Indexing languages

6.2.6. Natural language

Overview

Wouldn't it be wonderful if we could ask computers for information as easily as we ask people? And if the computers could understand our requests and retrieve relevant documents at just our level of knowledge? Obviously that is the ultimate—and some say unattainable—goal of natural-language information retrieval. It means IR systems would have to be able to:

- Interpret users' needs as expressed in free text.
- Represent the complete range of meaning conveyed in documents.
- Recognize a match between users' needs and the documents that meet them.

Natural language is the linguistic expressions people use in oral and written communications. As we know from everyday life, natural language is full of ambiguities. Sentences are often incomplete, with meanings that must be inferred within a larger context. Natural language is constantly changing in vocabulary, usage, and meanings. The challenges of natural-language indexing are enormous.

Natural-language indexing includes all vocabulary that is not controlled. Indexers use terms existing in a document,* any of its parts, any surrogate for the document, and sometimes in the indexers' own minds. Basically . . .

- The indexing language of each system is based on terms in its documents.
- Any term can function as an index term.
- Terms can be in anywhere in a full-text document or its metadata record.
- Composite subjects are postcoordinated (combined using boolean operators during searching).

*"Document" can refer to any text or nontext information object. Here it refers to target documents, not search queries.
This module explains how natural language works in indexing and searching. The three problem areas are the same as for controlled-vocabulary indexing:

1. **Naming single concepts**: What is the best term for a given concept?
2. **Showing relationships among single concepts**: What concept is related to a given concept?
3. **Showing relationships among multiple concepts**: What if the subject of a document contains two concepts?

Unlike situations with controlled vocabulary, however, searchers are largely responsible for figuring out the solutions. In addition to human solutions, this module describes two computerized approaches: automated indexing and natural-language processing (NLP).

---

### 1. Naming single concepts

**Problems**

What is the best term for a given concept? How does one choose among variant word forms for a concept?

**Solutions**

The searcher can . . . for example . . .

- Think of all spelling variants. online, on-line, on line
- Think of morphological variants. toxic, toxicity, toxicology, hypotoxic
- Truncate (use wildcard symbol: *, ?, etc.). toxic*, wom*n, _esthetic
- Consult a thesaurus for ideas.

### 2. Showing relationships among single concepts

**Problems**

What concept is related to a given concept? How is it related?

**Solutions**

These are semantic problems: determining alternatives for equivalent, hierarchical, and associative relationships. The searcher can . . .

- Think of synonyms.
- Think of broader and narrower concept terms.
- Think of related terms.
- Consult a thesaurus for ideas.
3. Showing relationships among multiple concepts

Problems

What if the subject of a document contains two concepts? What if it contains more than two concepts?

Solutions

The searcher can . . .

- Rearrange word orders (syntax).
  
  violence schools
  
  schools violence

- Use phrase matching.
  
  "violence in public schools"

- Use boolean operators (AND, OR, NOT).
  
  violence AND schools

- Use proximity operators (NEAR, ADJACENT, FAR, BEFORE).
  
  violence 5N schools
  
  (terms are within five words of each other)

- Consult a thesaurus for ideas.

Note: IR systems vary as to whether or how they use quotation marks and operators.

Computerized approaches

Automated indexing

Automated indexing is computerized creation of indexes based on natural language in documents. Software exists to analyze any unit from title to full text and extract single words. This is used to create back-of-the book indexes, periodical title indexes, and other kinds of indexes. Such software has existed since the advent of computers after World War II and, although it is now quite complex in design, its principles are simple:

- Count and sort words.
- Use stop list that ignores articles, conjunctions, prepositions, etc.
- Use stemming to reduce word variants to meaningful roots.
- Parse phrases for syntax.
- Identify words that are important enough to use, based on their frequency of occurrence.

An advantage of automated indexing is that it evolves as documents are added to the system. A disadvantage is that the computer has no real understanding of the concepts, and human labor is still required to refine the results.
Natural-language processing (NLP)

Natural-language processing (NLP) is directed toward the development of IR systems that can communicate with people as naturally as people communicate with each other. This area of research is closely related to research in artificial intelligence (AI). These systems are extremely complex and still do not approach the subtleties of normal human conversation. However, researchers have made much progress in recent years with sophisticated software that relies on analysis of natural-language texts at seven linguistic levels (Liddy, 1998):

1. Phonological: speech sounds
2. Morphological: parts of words, including roots, prefixes, suffixes
3. Lexical: whole words and their parts of speech
4. Syntactic: sentence order and grammar
5. Semantic: sentence meaning
6. Discourse: paragraph (and larger unit) meaning
7. Pragmatic: situational uses of language

Levels 1 through 4, syntactic analysis, are relatively easy for today's computers to perform. These levels of analysis are performed by the automatic indexing systems above and many other applications including Web search engines.

Levels 5 through 7 are progressively more difficult for computers to perform, as the size of the units of analysis and linguistic variability increases. The 7th, pragmatic level, according to Liddy, concerns the "purposeful use of language in situations, particularly those aspects of language which require world knowledge." She says her own DR-LINK (Document Retrieval using LINguistic Knowledge), to which she provides Web links, incorporates this highest level of natural-language processing.

Does this mean that the ultimate goal of natural-language IR is now within reach? Well, it's getting closer, but it still works best in narrow subject domains where the natural language is relatively consistent.

Natural language terminology

Speaking of natural language, the professional jargon associated with natural language is in itself naturally confusing! Here's some clarification:

- **Uncontrolled vocabulary** is rarely used to refer to natural language.
- **Keyword** usually refers to a natural-language search term but it may refer to a controlled-vocabulary search term. Caution!—usage varies from system to system!
- **Full-text searching** means that all/most of document text is available for searching using natural language. (Graphics, captions, symbols, etc. may not be searchable.)
- **Free-text searching** means that any term can be entered to search any field in bibliographic record or all/most of full-text document.
- **Free indexing language** is occasionally used to indicate that all or part of documents or surrogates are available for searching electronically using natural language.
Summary

In natural-language indexing, indexers assign terms to fields in database records based on their own best judgment. They must understand terms that authors use (literary warrant) and terms searchers are likely to use (user warrant) and be able to translate author terms into user terms if necessary.

In order to retrieve information successfully, searchers must try to think like both authors and indexers. This puts an equal, if not greater, burden on searchers, who like indexers must . . .

- have excellent language skills
- be knowledgeable in the subject domain
- be skilled in using the IR system

That's a lot to expect of many end users!

To add to the challenges of natural language, consider that many users rely exclusively on the World Wide Web, where most full-text documents are not stored in databases and have no accompanying metadata at all. Instead of being frustrated with results from search engines, perhaps we should be impressed by how remarkably successful they are!

You may want to compare the solutions in this module with those in the module on controlled vocabulary.

Cites & sites