Towards a Gold Standard Corpus for Variable Detection and Linking in Social Science Publications

Andrea Zielinski and Peter Mutschke
GESIS - Leibniz Institute for the Social Sciences
{[andrea.zielinski,peter.mutschke]}@gesis.org

Abstract

In this paper, we describe our effort to create a new corpus for the evaluation of detecting and linking so-called *survey variables* in social science publications (*e.g.*, *"Do you believe in Heaven?"*). The task is to recognize survey variable mentions in a given text, disambiguate them, and link them to the corresponding variable within a knowledge base. Since there are generally hundreds of candidates to link to and due to the wide variety of forms they can take, this is a challenging task within NLP. The contribution of our work is the first gold standard corpus for the variable detection and linking task. We describe the annotation guidelines and the annotation process. The produced corpus is multilingual – German and English – and includes manually curated word and phrase alignments. Moreover, it includes text samples that could not be assigned to any variables, denoted as negative examples. Based on the new dataset, we conduct an evaluation of several state-of-the-art text classification and textual similarity methods. The annotated corpus is made available along with an open-source baseline system for variable mention identification and linking.

Keywords: Text mining, semantic textual similarity, paraphrase detection, linking

1. Introduction

There is a growing trend to integrate research data into the scientific publication process. Open Science encourages scientific practices in which all research data should be interlinked and contextualized to enhance reproducibility and reusability of research results. Ideally, publications that report on a result of an empirical study should contain a direct link to the cited dataset and lead the reader directly to the research data that underlies the publication. However, in practice, this metadata is often missing. The potential of text and data mining technologies to automatically detect dataset citations has been addressed, *e.g.*, in the International Workshop on Mining Scientific Publications[1] and 4REAL Workshop[2](Cohen et al., 2016; Branco, 2012; Fokkens et al., 2013).

Interesting work in this direction has been carried out by Mariani et al. (2016) who seek to retrieve mentions of language resources (*e.g.*, corpora, lexica listed in the LRE map) by analyzing the content of the proceedings of the Language Resources and Evaluation Conference (LREC), *i.e.*, to discover the most relevant topics in this field. However, specialized solutions for specific use cases are still required within certain scientific areas.

For instance, social sciences publications often discuss the results of survey studies. A survey generally consists of several hundreds of variables, each of them representing a single survey question (*e.g.*, *Do you believe in Heaven?*). Social science papers, however, only focus on a particular selection of variables. In order to establish links between data and publications on a fine-grained level it is therefore necessary to link not only on study name level but also on the level of survey variables.

The problem of automatically linking a data citation text fragment to the corresponding dataset has been addressed in the INFOLIS project (Boland et al., 2012). However, advanced algorithms that are able to identify the survey variable mentions used in the underlying study and link them to a specific survey variable identifier in a knowledge base are still lacking.

We will refer to the problem as Variable Detection and Linking task, *i.e.*, given a set of variables of a particular survey and topic, all relevant mentions that refer to one of these variables are to be identified.

Within computational linguistics, the problem can be framed in two different ways. It can be conceptualized as an extension of the entity linking problem (Erbs et al., 2011; Rao et al., 2013), attempting to link citations to variables in a knowledge base or it can be phrased as a Recognizing Textual Entailment (RTE) problem, where the system should be able to identify whether a sentence entails a given candidate hypothesis or not (Dagan et al., 2013; Bentivogli et al., 2009).

Specifically the *textual entailment search* task (cf. (Harabagiu and Hickl, 2006), a variation of the RTE task, is adequate: Thus, the question and each answer option form the Hypothesis (H) and the system should be able to retrieve candidate entailing sentences from the document, defined as the Text (T). Also, an undirected relationship between the pairs of texts might hold, as in the related task of detecting Semantic Textual Similarity (STS) (Agirre et al., 2013).

The format of the dataset is released in the same fashion as the RTE data, consisting of pairs of Text (T) and Hypothesis (H), *i.e.*, in our scenario T corresponds to variable mentions in scientific documents and H corresponds to questions and answer sets obtained from the variable data catalog.

