Dynamic Programming
- A new paradigm for Information Retrieval-

Ryuichi Sawada Kyoji Umemura
Toyohashi University of Technology
ICS Department 1-1 Tempaku Toyohashi Aichi 441-8580 Japan
{sawada, umemura}@avenue.tutics.tut.ac.jp

Abstract

Information Retrieval systems usually use addition to aggregate similarity values of words in documents. In this paper, we propose a dynamic programming method to aggregate the similarity. We have found that this method is useful for Japanese-English cross lingual information retrieval, where the performance suffers from word-sense ambiguity and difficulty of capturing appropriate phrases. This method is especially effective where technical terms play important roles in the documents to be retrieved.

Keyword: Document Processing,Information Retrieval,Technical Term, Edit distance,Dynamic Programming

1 Introduction

IR (Information Retrieval) systems usually regard a document as a set of words. This assumes that the words are usually efficient handle to retrieve documents. This is not always the case for cross lingual IR.

A word in one language may correspond to several words in another language. This will degrade the precision of IR system. In addition to this, an important but new technical term may not exist in dictionary. This may make the system fail to retrieve the document. Moreover, appropriate word boundaries may not be apparent in some languages. Since Japanese does not have any separators between words, it is generally difficult to obtain correct words. This make the situation even worse.

Replacing a Japanese word with a corresponding English word may be regarded as one edit operation. We can define cross lingual distance between the two sequences of words using this operation with appropriate weights. This enables us to treat sequences of words, rather than just one word. Therefore, we propose to use this extended edit distance. We have verified the effectiveness of this method for the retrieval of English technical abstracts by Japanese query.

2 Approach

The technical term is the most important handle by which we retrieve technical documents. Since concepts are newly created everyday, it is hard to maintain a dictionary that includes all of them.

We observe that a technical term in English usually consists of several words. We also observe that the corresponding Japanese technical term usually preserve the order. In fact, 1943 pairs of terms out of 2000 preserve word order by randam sampling from a technical dictionary.

This does not imply that we can automatically generate an English technical term from a Japanese technical term by translating word by word. This is partly because one English word may have several senses, and because there are synonymous words in Japanese. For example, "machine translation system" may correspond to either "機械 翻訳 システム" or "マシントランスレーション システム", where "machine" corresponds to "機械" or "マシン", and "translation" corresponds to "翻訳", " 変換" and "トランスレーション". The dictionary may contain the entry "translation system" but may not always contain the entry of "machine translation system". In addition, the mapping between "translation" and "トランスレーション" may not exist in the dictionary. This is because "トランスレーション" is phonetic

conversion of "translation".

From this observation, we have decide to use an extend edit distance where replacing a Japanese word with a corresponding English word is regarded as unit operation, whereas ordinal edit distance considers one character to replace. The edit distance is usually calculated by DP (Dynamic Programming) method[1]. Our extended distance also uses DP. Therefore, we call our system a DP sytem.

3 Baseline System

We will define a simple baseline system before we define our system. The baseline system is a simple cross lingual IR system. First, the Japanese query string is segmented into Japanese words using existing system called "Chasen"[4]. Then each Japanese word is consulted by the dictionary and the corresponding English words are obtained. Then the document is scanned to determine whether these English words exist in the document.

If one of the words exists in the document, the inverse document frequency of the word is calculated and added to the score of the document. After scanning all of the documents, the baseline system outputs the documents according to the scores.

For example, if the query contains "最大共通部分グラフ", which corresponds to "Largest Common Subgraph", the base system gets the list of the string ("最大", "共通", "部分", "グラフ"), and then try to find words, "greatest, largest, maximum" from "最大", "common" from "共通", "part" from "部分", "graph" from "グラフ". This system does not try to find "subgraph" because "部分グラフ" is not considered.

Let α denote a string. Let $DIC(\alpha)$ denote set of English words corresponds to Japanese string α. Let DOC_k be a list of English word corresponding to kth document. Chasen is modeled as a function Q; $Q(\alpha)$ is list of segmented substrings of α whose elements are estimated as words by "Chasen" system.

The baseline system calculates the following $Score_1$ for each document and query. Then, the system sorts the document according to the $Score_1$.

$$Score_1(\alpha, DOC_k) = \sum_{q \in Q(\alpha)} \sum_{w \in DIC(q) \wedge w \in DOC_k} weight_1(w)$$

where $weight_1(w) = -\log_2(\frac{df(w)}{N})$,
N is the total number is documents,
and $df(w)$ is document frequency of word w.

