Expert Systems With Applications 60 (2016) 222–233

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework

Hyoungil Jeong[a], Youngjoong Ko[b,*], Jungyun Seo[a]

[a] Department of Computer Science and Engineering, Sogang University, 35, Baekbeom-ro, Mapo-gu, Seoul, 04107, Republic of Korea
[b] Department of Computer Engineering, Dong-A University, 37, Nakdong-daero 550beon-gil, Saha-gu, Busan, 49315, Republic of Korea

ARTICLE INFO

Article history:
Received 12 October 2015
Revised 30 April 2016
Accepted 1 May 2016
Available online 10 May 2016

Keywords:
Text summarization
Text classification
Binary independence model
Cluster-based language model
Support vector machine

ABSTRACT

Text summarization and classification are core techniques to analyze a huge amount of text data in the big data environment. Moreover, as the need to read texts on smart phones, tablets and television as well as personal computers continues to grow, text summarization and classification techniques become more important and both of them do essential processes for text analysis in many applications.

Traditional text summarization and classification techniques have individually been considered as different research fields in this literature. However, we find out that they can help each other as text summarization makes use of category information from text classification and text classification does summary information from text summarization. Therefore, we propose an effective integrated learning framework using both of summary and category information in this paper. In this framework, the feature-weighting method for text summarization utilizes a language model to combine feature distributions in each category and text, and one for text classification does the sentence importance scores estimated from the text summarization.

In the experiments, the performances of the integrated framework are better than ones of individual text summarization and classification. In addition, the framework has some advantages of easy implementation and language independence because it is based on only simple statistical approaches and POS tagger.

1. Introduction

Big data (White, 2010) is the term for a collection of data sets so large and complex that it becomes difficult to be processed by using on-hand database management tools or traditional data processing applications. The challenges for big data analysis are various such as capture, curation, storage, search, sharing, transfer, analysis, visualization, etc. In this environment, since the number of online texts has significantly increased, text summarization and classification techniques provide us efficient tools to easily recognize the content of a text without the waste of time. Text summarization is the task to reduce the length of a text in order to create a summary that retains the most important contents of the original text (Mani, 2001) and text classification is one to assign a text to one or more categories (Lewis, Schapire, Callan, & Papka, 1996). Whereas, these tasks have the same purpose to facilitate efficient management of information and reduction of handwork, they traditionally have been studied as different research fields.

In particular, as the need to read texts on a smart phone and a smart tablet continues to grow, the utilization of text summarization and classification is required more importantly (Radev, Hovy, & McKeown, 2000). Many web services such as news portals, blogs, question & answer sites, etc. provide many useful functions using summary and category information; predefined categories are commonly defined in these web services. Both of the summary and category information give effective ways for users to easily search information relevant to their information need and help them complete their task on mobile devices. Therefore, if a technique can be developed to improve both of text summarization and of the classification by mutual cooperation, it will be a really effective technique to these web services, especially, the mobile based services.

In this paper, we propose an integrated framework with four techniques to effectively combine text summarization and classification: (1) a statistical-extractive summarization technique using pseudo relevance feedback, (2) an enhanced text summarization technique with category information, (3) an enhanced text

[*] Corresponding address.
E-mail addresses: hijeong@sogang.ac.kr (H. Jeong), youngjoong.ko@gmail.com (Y. Ko), seojy@sogang.ac.kr (J. Seo).

http://dx.doi.org/10.1016/j.eswa.2016.05.001
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

classification technique with summary information and (4) a boosting technique for text summarization and classification on the integrated framework.

First of all, the statistical-extractive summarization technique can be augmented by a pseudo-relevance feedback (Salton, 1989) with a Binary Independence Model (BIM). The BIM and its accompanying query term weighting methods were developed for probabilistic text retrieval techniques (Yu & Salton, 1976 and Robertson & Sparck nes, 1976). The significant sentence extraction task for text summarization first divides a text into relevant and irrelevant sentences using title or the first sentence and then constructs a summary by BIM estimated from the relevant and irrelevant sentences. To enhance this summarization technique, we focus on the data sparseness problem observed in most of traditional statistical-extractive summarization approaches. In general, such approaches have used probabilistic information estimated in a text and its sentences. Unfortunately, most sentences have small number of features (words) and it is the main cause of data sparseness. A category-based language model is applied to estimate the importance of features to overcome the data sparseness. This is a feature probability estimation model from a category and a collection as well as a sentence and a text (Liu & Croft, 2004). For text classification, we attempt to improve the feature weighting method by reflecting sentence importance estimated from summarization to feature weighting. The feature weighting methods in traditional text classification have not considered the importance of each sentence and they always generate a same weight irrespective of the sentence in which the features appear. However, a sentence in a text has its own importance and the importance influences the importance of features that are included in the sentence (Ko, Park, & Seo, 2004). Thus, the text classification technique reflects the sentence importance from text summarization to the feature weight scheme. Consequently, an integrated framework was developed to mutually boost both summarization and classification and it will be effective because text summarization and classification have a reciprocal relationship.

