Meiji University HARD and Robust Track Experiments

Kazuya Kudo, Kenji Imai, Makoto Hashimoto and Tomohiro Takagi
Department of Computer Science, Meiji University
{kudo, imai, m-hasimo, takagi}@cs.meiji.ac.jp

1. Introduction

We participated in HARD Track and Robust Track at TREC2005. Our main challenge is to deal with expansion of a word by recognition of context. In HARD Track, we handled semantic expansion of a word. In Robust Track, we tried a challenge to new approach of "Document expansion" by context recognition.

In this paper, the next section presents HARD Track. Section 3 describes Robust Track.

2. HARD Track

We made Clarification Form (CF) using mainly two kinds of information for Query Expansion (QE). One is Local Information, which is obtained from retrieval results like Pseudo Relevance Feedback. The other is Global Information, which is implicit in a corpus. Our system showed the two kinds of information to a user as CF, and generated new query from the CF results.

We used high-ranked documents in a retrieval result as Local Information. High-ranked documents would relate to original query. Therefore, it is useful for QE in many cases. However, there is one defect that accuracy of an initial retrieval result greatly influences the quality of the information.

We used information extracted from a whole corpus as Global Information. This information is useful for difficult query which doesn't obtain a good retrieval result. However, in the research of recent years, it is assumed that the use of Global Information is generally more difficult than the use of Local Information.

In consideration of these factors, we proposed to use the two kinds of information at the same time and to cover each other's defect.

Time to evaluate items by a user in CF is limited. Consequently, to show more useful information to a user within a limited time, redundancy in question items should be weak. Our system makes sets of the documents or the words whose content or the meaning is close, and shows these sets to the user as CF.

Moreover, as a comparison with Global Information, we experimented using WordNet, the dictionary manually made.

2.1 Clarification Forms

Time to evaluate items by a user in CF is limited. Local Information and Global Information are too large to evaluate directly. So we think that it is necessary to classify each kind of information and to fix up question items in CF. In consequence, we classified the documents or the words, and made sets of documents whose content is similar or words whose meaning is close. In the approach of WordNet, Synonym set, node in the tree is considered like the cluster.

2.1.1 CF Generation using Local Information

We extracted effective information from high-ranked documents of a initial retrieval result. Our system retrieved a topic and got an initial retrieval result. Possibility that high-ranked documents are relevant to a topic is high. Therefore, we think that the documents include effective information for a topic.

Outline of CF generation using Local Information is shown in Figure2.1.

