# Summarizing Definition from Wikipedia

**Shiren Ye** and **Tat-Seng Chua** and **Jie Lu**
Lab of Media Search
National University of Singapore
{yesr|chuats|luj}@comp.nus.edu.sg

## Abstract

Wikipedia provides a wealth of *knowledge*, where the first sentence, infobox (and relevant sentences), and even the entire document of a wiki article could be considered as diverse versions of summaries (definitions) of the target topic. We explore how to generate a series of summaries with various lengths based on them. To obtain more reliable associations between sentences, we introduce wiki concepts according to the internal links in Wikipedia. In addition, we develop an extended document concept lattice model to combine wiki concepts and non-textual features such as the outline and infobox. The model can concatenate representative sentences from non-overlapping salient local topics for summary generation. We test our model based on our annotated wiki articles which topics come from TREC-QA 2004-2006 evaluations. The results show that the model is effective in summarization and definition QA.

## 1 Introduction

Nowadays, 'ask Wikipedia' has become as popular as 'Google it' during Internet surfing, as Wikipedia is able to provide reliable information about the concept (entity) that the users want. As the largest online encyclopedia, Wikipedia assembles immense human knowledge from thousands of volunteer editors, and exhibits significant contributions to NLP problems such as semantic relatedness, word sense disambiguation and question answering (QA).

For a given definition query, many search engines (e.g., specified by '*define:*' in Google) often place the first sentence of the corresponding wiki[1] article at the top of the returned list. The use of one-sentence snippets provides a brief and concise description of the query. However, users often need more information beyond such a one-sentence definition, while feeling that the corresponding wiki article is too long. Thus, there is a strong demand to summarize wiki articles as definitions with various lengths to suite different user needs.

The initial motivation of this investigation is to find better definition answer for TREC-QA task using Wikipedia (Kor and Chua, 2007). According to past results on TREC-QA (Voorhees, 2004; Voorhees and Dang, 2005), definition queries are usually recognized as being more difficult than factoid and list queries. Wikipedia could help to improve the quality of answer finding and even provide the answers directly. Its results are better than other external resources such as WordNet, Gazetteers and Google's *define* operator, especially for definition QA (Lita et al., 2004).

Different from the *free* text used in QA and summarization, a wiki article usually contains valuable information like infobox and wiki link. **Infobox** tabulates the key properties about the target, such as birth place/date and spouse for a person as well as type, founder and products for a company. Infobox, as a form of thumbnail biography, can be considered as a mini version of a wiki article's summary. In addition, the relevant concepts existing in a wiki article usually refer to other wiki pages by wiki internal links, which will form a close set of reference relations. The current Wikipedia recursively defines over 2 million concepts (in English) via **wiki links**. Most of these concepts are multi-word terms, whereas WordNet has only 50,000 plus multi-word terms. Any term could appear in the definition of a concept if necessary, while the total vocabulary existing in WordNet's glossary definition is less than 2000. Wikipedia addresses explicit semantics for numerous concepts. These special *knowledge* representations will provide additional information for analysis and summarization. We thus need to extend existing summarization technologies to take advantage of the *knowledge* representations in Wikipedia.

[1] For readability, we follow the upper/lower case rule on *web* (say, 'web pages' and 'on the Web'), and utilize 'wiki(pedia) articles' and 'on (the) Wikipedia', the latter referring to the entire Wikipedia.

The goal of this investigation is to explore summaries with different lengths in Wikipedia. Our main contribution lies in developing a summarization method that can (i) explore more reliable associations between passages (sentences) in huge feature space represented by wiki concepts; and (ii) effectively combine textual and non-textual features such as infobox and outline in Wikipedia to generate summaries as definition.

The rest of this paper is organized as follows: In the next section, we discuss the background of summarization using both textual and structural features. Section 3 presents the extended document concept lattice model for summarizing wiki articles. Section 4 describes corpus construction and experiments are described; while Section 5 concludes the paper.

## 2 Background

Besides some heuristic rules such as sentence position and cue words, typical summarization systems measure the associations (links) between sentences by term repetitions (e.g., LexRank (Erkan and Radev, 2004)). However, sophisticated authors usually utilize synonyms and paraphrases in various forms rather than simple term repetitions. Furnas et al. (1987) reported that two people choose the same main key word for a single well-known object less than 20% of the time. A case study by Ye et al. (2007) showed that 61 different words existing in 8 relevant sentences could be mapped into 16 distinctive concepts by means of grouping terms with close semantic (such as *[British, Britain, UK]* and *[war, fought, conflict, military]*). However, most existing summarization systems only consider the repeated words between sentences, where latent associations in terms of inter-word synonyms and paraphrases are ignored. The incomplete data likely lead to unreliable sentence ranking and selection for summary generation.

