Making sense out of searching

Stuart L. Pook
Jason Catlett

Basser Department of Computer Science
University of Sydney

ABSTRACT

The trouble with text retrieval using keywords is that English words are imprecise: there may be many words for the entity the user wants, and each of these may also have other (unintended) meanings. To get closer to the dream of ‘retrieve what I mean, not what I say,’ systems need to take into account the sense in which each word is used in the text stored, and to get from the user a more precise and complete statement of the concept sought. Recent research on machine-readable dictionaries and thesauri may soon make this a reality.

Published in the Online Information Conference, Sydney, Australia, 1988. This version printed 15 July 1998.

1. Introduction

Large databases of text containing information such as scientific papers, legal cases, newspaper archives and library catalogues have been common for many years. Such databases are typically too large to allow a user to scan the entire collection of text in search of an interesting item. For this reason more practicable and faster methods of retrieving pieces of text have been devised.

One of these methods is called keyword retrieval. In this method a user specifies a keyword (or perhaps a boolean combination of keywords) to be used in retrieval. For example a simple query such as `bridge` would retrieve all articles containing the word ‘bridge’. A boolean combination such as `juice and not (apple or orange)` might be used to find articles that mention ‘juice’ but are not concerned with fruit juices.

The standard method of text retrieval by keyword is fraught with difficulties caused by the imprecise nature of words. Section 2 explains and attempts to solve these problems using the information contained in machine-readable dictionaries and thesauri.

Other methods of retrieval from a large database allow the user to retrieve articles that concern topics of interest to the user. For example a user might want all the articles that are concerned with physiology. Previous implementations of this method of text retrieval have required the text to be classified by subject before this process can be used. This classification normally has to be performed manually. Section 3 gives a method by which text classification can be automated.

2. Text retrieval by keyword

Existing text retrieval systems allow the user to specify a list of keywords of interest. Each piece of text in the database is checked to see if it contains one of the keywords specified. If a keyword is found the piece of text is retrieved.

This method is easy to use and, by using algorithms such as hash tables and inverted indices, efficient to implement. However, keyword retrieval suffers from the restriction that an exact match between the keyword and a word in the text is required for retrieval. If the user is attempting to retrieve text concerned with road works, specifying the keyword *bridge* will only find articles that explicitly mention ‘bridge’. Articles that contain the words ‘walkway’, ‘pontoon’ or ‘footbridge’ will not be retrieved even though these are close synonyms of the keyword.

Keyword retrieval also suffers from the problem of retrieval of irrelevant articles that contain a keyword used in a different sense to that intended by the user. Someone who wants to know about road works and uses the word *bridge* will not be interested in an article relating to the card game. The new text retrieval system described in this section of the paper attempts to use the information contained in dictionaries and thesauri to solve the two problems outlined above. This new method is based on two algorithms: one that performs sense disambiguation in natural text; and one to perform dictionary to thesaurus sense matching. The way that these two algorithms can be used in text retrieval is outlined in the next two sections, followed by details on the algorithms.

2.1 How to retrieve text

The method used to retrieve text described here consists of three parts: the selection of the keywords by the user, the processing of the text to be searched, and the selection of relevant pieces of text.

2.1.1 Keyword selection

In a standard keyword retrieval program the user enters words to be matched against words in the text. In this new system the user is also prompted to obtain information on the intended sense of each keyword specified. There are two possible ways of obtaining this information. The first and preferred way is to look up the word provided by the user in the thesaurus, displaying the entries for the various senses of the word. The user can then specify which of the senses best captures the intended usage of the word. For example the entries for *juice* are:

alcohol, electricity, fuel, liquid, nitty-gritty, secretion, vigour.

The user can then indicate whether the sense meaning electricity or the sense meaning fruit juice is required. Using this method allows the user to expand the range of searching by including words at a higher level of the thesaurus by simply choosing a sense and asking for the region to be enlarged to include more remote synonyms of the keyword. Once this is done the system has a thesaurus paragraph number to be used in retrieval.

When using the second method the user is presented with the dictionary definition of the keyword and is asked which of the indicated senses is intended. This sense could then converted into a thesaurus paragraph or matched directly against occurences of that keyword in the stored texts.

2.1.2 Text Processing

Most text archives available are distributed by some organisation. For example press agencies distribute news stories to paying subscribers. To use this new method of text retrieval the distributor will have to run a computer program to preprocess the text. It will determine the sense of each non-function word in the text using Algorithm 1. As in the 'Keyword selection' section this dictionary sense has a corresponding thesaurus paragraph which is determined using Algorithm 2. This thesaurus paragraph has a unique number. Once this process is complete each article of text has a list of thesaurus paragraph numbers which are stored and transmitted with the article. These numbers can then be used for retrieval as follows.

