Sentence Topics Based Knowledge Acquisition for Question Answering

Hyo-Jung OH*, Student Member and Bo-Hyun YUN†a), Nonmember

SUMMARY This paper presents a knowledge acquisition method using sentence topics for question answering. We define templates for information extraction by the Korean concept network semi-automatically. Moreover, we propose the two-phase information extraction model by the hybrid machine learning such as maximum entropy and conditional random fields. In our experiments, we examined the role of sentence topics in the template-filling task for information extraction. Our experimental result shows the improvement of 18% in F-score and 434% in training speed over the plain CRF-based method for the extraction task. In addition, our result shows the improvement of 8% in F-score for the subsequent QA task.

key words: knowledge acquisition, machine learning, question answering

1. Introduction

In the Web, there exists a lot of information which can be constructed into knowledge base (KB) by the information extraction (IE) technique. Similarly, in the question answering (QA), knowledge-based QA (KBQA) is efficient in finding some fast and precise answers to questions of users [1]–[3]. Therefore, it is important to construct the knowledge base automatically, not manually, in the QA.

In constructing KB by information extraction, it is critical to define templates for information extraction and to extract the precise information through machine learning. For that purpose, we define templates for information extraction by the Korean concept network semi-automatically. Moreover, we propose the two-phase information extraction model by the hybrid machine learning such as maximum entropy and conditional random fields.

In the following sections, we: (1) present some brief characteristics of machine learning methods used in our method; (2) define templates and template attributes for information extraction by using Korean concept network; (3) construct KB by using template filling; (4) show the experimental results for each phase; (5) describe the conclusion and future works.

2. Machine Learning Algorithms

For knowledge acquisition, we combined two different machine learning algorithms: maximum entropy and conditional random field. This section explains the overview of each algorithm and reveals differences between them.

2.1 Maximum Entropy

Maximum Entropy (ME, [4], [5]) finds the most uniform distribution model given a set of constraints. In other words, it estimates the conditional probability that, given a context x, the random process will output y. A mathematical measure of the uniformity of the conditional distribution p(y|x) is expressed as the conditional entropy:

\[ H(p) = - \sum_{x,y} p(x)p(y|x) \log p(y|x) \]  (1)

The goal of ME is to select a model \(p^*\) from a set C of allowed probability distribution which maximizes entropy

\[ p^* = \arg \max \ H(p) \]  (2)

It has been shown that the ME solution \(p^*\) is unique and must be in the following form [6]:

\[ p^* = \frac{1}{Z(x)} \exp \left( \sum_{i} \lambda_i f_i(x,y) \right) \]  (3)

where each \(f_i(x,y)\) is a feature, \(\lambda_i\) is a parameter to be estimated and \(Z(x)\) is simply the normalizing factor to ensure a proper probability:

\[ Z(x) = \sum_{c} \exp \left( \sum_{i} \lambda_i f_i(x,y) \right) \]  (4)

When the constraints are estimated from labeled training data, the solution to the maximum entropy problem is also the solution to a dual maximum likelihood problem for models of the same exponential form. Additionally, it is guaranteed that the likelihood surface is convex, having a single global maximum and no local maxima [4].

2.2 Conditional Random Field

Conditional Random Fields (CRFs) [7], [8] is announced as a method using undirected graphs trained to maximize a conditional probability. A linear-chain CRF with parameters \(\Lambda = \{\lambda_1, \ldots\}\) defines a conditional probability for a state (such as a label "BP (birthplace)" in Fig. 3) sequence...
$s = \langle s_1, \ldots, s_T \rangle$ given an input sequence $o = \langle o_1, \ldots, o_T \rangle$ to be

$$p_A(s|o) = \frac{1}{Z_o} \exp \left( \sum_{t=1}^{T} \sum_{i} a_{i} f_{i}(s_{i-1}, s_i, o, t) \right) \quad (5)$$

The primary advantage of CRFs over hidden Markov models (HMMs) is their conditional nature, resulting in the relaxation of independence assumptions required by HMMs. Additionally, CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs). CRFs outperform both MEMMs and HMMs on a number of real-world sequence labeling tasks [8].