To recover the hidden associations between sentences, Ye et al. (2007) compute the semantic similarity using WordNet. The term pairs with semantic similarity higher than a predefined threshold will be grouped together. They demonstrated that collecting more links between sentences will lead to better summarization as measured by ROUGE scores, and such systems were rated among the top systems in DUC (document understanding conference) in 2005 and 2006. This WordNet-based approach has several shortcomings due to the problems of data deficiency and word sense ambiguity, etc.

Wikipedia already defined millions of multiword concepts in separate articles. Its definition is much larger than that of WordNet. For instance, more than 20 kinds of songs and movies called Butterfly , such as *Butterfly_(Kumi_Koda_song)*, *Butterfly_(1999_film)* and *Butterfly_(2004_film)*, are listed in Wikipedia. When people say something about butterfly in Wikipedia, usually, a link is assigned to refer to a particular butterfly. Following this link, we can acquire its explicit and exact semantic (Gabrilovich and Markovitch, 2007), especially for multi-word concepts. Phrases are more important than individual words for document retrieval (Liu et al., 2004). We hope that the wiki concepts are appropriate text representation for summarization.

Generally, wiki articles have little redundancy in their contents as they utilize encyclopedia style. Their authors tend to use wiki links and '*See Also*' links to refer to the involved concepts rather than expand these concepts. In general, the guideline for composing wiki articles is to avoid overlong and over-complicated styles. Thus, the strategy of '*split it*' into a series of articles is recommended; so wiki articles are usually not too long and contain limited number of sentences. These factors lead to fewer *links* between sentences within a wiki article, as compared to normal documents. However, the principle of typical extractive summarization approaches is that the sentences whose contents are repeatedly emphasized by the authors are most important and should be included (Silber and McCoy, 2002). Therefore, it is challenging to summarize wiki articles due to low redundancy (and links) between sentences. To overcome this problem, we seek (i) more reliable links between passages, (ii) appropriate weighting metric to emphasize the salient concepts about the topic, and (iii) additional guideline on utilizing non-textual features such as outline and infobox. Thus, we develop wiki concepts to replace 'bag-of-words' approach for better link measurements between sentences, and extend an existing summarization model on free text to integrate structural information.

By analyzing rhetorical discourse structure of aim, background, solution, etc. or citation context, we can obtain appropriate abstracts and the most influential contents from scientific articles (Teufel and Moens, 2002; Mei and Zhai, 2008). Similarly, we believe that the structural information such as infobox and outline is able to improve summarization as well. The outline of a wiki article using inner links will render the structure of its definition. In addition, infobox could be considered as topic signature (Lin and Hovy, 2000) or keywords about the topic. Since keywords and summary of a document can be mutually boosted (Wan et al., 2007), infobox is capable of summarization instruction.

When Ahn (2004) and Kor (2007) utilize Wikipedia for TREC-QA definition, they treat the Wikipedia as the Web and perform normal search on it. High-frequency terms in the query snippets returned from wiki index are used to extend query and rank (re-rank) passages. These snippets usually come from multiple wiki articles. Here the useful information may be beyond these snippets but existing terms are possibly irrelevant to the topic. On the contrary, our approach concentrates on the wiki article having the exact topic only. We assume that every sentence in the article is used to define the query topic, no matter whether it contains the term(s) of the topic or not. In order to extract some salient sentences from the article as definition summaries, we will build a summarization model that describes the relations between the sentences, where both textual and structural features are considered.

## 3 Our Approach

### 3.1 Wiki Concepts

In this subsection, we address how to find reasonable and reliable links between sentences using wiki concepts.

Consider a sentence: *'After graduating from Boston University in 1988, she went to work at a Calvin Klein store in Boston.'* from a wiki article *'Carolyn Bessette Kennedy'*[2], we can find 11 distinctive terms, such as *after, graduate, Boston, University,1988, go, work, Calvin, Klein, store, Boston*, if stop words are ignored.

However, multi-word terms such as *Boston University* and *Calvin Klein* are linked to the corresponding wiki articles, where their definitions are given. Clearly, considering the anchor texts as two ***wiki concepts*** rather than four words is more reasonable. Their granularity are closer to semantic content units in a summarization evaluation method Pyramid (Nenkova et al., 2007) and nuggets in TREC-QA . When the text is represented by wiki concepts, whose granularity is similar to the evaluation units, it is possibly easy to detect the matching output using a model. Here,

- Two separate words, *Calvin* and *Klein*, are meaningless and should be discarded; otherwise, spurious links between sentences are likely to occur.
- *Boston University* and *Boston* are processed separately, as they are different named entities. No link between them is appropriate[3].
- Terms such as *'John F. Kennedy, Jr.'* and *'John F. Kennedy'* will be considered as two diverse wiki concepts, but we do not account on how many repeated words there are.
- Different anchor texts, such as *U.S.A.* and *United States of America*, are recognized as an identical concept since they refer to the same wiki article.
- Two concepts, such as *money* and *cash*, will be merged into an identical concept when their semantics are similar.

