Available online at www.sciencedirect.com
ScienceDirect

Information Sciences 177 (2007) 3696–3717

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Semantic passage segmentation based on sentence topics for question answering

Hyo-Jung Oh [a,b,*], Sung Hyon Myaeng [b], Myung-Gil Jang [a]

[a] *Electronics and Telecommunications Research Institute (ETRI), 161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Republic of Korea*
[b] *School of Engineering, Information and Communications University, 119 Munjiro, Yuseong-gu, Daejeon 305-732, Republic of Korea*

Received 24 May 2006; received in revised form 28 February 2007; accepted 28 February 2007

Abstract

We propose a semantic passage segmentation method for a Question Answering (QA) system. We define a semantic passage as sentences grouped by semantic coherence, determined by the topic assigned to individual sentences. Topic assignments are done by a sentence classifier based on a statistical classification technique, Maximum Entropy (ME), combined with multiple linguistic features. We ran experiments to evaluate the proposed method and its impact on application tasks, passage retrieval and template-filling for question answering. The experimental result shows that our semantic passage retrieval method using topic matching is more useful than fixed length passage retrieval. With the template-filling task used for information extraction in the QA system, the value of the sentence topic assignment method was reinforced.

Keywords: Semantic passages; Passage segmentation; Topic assignment; Passage retrieval; Question answering; Sentence classification

1. Introduction

Segmenting a document into passages can be helpful in a variety of information access activities. When passages are identified based on semantics, they can be used as units of analysis and access, and are smaller and more coherent to handle than whole documents in a variety of text-based information systems.

Passage-level information access offers several advantages over document-based information access. First, the effectiveness of the retrieval system can be improved when passages are used as retrieval units first and their retrieval status values are combined later to rank documents. This is partly because the problem caused by varying document lengths can be avoided. Some similarity measures tend to favor shorter documents and thus may produce anomalous results for collections of documents of different lengths. For fixed-length passages, however, the problem of length normalization is less significant [19,36,38].

Second, segmented passages can help in the operation of other text-based information systems. For example, the identification of relevant passages of a document, which are smaller and more coherent than the whole

[*] Corresponding author. Tel.: +82 42 860 5405; fax: +82 42 860 4889.
E-mail addresses: ohj@etri.re.kr (H.-J. Oh), myaeng@icu.ac.kr (S.H. Myaeng).

0020-0255/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2007.02.038

document containing them, helps in locating the answers for the question answering (QA) system. Empirical studies on QA systems show that extracting answers from passages is more effective than extracting answers from documents [9,13,25,32].

Finally, passages are more convenient for presentation of retrieval results than full documents that may be overly long and contain several intertwined topics. That is, passages allow users to focus on relevant parts of a document rather than on the document as a whole.

There are three approaches to passage boundary identification. The first one uses the structural information of a document for passages, i.e. paragraphs or section boundaries [36]. The second defines passages of fixed length [5,7,25]. The last approach uses semantic clues or topicality for identifying passages [2,14,33]. Determining passages based on topicality has proven to be remarkably successful in QA systems.

In this paper, we propose a semantic passage segmentation method based on the topic of a sentence and examine its efficacy in several contexts. The topic of a sentence is defined to be the subject or event that is most often mentioned and deemed significant in a particular domain to which the document containing the sentence belongs. We then show how semantic passages can serve as useful and meaningful units for improving the accuracy of the QA system.

The paper is organized as follows. In Section 2 we introduce related work on passage segmentation and retrieval and its use for QA. Section 3 explains the notions of semantic passages and sentence topics in order to set the stage for semantic passage segmentation. In Section 4, we present our new semantic passage segmentation method in detail. Section 5 describes how we apply semantic passages in our QA system; and Section 6 presents our evaluation results and conclusions.

2. Related work

Text segmentation has been studied in various contexts, with or without specific applications in mind. We first introduce text segmentation methods developed for flat text with no structural information and those applied for segmentation of spoken dialogues. Contrasted to these are the techniques developed for situations in which some text structure information is available. We then introduce ways in which text segmentation is used for information retrieval and question answering. Throughout the section, the unique aspects of our proposed method are discussed in comparison with previous works.

Topic-based text segmentation or boundary detection for topical passages has been studied extensively [33]. One of the earliest is an attempt to determine a boundary by comparing text segments before and after each potential boundary [14]. Text is initially broken up into blocks of size N ranging from three to five sentences, and the resulting blocks are represented as vectors. A similarity curve of adjacent blocks is computed using cosine similarity to identify potential topic boundaries.

An extension of this approach can be found in the work of Brants and Chen [2] who use the Probabilistic Latent Semantic Analysis (PLSA) model in conjunction with the method of selecting segmentation points based on block similarity values. They use block vectors generated by folding in term vectors in the PLSA model and expand them again in order to exploit information on semantically similar words.

Also related is the work of Li and Yamanishi [27], where topic segmentation and identification are performed in a single framework. A topic is represented by means of a cluster of words that are closely related to the topic, and a stochastic model, called a finite mixture model, is employed to represent the word distribution within the text. A sequence of stochastic topic models is estimated, with an assumption that each block of the text is generated by an individual model. Significant differences between the models are detected to determine the boundaries. Caillet et al. [3] also introduced an approach with term clustering that is based on the Maximum Likelihood approach.

Ji and Zha [16] proposed a method that builds a sentence-distance matrix to represent the cohesion information of sentences—where each entry corresponds to the similarity between a sentence pair—and also adopt a dynamic programming technique to find the optimal topical boundaries. They convert the distance matrix into a gray-scale image and apply the anisotropic diffusion technique to the image representation to enhance semantic cohesion and sharpen topic boundaries.

Unlike the aforementioned approaches, which are considered unsupervised learning methods, Reynar [34] proposed a supervised learning approach in which the utilities of various cues are determined by the training corpus and used for segmentation. Instead of forming hard and fast rules by using clues, the algorithm combines the evidence contributed by clues, each of which either increases or decreases the likelihood of proposing a topic boundary between two regions of text. The cues used are: domain-specific words and phrases, word bigram frequency valued, repetition of named entities, and pronoun usages.

Text segmentation has been applied to spoken dialogues. Stokes and Carthy [39] proposed a lexical chaining-based approach to coarse-grained segmentation of CNN news transcripts. Christensen and Kolluru [7] presented a cascading Automatic Speech Reorganization (ASR) system with utterance and topic segmenters based on a Maximum Entropy model. Recently, Hsueh and Moore [15] investigated the problem of automatically predicating segment boundaries in spoken multiparty dialogues using a lexical cohesion-based model. Research for topic segmentation in a dialogue domain uses basically the same techniques as those used for text segmentation, but they are adapted to spoken language features.

While most of the text segmentation research has dealt with flat text with no structure in it, Matusov and Peters [30] concentrated on documents that follow a typical structure in the sequence and organization of individual sections. They proposed an algorithm that explicitly adds such structures as additional knowledge sources by modeling the document structure.

Previous segmentation methods differ from our proposed model in terms of topic distribution assumptions. Previous studies assume that each text block to be identified possesses a coherent topic with no need for modification of sentence sequences. On the other hand, we assume that there may be multiple topics in a particular block of text or passage and that several passages may have the same topic. This led us to devise a method that assigns a topic to each sentence in a document.

Topic segmentation has been utilized for various applications. Salton and Singhal [37] apply text structure knowledge to perform automatic text summarization by passage extraction. Using inter-document link generation techniques, they generate intra-document links between the passages of a document.

Automatically identified passages have been used for information retrieval by means of passage retrieval. Salton and Allan [36] showed that passage matching improves retrieval effectiveness and efficiency over unrestricted global text matching in many long, book-sized documents. They define passages by text paragraphs based on the number of common text components in the respective paragraphs rather than on sentences for the construction of text passages. Callan [4] defined bounded-paragraph passages by merging short paragraphs and dividing large paragraphs for information retrieval. A series of experiments suggested 20 words as a minimum paragraph size and 50 words a maximum paragraph size. Kaszkiel and Zobel [19] introduced a new type of passage, overlapping fragments of either fixed or variable length. They show that ranking these arbitrary passages substantially improves retrieval effectiveness in comparison to traditional document-ranking schemes, particularly for queries on collections of long documents.

