Automatic Question Generation for Learning Evaluation in Medicine

Weiming Wang[1,2], Tianyong Hao[2], and Wenyin Liu[2]

[1] School of Computer, Wuhan University, Wuhan, PR China
{whu.wweiming}@gmail.com
[2] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, PR China
{tianyong, csliuwy}@cityu.edu.hk

Abstract. An approach of automatic question generation from given learning material of medical text is presented in this paper. The main idea is to generate the questions automatically based on question templates which are created by training on many medical articles. In order to provide interesting questions, our research focuses on medical related concepts. This method can be used for evaluation of learner's comprehension after he/she finished a reading material. Different from traditional learning system the articles and questions are all prepared beforehand; participants can learn whatever new input medical articles with the help of automatic question generation.

Keywords: Question Generation, Nature Language Generation, Question-based, E-Learning, Learning Evaluation.

1 Introduction

With the continuing development of technical knowledge and expertise, people are required to update their knowledge timely. Continuing education is a good approach to keep up with technical development. Easy usage of web-based courses has potential value to learn and enhance personal knowledge. Though more and more websites provide easy access to useful learning material of specific domain, most of them fail to test the performance of the participants and thus result in poor usage. So, new technologies need to be developed to take full advantage of this learning environment.

Evaluation is an essential aspect of course learning. It is definitely helpful for any participant to test the reliability of his comprehension towards given material. Answering questions related to the course is an effective way of evaluation. One of the most important uses of questions is reflection, improving our understanding of things we have found out. To successfully respond to a question, the participant is often to integrate several skills and processes such as prior knowledge and knowledge from course material. As a result, seeking answers for specific questions leads to more in-depth understanding. Studies investigation found that more questions in the same topic can improve the level of understanding while answering different questions on the same topic area [1]. This occurs even no standard answers are given to participant.

H. Leung et al. (Eds.): ICWL 2007, LNCS 4823, pp. 242–251, 2008.

This paper describes a medical E-Learning system which generates the questions from free text documents as part of learning. We propose an approach to automatic question generation from medical documents based on question templates. The basic idea is to build question set automatically for given learning material. We have created lots of question templates on many medical articles with the help of some NLP tools. Questions are all generated based on these templates. Traditionally, learning articles in the evaluation system are prepared beforehand. Participants won't benefit from a learning system if the database does not contain desired content. In our system, participants can input whatever medical texts they want to study. And they will also benefit from automatic test since it provides immediate feedback, online grading, and online submission.

We have built a prototype and created 23 question templates. 100 medical articles on headache were tested and results show that our system achieves 88% accuracy in finding the relevant questions. And 83% of these generated questions were correctly answered.

The paper is organized as follows: Section 2 introduces some related work about question generation and learning evaluation. In Section 3 question templates which define rules of question matching and answering are presented. The question generation including processes such as article analysis and template matching are discussed in Section 4. We present the experimental results in Section 5 and the last section summarizes our current research and speculates on what the future may bring.

2 Related Work

With large number of courseware being delivered online, more and more testing and evaluation systems are currently available [2][3]. These systems normally come with database support and hence formulating different kinds of objective questions, automatic grading and records keeping for advices are all possible. Although they provide excellent support managing of related records to evaluate their understanding, but most of the tests are provided beforehand and none of them support automatic question generation.

Question generation has received great attention in recent year. It's a subclass of Natural Language Generation (NLG) which plays an important role in learning environments, data mining, information extraction and a myriad of other applications[4]. Some of them declare that they present domain-independent question generation system. However, it makes no difference with traditional Database[5]. Eriks Sneiders introduces a template-based approach to generate questions on four kinds of entities [6]. The entities in their system are person name, location, organization, or time which may get from Name Entity Recognizer (NER) tools, such as Stanford NER [7]. As a result, it fails to produce interesting questions to help improve comprehension.

This paper proposes a new evaluation system based on automatic question generation. Questions are generated from plain texts in medical area for learning.

