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

March 6, 2024

Supervised by Prof. Dr. Manfred Hauswirth
Additional guidance by Prof. Dr. Camille Roth*
and Dipl.-Math. 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 **5**
- 1.1. References from the Future . . . 5
- 1.2. Expose intro . . . 6

2. Fundamentals and Related Work **7**
- 2.1. Semantic Hypergraph . . . 7
  - 2.1.1. Structure . . . 7
  - 2.1.2. Syntax / Pattern Matching . . . 7
- 2.2. Semantic Similarity . . . 7
  - 2.2.1. Different Similarity Measures . . . 7
  - 2.2.2. Types of Semantic Similarity . . . 8
- 2.3. Embedding-based Similarity . . . 8
  - 2.3.1. Embedding Types . . . 8
  - 2.3.2. Distance Measures . . . 8

3. Problem Statement **9**
- 3.1. Research Questions . . . 11
  - 3.1.1. Primary Question . . . 11
  - 3.1.2. Secondary Questions . . . 11

4. Solution Approach **12**
- 4.1. Integration into the Pattern Matching Process . . . 12
  - 4.1.1. SemSim Functional Pattern . . . 12
  - 4.1.2. Sub-pattern Similarity Thresholds . . . 12
- 4.2. Fixed Word Embedding-based Matching . . . 12
- 4.3. Contextual Embedding-based Matching . . . 13
  - 4.3.1. Context References . . . 13
  - 4.3.2. Token Mapping . . . 13
- 4.4. Similarity Threshold Control . . . 13
  - 4.4.1. Breakpoint Discovery . . . 13

5. Implementation **14**
- 5.1. Relevant external Software Libraries used . . . 14
- 5.2. Modules newly added to the SH Framework . . . 14
- 5.3. Modifications of the SH Pattern Matching . . . 14
- 5.4. Modifications to the Hypergraph database . . . 14

6. Evaluation **15**
- 6.1. Case Study: Conflicts . . . 15
  - 6.1.1. Expressions of Conflict . . . 15
  - 6.1.2. Reddit Worldnews Corpus . . . 15
  - 6.1.3. Semantic Hypergraph Patterns . . . 16

6.2. Conflict Dataset . . . 18
6.2.1. Base Edge Set . . . 18
6.2.2. Desired Characteristics . . . 19
6.2.3. Construction Process . . . 19
6.2.4. Edge Set Comparison . . . 21
6.3. Evaluation Process . . . 21
6.3.1. Evaluation Run Configurations . . . 21
6.3.2. Evaluation Metrics . . . 23
6.4. Evaluation Results . . . 24
6.4.1. Best F1-Score based Evaluation Run Comparison . . . 25
6.4.2. Evaluation Metric vs. Similarity Threshold . . . 26
6.4.3. Best F1-Score vs. Number of Reference Edges . . . 28
6.4.4. Predicate Lemma based Evaluation Run Comparison . . . 28
6.5. Result Discussion . . . 31
6.5.1. Retrieval Performance Improvement . . . 33
6.5.2. Retrieval Precision Behaviour . . . 34
6.5.3. Contextual Differentiation Ability . . . 34

7. Conclusion **35**

8. Future Work **36**
8.1. Conceptual Improvements . . . 36
8.2. Implementation Improvements . . . 36
8.3. Further Evaluations . . . 36

A. Appendix **39**
A.1. Reference Edge Sets . . . 40
A.2. Best F1-score based Eval. Run Comparison Tables . . . 45
A.3. Evaluation Metric Scores vs. Similarity Threshold Plots . . . 47

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...

Great progress has been made in the made in NLP, IR and IE in the past decade. This advancement of the state-of-the-art can primarily be attributed *Deep Learning* based methods, often also referred to as *Neural Networks*. (Hirschberg and Manning 2015; Min et al. 2023; Young et al. 2018)

1.1. References from the Future

open-opaque / strict-adaptive categorisation dimensions

1.2. Expose intro

A significant part of the social world is nowadays being represented by digitally manifested text. Examples for this range from instant messaging, social media and any form of collective web activity to encyclopaedic websites, digitized libraries and government intelligence. The amount and richness of available social text makes it a valuable data source for social science research while simultaneously creating an interest in automatic systems to analyze these texts on a large scale (Evans and Aceves 2016). Such research can be understood as part of the domain of *Computational Social Science* (CSS) (Lazer et al. 2009).

Systems based on techniques from the field of *Natural Language Processing* (NLP), as well well as the interlinked fields of *Information Retrieval* (IR) and *Information Extraction* (IE), have demonstrated great success in a variety of task related to text analysis. This success is largly attributed to the advancements of applying of *Machine Learning* (ML) and especially *Deep Learning* (DL) methods to text (Hirschberg and Manning 2015) (Qiu et al. 2020). While being very effective at predicting or decision making, ML- and specifically DL-based systems generally do not deliver an explanation for their judgement and can mostly be viewed as "black box models" that are not transparent in their prediction or decision making process (Rudin 2019). Conversely this transparency and explainability is of high interest in CSS applications such as predicting political opinion based on social media activity (Wilkerson and Casas 2017).

The *Semantic Hypergraph* (SH) (Menezes and Roth 2021) is a framework for representing and analyzing Natural Language (NL). *NL* sentences can be modelled as an ordered, 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. It aims to provide an *open* and *adaptive* system to analyse text corpora, especially in the domain of CSS. The framework is *open* in the sense that its 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 built from adaptive, ML-based subsystems 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

synonymes: SH, hypergraph, graph.

Edge

Edge Content

Root edge / Sequencen (i use the term "sequence root edge")

2.1.2. Syntax / Pattern Matching

wildcard operator

Variables

Square bracket notation –> word lists

functional patterns -> lemma

> operator for innermost atom

—

Differences between the formal notation and the notation sued in the implementation (or should this be contained in the implementation chapter?)

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

similarity threshold (ST)

3. Problem Statement

CSS researches may typically be interested in extracting statements of a specific kind from a text corpus, such as expressions of sentiment of an actor towards some entity or expressions of conflicts between different actors. Two sensible ways to frame this task are as *text classification* (Kowsari et al. 2019) or *text retrieval* (Manning, Raghavan, and Schütze 2008). It can be addressed with a wide range of system, which will be generally referred to as *automatic text analysis* systems in the following. These systems are mostly based on techniques from the field of Natural Language Processing (NLP), as well well as the interlinked fields of Information Retrieval (IR) and Information Extraction (IE) (Chowdhary 2020).

A relevant perspective of categorising text analysis systems, especially from the point of view of CSS researchers, are the dimensions *open-opaque* and *adaptive-strict* (Menezes and Roth 2021). Here openness refers to the systems users ability to inspect and understand the processing, which we can also bee describes as transparency and explainability. These properties are of high interest in CSS applications such as predicting political opinion based on social media activity (Wilkerson and Casas 2017). An adaptive text analysis system does not (only) operate on strict rules, but is able to learn and modify its behaviour in some way. It is therefore in principle able to handle unforeseen variations in the text it processes. While both of these two properties are desirable for users are often found to be a trade-off. Current successful adaptive systems are most often bases on neural networks (Hirschberg and Manning 2015), which are opaque in the way how they represent and process text (Rudin 2019).

The Semantic Hypergraph (SH) framework aims to fulfil both the open as well as the adaptive property of a text analysis system. It offers an inspectable and understandable representation of text that is constructed by a parser based on machine learning components. The SH representation and its construction can be therefore considered to fulfil the open-adaptive properties. The SH pattern matching language can be used to define patterns that match a specific subset of hyperedges in a given hypergraph. The matching process is purely symbolic and follows a set of fixed rules. It can therefore be considered to be open-strict. In the context of the SH framework the CSS research task described above is better framed as text retrieval. The SH pattern acts as a *query* for which the most relevant items are retrieved. While the SH frameworks capabilities are not restricted to text retrieval, the work is focused on this application.

The SH pattern defined by a user 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). The SH pattern language allows it to describe different levels of generalisations for the structural matching. Additionally the actual words that should match need to be specified i.e. the edge content to match against, if the edge matches the pattern structurally. These words need to be given explicitly and the only way of generalising is via the lemma functional pattern. This lack of generalisation capability entails that a bias is introduced into the

matching process by the manual selection of words by the SH frameworks user, who defines the pattern.

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 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 in the SH matching process would allow users to define patterns, which describe generalisations of edge content. 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 openness 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.

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.

This has to be adapted based on chapter 2
-> add reference to FNESS/CNESS and derive relevancy of both for this work -> modify RQs
-> derive in more detail why embedding based SS is chosen
-> add discussing about effiency?

Integrating NESS measures into the pattern matching process would allow for edge content generalisation and therefore would make the process more adaptive. As illustrated earlier, NESS measures principally do not provide the openness that is inherent to the pattern matching process of the SH framework. In the sense of the open-opaque / strict-adaptive classification described above this integration would mean a shift from openness to opaqueness and from strictness to adaptivity. To counteract the opaqueness introduced by an NESS integration into the SH pattern matching, allowing user control over generalisation levels can maintain some openness while still benefiting from increased adaptivity.

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 generalisation and therefore maintaining some openness of the pattern matching process?

3.1.2. Secondary Questions

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

R.2 How can neural embedding based semantic similarity effectively (and efficiently?) be integrated into the Semantic Hypergraph pattern matching?

R.3 Does integration neural embedding-based semantic similarity improve the retrieval performance of the Semantic Hypergraph framework and how does it impact recall and precision when matching a pattern against a set of known desired matching results?

R.4 How can the level of edge content related generalisation in the pattern matching process be effectively and transparently 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 Embedding-based Semantic Similarity Matching, which is described in chapter 3 where the relevancy of this problem for has also been derived.

The system is described here will in the following be referred to as *Neural Embedding-based Semantic Similarity extended Semantic Hypergraph Pattern Matching* or:

NESS-SHPM aka ***NESSeSHyPaM***

The recall of NESS-SHMP in relation to ST $r(t_s$ is strictly monotonically decreasing, since the set of points in embedding space that is inside of the similarity boundary consistently gets smaller with increasing ST.

4.1. Integration into the Pattern Matching Process

4.1.1. SemSim Functional Pattern

`semsim`

pattern works only for atoms

4.1.2. Sub-pattern Similarity Thresholds

4.2. Fixed Word Embedding-based Matching

(FNESS)

word2vec via gensim

discussion about using transformer models for single word embeddings?

reference words: single-word and multi-word reference

Square bracket notation

4.3. Contextual Embedding-based Matching

Contextual Neural Embedding-based Semantic Similarity (CNESS)

all tokens option (AT) vs sub tokens

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.

reference edges

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

Semsim instances

reference edge sample modification parameter

5.3. Modifications of the SH Pattern Matching

skip semsim

5.4. Modifications to the Hypergraph database

is this really necessary? tok pos etc, but not actually specific to semsim

6. Evaluation

In this chapter the conceived concept (see chapter 4) and specific implementation (see chapter 5) of the NESS-SHPM system is being evaluated to answer the research question(s) posed in section 3.1. Therefore a a case study is conducted to evaluate the system for a specific use case.

refer to the RQs more specifically? how are they going to be answered?

6.1. Case Study: Conflicts

The conflicts case study follows the approach presented in Menezes and Roth 2021, where expressions of conflict are extracted from a given SH using a single SH pattern. In their work they build upon the information extracted by the pattern to conduct further analyses, which are not in the scope of this work. Here the evaluation is limited to the task of classifying whether the content of a given edge in the SH is an expression of conflict or not. Or framed differently, the task is to retrieve exactly all those edges whose content is an expression of conflict. The evaluation will compare the retrieval performance of a suitable set of different SH patterns and corresponding configuration of the NESS-SHMP system by matching them against a labelled dataset of hyperedges.

should I explain why specifically the conflicts and not some other case study (i.e. dataset) –> because there was none... but then I need to show why there was none and what are the criteria for a case study to be suitable to evaluate the system

6.1.1. Expressions of Conflict

An expression of conflict in the context of this case study is defined as a sentence which fulfils the following properties:

> There is a conflict between two explicitly named actors, wherever these actors are mentioned in the sentence; whereby a conflict is defined as antagonizing desired outcomes.