Negative pairs are created as combinations of verified Ts with *other* Hs, that is, Hs from the same topic but that cannot be linked to T. In correspondence to the real-life application setting, we also selected a high proportion of unrelated sentences. While the first set of positive and negative T/H pairs can be used for the subtask of *Variable Disambiguation*, the second set of unrelated sentences has been

[1] https://wosp.core.ac.uk/

[2] Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language

created for the subtask of *Variable Detection*.
Furthermore, we investigated the diverse types of citations in a qualitative and quantitative study. It is the first corpus for this task and we will make it available to the research community. It has a broad coverage of linkage types, showing that lexical semantics is important for obtaining good performance. The new corpus is intended to drive the development of NLP methods for the detection and linking of variable mentions and can be used for benchmarking them.

2. Use Case Description

The Variable Detection and Linking task is to automatically augment a plain text document with links to variables in order to annotate salient social science concepts. It assumes the existence of a knowledge base, covering all variables of interest. In social science, survey variables are generally listed in data catalogs such as ISSP[3] or ALLBUS[4]. While each survey (or questionnaire) is composed of a specific set of survey variables, only a subset of them might be cited in a publication. An illustration for the variable linking task is shown in Figure 1.
In our setting, the task focuses on classifying and linking mentions to one (or more) of the variable classes that are identified by a unique identifier in ALLBUS[5]. Variable-level information includes the question and subquestion text, an associated topic and a predefined set of answers (*i.e.*, the majority of survey questions are closed and respondents have to mark their choices w.r.t. the given response options). The wording of the questions and answers is generally well chosen according to common practices in survey design. An example of a survey variable is provided in Figure 2.

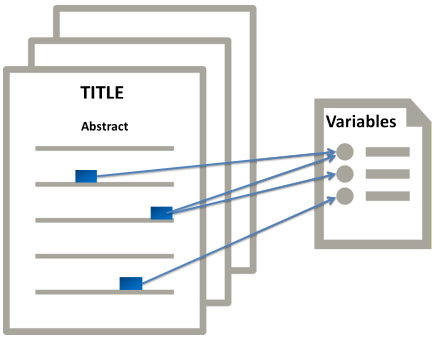

Figure 1: Linking Survey Variable Mentions in Scientific Publications

Identifying mentions of survey variables is a particular challenge in social science publications, because they usually appear in a wide variety of forms. Authors in this field often do not quote a variable literally but tend to paraphrase it, as exemplified in (1), or summarize more than one variable in a single sentence, as illustrated in example (2). For instance, the paraphrase in example (1) is used in our dataset to refer to variable v278 (cf. Fig. 2).

[3]International Social Survey Programme `http://www.issp.org/menu-top/home/`

[4]`https://www.gesis.org/en/allbus/`

[5]Variable IDs extracted from ALLBUS CUM. 1980-2014 for German and English (Allerbeck et al., 2016; Lepsius et al., 2016)

```
<variable v_id="v278" lang="English"
surey="ALLBUS_cumulated">
      <v_label>OPINION ON DUAL CITIZENSHIP
      </v_label>
      <v_topic>Ethnocentrism and
      Minorities</v_topic>
      <v_question>Using the scale on the card,
      please indicate the extent to which you agree
      with each statement. </v_question>
      <v_subquestion>Foreigners living in Germany
      should be able to acquire German citizenship
      without having to give up their own
      citizenship, dual citizenship should be
      possible. </v_subquestion>
      <v_answer a_id="1">Not available</v_answer>
      <v_answer a_id="2">Completely
      disagree</v_answer>
      <v_answer a_id="3">Disagree</v_answer>
      <v_answer a_id="4">Indifferent</v_answer>
      <v_answer a_id="5">Agree</v_answer>
      <v_answer a_id="6">Completely agree</v_answer>
      <v_answer a_id="7">No answer</v_answer>
   </variable>
```

Figure 2: Example of a Survey Variable

> Results of more recent public opinion such as the Allbus Survey conducted in 2006 show that the majority of Germans continue to reject the idea of generally granting dual citizenship. (1)

The text fragment in (2) can be linked to the Allbus variables v274, v275, v276 and v277. All of them have a common main question *Do you have any personal contact with foreigners living in Germany?* but differ in their respective subquestions, *e.g.*, *...at work?*, *...in your neighborhood?* etc.

