential function method in pattern recognition learning*. Automation and Remote Control. 25: pages 821–837.

Peters Matthew E et al. *Deep contextualized word representations*. arXiv preprint arXiv:1802.05365 (2018).

Kusner Matt, Sun Yu, Kolkin Nicholas, and Weinberger Kilian. 2015. From word embeddings to document distances. In *Proceedings of International Conference on Machine Learning*, pages 957-966.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeff Dean. 2013. *Distributed Representations of Words and Phrases and Their Compositionality*. Advances in Neural Information Processing Systems, pages 3111-3119.

Mohtarami Mitra et al. *Automatic stance detection using end-to-end memory networks*. arXiv preprint arXiv:1804.07581 (2018).

Achananuparp Palakorn, Hu Xiaohua and Shen Xiajiong. 2008. The Evaluation of Sentence Similarity Measures. In *Proceedings of Data Warehousing and Knowledge Discovery*, pages 305-316.

Jeffrey Pennington, Richard Socher and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532-1543.

Le Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *proceedings of International conference on machine learning*, pages 1188-1196.

Benjamin Riedel et al. *A simple but tough-to-beat baseline for the Fake News Challenge stance detection task*. arXiv preprint arXiv:1707.03264 (2017).

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the 6th*

International Joint Conference on Natural Language Processing. pages 1348–1356.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki and Kentaro Inui. 2016. Neural Joint Learning for Classifying Wikipedia Articles into Fine-grained Named Entity Types. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pages 535-544.

Vapnik Vladimir. 1998. *Statistical learning theory*. Wiley, New York.

Vapnik Vladimir. 2013. *The Nature of Statistical Learning theory*. Springer Science & Business Media.

Yue Wang, Jidong Ge, Yemao Zhou, Yi Feng, Chuanyi Li, Zhongjin Li, Xiaoyu Zhou and Bin Luo. 2017. Topic Model Based Text Similarity Measure for Chinese Judgment Document. In *Proceedings of ICPCSEE*, vol 728.

Kimura Yasumoto et al. 2019. Overview of the NTCIR-14 QA Lab-PoliInfo task. In P*roceedings of the 14th NTCIR Conference*.