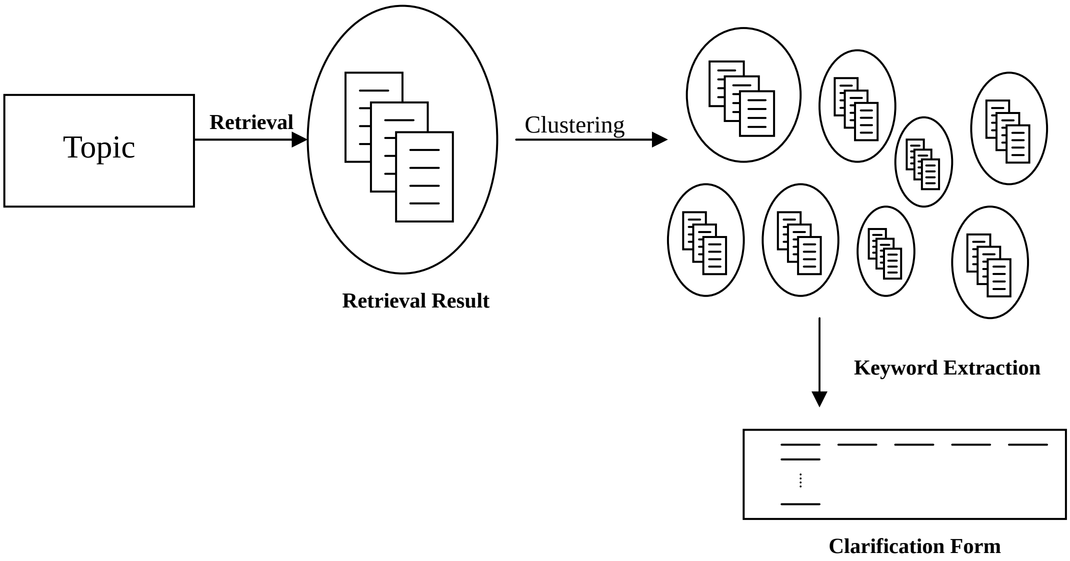


Figure 2.1 Outline of CF Generation using Local Information

We generated query using text of title and description to retrieve a topic. And we classified the top 200 documents in an initial retrieval result to divide the topic into smaller topics. We used K-means algorithm for clustering. We used cosine measure as the distance computation between a document and a cluster.

Then, we extracted keywords from each cluster to express the cluster well. We defined the formula of a degree how much a word in a cluster expresses the cluster as follows.

$$\mathrm{ClusterWordScore}(w,c) = \mathrm{ClusterPrbability}(w,c) \cdot \log \frac{\mathrm{ClusterPrbability}(w,c)}{\mathrm{OtherClusterPrbability}(w,c)}$$

$$\mathrm{ClusterPrbability}(w,c) = \frac{\mathrm{ClusterTF}(w,c)}{\mathrm{ClusterLength}(c)}$$

$$\mathrm{OtherClusterPrbability}(w,c) = \frac{\mathrm{OtherClusterTF}(w,c)}{\mathrm{OtherClusterLength}(c)}$$

$$\mathrm{OtherClusterTF}(w,c) = \sum_{c_i \neq c, c_i \in C} \mathrm{ClusterTF}(w,c)$$

$$\mathrm{OtherClusterLength}(c) = \sum_{c_i \neq c, c_i \in C} \mathrm{ClusterLength}(c_i)$$

$$\mathrm{ClusterTF}(w,c) = \sum_{d \in c} \mathrm{TF}(w,d)$$

$$\mathrm{ClusterLength}(c) = \sum_{d \in c} |d|$$

$\mathrm{TF}(w,d)$ ： TF value of the word w in document d

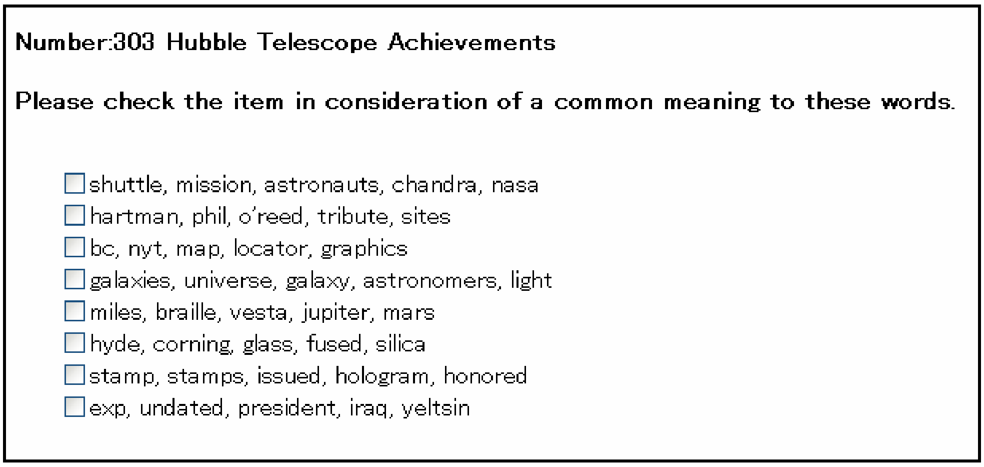
$|d|$ ： Document length

C ： Cluster set

c ： cluster c

Cluster Word Score used the ratio of appearance probability of the word w in one cluster and appearance probability of the word w in the other clusters. A word that appears very much in one cluster and doesn’t appear very much in the other cluster has the high score. We think the words are important and express the cluster well.

