The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Variables Extraction in Natural (English) Language Through Possessive Relationships

Danilo Eidy Miura
The University of Tokyo

Teruaki Hayashi
The University of Tokyo

Yukio Ohsawa
The University of Tokyo

The already highlighted importance of the 'flow' of data in the Market of Data brings needs of development of ways to better explore the utilization of data. Aware of the existence of rich knowledge stored and shared in text format, this paper aims to propose a form of representation of variable names that can be identified in natural language written knowledge. With the use of possessive relationships between words in Noun Phrases, we supported the representation of variable name relating a variable to a thing or event. A simple experiment was performed to demonstrate the efficacy of the proposed representation supported by Data Jacket Store, where we can find well-form variable names under the name of Variable Labels.

1. Introduction

The Chance Discovery in the Market of Data is a field of research that aims to design the flow of data in the society through creative methods and find hidden patterns. From the generation, through supply, processing, distribution to utilization of data, discoveries can enhance the demands and pull the consumption of data in the society. Therefore, the perception of the value of data through its utilization is the essential force to innovate in market. Although data is offered and advertised in online repositories, such as Data Jacket Store (Hayashi & Ohsawa, 2015), potential users of data may not be aware of the potential utilization of the offered data. Knowledge of data analysis and processing is required to explore the applications of data.

In order to support data users or suppliers to assess the value of their data, the exploration of potential utilization is the basis for valuing what is not in use yet. To support users in the exploration of potential use of data, we aim to recognize potential use of data from repositories of knowledge recorded in texts, such as research papers and data analytics reports. In this paper we explain the formal representation of data as a well formed variable name (WFVN) that can be used to discover potential new variable labels to name data.

In the field of Knowledge Engineering, the recognition of variables is an essential step to design the knowledge-based system. Implemented systems will contain well-formed variables and knowledge representation. But in natural language, the variables may be expressed in free style, enabling a direct codification to computer-readable language. In order to get the advantage of accumulated knowledge written in natural language, the variable identification and its formal representation may allow knowledge-based systems users to explore possibilities of data utilization.

Contact: Danilo Eidy Miura, The University of Tokyo, daneidyucb@gmail.com

2. Related Work

2.1.Knowledge Representation

In previous works in knowledge representation (Studer et al., 1998, Davis et al., 1993), we learned that the representation of the knowledge depends on the intended task to perform, giving the limited capacity to codify the complete reality of the represented knowledge. Therefore, the representation of the knowledge should be limited and defined according to convenience to the given task. In the definition of the representation, the level of formalism of the representation may enable the use of the represented knowledge in different ways (Guarino, 1995). In this study, the functional approach to representation was adopted to enable further analysis of the represented knowledge (Hayashi & Ohsawa, 2015), as defined in later section.

2.2.Named Entity Recognition

In Natural Language Processing (NLP), Named Entity recognition and Classification (NERC) is a kind of task that identifies entities according to predefined classes. The use of textual features support the identification of entities.

Possessive Noun Phrases (PNP) are the expressions for possessive relationships possessor-possessed. With the use of textual features, such as markers and syntax , relationships between elements of the sentence can be identified and named. According to WALS (Nichols & Bickel, 2013), there are 4 main locus of marking in PNPs:

1)Possessor is head-marked,
2)Possessor is dependent-marked,
3)Possessor is double-marked, and
4)Possessor has no marking.

Markings in PNP explain the existence of various forms of expression of possessive relationships, given the emphasis on of different elements in the phrase.

Given the nature of variables in data analysis in the Market of Data, possessive expressions including pronouns, which are relevant in narratives, are not relevant to our task. The focus of this study is the identification of variables of entities and events, and not possessions of people.

3. Variable Identification

Our aim is to identify and extract well-formed variable names (WFVN) from the natural language text and discover potential variable labels. Recognizing essential elements of a PNP, and defining their relationship and roles, we formally represent the knowledge of the variable.

In order to understand the WFVN, let's consider the variable as an abstract sense of varieties (attributes, properties, features, qualities, etc.) that needs a paired representative (thing of event) that provides a more concrete sense of what variables may vary to. Between the pair, should exist a possessive relation that places a variable as a qualifier of the concrete sense.

Let PNP be represented by the expression *has* (*x* , *y*) where *x* is the possessor noun and *y* is the possessed noun. For the identification of PNP as a WFVN, the Formal Representation of a Variable should satisfy the following conditions of possessor should be a thing (or event) and possessed should be a variable:

$$\text{has}(x, y) \wedge E(x) \wedge V(y) \rightarrow \text{WFVN}(\text{has}(x, y)) \quad (1)$$

1) has (x , y) : Possessor has possessed.
2) *E (x) : Possessor is a thing or an event.*
3) V (y) : Possessed is a variable.