> Encounters between Germans and foreigners can take place in different spheres of life, at work, in the neighborhood, in the family or in the circle of friends and acquaintances. (2)

Our dataset consists of German and English mentions and variables. For instance, sentence (3) originates from a German scientific publication (synonymous to (2)) and can be linked to the corresponding German survey variable v275 (*i.e.*, *Haben Sie persönlich Kontakte zu in Deutschland lebenden Ausländern, und zwar an Ihrem Arbeitsplatz?*).

> Begegnungen zwischen Deutschen und Ausländern können in verschiedenen Lebensbereichen stattfinden, bei der Arbeit, in der Nachbarschaft, in der Familie oder im Freundes- und Bekanntenkreis. (3)

The problem of identifying survey variable mentions in texts can be defined as a multi-label classification task (Zielinski and Mutschke, 2017): given a set of sentences S $\subseteq \{s_1, .., s_i\}$ and variables V $\subseteq \{v_1, .., v_j\}$, a function needs to be defined $h : S \rightarrow V$. Each sentence s is represented by a single instance which can be associated with one (or more) class label(s), including *unrelated* as a label, in case the mention cannot be assigned to any of the variables.

3. General Annotation Procedure

Our benchmark dataset contains data from 100 scientific publications compiled from the Social Science Open Access Repository[6] (SSOAR) which all carry an established link to the survey study ALLBUS. It covers 20 general-domain topics such as *economy*, *political attitudes and participation*, *attitudes towards marriage, family and partnership*, and *use of media*.[7]

The Variable Corpus is a development corpus of English and German data given as single structured XML files, one for each language. The corpus consists of 415 *positive* and 505 *negative* sentence pairs hand-tagged by two social science students when in their judgment a mention in the text can be linked to a variable or not. For example, given the variable "Do you believe in Heaven?" and a citation "Two million inhabitants believe in God and Heaven", the annotators picked the citation and the variable as a positive pair. The citations extracted from the scientific documents provide a set of roughly synonymous sentences representing different linguistic realizations of a particular variable in the knowledge base. Positive sentence pairs and negative pairs can be used jointly for *Variable Disambiguation*. Moreover, 865 *unrelated* sentences have been selected and can be used for the subtask of *Variable Detection*.

The annotation procedure consists of the following steps:

1. Annotators read all documents from beginning to end and search for text passages that refer to any of the variables of the underlying survey. If so, they look them up in the variable catalog and retrieve the variable ID, text, question and answer set.

2. Annotators build negative pairs from topically-related variables. The mention-variable pairs are cases where annotators disagreed and/or with a high degree of confusability. These cases might be particularly informative because they are near the decision boundary.

3. Annotators select additional sentences from the abstract and/or the same paragraph in which a mention occurs. In our setting, approximately three quarter of the corpus is made up of *unrelated* sentences.

4. Validation: In a final pass, all given sentence pairs are revisited to resolve any remaining inconsistencies. Any pdf-to-text errors in the text and line-breaks were removed as a part of the preformatting.

In step 2, annotators choose alternative variables such as, *e.g.*, *v1328*, *i.e.*, *"Have you had German citizenship since birth?"* and *v261* which relates to the question *"whether to grant German citizenship only to persons that were born in Germany"*. These are used to build negative variable-mention pairs.

Then, in step 3, sentences from the abstract and the same passage are extracted. For instance, sentence (4) immediately follows (1) and elaborates on the same topic.

