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

January 31, 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

- 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

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 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 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***

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

(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)

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. In this case study quantitative results as well as qualitative examinations of the behaviour of NESS-SHPM are conducted. The quantitative results reflect the systems performance using metrics which are established for retrieval and classification tasks. The qualitative results exemplary showcase detailed aspects of the systems behaviour in the given use case.

refer to the RQs more specifically?

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 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 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.

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 BSE contains $n_b = 71804$ edges. Due to the time constrains of this work and the limited availability of three annotators, the BSE needs to be subsampled to create the CD. 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.

Filtering

Every edge in the BES theoretically corresponds to a predicate lemma, as stated above. Still it can occur that it is not possible to determine the predicate lemma of a given edge to to implementation issues, which out of scope of this work. In these cases an edge is filtered from the BSE, which results in the *filtered base edge set* (FBES). The FBES contains $n_f = 69380$ edges.

Sampling

The edges in the full dataset 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 full dataset is statistically represented multiple times in the subsampled dataset.

add lemma distribution plot

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 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 full dataset 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 full dataset will be represented in the subsampled dataset, while still representing the distribution of the full dataset.

do I have to show this in some way? should I then mention it at all?

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 $n_s = 2000$, 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, FD and SD are listed in table 6.2 for comparison, along with the number and percentage of edges which are labelled as an expression of conflict and of those which are not.

add examples

Edge set name	Number of all edges	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	-	-
Conflict Dataset (CD)	2000	599 (29.95 %)	1401 (70.05 %)

Table 6.2.: Number of edges and proportion of labels for the different edge sets

6.3. Evaluation Process

move some of the content from below up here

6.3.1. Evaluation Runs

The evaluation process consists of multiple *evaluation runs*. An evaluation run is described by an *evaluation (run) configuration* consisting of a dataset (which is always the CD in this case study) and an SH pattern which is matched against the dataset. For the semsim conflict patterns the evaluation run is further described by an *NESS configuration*. These configurations specify the properties of the NESS matching. That means the *NESS model* and in the case of CNESS it is also specified which set of reference edges is used and if the *all-tokens* version of the matching is used.

add references to chapter 2

For each evaluation run the *evaluation metrics* are computed. If the evaluation run corresponds to a semsim conflict pattern, this is done for every similarity threshold $t_s \in r_t$, where r_t is the range of similarity thresholds. For all evaluation that are conducted in this case study $r_t = \{0, 0.01, ..., 0.99, 1\}$ is chosen.

Conflict Patterns

Section 6.1.3 describes the four conflict patterns used in this evaluation in detail and derives why they are chosen. A structured overview of the properties of these patterns can be seen in table 6.3.

explain which properties

Reference Edge Sets

The reference edges 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 account for the effect of different reference edge sets, multiple different samples are drawn. To compare the effect of different sample sizes five sets with one reference edge in each and five sets with ten reference edges in each are sampled. In the evaluation run name, the set of reference edges used is denoted by the last part of the form "r-N-X", with where $N \in \{1, 3, 10\}$ is the number of reference edges and $X \in \{1, ..., 5\}$ is the the ID of the set of references edges. The specific sets of references edges, which have been sampled can be seen in

add table to the appendix

Pattern name	Lemma based	SemSim 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 (pattern 7)	Yes	FIXED	Yes	No
semsim-ctx conflict (pattern 7)	No	CONTEXT	No	Yes

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

Evaluation Configurations

Table 6.4 describes the configurations for all evaluation runs that are conducted in this case study. For evaluation runs based on the semsim-ctx conflict pattern, the ID of the reference edge sets has been omitted in this table. All five reference edges sets for both sample sizes have been used for the evaluation runs.

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.

6.4. Evaluation Results

add section introduction

6.4.1. Quantitative Results

add intro

Result Tables

Table table 6.5 shows results for all evaluation runs conducted in this case study. For every evaluation run, the score for all three evaluation metric is given. In case of the evaluation runs based on semsim conflict patterns, the scores relate to the *best F1 score*. This means among the results for all similarity threshold in the given threshold range for a given evaluation run, the ST that results in the highest F1-score is selected. The scores of the other evaluation metrics also correspond to this ST.

For all evaluation runs based on the semsim-ctx conflict pattern, only the *mean-best* evaluation score is listed for each evaluation run configuration. This is the mean of all evaluation scores corresponding to the best F1 score for one of the five different sets of reference edges

Evaluation Run Name	Conflict Pattern	NESS Configuration		
		Model	Use *all-tokens*	Num. of Ref. Edges
original	original	-	-	-
semsim-fix w2v	semsim-fix	word2vec	-	-
semsim-fix cn	semsim-fix	conceptnet-numb.	-	-
semsim-fix-lemma w2v	semsim-fix-lemma	word2vec	-	-
semsim-fix-lemma cn	semsim-fix-lemma	conceptnet-numb.	-	-
semsim-ctx e5 r-1	semsim-ctx	e5	No	1
semsim-ctx gte r-1	semsim-ctx	gte	No	1
semsim-ctx e5-at r-1	semsim-ctx	e5	Yes	1
semsim-ctx gte-at r-1	semsim-ctx	gte	Yes	1
semsim-ctx e5 r-3	semsim-ctx	e5	No	3
semsim-ctx gte r-3	semsim-ctx	gte	No	3
semsim-ctx e5-at r-3	semsim-ctx	e5	Yes	3
semsim-ctx gte-at r-3	semsim-ctx	gte	Yes	3
semsim-ctx e5 r-10	semsim-ctx	e5	No	10
semsim-ctx gte r-10	semsim-ctx	gte	No	10
semsim-ctx e5-at r-10	semsim-ctx	e5	Yes	10
semsim-ctx gte-at r-10	semsim-ctx	gte	Yes	10

Table 6.4.: Evaluation Run Configurations

of a specific size. For example *semsim-ctx e5-at r-10-best-mean* refers to the mean of best F1 scores for all evaluation runs that use the semsim-ctx conflict pattern, the e5 model with all-tokens enabled and a reference edge set of size ten.

