NTCIR : NACSIS Test Collection Project

Noriko Kando, Teruo Koyama, Keizo Oyama, Kyo Kageura, Masaharu Yoshioka,
Toshihiko Nozue, Atsushi Matsumura, Kazuko Kuriyama
R & D Department, NACSIS, 3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-8640, JAPAN
E-Mail: {kando,koyama,oyama,kyo,yoshioka,nozue,atsushi,kuriyama}@rd.nacsis.ac.jp

1. Introduction

In this poster we introduce the project constructing NTCIR (NACSIS Test Collection for evaluation of Information Retrieval systems) currently carried out at the National Center for Science Information Systems (NACSIS), Japan[1]. In the following, the backgrounds and the aims of the project are firstly introduced, followed by the specification of the test collection. We then discuss some technical problems concerning the construction of the collection.

2. Background and Aims

The project is intended to provide sound infrastructure to evaluate the search effectiveness of information retrieval systems with Japanese language and facilitate the IR research with Japanese language and cross-lingual retrieval including Japanese.

The project is also motivated by the recognition of the following situations:

(1) Needs for a standard Japanese test collections

(2) Need for cross-lingual retrieval

(3) Need for the variety in text types

(4) Need for the various components in a documents

(5) Need for the fundamental data for research into the intersection of IR and NLP

The importance of the large-scale standard test collection in IR research are widely recognised. Stopping, stemming and query analysis are language depended procedures. Especially indexing texts written in Japanese or other East Asian languages like Chinese or Korean are quite different from those with English, French or other European languages since there is no explicit boundary (i.e. no space) between words in a sentence. Regarding other East Asian languages, there is large-scale test collection of TREC Chinese Collection and KORDIC Collection for Korean language, which contains ca.50,000 documents. For Japanese, there is only one standard test collection called BMIR-J2, which has just published in March, 1998 and consists of 5,080 Japanese newspaper articles and ca.60 queries. There are

[1] This is a part of the bigger research project "A Study on Ubiquitous Information Systems for Utilization of Highly Distributed Information Resources" supported by the Japan Society for the Promotion of Science (JSPS).

still acute needs for enhancement of the collection in the aspects of the variety of text types and the scale.

Need for cross-lingual retrieval is acute in the internet environment. Moreover in the scientific texts, foreign language terms, sentences, or abstracts are often appeared in a Japanese text in their original spelling. Therefore cross-linguistic strategies should be used not only for retrieval of the non-equivalent multilingual database but also retrieval of Japanese scientific documents [Kando, 1997]. Therefore we need a test-collection which can be used for cross-lingual retrieval and consists of scientific texts.

In order to respond the needs stated above, we aim to construct a large scale test collection which is also usable for cross-linguistic retrieval and application of NLP to IR.

3. The specifications of the collection

The Collection contains 300,000 exactly translated Japanese-English paired documents, 100 search topics for each subject domain, and relevance assessment for each search topic.

3.1 The Document

The documents are mainly abstracts of conference papers, and relatively small number of fulltext of research papers. Abstract records of conference papers are derived from NACSIS's Conference Paper Database. The fulltext record are also from NACSIS's Fulltext Database of Scientific Papers. Because we are still negociating with the publishers of the journals, we can not specify the exact number of fulltext records can be included in the collection. The format of a document record is SGML tagged plain text. An abstract record consists of ID number, title (Japanese (J) & Englisih (E)), name(s) of authors (J & E), name of conference (J), date of conference, hosting organization (J & E), abstract (segmented into paragraphs; J & E) and manually assigned keywords (J & E). In addition to this, a fulltext record contains tags for section, section heading, subsection, and so on. A example is shown below;

```
<REC>
・
<ACCN>0000010360</ACCN>
・
<TITL TYPE="kanji"><TITL.ORIG>機械翻訳における構造変換の干渉について
</TITL.ORIG></TITL>
・
<TITE TYPE="alpha">Interaction between Structural Changes in Machine
Translation</TITE>
<AUTH TYPE="kanji">木下 聡 / 辻井 潤一</AUPK>
<AUTE TYPE="alpha">Kinoshita,Satoshi / John,Phillips / Tsujii,Jun-ichi</AUPE>
<CONF TYPE="kanji"><CONF.ORIG>研究発表会(自然言語処理)</CONF.ORIG></CONF>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<ABST TYPE="kanji"><ABST.P><ABST.P.ORIG>語い項目によ
```