2.3 Parameter Estimation

Most of the cases where ME and CRFs are applied to text understanding take usually quite large quantities of unknown parameters because text may be represented by a lot of features. Estimation of such a large number of models is not only expensive but also sensitive to sparsely distributed features. To find an optimized solution for a set of constraints, various estimation methods have been used to compute the parameters of the model that contains the features.

Generalized Iterative Scaling (GIS) [9] and Improved Iterative Scaling (IIS) [10] are popular methods for iteratively refining the model parameters. GIS scales the probability distribution $p^*$ in Eq. (6) by a factor proportional to the ratio $E_{p^*}f_i$ to $E_{p^{(n)}}f_i$, with the restriction that $\sum_i f_i = C$. We can find the optimal parameter $\lambda_i^{(n+1)}$ with the update rule [9]:

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + \frac{1}{C} \log \frac{E_{p^*}f_i}{E_{p^{(n)}}f_i} \quad (6)$$

There are general-purpose optimization techniques such as gradient ascent and conjugate gradient. Recently, another general-purpose optimization technique, the Limited-Memory Variable Metric (L-BFGS) method, has been found to be especially effective for the parameter estimating problem [11]. L-BFGS is a limited-memory algorithm for solving large nonlinear optimization problems. Variable metric methods show excellent convergence properties and can offer substantial savings in storage requirements.

3. Sentence Topics Based Templates

We adopt the template-filling approach as the knowledge acquisition process. Our knowledge base for the encyclopedia consists of a number of templates, which were semi-automatically built to reflect the most frequently asked themes in user questions and the potential answers in the encyclopedia.

3.1 Sentence Topic Hierarchy for QA

The first step in filling templates is to determine sentence topics that represent important characteristics of a domain.

Sentence topics are defined as a set of subject or event categories that are most often mentioned in the text of a certain domain.

Sentences are considered to possess one or more topics that reveal their nature in a domain. In the “PERSON” domain, for example, topics include events that happened to the person being described. Similarly, user questions can be mapped to a sentence topic because they reflect the topics of the information being sought. For example, if someone is interested in knowing the winner of the Nobel Prize in 2004, the answer is likely to be found in the sentence labeled “award” in the “PERSON” articles.

With an assumption that sentence topics to be used for a particular domain are often mentioned in the text belonging to that domain, we devised a method that helps define a set of sentence topics by selecting candidate terms and referring to a lexical database called the Korean Lexical Concept Net for Nouns (LCNN), which was manually constructed by ETRI [12]. Candidate sentences for a sentence topic are selected based on term frequency information in the domain: terms ranked within the top 30% in frequency values are chosen and assumed to represent the topics [13].

The hierarchical relationships among the terms were determined by referring to the Korean LCNN [12]. Unlike WordNet [14], where words are grouped into synonym sets that are in turn related to each other through several semantic relationships, individual nouns in the Korean LCNN are related with each other. The Korean LCNN consists of 377,588 nodes (nouns) that are hierarchically organized with a maximum depth of 12. The semantic relations in the Korean LCNN are “IS-A,” “Part-of,” “Instance-of,” “Synonym-of,” and “Antonym-of,” among which the “IS-A” relation is used for the hierarchical relationships, as shown in Fig. 1.

We constructed a topic hierarchy by mapping the terms selected from a domain to the Korean LCNN. For example, hierarchical relationships among the candidate terms, “Name,” “Alias,” and “Pen Name,” can be determined as (“Name” (“Alias”) (“Pen Name”)) by their relative positions in the Korean LCNN. Figure 1 illustrates the positions of the candidate terms (in the shaded boxes) in the Korean LCNN.
Table 1  Sentence topic hierarchy in the “PERSON” domain.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>Activity</td>
<td>Contribution</td>
<td>Nationality</td>
</tr>
<tr>
<td>Death</td>
<td></td>
<td>Institution</td>
<td>Organization</td>
</tr>
<tr>
<td>Name</td>
<td>Alias</td>
<td>Record</td>
<td>Organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discovery</td>
<td>Record</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Succession</td>
<td>Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Assertion</td>
<td>None (default)</td>
</tr>
</tbody>
</table>

The final result was examined and tuned manually for quality assurance. Since the candidate terms were automatically generated, inappropriate terms existed for sentence topics. We eliminated these terms manually in order to build a suitable sentence topic hierarchy.