4 DP system

Our DP system tries to translate every substring of the given query, and find the best way to accumulate the score, preserving the order of words. In other words, it finds best translation alignment between the Japanese query and English document, and calculates the score for this alignment.

This calculation is interpreted as edit distance. We assign value 0.0 to insertion and deletion, and a positive value to replacement. The most similar pair of strings has the most successful replacement between them. Thus, they give the maximum score in this measure. If the pair has no successful replacement, it gives the score of 0.0. It is worth pointing out that obtained value is similarity and not distance. The more simliar is the pair, the larger is the score.

For example, if the query contains "最大共通部分グラフ", which corresponds to "Largest Common Subgraph", the DP system will try to find path using, "greatest, largest, maximum" from "最大", "large, big" from "大", "common" from "共通", "part" from "部分", "graph" from "グラフ", and "subgraph" from "部分グラフ". The figure 1 shows the more detail. We can use all substrings because of the robustness of DP. Since the final score is decided by the best alignment only, Accidental matching of small substring will not affect the final score. This is not the case in simple addition.

Let α denote a string. Let α_{ij} denote substring from the $(i+1)$th character to the jth character. Let $DIC(\alpha)$ denote set of English words corresponding to Japanese string α. Let DOC_k denote a list of English words in the kth document. Let $DOC_k(i,j)$ denote sublist of DOC_k from the $(i+1)$th word to the jth word. The DP system calculates following $Score_2$ for each document, and retrieves the document according to the $Score_2$.

$$Score_2(\alpha, DOC_k) = SIM_{DP}(length(\alpha), length(DOC_k))$$

$$SIM_{DP}(0, n) = 0.0$$
$$SIM_{DP}(m, 0) = 0.0$$

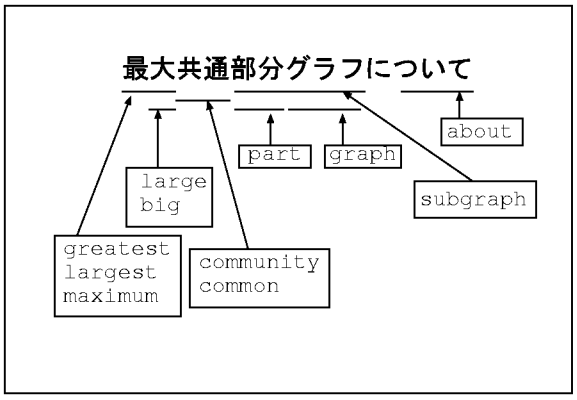

Figure 1: Japanese-English mapping in DP system

$$SIM_{DP}(m,n) =$$
$$MAX(SIM_{DP}(m-1,n),$$
$$SIM_{DP}(m,n-1),$$
$$SIM_{DP}(m-1,n-1),$$
$$SIM_{MATCH}(m,n))$$
$$SIM_{MATCH}(m,n) =$$
$$MAX(weight_2(w) + SIM_{DP}(i,j))$$
$$\forall i \in [0..m]; \forall j \in [0..n];$$
$$\forall w \in DIC(\alpha_{im}) \wedge w = DOC_k(j,n);$$
$$weight_2(w) = -length(w) \cdot \log_2(df(w)/N)$$

5 Experiment

The experiment uses NACSIS[5, 6] data collection. It contains 21 queries and 330,000 documents of technical abstracts with relevance judgments for cross lingual information retrieval. The documents are in English and queries are in Japanese. The base system and DP system are compared using standard 11-point average precision measure.

Our Japanese-English dictionary mainly contains technical terms in the fields of Computer Science, Medicine, Architecture, Electronics, and Chemistry. If there are several corresponding words for one Japanese term, the Japanese word will have plural entries for each different English term. The dictionary provides many to many mapping between Japanese and English. The example of the dictionary is shown in figure 2. The dictionary consists of 579116 entries. Both the baseline and DP systems use the same dictionary.