3 Question Template Generation

3.1 Question Template

Question templates are essential to this automatic question generation approach from which questions are generated. A template represents a class of questions with same structure. There are four components in a common template: question, entries, keywords and answer. Question and the answer are the output from medical documents for learning. Entries and keywords are used for template matching. The significant innovation is entity slots - free space for data instances of medical terms, which make it possible to tailor traditional questions and include data instances from the taxonomy database [6]. Once a template is created for a test, multiple equivalent questions can be generated according to this template as necessary.

In this paper, we focus on only common medical entities, such as <Disease>, <Medicine>, <Cause>, <Therapy>, <Symptom> and <Device>. Our purpose is to find the best template and apply it to given sentences. The rules for matching and answering associated to a template are given in advance. When a question is generated, the exact answers are also generated according to the rules defined in given question template. A sample template is given as follows:

Question:"what is the symptom for <Disease>?"
Required Entries: symptom, disease;
Keywords: feel, experience, accompany
Related Answer:<symptom>.

When a sentence matches this template, we generate the exact question by replace the <Disease> with the special disease. At the same time, the relevant symptoms in this sentence are generated as the answers to this question.

3.2 Template Generation

In our system, question template plays an important role for automatic question generation. Questions are generated for medical documents with the help of these templates.

We have created nearly one hundred templates in every aspect of medical domain. Question templates are mainly created by experts based on the parsed articles, but not absolutely. With some basic medical knowledge, participants can provide satisfactory templates themselves. We give an example as follows:

(1) The <Medicine>angiotensin converting enzyme (ACE) inhibitors </ Medicine > and the <Medicine> angiotensin receptor blocker (ARB) </ Medicine > drugs both affect the <Substance>renin-angiotensin hormonal </Substance> system, which, as mentioned previously, helps regulate the blood pressure.

(2) As an added benefit, <Medicine> ACE inhibitors </Medicine> may reduce an <finding> enlarged heart </finding> (left ventricular hypertrophy) in patients with hypertension.

(3) <Medicine>ARB</Medicine> drugs are also suitable as first line agents to treat <Disease> hypertension </Disease>.

(4) In patients who have <Disease>hypertension</Disease> in addition to certain second diseases, a combination of an <Medicine>ACE inhibitor </Medicine> and an <Medicine>ARB</Medicine> drug may be effective in controlling the hypertension and also benefiting the second disease.

(5)<Medicine>angiotensin converting enzyme (ACE) inhibitors </Medicine> or <Medicine> angiotensin receptor blocking (ARB)</Medicine> drugs are the drugs of choice in patients with <Disease>heart failure</Disease >, <Disease> chronic kidney failure</Disease> (in diabetics or non-diabetics), or <Disease> heart attack (myocardial infarction) </Disease> that weakens the heart muscle (systolic dysfunction).

(6) <Medicine>Acetaminophen</Medicine> is used for the relief of <Disease > fever </Disease> as well as aches and <Symptom>pains</Symptom> associated with many conditions.

(7) There are three types of cough medications available OTC for the temporary relief of <Symptom> cough </Symptom> due to a <Disease>cold </Disease>. They are oral cough < Substance > suppressants</Substance>, <Substance> oral expectorants </Substance>, and <Medicine>topical (externally applied) medicines</Medicine>.

These sentences are talking about medicine related to special disease, so we define the question template as follow. The **question**: What medications are used in <Disease>? <Medicine> and <Disease> are **required entities** in the template. <Substance >and <Symptom> are optional ones. In our analysis we also found that some keywords are tightly related to the template. The semantic interpretation is driven by syntactic phenomena that indicated semantic predicates including nouns, verbs and adjectives. In this template, we define these keywords for interpretation:

Nouns: drug, medicine, effect.

Verbs: help, use,release, improve.

Adjectives: suitable, effective.

The main role of keywords is to indicate relationships and relevant attributes of key concepts represented by entity slots. In the above template, we use the content of <Medicine> as the answer to these questions generated from this template.

3.3 Guideline of Question Generation

There are different kinds of questions. However, some criterions are necessary defined to prevent from arbitrarily generation. The most important points are answerable, interesting and medical-related. The questions we will less generate are shown as follows:

1. **Interesting but not easy to answer**
   There is a common question for every article: "What's the article are mainly about?" It does absolutely an interesting question for every article. However

it's hard to answer automatically though readers can confer the main idea from the abstract based on the findings of therapy, disease, Medicine and so on.