In table A.1 and table A.2 of appendix A.2 the complete evaluation results for the semsim-ctx conflict pattern based evaluation runs can be seen. For each evaluation run configuration there are five entries for the different sets of reference edges, the discussed *mean-best* entry and an additional *mean* entry. The mean evaluation runs are constructed from the mean value for all evaluation scores across all reference edge sets for a given evaluation run configuration, i.e. for every ST the mean of the corresponding evaluation scores of all reference edge sets is computed. These mean evaluation runs are primarily relevant for visualisation in the result plots (see section 6.4.1).

For the mean-best and the mean entries there is an additional column listed in table 6.5, table A.1 and table A.2, which shows the standard deviation of the F1 score for this entry. For the mean evaluation run, the standard deviation is computed across all reference edge sets for a specific evaluation run configuration. This results in a value for the standard deviation for each ST, whose mean is then computed. For the mean-best evaluation runs this is the standard deviation of the best F1 scores across all reference edge sets for a for a specific evaluation run configuration.

Result Plots

Eval. Metrics vs. ST Plots These plots visualise the resulting scores for the different evaluation metrics in relation to different value for the ST.

Evaluation Run Name			ST	Prec.	Rec.	(Best) F1-Score	
							Std. Dev.
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.: Best precision, recall and F1 scores for all evaluation runs

Figure 6.1 shows the best performing evaluation runs for each pattern. That means for each conflict pattern, this shows the evaluation run with the configuration that resulted in the highest best F1-score.

Figure 6.3 shows the evaluation runs for the best performing semsim-fix(-lemma) based evaluation configurations and the best performing semsim-ctx based evaluation configuration. For all evaluation runs the scores of all three evaluation metrics are shown.

More plots of this kind can be found in the appendix.

6.4.2. Qualitative Results

table with evaluation runs, their

6.4.3. Result Discussion

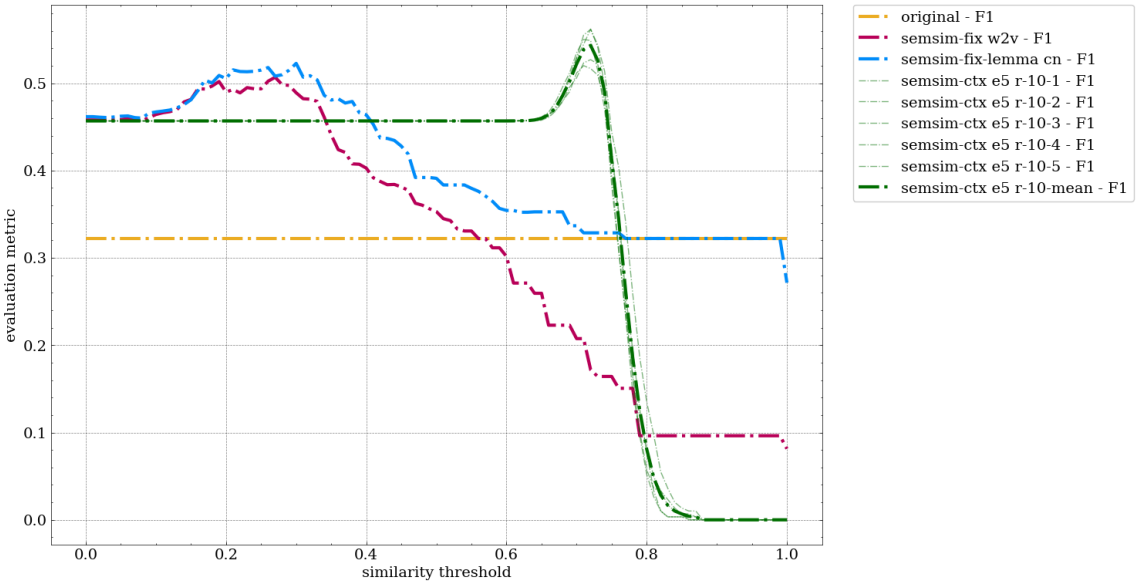

Figure 6.1.: F1-score vs. ST of best performing evaluation run configurations for each conflict pattern

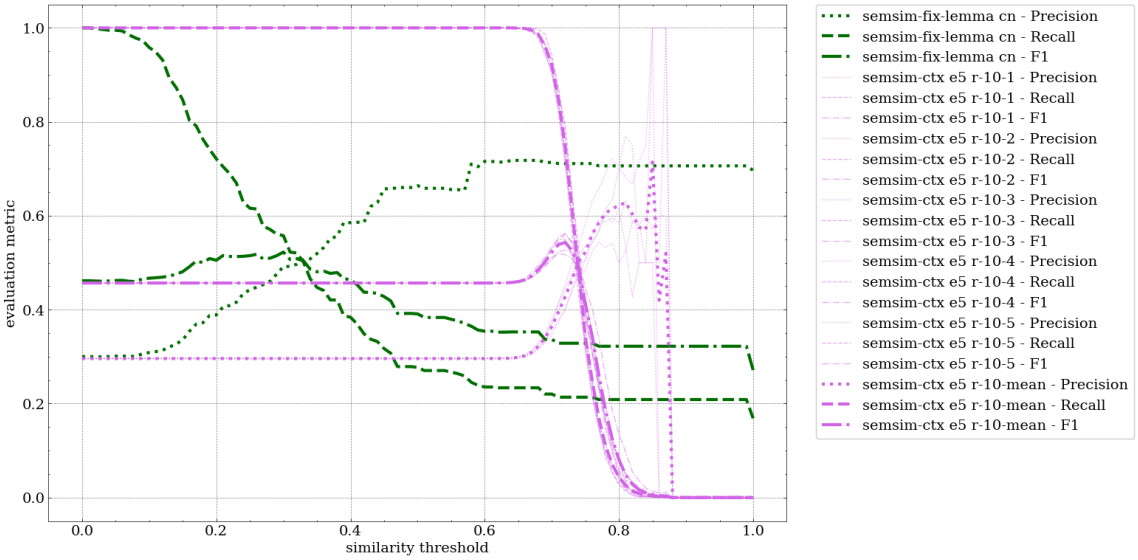

Figure 6.2.: Precision, recall and F1-score vs. ST for `semsim-fix-lemma cn` and `semsim-ctx e5 r-10`