Table 1 shows the 11-point average precision of both systems for different queries. ID is given by NACSIS, and there are missing numbers. We have

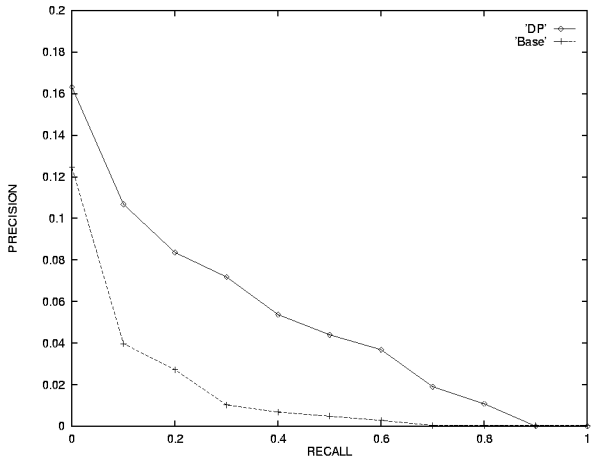

Figure 3: Overall Recall-Precision

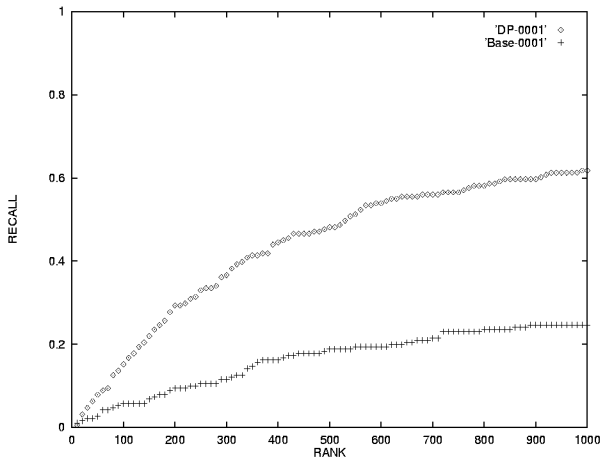

Figure 4: Query1: Autonomous Mobile Robot

excluded Query 8 and query 25 from this table because the keyword in the query is expressed in the English alphabet, and our system does not retrieve any thing because our dictionary does not have any English entry. Though it is easy to add a English entry in Japanese-English dictionary, we have decided not to tune the dictionary, and are reluctant to add English entries to the Japanese dictionary

A query consists of a title and description. The title is one line in length and describes what the query is about. The description is one paragraph in length in natural language. Figure 3 show the overall recall-precision. Figure 4, 5, 6 7 and 8 are samples of Recall-Rank graph. This graph plots the recall rate y for highest rank of x documents.

デジタルデータ交換網	DDX
デジタルデータ交換網	digital date exchange
デジタルデータ通信網	digital data communication network
デジタルプロセス制御	digital process control
デジタルプロセス制御システム	digital process control system
デジタルモデル	digital model
デジタル画像解析	digital image analysis
デジタル画像処理	digital image processing
デジタル画像処理システム	digital image processing system
デジタル解析システム	digital analysis system
デジタル計算機設計	digital computer design
デジタル姿勢制御システム	digital attitude control system
デジタル指令通信システム	DCCS
デジタル指令通信システム	digital command communication system

Figure 2: Entries in Japanese-English Dictionary.