While many passage retrieval techniques have used fixed length passages or other features such as paragraph boundaries, an exception to this is the work of Mittendorf and Schäuble [31] in which Hidden Markov Models (HMMs) were used to retrieve relevant passages of variable length. In their approach, a specific information need is modeled by a stochastic process which generates text fragments relevant to a particular query. Ponte and Croft [33] describe a method for segmentation which makes use of a query expansion technique to find common features for topic segmentations. These approaches are somewhat related to our work in that their methods generate passages of variable length.

Passage retrieval in the QA system is an intermediate step to identify text regions that are likely to contain an answer to the user question [13], with the goal of improving QA. Passages generated for QA are of either fixed length or variable length. In the latter case, passages with a varying number of sentences may overlap or disjoint [9,32] and are generated by semantic clues such as conjunctions, similarity between sentences, or relatedness to user question [20,25].

Unlike most passage retrieval, Clarke and Cormack [9] generated passages dynamically based on the user question, not using predefined passages. They developed a passage retrieval method that can identify passages of fixed size that cover as many question concepts as possible. The score of a passage depends on its length, the number of question concepts, and the relative concept weight. Lee and Hwang [25], on the other hand, generated passages of variable size, according to the user question. They obtained higher recall by using variable length rather than fixed length passages.

To the best of our knowledge, none of the past approaches to QA break a sentence or the order of sentences in generating passages. However, a passage may contain several clearly identifiable topics, and at the same time, several passages may have the same topic. This is especially true in encyclopedia text where sentences tend to be terse. Since words are not repeated in adjacent sentences, it is not easy to calculate similarity between adjacent sentences for coherence. Our approach is unique in that passages are formed by assigning a topic to individual sentences, or to sub-sentences if a sentence contains multiple topics. Essentially a passage is a collection of sentences with the same topic, regardless of their adjacency, and becomes a unit for retrieval based on the topic of a user question. It should be noted that the use of topics for passage generation is semantic in nature.

3. Semantic passages based on sentence topic

3.1. Semantic passages

The ultimate goal of our semantic passage identification is to help the QA system by providing small, meaningful text units as candidate answer passages. More specifically, we use this method in building our QA system that provides answers from a Korean encyclopedia. As such, we studied the characteristics of the encyclopedia and incorporated them in our semantic passage segmentation method.

Our encyclopedia, Pascal^{tm} Encyclopedia (http://www.epascal.co.kr), contains articles in various domains such as “Person”, “Art”, and “Science”, and each article consists of a title, a summary, and body text. One of the characteristics of the encyclopedia is that the size and the complexity of sentences vary depending on the nature of the articles containing them. When an article is a brief description of a fact, the sentences tend to be short and simple. On the other hand, sentences tend to have a complex structure, with many words, when the article explains a complex phenomenon (i.e. chemical reaction) or describes the process of a complex event (for example, World War I).

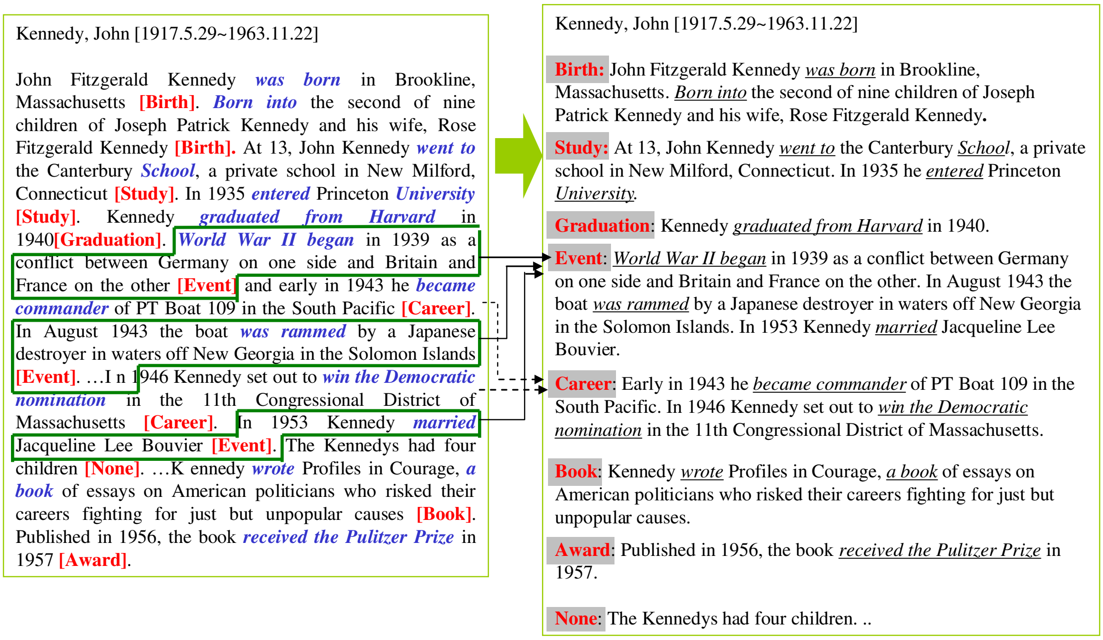
We also observed that an encyclopedia article consists of paragraphs of varying size, each of which may not have semantic coherence. The author might have decided on the paragraph boundaries based on perceived ease-of-reading. Consequently, a paragraph sometimes consists of sentences with different topics. In an article belonging to the “PERSON” domain, for example, adjacent sentences often describe different types of events in chronological order that happened to the person being described, usually from the individual’s birth to death. As a result, a paragraph, not to mention the article in which it is found, may have sentences with different subjects or topics that are all related to the person. Sometimes a single sentence contains multiple topics.

Fig. 1 shows an example of an encyclopedia article entitled “John Fitzgerald Kennedy”. The first sentence, “*John Fitzgerald Kennedy was born in Brookline, Massachusetts*”, is short and simple, explaining only his birth information. However, the fifth sentence consists of two simple phrases connected with a conjunction, whose topics are distinct from each other. While the first phrase, “*World War II began in 1939 as a conflict between Germany on one side and Britain and France on the other*” describes an event that happened during his lifetime, the second, “*early in 1943 he became commander of PT Boat 109 in the South Pacific*”, describes a career event. The left side of Fig. 1 shows sentences and their corresponding topics, while the right side shows the result of semantic passage segmentation based on the topics. As shown on the right-hand side, semantic passages can be discerned quite obviously and clearly.

Our semantic passage segmentation method is based on the aforementioned characteristics of encyclopedia articles. In other words, sentence topics are automatically determined and used for creation of passages. These passages are used as meaningful units not only for the indexing and retrieval of long texts, but also for extracting knowledge, with the goal of improving the accuracy of the QA system.

3.2. Sentence topics

The first step in forming semantic passages 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



Kennedy, John [1917.5.29~1963.11.22]

John Fitzgerald Kennedy ***was born*** in Brookline, Massachusetts **[Birth]**. ***Born into*** the second of nine children of Joseph Patrick Kennedy and his wife, Rose Fitzgerald Kennedy **[Birth].** At 13, John Kennedy ***went to*** the Canterbury ***School***, a private school in New Milford, Connecticut **[Study]**. In 1935 ***entered*** Princeton ***University*** **[Study]**. Kennedy ***graduated from Harvard*** in 1940**[Graduation]**. ***World War II began*** in 1939 as a conflict between Germany on one side and Britain and France on the other **[Event]** and early in 1943 he ***became commander*** of PT Boat 109 in the South Pacific **[Career]**. In August 1943 the boat ***was rammed*** by a Japanese destroyer in waters off New Georgia in the Solomon Islands **[Event]**. …I n 1946 Kennedy set out to ***win the Democratic nomination*** in the 11th Congressional District of Massachusetts **[Career]**. In 1953 Kennedy ***married*** Jacqueline Lee Bouvier **[Event]**. The Kennedys had four children **[None]**. …K ennedy ***wrote*** Profiles in Courage, ***a book*** of essays on American politicians who risked their careers fighting for just but unpopular causes **[Book]**. Published in 1956, the book ***received the Pulitzer Prize*** in 1957 **[Award]**.

Kennedy, John [1917.5.29~1963.11.22]

Birth: John Fitzgerald Kennedy *was born* in Brookline, Massachusetts. *Born into* the second of nine children of Joseph Patrick Kennedy and his wife, Rose Fitzgerald Kennedy**.**

Study: At 13, John Kennedy *went to* the Canterbury *School*, a private school in New Milford, Connecticut. In 1935 he *entered* Princeton *University*.