We extract top 5 words of Cluster Word Score. Then we lay out the words to CF. CF using Local Information that we submintted is shown in Figure2.2.



Number:303 Hubble Telescope Achievements

Please check the item in consideration of a common meaning to these words.

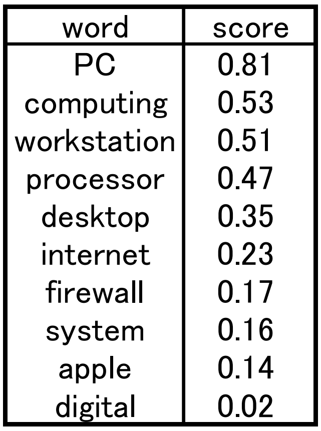
□ shuttle, mission, astronauts, chandra, nasa
□ hartman, phil, o’reed, tribute, sites
□ bc, nyt, map, locator, graphics
□ galaxies, universe, galaxy, astronomers, light
□ miles, braille, vesta, jupiter, mars
□ hyde, corning, glass, fused, silica
□ stamp, stamps, issued, hologram, honored
□ exp, undated, president, iraq, yeltsin

Figure 2.2　CF using Local Information

2.1.2 CF Generation using Global Information

First, we tried to extract information that is implicit in a corpus. We think that the meaning of a word would be explained by other words like dictionary. For instance, "Computer" can be semantically expressed by "PC", "Computing", and "Workstation". Relation between "Computer" and "Computer" is not treated. Additionally, a word is semantically expressible by giving the membership values of a related word. This example is shown in Table 2.1.

Table 2.1 Membership values of “computer"



word	score
PC	0.81
computing	0.53
workstation	0.51
processor	0.47
desktop	0.35
internet	0.23
firewall	0.17
system	0.16
apple	0.14
digital	0.02

Then, we defined this membership value by using the degree of a relation between words. The degree of a relation between words is calculated based on the co-occurrence in the same sentence. We think that if two words co-occurred frequently in the same sentence, a relation between these words is strong.

Outline of CF Generation using Global Information is shown in Figure 2.3. We calculated the degrees of relations of all words, and generated word vectors of the words.