With this semi-automatic method, we defined 35 sentence topics including the eight at the highest level for the “PERSON” domain as shown in Table 1. The sentence topic taxonomy for the entire encyclopedia consists of 14 domains. Table 2 shows the number of sub-topics (around 257), and some example sub-topics. Because some topics are used in more than one domain, there are 73 unique topic names.

3.2 Templates Construction

Each topic shown in Tables 1 and 2 was assumed to require a template for which we defined template elements for specific attributes. For example, the “PERSON” domain consists of templates for “birth,” “death,” “career,” “education,” and so on. The “birth” template has two attributes, “birth date” and “birth place” which can be candidate answers for user questions.

These attributes are usually appeared as specific answer types (ATs). For the purpose of defining templates and corresponding attributes, we split a complex sentence into simple ones after the part-of-speech (POS) tagging but before the AT tagging [8] procedures. In Korean, sentences can be split rather easily by rules. The rules determine whether or not a sentence is connected by a word with a special ending (suffix) which indicates the beginning of a new clause in a sentence. Figure 2 depicts our templates building steps.
in which the example sentence is composed of two simple sentences whose three attributes are "graduated school (ORG: Harvard University)" for "Education" and "start date (YEAR: 1972)" and "position (POS: Senator)" for "Career," respectively. We defined 110 templates and 268 template attributes for 14 domains. Table 3 shows the templates and template attributes for the "PERSON" domain. Comparing with Tables 2 and 3, among 257 sentence topics, 110 topics that have sub-attributes are considered as templates.

### 4. Knowledge Acquisition Using Template-Filling

We developed template-filling functionality for a KB QA system using the two-phase hybrid machine learning approach. We utilized ME model for restricting candidate labels and CRF model for assigning appropriate labels.

Our template-filling process consists of several steps as shown in Fig. 3. First, we generate the features of the words in a sentence of an encyclopedia article by attaching POS and AT tags (only the POS tags are shown in Fig. 3). Second, we assign a topic to each sentence by sentence classification using ME so that we can reduce the number of possible labels (i.e. template attributes) from which the correct one is assigned for each word. Figure 3 shows that there are only two template attributes, BD (birth date) and BP (birth place) derived from the sentence topic "birth."

Third, we apply CRF to tag the words in a sentence with their corresponding labels, such as BP, BD, and GS (graduated school). Finally we fill template attributes with words whose labels have been determined and save them in the knowledge base.

Compared to our previous template-filling system [3] that considers all possible labels at the labeling step using only CRF, the current work not only reduces the search space but also increases labeling accuracy by restricting possible template attributes (or labels) with the sentence topic and thereby avoiding mislabeling individual words. While building an accurate KB of templates is not a sufficient condition for a high quality QA, it is a necessary condition.

By the way, CRFs which have many classes are generally inefficient [8]. In our case, the total number of attributes is 268 as shown in Table 3. Our hybrid approach can reduce KB construction time in terms of both training and assigning speed.

### 5. Empirical Evaluation

To verify the efficacy of our proposed methods, we conducted a set of experiments. In a ground work, we first evaluated the accuracy of sentence topic classification. Second, we compared the efficiency of our hybrid method with using only CRF and then showed the effect of Knowledge base for QA.

#### 5.1 Data Set and Evaluation Measures

The ultimate goal of our information extraction method is construction more accurate KB to help improve QA performance. The encyclopedia used in our system currently consists of 100,382 entries (articles) and 1,017,807 sentences belonging to 14 domains. Each article in the encyclopedia contains a title, a summary, and descriptions that sometimes include multimedia resources.

To evaluate the performance of template filling for KB to be used in a QA, we used the F-score measure commonly used for text classification systems. We used both macro and micro F-scores [15] for analyzing the accuracy of semantic passage segmentation. Micro F-scores are used for comparing two systems based on their binary decisions on all the document/category pairs, whereas macro F-scores are used to compare two systems using the paired F values for individual categories, with no size distinctions made among them.