In possessive nouns phrases, we can identify the noun that represents possessor and the noun that represents possessed. To satisfy the condition 1, the possessive relationship will be identified with three patterns of PNPs, as follows:

a)Pd + of + Pr, (ex. temperature of water)
b)Pr + 's + Pd, (ex. water's temperature)
c)Pd + Pr (ex. water temperature)
Pd: Possessed noun
Pr: Possessor noun

In order to satisfy the conditions 2 and 3, it is needed to take in account the relation between the noun and their relative abstractness and concreteness. The relation between Pr and Pd should make a clear distinction between their senses and provide both abstract and concrete sense. This consideration regards the fact of PNPs without clear distinction may not a full sense of the variable:

A) Level of temperature (double abstract)
B) Pool water (double concrete)

In the example A, the level of abstraction of both nouns are high, and we don't have a full sense of the variable, missing the concrete sense of and event (ice, melting, boiling, condensing, …)

In the example B, the problem is on the lack of abstraction. Since both nouns provide concrete sense, we don't have a clear idea possession. This example allows us to have interpretation without possession:

Instead of *Water of pool*,
Interpret W*ater in pool*

4. Experiment on Data Jacket Store

An experiment was designed to verify the performance of the identification of WFVN. Possessive relationships can be identified with use of the three patterns of PNPs, defined before. And regarding the conditions 2 and 3 discussed before, we considered the use of Wordnet hypernyms (Scott & Matin, 1998) as features to establish the concrete-abstract relationship between the nouns in the phrase.

Using a database containing natural language contents and variable names, we could attribute a score to the candidates of variables. Data Jacket Store is a catalog of more than 1000 Data Jackets (Hayashi & Ohsawa, 2015), digest information about the datasets that contain natural language description of the data, as well as variable labels.

In the experiment, we tested the use of hypernyms as features to distinguish PNPs that represent WFVN from the others. Assuming features of WFVN supports the identification of new variable names with similar features, we defined the probability of a new PNP be a WFVN is defined by the probability of the new PNP to have similar hypernyms of WFVN.

Given *n* as a noun in a new PNP, H as a *i* number of hypernyms of *n*, extracted from Wordnet, *v* is the condition of being a variable, and *e* is the condition of being a thing (or event). We define the probability of role (*v* or *e*) of the noun in a new PNP as the average of the probabilities of each PNP's hypernym to be a variable's hypernym:

$$P(v|n) = \sum_{1}^{i} P(v|Hi) * P(Hi|n) \quad (2)$$

$$P(e|n) = \sum_{1}^{i} P(e|Hi) * P(Hi|n) \quad (3)$$

The probability of a given hypernym to be a variable's hypernym can be defined by the probability of the given hypernym be the VLs' hypernym.

4.1. Procedures

The experiment was design to demonstrate the performance of the Classifier with the use of trained data.

1) Define the probability of hypernyms to be variable's hypernyms by the distribution of hypernyms of VLs of DJ Store.

2) Identify and extract PNPs in the DJ's outlines.

3) Calculate the probability of the PNPs to be variables, according to the equation in previous session.

4) Classify PNPs according to the existence in VLs list and satisfaction of the following criteria:
   $P (v | n) > P (\neg v | n)$

5. Results of DJ Store Experiment

In total of 1032 DJ Outlines, 8660 PNPs were identified according to the three patterns of PNP. In total, DJ Store provided 5836 Variable Labels from which 3788 different representations were extracted.

The formal representation of the variable names solved the problem of variances in the expression of possessive relationships due to the markings of the language (Nichols & Bickel, 2013). The formal representation eliminates markings and different patterns of PNP are represented in the same way. Possessive Noun Phrases such as *temperature of water*, *water's temperature* and *water temperature* will be uniquely represented as has (water , temperature).

Regarding the performance of the experimental test to discover potential Variable Labels, discovery is shown as the identification of variables that does not exist in DJ Store VL list. Using the formal representation of variable, we could discover 2017 new PNPs that satisfied the defined criteria. It suggests the information from DJ Outlines shows latent Variable Labels that can be considered in the utilization of that data.

Examples of new identified variables:

Temperature of air,
Efficiency of fuel,
Number of climbers,
Number of surgeries, …

6. Discussion and Future Work

In this paper, the proposal of use of Possessive Relationships as feature of variable names was demonstrated through a simple experiment using DJ Store. And results point to a potential use of the possessive relations as feature to identify variables in text. But it suggests the need of improvements in the selection of candidate relations.

In future work, we aim to refine variable selection with better understanding of types of variables and more accurate algorithms. We also should consider variables that are not considered measurable, such as classes or unstructured data.

Acknowledgements

This work was made possible thanks to to the initiatives of Data Jacket Promotion Work Group which provided the access to the DJ Store data for experiments and it was funded by JST CREST No. JPMJCR1304, JSPS KAKENHI JP16H01836, and JP16K12428, and industrial collaborators.

References

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. International journal of human-computer studies, 43(5-6), 625-640.

Johanna Nichols, Balthasar Bickel. 2013. Locus of Marking in Possessive Noun Phrases. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/24, Accessed on 2019-01-28.)

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. Data and knowledge engineering, 25(1), 161-198.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation?. AI magazine, 14(1), 17.

Hayashi, T., & Ohsawa, Y. (2015). Knowledge Structuring and Reuse System Using RDF for Supporting Scenario Generation. Procedia Computer Science, 60, 1281-1288.

Scott, S., & Matwin, S. (1998). Text classification using WordNet hypernyms. Usage of WordNet in Natural Language Processing Systems.