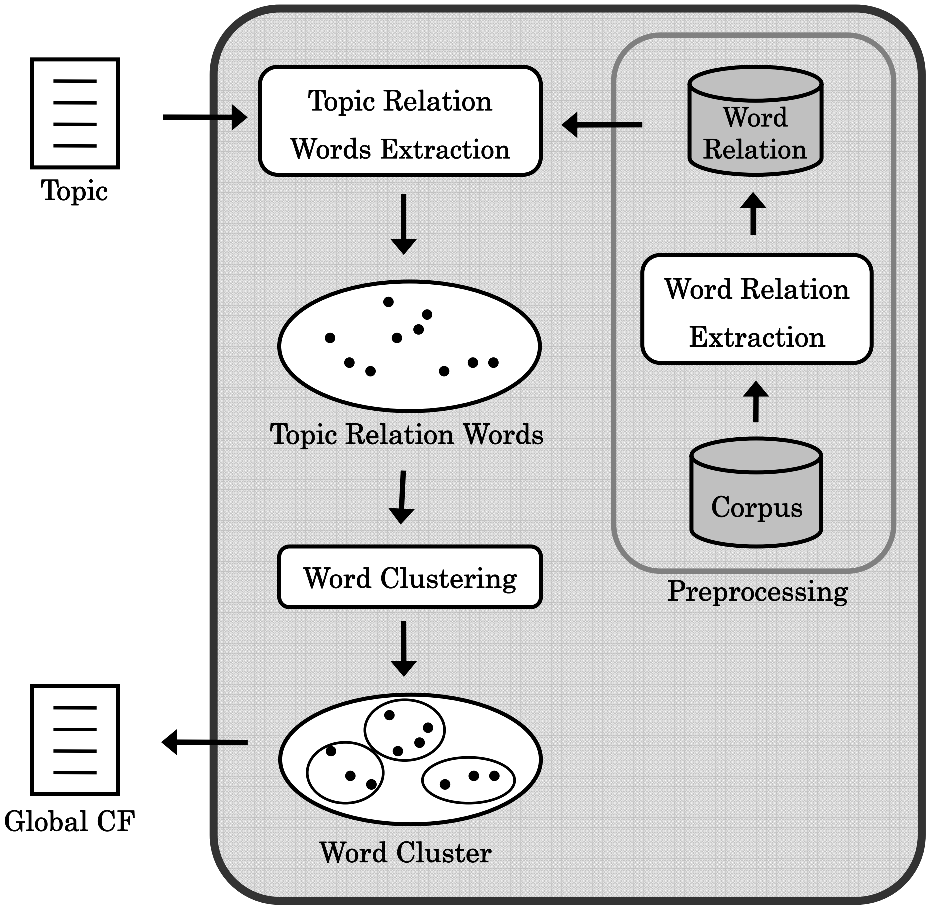


Figure 2.3 Outline of CF Generation using Global Information

There is a very serious problem to treat co-occurrence of words. This problem is that an actual relation between continuous words is infrequently weak even if these words co-occur frequently in the same sentence. One example is "Hot dog". "Hot dog" doesn't mean the relation between "Have a high temperature" and "Dog in animal", but means simply "Food". By treating these words as one word, a wrong relation between these words is not given. We called such continuous words Phrase.

We take note of the following properties to extract Phrase.

a) Words of Phrase mutually have weak meaning relation. Therefore, these tend not to appear discontinuously in the same document.
b) Phrase has generality.

The following Phrase Score is defined in consideration of these properties. Co-occurrence Probability shows a) and Generality shows b).

$$\mathrm{PhraseScore}(w_1, w_2) = \mathrm{Co{-}occur\,r\,enceProbability}(w_1, w_2) * \mathrm{Generality}(w_1, w_2)$$

$$\mathrm{Co{-}occur\,renceProbability}(w_1, w_2) = \frac{\mathrm{tf}\,(w_1 w_2)}{\mathrm{DocumentCo\text{-}occur\,r\,ence}(w_1, w_2)}$$

$$\mathrm{Generality}(w_1, w_2) = \log(\mathrm{df}\,(w_1 w_2)) - \alpha$$

$\mathrm{tf}\,(w)$: Term frequency of word w in a corpus

$\mathrm{DocumentCo - occur\ rence}\,(w_1, w_2)$: Frequency of word w_1 co-occurrence with word w_2 in the same document

$\mathrm{df}\,(w)$: Document frequency of word w

α : Constant

If the score is larger than threshold, the words are assumed to be Phrase and treated like one word.

Then, we defined the asymmetric degree of a relation between words. When word A strongly relates to word B, the following properties is effective.

a) The word B appears frequently when the word A appears.

b) The appearance probability of the word B rises by co-occurring with the word A. In other words, the appearance probability of the word B when the word A appears is larger than in a corpus.

In consideration of these properties we defined the following expressions.

$$\mathrm{RelationScore}(\mathrm{B} \leftarrow \mathrm{A}) = \mathrm{Prob}(\mathrm{B}|\mathrm{A}) * \log\left(\frac{\mathrm{Prob}(\mathrm{B}|\mathrm{A})}{\mathrm{Prob}(\mathrm{B})}\right)$$

$$\mathrm{Prob}(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{sentence{-}\,tf}\,(\mathrm{B}|\mathrm{A})}{\sum_{w} \mathrm{sentence{-}\,tf}\,(w|\mathrm{A})}$$

