A Semantic Annotation Tool to Extract Instances from Korean Web Documents

Hai-tao Zheng, Bo-Yeong Kang, Sang-Ok Koo, Hee-Chul Choi, Kwang-Sub Kim, Hong-Gee Kim[1]
Biomedical Knowledge Engineering Laboratory,Seoul National University,
Yeongeon-dong, Jongro-gu, Seoul, Korea, 110-749
{quickly,comeng99,ironyjk,kss,hgkim}@snu.ac.kr, sangokkoo@gmail.com

ABSTRACT

Although there has been extensive research on developing semantic annotation tools recently, only few systems support automatic information extraction. In this paper, we propose a semantic annotation system named SARM, which has an automatic instance extraction module based on two machine learning techniques, Bayesian Classifier and Support Vector Machine. SARM has been tested to make a Korean Restaurant ontology evolve by automatically extracting instances from Web documents in Korean. The automatic instance extraction module can accelerate the annotation work which is very time-consuming and involves a lot of human labor. We describe the implementation of our system and also compare the performances of the two machine learning methods we used.

Categories and Subject Descriptors

H.4.3 [Communications Applications], I.2.6 [Learning], I.2.7 [Natural Language Processing]

General Terms

Performance, Design, Experimentation

Keywords

SARM, Bayesian Classifier, SVM, Information Extraction, Korean Restaurant Ontology

1. INTRODUCTION

This paper proposes a semantic annotation system named SARM, which has the automatic instance extraction module based on two machine learning techniques, Bayesian Classifier and Support Vector Machine. SARM has been tested to make a Korean Restaurant ontology evolve by automatically extracting instances from Web documents in Korean.

Recently, there has been extensive research to develop ontology based annotation tools that facilitate annotation of web document items in manual or automatic ways. KIM Semantic Annotation Platform[2], for example, provides a Knowledge and Information Management (KIM) infrastructure and services for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured contents. Another tool, MnM[3] is an annotation tool which provides both automated and semi-automated support for annotating web pages with semantic contents as well. Compared to KIM and MnM, our system is more suitable for domain and language specific annotation task, thus has been tested for the task domain of searching restaurant information in Korean.

In this paper, we firstly propose the system architecture of SARM. In the next step, we elaborate the learning methods for instance extraction and mechanism of the instance extraction for the Korean restaurant ontology, which can accelerate the annotation work which is time-consuming. Thirdly, the experiment result of SARM will be described.

2. INSTANCE EXTRACTION FOR KOREAN RESTAURANT ONTOLOGY

In this section, we explain the proposed semantic annotation system (SARM), which has the automatic instance extraction module based on two machine learning techniques. SARM has been tested to make a Korean Restaurant ontology evolve by automatically extracting instances from Web documents in Korean

2.1 Overall Architecture

In the Fig 1, we describe the whole architecture of SARM. SARM consists of automatic instance extraction module (Web Document Crawler, Web RawDB, HTML Parser, Morphology Analyzer, Bayesian Classifier and SVM Classifier), Semantic Annotator (with the API for remote access, embedding, and integration), Domain Ontology , Annotated Results (the semantic annotated web contents, RDF or OWL statements) and front-ends (the user interface with web browser control and knowledge explore for Ontology navigation).

Firstly, the crawler extracts the domain specific web documents as input for the Bayesian classifier and SVM Classifier. After preprocessing of HTML parsing and morphology analysis on the crawled web documents (Web RawDB), the classifier learns the features of the restaurant instances in the preprocessed web documents. Thus, given the new web documents, the classifier can extract the restaurant domain instances based on the training data. The user interface with web browser control[1] supports the functionalities such as ontology import/export, manual annotation editing by user, annotation browsing with instance extraction etc.

[1] Corresponding author. Tel: +82-20-740-8796

Email:hgkim@snu.ac.kr(Hong-Gee Kim)