6.1.2. Reddit Worldnews Corpus

The corpus from which those expressions of conflict are retrieved consists 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*, which is described as: “A place

for major news from around the world, excluding US-internal news."[1] This corpus contains 479,384 news headers and is in the following referred to as the *Worldnews-Corpus*.

Each of these headers is comprised of a single sentence and is represented as a sequence root edge in the SH constructed from it. In the following this SH is referred to as the *Worldnews-SH*. Parsing errors that may potentially occur during this constructed and can obstruct a correct retrieval of a wrongly parsed edge i.e. wrongly represented sentence. These errors are out of scope of this work. All edges in the Worldnews-SH are assumed to be correctly parsed.

6.1.3. Semantic Hypergraph Patterns

The SH patterns that are used in this evaluation all have the same general form to isolate the effect of replacing a purely symbolic matching against a specific word or list of words with NESS-SHMP. In this section the general form of these pattern will be described, which entails consequences for the creation of the labelled dataset described in section 6.2.

Original Conflict Pattern

Pattern 3 is originally defined in Menezes and Roth 2021, p. 22 and is therefore referred to as the *original conflict pattern*. It is used to extract conflicts between two parties SOURCE and TARGET, potentially regarding some TOPIC. As mentioned before, the assignment of these variables is irrelevant for this case study.

The original conflict patterns contains two sub-patterns which utilize word lists. These sub-patterns match the trigger sub-edge and predicate sub-edge of a candidate edge respectively and are in following referred to as *trigger sub-pattern* and *predicate sub-pattern*. If not stated otherwise these terms will refer to pattern 3.

- **Trigger sub-pattern:** [against,for,of,over]/T
- **Predicate sub-pattern:** (PRED/P.so,x) ∧
  (lemma/J >PRED/P [accuse,arrest,clash,condemn,kill,slam,warn]/P)

In the trigger sub-pattern the content of the candidate trigger sub-edge is directly matched against a list of prepositions, which are in the following referred to as the *conflict prepositions*. In case of the predicate sub-pattern, the word list is matched against the lemma of the innermost atom of the candidate predicate sub-edge, which is always a verb. The list of verbs used here will in the following be referred to as the *conflict verbs*.

- **Conflict prepositions:** against, for, of, over
- **Conflict verbs:** accuse, arrest, clash, condemn, kill, slam, warn

(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.: Original conflict pattern

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

Wildcard Conflict Patterns

Replacing either the trigger sub-pattern, the predicate sub-pattern or both of them with a semsim function are the options for utilizing NESS-SHPM in a modified version of pattern 3 without modifying the general structure of the pattern. To evaluate which of these options are best suited to evaluate the retrieval performance of NESS-SHPM, three *wildcard conflict patterns* are constructed. In these patterns the predicate sub-pattern (pattern 4) or the trigger sub-pattern (pattern 5) are replaced by the wildcard operator.

(PRED/P.so,x SOURCE/C TARGET/C [against,for,of,over]/T TOPIC/[RS]) ∧
(PRED/P */P)

Pattern 4.: Predicate wildcard pattern

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

Pattern 5.: Trigger wildcard pattern

Preliminary Evaluation The three wildcard conflict patterns are matched against the Worldnews-SH and the number of matches is recorded. Comparing the number of matches of these patterns shows which of the sub-patterns is most influential for the retrieval performance of pattern 3. Table 6.1 shows the results of these preliminary evaluations as well as the number of matches that result from matching pattern 3 against the Worldnes-SH. It can be seen that the choice of conflict verbs is much more influential on the number of matches than the choice of conflict prepositions when compared to the number of matches resulting from the original conflict pattern. While replacing the predicate sub-pattern with a wildcard operator yields an increase with a factor of $12,45$, replacing the trigger sub-pattern with a wildcard operator only yields and increase with a factor of $1,07$.

Pattern name	Number of matches
Original conflict pattern	5766
Predicate wildcard pattern	71804
Trigger wildcard pattern	6154

Table 6.1.: Results of matching the wildcard patterns against the Worldnews-SH

SemSim Conflict Patterns

Based on the result of the preliminary evaluation in section 6.1.3, the predicate sub-pattern of pattern 3 is replaced by different forms of semsim functional patterns to construct different *semsim conflict patterns*. These patterns are then used to evaluate the effects of utilizing NESS-SHPM. The trigger sub-pattern is not modified to better isolate these effects in comparison to purely symbolic SHPM.

Pattern 6 describes the general form of a semsim conflict pattern. The `<SEMSIM-FUNCTION>` placeholder is replaced with one of the three implemented semsim functions to construct the *semsim-fix conflict pattern* (pattern 7), *semsim-fix-lemma conflict pattern* (pattern 8)

and the *semsim-ctx conflict pattern* (pattern 9). As <SEMSIM-ARGUMENT> the conflict verb list is used as similarity reference words in pattern 7 and pattern 8, which utilize FNESS. In the semsim-ctx conflict pattern, the wildcard operator is used as <SEMSIM-ARGUMENT> since the necessary reference edges can only be provided via an external parameter and not inside the pattern.

(PRED/P.so,x SOURCE/C TARGET/C [against,for,of,over]/T TOPIC/[RS]) ∧
(<SEMSIM-FUNCTION>/J PRED/P <SEMSIM-ARGUMENT>/P)

Pattern 6.: General SemSim conflict pattern

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

Pattern 7.: semsim-fix conflict pattern

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

Pattern 8.: semsim-fix-lemma conflict pattern

(PRED/P.so,x SOURCE/C TARGET/C [against,for,of,over]/T TOPIC/[RS]) ∧
(semsim-ctx/J PRED/P */P)

Pattern 9.: semsim-ctx conflict pattern

6.2. Conflict Dataset

To conduct an evaluation which assesses the retrieval performance of the NESS-SHPM system it is necessary to have a dataset of edges with labels that state whether an edge is an expression of conflict or not. Since such a dataset does not exists it needs to be constructed. In the following the construction process of this *conflict dataset* (CD), which is used for the evaluation in this case study, and the datasets characteristics are discussed.

6.2.1. Base Edge Set

The set of edges that can be retrieved by a conflict pattern, i.e. the original conflict pattern or a semsim conflict pattern is restricted the general form of these patterns. This entails that, given the same SH, every set of matching edges of a pattern of this form will be a subset of the matching edges of the predicate wildcard pattern (pattern 4). The set of edges resulting form matching this pattern against the Worldnews-SH are therefore used as the *base edge set* (BES) from which the conflict dataset is constructed, instead of the entirety of all the hypergraphs sequence root edges.

Predicate Lemma Every edge in the BES has a predicate sub-edge that has an innermost atom, which is a verb that has a lemma. In the following this is called the *predicate lemma* of an edge. Each of the edges matching pattern 3 or a pattern in the form of of pattern 6 therefore corresponds to a predicate lemma.

6.2.2. Desired Characteristics

To effectively evaluate the effectiveness of the application of NESS by matching a pattern in the form of pattern 6, the dataset used for this should have the following characteristics:

- Contain the largest possible number of unique predicate lemmas
- Contain the largest possible number of edges per unique predicate lemma

On the one hand it is desired to have as many different unique predicate lemmas as possible in the dataset to be able to evaluate whether NESS can differentiate if a predicate lemma indicates an expression of conflict or not. On the other hand it is desired to have as many different edges per unique lemma as possible in the dataset to be able to evaluate whether CNESS is able to differentiate if edges represent an expression of conflict or not, given that they correspond to the same predicate lemma.

6.2.3. Construction Process

To create the labelled CD, the edges of the dataset need to be manually labelled by human annotators, which is labor-intensive. The BES contains $n_b = 71804$ edges. Due to the time constrains of this work and the limited availability of three annotators, the BES needs to be subsampled to create the CD.

Filtering

Since the desired characteristics described above relate the the distribution of predicate lemmas, it is relevant to verify that is possible to determine the predicate lemma for all edges in the edge set from which the CD is sampled. In some cases it is not possible to determine the predicate lemma of a given edge due to to implementation issues, which out of scope of this work. In these cases an edge is filtered from the BES, which results in the *filtered base edge set* (FBES). The FBES contains $n_f = 69380$ edges.

Sampling

The edges in the FBES correspond to $n_l = 2195$ unique predicate lemmas. Attaining to the desired dataset characteristics, the number of samples n_s in the subsampled dataset should ideally be a multiple $m_l >= 2$ of n_l, so that $n_s = m_l \cdot n_l$. This would mean that every predicate lemma contained in the FBES is statistically represented multiple times in the subsampled dataset.

A dataset size of $n_s = 2000$ was chosen, wich means $m_l < 2$ and $n_s << n_f$. This entails that a trade-off between the desired dataset characteristics has to be made. To account for this, a sampling method is applied that offers more control over the distribution of

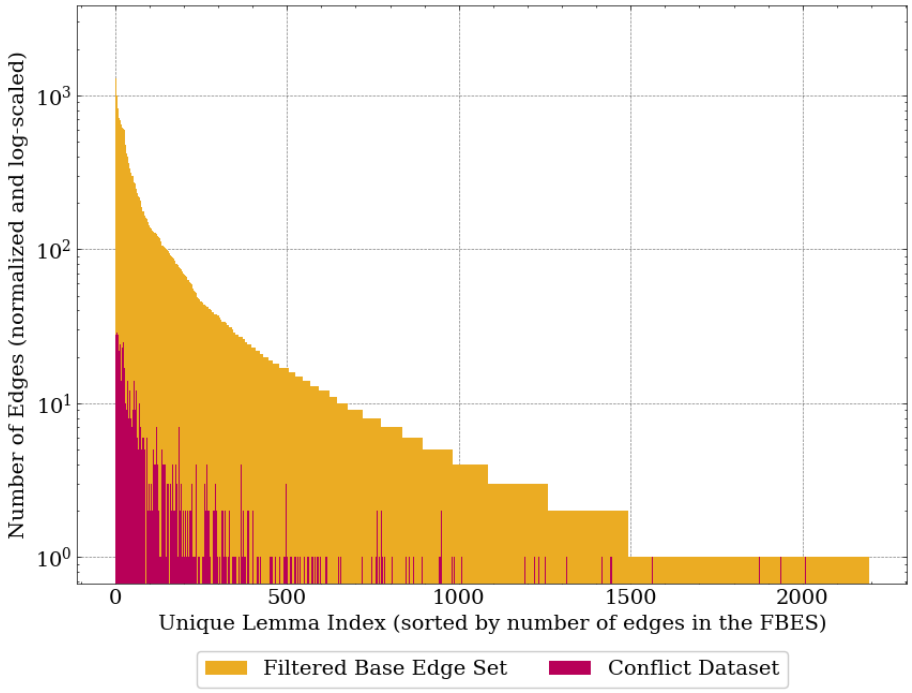


Figure 6.1.: Distribution of unique lemmas in the FBES and CD

predicate lemmas in the subsampled dataset than uniform random sampling does. This sampling method is based on the idea of *Stratified Sampling* (Parsons 2017) and is described in detail in algorithm 1.

is this correct?

The procedure splits the FBES into multiple bins after the edges are sorted by number of occurrence of their predicate lemma and then uniformly randomly samples from each bin. This method guarantees that predicate lemmas which correspond to a relatively small number of edges in the FBES will be represented in the subsampled dataset. The distribution of uniques lemmas in the FBES and the CD is compared visually in fig. 6.1.

Algorithm 1 Dataset sampling algorithm

1. Create a list of tuples t of edges and their corresponding predicate lemma:
   $L = [(l_k, e_i), ...]$ with $k \in \{0, ..., m\}$ and $i \in \{0, ..., n\}$
2. Sort this list by the number of tuples containing a predicate lemma to create the list:
   $L_{sort} = [(l_0, e_0), ...(l_m, e_n)]$, so that:
   - n_k is the number of tuples containing a lemma l_k
   - t_j with $j > i$ is a tuple wich sorted after tuple t_i
   - $n_o >= n_p$ if $t_i = (l_o, e_i)$ and $t_j = (l_p, e_j)$
3. Split the list L_{sort} into n_b bins.
4. Uniformly sample n_{sb} tuples from each bin.
5. Build a set of all edges e contained in the sampled tuples.

