Variable Detection and Linking – Annotation Guidelines

Background

The present annotation guidelines have been written for the task of Variable Detection and Linking, a Social Sciences Use Case, which seeks to identify the variables mentioned in free text.

The lack of representative datasets for the development and evaluation of tools for this task has been the major motivation for building up this pilot corpus. The corpus consists of text fragments from 100 randomly selected social sciences research publications[1] with established links[2] to the German General Social Survey Allbus[3].

The corpus is designed as a representative sample that reflects the various types of variables (*control* vs. *non-control variables*) and how they are mentioned (w.r.t. linguistic phenomena). The data set, henceforth referred to as *VariableDetectionCorpus*, contains 277 mentions labeled as referring to 406 (out of 1.936) variables for English and German, i.e. 126 English and 151 German mentions, respectively. The selection of text samples from within the publications seeks to exhaustively cover all identified links. A specific challenge of the task is that mentions can refer to more than one single variable and that the local context might be required for variable linking. A design decision was to restrict the length of text samples to a sentence and include 'unrelated' sentences.

Task Description

Our task is to create annotations in scientific papers, thus providing a direct link to the corresponding variables that the author was referring to. This way, it can be found out which variables are addressed most in social science research, and what topic(s) they belong to. Whenever an author seeks to support his/her central findings and claims by primary data, e.g. survey data, references to underlying datasets and variables are likely to occur.

Figure 1 represents the scenario where the reader is deciding whether there is a reference to a variable in the publication.

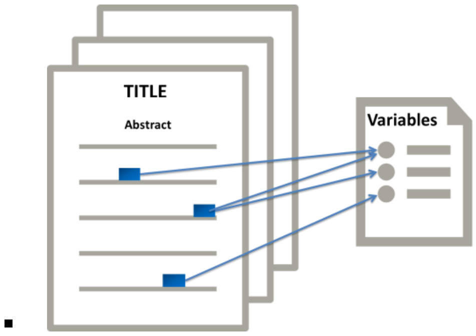

Table 1: Variable Linking Task

[1] Scientific publications are drawn from SSOAR (Social Science Open Access Repository) and further open repositories for Social Science publications.

[2] Dataset links are provided by Infolis, an automated method to identify a large number of data set mentions in publications.

[3] All ALLBUS variables are listed in ZACAT, a social science data portal provided by GESIS Leibniz Institute for the Social Sciences which offers published datasets and accompanying documentation. It provides useful information on variables like, e.g., variable texts and answer categories.

Introduction

The Variable Detection Annotation Task is defined as follows: For a given set of scientific publications that contain dataset references, identify all the text spans (i.e., mentions) that pertain to a particular variable.
Task 1 (*Variable Detection*): For each mention, identify the spans of text that most accurately reflect the contents of the variable. If there are several consecutive sentences that elaborate on one single variable, the most informative one will be chosen. Mentions occur in most cases as a text fragment inside a sentence. It might also happen that a single sentence includes mentions to several different variables.
Task 2 (*Variable Disambiguation*): For each detected text span, identify what variable it belongs to, from a predefined set of variables.

Key Concepts

Mention text: the sentence(s) in a scientific publication that contain a link to a variable.

References to Variables: A single mention may be referenced by one or more variables. An author may also discuss and refer to one or more fields of a particular variable: Currently, there are the variable fields *ID, question text, subquestion text, topic, and answer set.*

Annotation Procedure:

This paragraph describes the procedure that is to be followed when annotating a document.

Step 1: Get an overview

Read the title and keywords of the document and determine what the topic of the article is. Generally, the selected publications focus on a specific topic and therefore author's predominantly mention variables that corresponded to the respective topic.

Step 2: Identification of variables – Text segmentation

Start to read the document sentence by sentence from beginning to end. Decide whether there is a text fragment inside a paragraph that refers to a variable. If so, choose the sentence which is most informative.

The sentence containing the variable mention should, if possible, be self-contained and clear without the knowledge of any preceding or following sentences. It should be a suitable reference also when seen in isolation from the context.
Note: While individual sentences most often refer to a single variable, in some cases, there are also multiple occurrences of variable.

Note: Variables generally do not occur in the abstract and in the reference section.

Step 3: Identification of variables – Select Variable ID

The mention text may contain (part of) the variable text literally, or express the semantic content of the variable in other words, or be narrower/broader. Therefore, it is not always easy to select the variable that best summarizes the mention.

