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

April 13, 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

- 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...

Human life in times of widespread use of the internet and smartphones is most certainly more than ever interspersed with text-based communication...

A Semantic Hypergraphd [MR21] is a form of representation for Natural Language (NL) and therefore knowledge. *NL* sentences can be modelled as a recursive hypergraph which can be represented in a formal language. The framework allows to specify semantic patterns in this formal language which can be matched against an existing *SH*.

The aim of the SH framework is to provide a *open* and *adaptive* framework to analyse text corpora, especially in the domain of computational social science (CSS) [Laz+09]. The framework is *open* in the sense that it's representation formalism is inspectable and intelligible by humans and that the pattern matching follows explicit rules. The framework is adaptive in the sense that the parsing is based on adaptive subsystems (ML-based) and therefore allows for an error-tolerant parsing from *NL* to *SH* in regards to grammatical and syntactical correctness (???).

2 Fundamentals and Related Work

2.1 Semantic Hypergraph

2.1.1 Structure

2.1.2 Syntax

Square bracket notation

2.2 Semantic Similarity

2.2.1 Different Similarity Measures

String Similarity

Lievenstein distance, etc..

Lexical Similarity

tf-idf, etc.?

2.2.2 Types of Semantic Similarity

Lexical Databases

WordNet and alike (not the scope of this work)

2.3 Embedding-based Similarity

2.3.1 Embedding Types

Fixed Word Embeddings

Contextual Ebeddings

2.3.2 Distance Measures

Mean reference vector vs. pairwise distance

3 Solution Approach

Combining Semantic Hypergraphs with neural embeddings

3.1 `semsim` Functional Pattern

pattern works only for atoms

3.1.1 Pattern Matching Process

3.1.2 Pattern-wise Similarity Threshold

3.2 Fixed Word Embedding-based Matching

word2vec via gensim

3.2.1 Single Word

3.2.2 Multi Word

Square bracket notation

3.3 Contextual Embedding-based Matching

3.4 Similarity Threshold Discovery

detect change points in number of matches
see https://github.com/deepcharles/ruptures

4 Implementation

4.1 Integration into the SH Framework

Realisation as functional pattern

4.2 Similarity Threshold

4.3 Tokenization

4.3.1 SpaCy

SpaCy linguistic tokenization (https://spacy.io/usage/linguistic-features how-tokenizer-works) spacy (without transformers) uses an purely rule based (but language depended) tokenizer as far as I understand: https://spacy.io/usage/linguistic-features how-tokenizer-works (the call it linguistic tokenizer)

side note about using different transformer models than the provided one (because i was always confused about this): it it possible to exchange the underlying transformer component for basically every transformer model (as long as it follows the conventions that spacy expects), but you would have to retrain the spacy model to be able to use the task specific heads (like e.g. NER) footnote: https://github.com/explosion/spaCy/discussions/10327

an alignment is provided between the transformer-tokenizer and the spacy-tokenizer lib: https://github.com/explosion/spacy-alignments

footnote: https://explosion.ai/blog/spacy-transformers

4.3.2 WordPiece and SentencePiece

SentencePiece: https://github.com/google/sentencepiece

4.4 Matching edges to token embeddings

string matching

4.5 External Libraries and Models

Here list libs and models to be referenced later.

Word2Vec Gensim SentenceTransformers Transformers SpaCy

4.6 SH Notation

Bracket notation for multi-word Semsim

5 Results and Evaluation

In this chapter...