Fig. 1 illustrates an example to explain whole processes of the proposed framework to integrate text summarization and categorization. The left side of Fig. 1 is an example of the results from the proposed framework and the right side is one from the general text summarization and categorization. The input text of the example can be confused between two categories, ‘Sports’ and ‘IT/Science’, to be predicted because its keywords include ‘tournament’, ‘beat’, ‘computer’, and ‘Go’. Since text classification uses more information from summary in the integrated frame, it drives the predicted category of the classification task to correct if summarization works well. In the case of this example, its predicted category is changed into the correct label, ‘IT/Science’, in the second iteration.

In our experiments, the proposed framework achieved 0.614 (+4.7%p) and 0.664 (+5.2%p) of F_1-measure in the summarization task on the KORDIC and AbleNews data sets, respectively, and 0.879 (+1.7%p), 0.784 (+2.4%p), and 0.890 (+3.8%p) of F_1-measure in the classification task on the Newsgroups, KORDIC and AbleNews data sets, respectively (%p; percentage point: the arithmetic difference of F_1-measure with baseline). The scores in parentheses denote the extent of improvement when the performances of the proposed method is compared to those of the baselines.

The remainder of this paper is organized as follows. Section 2 surveys related work. In Section 3, we propose a new statistical summarization method by using BIM and pseudo-relevance feedback. Section 4 presents an enhancement technique for text summarization by using category information based on the category-based language model, and Section 5 presents an enhancement technique for text classification using summary information. An integrated framework of text summarization and classification is shown in Section 6. In Section 7, the proposed approaches are experimentally evaluated in comparison with other text summarization and classification approaches. Section 8 provides a summary and discussion about our work. Finally, conclusion and future work are presented in Section 9.

2. Related work

Text summarization is the process of reducing a text in order to create a summary including only the core contents of an original text. Technologies that can make the coherent summary take into account many factors such as length, writing style and syntax. Generally, there are two approaches to text summarization: extraction and abstraction. An extractive approach works by selecting a subset of existing words, phrases, or sentences in an original text as a summary. In contrast, an abstractive approach builds an internal semantic representation and then uses natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words that are not explicitly in the original text. The state-of-the-art abstractive approaches are still quite weak, so most research has focused on the extractive ones.

A lot of techniques for summarizing texts have been studied by many researchers. Berger and Mittal (2000) proposed a web-page summarization system to generate coherent summaries that are not excerpts from the original text. Harabagiu and Lacatusu (2002) proposed an information extraction based multi-document summarization procedure that incrementally adds information. They have shown that it is possible to obtain good quality multi-document summaries using extraction templates populated by a high performance information extraction system. Matsuo and Ishizuka (2004) developed an algorithm that extracts keywords from a single text. Their algorithm can extract keywords without requiring the use of a corpus. They stated that their method has a higher performance than the term frequency-inverse document frequency (TF-IDF) and it is useful in many applications, especially for domain-independent keyword extraction. Svore, Vanderwende, and Burges (2007) proposed an automatic summarization method based on neural nets, called NetSum. They extracted a set of features from each sentence that help to identify its importance in a text and then their extracted features are based on news search query logs and Wikipedia entities using the RankNet learning algorithm. Ko and Seo (2008) proposed an effective method for extracting salient sentences using contextual information and statistical approaches for text summarization. They combined two consecutive sentences into a bi-gram pseudo sentence so that contextual information is applied to statistical sentence-extraction techniques. Li, Wang, Shen, and Li (2010) proposed a text summarization approach, named ontology enriched multi-document summarization, for utilizing background knowledge to improve summarization results. Their proposed system can better capture the semantic relevance between a query and the sentences, and leads to better summarization results on the domain-related ontology. Recently, some researchers studied that text summarization can be formulated by some well-known optimization problems such as knapsack problem. Filatova and Hatzivassiloglou (2004) proposed a model for simultaneously performing the selection of important text passages and the minimization of information overlap. Takamura and Okumura (2009) also demonstrated that summarization can be considered as an optimization problem that is maximum coverage problem with a knapsack constraint. Hirao, Yoshida, Nishino, Yasuda, and Nagata (2013) proposed a summarization method based on the trimming of a discourse tree. But there are no challenges in these summarization methods to make use of the category information.

Text classification is a popular research field with existing and ongoing scientific research. Text classification is the task to