$$\mathrm{Prob}(\mathrm{B}) = \frac{\mathrm{tf}\,(\mathrm{B})}{\mathrm{N}}$$

$$\mathrm{N} = \sum_{w \in Corpus} \mathrm{tf}\,(w)$$

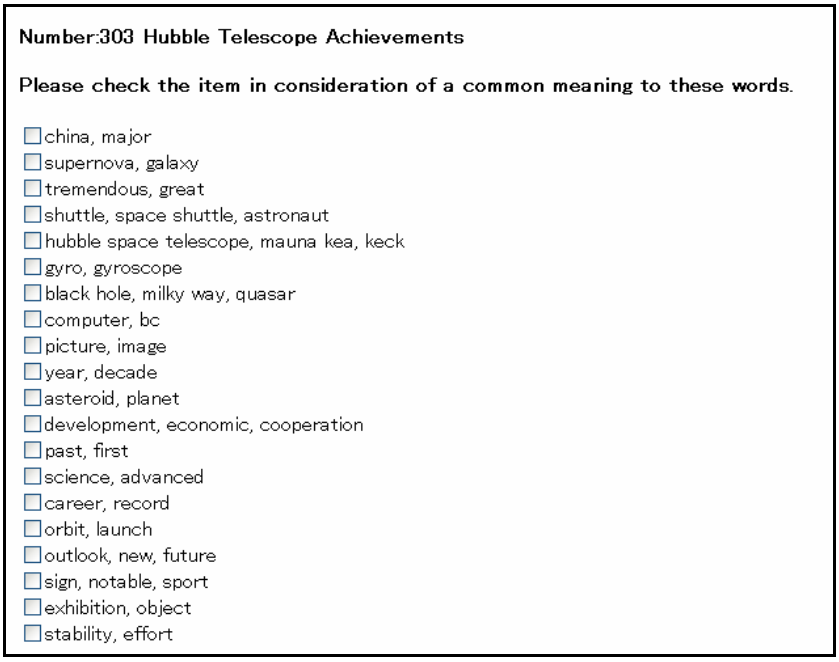
$\mathrm{tf}\,(\mathrm{B})$: Term frequency of word B in a corpus

$\mathrm{tf}\,(\mathrm{B}|\mathrm{A})$: Term frequency of word B when word A appears in the same sentence

As a result, the degree of a relation is calculated. After generating word vectors of noun words or adjective words in a corpus, the word vectors are normalized. An element in word vector corresponds to a related word.

And the word vector of a topic is calculated based on the sum of the word vectors generated from words in title and description. In consequence, a word in word vector of a topic whose value is larger than a threshold was extracted as a topic's related word.

Our system classified the extracted related words. The k-means algorithm and cosine measure were used for clustering. Centers of initial clusters were selected so that all centers might mutually become sparse. The degree of a cluster's relation to a topic is calculated from the words in the cluster. We extracted top-20 clusters, and top two or three words in each cluster and showed the words as CF. CF using Global Information that we submintted is shown in Figure2.4.



Number:303 Hubble Telescope Achievements

Please check the item in consideration of a common meaning to these words.

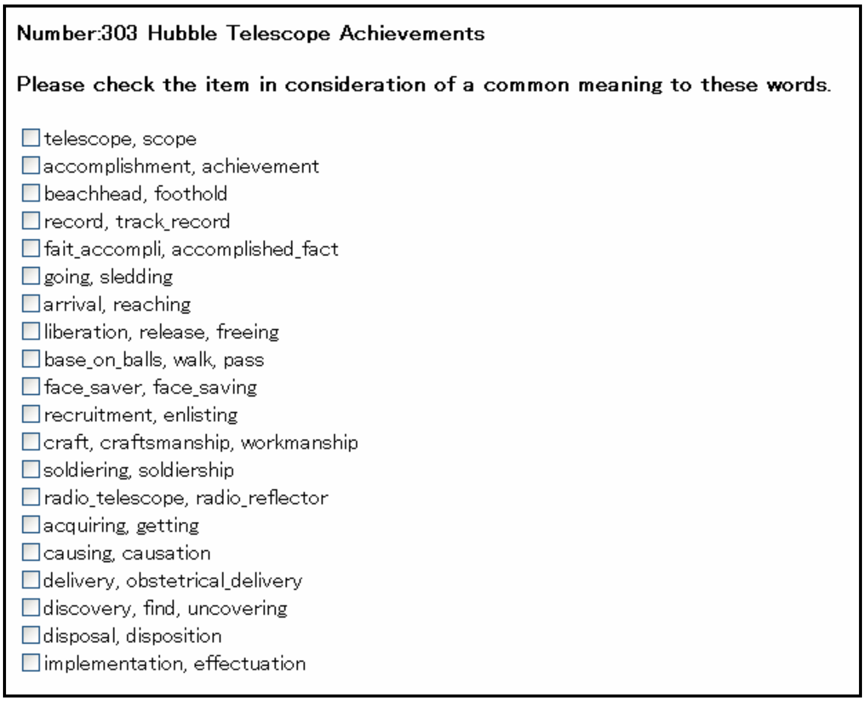
□china, major
□supernova, galaxy
□tremendous, great
□shuttle, space shuttle, astronaut
□hubble space telescope, mauna kea, keck
□gyro, gyroscope
□black hole, milky way, quasar
□computer, bc
□picture, image
□year, decade
□asteroid, planet
□development, economic, cooperation
□past, first
□science, advanced
□career, record
□orbit, launch
□outlook, new, future
□sign, notable, sport
□exhibition, object
□stability, effort