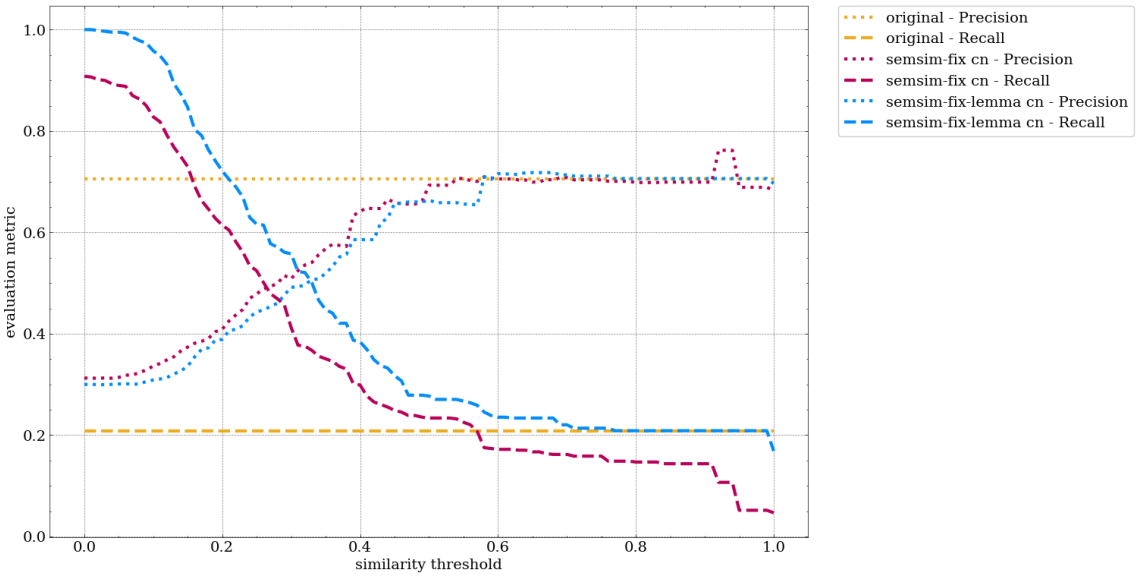

Figure 6.3.: Precision and recall vs. ST for `semsim-fix cn` and `semsim-ctx e5 r-10`

`semsim-fix-lemma cn`			`semsim-ctx e5 r-10-X`		
Lemma	F1-score	Num. of Samples	Lemma	F1-score	Num. of Samples
arrest	1.00	11	arrest	1.00	11
accuse	0.97	33	file	1.00	5
warn	0.93	24	attack	1.00	12
slam	0.92	7	accuse	0.95	33
attack	0.91	12	slam	0.92	7
criticize	0.91	6	criticize	0.91	6
detain	0.89	5	shoot	0.89	5
shoot	0.89	5	order	0.88	14
condemn	0.84	25	detain	0.86	5
dismiss	0.80	6	launch	0.86	14

Table 6.6.: Top ten lemmas regarding F1-score with number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` and `semsim-ctx e5 r-10-X`

Lemma	Abs. Prec. Diff.	F1 Eval. Run. A	F1 Eval. Run B	Num. of Samples
capture	0.71	0.29	1.00	7
accept	0.36	0.14	0.50	7
detain	0.20	0.80	1.00	5
attack	0.17	0.83	1.00	12
strike	0.17	0.67	0.50	9
suggest	0.13	0.20	0.33	5
deny	0.11	0.30	0.41	20
claim	0.10	0.27	0.38	22
threaten	0.09	0.35	0.44	20
say	0.08	0.29	0.38	31

Table 6.7.: Top ten lemmas regarding the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B)

Lemma	Prec Difference	F1 Eval. Run A	F1 Eval. Run B	Num. of Samples
shoot	0.00	0.80	0.80	5
criticize	0.00	0.83	0.83	6
dismiss	0.00	0.67	0.67	6
arrest	0.00	1.00	1.00	11
slam	0.00	0.86	0.86	7
seize	0.00	0.50	0.50	12
condemn	0.00	0.72	0.72	25
kill	0.01	0.49	0.49	77
urge	0.01	0.51	0.50	41
destroy	0.02	0.58	0.60	12

Table 6.8.: Top ten lemmas regarding the lowest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B)

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.

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).

