Verb Sense Disambiguation for Densifying Knowledge Graphs in Earth Science

Verb Sense Disambiguation for Densifying Knowledge Graphs in Earth Science

Ashish Acharya (1), Carson Davis (1), Derek Koehl (1), Muthukumaran Ramasubramanian (1), Shubhankar Gahlot (1), Iksha Gurung (1), Rahul Ramachandran (2)

1 - University of Alabama in Huntsville, 2 - National Aeronautics and Space Administration

Problem Statement

When sentences from Earth science journal articles are broken down into their semantic components and loaded onto a graph, the relationships among these semantic components are

Proposed Solution

In this presentation, we show a technique where we disambiguate the meaning of a verb in a given sentence using word2vec, with an aim to consolidate it into one among a limited number of synonym sets. This leads to a denser graph and more matches for a given query.

Methodology

Knowledge graphs have become an important tool, both for representing knowledge and for the retrieval of information. Fundamentally, they are semantic networks that represent entities and relationships in the form of nodes and edges. A large corpus of natural language text can be broken down into various entities and relationships to form a useful knowledge graph. Existing research breaks down text into a subject, object, and verb relationship triple. Although this is a useful first step, it loses much of the original contextual information encoded within the text. Our process uses a novel 7-tuple approach, in which elements of sentences are programmatically parsed into seven categories: initiator, impacted, receiver, beneficiary, result, and context. In this presentation, we show a knowledge graph built using this 7-tuple processing of an Earth science corpus. We explain the techniques used to create the graph and analyse its information retrieval capability while assessing the accuracy and limitations of the results.

Results

Future Tasks

. Use the verb disambiguation for verb classification in the earth science knowledge graph

. Check whether the queries are now better after applying verb disambiguation/classification

Acknowledgements

This would not have been possible without the support of software/libraries/algorithms such as:

. Python

. WordNet

. spacy

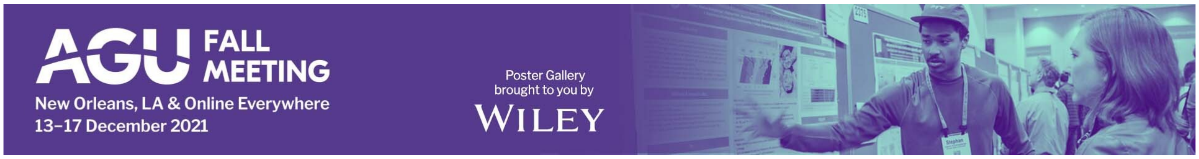
. matplotlib

. Jupyter notebook

Ashish Acharya (1), Carson Davis (1), Derek Koehl (1), Muthukumaran Ramasubramanian (1), Shubhankar Gahlot (1), Iksha Gurung (1), Rahul Ramachandran (2)

1 - University of Alabama in Huntsville, 2 - National Aeronautics and Space Administration

PRESENTED AT:



AGU FALL MEETING
New Orleans, LA & Online Everywhere
13–17 December 2021
Poster Gallery brought to you by
WILEY

PROBLEM STATEMENT

We begin with an ambitious goal: to create a knowledge graph that spans the entire discipline of Earth science. In order to achieve this, we need to apply Natural Language Processing (NLP) techniques on Earth science journal articles to extract their semantic components for the graph.

When sentences from Earth science journal articles are broken down into their semantic components and loaded onto a graph, the relationships among these semantic components are represented by the verbs in the sentences.

However, since there are multiple verbs in English that can be used to denote the same meaning, the knowledge graph can become sparse and so can the results when we query the graph. In order to ensure quality results, it would be desirable to consolidate similar verbs into a single "class".

So, this is the problem at hand: how do we make sure that multiple verbs that mean the same thing are represented as a single class of verb in the knowledge graph? Or in other words, how do we distinguish which meaning a particular verb takes given a particular sentence?

In this poster, we demonstrate a potential technique to solve this problem.

PROPOSED SOLUTION

To compare the numerical representation of the verb in a given sentence to the numerical representation of its potential meanings to figure out which meaning the verb is closest to given the context of the sentence.

METHODOLOGY

First, a brief list of definitions for the terms used in the solution is in order.

A **knowledge graph** is a graph constructed by extracting concepts and ideas from data sources (here: journal articles) using NLP techniques and connecting them together in a graphical data structure which can be traversed.

A **synset** is a potential meaning that a verb can take. For example, the word **break** has a large number of synsets, each of which has its own definition and example usage. The phrase ‘break into’ is completely different from ‘break down’ or just ‘break’. **Wordnet** gives us a library of possible synsets for each verb in English.

Word2vec is a mathematical technique that allows us to numerically represent linguistic features such as words and phrases. This allows us to compare two such linguistic features and figure out how close they are to each other.

The solution we propose here involves extracting the main verb from a given sentence using well-known NLP techniques. Then, we get a list of possible synsets for the verb using Wordnet.

Using word2vec, we then do two things: 1. We get the numerical representation of each synset and, 2. We get the numerical representation of the verb in the given sentence and add to it the numerical representation of the context that surrounds it (subject, object, and preposition).

Now, we compare the numerical representation of the verb to the numerical representation of the synsets to figure out which particular synset is closest to the verb. At this point, we have figured out the meaning of the verb in the given sentence and this can be used to consolidate verbs with similar meanings into well-defined classes.

The diagram below shows the process visually:

RESULTS

In order to quantify our results, we came up with a test set that includes 20 verbs in English with a very high number of synsets. The test set includes verbs such as **break**, **make**, and **give.** For each verb, we came up with two radically different sentences and hand-picked the correct synset for each sentence.

We compared the synsets predicted by our algorithm to the hand-picked correct synsets to come up with accuracy numbers.

The algorithm we've come up with is still in its nascent stage and not ready for use in production. While it performs well for certain verbs with fewer synsets, it does quite poorly for verbs where the synsets are numerous.

Using our current techniques, we were able to get a 26% accuracy rate, which seems low, but considering we are testing against the toughest to classify verbs in the English language, and the fact that these are preliminary results that can be improved, we believe we're on the right track. Some possible improvements to the algorithm are described in the future tasks section.

FUTURE TASKS

- Include the definition of a synset when creating its word2vec embedding in order to distinguish synsets from one another

- Pre-process phrasal verbs (break into, break down)

- Improve metrics for the results

- Use the verb disambiguation for verb classification in the earth science knowledge graph

- Check whether the queries are now better after applying verb disambiguation/classification

ACKNOWLEDGEMENTS

This would not have been possible without the support of software/libraries/algorithms such as:

- Python
- Word2vec
- spacy
- matplotlib
- Jupyter notebook

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.