> Asked to indicate the degree to which they agreed or disagreed with the idea of allowing immigrants to naturalize without relinquishing their former citizenship, 40 percent of respondents strongly agreed and 54 percent rejected the idea to some degree. (4)

4. Annotation Guidelines

Setting up a clear annotation guideline is important for defining the task properly. The disambiguation of variable mentions with respect to a predefined set of variables is sometimes difficult due to ambiguities, vagueness or identification of only partial matches. For this task, we have defined the following annotation guidelines for human annotators:

- *The sentence containing the variable mention should be self-contained*, *i.e.* it should be a suitable reference also when seen in isolation from the context. Annotators thus need to identify the spans of text that most accurately reflect the contents of the variable.
- *Mentions referring to more than one variable* should be assigned all valid variable IDs.
- *Linking mentions at the correct level of granularity*, *i.e.* if alternatively more general or more specific variables exist, they should not be selected.

Our aim was to exhaustively identify all links in the publications and include them as text samples in our gold standard corpus. However, there are some exceptional cases where samples were too vague or part of a table and have therefore not been included in the corpus. For instance, if the reference involves not a single variable but can only be achieved by selecting a whole set of variables, we opted not to include this sample in the corpus. Also, if a survey question cannot be understood without the previous question, we discard the sample. This might happen because the interviewer generally asks a standardized list of questions in consecutive order.

For a few control variables – these are generally used to describe the population selected for the study – the question text was missing and had to be added manually[8].

A design decision was to restrict the length of text samples to a sentence. However, in the variable corpus, the *local context* of the variable is provided, *i.e.*, the whole paragraph in which the mention occurs, so that the similarity of the context of the mention with the associated variable can be exploited. All mentions from the document define the *global context* of the variable, which make it possible to investigate the semantic coherence between co-occurring mentions in a document.

4.1. Sub-sentential Alignments

In order to assess the difficulty of the task, we also explore the dataset in relation to possible sub-sentential alignments and context dependencies for all positive pairs.

[6] http://www.ssoar.info/ssoar/

[7] Variables are assigned to thematic categories based on the CESSDA topic classification, cf. https://dbk.gesis.org/DBKSearch/Topics.asp

[8] For instance, for the variable assessing the respondent's age, we chose *Please tell me your age.*

4.1.1. Linguistic Annotation Layer

Two computational linguistics students annotated and aligned all positive sentence pairs using the paraphrase typology and tagset of Vila (2015) which has been created for addressing individual paraphrase phenomena. In their work, 24 paraphrase types have been defined, ranging from morpho-lexical changes (*e.g.*, derivational changes, lexical substitutions), to syntax-based (*e.g.*, negation switching, ellipsis), discourse-based (modality changes), and semantic-based changes.

An overview of the frequency of the different types of word and phrase alignments between corresponding sentence pairs in our English and German corpus is provided in Figure 3, including examples in Table 1. The most frequent types are identity mappings, followed by (local) lexical-semantic variations and (global) discourse-based modifications. In the case of identity mappings (*i.e.*, aligned phrases that are exactly the same in wording), token-level overlap is 16,1% and character-level overlap makes up 16,75%, when normalized by the length of the question text[9]. Note that discourse-related modifications such as conversion from direct to indirect speech are frequent in our use case and not relevant for the judgment on meaning preservation. They also imply changes regarding modality, tenses, adverbials, pronouns, and often go along with argument variation.