Graduation: Kennedy *graduated from Harvard* in 1940.

Event: *World War II began* in 1939 as a conflict between Germany on one side and Britain and France on the other. In August 1943 the boat *was rammed* by a Japanese destroyer in waters off New Georgia in the Solomon Islands. In 1953 Kennedy *married* Jacqueline Lee Bouvier.

Career: Early in 1943 he *became commander* of PT Boat 109 in the South Pacific. In 1946 Kennedy set out to *win the Democratic nomination* in the 11th Congressional District of Massachusetts.

Book: Kennedy *wrote* Profiles in Courage, *a book* of essays on American politicians who risked their careers fighting for just but unpopular causes.

Award: Published in 1956, the book *received the Pulitzer Prize* in 1957.

None: The Kennedys had four children. ..

Fig. 1. An example of an encyclopedia article and semantic passages.

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[1] and referring to a lexical database called the Korean Lexical Concept Net for Nouns (LCNN), which was manually constructed by ETRI [8]. Candidate sentences for a sentence topic are selected based on term frequency information in the domain: terms ranked within the top 30%[2] in frequency values are chosen and assumed to represent the topics [18].

The hierarchical relationships among the terms were determined by referring to the Korean LCNN [8]. Unlike WordNet [12], 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. 2.

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. Fig. 2 illustrates the positions of the candidate terms (in the shaded boxes) in the Korean LCNN hierarchy. Note that in the resulting topic hierarchy, the node with "Activity" is directly connected to the node with "Work", omitting the node with "Creation".

[1] A term is a single word. When it is a compound noun in Korean, however, it can be divided into multiple simple nouns.

[2] Most of the terms in the hierarchy at or above the right level of abstraction were found to be within the top 30% of the ranked list of terms based on their frequency in the text.

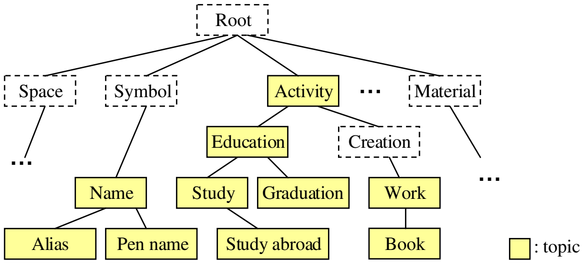


Fig. 2. Formation of a hierarchy for topics by mapping candidate terms to Korean LCNN.

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.

4. Semantic passage segmentation

Our semantic passage segmentation process consists of two phases: *the topic assignment phase*, where sentences are classified into sentence topics; and *the sentence reorganization phase*, where sentences are grouped

Table 1
Sentence topic hierarchy in the "PERSON" domain

Domain	Level 1	Level 2	Level 3
PERSON	Birth	Nationality	
	Death		
	Name	Alias	
		Pen name	
	Activity	Achievement	Contribution
			Institution
			Organization
			Record
			Discovery
			Succession
			Research
			Assertion
		History	Career
			Award
		Education	Study
			Graduation
			Study abroad
		Movement	
		Guilt	
		Fight	
		Work	Book
		Style	
		Valuation	
		Negative	
	Resistance		
	Event		
	Change		
	None (default)		

Table 2
Sentence topics for all the domains

Domain	No. of sub-topics	Examples of sub-topics
Organization & Company	26	Opening, establishment, location,...
Animal & Plant	18	Inhabitation, distribution, color, shape,...
Literature	14	Publication, foundation, premiere,...
Technical science	10	Construction, invention, proposition,...
Social science	25	Organization, establishment, name,...
Pure science	19	Matter, discovery, name, temperature,...
Language	5	Name, establishment, proposition,....
History & Cultural assets	39	Country, remains, location, name, renaming,...
Arts	15	Works, name, location, style,...
Medicine	11	Discovery, cause of disease, usage of medicine,...
Person (individual)	35	Birth, death, name, education, career,...
Religion	8	Proposition, origin, establishment,...
Geography & Region	27	Climate, location, population, race,...
Philosophy & Psychology	5	Proposition, name,...
Total 14 domains	257	Number of unique topic names: 73

into semantic passages. For the first phase, we use terms and other linguistic features in our learning-based classification approach. Compared with typical document classification methods using term distribution statistics, sentence classification needs additional features because the number of terms in a sentence is too small for training. This requirement led us to choose a classification method based on Maximum Entropy rather than other more popular statistical learning-based methods.

4.1. Learning-based classifier

By considering sentence topics as categories, we could define a sentence classifier for the encyclopedia articles. Among various classification methods based on machine learning algorithms, Naïve Bayesian (NB) and Support Vector Machine (SVM) classifiers have been most often mentioned in the text classification literature. NB uses the joint probabilities of words and categories to estimate the probability of each category for a given document. With its assumption on word independence, the NB classifier is quite efficient [26,41]. SVM is based on the structural risk minimization principle in computational learning theory whose main idea is to find a hypothesis h for which we can guarantee the lowest true error [17]. SVM classifiers have exhibited superior performance compared to NB classifiers. However, SVM has a problem with its speed relative to the size of data in both training and testing. Training an SVM classifier with very large datasets is an unsolved problem. Both NB and SVM methods are known to give good results only with a well-balanced data set. While both work well with a large number of features of one type (e.g. terms), it is not clear how they can handle multiple types of features.

The classifier we built is based on the Maximum Entropy (ME) concept [1,5], which is a general technique for estimating probability distributions or modeling a random process from data. The main principle in ME is that when nothing is known, the distribution should be as uniform as possible, which can be measured by entropy. In constructing a model or a classifier, labeled training data are used to derive a set of constraints that characterize the class-specific expectations for the distribution. Expectations on distribution are represented by the expected values of "features", any real-valued function of an example. In the context of text classification, ME estimates the conditional distribution of the class labels for a given document [28], represented as a set of term frequency features. Training data are used to estimate the expected value of these term frequencies on a class-by-class basis. One remarkable property of ME is its ability to combine multiple sources of features. For instance, ME in word sense disambiguation might consider morphological features such as part-of-speech tags, structural features like word windows, and semantic features like synset words [6] in WordNet. This is the main reason why we chose to use ME over NB or SVM.

4.1.1. Overview of maximum entropy

ME 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:

$$H(p) \equiv -\sum_{x,y} \tilde{p}(x)p(y|x)\log p(y|x) \tag{1}$$

Under the ME framework, constraints are represented by features known as feature function f in the form:

$$f(x,y) = \left\{ \begin{matrix} 1 & \text{if } (x,y) \text{ satisfies certain constraint} \\ 0 & \text{otherwise} \end{matrix} \right\}$$

The expected value of f with respect to the empirical distribution $\tilde{p}(x,y)$ is exactly the statistics we are interested in. We denote this expected value by

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x,y)f(x,y) \tag{2}$$

The expected value of f with respect to the model $p(y|x)$ is

$$p(f) = \sum_{x,y} \tilde{p}(x)p(y|x)f(x,y) \tag{3}$$

Suppose that we are given n feature functions f_i, we would like to find a model p^* to lie in the subset C of P defined by

$$C \equiv \{p^* \in P|p(f_i) = \tilde{p}(f_i) \quad \text{for} \quad i \in \{1,2,\ldots,n\}\} \tag{4}$$

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

$$p^* = \arg\max_{p \in C} H(p) \tag{5}$$

It has been shown that the ME solution p^* is unique and must be in the following form [11]:

$$p^* = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right) \tag{6}$$

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

$$Z(x) = \sum_c \exp\left(\sum_i \lambda_i f_i(x,y)\right) \tag{7}$$

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 [1].

Most of the cases where ME is 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 sparely distributed features. To find an optimized ME 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) [6] and Improved Iterative Scaling (IIS) [28] are popular methods for iteratively refining the model parameters. 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 [29].