Listed below are some examples of different types of references to variables one can find in a text. In this section we will describe the following annotation cases:

Examples for Annotations

1. Identification of variables – Text segmentation

1.1. Self-containing references within one single sentence

Different references to the same variable occur in one paragraph. Since both references are self-contained and valid, they are both selected and included in the corpus.

Reference 1: "To test this, we analyzed data on the strength of individuals' identification with their home town and its inhabitants from the German ALLBUS surveys.

Reference 2: "This is presented in figure II, which reports the marginal effect of being Catholic on the propensity to feel strongly attached to one's home town and its inhabitants...' "

Variable label: IDENTIFICATION WITH OWN COMMUNITY

Comment: As a pre-processing step, we eliminated spelling errors and errors from automatic sentence segmentation.

1.2. Context is necessary to determine the variable

While the sentences should be self-contained, the sentences before/after the reference sentence provide additional contextual information and valuable clues regarding the identification and disambiguation of variables.

Document Title "Exploring Sources of Punitiveness Among German Citizens

Section Title: "Covariates of punitive attitudes.

Sentence before: "For both, respondents were asked whether they agree or disagree with these statements, and the responses were recoded so that a positive response (1) indicates feelings of the designated type of cynicism."

Reference: "We also include a measure that we label as "life satisfaction", which is a four-category item asking respondents the following: "All things considered, have your ideas of what you wanted to achieve in life been (1) more than fulfilled, (2) fulfilled, (3) not quite fulfilled, and (4) not at all fulfilled?"

Variable label: PERSONAL AMBITIONS IN LIFE FULFILLED?

Document Title "RACISM IN SOCCER: ELIMINATING SOCCER RACISM AND USING SPORT AS A VEHICLE FOR NATIONAL CHANGE"

Sentence before: "This graph shows significantly lower levels of positive national pride and significantly higher levels of negative national pride in Germany compared to other comparable countries."

Reference: "There was an increase from 71% of the population stating they were "very proud" or "fairly proud" a few months before the games to 78% of the population stating they possessed these positive feelings during the games."

Variable label: v247 PROUD TO BE A GERMAN?

Comment: In general, the unit of a passage is selected as context.
Note: It can be helpful to pay attention to co-occurring variables in the same paragraph. Authors tend to discuss not a single variable but a whole set of variables. Such clustering of variables may be to express an entire social concept composed of different variables, or it may list dependent variables, in particular control variables.

2. Identification of variables – Selection of correct Variable ID

2.1. Linguistic variations of the variable mention

Due to the huge variability in language, it is not always easy to find mentions of variables. Semantic similarity and inferencing might be necessary to detect correspondences.

2.1.1. Quotation of the variable text

They easiest case is when there is a quote of the variable text in the publication:

Reference: "There is only one item measuring happiness which directly asks the respondents: 'If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole...' "

Variable question: "If you were to consider your life in general these days, how happy or unhappy would you say you are, on the whole..."

2.1.2. Paraphrase of the variable text

It is also possible that the variable text is paraphrased like in this example below, where the first and second variable text would be correct choices:

Reference: "The second and the third questions come from the ISSP research, where respondents were asked about the influence of religious leaders on people's votes and the government."

Variable question 1: "How much do you agree or disagree with each of the following: Religious leaders should not try to influence how people vote in elections."

Variable question 2: "How much do you agree or disagree with each of the following: Religious leaders should not try to influence government decisions."

2.1.3. Negative Polarity Item

It is also possible that the variable text is expressed by means of a different polarity marker. In the example below, the negative polarity item is used for the variable text, while the mention is expressed by means of a positive polarity item.

Reference: "Victimization and fear of crime are dichotomous, with "1" indicating positive responses to either of the two following questions: "Have you been a victim of theft in the past 3 years?" and "Is there any place in the immediate vicinity in which you fear walking alone at night?"

Variable question: "Is there any area in the immediate vicinity - I mean within a kilometer or so - where you would prefer not to walk alone at night?"

2.1.4. Lexical Inference

Lexical inferencing denotes a process of guessing the meaning of an unknown word by employing all linguistic cues available in the text together with the reader's world knowledge, his/her linguistic knowledge, and his/her awareness of the context.(Haastrup, 1991). The appropriate inference includes also hypernyms and hyponyms occurring as a lexical unit of the variable.