#### 5.2 Preliminary Experiment

To set the best condition of evaluations, we performed several preliminary experiments with different parameter estimation methods. Figure 4 shows changing accuracy during parameter estimation for ME and CRF in "PERSON" domain. ME and CRF were used in sentence topic classification stage and template-filling stage, respectively. In Fig. 4, dotted circles indicate the best conditions. For example, ME archived the best accuracy when we used GIS for parameter estimation and Gaussian Prior of 10 and 30 iterations.
5.3 Sentence Topic Classification Using ME

The main purpose of this part of the experiment was to establish the groundwork for further experiments. We separated knowledge acquisition process into two stages: sentence topic classification and attribute labeling. Thus, its effectiveness in turn depends on how well the sentence classification works.

A total of 6,223 documents were selected and prepared as the training corpus by human annotators, which is only 6% of all the articles in the encyclopedia. In order to decide on the amount of the training corpus that should be minimized, we started with 2% of the whole corpus and gradually increased the amount by 1% to see if any drastic changes occurred in the performance of sentence classification. We observed a steep increase in performance up to 5%, but from 5% to 6%, the difference was marginal. While it could be possible to see additional steeper increases beyond this point, our limited resources forced us to stop there.

The testing corpus consisted of 13,494 sentences. For detailed analysis, we selected 1,131 sentences belonging to the “PERSON” domain among them. For the training of the ME model, we used GIS for parameter estimation and Gaussian Prior of 10 and 30 iterations based on the result of our preliminary experiment as shown in Fig. 4.

The classification results are shown in Tables 4 and 5. Table 4 shows F-scores of individual topics in the “PERSON” domain. While the overall performance is reasonable with an average F-score of 0.877, there is wide variation among the topics. Some low scores (e.g. for “Discovery” and “Guilt”) were due to the sparseness of training data.

Table 5 shows 2-fold and 4-fold cross validation results for all the domains. We observed high performance variations across the domains, as in the result of the “PERSON” domain, with the highest precision being 0.92 (Geography & Region) and the lowest being only 0.28 (Religion). The fact that the micro-average score (0.838) is higher than the macro-average score (0.726) indicates that the domains with higher scores have a bigger portion (i.e. more documents) of the entire corpus.
5.4 Knowledge Acquisition with Sentence Topics

In this experiment, we attempted to observe how sentence topics improve the template-filling process. Instead of considering all possible labels as in the original KB construction [3], we broke down the process in two phases: restricting the sentence topic using ME to filter out possible noisy labels; and assign target labels using CRF. By this hybrid approach, the candidate labels are restricted to those in the template related with the sentence topic. For example, when a given sentence was assigned to “Education,” the candidates should be restricted to “GS (graduation school),” “GD (graduation date),” and “AD (academic degree).”

We built 1,000 and 950 tagged sentences for the “PERSON” and “Animal & Plants” domains, respectively. As shown in Fig. 2, one sentence can have multiple attributes. The number of attribute examples is 1,217 for the “PERSON” and 2,049 for “Animal & Plants”. It reflects that more sentences which contain multiple attributes within one sentence are appeared in “Animal & Plants” than “PERSON”.

For training of the template-filling task based on CRF, we used L-BFGS for parameter estimation, Gaussian Prior of 10 and 60 iterations.

Table 6 shows the comparison results. “w/o sentence topics” means that we used only CRF model for template-filling with 37 attributes, while “w/ sentence topics” means that we performed two-phase model. By using sentence topics, we obtained an improvement in both domains: 18.3% for the “PERSON” domain and 14% for the “Animal & Plant” domain. When we corrected the errors in topic assignments, we gained a further increase from 0.964 to 0.974 for the “PERSON” domain and from 0.976 to 0.994 for the “Animal & Plant” domain.

Moreover, the relative efficiency is dramatically improved. In “PERSON” domain, training time at two-phase process took four times fast (434%) because CRF without sentence topics considered all attributes (36 tags) at the same times. On the contrary, CRF in two-phase model considered related attributes (2~5 tags). Even two-phase model require sentence topic stage using ME, but CRF model usually take longer for training than ME.

