Technische Universität Berlin
Fakultät IV: Elektrotechnik und Informatik
Institut für Telekommunikationssysteme
Fachgebiet Verteilte offene Systeme

Masters Thesis in Computer Science

Extending Semantic Hypergraphs by neuronal semantic similarity matching to ???

Max Reinhard

January 27, 2023

Supervised by Prof. Dr. Manfred Hauswirth
Additional guidance by Prof. Dr. Camille Roth* and Dr. Thilo Ernst†

*Centre Marc Bloch (An-Institut der Humboldt-Universität zu Berlin)
†Fraunhofer-Institut für offene Kommunikationssysteme

Abstract Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Contents

1 Introduction 4

2 Fundamentals and Related Work 5

2.1 Semantic Hypergraph 5

2.1.1 Notation 5

2.2 Text Similarity Measures 5

2.3 Semantic Similarity 5

2.4 Embedding-based similarity 5

2.4.1 Embedding types 5

2.4.2 Embedding distance measures 5

3 Solution Approach 6

3.1 `semsim` Functional Pattern 6

3.1.1 Pattern-wise Similarity Threshold 6

3.2 Fixed word-vector based 6

3.2.1 Single Word 6

3.2.2 Multi Word 6

4 Implementation 7

4.1 Integration into the SH Framework 7

4.2 External Libraries and Models 7

4.3 SH Notation 7

5 Results and Evaluation 8

5.1 Case Study: Conflicts 8

6 Conclusion 11

1 Introduction

- Context: The big problem
- Problem statement: The small problem
- Methodology / Strategy
- Structure

Notes:

- Huge amounts of text, which can provide insight about stuff
- Automatic tools can provide assistance for humans to process all the text
- This generally means filtering the original text corpus or otherwise reducing amount of information the information that has to be processed by humans
- Filtering introduces a bias
- Especially for scientific purposes it is relevant to mitigate bias or at least understand what bias has been introduced (to make it transparent)
- Semantic Hypergraphs can be a valuable tool for that because...

A Semantic Hypergraph (SH) [MR21] is...

2 Fundamentals and Related Work

2.1 Semantic Hypergraph

2.1.1 Notation

Square bracket notation

2.2 Text Similarity Measures

tf-idf, etc.?

2.3 Semantic Similarity

2.4 Embedding-based similarity

2.4.1 Embedding types

Fixed word embeddings

Contextual embeddings

Sentence embeddings

2.4.2 Embedding distance measures

3 Solution Approach

Combining Semantic Hypergraphs with neural embeddings

3.1 `semsim` Functional Pattern

3.1.1 Pattern-wise Similarity Threshold

3.2 Fixed word-vector based

3.2.1 Single Word

3.2.2 Multi Word

4 Implementation

4.1 Integration into the SH Framework

Realisation as functional pattern

4.2 External Libraries and Models

Word2Vec Gensim

4.3 SH Notation

Bracket notation for multi-word Semsim

5 Results and Evaluation

5.1 Case Study: Conflicts

This case study follows the approach presented in [MR21, p. 22] where expressions of conflict are extracted from a SH constructed from a corpus of news titles that were shared on the social media platform *Reddit*. Specifically all titles shared between January 1st, 2013 and August 1st, 2017 on *r/worldnews*[1], which is described as: "A place for major news from around the world, excluding US-internal news."

Pattern 1 is used to extract conflicts between two parties, where the SOURCE shows some form of aggression against the TARGET, potentially regarding some TOPIC:

```
( PRED/P.so,x SOURCE/C TARGET/C [against,for,of,over]/T TOPIC/[RS] ) ∧
( lemma/J >PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P )
```

Pattern 1: Conflict pattern

To investigate whether it is possible to capture the abstract concept of a country using the multi-word `semsim` pattern introduced in 3.2.2, a list of the 20 most populous countries [Wik23] is used:

China, USA, Indonesia, Pakistan, Nigeria, Brazil, Bangladesh, Russia, Mexico, Japan, Philippines, Ethiopia, Egypt, Vietnam, Congo, Iran, Turkey, Germany, France