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 worlds 20 most populous countries [Wik23] is used (listed in descending order by population size):

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

To avoid the repetition of that list in the following pattern, we introduce a variable:

COUNTRIES = [india,china,usa,indonesia,pakistan,nigeria,brazil,bangladesh,russia,mexico,
ajapan,philippines,ethiopia,egypt,vietnam,congo,iran,turkey,germany,france]

Pattern 2: Countries variable

Pattern 3 shows the resulting pattern for conflicts between countries:

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

(PRED/P.so,x SOURCE/C TARGET/C semsim [against,for,of,over]/T TOPIC/[RS]) $\wedge$
(semsim/J >/PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P) $\wedge$
(semsim/J >SOURCE/C COUNTRIES/C) $\wedge$ (semsim/J >TARGET/C COUNTRIES/C)

Pattern 3: Country conflict pattern

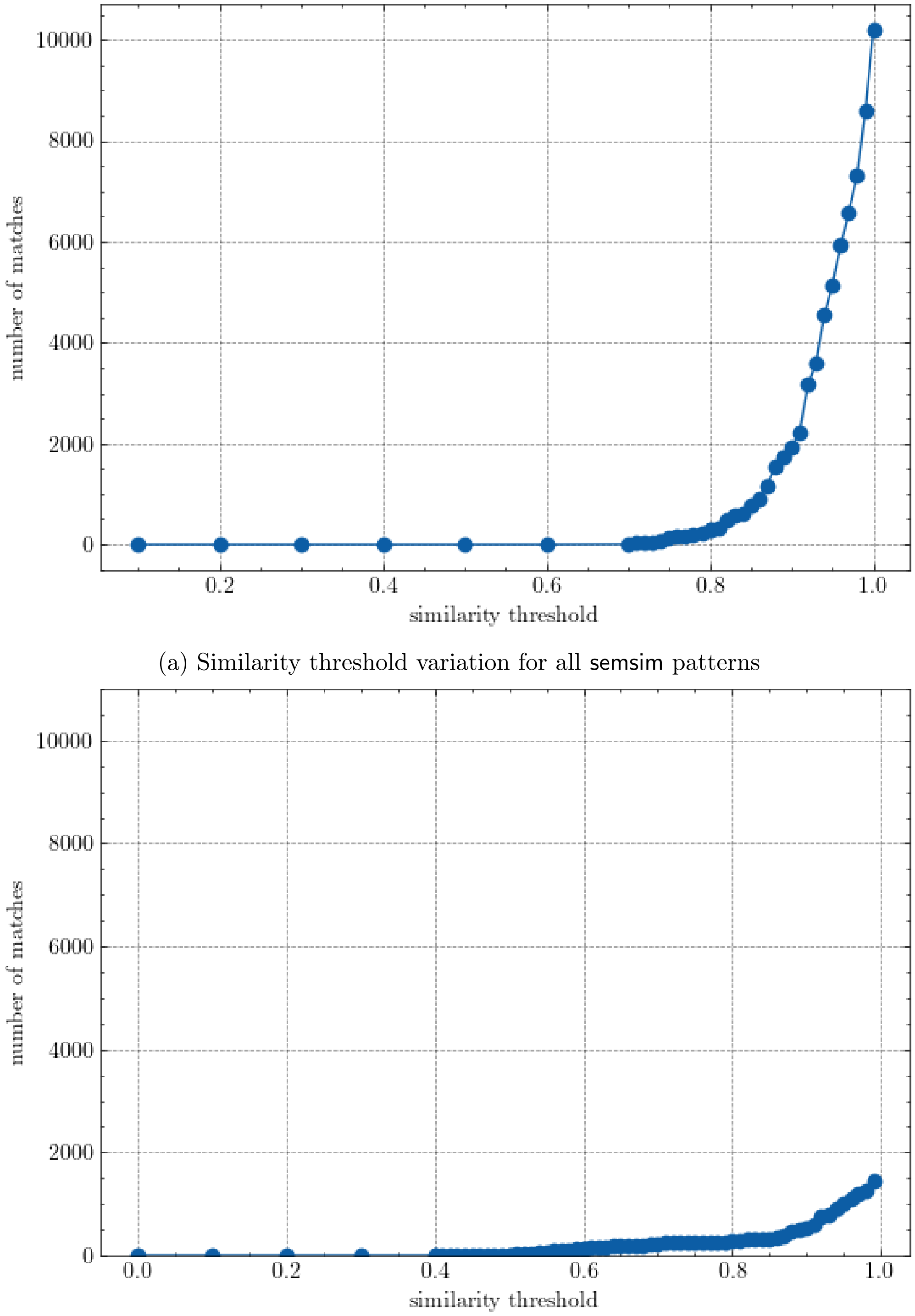
In pattern 4 a sub-pattern specific threshold $t_{sim}^{countries}$ for the countries `semsim` sub-pattern is introduced.

