How can a word be disambiguated in a set of documents: using recursive Lesk to select relevant records

Diego Chavarro* Yuxian Liu
diego.chavarro@sussex.ac.uk
SPRU, School of Business, Management and Economics, University of Sussex, Falmer, Brighton, BN1 9SL (UK)

Introduction

Retrieving data precisely is the first step to make a bibliometric analysis. Unfortunately, this may turn out to be one of the most difficult steps in an investigation. We denote by precision ratio (PR) the fraction of retrieved items that are relevant. Similarly, the recall ratio (RR) is the fraction of relevant items that are retrieved.

$$PR = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{retrieved\ items\}|}$$

$$RR = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{revelant\ items\}|}$$

Normally a high recall ratio brings a considerable number of items that are irrelevant. A high precision ratio normally implies that a number of relevant items are missed. PR and RR may depend on the words used in the search string. If the words used in a search string are specified, then the PR will be high. The RR will also be 100 percent if no other words but only those specified words in thesauruses are used to express the concept we want to retrieve. But natural language is complex: one word may have different meanings and different words may represent the same concept. Moreover, during the development of a concept its name may change over time.

This can be seen through an example. We consider the search for Human Epidermal Growth Factor 2. Its acronym is HER 2. It was found by several groups, each group naming this gene in a different way. Shih, Padhy, Murray, and Weinberg (1981) identified this kind of gene as a result of transfection studies with DNA from chemically induced rat neuroglioblastomas. Schechter et al (1984) called this gene neu, Coussens (1985) named the gene they isolated as Her-2; Semba, Kamata, Toyoshima, Yamamoto (1985) called it C-erbB-2. Later, C-erbB-2 and Neu, and Her-2 were revealed to be same (Coussens,1985; Schechter et al 1985; Fukushige et al 1985). Consequently, Slamon, Clark, Wong, Levin, Ullrich, and McGuire (1987) called this gene Her-2/Neu, which is nowadays the prevalent term to refer to this gene. In many cases scientists just use the term "her 2", "Her 2", "Her-2", "HER 2".

However, the search results from these different spellings in the web of science make no differentiation between the gene and other uses of the word. “Her 2” can be used as in “her 2 children” which has nothing to do with Human Epidermal Growth Factor 2. Worse, children can be replaced by any noun. Between her 2 and the noun, any adjective can be added in between. In addition, since the Web of Science (WoS) ignores all punctuations, any punctuation can be added in between. Also one item that has “her 2 children” does not necessarily mean it is not what we need. Even the articles dealing with Human Epidermal Growth Factor 2 don’t exclude the expression “her 2 children”. These configurations make it very difficult for us to formulate an effective search string.

It is said that using a controlled vocabulary would help to disambiguate the meaning of the word. However, since the development of science is uncontrollable, the vocabulary of science is also uncontrollable. Hence, a controlled system cannot be used. Medline is a system that combines a controlled vocabulary and natural language. However, this system can't give us a precise result. Using the terms from the knowledge domain as descriptors can specify these words that have general meaning, which will help to achieve high accuracy and completeness. However, we cannot pick up all the terms in one specific domain. Moreover, some terms are not specific enough to remove noise.
Then how can we develop our search strategy so that we can select as many relevant items as possible?

Data set

Firstly we use TS="her 2" to retrieve the data in the web of science, recalling 8,542 items. Among these items we exclude which is definitely related to the bio-marker Epidermal growth factor receptor 2. We then have 2,146 items that we cannot judge if they are related to the bio-marker her 2 or not. We tested these 2,146 records manually and found out that there are 98 items that are not related to her 2 bio-markers. The other 1,873 items are related to the her 2 bio-marker. In this paper we develop an algorithm that will be used to try to separate the relevant items from the irrelevant ones automatically. We will check the precision ratio and recall ratio of our program.