The subsampled dataset size resulting from this sampling method is $n_s = n_b * n_{sb}$. Given

Edge set name	Number of all edges	Number of un. lemmas	Number of conflict edges (% of all edges)	Number of no conflict edges (% of all edges)
Worldnews-SH	479384	-	-	-
Base Edge Set (BES)	71804	-	-	-
Filtered BES (FBES)	69380	2195	-	-
Conflict Dataset (CD)	2000	539	599 (29.95 %)	1401 (70.05 %)

Table 6.2.: Number of edges, number of unique lemmas and proportion of labels for the different edge sets

$n_s = 2000$, the values $n_b = 10$ and $n_{sb} = 200$ were chosen for sampling the CD.

Labelling

The labelling task is shared between the three annotators. A given edge will be either labeled as *conflict* or *no conflict* by an annotator following the definition given in section 6.1.1. Because of the aforementioned time constraints, every edge is only labeled by one annotator. To nonetheless ensure a consistent labelling among all annotators, a set of 50 edge is labelled by all three annotators. Every edge for which a disagreement in labelling occurs between at least two of the annotators, is inspected to reach an agreement on the label. Utilizing this process, the annotators understanding of what constitutes an expression of conflict is refined. Following this preliminary step, the n_s edges of the dataset are equally distributed among the three annotators and individually labelled by them.

6.2.4. Edge Set Comparison

The CD is the result of the filtering, sampling and labelling described above. The size of the Worldnews-SH, BES, FBES and CD are listed in table 6.2 for comparison. If applicable the the number of unique lemmas as well as the number and percentage of edges which are labelled as an expression of conflict and of those which are not are also noted.

add examples? (in appendix?)

6.3. Evaluation Process

In this evaluation multiple *evaluation runs* are conducted. Each evaluation run correspond to a SHPM process in which a pattern is matched against the CD. In the case of patterns utilizing NESS this requires that additional parameters in form of an *NESS configuration* are given to the matching process. An evaluation run is described by an *evaluation (run) configuration*. For each evaluation run the *evaluation metrics* are computed.

6.3.1. Evaluation Run Configurations

An evaluation configuration consists of the following parameters:

add references to chapter 4

- Conflict Pattern

- NESS Configuration (in case NESS-SHMP):
  - NESS model
  - Similarity Threshold
  - Use all tokens (in the case of CNESS)
  - Reference Edge Set (in the case of CNESS)

Conflict Patterns The four conflict patterns used in this evaluation are described in detail in section 6.1.3. An overview of the properties of these patterns can be seen in table 6.3.

Similarity Thresholds In this evaluation the similarity threshold t_s is always selected from a range of thresholds $r_t = \{0, 0.01, ..., 0.99, 1.00\}$, i.e. $t_s \in r_t$. This results in 101 different values of t_s.

Reference Edge Sets Multiple *reference edge sets* (RES) are randomly sampled from the set of edges in the CD, which are labelled as "conflict". These edges are then excluded from the dataset, to avoid introducing data from the test dataset to the system that is being evaluated. To compare the effect of different sample sizes, differently sized sets are drawn. To compare the effect of different samples, different samples are drawn. A RES with ID *N-X* has $N \in \{1, 3, 10\}$ samples and is from sample draw $X \in \{1, 2, 3, 4, 5\}$. This results in 15 different RES in total. The specific sets that have been sampled can be seen in appendix A.1.

Evaluation Run Names

An evaluation run name has the form: `CP NM-AT r-N-X t-TS`

In a specific evaluation run name, the placeholders (capitalised letters) are replaced with actual values. Such a name always begins with the conflict pattern (CP) name in its shortened form: *original*, *semsim-fix*, *semsim-fix-lemma* or *semsim-ctx*. In case of a NESS utilizing conflict pattern, the NESS model (NM) name is added in its shortened form: word2vec as *w2v* and conceptnet-numberbatch as *cn*. If the NESS type is CNESS, the usage of the all-tokens (AT) option is indicated by adding `-at` to the model name. If the option is not used, it is not added. Also in case of CNESS, the reference edge set ID N-X is indicated by appending `r-N-X`. For all NESS utilizing evaluation runs, the similarity threshold t_s is indicated by appending `t-TS`, where `TS` is the value of t_s.

Specification of Evaluation Run Configurations

The configurations for all evaluation runs that are conducted in this case study are specified in table 6.4. All possible parameter combinations of an evaluation configuration, i.e. the conflict pattern and the NESS parameters, are evaluated. The total number of conducted evaluation runs therefore amounts to 6465.[2] In table 6.4 the different values for the

[2] 1 (original) + 2 (semsim-fix and semxim-fix-lemma) * 2 (FNESS models) * 101 (STs) + 1 (semsim-ctx) * 2 (CNESS models) * 2 (all tokens) * 15 (ref. edge sets) * 101 (STs)

Pattern name	Lemma based	NESS type	Includes ref. words	Requires ref. edges
Original conflict pattern (3)	Yes	-	-	-
semsim-fix conflict pattern (7)	No	Fixed	Yes	No
semsim-fix-lemma conflict (8)	Yes	Fixed	Yes	No
semsim-ctx conflict patter (9)	No	Contextual	No	Yes

Table 6.3.: Properties of the conflict patterns used in the evaluation

Evaluation Run Name	Conflict Pattern	NESS Configuration	
		NESS Model	all tokens
original	original	-	-
semsim-fix w2v	semsim-fix	word2vec	-
semsim-fix cn	semsim-fix	conceptnet-numbatch	-
semsim-fix-lemma w2v	semsim-fix-lemma	word2vec	-
semsim-fix-lemma cn	semsim-fix-lemma	conceptnet-numbatch	-
semsim-ctx e5 r-N-X	semsim-ctx	e5	No
semsim-ctx gte r-N-X	semsim-ctx	gte	No
semsim-ctx e5-at r-N-X	semsim-ctx	e5	Yes
semsim-ctx gte-at r-N-X	semsim-ctx	gte	Yes

Table 6.4.: Evaluation Run Configurations

reference edge set ID and the ST are omitted. The *random* evaluation run configuration relates to a hypothetical evaluation run in which edges are uniformly randomly matched.

6.3.2. Evaluation Metrics

Using the information provided by the dataset labels it is determined whether a match is correct or not. If an edge is matches in a given evaluation run and is labeled as "conflict" in the dataset, it is considered a *true positive* (TP). If an edge matches but is labeled "no conflict", it is considered a *false positive* (FP). The *true negatives* (TN) and *false negatives* (FN) are determined analogously by examining the non-matching edges. Based on the TP, FP, TN and FN the metrics *precision*, *recall* and *F1-score* are computed.

show how this metrics are computed?

derive why these metrics were chosen:
accuracy is not interesting since the dataset is unbalanced, precision and recall both of interest, but are expected to be a trade-off (where recall should decline with rising ST). F1-score is an established metric that closely relates to precision and recall and represents this trade-off and therefore retrieval performance as a whole. MCC is arguably a better metric because it is symmetrical and incorporates true negatives. also the F1-score of the original pattern is worse than random, which indicates a metric mismatch. then again the close relation and equal value range of precision, recall and F1-score are a plus (for plotting especially).

6.4. Evaluation Results

In this section the results of the evaluation runs which are defined by the evaluation configurations in section 6.3.1 are examined. Different perspectives on the result data are constructed in the form of tables and plots to enable answering the research questions. The following subsections each represent one perspective and conclude with significant observations that can be made based on it.

Result Data Description Concepts

To facilitate constructing insightful perspectives on the result data, some novel concepts for its description are introduced in the following.

Evaluation Run Sets Multiple evaluation runs can be grouped into an *evaluation run set* (ERS) according to their shared configuration parameter values. The naming convention for an ERS follows the evaluation run naming convention described in section 6.3.1. The parameters values that are not shared among the evaluation runs in the ERS are omitted from the name or replaced by the wildcard symbol `*`. The placeholders (capitalised letters) are used to refer to an ESR of a generic form with fixed parameter values without specifying these values. By surrounding a part of the ERS name with parentheses `(*)`, it is indicated that this part is omitted if unsuitable.

Examples are given to illustrate this:

- An ERS of all evaluation runs utilizing NESS with $t_s = 0.5$ is named:
  `semsim-* t-0.5`
- An ERS of all evaluation runs utilizing FNESS with an unspecified but fixed NESS model has the form:
  `semsim-fix(-*) NM`
- An ERS of all evaluation runs utilizing NESS with all parameters (that are applicable) fixed but unspecified, except for the specific value $t_s = 0.5$, has the form:
  `semsim-* NM(-AT) (r-N-X) t-0.5`

Best F1-Score Evaluation run The *best F1-Score evaluation run* refers to the evaluation run with the highest F1-Score in an ERS corresponding to a NESS utilizing evaluation configuration where every parameter except for t_s is fixed. Such an ERS is generally named `semsim-* NM(-AT) (r-N-X)`. The corresponding F1-score is also simply referred to as *best F1-score*.

Mean Reference Edge Set Evaluation Runs A *mean reference edge set evaluation run* is constructed from the mean value of all evaluation scores for the evaluation runs in an ERS of the form `semsim-ctx NM-AT r-N-* t-TS`. This means for every t_s the mean of the corresponding evaluation scores of all reference edge sets of the same size is computed. In the following these synthetical evaluation runs are referred to in this form: `semsim-ctx NM-AT r-N-mean`

Mean Reference Edge Set Best F1-Score Evaluation Metric Scores The *mean reference edge set (RES) best F1-Score evaluation metric scores* are the mean values of all evaluation scores corresponding to the best F1 score for all evaluation runs in an ERS of the form `semsim-ctx NM-AT r-N-*`. In the following these evaluation metric scores will be referred to in this form: `semsim-ctx NM-AT r-N-mean-best`

6.4.1. Best F1-Score based Evaluation Run Comparison

Table 6.5 shows the evaluation scores for all evaluation metrics of the best F1-score evaluation runs for the original conflict pattern evaluation run and all evaluation runs utilizing FNESS. For the evaluation runs utilizing CNESS only the mean RES best F1-score evaluation metric scores and the standard deviation of the best F1-scores for the corresponding ERSs are shown. The t_s value listed for the mean RES best F1 score evaluation metrics is the mean of all t_s values for the best F1-scores in the corresponding ERSs.

In table A.5 and table A.6 of appendix A.2 the best F1-score evaluation run results can be seen for evaluation runs utilizing CNESS wit all tokens disabled and enabled respectively. These tables also list the hypothetical best F1-score evaluation for run the mean reference edge set evaluation runs. Additionally the mean standard deviation for these ERSs is shown, i.e. the mean of the standard deviations of the F1-score for every ERS of the form `semsim-ctx NM-AT r-N-* t-TS`.

