Terminology Extraction from Textbook

Yusi Fan

uOttawa
L'Université canadienne
Canada's university

Introduction

The number of digital collection is increasing with the development of Internet. A method to Automatically extract concepts from digital collection is becoming more and more important

C-value/NC-value [1] is one of the method to automatically extract multi-words terms from corpus using Automatic Term Recognition(ATR). The objective of this project is to implement a Java program using C-value/NC-value to Automatically extract important concepts from Textbooks.

Tagging

Use part-of-speech tagger to tag the corpus.

Before Tagging

Janathan and Tina are building an outdoor Skating rink.

After tagging

Jonathan and Tina are building an outdoor skating rink.
NNP CC NNP VBP VBG DT JJ NN NN .

Linguistic filter

Use linguistic filter to extract words and phrase from corpus.

1. Noun+Noun,
2. (Adj|Noun)+Noun
3. ((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun,

extractor

Rate of change
Instantaneous rate
Function
Graph
Graphing calculator
Parent function
Average rate
Slope of the tangent
…

extracted phrases

Stoplist

Stoplist is a list of words that should be avoided when extracting phrases from corpus.

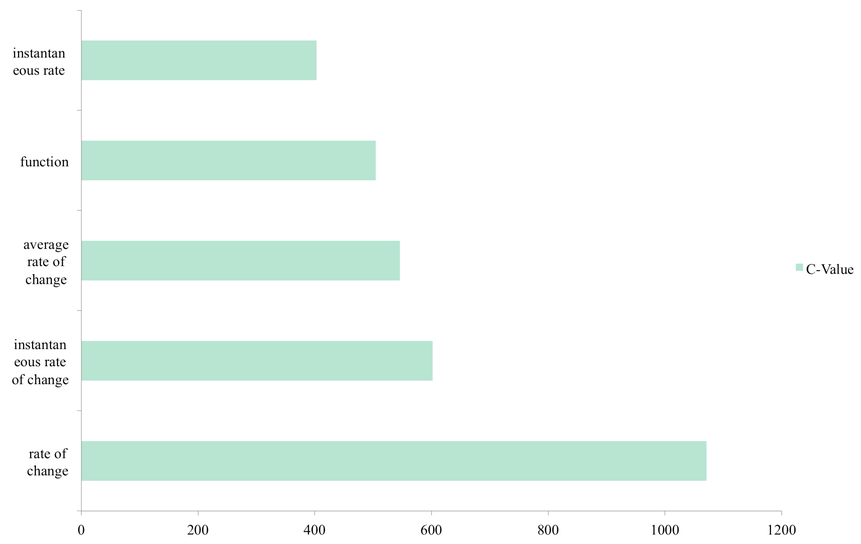
a,able,about,above,according,are,around…
be,became,because,become,been,before…
can, can't, cannot, cant, cause, could…
did, didn't, do, does, doing, don't, done…
each, edu, eg, e.g., either, else, etc…
far, few, for, from, furthermore…
had, hadn't, happen, hardly, has, hasn't…
I'd, I'll, ie, i.e., if, in, indeed, indicate…

C-value

$$C\text{-}value(a) = \begin{cases} log_2|a| \cdot f(a) & a \text{ is not nested,} \\ log_2|a|(f(a) - \frac{1}{P(T_a)} \sum_{b \epsilon T_a} f(b)) & otherwise \end{cases} \quad (3)$$

where
a is the candidate string,
$f(.)$ is its frequency of occurrence in the corpus,
T_a is the set of extracted candidate terms that contain a,
$P(T_a)$ is the number of these candidate terms.

[1]



Future work (NC-value)

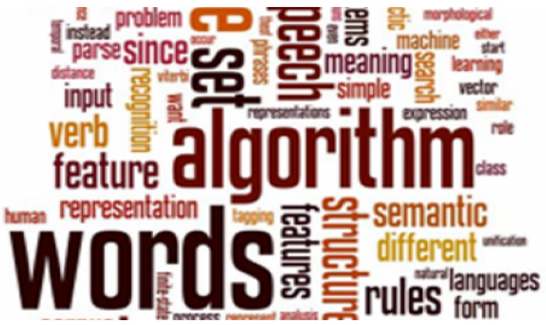
NC-value is used to assigned weight to context words of candidate terms. It will enhance accuracy of C-value.

Conclusion

We know that, the higher C-value candidate term has , the more likely it is a important term in corpus.

We test C-value with a high school math textbook.
In the result of our experiment, rate of change has the Highest C-value, which followed by instantaneous rate of change, average rate of change, function, instantaneous rate.

With the help of Automatic Term Recognition(ATR), we could help a student improve his/her learning process, especially when the learning is done in English as a second language.



[2]

Acknowledgements

[1] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3:115-130, 2000

[2] Coursera. *Natural language processing.* Available from: www.coursera.org (February 3 2013)

Thank you to Professor Diana Inkpen for her supervision and to the UROP project for this opportunity.

Poster made by: Yusi Fan (yfan094@uottawa.ca)