In wiki articles, the first occurrence of a wiki concept is tagged by a wiki link, but there is no such a link to its subsequent occurrences in the remaining parts of the text in most cases. To alleviate this problem, a set of heuristic rules is proposed to unify the subsequent occurrences of concepts in normal text with previous wiki concepts in the anchor text. These heuristic rules include: (i) edit distance between linked wiki concept and candidates in normal text is larger than a predefined threshold; and (ii) partially overlapping words beginning with capital letter, etc.

After filtering out wiki concepts, the words remaining in wiki articles could be grouped into two sets: close-class terms like pronouns and prepositions as well as open-class terms like nouns and verbs. For example, in the sentence *'She died at age 33, along with her husband and sister'*, the open-class terms include *die*, *age*, *33*, *husband* and *sister*. Even though most open-class terms are defined in Wikipedia as well, the authors of the article do not consider it necessary to present their references using wiki links. Hence, we need to extend wiki concepts by concatenating them with these open-class terms to form an extended vector. In addition, we ignore all close-class terms, since we cannot find efficient method to infer reliable links across them. As a result, texts are represented as a vector of wiki concepts.

Once we introduce wiki concepts to replace typical 'bag-of-words' approach, the dimensions of concept space will reach six order of magnitudes. We cannot ignore the data spareness issue and computation cost when the concept space is so huge. Actually, for a wiki article and a set of relevant articles, the involved concepts are limited, and we need to explore them in a small sub-space. For instance, 59 articles about *Kennedy family* in Wikipedia have 10,399 distinctive wiki concepts only, where 5,157 wiki concepts exist twice and more. Computing the overlapping among them is feasible.

Furthermore, we need to merge the wiki concepts with identical or close semantic (namely, building links between these synonyms and paraphrases). We measure the semantic similarity between two concepts by using cosine distance between their wiki articles, which are represented as the vectors of wiki concepts as well. For computation efficiency, we calculate semantic similarities between all promising concept pairs beforehand, and then retrieve the value in a Hash table directly. We spent CPU time of about 12.5 days preprocessing the se-

[2] All sample sentences in this paper come from this article if not specified.

[3] Consider new pseudo sentence: *'After graduating from Stanford in 1988, she went to work ... in Boston.'* We do not need assign link between *Stanford* and *Boston* as well.

mantic calculation. Details are available at our technical report (Lu et al., 2008).

Following the principle of TFIDF, we define the weighing metric for the vector represented by wiki concepts using the entire Wikipedia as the observation collection. We define the CFIDF weight of wiki concept $i$ in article $j$ as:

$$w_{i,j} = cf_{i,j} \cdot idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \log \frac{|D|}{|d_j : t_i \in d_j|}, \tag{1}$$

where $cf_{i,j}$ is the frequency of concept $i$ in article $j$; $idf_i$ is the inverse frequency of concept $i$ in Wikipedia; and $D$ is the number of articles in Wikipedia. Here, sparse wiki concepts will have more contribution.

In brief, we represent articles in terms of wiki concepts using the steps below.

1. Extract the wiki concepts marked by wiki links in context.
2. Detect the remaining open-class terms as wiki concepts as well.
3. Merge concepts whose semantic similarity is larger than predefined threshold (0.35 in our experiments) into the one with largest $idf$.
4. Weight all concepts according to Eqn (1).

### 3.2 Document Concept Lattice Model

Next, we build the document concept lattice (DCL) for articles represented by wiki concepts. For illustration on how DCL is built, we consider 8 sentences from DUC 2005 Cluster *d324e* (Ye et al., 2007) as case study. 8 sentences, represented by 16 distinctive concepts A-P, are considered as the base nodes 1-8 as shown in Figure 1. Once we group nodes by means of the maximal common concepts among base nodes hierarchically, we can obtain the derived nodes 11-41, which form a DCL. A derived node will annotate a local topic through a set of shared concepts, and define a sub concept space that contains the covered base nodes under proper projection. The derived node, accompanied with its base nodes, is apt to interpret a particular argument (or statement) about the involved concepts. Furthermore, one base node among them, coupled with the corresponding sentence, is capable of this interpretation and could represent the other base nodes to some degree.

In order to Extract a set of sentences to cover key distinctive local topics (arguments) as much as possible, we need to select a set of *important* non-overlapping derived nodes. We measure the importance of node $N$ in DCL of article $j$ in term of ***representative power (RP)*** as:

$$RP(N) = \sum_{c_i \in N} (|c_i| \cdot w_{i,j}) / \log(|N|), \tag{2}$$

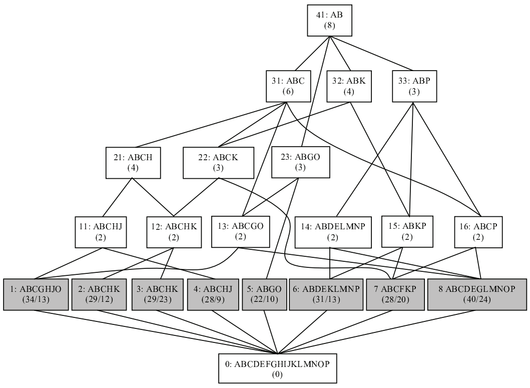
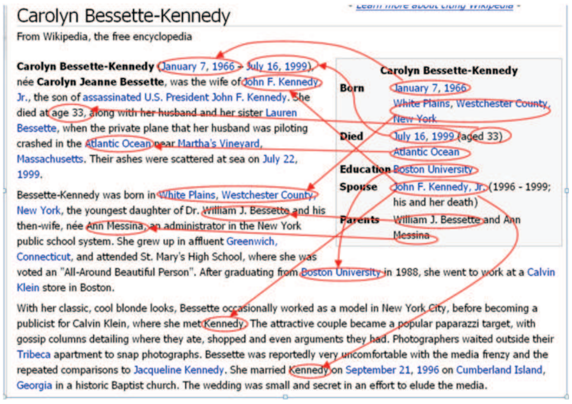


Figure 1: A sample of concept lattice

where concept $c_i$ in node $N$ is weighted by $w_{i,j}$ according to Eqn (1), and $|N|$ denotes the concept number in $N$ (if $N$ is a base node) or the number of distinct concepts in $|N|$ (if $N$ is a derived node), respectively. Here, $|c_i|$ represents the $c$'s frequency in $N$, and $\log(|N|)$ reflects $N$'s cost if $N$ is selected (namely, how many concepts are used in $N$). For example, 7 concepts in sentence 1 lead to the total $|c|$ of 34 if their weights are set to 1 equally. Its RP is $RP(1) = 34/log(7) = 40.23$. Similarly, $RP(31) = 6 * 3/log(3) = 37.73$.

By selecting a set of non-overlapping derived nodes with maximal RP, we are able to obtain a set of local topics with highest representativeness and diversity. Next, a representative sentence with maximal RP in each of such derived nodes is chosen to represent the local topics in observation. When the length of the required summary changes, the number of the local topics needed will also be modified. Consequently, we are able to select the sets of appropriate derived nodes in diverse generalization levels, and obtain various versions of summaries containing the local topics with appropriate granularities.

In the DCL example shown in Figure 1, if we expect to have a summary with two sentences, we will select the derived nodes 31 and 32 with highest RP. Nodes 31 and 32 will infer sentences 4 and 2, and they will be concatenated to form a summary. If the summary is increased to three sentences, then three derived nodes 31, 23 and 33 with maximal RP will render representative sentences 4, 5 and 6. Hence, the different number of actual sentences (4+5+6 vs. 4+2) will be selected depending on the length of the required summary. The uniqueness of DCL is that the sentences used in a shorter summary may not appear in a longer summary for the same source text. According to the distinctive derived nodes in diverse levels, the sentences with different generalization abilities are chosen to generate various summaries.



- Learn more about citing Wikipedia -

Carolyn Bessette-Kennedy

From Wikipedia, the free encyclopedia

**Carolyn Bessette-Kennedy** (January 7, 1966 – July 16, 1999), née **Carolyn Jeanne Bessette**, was the wife of John F. Kennedy, Jr., the son of assassinated U.S. President John F. Kennedy. She died at age 33, along with her husband and her sister Lauren Bessette, when the private plane that her husband was piloting crashed in the Atlantic Ocean near Martha's Vineyard, Massachusetts. Their ashes were scattered at sea on July 22, 1999.

Bessette-Kennedy was born in White Plains, Westchester County, New York, the youngest daughter of Dr. William J. Bessette and his then-wife, née Ann Messina, an administrator in the New York public school system. She grew up in affluent Greenwich, Connecticut, and attended St. Mary's High School, where she was voted an "All-Around Beautiful Person". After graduating from Boston University in 1988, she went to work at a Calvin Klein store in Boston.

With her classic, cool blonde looks, Bessette occasionally worked as a model in New York City, before becoming a publicist for Calvin Klein, where she met Kennedy. The attractive couple became a popular paparazzi target, with gossip columns detailing where they ate, shopped and even arguments they had. Photographers waited outside their Tribeca apartment to snap photographs. Bessette was reportedly very uncomfortable with the media frenzy and the repeated comparisons to Jacqueline Kennedy. She married Kennedy on September 21, 1996 on Cumberland Island, Georgia in a historic Baptist church. The wedding was small and secret in an effort to elude the media.

| **Carolyn Bessette-Kennedy** | |
|---|---|
| **Born** | January 7, 1966<br>White Plains, Westchester County, New York |
| **Died** | July 16, 1999 (aged 33)<br>Atlantic Ocean |
| **Education** | Boston University |
| **Spouse** | John F. Kennedy, Jr. (1996 - 1999; his and her death) |
| **Parents** | William J. Bessette and Ann Messina |

