Varanalysis@SV-Ident 2022: Variable Detection and Disambiguation Based on Semantic Similarity

Alica Hövelmeyer, Yavuz Selim Kartal[*]
GESIS – Leibniz Institute for the Social Sciences, Germany
{alica.hoevelmeyer,yavuzselim.kartal}@gesis.org

Abstract

This paper describes an approach to the *SV-Ident* Shared Task which requires the detection and disambiguation of survey variables in sentences taken from social science publications. It deals with both subtasks as problems of semantic textual similarity (STS) and relies on the use of sentence transformers. Sentences and variables are examined for semantic similarity for both detecting sentences containing variables and disambiguating the respective variables. The focus is placed on analyzing the effects of including different parts of the variables and observing the differences between English and German instances. Additionally, for the variable detection task a bag of words model is used to filter out sentences which are likely to contain a variable mention as a preselection of sentences to perform the semantic similarity comparison on.

1 Introduction

One important way of improving reproducibility and reusability of research is to make its results accessible and comparable. Besides the interlinking of scientific papers, researchers of different disciplines can also benefit from the interlinking of publications and primary data (Boland et al., 2012).

Social scientists often refer to the same survey datasets. Unfortunately, these are seldom properly linked in the publications and if they are, the different surveys and studies often contain a large amount of single questions, called variables which need to be found in the respective corpus (Zielinski and Mutschke, 2017). It would be really helpful to have an automized way of detecting and disambiguating survey variables in scientific papers. For this reason, the very first shared task for survey variable identification is organized as SV-Ident (Tsereteli et al., 2022) at the 3rd Scholarly Document Processing (SDP) workshop at COLING 2022.

We mainly approached the task as a problem of semantic textual similarity (STS) and used language-dependent sentence embedding models to detect and disambiguate variables. For variable detection we additionally used a *Bag of Words* (BoW) model. Although the results did not exceed the baselines, our approach gives insights into which parts of the variables provide the most useful information for semantic similarity based disambiguation.

This paper is structured as follows. In section 2, we present the task and the related data. In section 3, we describe our approach to both tasks. This section starts with an introduction on how the semantic similarity comparison is done, which is the same for both subtasks. Section 4 contains a presentation of experiments performed to find the best parameters for our system. In section 5 and 6 we present our results and we discuss lessons learned, respectively. The paper is concluded in section 7.

2 Task Description and Data

The shared task consists of two subtasks: variable detection (Task 1) and variable disambiguation (Task 2). Both subtasks relate to the same dataset consisting of examples of sentences taken from social science publications[1].

The provided training set consists of 3,823 sentences with labels that indicate whether they contain a variable or not. Each sentence has a document id referring to its source document and an unique id. If the sentence contains one or more variables, ids of these variables are also given, together with a research id which refers to the specific corpus or corpora the variables were taken from.

[*] As an organizer of SV-Ident, this author contributed to the model discussion and the preparation of the system description under the terms and conditions of this task.

[1] https://vadis-project.github.io/sv-ident-sdp2022/

Moreover, the sentences have a language label (see Table 1). There are 1,882 English and 1,941 German sentences in the training set. Additionally, a validation set containing 425 sentences (209 English and 216 German) was released. The test set consists of 1,724 sentences (944 English and 780 German). The test set was in the same format as the training set and it is expected to predict the value of the label indicating whether a sentence contains a variable or not for **Task 1** and the respective variables to the corresponding sentence for **Task 2**.

Attribute	Value
sentence	The probability of ‘never-membership’ is substantially lower if there is a union at the workplace.
is variable	1
variable	exploredata-ZA3700_VarV519
research data	ZA3700
document id	35933
uuid	e2428b76-28de-4b78-aa3f-6055c7d71a1e
lang	en

Table 1: Example of an instance from SV-Ident dataset.

For Task 2, a corpus of variables is also provided. It is divided into 329 sub-corpora labeled with different research ids. They contain variables with their unique ids. Each variable consists of the respective study title, a variable name, the question text in its original language, the question text in English, sub-questions, item categories, answer categories, the variable’s topic in its original language and the variable’s topic in English (see Table 2). Not every item is available for every variable. For 108,374 variables in total, the study titles, variable labels, variable names, topics in the original language and topics in English are missing 25 times. Question texts in the original language are missing 27,705 times, question texts in English 50,319 times, sub-questions 58,294 times, item categories 58,079 times and answer categories 8,783 times.