A. Appendix

A.1. Reference Edge Sets

A.2. Complete Result Tables

Evaluation Run Name			ST	Prec.	Rec.	(Best) F1-Score	
							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-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	0.69	0.392	0.846	0.536	-
semsim-ctx	e5	r-3-2	0.68	0.361	0.814	0.500	-
semsim-ctx	e5	r-3-3	0.68	0.435	0.752	0.551	-
semsim-ctx	e5	r-3-4	0.69	0.410	0.846	0.553	-
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	0.400	0.771	0.527	-
semsim-ctx	e5	r-10-3	0.72	0.435	0.793	0.562	-
semsim-ctx	e5	r-10-4	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	0.55	0.300	1.000	0.461	-
semsim-ctx	gte	r-1-3	0.61	0.339	0.829	0.482	-
semsim-ctx	gte	r-1-4	0.63	0.388	0.799	0.522	-
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	0.67	0.378	0.713	0.494	-
semsim-ctx	gte	r-3-2	0.67	0.385	0.740	0.506	-
semsim-ctx	gte	r-3-3	0.66	0.376	0.820	0.516	-
semsim-ctx	gte	r-3-4	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	0.69	0.399	0.730	0.516	-
semsim-ctx	gte	r-10-3	0.68	0.388	0.708	0.501	-
semsim-ctx	gte	r-10-4	0.67	0.361	0.917	0.518	-
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.1.: Best F1-score results for all semsim-ctx conflict pattern based evaluation runs not using all-tokens

Evaluation Run Name			ST	Prec.	Rec.	(Best) F1-Score	
							Std. Dev.
semsim-ctx	e5-at	r-1-1	0.71	0.365	0.798	0.501	-
semsim-ctx	e5-at	r-1-2	0.71	0.351	0.908	0.507	-
semsim-ctx	e5-at	r-1-3	0.71	0.434	0.701	0.536	-
semsim-ctx	e5-at	r-1-4	0.66	0.356	0.885	0.508	-
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-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	0.74	0.357	0.881	0.509	-
semsim-ctx	e5-at	r-10-3	0.74	0.371	0.890	0.523	-
semsim-ctx	e5-at	r-10-4	0.75	0.405	0.815	0.541	-
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	0.66	0.311	0.945	0.468	-
semsim-ctx	gte-at	r-1-2	0.68	0.386	0.774	0.515	-
semsim-ctx	gte-at	r-1-3	0.66	0.339	0.866	0.487	-
semsim-ctx	gte-at	r-1-4	0.63	0.300	1.000	0.461	-
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	0.67	0.317	0.930	0.473	-
semsim-ctx	gte-at	r-3-3	0.71	0.367	0.822	0.508	-
semsim-ctx	gte-at	r-3-4	0.72	0.348	0.852	0.494	-
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-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.2.: Best F1-score results for all semsim-ctx conflict pattern based evaluation runs using all-tokens