Figure 2: Properties in infobox and their support sentences

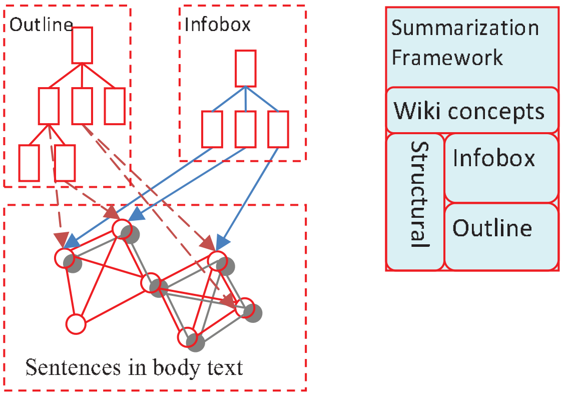


Figure 3: Extend document concept lattice by outline and infobox in Wikipedia

### 3.3 Model of Extended Document Concept Lattice (EDCL)

Different from free text and general web documents, wiki articles contain structural features, such as infoboxes and outlines, which correlate strongly to nuggets in definition TREC-QA. By integrating these structural features, we will generate better RP measures in derived topics which facilitates better priority assignment in local topics.

#### 3.3.1 Outline: Wiki Macro Structure

A long wiki article usually has a hierarchical **outline** using inner links to organize its contents. For example, wiki article *Cat* consists of a set of hierarchical sections under the outline of *mouth, legs, Metabolism, genetics*, etc. This outline provides a hierarchical clustering of sub-topics assigned by its author(s), which implies that selecting sentences from diverse sections of outline is apt to obtain a balanced summary. Actually, DCL could be considered as the composite of many kinds of clusterings (Ye et al., 2007). Importing the clustering from outline into DCL will be helpful for the generation of a balanced summary. We thus incorporate the structure of outline into DCL as follows: (i) treat section titles as concepts in the pseudo derived nodes; (ii) link these pseudo nodes and the base nodes in this section if they share concepts; and (iii) revise base nodes' RP in Eqn (2) (see Section 3.3.3).

#### 3.3.2 Infobox: a Mini Version of Summary

Infobox tabulates the key properties about the topic concept of a wiki article. It could be considered as a mini summary, where many nuggets in TREC-QA are included. As properties in infobox are not complete sentences and do not present relevant arguments, it is inappropriate to concatenate them as a summary. However, they are good indicators for summary generation. Following the terms in a property (e.g., spouse name and graduation school), we can find the corresponding sentences in the body of the text that contains such terms[4]. It describes the details about the involved property and provides the relevant arguments. We call it ***support sentence***.

Now, again, we have a hierarchy: Infobox + properties + support sentences. This hierarchy can be used to render a summary by concatenating the support sentences. This summary is inferred from hand-crafted infobox directly and is a full version of infobox; so its quality is guaranteed. However, it is possibly inapplicable due to its improper length. Following the iterative reinforcement approach for summarization and keyword extraction (Wan et al., 2007), it could be used to refine other versions of summaries. Hence, we utilize infobox and its support sentences to modify nodes' RPs in DCL so that the priority of local topics has bias to infobox. To achieve it, we extend DCL by inserting a hierarchy from infobox: (i) generate a pseudo derived node for each property; (ii) link every derived node to its support sentences; and (iii) cover these pseudo nodes by a virtual derived node called infobox.

#### 3.3.3 Summary Generation from EDCL

In DCL, sentences with common concepts form local topics by autonomous approach, where shared concepts are depicted in derived nodes. Now we introduce two additional hierarchies derived from outline and infobox into DCL to refine RPs of salient local topics for summarization, which will render a model named **extended document concept lattice** (EDCL). As shown in Figure 3, base nodes in EDCL covered by pseudo derived nodes will increase their RPs when they receive influence from outline and infobox. Also, if RPs of their covered base nodes changes, the original derived nodes will modify their RPs as well. Therefore, the new

[4] Sometimes, we can find more than one appropriate sentence for a property. In our investigation, we select top two sentences with the occurrence of the particular term if available.

RPs in derived nodes and based nodes will lead to better priority of ranking derived nodes, which is likely to result in a better summary. One important direct consequence of introducing the extra hierarchies is to increase the RP of nodes relevant to outline and infobox so that the summaries from EDCL are likely to follow human-crafted ones.

The influence of human effects are transmitted in a 'V' curve approach. We utilize the following steps to generate a summary with a given length (say $m$ sentences) from EDCL.

1. Build a normal DCL, and compute RP for each node according to Eqn 2.
2. Generate pseudo derived nodes (denoted by $P$) based on outline and infobox, and link the pseudo derived nodes to their relevant base nodes (denoted by $B_0$).
3. Update RP in $B_0$ by magnifying the contribution of shared concepts between $P$ and $N_0$[5].
4. Update RP in derived nodes that cover $B_0$ on account of the new RP in $B_0$.
5. Select $m$ non-overlapping derived nodes with maximal RP as the current observation.
6. Concatenate representative sentences with top RP from each derived node in the current observation as output.
7. If one representative sentence is covered by more than one derived node in step 5, the output will be less than $m$ sentences. In this case, we need to increase $m$ and repeat step 5-6 until $m$ sentences are selected.

