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

Masters Thesis in Computer Science

Extending Semantic Hypergraphs by Neural Embedding-based Semantic Similarity for Pattern Matching

Max Reinhard
Matrikelnummer: 359417

July 31, 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 (**menezes_semantic_2021**) 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) (**lazer2009computational**). 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 Problem Statement

CSS researches may typically be interested in retrieving statements of specific kind from a text corpus, such as expressions of sentiment of an actor towards some entity or expressions of conflicts between different actors. One approach for performing the retrieval would be to use a system which allows to specify some form of pattern which abstractly represents the statements they are trying to capture. This requires the definition of some form of formal pattern language[1] and possibly the prior transformation of the text corpus into some form of structured format to match against. Another approach is to use a system, which accepts example statements concretely representing the statements that are desired to be retrieved. Those systems may require a large number of positive and negative examples to be able to perform the retrieval. The two types of retrieval systems described here are in tendency situated in the realms of symbolic IR/IE and probabilistic ML/DL respectively.

The SH framework is more situated in the former symbolic realm. In SH text is represented in the form of *hyperedges* (in the following also referred to as *edges* only). These edges are either atomic or they consist of edges themselves, which essentially accounts for the recursive character of the SH. Each edge has a specific *type* from a set of eight different types of which the most trivial two types are probably *concept* (C) and *predicate* (P).

Users of the SH framework (e.g. CSS researchers) can define patterns in the SH formalism to match against a text corpus (e.g. a collection of news articles) that has previously been parsed as an SH. These patterns may among other things specify the structure of the edges that should match it as well as their type (and the types of possible sub-edges). Additionally the actual words that should match need to be specified i.e. the content to match against, if the structure of an edge matches the pattern. There are additional operators in the pattern language such as the wildcard operator *, which can be used e.g. to match every atomic edge edge of a specific type and therefore discard content.

To better illustrate the problem Hyperedge 2 and Hyperedge 1 demonstrate how NL sentences are parsed to SH based on this simplified introduction the the SH representation.

(likes/P ann/C apples/C)

Hyperedge 1: SH representation for the sentence "Ann likes apples"

(likes/P ann/C bananas/C)

Hyperedge 2: SH representation for the sentence "Ann likes bananas"

Hyperedge 1 and Hyperedge 2 both follow the same structure, but differ in the content of the last sub-edge. Both edges are hence matched by Pattern 1, which does not specify content for this sub-edge. The SH pattern language also allows to define a pattern that matches both Hyperedge 1 and Hyperedge 2 via a list of words as in Pattern 2. However

[1] The *Google Search* query language can be seen as a simple example of such a pattern language, albeit with a different use case focus: `https://support.google.com/websearch/answer/2466433?hl=en`

is not possible define a pattern that matches based on some form of *Semantic Relatedness* (SR) or *Semantic Similarity* (SS) (Harispe et al. 2015) regarding content. Referring to the example above this means using the SH framework it is not directly possible to to retrieve every sentences that declares that "Ann likes *some kind of fruit*" or that "Ann likes *fruits similar to apples*". This former would require to provide a comprehensive list of every fruit while the latter would require the user to specify all fruits he deems similar to apples.

(likes/P Ann/C */C)

Pattern 1: "Ann-likes-something" pattern

(likes/P ann/C [apples, bananas]/C)

Pattern 2: "Ann likes apples or bananas" pattern

Utilizing some form of SR/SS regarding to edge content for the matching step would allow users to define more generalising patterns. There exists a great variety of approaches for determining the SR/SS of text, which can generally be divided into *Corpus-based Measures* and *Knowledge-based measures* (Harispe et al. 2015, Section 1.3.2). The latter approaches may generally provide the explicitness in the measurement determination that is desired by CSS researchers. However among the former recent ML-based and especially DL-based approaches have been outperforming most other approaches (Chandrasekaran and Mago 2021). They generally rely on computing a vector space representation (or embedding) of texts which can then be used to calculate their similarity and will therefore be referred to as *Neural Embedding-based Semantic Similarity* (NESS) measures in the following.

Word semantics generally depend on textual context and hence does the SS between words (Harispe et al. 2015, Section 2.2.3). Incorporating contextuality when extending the SH pattern matching process by SS therefore poses a central challenge. Context-dependent SS would allow to specify matching edge content beyond isolated word semantics, although this may not always be desirable or necessary as in the example above.

As illustrated earlier, NESS measures principally do not provide the explicitness that is inherent to the pattern matching process of the SH framework. In the sense of the adaptive-open classification described above an integration of NESS would mean a shift from openness to adaptivity in this regard. While the SH framework generally can be situated in the realm of symbolic approaches, this integration would build a bridge between it and the realm of probabilistic approaches.

3.1 Research Questions

Based on the problem statement outlined above, we pose the following research questions:

3.1.1 Primary Question

R Can neural embedding-based semantic similarity regarding edge content be be integrated into the pattern matching of the Semantic Hypergraph framework to allow for more generalising patterns while providing control over the adaptiveness and therefore loss of explicitness in the matching process?

3.1.2 Secondary Questions

R.1 What neural embeddings model would be the most suitable for accurately assessing semantic similarity within the Semantic Hypergraph pattern matching process while effectively addressing the challenges posed by contextuality?

R.2 To what extent does incorporating neural embedding-based semantic similarity improve the generalization performance (recall) and how does it impact precision when matching a pattern against a set of known desired matching results?