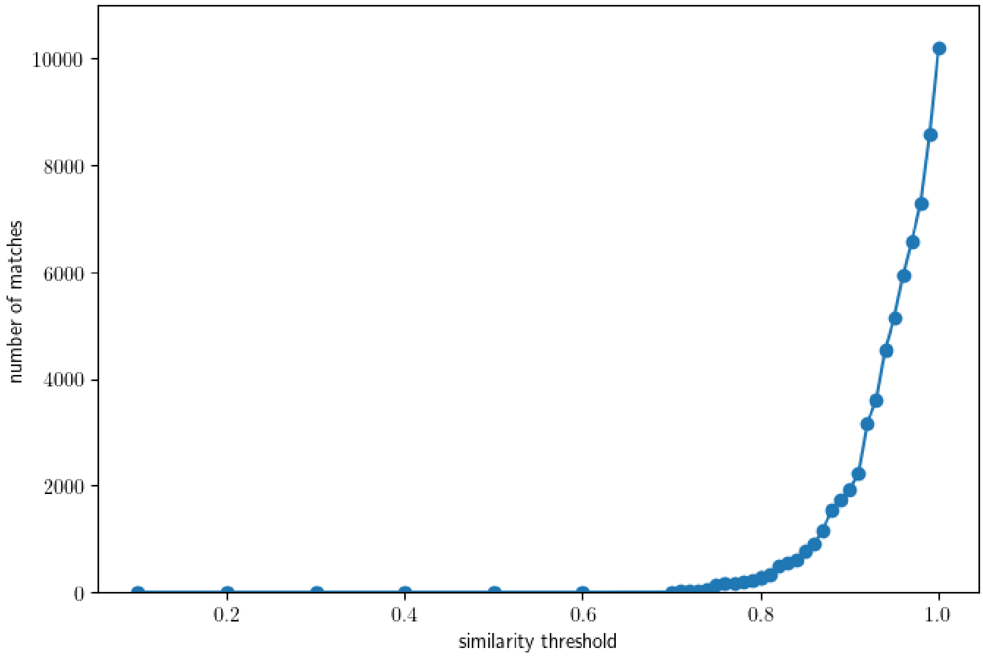
Pattern 2 shows the resulting pattern for conflicts between countries.

```
( PRED/P.so,x SOURCE/C TARGET/C semsim [against,for,of,over]/T TOPIC/[RS] ) ∧
( semsim/J >/PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P ) ∧
( semsim/J >SOURCE/C [china,usa,indonesia,pakistan,nigeria,brazil,bangladesh,russia,
mexico,japan,philippines,ethiopia,egypt,vietnam,congo,iran,turkey,germany,france]/C ) ∧
( semsim/J >TARGET/C [china,usa,indonesia,pakistan,nigeria,brazil,bangladesh,russia,
mexico,japan,philippines,ethiopia,egypt,vietnam,congo,iran,turkey,germany,france]/C )
```

