

Second International Symposium
on Universal Communication

December 15, 2008

Overcoming the Language Barrier

Makoto Nagao

National Diet Library

Language and Linguistics

Two Approaches to Linguistics

- (1) Discovery of rules in language expressions
 - phoneme, morpheme, syntax, semantics, pragmatics
 - limit of acquisition of large scale language data by manpower
- (2) Investigation of language generation mechanism
 - human brain competence of language generation does not come out from language data analysis (Chomsky)

Linguistics as A Natural Science

- (1) Huge amount of speech and text data can be accumulated in a computer database, and linguistic structure of a language will be clarified by the computer analysis of these data.
- (2) Semantics of a word, a phrase and a sentence will be clarified by automatic clustering of a word or a sentence in context.
- (3) From a linguistic theory constructed by human instinct to a linguistic theory by objective analysis of large scale language data.

Linguistic Resources

Linguistic Resources as A Source of Linguistic knowledge

- (1) Brown Corpus (1967, one million words)
- (2) The first dictionary based on a linguistic corpus was
“Longman Dictionary of Contemporary English (1978)”
- (3) This dictionary has a lot of collocational examples, which shows how a word is used in a sentence.
(A word cannot be used by reading semantic description alone)

Longman English-Japanese Dictionary

- (1) Corpus of National Institute of Japanese Language and National Institute of Information and Communications Technologies are used.
- (2) Text corpus : 40 million words (academic books 10 million, fiction books 10 million, newspapers 10 million, journals 10 million)
- (3) Speech corpus : 10 million words
- (4) Corpus analysis for the dictionary was done by Prof. Kurohashi who is a leading researcher in natural language processing.

Importance of A Balanced Corpus

- (1) Care must be taken about time and space extention and kind of community, when linguistic data are collected.
(synchronic and diachronic aspects).
- (2) Care must be taken about kind of text, such as literatures, poems, articles, journals, male/female, age,····
- (3) How much data must be collected depends on what kind of analysis is intended.

Corpus with Associated Information

- (1) Text corpus with word segmentation and parts of speech
- (2) Text corpus with syntactic structure for each sentence.
- (3) multilingual parallel corpus
- (4) speech corpus with corresponding texts
- (5) corpus analysis software

Multiple Language Corpus

- Parallel corpus
with links of corresponding words and phrases between multiple languages
- Parallel corpus
with corresponding syntactic structures between multiple languages
- Comparable corpus

Speech Corpus

- Small speech corpora of Japanese language are available at several research groups. These corpora must be enlarged.
- Consortium for speech resources were established in 2006 in Japan.
- Speech recognition system is going to be introduced in the House of Representatives, which rely heavily on the speech corpus.

Machine Translation (MT)

Machine Translation Methods

- (1) syntactic MT (Rule-based MT)
(RBMT)
- (2) Example-based MT (EBMT)
- (3) Statistical MT (SMT)
- (4) Combined system of the above three methods.
- (5) Speech translation system

Rule-based MT (RBMT)

- (1) The basic unit of translation is a word (word dictionary)
- (2) Three grammars : analysis grammar, transfer grammar, generation grammar, are to be prepared.
- (3) A lot of grammar rules are required, and the improvement of a grammar becomes quite difficult.
- (4) The analysis, for example, produces a lot of syntactic structures for a sentence.

Example-based MT (EBMT)

- (1) The basic unit of translation is a phrase (example) (database of example translations).
- (2) A grammar requires only a small set of basic grammar rules, because the analysis and synthesis are done by phrase unit.
- (3) Huge amount of example translations (example dictionary) is required, but the multiple meanings or ambiguity of meaning of a word can be avoided by the context of a phrase.
- (4) Improvement of an MT system can be done simply by increasing example translations.

Statistical MT (SMT)

- (1) This method can be regarded as a system which acquires example phrases by statistical method from a huge amount of parallel corpus.
- (2) As the grammatical knowledge is not used in the extraction of a phrase, word sequences with grammatically no meanings are often extracted, and therefore the quality of translation is not so good.

Improvement of Grammar Rules

- (1) Analysis of a huge amount of sentences by a set of basic grammar rules or by the principle of dependency will lead to the improvement and the refinement of a grammar system. The repetition of this process will produce better grammar system.
- (2) This recursive process will also clarify word and phrase usages.

Construction of Example Dictionary

- (1) Example dictionary with translation (EBMT dictionary) can be obtained by the analysis of bilingual parallel corpus.
- (2) By the analysis of comparable text corpus, paraphrasing expressions will be obtained. These expressions will be valuable for the automatic construction of word and phrase dictionaries.
- (3) By cascading the paraphrasing, an original expression will become more simplified, and will be able to traverse from an original language to another, thus achieving the machine translation.

Automatic Construction of a Thesaurus

- (1) By the analysis of dictionary texts which include synonyms, upper/lower concept words, a thesaurus will be formulated.
- (2) By the analysis of texts in a special field, a terminology dictionary and an ontology of the field will be formulated.