Significant Observations

1.1 All evaluation runs utilizing NESS achieve a best F1-score that is higher than the F1-score of the random evaluation run and the original evaluation run

1.2 CNESS achieves achieves a higher F1-score than FNESS by 4.0%, when comparing the highest F1-scores achieved among all FNESS utilizing evaluation runs and the highest mean RES best F1 score achieved among all CNESS utilizing evaluation runs (`semsim-fix-lemma cn t-0.30` and `semxim-ctx e5 r-10-mean-best`

1.3 Lemma based FNESS achieves a higher F1-score than non-lemma FNESS by 3.2%, when comparing the highest F1-scores achieved by evaluation runs utilizing one of the two variants (`semsim-fix w2v t-0.27` and `semsim-fix-lemma cn t-0.30`)

1.4 For lemma based FNESS, the conceptnet-numberbatch model achieves a higher best F1-score than the word2vec model by 4.2%
(`semsim-fix-lemma cn t-0.30` vs `semsim-fix-lemma w2v t-0.33`)

1.5 For not lemma based FNESS, the word2vec model achieves a higher best F1-score than the conceptnet-numberbatch model by 1.4%
(`semsim-fix w2v t-0.27` vs `semsim-fix cn t-0.25`)

1.6 CNESS with the AT option disabled achieves a higher or equal mean RES best F1-score than CNESS with AT option enabled in 6/6 (100%) direct comparisons
(`semsim-ctx NM r-N-mean-best` vs `semsim-ctx NM-at r-N-mean-best`)

1.7 CNESS with the e5 model achieves a higher or equal mean RES best F1-score than CNESS with the gte model in 6/6 (100%) direct comparisons
(`semsim-ctx e5-* r-N-mean-best` vs `semsim-ctx gte-* r-N-mean-best`)

Evaluation Run Name				Prec.	Rec.	(Best) F1-Score	
CP	NM	RES	t_s				Std. Dev.
random			-	0.300	0.500	**0.375**	-
original			-	0.706	0.209	**0.322**	-
semsim-fix	cn		0.25	0.479	0.524	0.500	-
semsim-fix	w2v		0.27	0.483	0.533	**0.507**	-
semsim-fix-l.	cn		0.30	0.492	0.558	**0.523**	-
semsim-fix-l.	w2v		0.33	0.460	0.553	0.502	-
semsim-ctx	e5	r-1-mean-best	0.65	0.392	0.772	0.518	+/- 0.025
semsim-ctx	gte	r-1-mean-best	0.59	0.336	0.879	0.483	+/- 0.025
semsim-ctx	e5	r-3-mean-best	0.68	0.399	0.818	0.536	+/- 0.021
semsim-ctx	gte	r-3-mean-best	0.65	0.365	0.799	0.499	+/- 0.016
semsim-ctx	e5	r-10-mean-best	0.72	0.416	0.790	**0.544**	+/- 0.020
semsim-ctx	gte	r-10-mean-best	0.68	0.382	0.812	0.517	+/- **0.010**
semsim-ctx	e5-at	r-1-mean-best	0.69	0.369	0.841	0.509	+/- 0.016
semsim-ctx	gte-at	r-1-mean-best	0.66	0.335	0.882	0.483	+/- 0.021
semsim-ctx	e5-at	r-3-mean-best	0.72	0.378	0.821	0.516	+/- 0.011
semsim-ctx	gte-at	r-3-mean-best	0.70	0.336	0.876	0.485	+/- 0.017
semsim-ctx	e5-at	r-10-mean-best	0.74	0.382	0.843	**0.525**	+/- 0.012
semsim-ctx	gte-at	r-10-mean-best	0.72	0.338	0.900	0.491	+/- **0.008**

Table 6.5.: Evaluation scores corresponding to best F1-scores for all evaluation runs

1.8 CNESS with the AT option enabled has a lower standard deviation of best F1-score than CNESS with the AT option disabled in 5/6 (83%) direct comparisons (`semsim-ctx NM-at r-N-mean-best` vs `semsim-ctx NM r-N-mean-best`)

6.4.2. Evaluation Metric vs. Similarity Threshold

These plots visualise the resulting evaluation scores for the different evaluation metrics in relation to different values for the similarity threshold.

Figure 6.2a shows the F1-score vs. the ST for the best performing evaluation runs for each conflict pattern. That means for every conflict pattern, this shows the evaluation run(s) with the configuration that resulted in the highest best F1-score. For the `random` and `original` evaluation runs, there is obviously no configuration to choose from.

For the FNESS utilizing evaluation runs, the evaluation run with the highest F1-score in the ERSs of the form `semsim-fix(-lemma) NM` is selected. This means for the semsim-fix and semxim-fix-lemma conflict patterns the corresponding best F1-score evaluation runs of the best performing model are selected..

For the CNESS utilizing evaluation runs, the evaluation runs which correspond to highest mean RES best F1-score are selected. This means the ERS of the form `semsim-ctx NM-AT r-N-*` which resulted in the highest mean value of best F1-scores. This ERS consists of five evaluation runs, wich each have the form `semsim-ctx NM-AT r-N-X`. The F1-scores for these runs are plotted with a lighter curve. The additional synthetical `semsim-ctx NM-AT r-N-mean` score is plotted with a normally bold curve.

Figure 6.2b follows the same concept as fig. 6.2a, but instead of the F1-score this plot shows the scores of precision and recall vs. the ST. Also in the selection of evaluation runs, the semsim-fix (not lemma) conflict pattern based evaluation runs are excluded.

The plots in this section are selected because they are deemed to be most relevant for the following. A more comprehensive comparison of the different evaluation runs from this perspective can be found in appendix A.3.

Active Similarity Threshold Range To facilitate the description of observation for this perspective on the result data, the concept of the *active similarity threshold range* (ASTR) is introduced. For a given ESR of the form `semsim-* NM(-AT) (r-N-X)`, the ASTR describes the range of t_s for which the recall (r) that is achieved by these evaluation runs is $r \neq r_{max} = 1$ and $r \neq r_{min}$. Here r_{min} and r_{max} are the lowest and highest recall values, which correspond to $t_{s1} <= t_{s2}$, since the function $r(t_s)$ is monotonically decreasing.

Significant Observations

2.1 The ASTR of `semsim-fix-lemma cn` is larger ($0.0 < t_s < 1.0$) than the ASTR of `semsim-ctx e5` (ca. $0.625 < t_s < 0.875$)

2.2 Generally the ASTR of ERSs of the form `semsim-fix-* NM` is larger than the the ASTR of ERSs of the form `semsim-ctx` (confer also fig. A.1, fig. A.2 fig. A.5 and fig. A.6)

2.3 The ASTRs are nearly equal for ERSs of the form `semsim-fix-* NM` (confer fig. A.1 and fig. A.2)

2.4 The ASTR of ERSs of the form `semsim-ctx gte-AT` begin at a lower value than for ERSs of the form `semsim-ctx e5-AT` (confer fig. A.5 and fig. A.4)

2.5 The ASTR of ERSs of the form `semsim-ctx NM` begin at a lower value than for ERSs of the form `semsim-ctx NM-at` (confer fig. A.6 and fig. A.4)

2.6 The ASTRs end at nearly the same value for ERSs of the form `semsim-ctx NM-AT` (confer fig. A.5, fig. A.6 and fig. A.4)

2.7 The evaluation runs in the ERS `semsim-fix-lemma cn` achieve a higher F1 value than the evaluation run in the ERS `semsim-fix w2v` for nearly all values of t_s

2.8 The precision of the evaluation run which achieves the highest precision among those in the synthetic ESR `semsim-ctx e5 r-10-mean` ($p_{\max}$) and the precision of the original evaluation run (p_{og}) are nearly equal ($p_{\max} \approx p_{\text{og}}$)

2.9 The precision of evaluation run `semsim-fix-lemma cn` correlates with the ST until it reaches the value achieved by the `original` evaluation run, where it plateaus

2.10 The precision of the evaluation runs in ERS `semsim-ctx e5` correlate with the ST until precision (p) and recall (r) reach approximately the same value ($p \approx r \approx 0.5$), after which it fluctuates (the specific fluctuation varies with the specific RES)

2.11 The precision achieved by evaluation runs in ERSs of the form `semsim-ctx e5 r-10-* t-TS` has a higher variation for $t_s \geq 0.75$ than for $t_s < 0.75$

2.12 The precision of the evaluation run which achieves the highest precision among those in the synthetic ESR `semsim-ctx e5 r-10-mean` ($p_{\max}$) and the precision of the original evaluation run (p_{og}) are nearly equal ($p_{\max} \approx p_{\text{og}}$)

2.13 The evaluation metric scores of `semsim-fix-lemma cn t-1.00` are lower than those of the `original` evaluation run

should i quantify all these observations?

would a more detailed analysis of precision and recall make sense? maybe a precision-recall curve or an roc? maybe recall at best precision and precision at random recall?

6.4.3. Best F1-Score vs. Number of Reference Edges

Figure 6.3 visualises the relation of the number of reference edges and the best-F1 score. For this purpose the number of reference edges `N` is plotted versus the mean RES best F1-score for all ERSs of the form `semsim-ctx NM-AT r-N-*`. The standard deviation of the best F1 scores for these ERSs is visualised by the shaded areas around the curves of the mean RES best F1-scores.

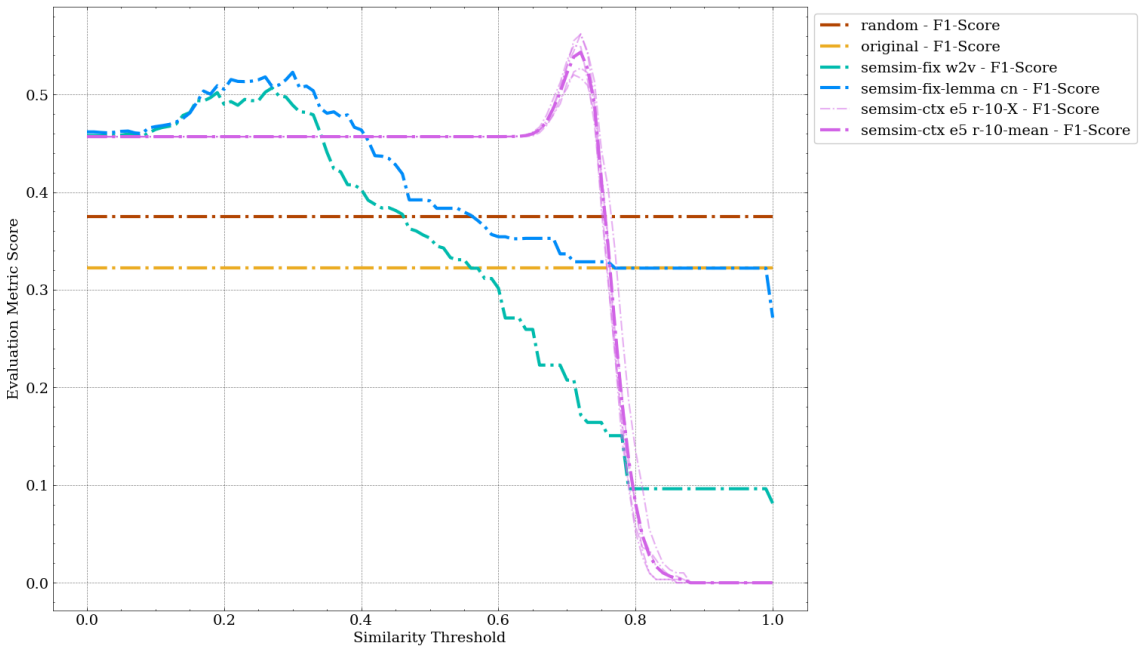
Significant Observations

3.1 The mean RES best F1-score of the evaluation runs in an ERS of the form `semsim-ctx NM(-AT) r-`N_1`-*` is higher than the mean RES best F1-score of the evaluation runs in an ERS of the form `semsim-ctx NM(-AT) r-`N_2`-*`, if $N_2 > N_1$ for $N_1, N_2 \in \{1, 3, 10\}$

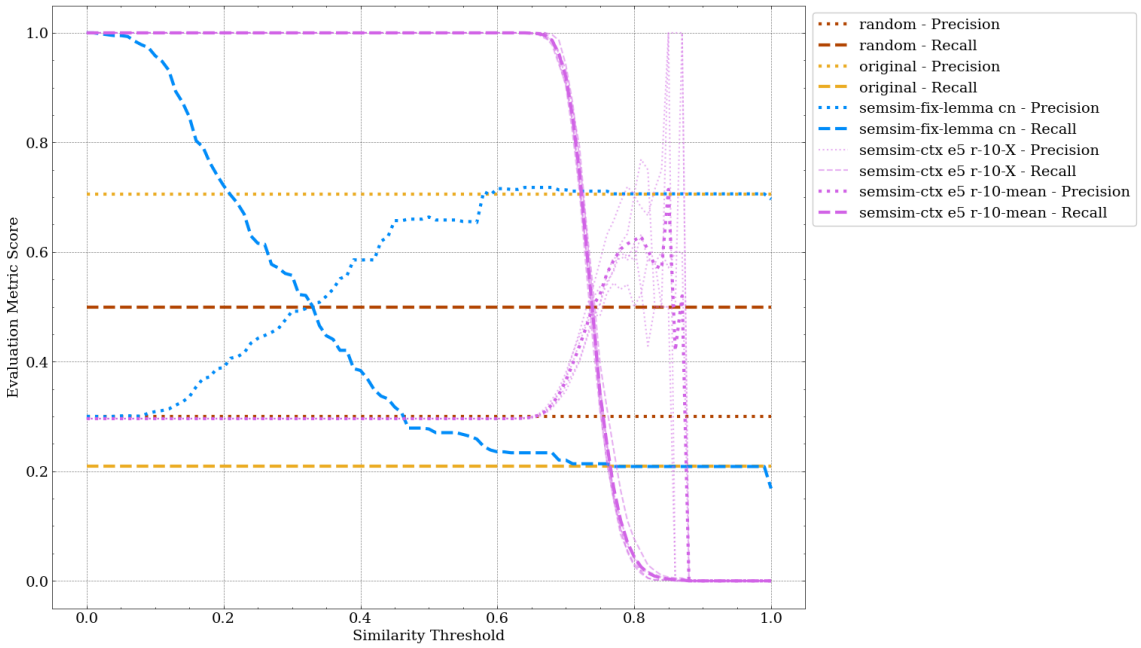
3.2 The standard deviation of the mean RES best F1-score of the evaluation runs in an ERS of the form `semsim-ctx NM(-AT) r-`N_1`-*` is lower than the standard deviation of the mean RES best F1-score of the evaluation runs in an ERS of the form `semsim-ctx NM(-AT) r-`N_2`-*`, if $N_2 > N_1$ for $N_1, N_2 \in \{1, 3, 10\}$, except for `semsim-ctx et-at r-3-*` ($sd_1 = 0.11$) and `semsim-ctx et-at r-10-*` ($sd_2 = 0.12$)

6.4.4. Predicate Lemma based Evaluation Run Comparison

In this section it is explored how the different NESS systems differ in which edges they match. This is done by following up on the concept of the predicate lemma introduced in section 6.2.1. In section 6.2.2 one of the two desired characteristics of the dataset states that it should contain the largest possible number of edges per unique predicate lemma. Specifically it is of interest, how the CNESS system performs in comparison to the FNESS system for subsets of edges which share the same predicate lemma. Such a subset of edges of the conflict dataset is in the following referred to as *predicate lemma edge set* (PLES).