R.3 How can adaptiveness and explicitness of the matching process be effectively and transparently balanced and controlled?

4 Solution Approach

In this chapter we present the approach that was developed to answer the research questions (see section 3.1). Therefore trying to provide a solution to the problem of extending the SH framework by Neural Semantic Similarity Matching, which is described in chapter 3 where the relevancy of this problem for has also been derived.

Combining Semantic Hypergraphs with neural embeddings

4.1 Integration into the Pattern Matching Process

4.1.1 `semsim` Functional Pattern

pattern works only for atoms

4.1.2 Sub-pattern Similarity Thresholds

4.2 Fixed Word Embedding-based Matching

word2vec via gensim

discussion about using transformer models for single word embeddings?

single-word and multi-word reference

Square bracket notation

4.3 Contextual Embedding-based Matching

i generally like your idea of contrasting the discrete and continous space as it allows to point out that there cant be one single point, also for a set of words which represents the meaning, but rather some subspace depending on the specific context Regarding the point of the semantic entities in continuous space being either word- or phrase based, the important difference is, that in case of semsim with context we do not compare the embedding representation of the phrases themselves. rather the sententences/phrases influence the embedding representations of the word (or maybe phrases) I tend to see this a bit like a blurring algo. The meaning of each token starts bleeding into its neighbours.

4.3.1 Context References

4.3.2 Token Mapping

4.4 Similarity Threshold Control

4.4.1 Breakpoint Discovery

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

half-max point and quarter/three-quarter points (percentiles, not quantiles) fit function and search for infliction as well as maximum derivative points, problematic in cases with less continuous change in number of matches.

how to approach this for practical applications?

5 Implementation

5.1 Relevant external Software Libraries used

Here list libs and models to be referenced later.

Word2Vec Gensim SentenceTransformers Transformers SpaCy

5.2 Modules newly added to the SH Framework

5.3 Modifications of the SH Pattern Matching

5.4 Modifications to the Hypergraph database

6 Results and Evaluation

In this chapter...

6.1 Case Study: Conflicts

This case study follows the approach presented in (Menezes and Roth 2021, 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."

Number of news headers in corpus: 479384

Add dataset statistics

Pattern 3 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 3: Conflict pattern

To investigate whether it is possible to capture the abstract concept of a country using the multi-word `semsim` pattern introduced in 4.2, a list of the worlds 20 most populous countries [2] 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 4: Countries variable

Pattern 5 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 COUNTRIES/C) ∧ (semsim/J >TARGET/C COUNTRIES/C)

Pattern 5: Country conflict pattern

[1] `http://reddit.com/r/worldnews`
[2] Based on `https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population`, Accessed on XYZ

In pattern 6 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]) ∧
(semsim/J >/PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P) ∧
(semsim/J >SOURCE/C COUNTRIES/C $t_{sim}^{countries}$) ∧
(semsim/J >TARGET/C COUNTRIES/C $t_{sim}^{countries}$)

Pattern 6: Country conflict pattern

6.1.1 Quantitative Results

Pattern 5 is matched against the described *Reddit r/worldnews* hypergraph. The similarity threshold t_{sim} for the `semsim` function (see **??**) is varied. t_{sim} is either varied for the entire pattern or for a specific `semsim` sub-pattern.

6.1.2 Qualitative Results

Table 6.1: hyper dyper table

Scenario Name	Pattern	Samples	Variable Threshold	Reference Edges	Ref. Edges Source
1_original-pattern	Pattern X	Erdogan slams ridicule of 'Muslims discovered Americas' claim Iran forces 'kill Kurdish rebels on Iraq border Ukraine Accuses Russia of Invasion	-/-	-/-	-/-
2-1_semsim-fix_preds	Pattern X	Pakistani police kill feared militant leader in mysterious pre-dawn shootout Al-Shabaab militants claim responsibility for deadly attack on Garissa University College in Kenya Casualties as Congo troops, UN forces fight rebels	'preds': 0.19 Percentile: 50	-/-	-/-
2-2_semsim-fix_preps	Pattern X	Iranian police have arrested merchants for selling clothing that featured the flags of the United States and Britain, two longtime foes of the Islamic republic Syrian Air Force Strikes kill 38 ISIS fighters Seven Libyan soldiers killed fighting off Islamists near Benghazi: source	'preps': 0.54 Percentile: 50	-/-	-/-

7 Conclusion

8 Future Work

8.1 Implementation Improvements

implemnt multiprocessing, i.e. server process for both hypergraph and semsim matchers. other option would be to leverage python shared memory capabilities but is likely to be less stable and has less scaling potential

8.2 Further Evaluations

Bibliography

Chandrasekaran, Dhivya and Vijay Mago (Feb. 18, 2021). "Evolution of Semantic Similarity—A Survey". In: *ACM Computing Surveys* 54.2, 41:1–41:37. ISSN: 0360-0300. DOI: 10.1145/3440755. URL: https://dl.acm.org/doi/10.1145/3440755 (visited on 06/17/2023).

Harispe, Sébastien et al. (2015). *Semantic Similarity from Natural Language and Ontology Analysis*. DOI: 10.2200/S00639ED1V01Y201504HLT027. arXiv: 1704.05295 [cs]. URL: http://arxiv.org/abs/1704.05295 (visited on 06/19/2023).

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