4.1.2. Maximum entropy for sentence classification

In order to apply ME to sentence classification, we let a real-valued function of a sentence s and a topic t be a *feature*, $f(s,t)$. We use terms as our features, including nouns (N), verbs (V), and some named entities (AT: Answer type[3]). Also used are linguistic features that are discussed in Section 4.2. Suppose that the context of a sentence s is the set of all the features in s containing a term x and a named entity y, and that s is assigned a topic label z. To express this event, we can introduce the feature function $f(s,t)$ as follows:

$$f(s,t) = \begin{Bmatrix} 1 & \text{if } t = z \text{ and context(s)} \supset \{x,y\} \\ 0 & \text{otherwise} \end{Bmatrix}$$

We can have n feature functions f_i from the training data. For instance, when a sentence s_1 is "He was born in Mexico" and its topic is "Birth", the feature function $f_1(s,t)$ is defined as follows.

$$f_1(s,t) = \begin{Bmatrix} 1 & \text{if } t = \mathit{Birth} \text{ and context } (s_1) = \{\text{N} = \text{Mexico}, \text{V} = \text{born in}, \text{AT} = \text{location}\} \\ 0 & \text{otherwise} \end{Bmatrix}$$

Then we can envision an ME model, $p^* = p(t \mid s)$, where $p(t \mid s)$ represents the probability that a sentence is classified into topic t given the surrounding context s:

$$p^* = \frac{1}{Z(s)} \exp\left(\sum_i \lambda_i f_i(s,t)\right) \tag{8}$$

The model p^* constructed by adjusting the parameter λ is subject to the constraints imposed by the n feature functions. In our work, we have to deal with several types of features as well as a large number of features for sentence or text classification. This means that our ME takes large quantities of unknown parameters. In this paper, we consider several algorithms for estimating the parameters of the maximum entropy model, including GIS and L-BFGS algorithms [29]. While the sources of our training features are heterogeneous, the flexibility of the ME model allows for various features to be combined and trained easily.

4.2. Additional linguistic features

Our sentence classification based on ME is no different from traditional text classification methods in that it uses terms as training data. However, sentence classification for semantic passage segmentation has an unusual requirement. Whereas text classification deals with full documents that are usually sufficiently large for training as well as for assignment, sentence classification must deal with sentences that contain much fewer features.

To alleviate the problem in sentence classification, we decided to use other linguistic information in addition to terms, i.e. sentence patterns and extended verbs. The decision to use sentence patterns is based on our observation that a sentence belonging to a topic in an encyclopedia has a particular ⟨predicate-argument⟩ pattern. For instance, sentences labeled with the topic "Birth" always have "be born" as the predicate and "birth location" as the argument. The first sentence in Fig. 1, "*케네디는 '메세추세츠, 브룩클린'에서 '태어났다'*. (*John Fitzgerald Kennedy 'was born' in Brookline, Massachusetts.*)" is an example. By converting various expressions into a set of sentence patterns with abstracted arguments, we reduce the number of features used in training (i.e. increase the frequency of individual features), thereby increasing the possibility of matching the features of the sentence to be classified and class representations.

In a sentence pattern, the verb plays the most important role as a good indicator of a sentence topic. Unfortunately, only a small number of unique verbs were extracted from the training data because our limited resources allowed us to use only 6% of the sentences for training. Extended verbs were used to handle "unknown" verbs not found in the training data in classifying a sentence, and were generated from a verb hierarchy.

[3] In question answering, answer type (AT) is the label for the named entities that a question seeks, such as "location", "date", "person name", and so on.

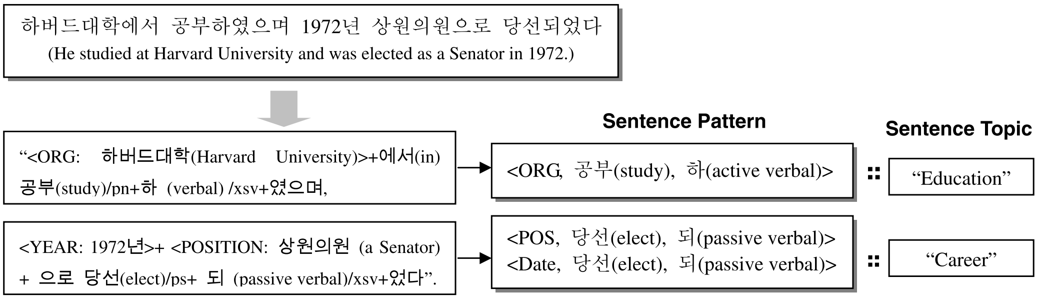

하버드대학에서 공부하였으며 1972년 상원의원으로 당선되었다
(He studied at Harvard University and was elected as a Senator in 1972.)
Sentence Pattern
Sentence Topic
“<ORG: 하버드대학(Harvard University)>+에서(in) 공부(study)/pn+하 (verbal) /xsv+였으며,
<ORG, 공부(study), 하(active verbal)>
“Education”
<YEAR: 1972년>+ <POSITION: 상원의원 (a Senator) + 으로 당선(elect)/ps+ 되 (passive verbal)/xsv+었다”.
<POS, 당선(elect), 되(passive verbal)>
<Date, 당선(elect), 되(passive verbal)>
“Career”

Fig. 3. Sentence pattern extraction.

4.2.1. Sentence patterns

Given the observation described above, the combination of the sentence pattern and the associated topic can serve as a new feature. While a sentence pattern can be defined just as the combination of a predicate and an argument (the skeleton of a sentence), in general, it can be more elaborated in Korean.

A predicate in Korean is divided into two parts, a verb and a predicative noun. For example, a predicate “공부하다 (to study)” has “공부 (study)” as the predicative noun and “하다 (do)” as the functional verb [18]. Therefore, a sentence pattern is expressed as a triplet consisting of a verb (V), a predicated noun (PN), and one or more nearest neighbor nouns (NN) of the verb, which is essentially the argument:

$$\textit{Sent-Pattern} = \langle \mathrm{NN}(\mathrm{AT}, \mathrm{N}), \mathrm{PN}, \mathrm{V} \rangle$$

NN is expressed as either AT (for an answer type tag) or N (for a general noun), depending on whether it is categorized as a named entity or not. If it is a specific named entity, it is expressed with a tag taken from the set of tags defined for answer types (AT). Otherwise it is expressed with a general noun (N). When a verb is not divided into a predicate noun and a functional verb, PN becomes “NULL”. From the example sentence, “*케네디는‘메세추세츠, 브룩클린’에서‘태어났다.’ (J.F. Kennedy ‘was born’ (verb) in ‘Brookline, Massachusetts.’ (location)*)”, we can extract ⟨LOC, NULL, be-born⟩ as the sentence pattern because the predicate “태어나다 (be born)” is a general verb.

As noted earlier in Section 3.1, a sentence in an encyclopedia article may have a complex structure with several topics embedded. For the purpose of extracting patterns, we split a complex sentence into simple ones after the part-of-speech (POS) tagging but before the AT tagging 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. Fig. 3 depicts our sentence pattern extraction steps in which the example sentence is composed of two simple sentences whose topics are “Education” and “Career”, respectively. Three sentence patterns are generated because the second part has two arguments, position and date.

4.2.2. Extended verbs

Extended verbs are generated from the Korean Lexical Concept Net for Verbs (LCNV) [8]. This lexical database consisting of 30,000[4] verbs, some of which have multiple meanings, also provides relations, such as hyponymy, synonyms and antonyms, among the verbs. For example, “죽다 (die)” has an extended verb set containing verbs like “돌아가다 (pass away)” and “급사하다 (die suddenly)” as synonyms and “암살당하다 (be assassinated)” as a passive relation. By substituting a verb from the extended verb set for the verb in a sentence pattern, we increased the number of sentence patterns from 22,916 to 38,214. In other words, more than one third of the sentence patterns were generated with the help of extended verbs.

[4] It will soon be expanded to include up to 50,000 verbs.

4.3. Semantic passage generation

To generate semantic passages, each of the sentences in an article is first divided into simple sentences that can be converted into sentence patterns. This process is preceded by linguistic analyses that includes morphological analysis, POS tagging [20], word sense disambiguation (WSD) [8], and AT tagging (i.e. named entity recognition) [24]. Our approach to WSD is based on the Korean LCNN and LCNV [8], and the location of a word to be disambiguated in either hierarchy is determined by the context of the word. Even though we had no statistics about the error rate of the WSD method,[5] we decided to use WSD because the performance of the sentence classifier performed better when it was used. The AT tagger [24] determines the label of named entities among such labels as "location", "date", "person name", and so on. The AT tagger is a special case of a general named entity tagger with a subset of the tag set used for a general entity name tagger.

After pre-processing, we apply the sentence classifier to generate a list of simple sentences with their topic tags in the order they appear in the original encyclopedia article. In order to generate coherent passages, simple sentences with the same topic tag are clumped together. An example is shown in Fig. 1 where the left box shows the original article with the topic tags enclosed in square brackets, and the right box shows the generated passages with tag names like "Birth" and "Study". The underlined phrases are the elements of sentence patterns by which the topic tags are determined.