(a) F1-score vs. ST for the evaluation runs `random`, `original`, `semsim-fix w2v`, `semsim-fix-lemma cn` and `semsim-ctx e5 r-10-X`



(b) Precision and recall vs. ST for the evaluation runs `random`, `original`, `semsim-fix-lemma cn` and `semsim-ctx e5 r-10-X`

Figure 6.2.: Evaluation metric scores vs. similarity threshold values

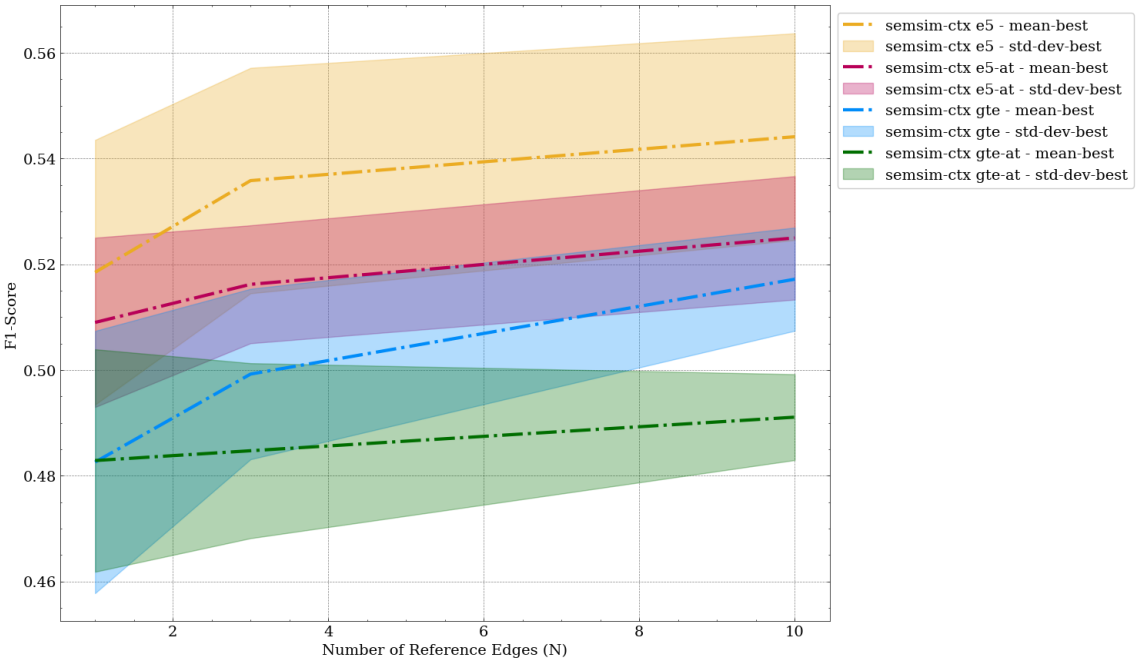


Figure 6.3.: Mean reference edge set best F1-score for the ERSs `semsim-ctx e5 r-N-*`, `semsim-ctx e5-at r-N-*`, `semsim-ctx gte-at r-N-*` and `semsim-ctx gte-at r-N-*`

NESS Type Representatives Two evaluation runs are selected to represent the two versions of the NESS systems for the comparison. The lemma-based version of FNESS is chosen here, because of its superior performance regarding best F1-score that is observed in section 6.4.1 and section 6.4.2. Specifically, the evaluation runs `semsim-fix-lemma cn t-0.30` (evaluation run A) and `semsim-ctx e5 r-10-2 t-0.72` (evaluation run B) are selected as representatives of the FNESS and CNESS system respectively. These are the best performing evaluation runs regarding F1-score for the respective semsim conflict patterns (i.e. NESS system), which can be seen in table 6.5 and table A.5.[3]

PLES Evaluation Score based Evaluation Run Comparison

Label Balance Ratio The *label balance ratio* (LBR) measures how balanced the labels in a given set of labelled edges are. It is calculated by eq. (6.1). Here n_{pos} and n_{neg} are the number of positively ("conflict") and negatively ("no conflict") labeled edges in the edge set. An edge set with fully balanced labels has $LBR = 1$ and completely unbalanced labeled edge set has $LBR = 0$.

$$LBR = 1 - \left(\frac{|n_{\text{pos}} - n_{\text{neg}}|}{n_{\text{pos}} + n_{\text{neg}}} \right) \tag{6.1}$$

[3]The latter table specifically shows that the evaluation runs `semsim-ctx e5 r-1-2 t-0.67`, `semsim-ctx e5 r-10-2 t-0.72` and `semsim-ctx e5 r-10-4 t-0.72` all correspond to an F1 score $s_{\text{F1}} = 0.56$. The evaluation runs utilizing the RESs of size $N = 10$ are chosen over the one utilizing an RES of size $N = 1$, because of their generally superior performance regarding F1 score, which is observed in section 6.4.3. Among the two remaining evaluation runs, `semsim-ctx e5 r-10-2 t-0.72` is selected randomly.

The evaluation metrics are computed for evaluation run A and B for each PLES, along with metrics that measure distributional properties of a PLES. Namely the number of edges n_e, the number of positively labeled and negatively labeled edges (n_{pos} and n_{neg}), the LBR and the entropy of a PLES.

Table 6.6 lists the ten predicate lemmas for whose PLES the absolute difference in F1-score, which is achieved in the two evaluation runs, is the highest. Conversely table 6.8 lists the ten predicate lemma for whose PLES the difference in F1-score which is achieved in the tow evaluation runs is the lowest. In both tables the predicate lemmas have been filtered beforehand, so that only PLES with $n_s >= 5$ samples are considered. Additionally the recalls (r_A, r_B) achieved by both evaluation runs regarding a PLES must fulfil the condition that $r_A + r_B > 0$, i.e. at least the recall achieved by one of the evaluation runs must be non-zero. Table 6.7 follows the same concept as table 6.6, except for the condition regarding the recalls being $r_A * r_B > 0$, i.e. both recalls achieved by the tow evaluation runs must be non-zero. This second variant of the table is not shown for table 6.8, because it only lists lemmas for whose PLES the F1-score achieved by both evaluation runs is zero.

should I add a table with the actual labels produced by the evaluation runs?

Significant Observations

4.1 The CNESS utilizing evaluation run achieves a higher F1-score than the FNESS utilizing evaluation run for every PLES of the the top ten PLES regarding highest difference in F1-score between the two evaluation runs (independently of the recall condition)

4.2 The mean LBR and mean entropy of the top ten PLESs regarding the highest difference in F1-score between the two evaluation runs are higher than the mean LBR and mean entropy of the top ten PLESs regarding the lowest difference in F1-score between the two evaluation runs (independently of the recall condition)

4.3 The differences in F1-score achieved by the two evaluation runs is higher for the recall condition $r_A + r_B > 0$ than for $r_A \cdot r_B > 0$, because for the first condition the F1-score of the FNESS evaluation run is zero for all PLESs

4.4 The lemmas corresponding to the the top ten PLES regarding highest difference in F1-score between the two evaluation runs are all not included in the conflict verbs: "accuse", "arrest", "clash", "condemn", "kill", "slam"

4.5 Of the lemmas corresponding to the top ten PLES regarding lowest difference in F1-score between the two evaluation runs, five of six are included in the conflict verbs: "arrest", "condemn", "kill", "slam" ("clash" is not included)

6.5. Result Discussion

In this section the previously presented evaluation results are discussed. It synthesizes the major insights derived from the observations of the different result data perspectives. The discussion is organized into categories and sub-categories, which relate to the research questions outlined in section 3.1. Here retrieval performance generally refers to joined measure of precision and recall and therefore means F1-score, as stated above in section 6.3.2.

Lemma	F1 Diff.	F1 A	F1 B	n_e	$n_{\text{pos}}/n_{\text{neg}}$	LBR	Entropy
file	1.00	0.00	1.00	5	5/0	0.00	0.00
order	0.88	0.00	0.88	14	7/7	1.00	1.00
launch	0.86	0.00	0.86	14	8/6	0.86	0.99
step	0.80	0.00	0.80	5	2/3	0.80	0.97
target	0.77	0.00	0.77	8	6/2	0.50	0.81
use	0.75	0.00	0.75	14	5/9	0.71	0.94
block	0.73	0.00	0.73	8	4/4	1.00	1.00
take	0.71	0.00	0.71	19	6/13	0.63	0.90
open	0.67	0.00	0.67	8	1/7	0.25	0.54
suspend	0.67	0.00	0.67	6	2/4	0.67	0.92
build	0.67	0.00	0.67	6	1/5	0.33	0.65
						mean	
						0.61	0.79

Table 6.6.: Top ten lemmas regarding the highest difference in F1-score between the evaluation runs with recalls $r_A + r_B > 0$ and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn t-0.30` (A) and `semsim-ctx e5 r-10-2 t-0.72` (B)

Lemma	F1 Diff.	F1 A	F1 B	n_e	$n_{\text{pos}}/n_{\text{neg}}$	LBR	Entropy
accept	0.42	0.25	0.67	7	1/6	0.29	0.59
strike	0.30	0.80	0.50	9	6/3	0.67	0.92
capture	0.22	0.44	0.67	7	2/5	0.57	0.86
seize	0.17	0.67	0.50	12	6/6	1.00	1.00
deny	0.17	0.50	0.67	6	2/4	0.67	0.92
suggest	0.17	0.33	0.50	5	1/4	0.40	0.72
warn	0.12	0.93	0.81	24	21/3	0.25	0.54
claim	0.12	0.43	0.55	22	6/16	0.55	0.85
attack	0.09	0.91	1.00	12	10/2	0.33	0.65
threaten	0.09	0.52	0.61	20	7/13	0.70	0.93
approve	0.08	0.12	0.20	16	1/15	0.12	0.34
						mean	
						0.50	0.76

Table 6.7.: Top ten lemmas regarding the highest difference in F1-score between the evaluation runs with recalls $r_A \cdot r_B > 0$ and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn t-0.30` (A) and `semsim-ctx e5 r-10-2 t-0.72` (B)

Lemma	F1 Diff.	F1 A	F1 B	n_e	$n_{\text{pos}}/n_{\text{neg}}$	LBR	Entropy
arrest	0.00	1.00	1.00	11	11/0	0.00	0.00
slam	0.00	0.92	0.92	7	6/1	0.29	0.59
criticize	0.00	0.91	0.91	6	5/1	0.33	0.65
shoot	0.00	0.89	0.89	5	4/1	0.40	0.72
condemn	0.00	0.84	0.84	25	18/7	0.56	0.86
dismiss	0.00	0.80	0.80	6	4/2	0.67	0.92
tell	0.01	0.49	0.50	28	9/19	0.64	0.91
accuse	0.02	0.97	0.95	33	31/2	0.12	0.33
kill	0.02	0.66	0.64	77	38/39	0.99	1.00
say	0.03	0.45	0.48	31	9/22	0.58	0.87
						mean	
						0.46	0.68