3 Approach and System Description

The task dataset features several difficulties. One of them is that it is multi-lingual containing both German and English sentences. The fact that the variables consist of different parts with different semantic structures which are not available for all of the variables is another one. Our approach focuses on the analysis of how the diverse information available can be beneficial for semantic similarity comparison.

Attribute	Value
research id	ZA3950
variable id	exploredata-ZA3950_VarV31
study title	International Social Survey Programme: Citizenship - ISSP 2004
variable label	Q7b Rights in democr: Gov respect minorities
variable name	V31
question text	There are different opinions about people’s rights in a democracy. On a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it:
question text en	There are different opinions about people’s rights in a democracy. On a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it:
sub question	Q.7b - ... that government authorities respect and protect the rights of minorities
item category	... that all citizens have an adequate standard of living;... that government authorities treat everybody equally regardless of their position in society;... that politicians take into account the views of citizens before making decisions;... that people be given more opportunities to participate in public decision-making
answer category	Not at all important;2;3;4;5;6;Very important;Can’t choose, don’t know;No answer, refused
topic	[’Soziales Verhalten und soziale Einstellungen’, ’Internationale Politik und Institutionen’, ’Politische Verhaltensweisen und Einstellungen/Meinungen’, ’Regierung, politische Systeme, Parteien und Verbände’]
topic en	[’Social behaviour and attitudes’, ’International politics and organisation’, ’Mass political behaviour, attitudes/opinion’, ’Government, political systems and organisation’]

Table 2: Example of a survey variable.

3.1 Semantic Textual Similarity

We treated Task 1 partly and Task 2 fully as a problem of semantic textual similarity (Agirre et al., 2013). We used language-dependent sentence encoders (Conneau et al., 2017; Reimers and Gurevych, 2019) to create fixed-sized vector representations of the input sentences and some parts of the variables. For this purpose, we experimented with different sentence embedding models.

For the English data, we used *Sentence T5* (Ni et al., 2021) as a sentence embedding model. The variable parts that led to the best results for the English variables were the label, the question text, the question text in English and the topic in English. For the German data, we used the sentence embedding model *"Sahajtomar/German-semantic"* [2]. It is one of the few available German sentence embedding models hosted by HuggingFace to be used out of the box and the one we achieved the best results with[3]. We applied it on all variable parts, except the English translation ones.

We then computed the cosine similarity of all possible pairs of sentences and variables with the same research id. Afterwards the sentence-variable pairs were ranked by their similarity scores. This procedure was the same for both Task 1 and Task 2.

3.2 Task 1 – Variable Detection

The Variable Detection Task basically is a binary classification task. Since this task aims to detect only sentences containing any survey variable, the vocabulary of variables is not essential to use. We tried out two different approaches: one that is independent of the vocabulary and focuses on lexical features of the input sentences only (**BOW Model**) and one that is dependent on the vocabulary and focuses on semantic similarity (**STS Model**). A variation of the first one is used as a preparation for the latter.

3.2.1 BOW Model

For the vocabulary-independent approach, we trained a BoW model similar to (Zielinski and Mutschke, 2017). The input sentences were cleaned of special characters, converted to lowercase, tokenized and stop words were removed using *Natural Language Toolkit* (Bird et al., 2009). Then they were lemmatized.[4]

We used Logistic Regression for the English sentences and Multinomial Naive Bayes for the German sentences to predict whether a sentence contains a variable or not.

3.2.2 STS Model

The variable-independent approach relies partly on a variation of the vocabulary-dependent approach. We tried to increase the recall to ensure to classify all true positives. This way we got a candidate list to further exclude false positives from (similar to (Zielinski and Mutschke, 2017)).

In order to increase the number of true positives in the training data we used *random undersampling* and balanced the distribution of positive and negative instances as explained in section 4.1.1.

We used the STS settings described above to get the similarity scores for the positively labeled sentences and all possible variables. We discarded all sentences that did not exceed a certain threshold. This threshold was computed taking the mean of all true sentence-variable pairs of the training data which showed to be more successful than considering the mean subtracted by the standard deviation of the pairs.