A.3. Additional Result Plots

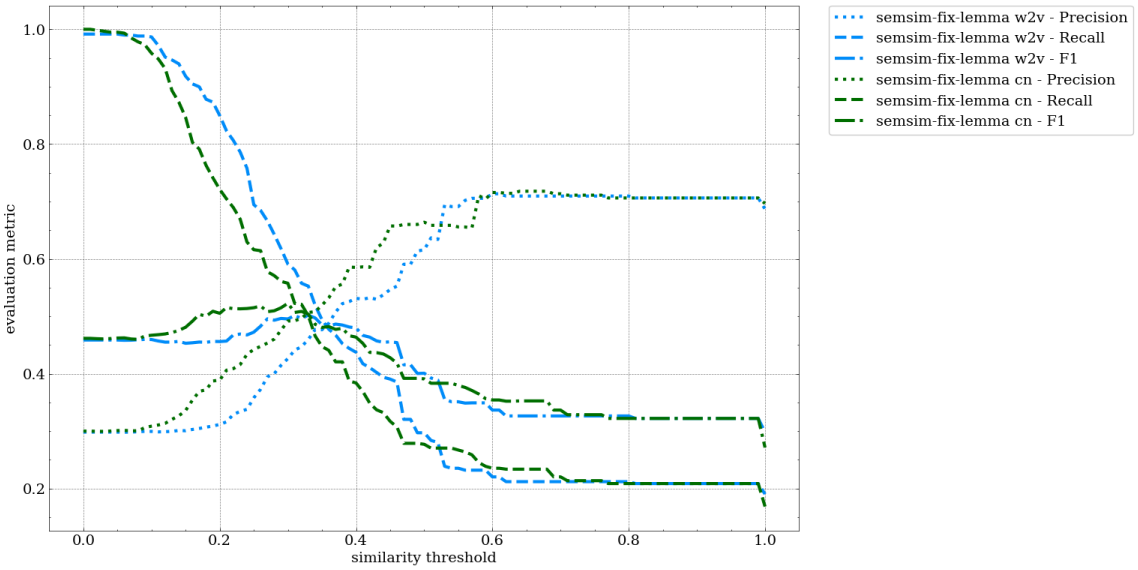

Figure A.1.: Caption for the figure

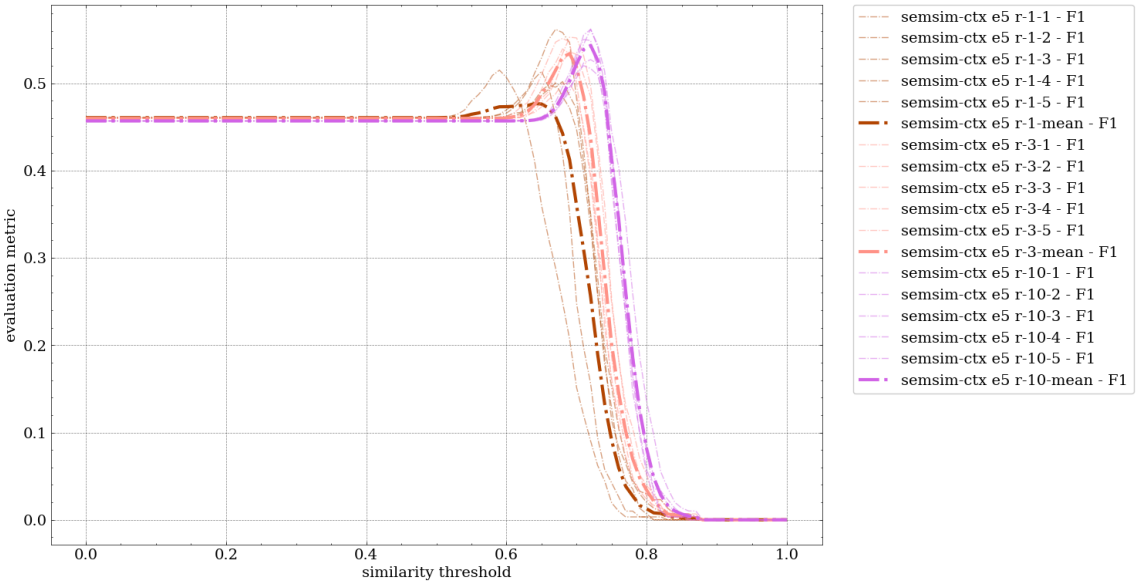

Figure A.2.: Caption for this figure

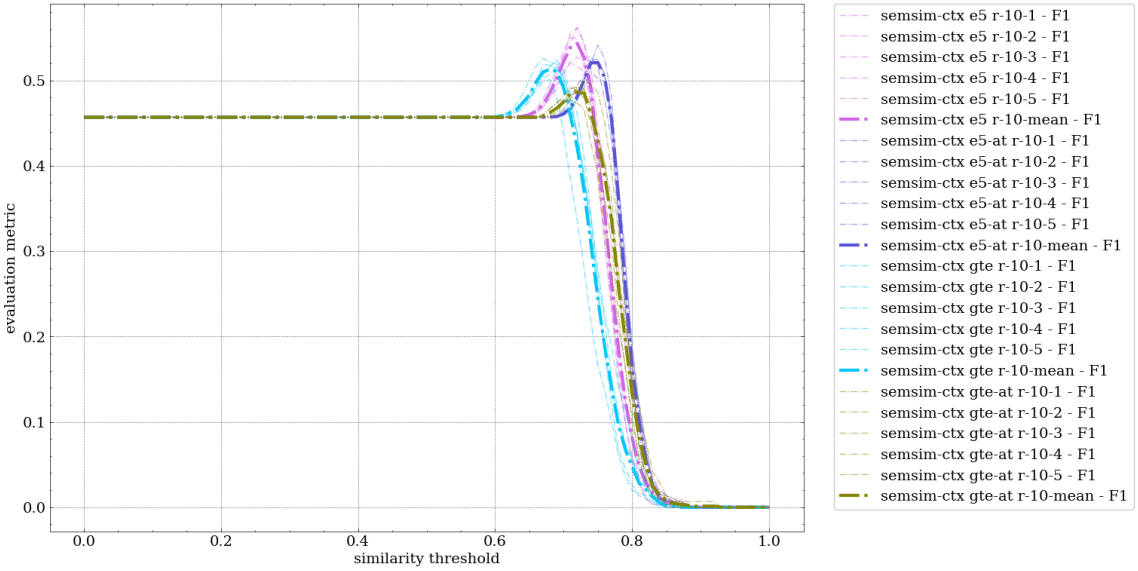

Figure A.3.: Another caption

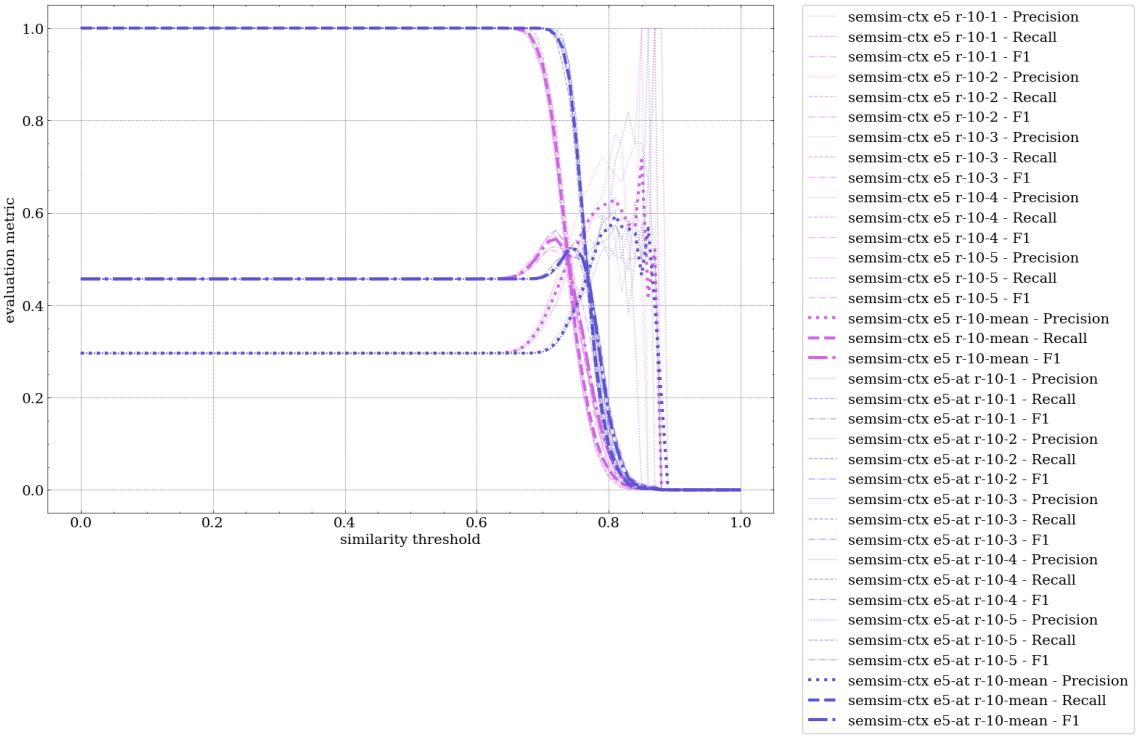

Figure A.4.: Another caption

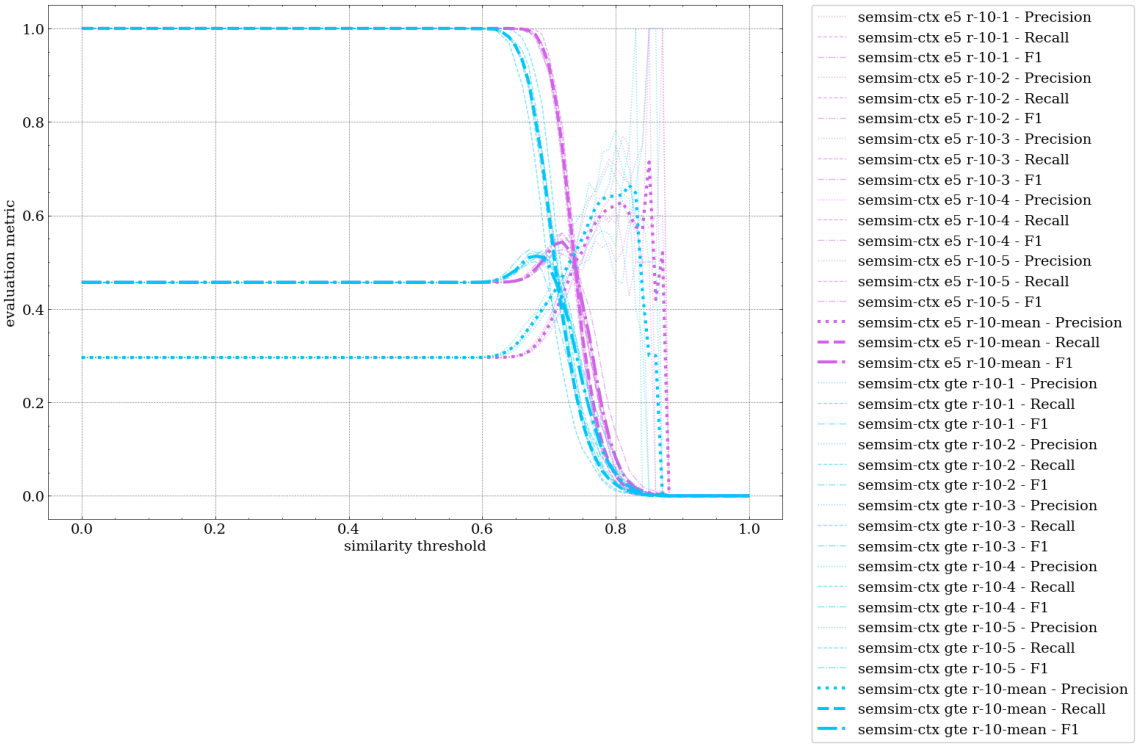

Figure A.5.: Another caption

A.4. Edge Predicate Lemma Evaluation Labels

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
capture (7)	Video captures horror of Istanbul airport attack	no conflict	conflict	no conflict)
	Syrian Troops Capture Village Near Northern City of Aleppo	no conflict	conflict	no conflict
	Iraqi forces capture key parts of Mosul neighborhood	no conflict	conflict	no conflict
	Iraqi forces capture west Mosul's main buildings in pre-dawn raid	no conflict	conflict	no conflict
	The videos capture 10 seconds of movement, just enough to catch trucks' license plates—the operation's main objective	no conflict	conflict	no conflict
	Security Service of Ukraine Captured Russian Spy	conflict	conflict	conflict
	Syrian Army strikes back in Deir Ezzor: Tal Barouk captured (Report + Video	conflict	conflict	no conflict
accept (7)	Saudi blogger Badawi's wife accepts Sakharov Prize	no conflict	conflict	no conflict
	UK accepts 1,500 asylum seekers from Syria	no conflict	conflict	no conflict
	Syria rebel chief rejects US-Russia chemical arms deal — "We cannot accept any part of this initiative	conflict	conflict	conflict
	Catholic hospitals accept birth control compromise	no conflict	conflict	conflict
	Japan accepts less than 1% of asylum applications	no conflict	conflict	no conflict
	Nato accepts Afghan leader Karzai's air strikes decree	no conflict	conflict	no conflict
	S. Korea accepts North's Sunday talks proposal, calls for Panmunjom meeting	no conflict	conflict	no conflict
detain (5)	Syrian rebels detain U.N. peacekeepers	conflict	conflict	no conflict
	Russian police detain several gay activists	conflict	conflict	conflict
	Mexico detains growing number of undocumented Cubans	conflict	conflict	conflict
	India releases detained Iranian ship	no conflict	conflict	no conflict
	Police in China have detained six executives of a meat supply company that supplied long-expired meat to foreign fast food chains McDonald's and KFC-parent Yum Brands Inc, among many others	conflict	conflict	conflict