Using recursive Lesk and keywords distance metrics to disambiguate the meaning of a word

The Lesk algorithm is a classical algorithm for word sense disambiguation introduced by Lesk (1986). The Lesk algorithm is based on the assumption that words in a given "neighbourhood" tend to share a common topic. A simplified version of the Lesk algorithm is to compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood. Our idea of a recursive lesk is to increase the terms in the dictionary after each similarity calculation. The next iteration we calculate the similarity between the unclassified articles and the new augmented dictionary. After all calculations, we get two sets: the items that have a high similarity in their topics (we denote this set as set S) and the other set that are distant (we denote this set as set D).

Applying the recursive lesk algorithm (rLesk) to compare its recall and precision to manual identification

The lesk algorithm relies on a dictionary composed of a set of words that define a term (A). Based on this dictionary, the algorithm classifies a given text (B(i)) according to its similarity with the dictionary. The algorithm iterates over a corpus and grows the dictionary in each iteration by adding the records that meet a certain percentage of string similarity. The general procedure is as follows:

1. load the dictionary with a starting definition of the term (A).
2. load the corpus with the records to classify (B).
3. for each record in B
   3.1. remove stopwords from record B(i).
   3.2. clean string B(i): colon, non printable characters, other non-significant characters.
   3.3. split string B(i) into an array of words
   3.4. For each word in array of words
      3.4.1. check if word is in dictionary
      3.4.2. count number of words that matched dictionary
      3.4.3. calculate percentage of words that matched the dictionary: number of matched words / number of words in string B(i)
      3.4.4. If percentage >= threshold, classify record as relevant; Otherwise continue.
4. Add identified records to the dictionary
5. start again, until reaching max. number of iterations.

the dictionary definition of Her 2

In order to identify the records related to Her 2 as a bio-marker, the following paper which has been cited more than 6,000 times was used as a seed and a dictionary.

> Slamon, Clark, Wong, Levin, Ullrich, and McGuire (1987). Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2lneu Oncogene. Science. 235. 177-181, 235.

We compare the precision rate of the records that picked up by the algorithm with what we have manually confirmed.

Table 1: The precision ratio of algorithm

	algorithm yes	algorithm no
Manually- yes	0.56	0.44
Manually no	0.06	0.94

We can see our algorithm is not very accurate finding the records that are related to the bio-marker her 2. But it is very accurate on the records that are not related to the bio-marker her 2. It is not unexpected, since our algorithm is based on the similarity of topics and there are different aspects that deal with the bio-marker her 2. In addition, we started only with one article as the seed to define the dictionary. So, the records that are picked up by our algorithm form just one aspect of her 2 that is related to the correlation of relapse and survival with amplification of her-2 oncogene. The other records concerning the medicine cannot be picked up by our algorithm, but could be increased if we add more articles to the seed at the first stage.

After running the algorithm we use VoSviewer to see how many aspects the research topic her 2 have concerned. And then select representative articles from each aspect.

Discussion

Including different articles in the seed we should pick up all articles in different aspects of a research topic. We also can scrutinize how these different aspects are related with each other so that we can understand how a research topic is recognized. After several loops, we have got some words that are related to the research topic. We can use these words as descriptors to specify these words that have general meaning, which will help to achieve high accuracy and completeness.

Future work

This is only the first stage in a more comprehensive test of algorithms to disambiguate words in data gathering for research. Other techniques such as naïve Bayes, neural networks, and support vector machines will be used in later stages to compare their advantages and disadvantages. We believe that the automation of this process can help to ensure the accuracy of research datasets in situations in which big datasets are used. Producing a reliable algorithm to disambiguate words will ensure the validity of the data. This is important because the results used as evidence in scientific papers and policy documents are key to make decisions that have an impact on society.

References

Coussens L. et al., (1985).Sacc230. 1132

Downward J. et al., (1984).Nature. 307, 521

Schechter A. L.et al., (1984).Nature. 312, 513

Semba K., Kamata N., Toyoshima K., YamamotoT. (1985). Proc. Natl. Acad. Sci. U.S.A.82, 6497.

Shih C., Padhy L., Murray M., Weinberg, R. A. (1981). Nature. 290, 261

http://en.wikipedia.org/wiki/Lesk_algorithm

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, pages 24-26, New York, NY, USA. ACM.