3.3 Task 2 – Variable Disambiguation

Task 2 aims to provide the id of the variable which is referenced in a given sentence. So for this task we directly computed the most similar variables for each input sentence. We used the setup described in 3.1 using language-dependent sentence embedding models to encode the input sentences and specified parts of the variables and then ranked the most similar pairs based on cosine similarity.

4 Experiments

Most of the settings described above were chosen because they proved to be successful in experiments on the validation data. This section provides the results of different experimental setups for BOW and STS models, respectively.

4.1 BOW Model

4.1.1 Random Undersampling

For the preselection of sentences likely to contain a variable, the aim was to exclude false negatives from the prediction by decreasing the number of negative instances in the training data. This was achieved by undersampling negative samples such that the ratio of negative ones to positives decreased from *3.865 to 1* (see Table 3).

4.1.2 Classifier Selection

We treated the languages separately since the lexical distribution of the English and the German

[2] https://huggingface.co/Sahajtomar/German-semantic

[3] The model is based on *German BERT large* (https://huggingface.co/deepset/gbert-large), but unfortunately we could not contact the author to find out which dataset it was further trained on.

[4] This approach is strongly aligned with the BOW Jupyter Notebook, which has been made available in the GitHub repository of the Shared Task as a starting point. https://github.com/vadis-project/sv-ident/tree/main/notebooks/variable_detection

Class Balance	F. Negatives	F. Positives
0: 400, 1: 773	7	63
0: 300, 1: 773	4	75
0: 200, 1: 773	0	81

Table 3: False negatives and false positives for different ratios of positive (1) and negative samples (0) in the training set using Multinomial Naive Bayes.

language differ significantly. The best predictions for variable detection were made using Logistic Regression for the English data and Multionomial Naive Bayes for the German data (see Table 4).

Classifier	English	German
Logistic Regression	**0.780**	0.703
Multinomial Naive Bayes	0.749	**0.745**
KNN	0.520	0.501
Linear SVM	0.757	0.701

Table 4: F1 Scores for Different Classifiers

4.2 STS Model

4.2.1 Variable Parts

Some parts of the variables, like the variable label and name, at first glance do not seem to contain a lot of useful semantic information. Thus, we experimented with using different parts of the variables. Tables 5 and 6 show the impact of these experiments. While using only some parts is effective for the English data, using all parts without English ones yields the best results for the German data.

4.3 Pre-Processing

Different methods of pre-processing were used for both subtasks (see Table 7 and 8, 9). For Task 2, we differentiated between pre-processsing all variable parts and pre-processing only those that do not consist of natural language sentences. Sentence transformers are designed to encode the meaning of whole sentences and pre-processing destroys their syntactical structure. Interestingly, the best result was achieved pre-processing all variable parts, including full sentences.

Variable Parts	Map@10
All	0.127
variable label + question text + question text en + topic en	**0.167**
variable label + question text + topic en	0.143

Table 5: Impact of including different parts of the variables for the English data. The variable parts ‘question text’ and ‘question text en’ are the same in this setting.

Variable Parts	Map@10
All (except from English)	**0.091**
variable label + question text + question text en + topic en	0.050
variable label + question text	0.077

Table 6: Impact of including different parts of the variables for the German data.

Pre-Processing Method	F1
No Pre-Processing	0.756
Pre-Processing without Lemmatization	0.761
Pre-Processing with Lemmatization	**0.765**

Table 7: Impact of pre-processing the English sentences for Task 1. The pre-processing with lemmatization is described in section 3.2.1

Pre-Processing Method	MAP@10
No Removal	0.163
Stop Words *	**0.169**
Duplicates *	0.114
Stop Words and Duplicates *	0.108
Stop Words †	0.164
Duplicates †	0.146
Stop Words and Duplicates †	0.136

Table 8: Impact of removing stop words and duplicates from every part of the variable * and from every part except those including full sentences † for the English instances of Task 2.

Pre-Processing Method	MAP@10
No removal	0.092
Stop Words *	0.081
Duplicates *	0.095
Stop Words and Duplicates *	**0.140**
Stop Words and Duplicates †	0.116

Table 9: Impact of removing stop words and duplicates from every part of the variable * and from every part except those including full sentences † for the German instances of Task 2.