5. Use of semantic passages in QA

Many studies on QA show that extracting answers from passages is more effective than from documents [9,13,25]. With the use of information extraction (IE) techniques, some QA systems use knowledge bases containing pre-acquired answers. In this section, we describe how the semantic passages are used for passage retrieval and information extraction for QA. Semantic passages can be used as more meaningful units than whole documents for indexing and retrieval as well as extracting facts to result in higher accuracy of the QA system.

5.1. Semantic passage retrieval for QA

The goal of semantic passage retrieval is to find candidate passages that are likely to contain an answer to a user's question. Fig. 4 illustrates the architecture of our semantic passage retrieval system in the context of

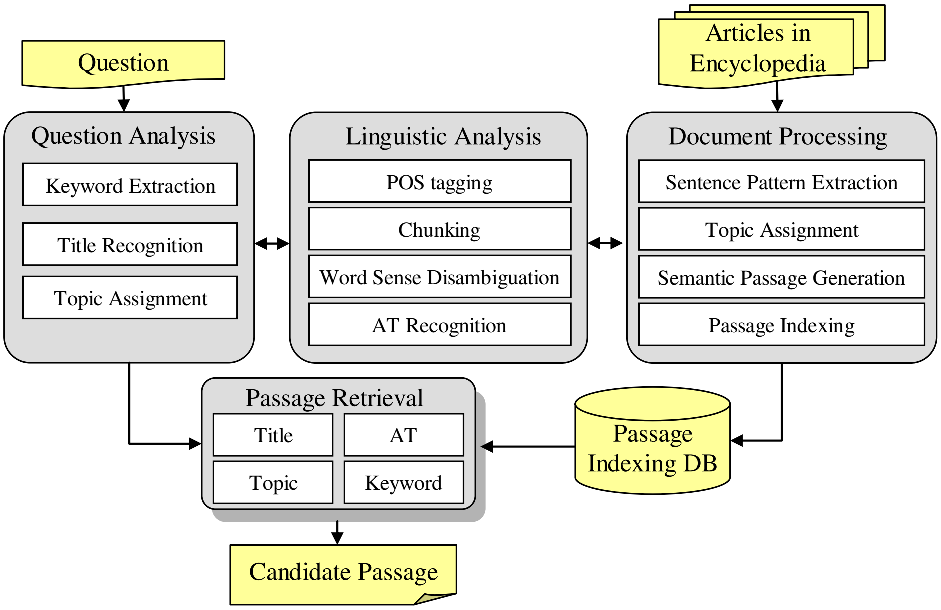


Fig. 4. Overview of passage retrieval.

QA, which retrieves the top five passages relevant to the query. In the Question Analysis module, four pieces of information are extracted by linguistic analysis: a title, an answer type (AT), a question topic, and keywords. A keyword can be a single noun, a compound noun, or a genitive noun phrase.

The features used in passage retrieval are *Title*, *Passage Topic*, *AT* (*Answer Type*), *Sentence Weight*, and *Keyword*. Each feature type has a weight that reflects its perceived importance in passage retrieval. The similarity between a question Q and a passage P is computed by summing the five feature values.

$$Score(P,Q) = Passage\_score + Keyword\_score$$
$$\begin{cases} Passage\_score = \sum_{i=1}^{3} w_i * f_i \quad \text{where } w_i = weight\ for\ Title, Passage\_topic, or\ \text{AT} \\ f_i = \begin{cases} 1 & \text{if feature } i \text{ exist in the passage} \\ 0 & \text{otherwise} \end{cases} \\ Keyword\_score = \sum_{k=1}^{|Q|} w_4 * kf_k \quad \text{where } w_4 = Sentence\ weight,\ kf_k = keyword\ weight \end{cases} \tag{9}$$

Based on our preliminary experiments, the three features used for Passage_score are ranked as follows:

$$Title > Passage\ topic > \text{AT}$$

The actual weights used for the experiments in Section 6 were 7, 5, and 3, respectively.

The *Title* feature is somewhat unique and useful in an encyclopedia in that the answer for a question about a certain entity is likely to occur in an article whose title is the same as the entity specified in the question. For example, when the user question is "*Where was J. F. Kennedy born?*" the answer might be found in an article entitled "*J.F. Kennedy*".

Passage topic can be used only when a question is assigned a certain topic. We assume that the topic of the question is likely to match that of the sentence containing the answer. In the previous example, the answer is likely to be found in a sentence labeled with the topic "Birth". This feature is distinct from a conventional passage retrieval system.

In our system, 164 Answer Types (AT) were defined and organized in a hierarchical structure. The subtype/super-type relations among the Answer Types give flexibility in matching. When a user wants to find an answer for a "location", for example, it can be matched with "city", "province", or "country" in a passage.

The *Sentence Weight* reflects the extent to which the sentence is important, and is determined by the number of query terms that appear in the sentence. Currently the Sentence Weight value for w_4 is set to 1.5 if the number of terms appearing in a sentence is more than one; otherwise it is set to 1.

The *Keyword* feature can be computed by term frequency statistics. We calculate the weight of a *keyword* using the BM25 weighting scheme [22,35]:

$$kf_k = \frac{tf_k}{K\left[(1-b) + b\frac{document\ \ length}{avg.document\ \ length}\right] + tf_k} \times \log\frac{N - df + 0.5}{df + 0.5} \tag{10}$$

where N is the number of entire passages in our encyclopedia, K is 1.5, and b is 0.5. For example, when "graduation", "high-school", and "university" have keyword weights 2.4, 1.5, and 0.5, respectively, the *Keyword_score* is calculated as follows:

$$\begin{aligned} Keyword\_score &= \sum_{k=1}^{|Q|} w_4 * kf_k = (1.5 * kf_{graduation}) + (1.5 * kf_{high\text{-}school}) + (1 * kf_{university}) \\ &= 1.5 * (2.4 + 1.5) + 0.5 \end{aligned} \tag{11}$$

assuming that "university" occurs once, whereas the other keywords occur more than once.

Fig. 5 shows an example of a user question and extracted features for passage retrieval. From the user's question asking for the birth place of Gauss, the system can compute the target title as "Gauss", the question topic as "Birth", and the target AT as "Location". The example passage can be indexed similarly so that it can

Question Processing:

Question: *가우스가 태어난 곳은?* (Where was Gauss born?)

Title: Gauss
Question Topic: Birth
Question AT: LOCATION

Passage Indexing:

Answer passage: …*브룬스비크에서 노동자의 아들로 태어나* …
(…was born in Brunswick as a son of a manual worker …)

Title: Gauss
Answer Topic: Birth
Answer AT: *비룬스비크*(Brunswick)-CITY, *노동자*(labor)-POSITION,
아들(son)-POSITION, …

Fig. 5. An example for semantic passage retrieval.

match the question representation. The target AT of the question can be extended with subtypes of the query AT so that "City" can be substituted for "Location" in a candidate answer passage.

5.2. Template-filling for knowledge-based QA

For encyclopedia-based QA, knowledge-based QA (KB QA) can be quite effective. KB QA uses a database containing domain-specific knowledge. Natural language questions are analyzed and converted into database queries to retrieve the answers. The performance of KB QA is mostly influenced by the quality of the knowledge base. Such a knowledge base is constructed by a combination of information retrieval (IR) and information extraction (IE). Candidate documents are retrieved for a database template and processed to extract values to fill in the slots in the template.

Our knowledge base for the encyclopedia consists of a number of templates, which were manually built to reflect the most frequently asked themes in user questions and the potential answers in the encyclopedia. Each topic 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 elements, "birth date" and "birth place". We defined 110 templates and 268 template elements for 14 domains. Table 3 shows the templates and template elements for the "PERSON" domain.

Our template-filling process consists of several steps as shown in Fig. 6. 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. 6). Second, we assign a topic to each sentence by sentence classification so that we can reduce the number of possible labels (i.e. template elements) from which the correct one is assigned for each word. Fig. 6

Table 3
Template for the "PERSON" domain

Domain	Template (sentence topic)	Template elements
PERSON	Birth	birth data, birth place
	Death	death date, death place, death reason
	Name	family origin, pseudonym
	Career	start date, position
	Discovery	discovered object, discovery date, discovery place
	Assertion	asserted theory
	Education	graduated school, graduation date
	Record	championship title, championship date
	Work	debut journal, debut date, debut work
	. . .	
Total 14 domains	110	268

Original sentence in the encyclopedia:

Title: Park Chung-Hee

Sentence 1: He was born in Seonsan, Kyungbuk Province.