Reference: "First, winners are more politically satisfied compared with losers, including those who voted for the Black–Red Grand Coalition."

Variable question: "Which party did you vote for with your second vote ("Zweitstimme")?

Variable answers:

Respondent didn't vote
The Christian Democratic/Christian Social Union CDU/CSU
The Social Democratic Party SPD

2.2. Variable in the Knowledge Base ZACAT

A variable is composed of various fields: question, subquestion, label, response categories, and topic. The text provided in the fields is largely complementary to one another and all of them can be exploited for establishing a link.

Note: In general, the (sub)question text is most informative, while the variable names and labels are often cryptic. To make sense of the data, the researcher often uses a code book in addition that contains a list of variable names, their description, and an interpretation of the variables.

2.2.1. Variable Text

2.2.1.1. Elliptical question and subquestion

There are certain variables which have a generic question and can only be distinguished from other variables at the level of subquestion which the survey participant has to answer. This often involves ellipses, as can be seen in the example below:

Reference: "To measure anti-immigrant sentiments, a four-item scale was created (α = .72) from responses to questions regarding citizens' beliefs about immigration for four groups: asylum seekers, EU workers, non-EU workers, and ethnic Germans."

Variable question: "The following questions deal with the entry as immigrants of various groups of people into Germany. What is your opinion about this?

Variable subquestion: What about ethnic Germans from Eastern Europe?

2.2.1.2. Various variable candidates at different levels of specificity

The mention should be linked to the variable at the correct level of granularity. Alternatives like 'identification with federal state' would be incorrect linkages and thus only v321 should be selected.

Reference: “This is presented in figure II, which reports the marginal effect of being Catholic on the propensity to feel strongly attached to one’s home town and its inhabitants.“

Variable ID: v321

Variable label: IDENTIFICATION WITH OWN COMMUNITY

Variable question: Now we would like to know how strongly you identify with your own town (community) and its inhabitants. Please use the card for your answers

Variable Subquestion: Do you identify emotionally with your town very strongly, pretty strongly, only weakly or not at all?

Variable ID: v322

Variable label: IDENTIFICATION WITH FEDERAL STATE

Variable question: Now we would like to know how strongly you identify with your own town (community) and its inhabitants. Please use the card for your answers

Variable Subquestion: Do you identify emotionally with your federal state very strongly, pretty strongly, only weakly or not at all?

2.2.2. Variable Response categories

Sometimes it is not possible to clearly identify a variable without using the corresponding response categories, as shown in the following example:

Reference: “Statistics on the degree of spirituality provide a clearer picture: 8.4% describe themselves as very spiritual, 29.7% as moderately spiritual, 32.8% as slightly spiritual, and 29.1% as not spiritual at all.”

Variable question: “What best describes you: “

Answer categories:

- I follow a religion and consider myself to be a spiritual person interested in the sacred or the supernatural.

- I follow a religion, but don't consider myself to be a spiritual person interested in the sacred or the supernatural.

- I don't follow a religion, but consider myself to be a spiritual person interested in the sacred or the supernatural.

- I don't follow a religion and don't consider myself to be a spiritual person interested in the sacred or the supernatural.

2.2.3. Variable Label

In a few cases the crucial information to clearly determine a variable can be in the label of the variable. This is particularly the case for indirect agents, i.e. the respondents.

Reference: “We run least squares regressions of attractiveness on anthropometric measures and several groups of control variables, including age, region, year, interviewer fixed effects, number of children, and health status.

Variable label: RESPONDENT: AGE

Variable question: -

Answer categories:

- Refused
- No answer

Another example is the variable v925 with label ‚AEQUIVALENZEINKOMMEN OECD - NEU, KAT' which does not have any question text in Allbus. If the variable description contains explanatory text about the definition of the variable instead, this text can be used alternatively.

3. Limitations

3.1.1. Variable – Not a single Variable but a set of Variables

If the reference involves not a single variable but can only be achieved by selecting a whole set of variables, we do not include this sample in the corpus. For instance, answers would be the set of nine variables v261 – 269 in the example below:

Reference: "By analyzing the data from ALLBUS (2006) which gathered information about characteristics considered important for awarding German citizenship, I got a first insight into possible boundaries. The result was that previous concepts such as being Christian, being of German origin or being born in Germany were losing in importance against much more flexible concepts such as German language skills, respecting the constitution and not being a criminal."