Figure 2.4 CF using Global Information

2.1.3 CF Generation using WordNet

In this section, we tried to expand words in a topic using WordNet, the dictionary manually made. WordNet has nodes, synonym set and tree structure. We attempted to specify the meaning by showing hyponym sets of the synonym set to a user.

Our system extracted hyponym sets of the synonym sets that respond to the words in a topic. Finally, two or three words of 20 hyponym sets were shown to a user as CF. CF using WordNet that we submintted is shown in Figure2.5.



Number:303 Hubble Telescope Achievements

Please check the item in consideration of a common meaning to these words.

□ telescope, scope
□ accomplishment, achievement
□ beachhead, foothold
□ record, track_record
□ fait_accompli, accomplished_fact
□ going, sledding
□ arrival, reaching
□ liberation, release, freeing
□ base_on_balls, walk, pass
□ face_saver, face_saving
□ recruitment, enlisting
□ craft, craftsmanship, workmanship
□ soldiering, soldiership
□ radio_telescope, radio_reflector
□ acquiring, getting
□ causing, causation
□ delivery, obstetrical_delivery
□ discovery, find, uncovering
□ disposal, disposition
□ implementation, effectuation

Figure 2.5 CF using WordNet

2.2 Query Selection

Our system extracted new query from items judged to be truth in CF. Each item corresponds to a cluster or a synonym set. In the following, keyword extraction from the clusters or synonym sets judged to be truth in CF is treated.

2.2.1 Using Local Information

After our system received the CF results, we extracted the important words from clusters judged to truth in CF to express topic well. We defined the formula of a degree how much a word expresses topic as follows.

$$\mathrm{TopicWordScore}(w,\mathit{TrueClusters}) = \prod_{c\in TrueClusters} \mathrm{ClusterWordScore}(w,c)$$

$$\mathrm{ClusterWordScore}(w,c) = \mathrm{ClusterPrbability}(w,c)\cdot \log\frac{\mathrm{ClusterPrbability}(w,c)}{\mathrm{CorpusPrbability}(w)}$$

$$\mathrm{ClusterPrbability}(w,c) = \frac{\mathrm{ClusterTF}(w,c)}{\mathrm{ClusterLength}(c)}$$

$$\mathrm{ClusterTF}(w,c) = \sum_{d\in c}\mathrm{TF}(w,d)$$

$$\mathrm{ClusterLength}(c) = \sum_{d\in c}|d|$$

$$\mathrm{CorpusPrbability}(w) = \frac{\mathrm{CorpusTF}(w)}{\mathrm{CorpusLength}}$$

$$\mathrm{CorpusTF}(w) = \sum_{d\in Corpus}\mathrm{TF}(w,d)$$

$$\mathrm{CorpusLength} = \sum_{d\in Corpus}|d|$$

$\mathrm{TF}(w,d)$ ： TF value of the word w in document d

$|d|$ ： Document length

TrueClusters ： clusters judged to be truth in CF

c ： cluster c

After Topic Word Score was comptered, Top 30 words were used as Query. The word vectors of the words were normalized.