Sentence 2: In 1937, he graduated from Taegu Normal School.

Step 1. Generating features for a given sentence:

Sentence 1: He was born in Seonsan, Kyungbuk Province.

- Feature: {word-2=born word-1=in word=Seonsan word+1=, word+2=Kyungbuk tag-2=VBD tag-1=IN tag=NN tag+1=, tag+2=NNP,…, verb=born}

Step 2. Determining the sentence topic to restrict labels:

Sentence 1: He was born in Seonsan, Kyungbuk Province.

- Sentence Topic: Birth
- Restrict target labels to *birth date(BD), Birth place (BP)*

Step 3. Labeling sentences using CRF:

Sentence 1: He was born in Seonsan , Kyungbuk Province .

- Label(s_1): _ _ _ _ BP BP BP BP _

Sentence 2: In 1937 , he graduated from Taegu Normal School .

- Label(s_2): _ GD _ _ _ _ GS GS GS _

* BP: Birth Place; GD: Graduation Date; GS: Graduation School

Step 4. Filling in the template:

<Knowledge Base>

Title	Template-Element	Record	Score
Park Chung-Hee	Birth-birth place (BP)	**Seonsan, Kyungbuk Province**	0.72
	Education-graduation school (GS)	Taegue Normal School	0.89
	Education-graduation date (GD)	1937	0.89

* Scores are computed using the Conditional Random Field (CRF) model.

Fig. 6. The process of filling a template.

shows that there are only two template elements, BD (birth date) and BP (birth place) derived from the sentence topic “birth”. Third, we apply Conditional Random Field (CRF) [23,24], which is a method using undirected graphs trained to maximize a conditional probability to tag the words in a sentence with their corresponding labels, such as BP, BD, and GS (graduated school). Finally we fill template elements with words whose labels have been determined and save them in the knowledge base.

Compared to our previous template-filling system [23] that considers all possible labels at the labeling step, the current work not only reduces the search space but also increases labeling accuracy by restricting possible template elements (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.

6. Empirical evaluation

To verify the efficacy of our proposed methods, we conducted a set of experiments. In a preliminary experiment, we first compared several classification algorithms to find the most suitable one for sentence topic classification. After fixing the topic classification method, we evaluated the performance of semantic passage segmentation and then showed the effects of sentence topics in passage retrieval and KB construction for QA.

6.1. Data set and evaluation measure

For evaluation, we developed a system called "AnyQuestion" (http://anyq.etri.re.kr). The encyclopedia used in our system currently consists of 100,373 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 semantic passage segmentation and information extraction 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 [41] 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.

For evaluation of the passage retrieval task, we used MRR (Mean Reciprocal Rank) and Top 5 Precision. MRR is a measure to reflect the answer order [40]. We considered the top five of the ranked list of answer candidates. If the correct answer is found within the top five candidates, the system is judged to have provided the correct answer.

6.2. Preliminary experiment

Since our ultimate goal was to build an operational QA system, we tested several classification models to choose the one with the best effectiveness and efficiency. The classification methods we compared were Naïve Bayesian (NB), Support Vector Machine (SVM), and Maximum Entropy (ME). In addition, a voting algorithm was employed to combine the three. If a topic is chosen by any two of the three methods, it is chosen as the final topic label. If all the three methods assign different topics, however, the one with the highest score is chosen. Effectiveness was measured with precision, recall, and *F*-measure while efficiency was measured with training time.

In the case of the ME model, we tested two parameter estimation methods, GIS and L-BFGS. GIS scales the probability distribution p^* in Eq. (8) by a factor proportional to the ratio $E_{\tilde{p}}f_i$ to $E_{p^{(n)}}f_i$, with the restriction that $\sum_i f_i = C$. We can find the optimal parameter λ_i^* with the update rule [10]:

$$\lambda_i^{(n+1)} = \lambda_i^{(n)} + \frac{1}{C}\log\left[\frac{E_{\tilde{p}}f_i}{E_{p^{(n)}}f_i}\right] \tag{12}$$

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 [29]. To avoid the problem of overfitting, a number of iterations and Gaussian Prior smoothing were required [21], and these were set to 30 and 10, respectively.

Table 4 shows the results obtained from the "PERSON" domain test set that was also used for the results in Section 6.3. NB gave the worst result among the three. SVM was as good in effectiveness as the ME methods,

Table 4
Comparisons among the three classification models

	Effectiveness			Relative efficiency
	Precision	Recall	*F*-score	
NB	0.718	0.788	0.751	1.3
SVM	0.852	0.839	0.845	2
ME with GIS	**0.884**	**0.870**	**0.877**	**1**
ME with L-BFGS	0.879	0.862	0.870	0.8
Voting (NB, SVM, ME)	0.882	0.868	0.875	2.1

but it took twice as much time as the ME model. ME with GIS was slightly better and much faster than the voting model. We conjecture that the reason why NB and SVM showed lower performance was because sentence topics were distributed unevenly. In the "PERSON" domain, for example, sentences labeled with the topic "Birth" appeared 10 times more than those with the topic "Guilty". In fact, the domains in our encyclopedia are unbalanced. For instance, the number of training examples for the "PERSON" domain is 6,329, whereas the number for the "RELIGION" domain is only 19.

For sentence classification based on topics, we used not only traditional term features, but also additional linguistic information such as *sentence patterns* and *extended verbs*. Table 5 shows the effects of using these for both the SVM and ME cases. In the case of ME, we obtained an approximate 7% improvement by using the additional linguistic information. While SVM was slightly better than ME in *F*-score when only terms were used, the situation was reversed when the linguistic information was used. This result is consistent with the theoretical argument that ME is more flexible in accommodating a new feature type. Since ME with GIS gave the best result, it was used for later experiments.

6.3. Performance of semantic passage segmentation

The main purpose of this part of the experiment was to establish the groundwork for further experiments. Semantic passage segmentation corresponds to the indexing stage of a semantic passage retrieval system, and its effectiveness in turn depends on how well the sentence classification works. As such, the performance of semantic passage segmentation and a part of the subsequent passage retrieval in a general setting is estimated by the sentence classification task in this experiment using a selected training and testing corpora.

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. The classification results are shown in Tables 6 and 7.

Table 6 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 errors in sentence patterns caused by the sparseness of training data. Therefore we decided to lower the weight of sentence patterns for those topics that showed low accuracy, minimizing the effect of errors in sentence categorization in semantic passage retrieval.

Table 7 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.

Table 5
Effects of linguistic information

Model	Features	Effectiveness		
		Precision	Recall	*F*-score (improvement)
SVM	Terms only	0.836	0.822	0.829
	Sent-pattern with extended-verb	0.852	0.839	0.845 (+1.9%)
ME with GIS	Terms only	0.824	0.811	0.817
	Sent-pattern	0.867	0.853	0.860 (+5.3%)
	Sent-pattern with extended-verb	**0.884**	**0.870**	**0.877** (**+7.3**%)

Table 6
Result of sentence classification for the "PERSON" domain

Domain	Level 1	Level 2	Level 3	Precision	Recall	*F*-score
PERSON	Birth			0.756	0.949	0.841
		Nationality		0.890	0.620	0.731
	Death			1.000	0.894	0.944
	Name			0.200	0.200	0.200
		Alias		0.550	0.690	0.612
		Pen name		0.300	0.300	0.300
	Activity			0.391	0.783	0.522
		Achievement		0.851	0.778	0.813
			Contribution	0.857	1.000	0.923
			Institution	1.000	1.000	1.000
			Organization	0.833	0.833	0.833
			Record	0.667	1.000	0.800
			Discovery	0.333	1.000	0.500
			Succession	0.500	0.750	0.600
			Research	0.364	1.000	0.522
			Assertion	1.000	0.933	0.966
		History		0.938	0.625	0.750
			Career	0.696	9.964	0.808
			Award	0.875	0.700	0.778
		Education		1.000	0.938	0.968
			Study	0.846	1.000	0.917
			Graduation	1.000	0.889	0.941
			Study abroad	1.000	1.000	1.000
		Movement		0.974	0.949	0.961
		Guilt		0.065	0.667	0.118
		Fight		1.000	1.000	1.000
		Work		0.744	1.000	0.853
			Book	0.972	1.000	0.986
		Style		0.714	0.833	0.769
		Valuation		0.571	0.800	0.667
		Negative		0.750	0.857	0.800
	Resistance			1.000	1.000	1.000
	Event			0.944	0.944	0.944
	Change			0.800	0.800	0.800
	None			0.600	0.448	0.513
Micro-average				**0.884**	**0.870**	**0.877**

6.4. *Use of semantic passages in passage retrieval for QA*

In this experiment, we evaluated the impact of semantic passages in passage retrieval for QA. To provide an appropriate answer to user questions, the first step is to find appropriate passages in which the answer is likely to be found. The goal of semantic passage retrieval is to find such candidate passages. Our method was compared against a simpler method of using fixed-length passages in the setting of QA and was generated under the following conditions: varying the number of sentences (three, five, and seven), overlapping between paragraphs, and no overlapping. The result shown in this paper is based on the case where fixed length passages consisting of three sentences with no overlaps were used as it gave the best result among all the variations.