Table A.3.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. one to five of top ten lemmas)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
deny (20)	Syrian officials deny chemical weapon use	no conflict	conflict	conflict
	Germany, EU deny report on European solidarity tax	no conflict	conflict	conflict
	Tebus deny rumours about handing over the last base	no conflict	conflict	no conflict
	Toronto mayor Rob Ford reportedly denied entry to US	conflict	conflict	conflict
	Kenya to deny entry to Ebola states	conflict	conflict	conflict
	The Vatican has categorically denied a report in an Italian newspaper that Pope Francis might have a small, yet curable brain tumor	no conflict	conflict	no conflict
	Indian government denies renewal of passport opposition lawyer over minor traffic rule violation	conflict	conflict	no conflict
	Mexican judge denies El Chapo's appeals against extradition	conflict	conflict	conflict
	Syria denies use of any chemical weapons	no conflict	conflict	conflict
	Toronto mayor denies existence of crack smoking video	no conflict	conflict	no conflict
	North Korea denies responsibility for Sony cyberattack	conflict	conflict	conflict
	Russia denies ground troop in Syria	no conflict	conflict	conflict
	New Pro-Kremlin Crimean Prime Minister Aksyonov denies allegations of criminal past	no conflict	conflict	no conflict
	Free Syrian Army denies responsibility	no conflict	conflict	conflict
	UK defense minister denies report of divisions in PM May's team	no conflict	conflict	no conflict
	Iran denies allegations of organizing spy cell in Nigeria	no conflict	conflict	no conflict
	Russia denies knowledge of arrest ISIS leader Baghdadi	no conflict	conflict	conflict
	Turkish FM denies investigation claims on German firms	no conflict	conflict	no conflict
	Russia Denies Any Role in Deadly Convoy Attack Syria	no conflict	conflict	conflict
	China denies Hong Kong visit request by U.S. carrier group: Pentagon	conflict	conflict	conflict

Table A.4.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. four of top ten lemmas)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
attack (12)	Pussy Riot members attack bandmates for appearing at Amnesty concert	conflict	conflict	conflict
	Israelis Attack Palestinian Farmers near Hebron	conflict	conflict	conflict
	Yemeni forces attack sixth Saudi warship	conflict	conflict	conflict
	Passengers attack Edinburgh taxi driver	conflict	conflict	conflict
	Angry Palestinians Attack Hamas Official Over Gaza Destruction	conflict	conflict	conflict
	Israeli Air Force attacks Assad army targets	conflict	conflict	conflict
	They Beat Me': Univision Reporter Attacked Outside Venezuelan Supreme Court	no conflict	conflict	no conflict
	Locals attack police to foil casino raid	conflict	conflict	conflict
	Afghans attack Indian consulate	conflict	conflict	conflict
	Suspected PKK supporters attack Turkish building in Germany	conflict	conflict	conflict
	Paris attacks a reaction to US actions in Syria, Iraq: Indian Minister	no conflict	conflict	no conflict
	Muslim mob attacks Mosque in Pakistan	conflict	conflict	conflict
strike (9)	Ex-Muslim poet strikes fear in PC Denmark	conflict	conflict	no conflict
	Double hotel bombing strikes Iraqi capital	no conflict	conflict	conflict
	Magnitude 6.9 quake strikes Sichuan region of China	no conflict	conflict	conflict
	Syrian Israel strikes Syrian targets	conflict	conflict	conflict
	Israeli warplanes strike Syrian weapons facility	conflict	conflict	conflict
	U.S. planes strike militants near Iraq's Amreli, airdrop aid	conflict	conflict	no conflict
	Fake threats strike fear for Sochi Olympic Security	no conflict	conflict	conflict
	UK strikes first ISIS targets	conflict	conflict	no conflict
	Iraqi Army strikes 3 Islamic State strongholds	conflict	conflict	conflict
suggest (5)	Turkish state news agency suggests link between Boko Haram and Western interests in Nigerian oil	no conflict	conflict	conflict
	New study suggests more and longer atmospheric stagnation events due to global warming	no conflict	conflict	no conflict
	A prominent Iranian official recently suggested a new drug policy for the country that includes taking steps toward the legalization of cannabis and opium	no conflict	conflict	no conflict
	U.N. aid chief suggests more intervention in humanitarian emergencies	no conflict	conflict	conflict
	North Korea's behavior suggests a possible EMP strike against the U.S.	conflict	conflict	conflict

Table A.5.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. five to seven of top ten lemmastop ten lemmas)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
claim (22)	Chevron claims new proof of fraud in Ecuador pollution ruling	no conflict	conflict	no conflict
	Iran claims new generation of 15-times-faster centrifuges	no conflict	conflict	no conflict
	A prolonged drought in Brazil has already claimed about half of Jose Francisco Pereira's coffee crop	no conflict	conflict	no conflict
	ISIS backers claim responsibility for Paris-style terror attack in Jakarta	no conflict	conflict	conflict
	Salafist group claims responsibility for bombing French center in Gaza	no conflict	conflict	conflict
	Rockets fired on southern Israel from Sinai, ISIS claims responsibility	conflict	conflict	conflict
	ISIS claim responsibility for Paris terror attacks in online statement	no conflict	conflict	conflict
	Isis claims responsibility for Paris shooting attack that left one police officer dead	no conflict	conflict	conflict
	Convicted Israeli trainer of Colombia paramilitaries claims CIA ties	no conflict	conflict	no conflict
	Pakistan Taliban faction claims park attack on Lahore Christians	conflict	conflict	conflict
	Burundi Rebels Claim Rwanda Military Training	no conflict	conflict	conflict
	Isis claims responsibility for killing of Hindu priest in Bangladesh	conflict	conflict	conflict
	PKK-linked TAK claims responsibility for Ankara attack	no conflict	conflict	conflict
	Ansar Bayt al-Maqdis claims responsibility for opening fire on Israeli soldiers near Egyptian border	conflict	conflict	conflict
	Sectarian war an aim of Gulf states, Israel and Turkey, claims Hezbollah	conflict	conflict	conflict
	Iran Sunni group Jaish al-Adl claims border attack	no conflict	conflict	conflict
	ISIS Claims Responsibility In Turkish Nightclub Attack; U.S. Man Among The Wounded	conflict	conflict	conflict
	Bangkok police claim reward in Erawan Shrine bomber hunt	no conflict	conflict	no conflict
	Taliban claim Kabul attack	no conflict	conflict	conflict
	Denmark claims north pole	no conflict	conflict	no conflict
	WikiLeaks' Julian Assange Claims Vindication in UN Ruling	no conflict	conflict	conflict
	ISIS claims responsibility for Eilat rockets	no conflict	conflict	conflict