The words in title and in description were also used as Query. The word vectors of the words was TF values of the words in title and in description.

2.2.2 Using Global Information

We extracted the words from CF using Global Information based on two methods. One method is using the words chosen by a user. The other is using information on the chosen clusters. Topic Word score is calculated by product of the elements in the center vectors of chosen clusters. Words whose score were higher than a threshold were adopted as Query.

2.2.3 Using WordNet

We think that Synonym sets judged to be the truth in CF is relevant to the topic. Then, the words in this Synonym set were added to Query.

Table 2.2 HARD Track Result

Run	MAP	R-precision	Improvement (Comparison with BL1)	Detail
MeijiHilBL1	0.1370	0.1966	Baseline	Title, Description
MeijiHilBL2	0.1654	0.2236	Baseline	Title, Description, Narrative
MeijiHilCWE1	0.1855	0.2326	18.3%	Local Information
MeijiHilRW	0.1451	0.2007	2.1%	Global Information (Word)
MeijiHilRC	0.1467	0.1972	0.3%	Global Information (Cluster)
MeijiHilRWC	0.1516	0.2021	2.8%	Global Information (Word + Cluster)
MeijiHilWN	0.1371	0.1993	1.4%	WordNet
MeijiHilMrg	0.1847	0.2312	17.6%	Local Information + Global Information
TREC Baseline Median	0.1901	0.2518		
TREC Median	0.2071	0.2639		

2.3 Result and Discussion

We submitted two baseline runs and seven final runs. We showed the results in Table2.2. Our system used Jakarta Lucene. Query using both Local Information and Global Information is generated from query using Local Information and query using Global Information.

In the run using Local Information, R-precision improved 18.3% in comparison with baseline run (MeijiHilBL1). In the run using Global Information, R-precision improved 2.8%. This is higher than 1.4% in the run using WordNet. So Global Information is more useful than WordNet, the dictionary manually made. In the run using Global Information and Local Information, R-precision improved 17.6%. This is lower than 18.3% in the run using Local Information. Therefore, our system using both Global Information and Local Information can’t get the good performance.

3. Robust Track

This is the first year that our group participates in the Robust Track of the TREC. Here we report our system and a method on the Ad Hoc Task. Our method employs Document Expansion Model.

3.1Document Expansion Model

Everyone characterizes a document with words actually occurring in the text of the document. We use all words related to the document in “Document Expansion Model”. In other words, some words related to a document are added to the document. The association with a document is presented by “word context”. We call the model “Document Expansion Model”. The model is word vector model.

3.2 Document Representation

Document represents words actually occurring in the text of the document and synonymy. Synonymy is described later.

3.2.1 Context

In the paper, we define word sequence to appear before a word “word context”. This is based on a thought of N-gram which occurrence of a word depends on only words just before. Figure3.1 shows an example of word context as we used four-word sequence. As Yarowsky suggested that local ambiguities need only a window of 3, or 4. It is general to use conditional probability to make a guess at next word from an appearance of word sequence. We use Mutual Information. The reason is we got good experimental result for relation between a word and word sequence just before. Document is characterized at word context. We call a word occurring in the document with the word context “original word”

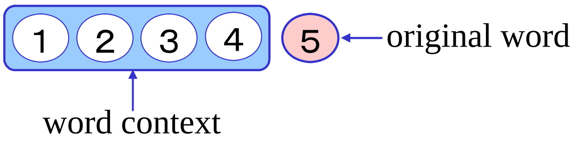


Figure 3.1 word context

3.2.2 Synonymy

Synonymy in an accurate meaning is such as America, U.S.A. ,United States of America, and U.S.. We define “Synonymy” as a set of words provided from the same word context.

ex)

