Semantic textual similarity

Kristijan Biščanić, Karlo Dumbović, Nino Jagar

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{kristijan.biscanic,karlo.dumbovic,nino.jagar}@fer.hr

Abstract

This paper focuses on building the system that can assess the semantic similarity between two texts. Each text consists of only one sentence, so more sofisticated approach should be used than just simple word overlap. The sentences are first being preprocessed and then specific features are extracted. The features are being used to train support vector regression model. The model outputs similarity score which is then compared with human similarity judgements. Performance of the model is being evaluated using Pearson and Spearman correlation coefficients.

1. Introduction

Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between them is based on likeness of their meaning or semantic content as opposed to similarity which can be estimated regarding their syntactical representation. We are going to build a system that can measure the semantic similarity between two sentences whose context is previously unknown and of no matter to us. The idea of building such system comes from the necessity of automated similarity estimating in a number of study fields such as biomedical informatics, geoinformatics, linguistics and natural language processing (NLP).

The main task of the system is to compute certain features which act as a representation of sentence similarity. The features are being computed on preprocessed sentences. Preprocessing is done kako?... Exact features used are: ngram overlap, wordnet-augmented overlap, weighted word overlap, vector-space similarity, 2 types of normalized differences, shallow NERC and numbers overlap. Calculation of the features will be briefly described in forthcoming sections. These features are used as an input to a support vector regression model. The model is being hyperoptimized by using grid search over different values of regularization factor, nesto i nesto. On it's output, the model gives us the similarity judgement based on the features extracted. The similarity is scored as a real number from 0 to 5, where 0 represents no similarity, while 5 represents maximum similarity. Model estimation is being compared with human judgements, and the accuracy of the model is measured using Pearson and Spearman correlation coefficients.

2. Overview of the field

3. Description of the model

This is the second section. In scientific papers this is usually (but not necessarily) the section in which related research is (briefly) described.

3.1. Preprocessing

This is a subsection of the second section.

Figure 1: This is the figure caption. Full sentences should be followed with a dot. The caption should be placed *below* the figure. Caption should be short; details should be explained in the text.

3.2. Second section

This is the second subsection of the second section. Referencing the (sub)sections in text is performed as follows: "in Section 2.1. we have shown . . . ".

3.2.1. Sub-subsection example

This is a sub-subsection. If possible, it is better to avoid sub-subsections.

4. Extent of the paper

The paper should have at least. The paper should have a minimum of 3 and a maximum of 5 pages plus an additional page for references.

5. Figures and tables

5.1. Figures

Here is an example on how to include figures in the paper. Figures are included in LaTeXcode immediately *after* the text in which these figures are referenced. Allow LaTeXto place the figure where it believes is best (usually on top of the page of at the position where you would not place the figure). Figures are references as follows: "Figure 1 shows . . . ". Use tilda (˜) to prevent separation between the word "Figure" and its enumeration.

5.2. Tables

There are two types of tables: narrow tables that fit into one column and a wide table that spreads over both columns.

Table 1: This is the caption of the table. Table captions should be placed *above* the table.

Heading1	Heading2
One	First row text
Two	Second row text
Three	Third row text
	Fourth row text

5.2.1. Narrow tables

An example of the narrows table is the Table 1. Do not use vertical lines in tables – vertical tables have no effect and they make tables visually less attractive.

5.3. Wide tables

Table 2 is an example of a wide table that spreads across both columns. The same can be done for wide figures that should spread across the whole width of the page.

6. Math expressions and formulas

Math expressions and formulas that appear within the sentence should be writen inside the so-called *inline* math environment: $2 + 3$, $\sqrt{16}$, $h(x) = \mathbf{1}(\theta_1 x_1 + \theta_0 > 0)$. Larger expressions and formulas (e.g., equations) should be written in the so-called *displayed* math environment:

$$b_k^{(i)} = \begin{cases} 1 & \text{ako } k = \text{argmin}_j \|\mathbf{x}^{(i)} - \mu_j\| \\ 0 & \text{inače} \end{cases}$$

Math expressions which you reference in the text should be written inside the *equation* environment:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} b_k^{(i)} \|\mathbf{x}^{(i)} - \mu_k\|^2 \tag{1}$$

Now you can reference equation (1). If the paragraphs continues right after the formula

$$f(x) = x^2 + \varepsilon \tag{2}$$

like this one does, then use the command *noindent* after the equation to prevent the indentation of the row starting the paragraph.

Multiletter words in the math environment should be written inside the command *mathit*, otherwise LaTeXwill insert spacing between the letters to denote the multicplication of values denoted by symbols. For example, compare $Consistent(h, \mathcal{D})$ and
$\mathit{Consistent}(h, \mathcal{D})$.

If you need a math symbol, but you don't know the command for it in LaTeX, try *Detexify*.[1]

7. Referencing literature

References to other publications should be written in brackets with the last name of the first author and the year of publication, e.g., (Chomsky, 1973). Multiple references are written in sequence, one after another, separated by semicolon and without whitespaces in between, e.g., (Chomsky, 1973; Chave, 1964; Feigl, 1958). References are typically written at the end of the sentence and necessarily before the sentence punctuation.

If the publication is authored by more than author, only the name of the first author is written, after which abbreviation *et al.*, meaning *et alia*, i.e., and others is written as in (Johnson et al., 1976). If the publication is authored by only two authors, then the last names of both authors are written (Johnson and Howells, 1974).

If the name of the author is incorporated into the text of the sentence, it should be out of the brackets (only the year should be in the brackets). E.g., "Chomsky (1973) suggested that ...". The difference is whether you reference the publication or the author who wrote it.

The list of all literature references is given alphabetically at the end of the paper. The form of the reference depends on the type of the bibliographic unit: conference papers, (Chave, 1964), books (Butcher, 1981), journal articles (Howells, 1951), doctoral dissertations (Croft, 1978) and book chapters (Feigl, 1958).

All of this is produced for you automatically by using BibTeX. Sve ovo dobivate automatski ako. In the file `tar2014.bib` insert the BibTeX entries, and then reference them via their symbolic names.

8. Conclusion

Conclusion is the last enumerated section of the paper. Conclusion should not exceed half of the column and is typically be split into 2–3 paragraphs.

Acknowledgements

If suited, before inserting the literature references you can include the Acknowledgements section in order to thank those who helped you in any way to deliver the paper, but are not co-authors of the paper.

References

Judith Butcher. 1981. *Copy-editing*. Cambridge University Press, 2nd edition.

K. E. Chave. 1964. Skeletal durability and preservation. In J. Imbrie and N. Newel, editors, *Approaches to paleoecology*, pages 377–87, New York. Wiley.

N. Chomsky. 1973. Conditions on transformations. In S. R. Anderson and P. Kiparsky, editors, *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

W. B. Croft. 1978. *Organizing and searching large files of document descriptions*. Ph.D. thesis, Cambridge University.

F. Feigl, 1958. *Spot Tests in Organic Analysis*, chapter 6. Publisher publisher, 5th edition.

W. W. Howells. 1951. Factors of human physique. *American Journal of Physical Anthropology*, 9:159–192.

G. B. Johnson and W. W. Howells. 1974. Title title title title title title title title title title. *Journal journal journal*.

G. B. Johnson, W. W. Howells, and A. N. Other. 1976. Title title title title title title title title title title. *Journal journal journal*.

[1] `http://detexify.kirelabs.org/`

Table 2: Wide-table caption

Heading1	Heading2	Heading3
A	A very long text, longer that the width of a single column	128
B	A very long text, longer that the width of a single column	3123
C	A very long text, longer that the width of a single column	-32