2. **Easy to answer but not interesting enough**
   We may generate this question: what's the relation between <Disease> and <Therapy> ? It does really reveal something. However, it is too common to find the exact knowledge from this question. So we pay less attention on this in this paper.
3. **Less content-related**
   It's easy to get the person name, location, organization, time from articles with the help of the common NER tools. We can also generate questions through "who" in place of person name, "where" in place of location, and "when" to time. However, these questions are less related to medical so that they should no be selected. Question is approach for learning and they should reveal the level of understanding and help participants to build more reliable understanding of the articles effectively.

3.4 Scoring Templates

In general, a good question template should have ability of distinguishing relevant data instances and ignore irrelevant data instances. In the above, we just show some intuitive impression about interesting, answerable and content-related. We have introduced a method to calculate the weight of the templates for template selection. The higher the weight, the more possibility the templates are selected for questions. There are two factors affect the weight:

1. Number of concepts and what kinds of concept exist in the sentence. The more concepts, the higher weigh.
2. Evaluation to questions generated from this template. These weights would change according to students' evaluation toward questions. There are many reasons for this: generated questions are not interesting or useful, or the answers of this template are often disappointed.

$$w_m = (1 - \frac{b}{n})(1 + \lg k) \prod_{i=1}^{k} (1 + \lg n_i) w_0$$

w_m: Weight of the template.
w_0: Initial weight, usually set 1.0.
b: Number of negative evaluations towards this template;
n: Number of total questions generated from this template;
k: Number of total concept kinds;
n_i: Number of total words of the i-th concept.

4 Automatic Question Generation

The learning system is a new testing system that allows for dynamic generation of questions, from which participants can make a good comprehension by seeking the true answer and thus improve the effect of learning. General architecture of automatic question generation is shown in in Figure 1. Medical articles were parsed by MMTx [8] to identify the medical terms and classify them to different classes. We extract medical concepts according to UMLS (Unified Medical Language System [9] [10]).

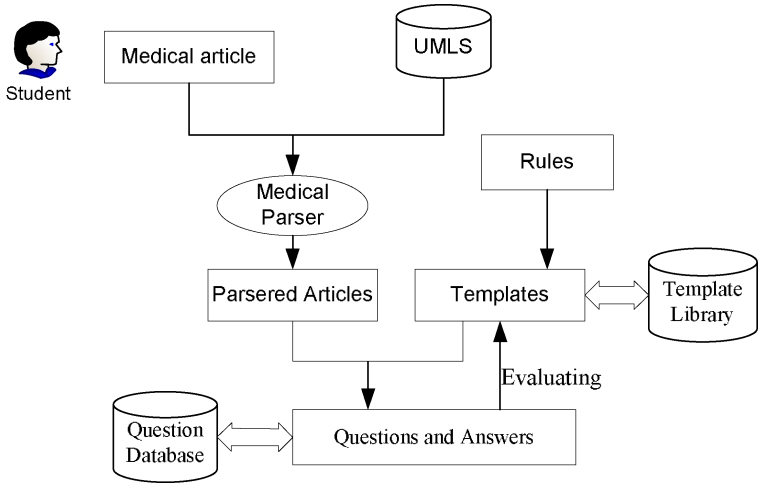

Fig. 1. Architecture of Automatic Question Generation

The research is based on UMLS. It is constructed to "understand" the meaning of the language of biomedicine and health. There are three UMLS Knowledge Sources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon, in which concepts, categories and Lexicon for indices are described separately. The current release contains over 1 million biomedical concepts and 5 million concept names, 135 semantic types and 54 relationships.

MetaMap, a component of the Semantic Knowledge Representation (SKR) project, was designed to map arbitrary text to concepts in the UMLS Metathesusus. MetaMap Transfer(MMTx) is a JAVA implementation of the MetaMap, which maps the noun phrases in the text to the best matching UMLS concept or set of concepts that best cover each phrase [8][11]. For each medical phrase, MetaMap generates synonyms, acronyms, abbreviations, spelling variants using the UMLS SPECIALIST lexicon. In our clinical QA system, we use these terms are for query expansion.