Table 6.8.: Top ten lemmas regarding the lowest absolute difference in F1-score between the evaluation runs with recalls $r_A + r_B > 0$ and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn t-0.30` (A) and `semsim-ctx e5 r-10-2 t-0.72` (B)

6.5.1. Retrieval Performance Improvement

- NESS-SHMP can achieve a better retrieval performance than the original conflict pattern, independent of NESS type and configuration (depending on the ST)
  Supporting Observations: Item 1.1
- CNESS-SHPM using the sub tokens embedding can achieve the overall best retrieval performance (in comparison to FNESS-SHMP and the original conflict pattern)
  Supporting Observations: Item 1.1, Item 1.2
- Using lemma-based FNESS-SHMP instead of word-based FNESS-SHMP can achieve a better retrieval performance
  Supporting Observations: Item 1.3, Item 2.7

Similarity Threshold Impact

- The relation of ST and NESS-SHMP retrieval performance and therefore the relevant ST range depends primarily on the NESS type
  Supporting Observations: Item 2.1 Item 2.2 Item 2.3
- The relation of ST and CNESS-SHMP retrieval performance depends secondarily on the NESS model and the usage of the all tokens option
  Supporting Observations: Item 2.4 Item 2.5 Item 2.6

NESS Configuration Impact

- Using a generally better performing NESS model (regarding established benchmarks) does not generally improve the NESS-SHMP retrieval performance
  Supporting Observations: Item 1.4, Item 1.5, Item 1.7

- Using the sub tokens embedding instead of the all tokens embedding improves retrieval performance, but using the all tokens embedding makes it less sensible to the selection of the reference edges
  Supporting Observations: Item 1.6, Item 1.8
- CNESS-SHPM retrieval performance improves with a higher number of reference edges and is less sensible to the specific selection of reference edges
  Supporting Observations: Item 3.1, Item 3.2

6.5.2. Retrieval Precision Behaviour

- The precision of NESS-SHMPM correlates with the ST until a specific value of the ST, which itself is specific to the NESS type, NESS model (and other CNESS parameters, especially the selection of reference edges)
  Supporting Observations: Item 2.9, Item 2.10, Item 2.13
- CNESS-SHMP achieves on average the same precision as the original conflict pattern and lemma-based FNESS, although CNESS-SHMP can achieve a higher precision, it depends on the selection of the reference edges
  Supporting Observations: Item 2.11, Item 2.12

6.5.3. Contextual Differentiation Ability

- CNESS-SHPM is able differentiate when matching a set of edges where purely symbolic SHPM and FNESS-SHMPM cannot, i.e. cases where context is needed to determine the correct semantics of word
  Supporting Observations: Item 4.1, Item 4.2, Item 4.4, Item 4.5
- While CNESS-SHPM achieves a highest difference in retrieval performance for sets of edges, where FNESS-SHMP does not match, it also achieves a better retrieval performance in cases where FNESS-SHPM does match
  Supporting Observations: Item 4.1 Item 4.3

Add or integrate more direct answer to the research questions

7. Conclusion

8. Future Work

8.1. Conceptual Improvements

Somehow extend the token span used for CNESS beyond the word tokens but not to all tokens. Use the tokens of the next best sub-edge e.g. although in the case of a predicate this is probably the entire sentence most of the time.

8.2. 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

8.3. 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).

Chowdhary, K. R. (2020). "Natural Language Processing". In: *Fundamentals of Artificial Intelligence*. Ed. by K.R. Chowdhary. New Delhi: Springer India, pp. 603–649. ISBN: 978-81-322-3972-7. DOI: 10.1007/978-81-322-3972-7_19. URL: https://doi.org/10.1007/978-81-322-3972-7_19 (visited on 03/05/2024).

Evans, James A. and Pedro Aceves (July 1, 2016). *Machine Translation: Mining Text for Social Theory*. DOI: 10.1146/annurev-soc-081715-074206. URL: https://papers.ssrn.com/abstract=2822747 (visited on 06/15/2023). preprint.

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).

Hirschberg, Julia and Christopher D. Manning (July 17, 2015). "Advances in Natural Language Processing". In: *Science* 349.6245, pp. 261–266. DOI: 10.1126/science.aaa8685. URL: https://www.science.org/doi/abs/10.1126/science.aaa8685 (visited on 06/15/2023).

Kowsari, Kamran et al. (Apr. 2019). "Text Classification Algorithms: A Survey". In: *Information* 10.4 (4), p. 150. ISSN: 2078-2489. DOI: 10.3390/info10040150. URL: https://www.mdpi.com/2078-2489/10/4/150 (visited on 03/06/2024).

Lazer, David et al. (Feb. 6, 2009). "Computational Social Science". In: *Science* 323.5915, pp. 721–723. DOI: 10.1126/science.1167742. URL: https://www.science.org/doi/full/10.1126/science.1167742 (visited on 06/15/2023).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (July 7, 2008). *Introduction to Information Retrieval*. Higher Education from Cambridge University Press. DOI: 10.1017/CBO9780511809071. URL: https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C (visited on 03/06/2024).

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.

Min, Bonan et al. (Sept. 14, 2023). "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey". In: *ACM Computing Surveys* 56.2, 30:1–30:40. ISSN: 0360-0300. DOI: 10.1145/3605943. URL: https://dl.acm.org/doi/10.1145/3605943 (visited on 03/05/2024).

Parsons, Van L. (2017). "Stratified Sampling". In: *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, pp. 1–11. ISBN: 978-1-118-44511-2. DOI: 10.1002/9781118445112.stat05999.pub2. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05999.pub2 (visited on 01/11/2024).

Qiu, XiPeng et al. (Oct. 1, 2020). "Pre-Trained Models for Natural Language Processing: A Survey". In: *Science China Technological Sciences* 63.10, pp. 1872–1897. ISSN: 1869-

1900. DOI: 10.1007/s11431-020-1647-3. URL: https://doi.org/10.1007/s11431-020-1647-3 (visited on 06/15/2023).

Rudin, Cynthia (May 2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5 (5), pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. URL: https://www.nature.com/articles/s42256-019-0048-x (visited on 06/17/2023).

Wilkerson, John and Andreu Casas (May 1, 2017). *Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges*. DOI: 10.1146/annurev-polisci-052615-025542. URL: https://papers.ssrn.com/abstract=2968080 (visited on 06/15/2023). preprint.

Young, Tom et al. (Nov. 24, 2018). *Recent Trends in Deep Learning Based Natural Language Processing*. DOI: 10.48550/arXiv.1708.02709. arXiv: 1708.02709 [cs]. URL: http://arxiv.org/abs/1708.02709 (visited on 06/15/2023). preprint.

A. Appendix

A.1. Reference Edge Sets

Num. of Ref. Edges	Ref. Edges Set ID	Reference Edge Content
1	1-1	Israeli gunfire wounds Gaza fisherman: ministry
1	1-2	Ukraine's Opposition Accuses Government of Provoking Violence
1	1-3	Thursday's attack by armed youths on the base in Bor left at least 58 dead, including children
1	1-4	Turkey suspends 15,200 education staff
1	1-5	Kurdish protesters storm the Conservative Party's campaign headquarters in London
3	3-1	Israeli gunfire wounds Gaza fisherman: ministry Kuwait Rejects Saudi Request for War Subvention Researchers Accuse Canadian Internet Company of Helping Yemen Censor the Web
3	3-2	Ukraine's Opposition Accuses Government of Provoking Violence Chinese island-building in the South China Sea is causing "irreversible and widespread damage to biodiversity and ecological balance," according to the Philippines; Manila accused China of disregarding the people who rely on the sea by destroying coral reefs to create new islands The government's opposition and various refugee organizations have harshly criticized the reforms
3	3-3	Thursday's attack by armed youths on the base in Bor left at least 58 dead, including children Venezuela expels 3 US consular officials Russia, China nix U.S. human rights claims: Russia and China disputed U.S. Ambassador Nikki Haley's contention that human rights violations are a main driver of conflicts
3	3-4	Turkey suspends 15,200 education staff Leading Muslim groups condemn ISIS killing of US journalists PayPal freezes Canadian media company's account over story about Syrian family
3	3-5	Kurdish protesters storm the Conservative Party's campaign headquarters in London Obama seeks new Syria strategy review to deal with ISIS Tougher Canadian visa policy hits foreign workers, protects Canadian jobs

Table A.1.: Edge content for the reference edge sets of size $N \in \{1, 3\}$

Num. of Ref. Edges	Ref. Edges Set ID	Reference Edge Content
10	10-1	Israeli gunfire wounds Gaza fisherman: ministry Kuwait Rejects Saudi Request for War Subvention Researchers Accuse Canadian Internet Company of Helping Yemen Censor the Web Afghanistan President Ashraf Ghani slams Pakistan for harbouring terrorists, praises India 5-year-old Kentucky boy fatally shoots 2-year-old sister Bangladesh police kill 'mastermind' of Dhaka cafe attack Philippines President Duterte orders army to destroy Islamic militants or risk ISIS disease Turkish jets kill 18 Daesh terrorists in northern Syria Report slams Israel's military law enforcement system Iran Pursuing Release of Sailors Abducted by Somalian Pirates
10	10-2	Ukraine's Opposition Accuses Government of Provoking Violence Chinese island-building in the South China Sea is causing "irreversible and widespread damage to biodiversity and ecological balance," according to the Philippines; Manila accused China of disregarding the people who rely on the sea by destroying coral reefs to create new islands The government's opposition and various refugee organizations have harshly criticized the reforms Syria and Russia oppose unilateral US strikes against ISIL in Syria Taliban attack in Afghanistan kills six policemen Turkish President condemns US commandos photographed sporting Kurdish militia insignia France could ease ban on gay men giving blood after ECJ ruling Libyan smuggler fighting kills 22 migrants Kazakhstan jails online editor for 'spreading false information Russia urges Assad to give up chemical weapons

Table A.2.: Edge content for the reference edge sets of size $N = 10$ (Part 1/3)

Num. of Ref. Edges	Ref. Edges Set ID	Reference Edge Content
10	10-3	Thursday's attack by armed youths on the base in Bor left at least 58 dead, including children Venezuela expels 3 US consular officials Russia, China nix U.S. human rights claims: Russia and China disputed U.S. Ambassador Nikki Haley's contention that human rights violations are a main driver of conflicts Turkish PM tells female reporter to 'know your place South Korean prosecutors seek arrest of ex-President Park in corruption probe Taliban Announce Spring Offensive in Afghanistan Spain dismantles 'jihadist cell U.S. conducts 'counter terrorism strike' against al Qaeda-linked target in Libya Swiss prosecutors launch money-laundering probe against fugitive Ukrainian President Yanukovych, and son UK Wants 10 Year Prison Sentence For Online Pirates
10	10-4	Turkey suspends 15,200 education staff Leading Muslim groups condemn ISIS killing of US journalists PayPal freezes Canadian media company's account over story about Syrian family U.S. urges China's Xi to extend non-militarization pledge to all of South China Sea Sri Lanka accuses Canada of holding Commonwealth 'to ransom Russia and pro-Moscow rebels on Wednesday condemned Ukraine for ratifying two bills on greater autonomy for the separatist east, saying they violated a peace deal and threatened a shaky month-long truce Malaysia turns away 800 boat people; Thailand spots 3rd boat US expresses concern over security of Pakistan's Nuclear weapons Health experts accuse WHO of 'egregious failure' on Ebola Sunni militants accuse the army, perhaps the only widely respected public institution in Lebanon, of siding with Hezbollah

Table A.3.: Edge content for the reference edge sets of size $N = 10$ (Part 2/3)