For evaluating the performance of QA, we used the ETRI Test Set [20,23] consisting of ⟨question, answer⟩ pairs for all domains in the encyclopedia. We selected 195 pairs corresponding to the "PERSON" domain. As in Table 8, our semantic passage retrieval gave a slight improvement of about 4% in MRR over the baseline. For Top 5 precision, no improvement was observed. However, it should be noted that the semantic passage retrieval method gave more correct answers (96) in the top rank than the fixed-length passage retrieval method (85).

The result in Table 8 was obtained without topic matching between questions and answer sentences. This means the passage topic features were not considered in computing the score of a passage. Sentence topics

Table 7
Result of sentence classification for all domains

Domain	No. of sub-topics	No. of examples	Precision	
			2-Fold	4-Fold
Organization & Company	26	241	0.570	0.583
Animal & Plant	18	1476	0.804	0.819
Literature	14	249	0.881	0.879
Technical science	10	80	0.713	0.738
Social science	25	457	0.652	0.691
Pure science	19	472	0.715	0.748
Language	5	33	0.938	0.813
History & Cultural assets	39	737	0.579	0.700
Arts	15	88	0.477	0.739
Medicine	11	77	0.938	0.778
Person (individual)	35	6329	0.874	0.884
Religion	8	19	0.278	0.278
Geography & Region	27	3178	0.915	0.917
Philosophy & Psychology	5	58	0.828	0.821
Macro-avg.	257	13,494	**0.726**	0.742
Micro-avg.			**0.838**	0.854

Table 8
Result of QA with fixed-length passages and semantic passages

Ranking	A. QA with fixed-length passages					B. QA with semantic passages				
	1	2	3	4	5	1	2	3	4	5
Number of correct answers	**85**	111	121	127	130	**96**	108	122	127	128
MRR	0.53 (baseline)					0.55 (+3.8%)				
Top 5 precision	0.67					0.66				

were used only for semantic passage generation, not for semantic passage retrieval. Since the passage topic feature is a differentiating factor in our semantic passage retrieval compared to other systems, we ran an experiment using the 35 topics of the "PERSON" domain to evaluate the performance of semantic passage retrieval. Table 9 shows the result.

Compared to the result shown in Table 8A, the performance with topic matching, as shown in Table 9A, increased by about 13% (0.67–0.76) in Top 5 precision, and 15% (0.53–0.61) in MRR. In other words, we gained an improvement by using sentence topics at both the semantic passage generation phase and the retrieval phase. Table 9B is the case where only 25 topics, after eliminating some error-prone topics iscussed in 6.3, were used. While the Top 5 precision stayed the same, the MRR results increased from 0.61 to 0.66.

In the proposed method, questions were processed with a sequence of linguistic analyses: morphological analysis, chunking, word sense disambiguation, AT recognition, and topic assignment. The errors incurred by each of the analyses resulted in inaccurate query representations that had a negative impact on retrieval effectiveness. Table 10 shows the error types.

Table 9
Improvement with topic matching

Ranking	A. Semantic passage-based QA with topic matching (all topics)					B. Semantic passage-based QA with topic matching (25 selected topics)				
	1	2	3	4	5	1	2	3	4	5
Number of correct answers	**100**	120	134	146	148	**116**	128	138	148	149
MRR	0.61 (+15%)					0.66 (+24.5%)				
Top 5 precision	0.76 (+13.4%)					0.76 (+13.4%)				

Table 10
Error types in semantic passage

Error type	Topic	Titles	Work name	Keyword	Homonym	AT	Others
Proportion	40%	16%	9%	6%	7%	4%	18%

The largest proportion of errors was due to inaccuracy in the topic assignment process. The topics that are accurately assigned resulted in high accuracy of passage retrieval (e.g. "Death", "Contribution", "Award", "Education", "Movement", and "Work"), whereas those not assigned accurately (e.g. "Activity", "Discovery", "Guilty") ended up lowering the performance of passage retrieval.

Among the error types in Table 10, "Work name" and "Keyword" errors were most affected by the morphological analysis. For "Work name", for example, "wind" was mistakenly extracted as a keyword when the title was "Gone with the Wind". Even when a keyword is correctly extracted, a homonym error could occur when sense disambiguation is not done correctly, causing an error with synonymy extension. Errors with "Title" were caused mainly by variations in user questions, especially with named entities. For example, "President Kennedy" can be expressed as "J.F. Kennedy", "John Fitzgerald Kennedy", or "J.F.K". Errors made in the morphological analysis phase or chunking phase sometimes have a negative impact on the AT tagging step at a later stage.

6.5. Use of sentence topics for template filling for KB construction

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 [23], we added a restricting step to filter out possible noisy labels. By this step, 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 1217 and 2049 tagged sentences for the "PERSON" and "Animal & Plants" domains, respectively. 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 11 shows that 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.

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. In fact, any sentence with a no topic label in the training data should be considered as having the special label "NONE". The template-filling method based on CRF tends to assign a label even when it has a lower weight.

Table 11
Template-filling with sentence topics

	Effectiveness		
	Precision	Recall	*F*-score (improvement)
PERSON			
w/o sentence topics	0.817	0.813	0.815
w/ sentence topics	0.984	0.945	0.964 (+18.3%)
w/ correct sentence topics	**0.994**	**0.954**	**0.974 (+19.5%)**
Animal & Plant			
w/o sentence topics	0.849	0.859	0.856
w/ sentence topics	0.979	0.973	0.976 (+14.0%)
w/ correct sentence topics	**0.997**	**0.991**	**0.994** (**+16.3%**)

Table 12
Effect of sentence topics in an IE-based QA

IE-based QA	Questions	Retrieved	Correct	Top 5		
				Precision	Recall	*F*-score
w/o sentence topic	66	57	**48**	0.842	0.727	**0.780**
w/ sentence topic	66	55	**51**	0.927	0.773	**0.843** (**+8.1%**)

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. It is critical to ensure that the accuracy of topic assignments is high enough before the method is used for this purpose.

To see the effect of sentence topics in an IE-based QA, we compared two different cases: with sentence topics and without sentence topics. Out of the 562 pairs of questions and answers in the ETRI Test Set [23], we used 66 pairs pertinent to IE-based QA. Table 12 shows the results. As expected, we obtained an approximate 8% improvement when the KB was constructed by using sentence topics.

7. Conclusion

We proposed a semantic passage segmentation method based on the notion of sentence topics to enhance the performance of our question answer system. We defined a semantic passage as a set of sentences grouped by semantic coherence determined by the topic assigned to individual sentences. We built a sentence topic classifier based on the Maximum Entropy (ME) model using terms and additional linguistic information called sentence patterns as features. Finally, we showed experimental results on the efficacy of the proposed method and its impact on application tasks, passage retrieval and template-filling for question answering.

Our experimental work consisted of three parts. In the first part, we evaluated the effectiveness of the sentence classification method to be used for semantic passage segmentation. The result was encouraging with a 0.838 micro-average score. In the second part, we compared the proposed semantic passage retrieval method based on sentence classification against the baseline of using fixed-length passages. The experimental result showed that our semantic passage retrieval method using topic match is more useful than the fixed length passage retrieval. With the selective topic match, we gained a 24.5% increase in MRR and a 13.4% increase in Top 5 precision. In the final part, we investigated the role of sentence topics in the template-filling task for information extraction as well as the subsequent QA task. We obtained about 18% and 8% improvements over the plain CRF-based method in *F*-score for the extraction task and for the QA task, respectively.