## 4 Experiments

The purposes of our experiment are two-fold: (i) evaluate the effects of wiki definition to the TREC-QA task; and (ii) examine the characteristics and summarization performance of EDCL.

### 4.1 Corpus Construction

We adopt the tasks of TREC-QA in 2004-2006 (TREC 12-14) as test scope. We retrieve articles with identical topic names from Wikipedia[6]. Non-letter transformations are permitted (e.g., from *'Carolyn Bessette-Kennedy'* to *'Carolyn_Bessette-Kennedy'*). Because our focus is summarization evaluation, we ignore the cases in TREC-QA where the exact topics do not exist in Wikipedia, even though relevant topics are available (e.g., *'France wins World Cup in soccer'* in TREC-QA vs. *'France_national_football_team'* and *'2006_FIFA_World_Cup'* in Wikipedia). Finally, among the 215 topics in TREC 12-14, we obtain 180 wiki articles with the same topics.

We ask 15 undergraduate and graduate students from the Department of English Literature in National University of Singapore to choose 7-14 sentences in the above wiki articles as extractive summaries. Each wiki article is annotated by 3 persons separately. In order for the volunteers to avoid the bias from TREC-QA corpus, we do not provide queries and nuggets used in TREC-QA. Similar to TREC nuggets, we call the selected sentences ***wiki nuggets***. Wiki nuggets provides the ground truth of the performance evaluation, since some TREC nuggets are possibly unavailable in Wikipedia.

Here, we did not ask the volunteers to create snippets (like TREC-QA) or compose an abstractive summary (like DUC). This is because of the special style of wiki articles: the entire document is a long summary without trivial stuff. Usually, we do not need to concatenate key phrases from diverse sentences to form a recapitulative sentence. Meanwhile, selecting a set of salient sentences to form a concise version is a relatively less time-consuming but applicable approach. Snippets, by and large, lead to bad readability, and therefore we do not employ this approach.

In addition, the volunteers also annotate 7-10 pairs of question/answer for each article for further research on QA using Wikipedia. The corpus, called ***TREC-Wiki collection***, is available at our site (*http://nuscu.ddns.comp.nus.edu.sg*). The system of Wikipedia summarization using EDCL is launched on the Web as well.

### 4.2 Corpus Exploration

#### 4.2.1 Answer availability

The availability of answers in Wikipedia for TREC-QA could be measured in two aspects: (i) how many TREC-QA topics have been covered by Wikipedia? and (ii) how many nuggets could be found in the corresponding wiki article? We find that (i) over 80% of topics (180/215) in the TREC 12-14 are available in Wikipedia, and (ii) about 47% TREC nuggets could be detected directly from Wikipedia (examining applet modified from Pourpre (Lin and Demner-Fushman, 2006)). In contrast, 6,463 nuggets existing in TREC-QA 12-14 are distributed in 4,175 articles from AQUAINT corpus. We can say that Wikipedia is the answer goldmine for TREC-QA questions.

When we look into these TREC nuggets in wiki articles closely, we find that most of them are embedded in wiki links or relevant to infobox. It suggests that they are indicators for sentences having nuggets.

[5] We magnify it by adding $|c_0| * w_c * \eta$. Here, $c_0$ is the shared concepts between $P$ and $N_0$, and $\eta$ is the influence factor and set to 2-5 in our experiments.

[6] The dump is available at `http://download.wikimedia.org/`. Our dump was downloaded in Sept 2007.

#### 4.2.2 Correlation between TREC nuggets and non-text features

Analyzing the features used could let us understand summarization better (Nenkova and Louis, 2008). Here, we focus on the statistical analysis between TREC/wiki nuggets and non-textual features such as wiki links, infobox and outline. The features used are introduced in Table 1. The correlation coefficients are listed in Table 2.

**Observation:** (1) On the whole, wiki nuggets exhibit higher correlation to non-textual features than TREC nuggets do. The possible reason is that TREC nuggets are extracted from AQUAINT rather than Wikipedia. (2) As compared to other features, infobox and wiki links strongly relate to nuggets. They are thus reliable features beyond text for summarization. (3) Sentence positions exhibit weak correlation to nuggets, even though the first sentence of an article is a good one-sentence definition.

| Feature | Description |
|---|---|
| Link | Does the sentence have link? |
| Topic rel. | Does the sentence contain any word in topic concept? |
| Outline rel. | Does the sentence hold word in its section title(s) (outline)? |
| Infobox rel. | Is it a support sentence? |
| Position | First sentence of the article, first sentence and last sentence of a paragraph, or others? |

Table 1: Features for correlation measurement

| Feature | TREC nuggets | Wiki nuggets |
|---|---|---|
| Link | 0.087 | 0.120 |
| Topic rel. | 0.038 | 0.058 |
| Outline rel. | 0.078 | 0.076 |
| Infobox rel. | 0.089 | 0.170 |
| Position | -0.047 | 0.021 |

Table 2: Correlation Coefficients between non-textual features in Wiki and TREC/wiki nuggets

### 4.3 Statistical Characteristics of EDCL

We design four runs with various configurations as shown in Table 3. We implement a sentence reranking program using MMR (maximal marginal relevance) (Carbonell and Goldstein, 1998) in Run 1, which is considered as the test baseline. We apply standard DCL in Run 2, where concepts are determined according to their definitions in WordNet (Ye et al., 2007). We introduce wiki concepts for standard DCL in Run 3. Run 4 is the full version of EDCL, which considers both outline and infobox.