4.4 Trial Data

Additional to the training and validation data, some trial data was released by the organizers [5]. This data set contains a smaller vocabulary of variables. Results on this data were overall better for both subtasks and significantly better for Task 2. Using a similar setup as described above, we achieved an F1 score of **0.823** for the English data on Task 1 and a MAP@10 score above **0.674** for Task 2.

5 Results

While the official evaluation metric for Task 1 is F1-macro, or averaged F1 (averaged harmonic mean of precision and recall), it is MAP@10 (mean average precision of the ten top ranked items) for Task 2.

We achieved the best results using the STS model for Task 1. It scored *0.6016* on the test data (compared to *0.58* for the BOW model) which is still beneath the baseline[6] of *0.6609*, but the best result provided by participants.

In Task 2, our model achieved a result of *0.1359*, which is also beneath the baseline[7] of *0.1893* and it was the only submission made by participants.

6 Lessons Learned

The task proved to be challenging. This can partly be explained by the challenging nature of the data in general. Variable mentions in social science publications typically vary a lot on the linguistic level (Zielinski and Mutschke, 2018). Additionally, dealing with a very large corpus of variables might explain why the results on the test data were so much worse than the results on the trial data..

Since the pre-processing and evaluation of taking into account different variable parts were the main factors improving the results, it would be beneficial to further concentrate on these approaches for future work.

One step into this direction could be the use of data augmentation. This already showed to be successful implicitly, since for the English data better results were achieved including the *question text* and *question text en*, which are the same sentences (see Table 2 and Table 5).

The sentence embedding models used in our approach have the advantage of being suitable for general STS tasks and perform competitively for a variety of such tasks without further fine-tuning (Hövelmeyer et al., 2022). Nevertheless, the baseline models of which one is fine-tuned on social science literature and the other is multilingual achieved better results for this task. For future work, it therefore would be interesting to experiment with models fine-tuned on data similar to the data at hand and multilingual models.

7 Conclusion

We presented a solution to the *SV-Ident* Shared Task relying on semantic similarity and basically treating the subtasks of variable detection and variable disambiguation as the same problem. We encoded the input sentences and parts of the variables using sentence transformer models and treating English and German sentences separately. For Task 1, we used a BOW model with random undersampling in order to create a preselection of likely candidates to contain a variable and then looked for sufficiently similar variables to decide whether a sentence contains a variable or not. For Task 2, we ranked the most similar variables to every input sentence. Throughout, we experimented with different preprocessing methods and different variable parts which proved to be beneficial. It showed that a promising approach for future work could be the consideration of data augmentation techniques.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. *Proceedings of *sEM*, pages 32–43.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Katarina Boland, Dominique Ritze, K. Eckert, and Brigitte Mathiak. 2012. Identifying references to datasets in publications. In *TPDL*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data.

Alica Hövelmeyer, Katarina Boland, and Stefan Dietze. 2022. Simba at checkthat! 2022: Lexical and semantic similarity based detection of verified claims

[5] https://github.com/vadis-project/sv-ident/tree/main/data/trial

[6] The baseline model is the fine-tuned SciBERT model *SSCI-SciBERT* that was further trained on English Social Science abstracts.

[7] The baseline model is a pre-trained multilingual sentence representation model in a zero-shot setting.

in an unsupervised and supervised way. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*, CLEF ’2022, Bologna, Italy.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *CoRR*, abs/2108.08877.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Tornike Tsereteli, Yavuz Selim Kartal, Simone Paolo Ponzetto, Andrea Zielinski, Kai Eckert, and Philipp Mayr. 2022. Overview of the SV-Ident 2022 Shared Task on Survey Variable Identification in Social Science Publications. In *Proceedings of the Third Workshop on Scholarly Document Processing*. Association for Computational Linguistics.

Andrea Zielinski and Peter Mutschke. 2017. Mining social science publications for survey variables. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 47–52. Association for Computational Linguistics (ACL).

Andrea Zielinski and Peter Mutschke. 2018. Towards a gold standard corpus for variable detection and linking in social science publications. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).