2.1.3 Retrieval

When a user is retrieving text, either interactively or in a batch procedure, each paragraph number generated by the user is compared against those associated with each article. If a match occurs the article is retrieved and presented to the user. Should this method produce too many incorrectly retrieved articles the user can specify more than one keyword for each topic of interest and require that an article be selected only if some number of paragraph numbers match.

2.2 Example

Suppose an ASIO agent wishes to look up or retrieve articles concerned with entrapment. He would enter the keyword *entrap.* The system presents him with the various senses for 'entrap' listed in the thesaurus. He would then choose the sense containing the following synonyms:

> endanger, compromise, entrap, expose, imperil, jeopardise, peril, put the skids under.

The following article would be retrieved:

> Soviet foreign ministry security agents showed off some of the bugging devices they say were discovered in Soviet diplomatic missions throughout the U.S. Ivan Miroshkin of the Soviet foreign ministry security service said that several bugs with connections to radio transmitters were uncovered. The presentation was designed to counter U.S. charges of Soviet spying. The Soviets displayed photographs and devices they called 'Violations of their sovereign territory.' They said the devices were taken out of the Soviet residence in New York City, the mission in Washington and consulate in San Francisco. The Soviets did not say if any of their secrets had been compromised by the presence of the listening devices.

When this article was processed the word 'compromise' had its sense determined using Algorithm 1. The correct sense was:

> **7.** *Mil.* to subject (classified material) to the risk of passing to an unauthorised person.

as opposed to the more common meaning of making mutual concessions. Algorithm 2 was then used to convert this word sense into a thesaurus parargraph. This was the same as the one chosen by the agent, so the article was retrieved.

3. Text retrieval by classification

Text retrieval by classification is an alternative to text retrieval by keyword. Keyword retrieval searches for individual words in the text that match a particular pattern. One particular word occurring in the text is enough to conclude that an article should be retrieved. Text retrieval by classification takes a more global approach. The entire text of an article is used to determine the subject matter of an article and the user is then able to retrieve articles that are about a subject of interest. As the entire text is used this works best on databases where each article is on one topic. Newswire stories are a good example of such text.

3.1 Previous Work

A program called FORCE4 [3, 2] has been implemented by Donald Walker and Robert Amsler of Bell Communications Research. This program performs text retrieval by classification. It uses the Longmans Dictionary of Contemporary English (LDOCE). This dictionary includes the special feature of subject codes for those word senses that are indicative of a particular subject area. For example the word ‘wind’ has the following subject codes for its various senses that apply to a particular subject area [3 p 76]:

ML	meteorology
DZP	physiology
MU	music
NA	nautical
HFZH	hunting

The FORCE4 program uses these subject codes to classify pieces of text. This is done quite simply by looking up each non-function word[1] in an article in the LDOCE. A record is kept of each subject code that is attached to a sense of this word. After the whole article has been processed, the subject whose code appears most frequently is deemed to be the subject of the article. Using the LDOCE articles can be classified by FORCE4 as being about, for example, *meteorology* or *physiology*. The successful use of this method requires a dictionary with a meaningful set of subject codes. A simple extension to FORCE4 would be to attempt to determine the sense of each word in the text and only use the subject code associated with that sense, rather than recording the subject codes given

[1] A function word adds meaning to or shows the relationships between content words in a sentence. Two examples are *the* and *of*.

for all the senses of the word. The sense disambiguation would be carried out using Algorithm 1. As the Basser Department of Computer Science does not have access to the LDOCE this modified method could not be tested.

3.2 Classification

Classification of text can be accomplished with a standard dictionary without subject codes if a thesaurus is available to supply subject groupings. The resulting subject classifications will depend on the subjects under which the words in the thesaurus are classified. Best performance will be obtained by using a thesaurus that reflects the subject classifications that a user would be expected to employ. Text can be classified by looking up each word of the text in the thesaurus and remembering under which thesaurus paragraphs it falls. The subject of the thesaurus paragraph that occurred most frequently is deemed to be the subject of the article.

A standard thesaurus, such as the Macquarie, has several levels of classification. These group words into small sets of synonyms and larger groups of loosely related words. We have not yet determined which of the four levels of classification is best suited for retrieval purposes.

This method can, in a similar way to FORCE4, be modified to use a dictionary to determine the sense in which a word is being used in text. Once the sense is known Algorithm 2 can be used to map the word into a single thesaurus paragraph. Finding only one paragraph number for each word leads to better accuracy in classification.

Another extension that would help to solve the problem of misclassification is to assume that articles can have more than one subject. This can be achieved by saying that the subjects of a piece of text are those subjects whose frequency of occurrence is within some percentage (specified by the user) of the most common subject.

3.3 Retrieval

Once all the pieces of text in a collection have been classified a user can retrieve articles about any desired subject by simply specifying the required subject.