**Observations**: (1) In Run 1, the average number of distinctive words per article is near to 1200 after stop words are filtered out. When we merge diverse words having similar semantic according to WordNet concepts , we obtain 873 concepts per article on average in Run 2. The word number decreases by about 28% as a result of the omission of close-class terms and the merging of synonyms and paraphrases. (2) When wiki concepts are introduced in Run 3, the number of concepts continues to decrease. Here, some adjacent single-word terms are merged into wiki concepts if they are annotated by wiki links. Even though the reduction of total concepts is limited, these new wiki concepts will group the terms that cannot be detected by WordNet. (3) DCL based on WordNet concepts has less derived nodes (Run 3) than DCL based on wiki concepts does, although the former has more concepts. It implies that wiki concepts lead to higher link density in DCL as more links between concepts can be detected. (4) Outline and infobox will bring additional 54 derived nodes (from 1695 to 1741). Additional computation cost is limited when they are introduced into EDCL.

| | |
|---|---|
| Run 1 | Word co-occurrence + MMR |
| Run 2 | Basic DCL model (WordNet concepts) |
| Run 3 | DCL + wiki concepts |
| Run 4 | EDCL (DCL + wiki concepts + outline + infobox) |

Table 3: Test configurations

| | Concepts | Base nodes | Derived nodes |
|---|---|---|---|
| Run 1 | 1173 (number of words) | | |
| Run 2 | 873 | 259 | 1517 |
| Run 3 | 826 | 259 | 1695 |
| Run 4 | 831 | 259 | 1741 |

Table 4: Average node/concept numbers in DCL and EDCL

### 4.4 Summarization Performance of EDCL

We evaluate the performance of EDCL from two aspects such as contribution to TREC-QA definition task and accuracy of summarization in our TREC-Wiki collection.

Since factoid/list questions are about the most essential information of the target as well, like Cui's approach (2005), we treat factoid/list answers as essential nuggets and add them to the gold standard list of definition nuggets. We set the sentence number of summaries generated by the system to

12. We examine the definition quality by nugget recall (NR) and an approximation to nugget precision (NP) on answer length. These scores are combined using the $F_1$ and $F_3$ measures. The recall in $F_3$ is weighted three times as important as precision. The evaluation is automatically conducted by Pourpre v1.1 (Lin and Demner-Fushman, 2006).

Based on the performance of EDCL for TREC-QA definition task listed in Table 5, we observe that: (i) When EDCL considers wiki concepts and structural features such as outline and infobox, its F-scores increase significantly (Run 3 and Run 4). (ii) Table 5 also lists the results of Cui's system (marked by asterisk) using bigram soft patterns (Cui et al., 2005), which is trained by TREC-12 and tested on TREC 13. Our EDCL can achieve comparable or better F-scores on the 180 topics in TREC 12-14. It suggests that Wikipedia could provide high-quality definition directly even though we do not use AQUAINT. (iii) The precision of EDCL in Run 4 outperforms that of soft-pattern approach remarkably (from 0.34 to 0.497). One possible reason is that all sentences in a wiki article are oriented to its topic, and the sentence irrelevant to its topic hardly occurs.

| | **NR** | **NP** | $F_1$ | $F_3$ |
|---|---|---|---|---|
| Run 1 | 0.247 | 0.304 | 0.273 | 0.252 |
| Run 2 | 0.262 | 0.325 | 0.290 | 0.267 |
| Run 3 | 0.443 | 0.431 | 0.431 | 0.442 |
| Run 4 | 0.538 | 0.497 | 0.517 | 0.534 |
| Bigram SP* | 0.552 | 0.340 | 0.421 | 0.510 |

Table 5: EDCL evaluated by TREC-QA nuggets

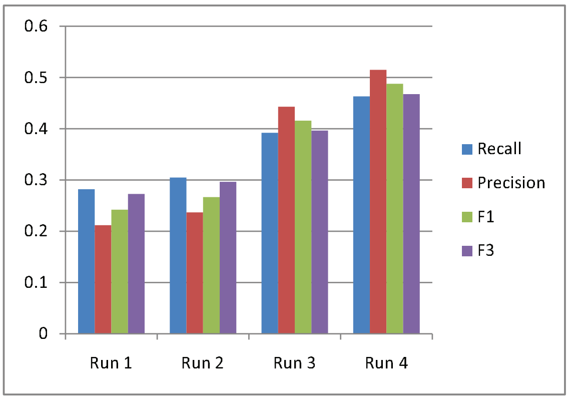


Figure 4: Performance of summarizing Wikipedia using EDCL with different configurations

We also test the performance of EDCL using extractive summaries in TREC-Wiki collection. By means of comparing to each set of sentences selected by a volunteer, we examine how many exact annotated sentences are selected by the system using different configurations. The average recalls and precisions as well as their F-scores are shown in Figure 4.