In template-filling, it is important not to extract incorrect information from the text as the value of a slot. It is also important not to assign a topic label to a sentence that is not supposed to have a specific topic. As one example, the sentence, “Fabre was born in St. Léons in Aveyron, France”, has the “Birth” topic. On the other example, the sentence, “Fabre went on to accomplish many scholarly achievements.”, does not belong to any particular topic. Accordingly, there is no appropriate attribute for the sentence. In fact, any sentence with a no topic label in the training data should be considered as having the special label “NONE”. Table 7 shows the result including “NONE” topic. The template-filling method based on CRF tends to assign a label even when it has a lower weight. A potential benefit of using sentence topics is the ability to reduce the number of candidate labels, thereby increasing the probability of assigning the “NONE” label instead of an incorrect one. Despite the benefit, erroneous topic assignments may be detrimental to the performance of the KB extraction. As compared Table 6 with Table 7, the improvement was only 10.7% due to misclassified sentences which belong to “NONE” topics. It is critical to ensure that the accuracy of topic assignments is high enough before the method is used for this purpose.

5.5 KB QA Performance

To see the effect of sentence topics in KB QA, we compared two different cases: hybrid approach with sentence topics and plan CRF without sentence topics. For evaluating the performance of QA, we used the ETRI Test Set [3] consisting of (question, answer) pairs for all domains in the encyclopedia. Out of the 562 pairs of questions and answers in the ETRI Test Set, we used 66 pairs pertinent to KB QA. Table 8 shows the results. As expected, we obtained an approximate 8% improvement when the KB was constructed by using sentence topics.

<table>
<thead>
<tr>
<th>Table 6 Template-filling with sentence topics.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effectiveness</strong></td>
</tr>
<tr>
<td><strong>Prec.</strong></td>
</tr>
<tr>
<td>PERSON (17 topics/36 attributes)</td>
</tr>
<tr>
<td>w/o sentence topics</td>
</tr>
<tr>
<td>w/ sentence topics</td>
</tr>
<tr>
<td>w/ correct sentence topics</td>
</tr>
<tr>
<td>Animal &amp; Plant (8 topics/20 attributes)</td>
</tr>
<tr>
<td>w/o sentence topics</td>
</tr>
<tr>
<td>w/ sentence topics</td>
</tr>
<tr>
<td>w/ correct sentence topics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7 Template-filling including “NONE” topic.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effectiveness</strong></td>
</tr>
<tr>
<td><strong>Prec.</strong></td>
</tr>
<tr>
<td>PERSON (18 topics/36 attributes)</td>
</tr>
<tr>
<td>w/o sentence topics</td>
</tr>
<tr>
<td>w/ sentence topics</td>
</tr>
<tr>
<td>w/ correct sentence topics</td>
</tr>
<tr>
<td>Animal &amp; Plant (9 topics/20 attributes)</td>
</tr>
<tr>
<td>w/o sentence topics</td>
</tr>
<tr>
<td>w/ sentence topics</td>
</tr>
<tr>
<td>w/ correct sentence topics</td>
</tr>
</tbody>
</table>
6. Conclusion

This paper has presented a knowledge acquisition method using sentence topics for question answering. We defined templates for information extraction by the Korean concept network semi-automatically. We also proposed the two-phase information extraction model by the hybrid machine learning.

In our experiments, we examined the role of sentence topics in the template-filling task for information extraction. Our experimental result shows the improvement of 18% in F-score and 434% in training speed over the plain CRF-based method for the extraction task. In addition, our result shows the improvement of 8% in F-score for the subsequent QA task.

Through analyzing the experimental results on the proposed topic assignment method, we can refine and adjust the topic categories based on their roles and ease-of-use. Moreover, we can refine the label (attribute) assigning method for the better performance.

However, due to the lack of the patterns and features in the training data, some attributes can not be recognized by our classification method. In the future works, we will devise an efficient method to overcome these data sparseness problems.

References


Hyo-Jung Oh received the B.S. and the M.S. degrees in computer science from Chungnam National University, Daejeon, South Korea, in 1998 and 2000, respectively. Currently she is a senior researcher in Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. Her research interests include machine learning, question answering, and dialogue management.

Bo-Hyun Yun received the B.S. in Mokpo University in 1992, and the M.S. and Ph.D. degrees in computer science from Korea University, Seoul, South Korea, in 1995 and 1999, respectively. Currently he is a professor in the department of computer education, Mokwon University, Daejeon, South Korea. His research interests include Question Answering, Semantic Web, and e-Learning.