(PRED/P.so,x SOURCE/C TARGET/C semsim [against,for,of,over]/T TOPIC/[RS]) $\wedge$
(semsim/J >/PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P) $\wedge$
(semsim/J >SOURCE/C COUNTRIES/C $t_{sim}^{countries}$) $\wedge$
(semsim/J >TARGET/C COUNTRIES/C $t_{sim}^{countries}$)

Pattern 4: Country conflict pattern

5.1.1 Quantitative Results

Pattern 3 is matched against the described *Reddit r/worldnews* hypergraph. The similarity threshold t_{sim} for the `semsim` function (see 4.2) is varied. t_{sim} is either varied for the entire pattern or for a specific `semsim` sub-pattern.



(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

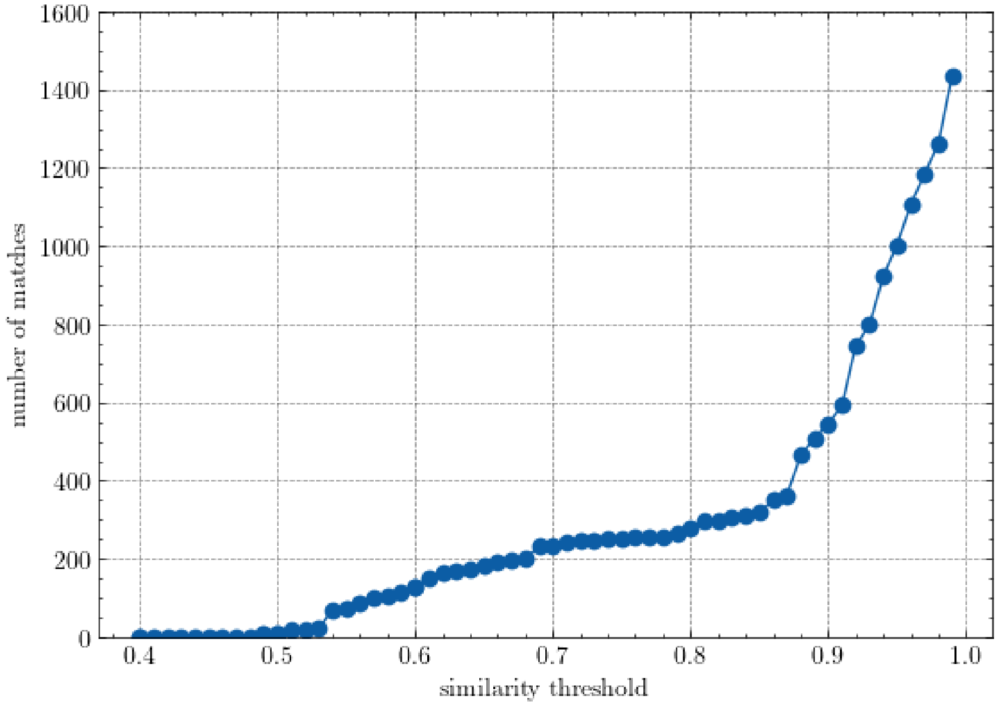


Figure 5.2: Number of matches for conflict pattern in relation to similarity threshold with threshold variation for SOURCE and TARGET (i.e. COUNTRIES) `semsim` patterns in the range 0.4 >= threshold < 1.0 (and y-axis limited to 2000 results)

6 Conclusion

Bibliography

[Laz+09] David Lazer et al. “Computational social science”. In: *Science* 323.5915 (2009), pp. 721–723.

[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.