Variable ID: v261

Variable label: NATURALIZATION:SHOULD BE BORN IN GERMANY

Variable question: I will tell you a few things which may play a role in the decision whether or not to grant German citizenship. Using the scale, please tell me how important these things should be IN YOUR OPINION.

Variable Subquestion: Whether the person was born in Germany.

Variable ID: v262

...

3.1.2. Variable – Unclear which Variable ID

If the mention cannot be clearly attributed to a specific variable, we exclude the sample from the corpus. For instance, for the reference sentence below, it is not clear if the variable is v1275 or v740:

Reference: "A connection was also found, albeit a weaker one, between fathers' religiosity and that of their children. 54% of those who characterized their fathers as "very religious" also characterized themselves as such, and 38% characterized themselves as "religious." 60% of the children of fathers who were "not religious" themselves remained "not religious," and 21.5% more characterized themselves as "not particularly religious." And only 8% of those with "very religious" fathers completely or almost completely abandoned religion.

Variable ID: v1275 or v740?

Variable label v1275: RELIGIOUS DENOMINATION: FATHER

Variable lablel v740: WHAT ROLE RELIGION IN UPBRINGING?

3.1.3. Consecutive Questions

The interviewer has a standardized list of questions, each respondent being asked the same questions in the same order. Some questions can only be answered given the previous question. For instance, an interviewer might first ask question v11 and then v12. However, the v12 survey question cannot be understood without v11. The crucial information to clearly determine the correct variable ID is hidden (available only in the label of the variable).

Reference: "We run least squares regressions of attractiveness on anthropometric measures and several groups of control variables, including age, region, year, interviewer fixed effects, number of children, and health status.

Variable ID: v11

Variable label: CURRENT ECONOMIC SITUATION IN FED. STATE

Variable question: And what about the economic situation in your federal state?

Variable ID: v12

Variable label: RESP. OWN CURRENT FINANCIAL SITUATION

Variable question: And your own current financial situation?

3.1.4. Control Variables

In each survey, a sample of people from a pre-determined population is selected. This is tackled by means of control variables (e.g., respondent's age or gender), which – in contrast to other survey variables – need not be posed as a direct question to the respondent.

Reference: "We run least squares regressions of attractiveness on anthropometric measures and several groups of control variables, including age, region, year, interviewer fixed effects, number of children, and health status.

Variable label: ATTRACTIVENESS OF R., END OF INTERVIEW

Variable question: (Int.: For the interviewer only. Please assess the attractiveness of the respondent one more time. Please come to a spontaneous decision again. Only ONE choice possible.)

Reference: To control the influence of age on these dislikes, age was recoded into three generations as in the case of the ALLBUS (2006) survey.

Variable label: AGE, RESPONDENT

Such control variables are dependent variables and describe the setting of the survey, in particular which groups of people have been interviewed. Generally, we would assume that the information is obligatory and each author reporting on a survey gives information on the underlying control variables as well. They are the most frequent types we observed in our corpus (English: 47 variables with 60 occurrences), but not always clearly stated in the publication and often might refer to one of several variables (see example below).

Reference: The juxtaposition of the respondents' estimation of their own religiosity from 1 (not religious) to 10 (very religious) and the number of their children reveals a clear connection.

Variable label: v1504: Respondent: Number of children

Variable label: v1505: Number of biological children

Variable label: v1507: Number of biological + other children

3.1.5. Identification of Variables vs. Non-Variables

We would like to annotate the mention that best matches the variable. However, sometimes 'vague' variable mentions occur in the vicinity of a variable that elaborate on a certain aspect of the variable only. These should be excluded from the task of "Variable Detection". However, since they provide valuable context information, we decided not to eliminate them completely from the corpus. We marked them by an attribute so that they can be skipped, if desired. An example is listed below:

Reference (Noskip): As our source of data, we chose the ALLBUS-survey from 2002, which was conducted throughout Germany and contained questions pertaining to family as well as to religiosity, income and educational circumstances.

Likewise, the reference to v18 "Political Satisfaction with federal government" is only partly covered in the sentence below:

Reference (Noskip): The quantitative models develop the analysis. In every model, the dependent variable is political satisfaction