Table A.6.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. eight of top ten lemma)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
threaten (20)	North Korea Threatens To "Invade USA," Use Weapons "Unknown To The World	conflict	conflict	conflict
	FATCA threatens Russia's financial system – official	conflict	conflict	conflict
	Wartime economic crisis threatens education of millions Yemeni children	no conflict	conflict	no conflict
	Syria opposition threatens withdrawal from Geneva talks	no conflict	conflict	conflict
	Al-Qaeda affiliates are threatening West Africa's most peaceful cities	conflict	conflict	conflict
	Video threatens Sochi Winter Olympics	no conflict	conflict	conflict
	Ukraine crisis threatens Transnistria	no conflict	conflict	conflict
	Giant comets may threaten Earth	no conflict	conflict	no conflict
	Protests in Turkey Threaten Erdogans Political Future	conflict	conflict	conflict
	Brazil Rejects Israel's Ambassador; Israel Threatens Relations Downgrade : NPR & conflict	conflict	conflict	
	North Korea Threatens 'Merciless' Strike Against US-South Korea Drill	conflict	conflict	conflict
	Kim Dotcom case threatens New Zealand Government	no conflict	conflict	conflict
	Proposed German legislation threatens broad internet censorship	no conflict	conflict	conflict
	Abe government threatens NHKs credibility	no conflict	conflict	conflict
	British ISIS fighter 'Al-Britani' threatens executions in Trafalgar Square	no conflict	conflict	conflict
	How one racy condom ad may threaten gains female sex education in Pakistan	no conflict	conflict	no conflict
	Police Use of Drones May Threaten Human Rights: UN Expert & no conflict	conflict	conflict	
	North Korea threatens pre-emptive nuke strike against U.S. & S. Korea	conflict	conflict	conflict
	Asylum seekers threaten hunger strike	no conflict	conflict	conflict
	Withdrawal of foreign troops from Afghanistan threatens rollback of women's gains	no conflict	conflict	no conflict

Table A.7.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. nine of top ten lemma)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
say (16 of 31)	Iran says no snap inspections of nuclear sites	no conflict	conflict	no conflict
	Child abuse image investigation leads to 660 arrests: UK National Crime Agency said the 660 arrested included doctors, teachers, scout leaders, care workers and former police officers	no conflict	conflict	no conflict
	Former Cuba Leader Fidel Castro Says 'Israel and US Fathered Isis	no conflict	conflict	conflict
	Malaysian court to Christians: You can't say 'Allah	conflict	conflict	conflict
	U.S. State Dept. says "confident" Russian gov	no conflict	conflict	conflict
	State Border Service says Russian troops still near Ukraine's border	no conflict	conflict	no conflict
	France military says Mali town Konna 'not recaptured' from Islamists denying an earlier claim by the Malian army	conflict	conflict	no conflict
	Thai PM says occupation of state buildings threat to stability	no conflict	conflict	conflict
	Australia says Chinese spy ship near war games	no conflict	conflict	no conflict
	Denmark Says Microsoft Owes £660 Million in Tax	conflict	conflict	no conflict
	Japan Says Armed Chinese Ship Infiltrates Its Territorial Waters	conflict	conflict	conflict
	EU report says LGBT face discrimination	no conflict	conflict	conflict
	Group says boycott Qatar Airways, cites poor rights record	conflict	conflict	conflict
	WHO incapable of reacting to crises such as Ebola, says report	no conflict	conflict	no conflict
	Migrant ship off Greece says armed people on board	no conflict	conflict	conflict
	Turkey says attacks on Aleppo a crime against humanity	no conflict	conflict	conflict

Table A.8.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. ten of top ten lemma)

Lemma (n_s)	Edge Content	Label Dataset	Label Eval. A	Label Eval. B
say (15 of 31)	Mystery man in Bangkok bomb probe 'never said a word	no conflict	conflict	no conflict
	I Recant Says Author of Infamous Seventies Newsweek Global Cooling Article	no conflict	conflict	no conflict
	Sister says don't make missing Flight 370 pilot the fall guy	no conflict	conflict	no conflict
	VW works council says talks over strategy pact broken off without results	no conflict	conflict	no conflict
	Head of Russia's Rosneft says U.S., not OPEC, Rules Oil Markets	no conflict	conflict	no conflict
	Indian textbook says "Japan Nuked US	no conflict	conflict	no conflict
	North Korea says US sanctions on leader "a declaration of war	conflict	conflict	conflict
	President of Gambia says homosexuality one of three greatest threats to humanity	conflict	conflict	conflict
	China Says Its Working with Latin America for a New World Order	no conflict	conflict	conflict
	Myanmar Parliament Chairman Says Nominees for Country's Next President, 2 Vice Presidents Will Be Revealed March 17	no conflict	conflict	no conflict
	G7 says no sanctions on Russia over Syria	no conflict	conflict	conflict
	South Korea's Lotte Duty Free says China cyber attacks crashed website	conflict	conflict	no conflict
	Kuwait says stateless to be offered Comoros citizenship	no conflict	conflict	conflict
	Syria Says Israel Attacked Military Airport	conflict	conflict	conflict
	Bainimarama says NZ easing of sanctions 'insincere	no conflict	conflict	conflict

Table A.9.: Dataset labels and evaluation labels for edges corresponding to predicate lemmas with the highest abs. diff. in precision between the evaluation runs with recall > 0 and number of samples per lemma $n_s >= 5$ for the evaluation runs `semsim-fix-lemma cn` (A) and `semsim-ctx e5 r-10-X` (B) (Num. ten of top ten lemma)