Duplicate Question Identification in Quora

Marko Opačić, Javier Salvatierra

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
autor1@xxx.hr, {autor2,autor3}@zz.com

Abstract

This document provides the instructions on formatting the TAR system description paper in LaTeX. This is where you write the abstract (i.e., summary) of the work you carried out within the project. The abstract is a paragraph of text ranging between 70 and 150 words.

1. Introduction

This section is the introduction to your paper. Introduction should not be too elaborate, as that is what other sections are for (the Introduction should definitely not spill over to the second page).

This is the second paragraph of the introduction. In LaTeX, paragraphs are separated by inserting an empty line in between them. Avoid very large paragraphs (larger than half of the page height), but also avoid tiny paragraphs (e.g., one-sentence paragraphs).

2. Related Works

This topic has already been worked on previously in 2017 as a part of a Kaggle competition where the aim of the participants was to predict which of the provided pairs of questions contained two questions with the same meaning attending to their semantic. Therefore, we should take in account the model provided by the winners of the competition:

2.1. Siamese LSTM

As a variant of recurrent neural networks (LSTM), it is an approach that compares the similarity between pairs of data sequences. It employs two identical LSTMs that process two input sequences simultaneously and then compares their final vector representations to determine their similarity.

3. Data

For the task of duplicate question identification based on their semantic similarity, we have used a dataset from the prior Kaggle competition from 2017 which handled the same topic.

3.1. Quora public dataset

The Quora duplicate questions public dataset contains 404k pairs of Quora questions, which wehave divided in three sets: a training set with 70% of the questions, a validation set with a 20% and a set with the last 10% of questions for comparing the predicted results with the actual labels.

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.

4. Models

4.1. Sentence BERT

Sentence-BERT (SBERT) is a method for creating semantically meaningful sentence embeddings (vector representations) using transformer-based models like BERT. It fine-tunes BERT to generate fixed-size representations of input sentences that capture their semantic meaning. SBERT is trained using a siamese or triplet network architecture, where it learns to map similar sentences closer together in the embedding space.

5. Experimental Setup

6. Results

7. Conclusion

8. Reference

8.1. Tables

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

8.1.1. Narrow tables

Table 1 is an example of a narrow table. Do not use vertical lines in tables – vertical tables have no effect and they make tables visually less attractive. We recommend using *booktabs* package for nicer tables.

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

8.2. 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.

9. Math expressions and formulas

Math expressions and formulas that appear within the sentence should be written 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:

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

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 paragraph continues right after the formula

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

like this one does, use the command *noindent* after the equation to remove the indentation of the row.

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

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

10. 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 one 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 not be in the brackets (only the year should be there). 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 automatically produced when using BibTeX. Insert all the BibTeX entries into the file `tar2023.bib`, and then reference them via their symbolic names.

11. Conclusion

Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

Acknowledgements

If suitable, you can include the *Acknowledgements* section before inserting the literature references 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