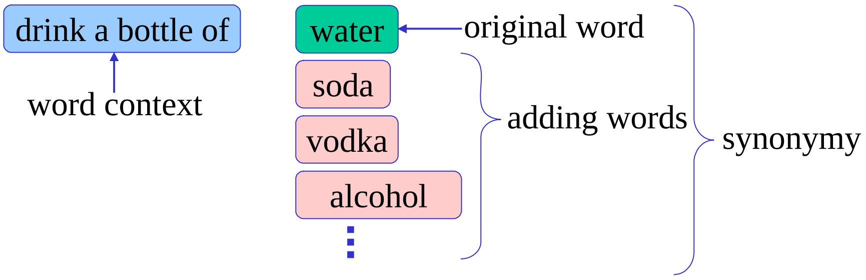


Figure 3.2 Synonymy

3.2.3 Context Weight

It needs to weight word context. Synonymy which word context created, depends on a word actually occurring in the document with the word context. As word context’s weight is original word’s IDF.

3.2.4 Relation of Synonymy and Word Context

Word’s weight is defined by relation between synonymy and word context, and show as follows:

$$Word's\ weight\ (b) = \sum_{a \in A} \log \frac{P(a,b)}{P(a)P(b)} * IDF(c)$$

a: a word context , A: a set of word context in a document

b: a word of synonymy , c: a original word

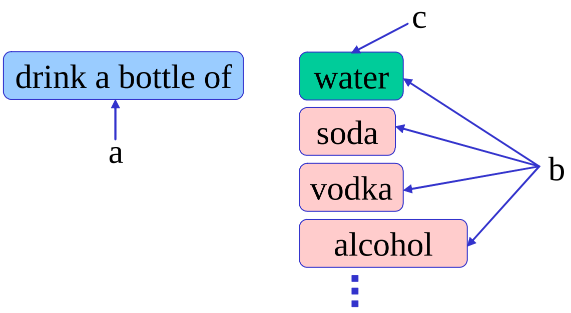


Figure 3.3

We use this score to characterize a document.

3.3 System

Our system is illustrated in Figure 3.4.

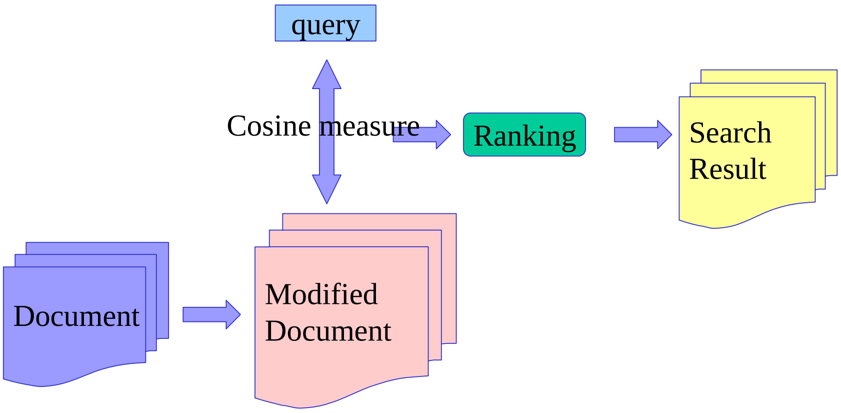


Figure 3.4 System Outline

Reference

I. Olga Vechtomova, Stephen Robertson, Susan Jones, “Query expansion with long-span collocates”, Information Retrieval, Volume 6 Issue 2 , April 2003

II. Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li,Mark D. Smucker, Courtney Wade, “UMass at TREC 2004: Novelty and HARD”, NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)

III. Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, Stephen Robertson, “Microsoft Cambridge at TREC–13: Web and HARD tracks”, NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)

IV. David J Harper, Gheorghe Muresan1, Bicheng Liu, Ivan Koychev, Dietrich Wettschereck, Nirmalie Wiratunga, “The Robert Gordon University’s HARD Track Experiments at TREC 2004”, NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)

V. Kenji Kita, “Computation and Language in Japanese”, 1999

VI. Ide, N. and Véronis, J., “Word Sense Disambiguation: The State of the Art”, Computational Linguistics (24)-1, pp. 1-41, 1998