

Time series from textual instructions for causal relations discovery

Documentation

15 July 2015

1 General information

Experiment title	Time series from textual instructions for causal relations discovery
Experiment id	D20150330-Causal_Relations-KY
Principal investigators	Kristina Yordanova
Affiliation	Mobile Multimedia Information Systems, Computer Science, University of Rostock
E-Mail	kristina.yordanova@uni-rostock.de
Date	30 March 2015
Type	Measurement
Location	Albert-Einstein-Straße 21, 18059 Rostock, Germany
Keywords	textual instructions, causal relations, time series
Language	English
Rights	CC BY 4.0

1.1 Objective

Identify causal relations in textual instructions with the help of time series analysis.

1.2 Problem Statement

One aspect of ontology learning methods is the discovery of relations in textual data. One kind of such relations are causal relations. Our aim is to discover causations described in texts such as recipes and manuals. There is a lot of research on causal relations discovery that is based on grammatical patterns. These patterns are, however, rarely discovered in textual instructions (such as recipes) with short and simple sentence structure. Therefore we use time series to discover causal relations. To do that, each word of interest in the text is converted into time series that represent how often and in which time stamp this word appears in the text. Then a time series analysis can be applied to discover causal relations.

2 Description

This Dataset was generated from textual instructions that describe recipes, manuals, and experiment instructions. It contains 20 datasets, each of which consists of time series representing different text elements. The textual instructions from which the dataset was obtained are as follows:

- The cooking recipes texts were found at <http://www.bbc.co.uk/food/recipes/>
- The washing machine instructions were found at <http://www.miele.co.uk/Resources/OperatingInstructions/W%203923%20WPS.pdf>
- The coffee machine instructions were found at http://www.cn.jura.com/service_support/download_manual_jura_impresa_e10_e20_e25_english.pdf
- The pizza, brownies, and salad instructions are based on the video log of the CMU Grand Challenge dataset <http://kitchen.cs.cmu.edu/>
- The carrots instructions are based on the video log from the kitchen experiment described in: Frank Krüger, Martin Nyolt, Kristina Yordanova, Albert Hein, and Thomas Kirste. "Computational state space models for activity and intention recognition. a feasibility study." PLoS ONE, 9(11):e109381, 11 2014. [1]

The results from the dataset can be found in the paper: Kristina Yordanova. "Discovering causal relations in textual instructions". In Recent Advances in Natural Language Processing, Hissar, Bulgaria, September 2015 [2].

2.1 Data format

The `Raw/` folder contains 20 folders, one for each instruction. It contains the time series generated from the textual instructions. For each dataset there are the following files:

- *actionsCorpusMatrix_*.txt*. It contains the time series for all actions identified in the text
- *actionObjMatrix_*.txt*. This file is generated for each object in the textual instructions. It contains the time series for each action that was in "direct object" relation with the identified object
- *nounsCorpusMatrix_*.txt*. This file contain the time series for all nouns in the text.
- *wordsCorpusMatrix_*.txt*. This file contains the time series for all adjectives or past tense verbs in the text. They describe the state of the action.

The `annotation/` folder contains the manually identified causal relations between words in the text. The arrow shows the direction of the causation (e.g. "take" -> "put" means take causes put).

2.2 Dataset generation

The dataset has been obtained based on the algorithm described in "Discovering causal relations in textual instructions" by K. Yordanova. More precisely, the textual instructions were annotated with POS-tags and dependencies. These were then used to identify the elements of interest (actions, objects, states, and the relations between object and action). Each of this elements was then converted into time series by looking at the number of occurrences of this element in each sentence. In that way a time series has length that equals the number of sentences in the dataset and where each time stamp is represented by a sentence.

3 Bibliography

References

- [1] Frank Krüger, Martin Nyolt, Kristina Yordanova, Albert Hein, and Thomas Kirste. Computational state space models for activity and intention recognition. a feasibility study. *PLoS ONE*, 9(11):e109381, 11 2014. URL: <http://dx.doi.org/10.1371/journal.pone.0109381>, doi:10.1371/journal.pone.0109381.
- [2] Kristina Yordanova. Discovering causal relations in textual instructions. In *Recent Advances in Natural Language Processing*, Hissar, Bulgaria, September 2015. accepted.