Num. of Ref. Edges	Ref. Edges Set ID	Reference Edge Content
10	10-5	Kurdish protesters storm the Conservative Party's campaign headquarters in London Obama seeks new Syria strategy review to deal with ISIS Tougher Canadian visa policy hits foreign workers, protects Canadian jobs U.S. preparing new sanctions against Chinese entities over financial support to North Korea Turkey's Erdogan makes Nazi jibe over Germany rally ban: "Your practices are not different from the Nazi practices of the past Brazil committee recommends Dilma Rousseff's impeachment U.S. dismisses Russian concern about missile defense system in South Korea South Korea mulls ban on bosses messaging employees at home Israel to destroy homes of Palestinian Jerusalem car attackers Majority of Finns reject NATO membership

Table A.4.: Edge content for the reference edge sets of size $N = 10$ (Part 3/3)

A.2. Best F1-score based Eval. Run Comparison Tables

Evaluation Run Name				Prec.	Rec.	(Best) F1-Score	
CP	NM	RES	t_s				Std. Dev.
semsim-ctx	e5	r-1-1	0.65	0.386	0.768	0.514	-
semsim-ctx	e5	r-1-2	0.67	0.442	0.769	0.562	-
semsim-ctx	e5	r-1-3	0.59	0.372	0.838	0.515	-
semsim-ctx	e5	r-1-4	0.67	0.363	0.804	0.500	-
semsim-ctx	e5	r-1-5	0.68	0.397	0.681	0.502	-
semsim-ctx	e5	r-1-mean	0.64	0.356	0.823	0.477	+/- 0.046
semsim-ctx	e5	r-1-mean-best	0.65	0.392	0.772	0.518	+/- 0.025
semsim-ctx	e5	r-3-1	0.68	0.398	0.836	0.539	-
semsim-ctx	e5	r-3-2	0.68	0.435	0.752	0.551	-
semsim-ctx	e5	r-3-3	0.69	0.410	0.846	0.553	-
semsim-ctx	e5	r-3-4	0.69	0.392	0.846	0.536	-
semsim-ctx	e5	r-3-5	0.68	0.361	0.814	0.500	-
semsim-ctx	e5	r-3-mean	0.69	0.417	0.753	0.534	+/- 0.024
semsim-ctx	e5	r-3-mean-best	0.68	0.399	0.818	0.536	+/- 0.021
semsim-ctx	e5	r-10-1	0.71	0.415	0.817	0.550	-
semsim-ctx	e5	r-10-2	0.72	0.435	0.793	0.562	-
semsim-ctx	e5	r-10-3	0.72	0.400	0.771	0.527	-
semsim-ctx	e5	r-10-4	0.72	0.454	0.735	0.562	-
semsim-ctx	e5	r-10-5	0.71	0.378	0.835	0.520	-
semsim-ctx	e5	r-10-mean	0.72	0.428	0.746	0.543	+/- 0.021
semsim-ctx	e5	r-10-mean-best	0.72	0.416	0.790	0.544	+/- 0.020
semsim-ctx	gte	r-1-1	0.60	0.349	0.794	0.485	-
semsim-ctx	gte	r-1-2	0.63	0.388	0.799	0.522	-
semsim-ctx	gte	r-1-3	0.57	0.303	0.972	0.462	-
semsim-ctx	gte	r-1-4	0.55	0.300	1.000	0.461	-
semsim-ctx	gte	r-1-5	0.61	0.339	0.829	0.482	-
semsim-ctx	gte	r-1-mean	0.60	0.327	0.870	0.474	+/- 0.013
semsim-ctx	gte	r-1-mean-best	0.59	0.336	0.879	0.483	+/- 0.025
semsim-ctx	gte	r-3-1	0.64	0.364	0.826	0.505	-
semsim-ctx	gte	r-3-2	0.67	0.378	0.713	0.494	-
semsim-ctx	gte	r-3-3	0.67	0.385	0.740	0.506	-
semsim-ctx	gte	r-3-4	0.66	0.376	0.820	0.516	-
semsim-ctx	gte	r-3-5	0.62	0.322	0.896	0.474	-
semsim-ctx	gte	r-3-mean	0.66	0.374	0.714	0.486	+/- 0.040
semsim-ctx	gte	r-3-mean-best	0.65	0.365	0.799	0.499	+/- 0.016
semsim-ctx	gte	r-10-1	0.67	0.377	0.869	0.526	-
semsim-ctx	gte	r-10-2	0.67	0.361	0.917	0.518	-
semsim-ctx	gte	r-10-3	0.69	0.382	0.834	0.524	-
semsim-ctx	gte	r-10-4	0.69	0.399	0.730	0.516	-
semsim-ctx	gte	r-10-5	0.68	0.388	0.708	0.501	-
semsim-ctx	gte	r-10-mean	0.68	0.378	0.803	0.513	+/- 0.009
semsim-ctx	gte	r-10-mean-best	0.68	0.382	0.812	0.517	+/- 0.010

Table A.5.: Best F1-score evaluation runs for all ERS of the form `semsim-ctx NM r-N-X` (CNESS with AT option disabled)

Evaluation Run Name				Prec.	Rec.	(Best) F1-Score	
CP	NM	RES	t_s				Std. Dev.
semsim-ctx	e5-at	r-1-1	0.68	0.338	0.915	0.494	-
semsim-ctx	e5-at	r-1-2	0.71	0.434	0.701	0.536	-
semsim-ctx	e5-at	r-1-3	0.66	0.356	0.885	0.508	-
semsim-ctx	e5-at	r-1-4	0.71	0.365	0.798	0.501	-
semsim-ctx	e5-at	r-1-5	0.71	0.351	0.908	0.507	-
semsim-ctx	e5-at	r-1-mean	0.69	0.350	0.847	0.487	+/- 0.021
semsim-ctx	e5-at	r-1-mean-best	0.69	0.369	0.841	0.509	+/- 0.016
semsim-ctx	e5-at	r-3-1	0.72	0.363	0.846	0.508	-
semsim-ctx	e5-at	r-3-2	0.72	0.389	0.765	0.516	-
semsim-ctx	e5-at	r-3-3	0.73	0.399	0.780	0.528	-
semsim-ctx	e5-at	r-3-4	0.73	0.386	0.829	0.526	-
semsim-ctx	e5-at	r-3-5	0.72	0.351	0.886	0.503	-
semsim-ctx	e5-at	r-3-mean	0.73	0.392	0.740	0.511	+/- 0.016
semsim-ctx	e5-at	r-3-mean-best	0.72	0.378	0.821	0.516	+/- 0.011
semsim-ctx	e5-at	r-10-1	0.74	0.382	0.849	0.527	-
semsim-ctx	e5-at	r-10-2	0.75	0.395	0.781	0.525	-
semsim-ctx	e5-at	r-10-3	0.75	0.405	0.815	0.541	-
semsim-ctx	e5-at	r-10-4	0.74	0.371	0.890	0.523	-
semsim-ctx	e5-at	r-10-5	0.74	0.357	0.881	0.509	-
semsim-ctx	e5-at	r-10-mean	0.75	0.395	0.766	0.521	+/- 0.016
semsim-ctx	e5-at	r-10-mean-best	0.74	0.382	0.843	0.525	+/- 0.012
semsim-ctx	gte-at	r-1-1	0.65	0.342	0.826	0.483	-
semsim-ctx	gte-at	r-1-2	0.68	0.386	0.774	0.515	-
semsim-ctx	gte-at	r-1-3	0.66	0.311	0.945	0.468	-
semsim-ctx	gte-at	r-1-4	0.63	0.300	1.000	0.461	-
semsim-ctx	gte-at	r-1-5	0.66	0.339	0.866	0.487	-
semsim-ctx	gte-at	r-1-mean	0.66	0.328	0.882	0.476	+/- 0.018
semsim-ctx	gte-at	r-1-mean-best	0.66	0.335	0.882	0.483	+/- 0.021
semsim-ctx	gte-at	r-3-1	0.69	0.340	0.827	0.482	-
semsim-ctx	gte-at	r-3-2	0.69	0.309	0.950	0.467	-
semsim-ctx	gte-at	r-3-3	0.72	0.348	0.852	0.494	-
semsim-ctx	gte-at	r-3-4	0.71	0.367	0.822	0.508	-
semsim-ctx	gte-at	r-3-5	0.67	0.317	0.930	0.473	-
semsim-ctx	gte-at	r-3-mean	0.69	0.324	0.883	0.472	+/- 0.013
semsim-ctx	gte-at	r-3-mean-best	0.70	0.336	0.876	0.485	+/- 0.017
semsim-ctx	gte-at	r-10-1	0.71	0.342	0.879	0.492	-
semsim-ctx	gte-at	r-10-2	0.72	0.336	0.910	0.491	-
semsim-ctx	gte-at	r-10-3	0.73	0.349	0.891	0.501	-
semsim-ctx	gte-at	r-10-4	0.72	0.341	0.885	0.492	-
semsim-ctx	gte-at	r-10-5	0.70	0.322	0.934	0.478	-
semsim-ctx	gte-at	r-10-mean	0.72	0.341	0.854	0.486	+/- 0.002
semsim-ctx	gte-at	r-10-mean-best	0.72	0.338	0.900	0.491	+/- 0.008

Table A.6.: Best F1-score evaluation runs for all ERS of the form `semsim-ctx NM-at r-N-X` (CNESS with AT option enabled)

A.3. Evaluation Metric Scores vs. Similarity Threshold Plots

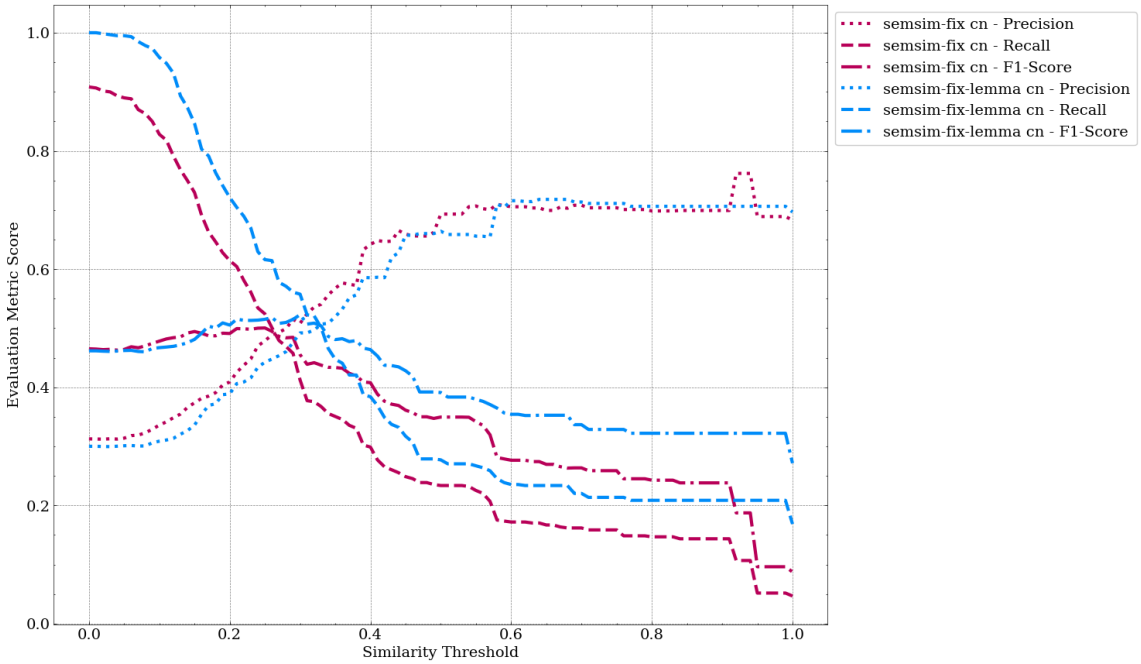


Figure A.1.: Precision, recall and F1-score vs. ST for the evaluation runs `semsim-fix cn` and `semsim-fix-lemma cn`