って引き起される特異的(idiosyncratic)な構造変換を規則の形で記述すること自体は、それほど困難ではない。しかし、そのような構造変換が単一の文に複数個同時に存在すると、それらの干渉によって予期せぬ問題を引き起こす。さらに、等位接続構造のような一般的な言語現象と組み合わさった場合にも、問題を引き起こすのである。本報告では、機械翻訳におけるトランスファーの枠組として、原言語側の言語構造を変更することなしに目標言語側の構造を作り出す非破壊的(non-destructive)な処理モデルを提案し、その枠組の下で、構造変換の干渉によって引き起こされる問題を解決するための手段を示す。
</ABST.P.ORIG></ABST.P></ABST>
<ABSE TYPE="alpha"><ABSE.P>This paper discusses complex structural changes during transfer in machine translation with a non-destructive transfer framework. Though the description of each individual idiosyncratic structural change, which is mainly caused by lexical items, is not difficult, special provision must be made when they are combined, because interaction between them sometimes causes unexpected problems. Transfer of coordinate structures is also discussed as this sometimes necessitates a structural change and interacts with other structural changes in a problematic way. We give solutions to this problem in our logic-based transfer model.</ABSE.P></ABSE>
<KYWD TYPE="kanji"><KYWD.ORIG>自然言語処理 // 機械翻訳 // 論理 // トランスファー</KYWD.ORIG></KYWD>
<KYWE TYPE="alpha">Natural Language Processing // Machine Translation // Logic // Transfer</KYWE>
<SOCN TYPE="kanji"><SOCN.ORIG>情報処理学会</SOCN.ORIG></SOCN>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>
<SER>0001</SER>
<TXTL>ENG</TXTL>
</REC>

3.2 Topics

Search topics are collected from users, and the analysts can rewrite them to make them more clear and objective. Format of the topics is similar to the one once used in TREC and contains SGML-like tags. A topic consists of title of the topic, description, detailed narrative, and list of concepts. Each narrative may contain detailed explanation of the topic, term definition, background knowledge, purpose of the search, expected number of relevant documents, preference in text types, criteria of relevance judgement, and so on. They may be used as a profile of user's information need in a specific situation to evaluate interactive retrieval systems. The lists of concepts may be used for administrative purpose only. The example of the topic was translated into English and shown below.

Examples of topics with English translation

ex. 1:

<topic>
<title>bibliometrics</title>
<description>

<J>計量書誌学分野で、標本データでは観察されない研究者あるいは論文の
扱いに付いて論じたものはないか。</J>
<E>Are there any documents of bibliometrics that deal with the proper
treatment of the types which are unseen in the given sample? </E>
</description>
<narrative>
<J>収集された標本データについての属性を統計的に論じたものは多いが、
計量書誌学（研究者と論文の関係、論文と雑誌の関係を扱っているものが
よい）では、実際に標本には現れないデータをいかにして扱えばよいかを
論じたものがあるのかどうか知りたい。ロトカの法則やブラッドフォード
の法則の、母集団モデルへの展開を考えたものがあるといちばん良い。</J>
<E>There are many studies that deal with the mathematical structure of
given samples in the field of bibliometrics. I wonder whether there
are studies that deal with the unseen types. Especially, I would
like to know the extention of Lotka's law or Bradford's law to the
theoretical population.</E>
</narrative>
<concepts>
<J>計量書誌学　未知データ　母集団　標本　ロトカの法則
ブラッドフォードの法則</J>
<E>bibliometrics, unseen types, population, sample, Lotka's law
Bradford's law</E>
</concepts>
</topic>

ex.2:

<topic>
<title>complex nouns</title>
<description>
<J>複合名詞解析において、シンボリックな手法と統計的な手法を組み合わせた
アプローチを取る研究はないか。</J>
<E>Are there any research concerning the automatic anaylsis of complex nouns
using both statistical and symbolic method together?</E>
</description>
<narrative>
<J>複合名詞の解析であれば、一応、要求に半分くらいレレバントと考える。言語
は日本語でもその他の言語でもよいが、「解析」としては、単なる分割だけでなく、
構造付与まで行っている必要がある。また、「複合名詞」の中で漢字列部分のみ、
カタカナ列のみ、というのはどちらでもよい。</J>
<E>Any work on the automatic analysis of complex nouns is half relevant to
my need. The target languages may be any, e.g. English, Japanese, etc.
But the analysis should cover parsing (bracketing), and the decomposition
is not sufficient.</E>
</narrative>
<concepts>
<J>複合名詞　言語処理　統計的手法　記号的手法　パージング　構造解析</J>
<E>complex nouns, natural language processing, statistical method, symbolic
method, parsing, structural analysis</E>

</concepts>
</topic>

3.3 Relevance assessment

The relevance assessment is done in three grades, i.e., relevant, partially relevant, non-relevant. The top ranked documents in the search results of various search strategies on content based IR systems and by searchers are pooled and form a set of candidates of relevant documents. The human analysts of each subject domains reviews the candidate set and assess the relevance of each document in it. The same analyst who create or rewrite the topic assess the relevance.

3.4. Linguistic Analysis

A part of the collection contains detailed part-of-speech tags [Kageura et al, 1997; Koyama et al, 1998]. Because of absence of explicit boundary between words in Japanese sentences, we set the three levels of lexical boundaries (i.e., word boundary, strong morpheme boundary, and week morpheme boundary), and assigned detailed POS tags based on the boundaries and types of origin, so that the collection can be used to examine the suitable term segmentation of Japanese texts for retrieval purpose.

Example of Lexical Boundaries

w ベクトル m 空間 m モデル w に w 基づく w 情報 m 検索 m システム w は...
vector space model (of) based information retrieval system (-) ...

(whereas: w : word boundary m : morpheme boundary)

4. Summary

This paper introduced the on-going project to construct a large-scale English-Japanese bilingual test collection. It has tried to follow the tradition of standard test collection for IR systems evaluation and at the same time it aims to applicable to different settings, for example, IR with Japanese language, retrieval of scientific documents, cross-lingual retrieval, construction of tools or vocabularies for cross-lingual retrieval, combination of manually assigned keywords and content-based IR, interactive IR systems, terminological processing, and terminological research, and so on.

For the further information about the project, please contact ***Noriko Kando*** on kando@rd.nacsis.ac.jp. Any discussion, comments, and leads about the project are greatly appreciated.

Reference

Kageura, K. et al. “NACSIS Corpus Project for IR and Terminological Research”. Natural Language Processing Pacific Rim Symposium '97, 2-5 December, 1997 at Phuket, Thailand, p. 493-496. (available from http://www.rd.nacsis.ac.jp/~kyo/)

Kando, N. “Cross-Linguistic Scholarly Information Transfer and Database Service in Japan”. Panel on “Multilingual Database” in the 1997 Anuual Meeting of the American Society for Information Science, 1-5 November, 1997 at Washington,D.C, U.S.A.

Koyama, T. et al. “The Construction of a Lexically Motivated Corpus : The Problem with Defining Lexical Units.” [to appear in] The First International Conference on Language Resources and Evaluation. May, 1998, Granada, Spain. (please see http://cerec.ugr.es/~rubio/elra.html)