Problems in MT

- (1) Development of multilingual MT system
- (2) Development of language grid system (by cascading MT from A language to B, and from B language to C, an MT from A language to C is achieved. By cascading MT systems of different languages in the form of a grid, new MT systems can be realized)
- (3) Improvement of an MT system.
- (4) Construction of ultra huge example dictionaries.

Problems in Information Retrieval

Information Retrieval

search field by keywords	text portion to be retrieved
bibliographic information part	full text
full text	Paragraphs/sentences which include keywords
Table of contents of a book/article	Text section pointed by the table of contents which is matched by keywords

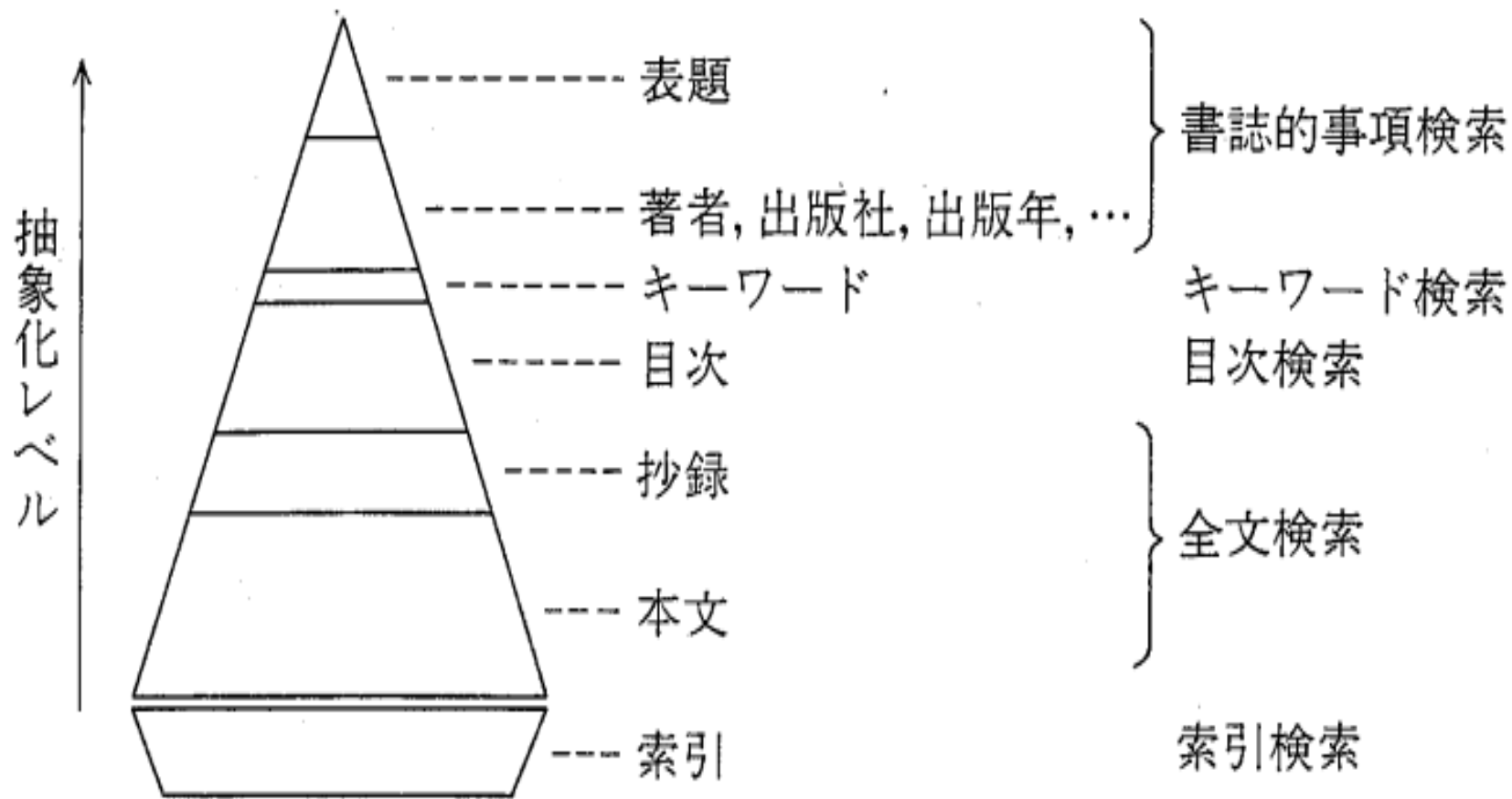


図 14 図書構造と検索対象

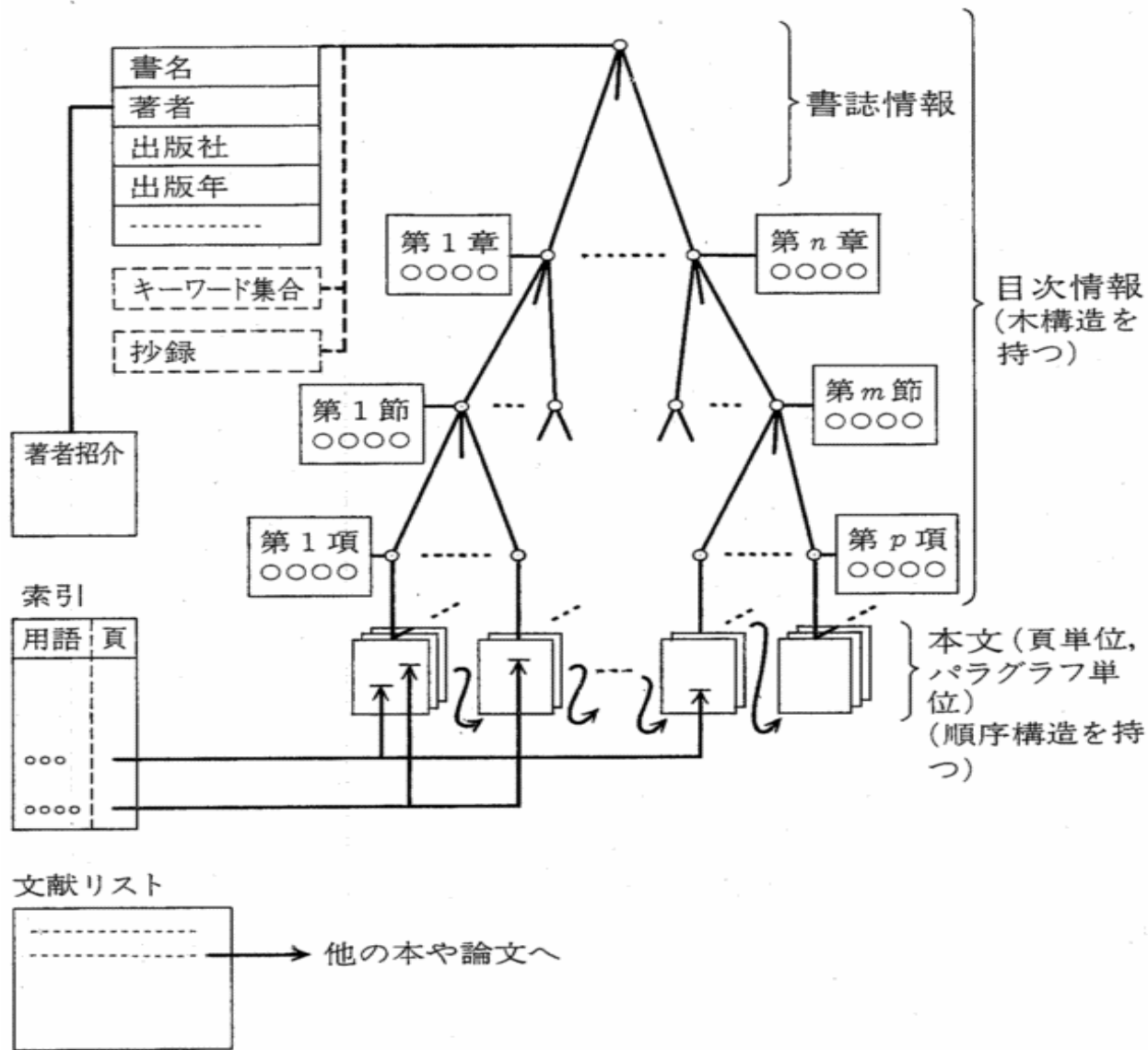


図8 図書の構造

Deconstruction and Resynthesis of Text

- Structuring of a text according to the table of contents.
- Retrieve text sections pointed by the table of contents from different books, and connect them from a new aspect, and create a new book.

Text Summarization

- (1) Extracting, abstracting
- (2) Summarize multiple text contents
- (3) Clustering similar texts, and summarize them into one.
- (4) Summarization from different aspects
- (5) Length of summary

Reliability and Credibility of Information

- (1) Estimation of reliability and credibility of retrieved Internet information is important.
- (2) How reliable is the information which comes to the top of Google search?
- (3) It is important to check the existence of opposing information in the long tail of the retrieval.
- (4) It possible, credibility check is recommended to the retrieved information by referring the solid academic knowledge or well-known facts.

Detection of Dangerous or Criminal Information

- (1) Warning of suicide, murder, attacks, on the Internet must be detected immediately so that such accidents will be prevented.
- (2) Detection of suspicious information must be detected and tracking of the information in time sequence must be performed.
- (3) Detection of a burst of similar information must be detected immediately.

Dialogue System

- (1) Dialogue between user and system is unavoidable to understand real intension of a user for information acquisition.
- (2) System must have knowledge and current information so that it can understand the intension of a user.
- (3) System must estimate the knowledge, intension and so on of a user to give back suitable response.

Development of language technology is essential in the future information society (apology for non-inclusion of speech technology).