Having seen the effect of the proposed topic assignment method and the sensitivity of the topic categories on semantic passage retrieval and template filling tasks, we plan to refine and adjust the topic categories based on their roles and ease-of-use. At the same time, the sentence classification method can be tuned for better results, especially for operational systems. Given that some topics were hardly recognized by the classification method, primarily due to the lack of training data containing the patterns and features, we will have to devise a way to overcome the data sparseness problem. We also plan to compare the proposed passage retrieval method against others that have been developed, mostly in the context of the TREC work.

References

[1] A.L. Berger, S.A. Pietar, V.J. Pietra, Maximum entropy approach to natural language processing, Computational Linguistics 22 (1) (1996) 39–71.

[2] T. Brants, F. Chen, I. Tsochantaridis, Topic-based document segmentation with probabilistic latent semantic analysis, in: Proceedings of the 11th International Conference on Information and Knowledge management (CIKM-02), 2002, pp. 211–218.

[3] M. Cllilet, J. Pessiot, M. Amini, P. Gallinari, Unsupervised learning with term clustering for thematic segmentation of texts, in: Proceedings of the 7th Recherche d'Information Assistée par Ordinateur (RIAO-04), 2004, pp. 1–11.

[4] J.P. Callan, Passage-retrieval evidence in document retrieval, in: Proceedings of 17th annual international ACM-SIGIR, 1994, pp. 302–310.

[5] F.E. Chakik, A. Shahine, J. Jaam, A. Hasnah, An approach for constructing complex discriminating surfaces based on Bayesian interference of the maximum entropy, Information Sciences 163 (4) (2004) 275–291.

[6] G. Chao, M.G. Dyer, Maximum entropy models for word sense disambiguation, in: Proceedings of 19th International Conference on Computational Linguistics (COLING-02), 2002, pp. 1–7.

[7] H. Christensen, B. Kolluru, Y. Gotoh, S. Renals, Maximum entropy segmentation on broadcast news, in: Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP-05), 2005, pp. 1029–1032.

[8] M.R. Choi, J. Hur, M.G. Jang, Constructing Korean lexical concept network for encyclopedia question-answering system, in: Proceedings of IECON 2004 – 30th Annual Conference of IEEE Industrial Electronics Society, 2004, pp. 3115–3119.

[9] C.L.A. Clarke, G.V. Cormack, T.R. Lynam, C.M. Li, G.L. McLearn, Web reinforced question answering (MultiText experiments for TREC 2001), in: Proceedings of Text Retrieval Conference (TREC), 2001, pp. 673–679.

[10] J. Darroch, D. Ratcliff, Generalized iterative scaling for log-linear models, The Annals of Mathematical Statistics 43 (5) (1972) 1470–1480.

[11] S. Della Pietra, V. Della Pietra, J. Lafferty, Inducing features of random fields, IEEE Transaction on Pattern Analysis and Machine Intelligence 19 (4) (1997) 380–393.

[12] Christiane Fellbaum, WordNet, an electronic lexical database, The MIT Press, 1998.

[13] S.M. Harabagiu, S.J. Maiorano, Finding answers in large collections of texts: paragraph indexing + abductive inference, in: Proceeding of AAAI, 1999, pp. 63–71.

[14] M. Hearst, Multi-paragraph segmentation of expository text, in: Proceedings of the 32nd Annual meeting of the Association of Computational Linguistics (ACL-94), 1994, pp. 9–16.

[15] P. Hsueh, J. Moore, S. Penals, Automatic segmentation of multiparty dialogue, in: Proceeding of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 2006, pp. 273–280.

[16] X. Ji, H. Zha, Domain-independent text segmentation using anisotropic diffusion and dynamic programming, in: Proceedings of 26th annual international ACM SIGIR, 2003, pp. 322–329.

[17] Thorsten Joachims, Learning to classify text using support vector machines, Kluwer, 2002.

[18] B.Y. Kang, S.H. Myaeng, Theme assignment for sentences based on head-driven patterns, in: Proceedings of 8th Conference on Text, Speech and Dialogue (TSD), 2005, pp. 187–194.

[19] M. Kaszkiel, J. Zobel, Effective ranking with arbitrary passage, Journal of the American Society of Information Science 52 (4) (2001) 344–364.

[20] H.-J. Kim, H.-J. Oh, C.-H. Lee, M.-G. Jang, The 3-step Answer Processing Method for Encyclopedia Question-Answering System: AnyQuestion 1.0, in: Proceedings of Asia Information Retrieval Symposium (AIRS-04), LNCS, vol. 3689, 2004, pp. 309–312.

[21] Z. Le, Y. Tain-shun, Filtering junk mail with a maximum entropy model, in: Proceedings of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03), 2003, pp. 446–453.

[22] C.K. Lee, J.H. Wang, H.J. Kim, M.G. Jang, Extracting template for knowledge-based question-answering using conditional random fields, in: Proceedings of 28th annual international ACM SIGIR Workshop on MF/IR, 2005.

[23] C.K. Lee, G.G. Lee, M.G. Jang, Dependency structure applied to language modeling for information retrieval, ETRI Journal 28 (3) (2006) 337–346.

[24] C.K. Lee, H.J. Oh, Y.G. Hwang, S. J. Lim, M.G. Jang, et al., Fine-grained named entity recognition using conditional random fields for question answering, in: Proceedings of Asia Information Retrieval Symposium (AIRS-06), LNCS, vol. 4182, 2006, pp. 581–587.

[25] Y.S. Lee, Y.S. Hwang, H.C. Rim, Variable length passage retrieval for Q&A system, in: Proceedings of the 14th Hangul and Korean Information Processing, 2002, pp. 259–266.

[26] David D. Lewis, Representation and Learning in Information Retrieval, Ph. D. thesis, Department of Computer Science, University of Massachusetts, 1992.

[27] L. Li, K. Yamanishi, Topic analysis using a finite mixture model, in: Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 2000, pp. 35–44.

[28] K. Nigam, J. Lafferty, A. McCallum, Using maximum entropy for text classification, in: Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering, 1999, pp. 61–67.

[29] R. Malouf, A comparison of algorithms for maximum entropy parameter estimation, in: Proceedings of the 6th Conference on Natural Language Learning, 2002, pp. 49–55.

[30] E. Matusov, J. Paters, C. Meyer, H. Ney, Topic segmentation using Markov models on section level, in: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-03), 2003, pp. 471–476.

[31] E. Mittendorf, P. Schäuble, Document and passage retrieval based on hidden Markov models, in: Proceedings of 17th annual international ACM-SIGIR, 1994, pp. 318–327.

[32] D. Moldovan, M. Pasca, S. Harabagiu, M. Surdeanu, Performance issue and error analysis in an open-domain question answering system, ACM Transactions on Information Systems (TOIS) 21 (2) (2003) 113–154.

[33] J.M. Ponte, B. Croft, Text segmentation by topic, in: Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries, 1997, pp. 113–125.

[34] J. Reynar, Statistical models for topic segmentation, in: Proceeding of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), 1999, pp. 1046–1053.

[35] S.E. Roberson, S. Walker, Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, in: Proceedings of 17th Annual International ACM-SIGIR, 1994, pp. 345–354.

[36] G. Salton, J. Allan, C. Buckely, Approaches to passage retrieval in full text information systems, in: Proceedings of 16th Annual International ACM-SIGIR, 1993, pp. 49–58.

[37] G. Salton, A.K. Singhal, M. Mitra, C. Buckely, Automatic text structuring and summarization, Information Processing and Management 33 (2) (1997) 193–207.

[38] A.K. Singhal, G. Salton, M. Mitra, C. Buckely, Document length normalization, Information Processing and Management 32 (5) (1996) 619–633.

[39] N. Stokes, J. Carthy, A.F. Smeaton, SeLeCT: A lexical cohesion based news story segmentation system, Journal of AI Communications 17 (1) (2004) 3–12.

[40] E.M. Voorhees, Overview of TREC 2003 Question Answering Track, in: Proceedings of Text REtreival Conference (TREC-12), 2003, pp. 1–8.

[41] Y. Yang, X. Liu, A re-examination of text categorization methods, in: Proceedings of 22nd annual international ACM-SIGIR, 1999, pp. 42–49.