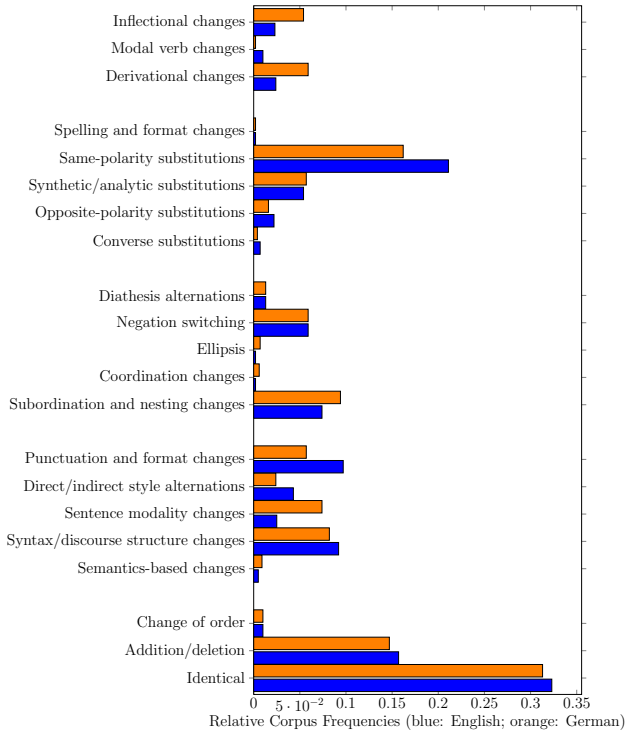

Figure 3: Linguistic Phenomena identified in the annotated corpus (in percentage terms)

4.1.2. Semantic Types of Linkages

Different phenomena regarding the semantic type of linkage could be observed at the sentence level. Special to our setting is that the relation between mention and variable is not necessarily symmetric, *e.g.*, an entailment might hold.

- EQUIVALENT: [fare dodging] ⇔ [use public transport without buying a valid ticket]
- MORE-SPECIFIC: [income tax return] ⇔ [tax return]
- MORE-GENERAL: [lead to problems] ⇔ [reason for shortages]

Moreover, the alignment is mono-lingual and non-exhaustive, *i.e.* it is not required that the entire reference sentence is semantically equivalent to the variable text.

[9]We compute the percentage of lexical overlap between two sentences T and H as: $|T \cap H|/|H|$

Phenomena	Example
Same polarity substitution	make lower - reduce
Opposite polarity substitution	fear - prefer not to
Converse substitution	to work hard - for individual achievement
Derivational change	immigrants - immigration
Inflectional change	walk - walking
Modal verb change	must - has to
Syntax/discourse structure	priorization - choose most important
Diathesis alternation	should have what they need - should get the money they need
Subordination and nesting	belong to a Christian denomination - being Christian
Spelling and format	quotations (”)
Direct/indirect style	What is your opinion? Should social benefits be cut in the future, or should they be extended? - The first question asks whether “social benefits” should be cut or extended [..].
Sentence modality change	How often do you pray? - The average number of children for respondents who never pray was only 1.39 compared to 2.06 for those who pray daily.

Table 1: Linguistic Phenomena in the English Dataset

4.2. Context Clues

Generally, classifying short text is a challenging task because only little context is available and word co-occurrence information cannot be reliably exploited. In our dataset, however, cue words often co-occur with a variable mention in the same sentence or text passage, as shown in example (5) and (6), respectively. Such trigger terms might introduce the speaker (*e.g.*, respondents of a survey), reporting verbs (*e.g.*, verbs expressing opinion or factuality),

measure and assessment verbs, or figures and percentages.

In this case the *figures from Eurobarometer show* that those with worst expectations *report* on average a *7.33%* lower life satisfaction [..]. (5)

In 2009 the *respondents* from *33 European countries were asked by Eurobarometer* whether they expected [..]. (6)

4.3. Corpus Statistics

Table 2 reports some key statistics about our collected datasets. Our benchmark corpus comprises 504 English and 638 German sentences from 35 English and 34 German documents (out of 50 English and 50 German documents) which contained variable mentions. The average length of the sentences extracted from scientific publications is 28 tokens for English and 24 for German, while the average length of variables is 24 tokens for English and 19 tokens for German, considering the question and subquestion text. Pairing the mentions with respective variables results in 466 English and 454 German sentence pairs.