**Observations:** (i) The structural information of Wikipeida has significant contribution to EDCL for summarization. We manually examine some summaries and find that the sentences containing more wiki links are apt to be chosen when wiki concepts are introduced in EDCL. Most sentences in output summaries in Run 4 usually have 1-3 links and relevant to infobox or outline. (ii) When using wiki concepts, infobox and outline to enrich DCL, we find that the precision of sentence selection has improved more than the recall. It reaffirms the conclusion in the previous TREC-QA test in this subsection. (iii) In addition, we manually examine the summaries on some wiki articles with common topics, such as *car, house, money*, etc. We find that the summaries generated by EDCL could effectively grasp the key information about the topics when the sentence number of summaries exceeds 10.

## 5 Conclusion and Future Work

Wikipedia recursively defines enormous concepts in huge vector space of wiki concepts. The explicit semantic representation via wiki concepts allows us to obtain more reliable links between passages. Wikipedia's special structural features, such as wiki links, infobox and outline, reflect the hidden human *knowledge*. The first sentence of a wiki article, infobox (and its support sentences), outline (and its relevant sentences), as well as the entire document could be considered as diverse summaries with various lengths. In our proposed model, local topics are autonomously organized in a lattice structure according to their overlapping relations. The hierarchies derived from infobox and outline are imported to refine the representative powers of local topics by emphasizing the concepts relevant to infobox and outline. Experiments indicate that our proposed model exhibits promising performance in summarization and QA definition tasks.

Of course, there are rooms to further improve the model. Possible improvements includes: (a) using advanced semantic and parsing technologies to detect the support and relevant sentences for infobox and outline; (b) summarizing multiple articles in a wiki category; and (c) exploring the mapping from close-class terms to open-class terms for more links between passages is likely to forward some interesting results.

More generally, the *knowledge* hidden in non-textual features of Wikipedia allow the model to harvest better definition summaries. It is challenging but possibly fruitful to recast the normal documents with wiki styles so as to adopt EDCL for free text and enrich the research efforts on other NLP tasks.

## References

[Ahn et al.2004] David Ahn, Valentin Jijkoun, et al. 2004. Using Wikipedia at the TREC QA Track. In *Text REtrieval Conference*.

[Carbonell and Goldstein1998] J. Carbonell and J. Goldstein. 1998. The use of mmr, diversity-based re-ranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336.

[Cui et al.2005] Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 384–391, New York, NY, USA. ACM.

[Erkan and Radev2004] Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Artificial Intelligence Research*, 22:457–479.

[Furnas et al.1987] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

[Gabrilovich and Markovitch2007] Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

[Kor and Chua2007] Kian-Wei Kor and Tat-Seng Chua. 2007. Interesting nuggets and their impact on definitional question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–342, New York, NY, USA. ACM.

[Lin and Demner-Fushman2006] Jimmy J. Lin and Dina Demner-Fushman. 2006. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587.

[Lin and Hovy2000] Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA. ACL.

[Lita et al.2004] Lucian Vlad Lita, Warren A. Hunt, and Eric Nyberg. 2004. Resource analysis for question answering. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 18, Morristown, NJ, USA. ACL.

[Liu et al.2004] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. 2004. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 266–272, New York, NY, USA. ACM.

[Lu et al.2008] Jie Lu, Shiren Ye, and Tat-Seng Chua. 2008. Explore semantic similarity and semantic relatedness via wikipedia. Technical report, National Univeristy of Singapore, http://nuscu.ddns.comp.nus.edu.sg.

[Mei and Zhai2008] Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio, June. ACL.

[Nenkova and Louis2008] Ani Nenkova and Annie Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-08: HLT*, pages 825–833, Columbus, Ohio, June. ACL.

[Nenkova et al.2007] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2):4.

[Silber and McCoy2002] H. Grogory Silber and Kathleen F. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.

[Teufel and Moens2002] Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, December.

[Voorhees and Dang2005] Ellen M. Voorhees and Hoa Trang Dang. 2005. Overview of the trec 2005 question answering track. In *Text REtrieval Conference*.

[Voorhees2004] Ellen M. Voorhees. 2004. Overview of the trec 2004 question answering track. In *Text REtrieval Conference*.

[Wan et al.2007] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic, June. ACL.

[Ye et al.2007] Shiren Ye, Tat-Seng Chua, Min-Yen Kan, and Long Qiu. 2007. Document concept lattice for text understanding and summarization. *Information Processing and Management*, 43(6):1643–1662.