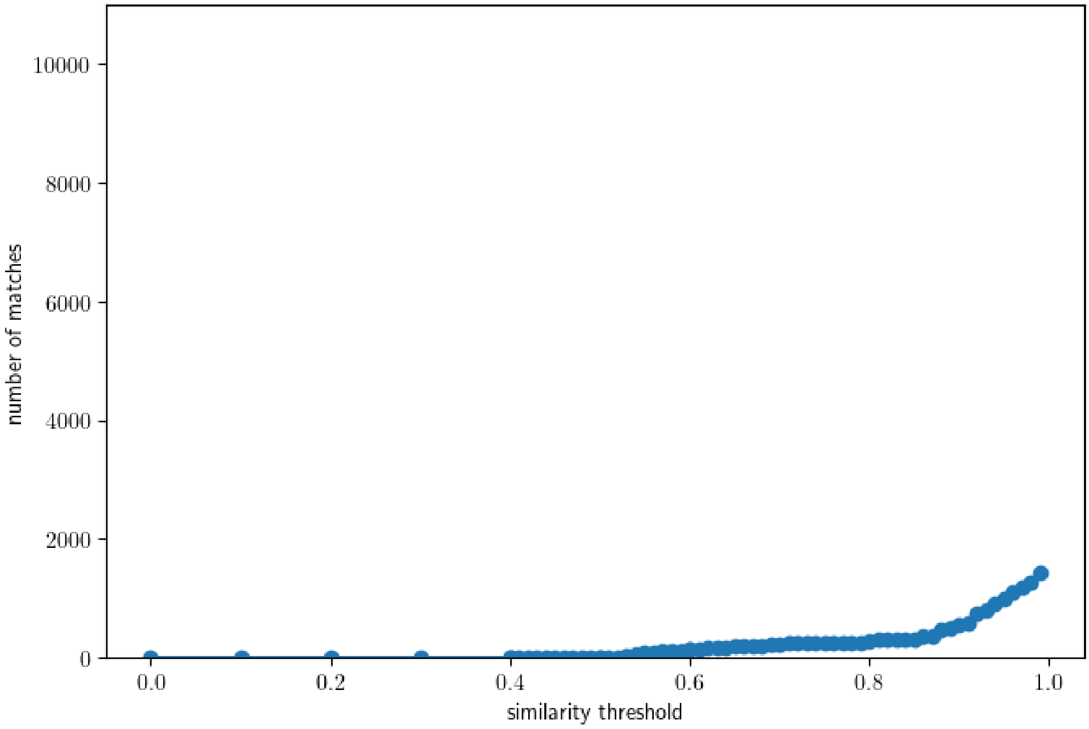
Pattern 2: Country conflict pattern

In **??** the number of matches that result from using this pattern is plotted against the similarity threshold for the `semsim` pattern.

[1] `http://reddit.com/r/worldnews`



(a) Similarity threshold variation for all semsim patterns



(b) Similarity threshold variation only for SOURCE and TARGET (country) `semsim` patterns

Figure 5.1: Number of matches for conflict pattern in relation to similarity threshold

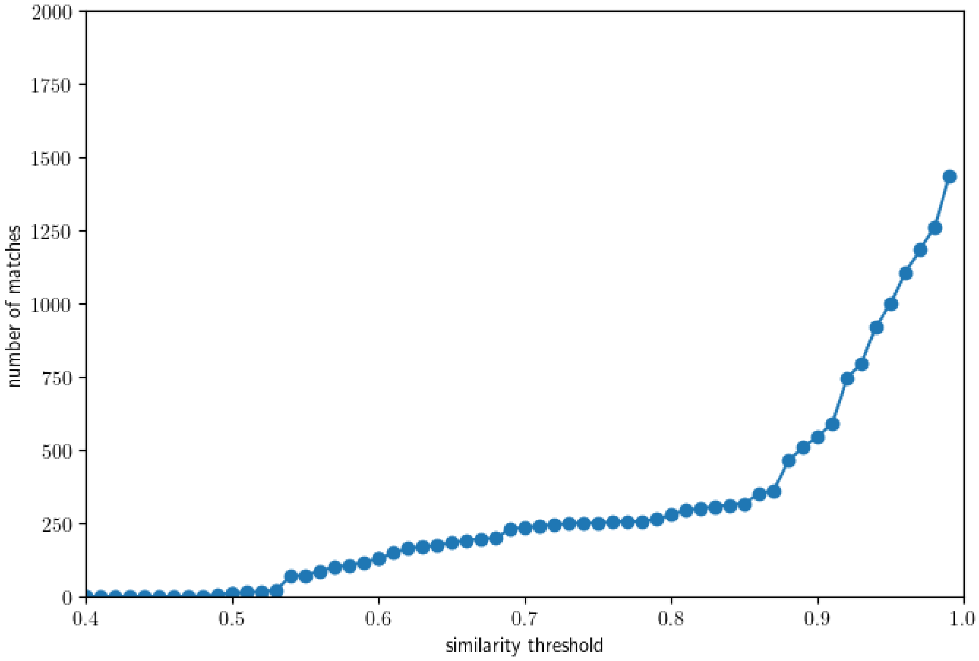


Figure 5.2: Number of matches for conflict pattern in relation to similarity threshold with threshold variation for SOURCE and TARGET (country) `semsim` patterns in the range 0.4 >= threshold < 1.0 (and y-axis limited to 2000 results)

6 Conclusion

Bibliography

[MR21] Telmo Menezes and Camille Roth. *Semantic Hypergraphs.* Feb. 18, 2021. DOI: 10.48550/arXiv.1908.10784. arXiv: 1908.10784[cs]. URL: http://arxiv.org/abs/1908.10784 (visited on 07/19/2022).

[Wik23] Wikipedia. *List of countries and dependencies by population — Wikipedia, The Free Encyclopedia.* http://en.wikipedia.org/w/index.php?title=List%20of%20countries%20and%20dependencies%20by%20population&oldid=1135750154. [Online; accessed 26-January-2023]. 2023.