Information Processing and Management 49 (2013) 330–340

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Automatically building templates for entity summary construction

Peng Li [a], Yinglin Wang [a,*], Jing Jiang [b]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[b] School of Information Systems, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 6 February 2011
Received in revised form 7 January 2012
Accepted 21 March 2012
Available online 31 May 2012

Keywords:
Summary template
LDA
Pattern mining

ABSTRACT

In this paper, we propose a novel approach to automatic generation of summary templates from given collections of summary articles. We first develop an entity-aspect LDA model to simultaneously cluster both sentences and words into aspects. We then apply frequent subtree pattern mining on the dependency parse trees of the clustered and labeled sentences to discover sentence patterns that well represent the aspects. Finally, we use the generated templates to construct summaries for new entities. Key features of our method include automatic grouping of semantically related sentence patterns and automatic identification of template slots that need to be filled in. Also, we implement a new sentence compression algorithm which use dependency tree instead of parser tree. We apply our method on five Wikipedia entity categories and compare our method with three baseline methods. Both quantitative evaluation based on human judgment and qualitative comparison demonstrate the effectiveness and advantages of our method.

1. Introduction

In this paper, we study the task of automatically generating templates for entity summaries. An entity summary is a short document that gives the most important facts about an entity. In Wikipedia, for instance, most articles have an introduction section that summarizes the subject entity before the table of contents and other elaborate sections. These introduction sections are examples of entity summaries we consider. Summaries of entities from the same category usually share some common structure. For example, biographies of physicists usually contain facts about the nationality, educational background, affiliation and major contributions of the physicist, whereas introductions of companies usually list information such as the industry, founder and headquarter of the company. Our goal is to automatically construct a summary template that outlines the most prominent types of facts for an entity category, given a collection of entity summaries from this category.

Such kind of summary templates can be very useful in many applications. First of all, they can uncover the underlying structures of summary articles and help better organize the information units, much in the same way as infoboxes do in Wikipedia. In fact, automatic template generation provides a solution to induction of infobox structures, which are still highly incomplete in Wikipedia Wu and Weld (2007). A template can also serve as a starting point for human editors to create new summary articles. Furthermore, with summary templates, we can potentially apply information retrieval and extraction techniques to construct summaries for new entities automatically on the fly, improving the user experience for search engine and question answering systems.

Despite its usefulness, the problem has not been well studied. The most relevant work is by Filatova, Hatzivassiloglou, and McKeown (2006) on automatic creation of domain templates, where the definition of a domain is similar to our notion of an entity category. Filatova et al. (2006) first identify the important verbs for a domain using corpus statistics, and then find

* Corresponding author.
E-mail addresses: jerryli1981@gmail.com (P. Li), ylwang@sjtu.edu.cn (Y. Wang), jingjiang@smu.edu.sg (J. Jiang).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.03.006

frequent parse tree patterns from sentences containing these verbs to construct a domain template. There are two major limitations of their approach. First, the focus on verbs restricts the template patterns that can be found. Second, redundant or related patterns using different verbs to express the same or similar facts cannot be grouped together. For example, "*won X award*" and "*received X prize*" are considered two different patterns by this approach. We propose a method that can overcome these two limitations. Automatic template generation is also related to a number of other problems that have been studied before, including unsupervised IE pattern discovery Sudo, Sekine, and Grishman (2003), Shinyama and Sekine (2006), Sekine (2006), Yan, Okazaki, Matsuo, Yang, and Ishizuka (2009) and automatic generation of Wikipedia articles Sauper and Barzilay (2009). We discuss the differences of our work from existing related work in Section 7.

In this paper we propose a novel approach to the task of automatically generating entity summary templates. We first develop an entity-aspect model that extends standard LDA to identify clusters of words that can represent different aspects of facts prominent in a given summary collection (Section 3). For example, the words "received," "award," "won" and "Nobel" may be clustered together from biographies of physicists to represent one aspect, even though they may appear in different sentences from different biographies. Simultaneously, the entity-aspect model separates words in each sentence into background words, document words and aspect words, and sentences likely about the same aspect are naturally clustered together. After this aspect identification step, we mine frequent subtree patterns from the dependency parse trees of the clustered sentences (Section 4). Different from previous work, we leverage the word labels assigned by the entity-aspect model to prune the patterns and to locate template slots to be filled in.

We evaluate our method on five entity categories using Wikipedia articles (Section 5). Because the task is new and thus there is no standard evaluation criteria, we conduct both quantitative evaluation using our own human judgment and qualitative comparison. Our evaluation shows that our method can obtain better sentence patterns in terms of f1 measure compared with two baseline methods, and it can also achieve reasonably good quality of aspect clusters in terms of purity. Compared with standard LDA and K-means sentence clustering, the aspects identified by our method are also more meaningful. Finally, we build an application which leverage our generated templates to construct entity summarization system. The results further prove that our automatically generated entity templates are useful (Section 6).

2. The task

Given a collection of entity summaries from the same entity category, our task is to automatically construct a summary template that outlines the most important information one should include in a summary for this entity category. For example, given a collection of biographies of physicists, ideally the summary template should indicate that important facts about a physicist include his/her educational background, affiliation, major contributions, awards received, etc.

However, it is not clear what is the best representation of such templates. Should a template comprise a list of subtopic labels (e.g. "education" and "affiliation") or a set of explicit questions? Here we define a template format based on the usage of the templates as well as our observations from Wikipedia entity summaries. First, since we expect that the templates can be used by human editors for creating new summaries, we use sentence patterns that are human readable as basic units of the templates. For example, we may have a sentence pattern "*ENT* graduated from? University" for the entity category "physicist," where *ENT* is a placeholder for the entity that the summary is about, and '?' is a slot to be filled in. Second, we observe that information about entities of the same category can be grouped into subtopics. For example, the sentences "Bohr is a Nobel laureate" and "Einstein received the Nobel Prize" are paraphrases of the same type of facts, while the sentences "Taub earned his doctorate at Princeton University" and "he graduated from MIT" are slightly different but both describe the person's educational background. Therefore, it makes sense to group sentence patterns based on the subtopics they pertain to. Here we call these subtopics the *aspects* of a summary template.

Formally, we define a summary template to be a set of sentence patterns grouped into aspects. Each sentence pattern has a placeholder for the entity to be summarized and possibly one or more template slots to be filled in. Table 1 shows some sentence patterns our method has generated for the "physicist" category.

Table 1
Examples of some good template patterns and their aspects generated by our method.

Aspect	Pattern
1	*ENT* received his phd from? university *ENT* studied? under? *ENT* earned his? in physics from university of?
2	*ENT* was awarded the medal in? *ENT* won the? award *ENT* received the nobel prize in physics in?
3	*ENT* was? director *ENT* was the head of? *ENT* worked for?
4	*ENT* made contributions to? *ENT* is best known for work on? *ENT* is noted for?