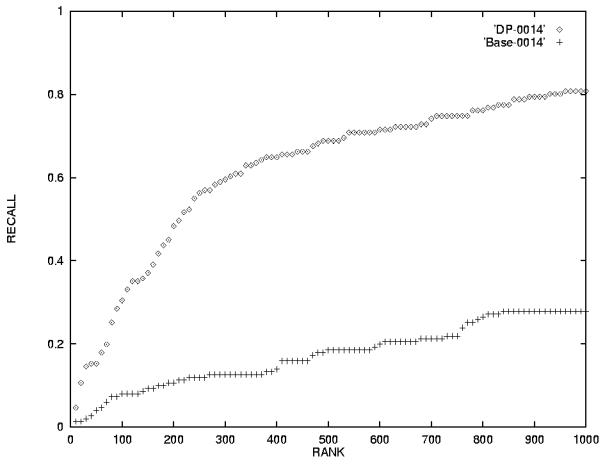

Figure 5: Query14:Fault Diagnostic System

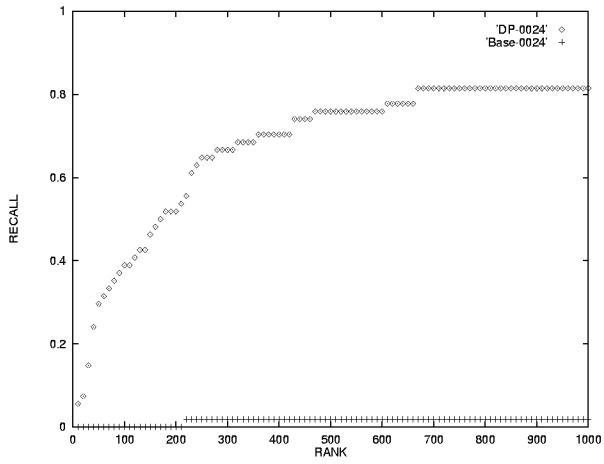

Figure 6: Query14:Machine Translation System

6 Analysis

The overall precision-recall (figure 3) shows that DP is effective. The overall score is rather low because many queries need more processing than the baseline system does. Still, the comparison with DP is meaningful.

The DP system looks up all substrings of a query in the dictionary. It is usually an extreme strategy to improve recall rate at the expense of precision. The DP aggregation takes advantage of this extreme strategy and prevents the disadvantage of this strategy at the same time by selecting only the best alignment. Table 1 also shows that it improves the performance in the order of magnitude when DP method is effective (figure 4, 5, 6).

The performance of the all-substring method can suffer when the keyword contains some unrelated keyword as a substring. It is problematic when the topic is expressed in the Japanese phonetic translation of English, Katakana. “Data Mining Method” (query 12) and “Neural Networks” (query 28) are the examples of this case.

The performance of the DP method also suffers when the query consists of poor keyword. In other word, every word in the topic have small information quantity. In this case, the DP fails to capture even a single word in topic, and fails to retrieve the documents. “Newspaper Article” (query 23) and “Position Measurement of Object ” (query 29, figure 8) are the examples of this

Table 1: 11 point average precision for each query

ID	Subject	DP	Base
1	Autonomous Mobile Robot	0.1477	0.0269
4	Document Image Understanding	0.0082	0.0374
5	Character Dimension Reduction	0.0112	0.0021
6	Intelligent Agent	0.0069	0.0065
10	Automatic Keyword Detection	0.1273	0.0288
12	Data Mining Method	0.0000	0.0107
13	Loop Region Analysis	0.0029	0.0162
14	Fault Diagnostic System	0.3443	0.0236
15	Collocation	0.0136	0.0006
16	Largest Common Subgraph	0.0266	0.0046
18	Quality of Service	0.0143	0.0067
19	Phrase Attachment Analysis	0.0243	0.0017
20	Katakana Loan Word	0.0013	0.0019
22	Knowledge Acquisition Method	0.0769	0.0109
23	Newspaper Article	0.0000	0.0026
24	Machine Translation System	0.1681	0.0001
26	Word Function Grammar	0.0004	0.0000
28	Neural Network	0.0001	0.0028
29	Position Measurement of Object	0.0010	0.0095

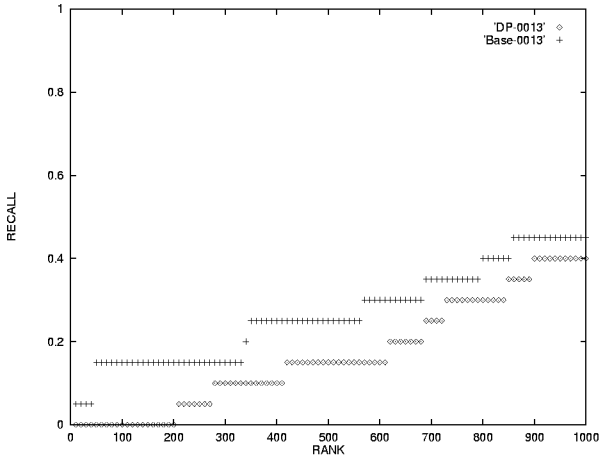

Figure 7: query13:loop region analysis

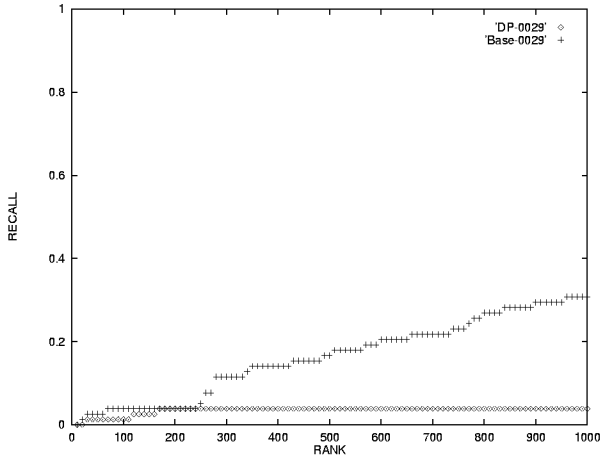

Figure 8: query29:position measurement

case.