Corpus for Variable Mention Detection

Sentences	#Related	#Unrelated
English	126	378
German	151	487

Corpus for Variable Mention Disambiguation

Sentence Pairs	#Positive	#Negative
English	194	272
German	221	233

Table 2: Corpus Statistics

We also computed the cardinality of the dataset S (*i.e.*, the mean of the number of labels of the instances that belong to S) and the density of S (*i.e.*, the mean of the number of labels of the instances that belong to S divided by $\mathbf{card}(L)$). Label density is between 1 and 7, with a mean of 1.54 and 1.46 as shown in Table 3.

Sentences	#English	#German
Label Density	1.54	1.46
Cardinality	3.69	3.00

Table 3: Label Density and Cardinality

4.4. Inter Annotator Agreement

Inter-Annotator Agreement (IAA) measured with Cohen's Kappa is relatively high, *i.e.*, 80% on sentence level[10] and 91.5% after a reconciliation among the annotators.
On the sub-sentential level, we focused on the agreement w.r.t. the paraphrase type rather than the phrase boundaries, yielding an agreement of approx. 85%.

[10]Average Kappa level of 0.78 corresponding to 'substantial agreement' (Landis and Koch, 1997).

5. Experiments

This section presents results of the baseline approach on our German and English corpus. We evaluate the two steps *Variable Detection* and *Variable Disambiguation* separately: In the first step, we seek to detect occurrences of variable mentions. In the second step, we consider a given set of variables as candidates for all relevant sentences and seek to assign the proper variable ID. Automatically produced annotations are then compared to ground-truth data. We experimented with prominent NLP and ML algorithms adopted to the related tasks RTE, STS and Entity Linking, and tested their effectiveness on out task. We used DKPro Core (de Castilho and Gurevych, 2014), a linguistic pipeline based on UIMA to pre-process the corpus. To facilitate further research on the new resource, we provide a baseline variable linking system based on DKPro-TC (Daxenberger et al., 2014) and DKPro-Similarity (Bär et al., 2012) with models trained on several standard text similarity datasets, *e.g.*, the Microsoft Research Paraphrase Corpus (MSR-Paraphrase)(Dolan et al., 2004)[11].

5.1. Baseline for Variable Detection: Classification using VSM

As a baseline to the Variable Detection task, we adopt a shallow approach based on machine learning applied to lexical features extracted from the dataset. Accordingly, linguistic expressions are treated as a bag of words, using the variable questions and subquestions for training and the mentions for testing. In order to overcome the problem of a lack of training data – in particular because there is only one example for each class – the training dataset is augmented with additional features from WordNet and GermaNet (*i.e.*, synonyms, hypernyms and derivational forms) and keyword terms from *TheSoz* (Zapilko et al., 2013)[12].
For the two-way classification task, *i.e.*, *related* class (any variable ID) versus *unrelated* class, we focus on a high recall which makes it possible to filter out false positives in a later processing step. Table 4 shows the performance in terms of (macro-averaging) precision and recall. As we hypothesized, including only the lemmas in the feature vector yields relatively low recall on the minority class. Best results in this regard are achieved when training the model by expanding the feature space with semantic relations from the lexical databases, and using it to classify the test data, which is based on lemmas. Moreover, results are constistently better for English than for German, mainly due to the high rate of German compounds that have not been splitted into their component parts.

5.2. Baseline for Variable Disambiguation: Similarity Metrics

We have also conducted experiments based on text similarity scores, including *e.g.*, *greedy-string tiling*, *levenshtein*, *longest common subsequence*, *character n-gram* and *BLEU*. In this configuration, similarity scores for pairs

[11]https://github.com/openminted/uc-tdm-socialsciences

[12] Thesaurus for the Social Sciences http://lod.gesis.org/thesoz/

English Baseline for Variable Detection

	P	R	P	R	MAP	MAR
	related		*unrelated*			
Lemma	0.81	0.18	0.59	0.97	0.70	0.57
+Lexical Resources	0.63	0.47	0.65	0.78	0.64	0.63
Train/Test on F_{Lex}/F_{Lem}	0.48	**0.91**	0.76	0.21	0.62	0.56