Based on these specific medical concepts from UMLS extracted by MMTx, we select the most similar templates related to this sentence stored in Template

library created beforehand. We then generate questions and answers according to the selected template and the semantic interpretation of the medical sentences. When a participant learns an article, the related questions are automatically generated for learning.

We also calculate interesting weight for each question. The higher the template weight, the more chances to be selected for question generation. The participants can determine the questions whether it is interesting or not and thus affect the weight of the question templates. Reversely, this action will affect the question generation.

4.1 Sentence Analysis

he role of sentence analysis is to produce semantic interpretations which can be used by the rest of the system. In our system, it produces the relevant medical concepts for template matching whilst convey the relations identified for answer generation. The domain knowledge is acquired through MMtx, which use the UMLS lexicon along with some rules to determine the best mapping between the text of a noun phrase and concepts in the UMLS Metathesaurus. The relevant UMLS concepts are obtain from each sentence in the medical documents, such as Influenza identified as <Disease>, Hemodialysis identified as <Therapy>. The examples are given as follows:

(8) <Therapy>Hemodialysis</Therapy> is the most common method used to treat advanced and permanent <Disease>Kidney Failure</Disease>.

(9) Because there appears to be a connection between <Medicine>aspirin </Medicine> and <Disease> Reye's syndrome <Disease>.

(10) <Doctor>Anesthesiologists</Doctor> in the ICU help patients recover from ARDS by using a <Device>mechanical ventilator </Device> to help oxygenate and ventilate the patient. Sometimes this requires sedating patients into a drug-induced <Symptom> coma </Symptom> so that they aren't too anxious and "fight" the ventilator.

(11) <Medicine>aspirin</Medicine> is no longer used to control flu-like symptoms or the symptoms of <Disease> chickenpox <Disease> in minors.

(12)<Disease>Influenza</Disease>, commonly known as the <Disease> flu </Disease>, is an infectious disease of birds and mammals caused by an <Virus> RNA virus</Virus> of the <Virus>family Orthomyxoviridae </Virus> (the influenza viruses).

After this process, medical concepts are identified from medical sentences. We discard sentences with less than two medical entity kinds for they do little help for question generation. The left sentences are further processed for semantic interpretation to generate the phrase-level answer.

4.2 Templates Matching

As we know, the questions are generated based some created question template in our system. Each template has defined its required entities and relevant keywords for matching. After the sentence analysis stage, we have obtained all

medical concepts in the sentences. A sentence will match the template if the sentence contains all required entities and at least exists one of the keywords. By matching the entities and keywords, the system makes subtler conclusion about the closeness between the question templates and the sentences. The algorithm to find candidate template for a sentence is as follows:

```
For each template in template base
 {
   If (exist all required entities) and (!exist forbidden entities)
     {
       If (sentence exist one of the required keywords)
          Add this template to candidate templates
     }
}
```

4.3 Answer Generation

In this stage, we generate the phrase-level answers for the generated questions. Semantic interpretation of the original sentence serves as the bridge. The UMLS Semantic Networks has defined a set of useful and important relationships, or semantic relations that exist between medical concepts. Such as, Therapeutic or Preventive Procedure **treats** Disease or Syndrome. We have defines some mapping rules between the keywords and the UMLS Semantic relations and generate the semantic interpretation based on these semantic relations. For example, in sentences (8), the semantic interpretation is given as follow: Hemodialysis **treats** Kidney Failure.

The semantic types of the answer are clearly defined in the question templates. To generate the exact answer, we match the semantic type and the semantic interpretation.

4.4 Question Selection

Questions are generated for learner to test the reliable of their understanding and make online assessment possible. For evaluate test we have made some rules for that:

1. Syntax difference
   Different types of questions are encouraged. A test shouldn't be full of the same questions: what's the symptom of ¡Disease¿? Even the diseases may be different. We will choose different templates for different questions.
2. Semantic difference
   Question should not ask the same point. There may be several relevant candidate templates, but we won't generate different questions for a single sentence. The template with max weight will be selected as the final choice which will make much completer use of all knowledge from the sentence.
3. Number of questions
   The questions in a test shouldn't too many or too few. We generate 5-10 questions for a articles, usually about 1 sentence for each 100 words.