When the query consists of independent words (figure 7), the DP system is not effective. The performance decrease is acceptable.

7 Discussion

The several queries in table 1 show that it sometimes fails to work at all. This means that using all substrings in a query is still too extreme a method even for DP. This will need more consideration. In fact, several queries get considerably worse scores than simple baseline system. We need more detailed analysis of the DP system failure.

Query expansion is another way to improve recall rate[3]. This method uses statistical dependency whereas DP method uses textual similarity, that is, heuristics by which new technical terms may be created. If a new technical term is important to retrieve documents, the new term may not have enough occurrence for statistical processing.

In that case, the DP method provides another way to handle the situation.

Improving the dictionary is another way to improve the recall rate of cross lingual information retrieval[2]. However, dictionaries are hard to maintain, especially for new technical terms. It is desirable to have some heuristics for unknown terms. Moreover, an improved dictionary is also benefits the DP system. Therefore we will get the best results using DP and the improved dictionary.

Learning corresponding string pairs from parallel copora [7] is more a language independent approach. Our method is based on the observation of the technical terms in English and Japanese. If the language pair is different, we may have different strategy for DP to calculate natural distance between some other pair of language. However, if the key term is new, we may suffer from the rare occurrence of the term and thus fail to learn the term.

A good dictionary for computer science may behave poorly in chemistry. While the DP method may not work as well as the trained dictionary in some fields, it will work wide range of fields because DP does not contain field specific heuristics. In this sense, DP method is more robust than the trained dictionaries.

The DP method reflects the order in which the substring appears. This information is generally discarded by standard information retrieval systems since it models document as a bag of words. DP method does not use "bag of words" model, and thus is capable of utilizing the order of occurrence. Thus, we might call this method as a new paradigm.

8 Conclusion

In this paper, we have explained DP similarity that aggregates the contribution of all substrings of query. We have verified the effectiveness using a cross lingual information retrieval system. We have explained that this method essentially provides more search space. This system does not use a "bag of words" model, or addition to aggregate the similarity.

Acknowledgement

This research is funded by Nippon Telegraph & Telephone and Sumitomo Electronic Corporation. The patent is being processed with number of 11-176477 in Japan.

References

[1] Petteri Jokinen, Forma Tarhio and Esko Ukkonen:
"A comparison of Approximate String Matching Algorithms". *Software-Practice and Experience*,VOL.26(12),1439-1458 (DECEMBER 1996)

[2] Lisa Ballesteros and W.Bruce Croft:
"Dictionary-based methods for cross-lingual information retrieval". *In Proceedings of the 7th International DEXA Confirence on Database and Expert Systems Applications*,pages 791-801,1996

[3] Lisa Ballesteros and W.Bruce Croft:
"Phrasal Translation and Query Expansion Techniques for Cross-Language Information ". *In Proceedings of SIGIR'97*,Philadelphia PA,USA, pages 84-91

[4] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Osamu Imaichi, and Tomoaki Imamura:
"Japanese Morphological analysis System ChaSen Manual," *NAIST Technical Report, NAIST-IS-TR97007*, February 1997, http: //cactus.aist-nara.ac.jp/ lab/ nlt/ chasen.html.

[5] Kando, N. et al.:
"NTCIR:NACSIS Test Collection Project," *20th Annual Colloquium of BCSIRSG*, Autrans, France, March 25-27, 1997.

[6] Kageura, K. et al.:
"NACSIS Corpus Project for IR and Terminological Research," *Natural Language Proceeding Pacific Rim Symposium'97*, Phuket, Thailand, pp.493-496, December 2-5, 1997.

[7] Lisa Ballesteros and W.Bruce Croft:
"Resolbing Ambiguity for Cross-language Retrieval". *In Proceedings of SIGIR'98*,Melbourne,Australia, pages 64-71