German Baseline for Variable Detection

	P	R	P	R	MAP	MAR
	related		*unrelated*			
Lemma	0.37	0.16	0.66	0.85	0.51	0.51
+Lexical Resources	0.37	0.16	0.66	0.85	0.51	0.51
Train/Test on F_{Lex}/F_{Lem}	0.36	**0.82**	0.71	0.23	0.54	0.53

Table 4: Baseline Results for Variable Detection based on the Naïve Bayes Classifier with 3 different configuations. *Lemma*: The classifier is trained and tested on lemmas; *+Lexical Resources*: The classifier is trained and tested on lemmas enriched with features from lexical resources; and F_{Lex}/F_{Lem}: When training the classifier, features from lexical resources are integrated, while testing is carried out on the lemma forms.

of variable description and their mention are calculated, combing the set of extracted features into one feature vector and feeding it into a Simple Logistic Regression classifier. For comparison, we also ran the Sequential Minimal Optimization (SMO) classifier which performed slightly better on our datasets. Evaluation results for the task of *Variable Disambiguation* based on 10-fold cross-validation are reported in terms of accuracy, *i.e.*, only for instances that belong to the *related class*. The accuracy in our multi-label classification setting is the proportion of labels correctly classified of the total number (predicted and actual) of labels for that instance averaged over all instances.

Algorithm	English	German
Logistic Regression Classifier	60.39%	65.25%
SMO	63.68%	66.81%

Table 5: Baseline Evaluation Results for Variable Disambiguation in terms of accuracy, based on the precision/recall for each class label over the *related class* dataset

6. Related Work

While various benchmark datasets have been developed for the shared tasks in semantic relatedness and textual entailment (Bentivogli et al., 2017) (Agirre et al., 2013), no resource exists so far for the Variable Detection and Linking task. There are major differences which makes the task interesting, summed up in Table 6, w.r.t. the following characteristics:

a) Context Dependency: Should information outside the sentence pairs be taken into account?

b) Class Distribution: Is the corpus balanced or unbalanced in terms of related and unrelated sentences? Is it balanced in terms of positive and negative sentence pairs?

c) Partial Entailment: Only some 'facets' within the sentences match

d) Domain: Formal (*e.g.*, scientific publications, news) versus informal domains (*e.g.*, forums, blogs)

Our corpus differs from other corpora in related applications in various ways: a) the local context (*i.e.*, paragraph) in which the mention occurs is provided so that context similarity clues can be taken into account; b) while the majority of sentences belong to the *unrelated* class, class distribution according to the *positive* and *negative* class is almost balanced; c) semantic equivalence or entailment relationships can generally be observed only in parts of the sentences; and d) the corpus has been compiled from German and English scientific publications.

Because of the fact that the major bottleneck of our use case is the high variability due to different linguistic realizations of the same variable, the RTE semantic search scenario seems most appropriate to our use case. Due to the lack of training data, no prior knowledge on the likelihood of a link can be exploited, as is usually done in entity linking.

7. Conclusion

We have introduced a new dataset which has been created for the *Variable Detection and Linking* task and originates from the needs within the social sciences. We intend to make the corpus freely available to the research community under a Creative Commons license, along with the annotation guidelines. We have proposed a pipeline that includes several stages: a) pre-processing, b) Variable Detection and c) Variable Disambiguation, and evaluated it on our German and English datasets. We first applied a Naïve Bayes Classifier on BoW lexical features extended with *WordNet/GermaNet* and *TheSoz* terms to achieve a high recall and then ran a more precision-oriented SMO classifier based on string similarity features. While this approach is flexibel and can easily adapt to any new repertoire of survey variables, experimental results show that due to the small number of available training instances this is a challenging task within NLP. Yet, we think that the dataset will foster research in this field and lead to enhanced solutions that might also take into account the local and global context of the variable mention, and exploit the answer set of the variables.

8. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021 (OpenMinTeD)[13].

[13] http://openminted.eu/

Datasets	RTE1-4	RTE5-6	RTE8	WikiQA	STS-par	SEM-QA	iSTS'16	VDL'18
Context Given	No	Yes	No	No	No	No	No	Yes
Balanced Class Distribution	Yes	No	No	No	Yes	Yes	Yes	No
Partial Entailment	No	No	Yes	No	No	No	Yes	Yes
Domain	News	News	Scholar	Wikipedia	News	Forum	News	Scholar

Table 6: Dataset Characteristics *(Bentivogli, 2017)*. VDL'18 is our Variable Detection and Linking Corpus.

9. Language Resource References

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). SEM 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics*.

Allerbeck, K., Allmendinger, J., Andreß, H.-J., Bürklin, W., Diekmann, A., Feger, H., Fetchenhauer, D., Huinink, J., Kiefer, M. L., Kreuter, F., Kühnel, S., Kurz, K., Lepsius, M. R., Liebig, S., Mayer, K. U., Meulemann, H., Müller, W., Opp, K. D., Pappi, F. U., Scheuch, E. K., Schmitt-Beck, R., Solga, H., Trappe, H., Wagner, M., Westle, B., and Ziegler, R. (2016). Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS - Kumulation 1980-2014.

Bär, D., Biemann, C., Gurevych, I., and Zesch, T. (2012). UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics (SemEval 2012)*, pages 435–440, Montréal, Canada, 7-8 June.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. (2009). The Fifth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Bentivogli, L., Dagan, I., and Magnini, B. (2017). The recognizing textual entailment challenges: Datasets and methodologies. In *Handbook of Linguistic Annotation*, pages 1119–1147. Springer.

Boland, K., Ritze, D., Eckert, K., and Mathiak, B. (2012). Identifying references to datasets in publications. In *International Conference on Theory and Practice of Digital Libraries*, pages 150–161. Springer.

Branco, A. (2012). Reliability and Meta-reliability of Language Resources: Ready to initiate the integrity debate? In *The 12th workshop on treebanks and linguistic theories*.

Cohen, K. B., Xia, J., Roeder, C., and Hunter, L. (2016). Reproducibility in natural language processing: A case study of two R libraries for mining PubMed/MEDLINE. In *LREC 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6–12.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). DKPro TC: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland, June.

de Castilho, R. E. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT at COLING*, pages 1–11.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04.

Erbs, N., Zesch, T., and Gurevych, I. (2011). Link discovery: A comprehensive analysis. In *Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 83–86. IEEE.

Fokkens, A., Van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. (2013). Offspring from Reproduction Problems: What replication failure teaches us. In *ACL (1)*, pages 1691–1701.

Harabagiu, S. and Hickl, A. (2006). Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics*, pages 905–912.

Lepsius, M. R., Kreuter, F., Liebig, S., Kurz, K., Kühnel, S., Kiefer, M. L., Huinink, J., Fetchenhauer, D., Solga, H., Scheuch, E. K., Schmitt-Beck, R., Pappi, F. U., Müller, W., Opp, K. D., Meulemann, H., Trappe, H., Mayer, K. U., Feger, H., Diekmann, A., Bürklin, W., Andreß, H.-J., Allmendinger, J., Allerbeck, K., Wagner, M., Westle, B., and Ziegler, R. (2016). German General Social Survey (ALLBUS) - Cumulation 1980-2014.

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2016). Rediscovering 15+ 2 years of discoveries in language resources and evaluation. *Language Resources and Evaluation*, 50(2):165–220.

Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Vila, M., Bertran, M., Martí, M. A., and Rodríguez, H. (2015). Corpus annotation with paraphrase types: New annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*, 49:77–105.

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2013). Thesoz: A SKOS representation of the thesaurus for the social sciences. *Semantic Web*, 4(3):257–263.

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