4.5 Grading and Evaluation

Immediate feedback is important for the participants to know whether their understandings are reliable. The levels of comprehension are reflected by the answering given by participants. After the participants finished the test, we present the scores according to their answering. The scores are calculated the semantic similarity between the answering and generated answers.

5 Experiments and Results

We have built an automatic question generation system and test 100 medical articles using 23 created templates on every aspect of headache. The experiment was conducted on six people with different levels of medical knowledge and the terminology. Each participant answers more than 10 articles to test the accuracy of the question generation system and the ability to help enhance the comprehension.

The experiment results are listed in Table 1, in which the columns are: articles, left medical sentences, accurate questions, and correctly answered.

Table 1. Experiment result

Articles	Question Templates	Accurate Questions	Correctly Answered
100	23	88%	83%

From the result, we can find that most of the medical sentences are correctly questioned and answered. The mistakes are mainly from the created templates. The defined entries and keywords are insufficient to represent the relationships in these templates. There are some sentences satisfying the matching rules, which are derived from associated entries and keywords, cannot apply this template. And we find that low-ability participants profit more from the generated questions than the average-ability participants, which may because the questions generated are factual one.

6 Conclusions and Future Work

The main contribution of this research is online assessment through automatic question generation, answering and grading.

It would be very easy to construct a medical learning system with the help of this research. No additional personal work is needed to build the question database or grading. Participants will benefit from this automatic question generation system for the online testing and immediate feedback.

The disadvantage of our method is that the generated questions are factual and may be less meaningful than the manual questions. Furthermore, it is time consuming to parse the articles and obtain the semantic interpretation for each

sentence, especially for long articles. The current method of question generation is mainly based on a single sentence. This would result in missing some important information and might bring weird questions. More works will be done in the future, including:

- more helpful questions automatic generation and giving convictive answers.
- giving more relevant mapping and answering rules with the help of UMLS.

Acknowledgements

The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7002137), the National Grand Fundamental Research 973 Program of China under Grant No.2003CB317002.

References

1. Shavelson, R.J., Berliner, D., Ravitch, M., Loeding, D.: Effects of Position and Type of Question on Learning from Prose Material: Interaction of Treatments with Individual Differences. Journal of Educational Psychology 65(1), 40–48 (1974)
2. Tinoco, L.C., Fox, E.A., Barnette, N.D.: Online evaluation in WWW-based courseware. In: Proceedings of the twenty-eighth SIGCSE technical symposium on Computer science education, pp. 194–198. ACM Press, New York (1997)
3. Bade, D., Nüssel, G., Wilts, G.: Online Feedback by Tests and Reporting for eLearning and Certification Programs with TCmanager. In: Proceedings of the 13 thInternational World Wide Web Conference on Alternate track papers, pp. 432–433 (2004)
4. Lauer, T.W.: Questions and information: Contrasting metaphors. Information Systems Frontiers 3(1), 41–48 (2001)
5. Merzbacher, M.: Automatic Generation of Trivia Questions. In: Hacid, M.-S., Raś, Z.W., Zighed, A.D.A., Kodratoff, Y. (eds.) ISMIS 2002. LNCS (LNAI), vol. 2366, pp. 123–130. Springer, Heidelberg (2002)
6. Sneiders, E.: Automated Question Answering: Template-Based Approach. PhD thesis, Royal Institute of Technology and Stockholm University (2002)
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Ann Arbor, Michigan, Association for Computational Linguistics, June 2005, pp. 363–370 (2005)
8. Meystre, S., Haug, P.J.: Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx). Stud Health Technol Inform 116, 823–828 (2005)
9. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The Unified Medical Language System. Methods Inf. Med. 32(4), 281–291 (1993)
10. Campbell, K.E., Oliver, D.E., Shortliffe, E.H.: The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminologic Problems. Journal of the American Medical Informatics Association 5(1), 12–16 (1998)
11. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AIMA, pp. 17–21 (2001)