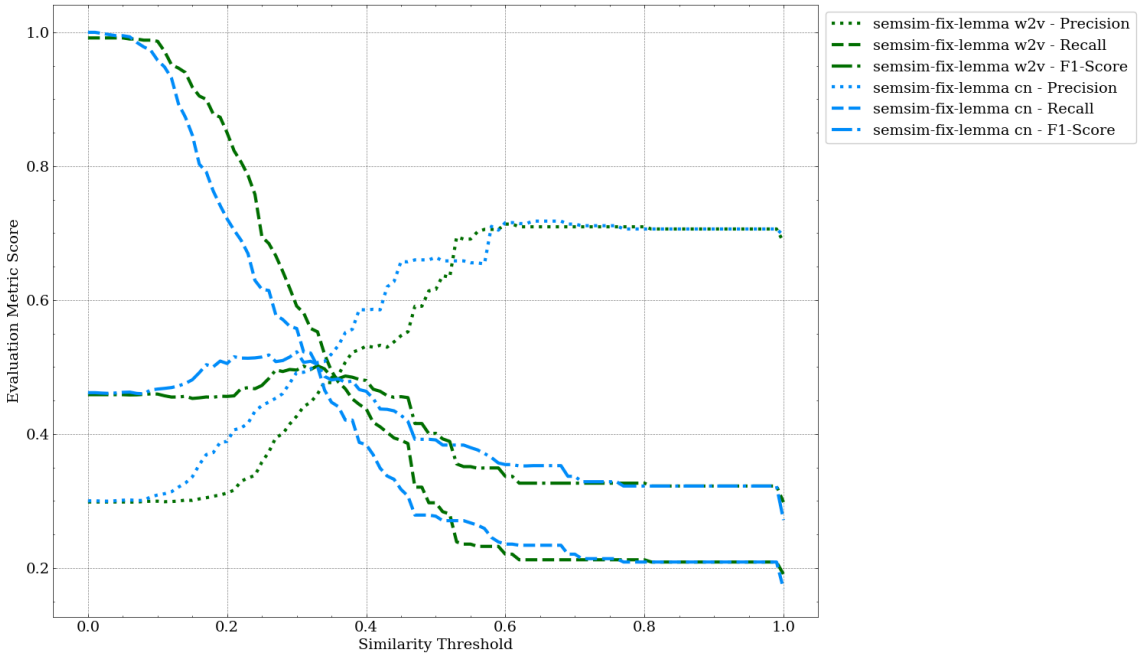


Figure A.2.: Precision, recall and F1-score vs. ST for the evaluation runs `semsim-fix-lemma w2v` and `semsim-fix-lemma cn`

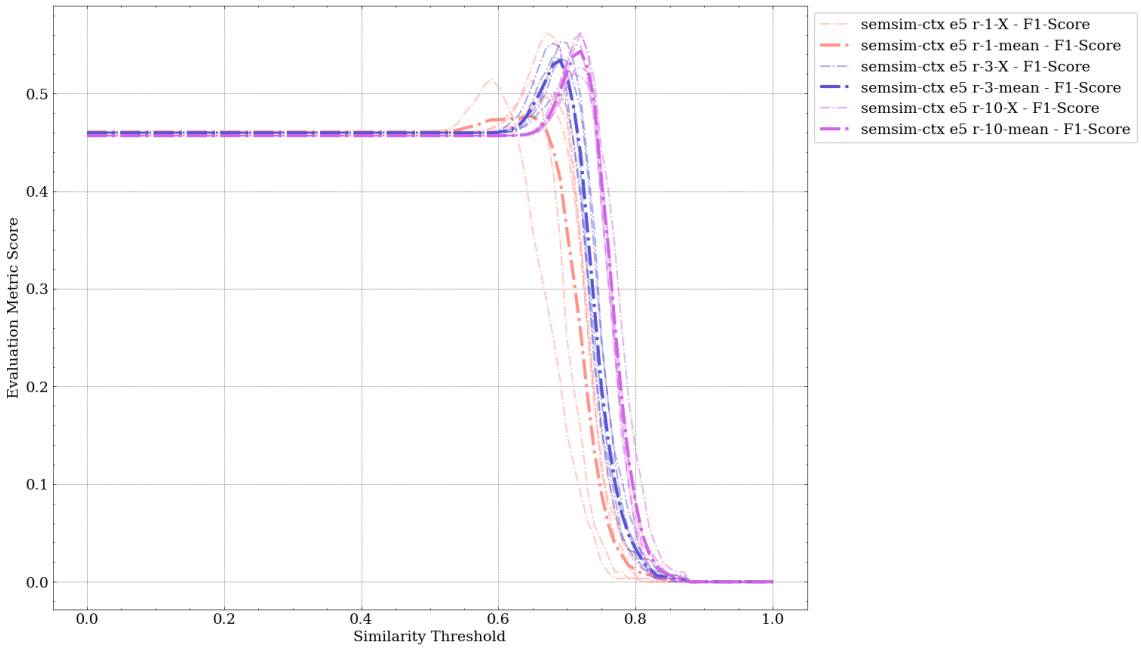


Figure A.3.: F1-score vs. ST for the evaluation runs `semsim-ctx e5 r-1-X`, `semsim-ctx e5 r-3-X` and `semsim-ctx e5 r-10-X`

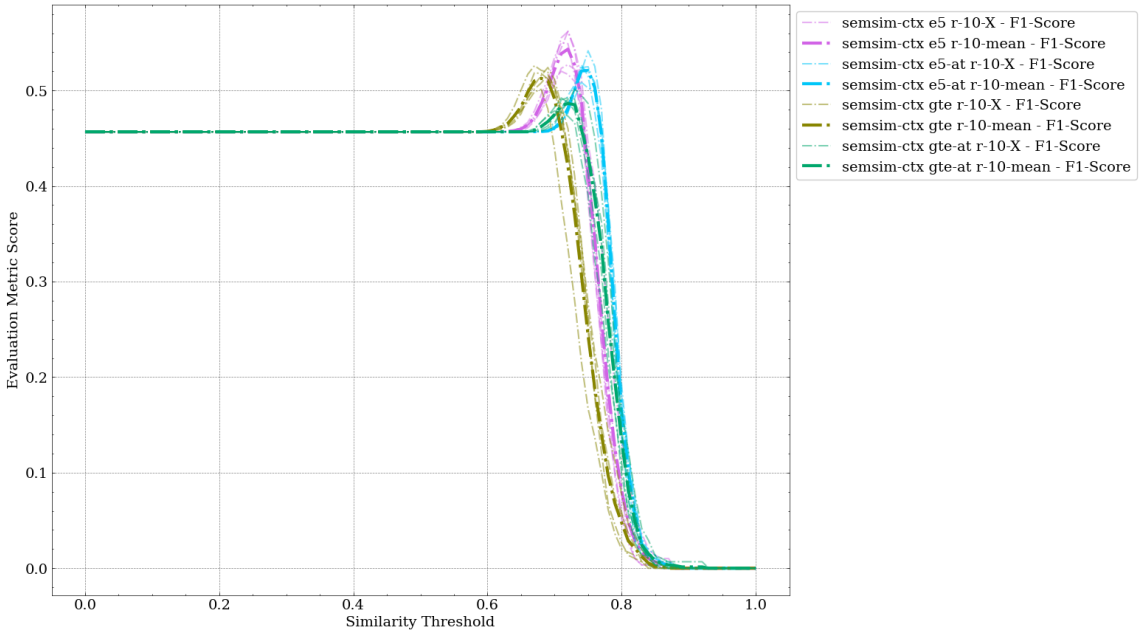


Figure A.4.: F1-score vs. ST for the evaluation runs `semsim-ctx e5 r-10-X`, `semsim-ctx e5-at r-10-X`, `semsim-ctx gte r-10-X` and `semsim-ctx gte-at r-10-X`

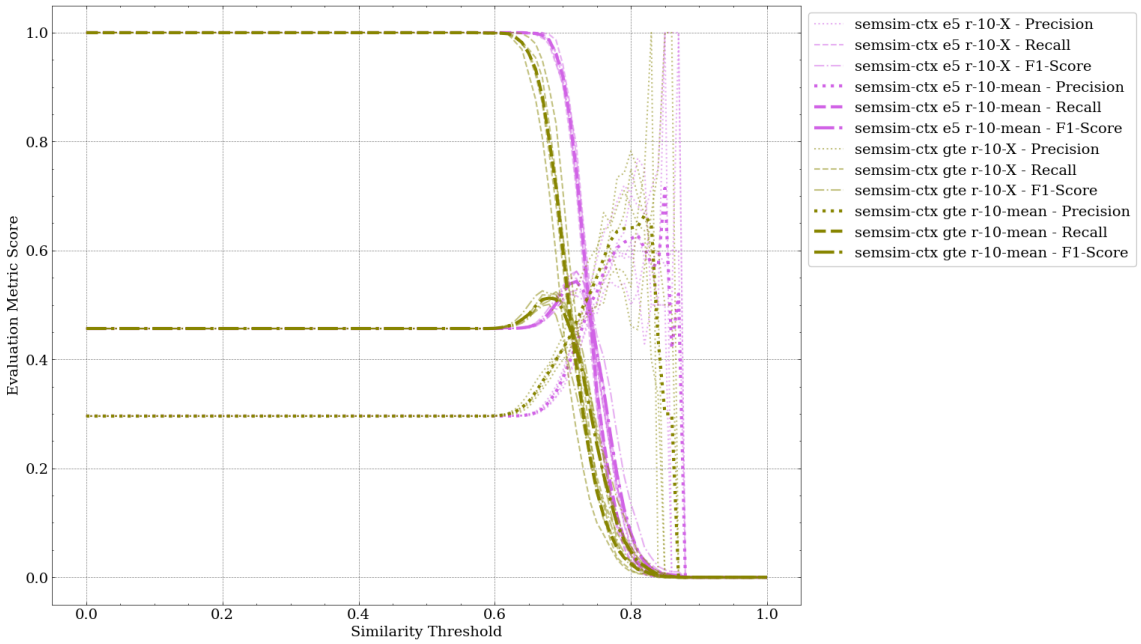


Figure A.5.: Precision, recall and F1-score vs. ST for the evaluation runs `semsim-ctx e5 r-10-X` and `semsim-ctx gte r-10-X`

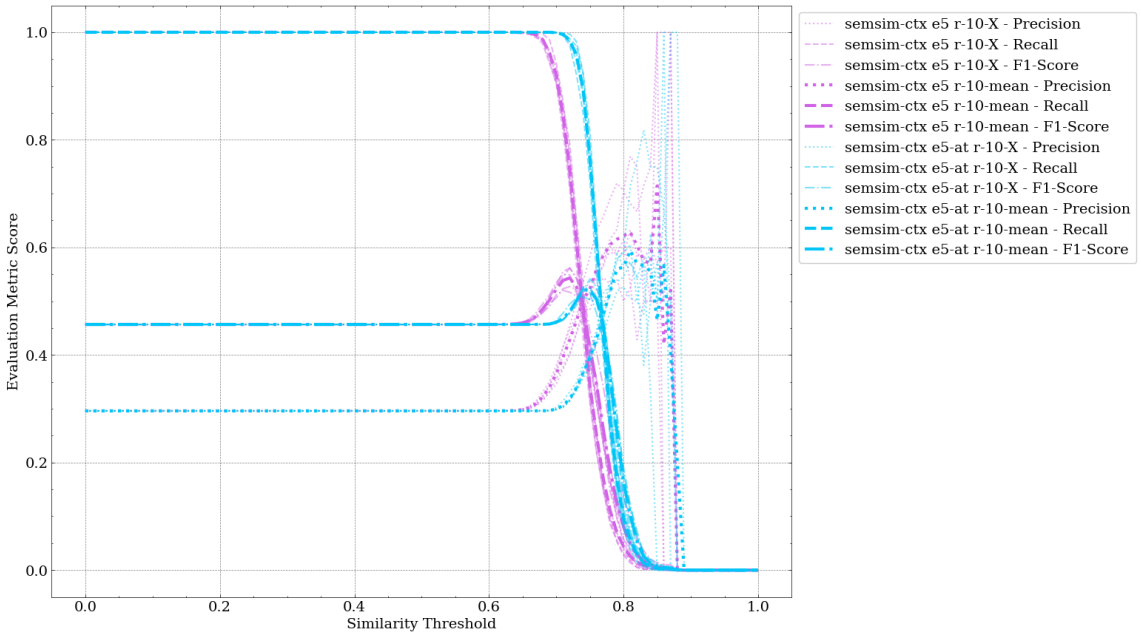


Figure A.6.: Precision, recall and F1-score vs. ST for the evaluation runs `semsim-ctx e5 r-10-X` and `semsim-ctx e5-at r-10-X`