論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Research on Context-based Document Topic Analysis
著者(和文)	呂有為
Author(English)	Youwei Lu
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10664号, 授与年月日:2017年9月20日, 学位の種別:課程博士, 審査員:新田 克己,樺島 祥介,渡邊 澄夫,小野 功,高村 大也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10664号, Conferred date:2017/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	知能システム科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学生氏名： Student's Name	呂有為	指導教員（主）： Academic Supervisor(main)	新田克己
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary （approx.800 English Words ）

In the thesis titled "Research on Context-based Document Topic Analysis", the weaknesses of LDA, which is one of the most typical topic models, in terms of modeling context information in texts are studied. Correspondingly, a few novel models are proposed to overcomes these problems in different aspects.

First, regarding the fact that topics inferred by LDA are unpredictable due to its unsupervised nature, the author developed the semi-supervised latent Dirichlet allocation (ssLDA), so that text labels can be incorporated for learning and in turn for prediction of labels in new texts. Different from some existing supervised topic models, ssLDA is semi-supervised in the sense that it can utilize labeled and unlabeled texts when training a model. It has a flexible training scheme allowing both word-level and text-level topic assignments, making it applicable for different forms of human labels we may encounter in different application scenarios. The ssLDA directly associates topics with labels by 1-1 correspondences. Building such associations not only helps us interpret the topics learned with unsupervised topic models like LDA, but also can predict the labels for unlabeled texts, namely multi-label classification. The author performed experiments under different topic assignment schemes using different data sets. The experiment results demonstrate that ssLDA can achieve very competitive results compared with other methods such as transductive support vector machines and the nearest neighbor algorithm.

Second, for LDA's limited capability of representing document structures in terms of latent topics, the author proposed the RUP-HDP-HSMM, which is a nonparametric Bayesian model built upon the HDP-HSMM, for modeling the content structure of domain-specific texts. RUP-HDP-HSMM incorporates RUP to tackle the rapid switching, the problem of generating redundant hidden states rapidly switching between one another. Different types of experiment are performed to illustrate applications of RUP-HDP-HSMM. In the text retrieval experiments, RUP-HDP-HSMM could properly model the content structures of domain-specific texts, in the light of the experiment results on our data set. In particular, RUP-HDP-HSMM successfully avoided the rapid switching between hidden states, while typical HDP-HSMMs did not. Besides, RUP-HDP-HSMM greatly outperformed the others in terms of ranking performance, while using less local features than HDP-HSMM. The author also demonstrated that modeling content structures of texts could provide useful features for text classification tasks. In our experiments, the content structures discovered with RUP-HDP-HSMM could lead SVM to achieve higher F1 scores than other methods. Furthermore, de- pending on RUP-HDP-HSMM, a method of extracting the relevant local contexts accounting for why a text is classified as a positive instance is presented. This method is efficient at finding the relevant local contexts, only relying on the statistical information underlying the current data set; therefore, it is considered a promising tool for discourse analysis, automatic summarization, etc., A specific application of our method remains as future work, however.

Third, as an extension of RUP-HDP-HSMM, WPM-HDP-HSMM by combining the WPM with HDP-HSMM the Weibull partition model (WPM) is developed. WPM defines a nonparametric stochastic process over distributions of partitions of sequential data. Compared to RUP, it can generate partitions according to more complicated segment duration distributions. In our synthetic experiments, WPM-HDP- HSMM has achieved very competitive predictive performance compared with strong baselines, indicating that the WPM can be used as a good prior for generating partitions of sequences. However, it has some challenges that need further consideration, including the selection of key positions, which plays an important role in WPM, and more efficient sampling methods for long sequences.

Last, the future work of this thesis is also given. Apart from content-structure extraction, the author also wants to find the logical structures--logical relations between the topics--of texts by making use of their content structures. In our daily conversations and arguments, people convey their opinions by claiming a series of statements in a logical way. Apparently, there are logical relations between these statements, such as, "support", "attach", etc. Recognizing these statements and furthermore their

relations constitutes the goal of finding the logical structures of texts. Clearly, this task is very challenging. However, at least we can model the content structures of texts by making use of RUP- HDP-HSMM; besides, combining with some NLP techniques, the author thought they might achieve this goal. Discovering the logical structures of texts is important in making computers think logically, because modeling logical structures of texts leads to a more concrete form of knowledge discovery. The long-term research plan comprises the development of automatic logical argument dialog systems. What is important here is "logical". Currently, there are some dialog and conversational systems, which are based on deep learning approaches. However, logical conversational systems cannot completely rely on deep learning no matter how big the learning data can be. The reason is simple, a trained neural network cannot automatically adopt itself with different arguments to create a logically coherent argument strategy, by learning from a large amount of logically inconsistent data. However, deep learning approaches (CNN, RNN, etc.) are promising in terms of pattern recognition and generating data. Therefore, we can combine deep learning with the knowledge of logical relations. The former is mainly responsible for generating responses, and the latter can compute and organize logical arguments, where the theory of argumentation frameworks may be useful.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。

Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).