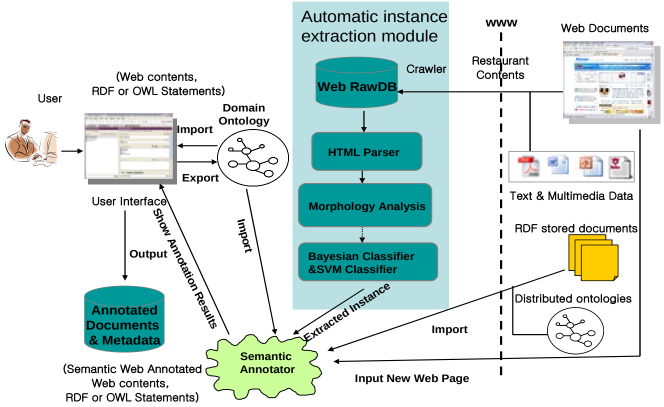

Fig. 1. The Architecture of the proposed semantic annotation

2.2 Restaurant Domain Instance Extraction Based on Compound Word Learning

Most restaurant names in Korean are composed of compound words. For example, a restaurant name, 강-변-식당(rive-nearby-restaurant) is composed of three words: 강(river), 변(nearby), and 식당(restaurant). Another example is 서울-집(Seoul-house) that is composed of two words, 서울(Seoul) and 집(house). We also found that most Korean restaurant instances are composed of some combination of concepts such as house, location and dish. Therefore, restaurant instances can be recognized successfully by applying a machine learning technique on the known combination of concepts for restaurant names. To annotate compound words as a restaurant instance, we used an ontology that contains concepts of restaurant, dish, beverage, and food stuff. The restaurant class instantiates a set of instances which represent the restaurant name, and the dish class instantiates a set of instances which represent the dish name.

Based on the observation of data, the multi-word decomposition was processed for each word in training data. Then we make a word vector $V = (w_1, w_2, ..., w_n)$ for representing the name of restaurant instances. Here, w_i is a single noun that is decomposed by a Korean noun dictionary. We can apply naïve Bayesian classification using *MAP(Maximum a posteriori)* decision rule as following equation.

$$classify(V_k) = \arg\max_{c} p(C=c)\prod_{i=1}^{n} p(w_i \mid C=c) \qquad \textbf{(3)}$$

We can also apply the word vectors V , into SVM classifier to find the OHP that best separates a set of training examples as following equation. Here, the OHP can be achieved by minimizing the objective function O_l. Then V_k ($\in R^N$) is the k-th input vector and $y_k \in \{+1,-1\}$ is the corresponding label for V_k in a two-class classification problem. W denotes the perpendicular vector to the OHP.(cf. Equation 4)

$$O_l = \frac{1}{2} w \cdot w, \qquad \textbf{(4)}$$

$$\textit{subject} \quad \textit{to} \quad y_k(w \cdot V_k + b) - 1 \geq 0, \quad k = 1,..., \ m$$

3. EXPERIMENT RESULTS

The proposed method was applied to the 467 web pages crawled[6] from the JoyFood.Com[2] of restaurant domain. Before restaurant instances were extracted for learning data construction, a series of preprocessing steps had to be done: morphological analysis[4] and html parsing[5] of the 467 web pages. Then we extracted a set of 1,260 restaurant instances that were divided into two sets: one for training the classifier and the other for actual validation. The performance of the proposed method was evaluated by accuracy micro average after conducting 5-fold cross validation on the 1,260 extracted restaurant instances. The train and the test sets used for the 5-fold cross validation consist of 1,008 restaurant instances and 252 restaurant instances, respectively.

In Bayesian Classifier, the performance was 98% in accuracy when the train set was used for both learning and testing. And the performance decreased to 94% when the test set was used after learning on the train set. In SVM, when the train set was used for both learning and testing, the performance was 96% in accuracy whereas it decreased to 92% when the test set was used after learning on the train set.

4. CONCLUSION AND FUTURE WORK

In this paper, we provide a system of semantic annotation with instance extraction for Korean restaurant ontology. Although there are some semantic annotations tools, there is few semantic annotation with automatic instance extraction that is suitable for domain and language specific annotation task. Our annotation tool, SARM is expected to help the user to make the annotation more effectively with respect to the restaurant domain in Korean language.

5. ACKNOWLEDGMENTS

This research is supported by Ministry of Information and Communication Republic of KOREA - National Project (Project management of Institute for Information Technology Advancement).

6. REFERENCES

[1] Microsoft WebBrowser control. http://msdn.microsoft.com/workshop/browser /webbrowser/browser_control_ovw_entry.asp

[2] Borislav Popov, A.K., Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, Miroslav Goranov. Kim-Semantic Annotation Platform. *2nd International Semantic Web Conference* (ISWC2003), Vol. 2870. Springer, Verlag Berlin Heidelberg (2003) 834-849

[3] Maria Vargas-Vera, Enrico Motta, John Domingue, Mattia Lanzoni, Arthur Stutt and Fabio Ciravegna :MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup,*The 13th International Conference on Knowledge Engineering and Management (EKAW 2002),* ed Gomez-Perez, A., Springer Verlag, 2002

[4] HAM. http://nlp.kookmin.ac.kr/HAM/kor/index.html

[5] HTML parser. http://htmlparser.sourceforge.net/

[6] WIRE.http://www.cwr.cl/projects/WIRE/index.htm

[2] Joyfood.Com. http://www.joyfood.com