4. Algorithms

The two algorithms described here are based on work by Michael Lesk of Bell Communications Research [1]. He describes how counting word overlaps between dictionary definitions can be used to determine the sense of words as they appear in natural text.

4.1 Algorithm 1: Sense disambiguation

Algorithm 1 determines the dictionary definition of a word (called the *target* word) in a piece of natural text. This is accomplished by looking up the definitions of the words surrounding the target word. These definitions are formed into one list of words. The definition of the target word is looked up in the dictionary and separated into a list of words for each sense. Then the number of word overlaps between 1) the list of words for each sense and, 2) the large list of words from the other definitions, is computed. The sense with the largest number of overlaps is deemed to be the sense in which the target word is being used in this piece of text.

This algorithm was tested on the clause ‘all hands to reef topsails’. The aim is to determine in what sense the word ‘reef’ is being used. The dictionary definitions of ‘reef’ are:

> **reef**
> 1 *n.* **1.** a narrow ridge of rocks or sand, often of coral debris, at or near the surface of water.
> 2 **2.** *Mining.* a lode or vein.
> **reef**
> 3 *n.* **1.** a part of a sail which is rolled and tied down to reduce the area exposed to the wind.
> 4 −*v.t.* **2.** to shorten (sail) by tying in one or more reefs.
> 5 **3.** to reduce the length of (a topmast, a bowsprit, etc.), as by lowering, sliding inboard, or the like.
> 6 −*v.i.* **4.** (of a horse) to throw its head up, thereby pulling against the reins.
> **reef**
> 7 *v.t. Colloq.* **1.** to remove, usu. by force (fol. by *out*).
> 8 **2.** to steal (fol. by *off*).

Algorithm 1 choose sense 4, ‘to shorten (sail) by tying in one or more reefs’, which is the correct sense for this use of the word reef.

4.2 Algorithm 2: Dictionary to thesaurus sense matching

Algorithm 2 matches a word sense in a dictionary to the corresponding section in a thesaurus. For example, the sense ‘petrol, fuel, oil, etc., used to run an engine’ of the word *juice* corresponds to the section of the thesaurus that contains ‘fuel, combustible, feed, juice’. This algorithm allows a dictionary word sense to be converted into a list of synonyms in a thesaurus.

Dictionary to thesaurus sense matching is carried out in a similar way to sense disambiguation. The target word is looked up in the thesaurus. If it does not appear in the thesaurus the algorithm reports an error. Should the target word appear only once there is only one choice for the matching thesaurus paragraph. This choice will be correct if the thesaurus and dictionary are consistent.

When the target word appears more than once in the thesaurus the algorithm must decide which of the lists of synonyms found contains the target word used in the sense required. Each of the lists of synonyms has the target word removed. Next, each of the remaining words in the lists is looked up in a dictionary. The definitions for each of the words in a synonym list are combined together to give a long list of words. This process results in a long list of words corresponding to each occurrence of the target word in the thesaurus. The text of the dictionary sense of the target word is then compared with each of the word lists. The thesaurus entry corresponding to the word list with the most overlaps is assumed to contain a list of synonyms of the word sense.

An example of this algorithm on the word *standard* is now presented. The word sense that is to be matched with a thesaurus section is:

> **11.** a flag, emblematic figure, or other object raised on a pole to indicate the rallying point of an army, fleet, etc.

The various possibilities that exist in the thesaurus are:

> average, classic, control group, ethic, flag, flower, organ, fossil fuel, musical piece, rank, standard (model), standard (money), standard (rule), ordinary, standard.

The algorithm chose:

> **flag**, banderol, banner, bannerette, burgee, dogvane ensign, fanion, gonfalon, guidon, hoist, jack, labarum, pennant, pennon, standard, streamer, vexillum.

That is the correct thesaurus section for this sense of the word *standard*.

5. Conclusion

This paper has described two new methods of performing retrieval from large bodies of text. These methods offer the promise of more efficient and successful retrieval of text.

Both the methods consist of two parts: the two underlying algorithms that perform sense disambiguation and dictionary to thesaurus sense matching; and a way of using these algorithms effectively. These two parts are both important to the success of the overall system. Improvements in both are necessary to produce a commercially viable system.

6. References

1. Lesk, Michael E., *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone,* Bell Communications Research, Morristown NJ (1986).

2. Walker, Donald E., ‘‘Knowledge Resource Tools for Accessing Large Text Files,’’ TR-85-21233-25, Bell Communications Research, Morristown NJ.

3. Walker, Donald E. and Robert A. Amsler, ‘‘The Use of Machine-Readable Dictionaries in Sublanguage Analysis,’’ pp. 69-83 in *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, ed. Ralph Grishman and Richard Kittredge, Lawrence Erlbaum Associates, Hillsdale NJ (1986).