*Natural Language Engineering* (2022), 1–24
doi:10.1017/S1351324922000092

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# Explainable lexical entailment with semantic graphs

Adam Kovacs[1,2], Kinga Gemes[1,2], Andras Kornai[3] and Gabor Recski[2,*]

[1]Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary, [2]TU Wien, Vienna, Austria, and [3]SZTAKI Institute of Computer Science, Budapest, Hungary
*Corresponding author. E-mail: gabor.recski@tuwien.ac.at

(Received 21 January 2021; revised 27 January 2022; accepted 31 January 2022)

**Abstract**

We present novel methods for detecting lexical entailment in a fully rule-based and explainable fashion, by automatic construction of semantic graphs, in any language for which a crowd-sourced dictionary with sufficient coverage and a dependency parser of sufficient accuracy are available. We experiment and evaluate on both the Semeval-2020 lexical entailment task (Glavaš *et al*. (2020). *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 24–35) and the `SherLIiC` lexical inference dataset of typed predicates (Schmitt and Schütze (2019). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 902–914). Combined with top-performing systems, our method achieves improvements over the previous state-of-the-art on both benchmarks. As a standalone system, it offers a fully interpretable model of lexical entailment that makes detailed error analysis possible, uncovering future directions for improving both the semantic parsing method and the inference process on semantic graphs. We release all components of our system as open source software.

**Keywords:** Semantics; Textual entailment

## 1. Introduction

The ability to model lexical entailment is a test of the adequacy of any theory of lexical semantics. We study lexical entailment not merely as another engineering task in the field of natural language processing (NLP), but as a central goal of our efforts towards building generic and actionable representations of natural language semantics. We present a rule-based, language-independent system for detecting entailment between pairs of words using `4lang` semantic graphs (Kornai *et al*. 2015; Recski 2018) built from dependency parses of dictionary definitions. We also present a novel architecture for constructing and manipulating graphs with synchronous grammars and use it to implement improvements of previous methods for building and expanding concept graphs from dictionaries. Our goal is to establish symbolic representations of lexical semantics that allow us to model entailment stragihtforwardly, as overlaps between premise and hypothesis graphs.

Stopping conditions of our system can be adjusted to multiple formulations of the entailment task. We present experiments on two substantially different datasets, the Semeval-2020 task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment" (Glavaš *et al*. 2020) and the `SherLIiC` benchmark for context aware-typed lexical inference (Schmitt and Schütze 2019). On the Semeval task of detecting binary lexical entailment, we achieve new state-of-the- art results on three languages, two of which were held by our earlier system (Kovács *et al*. 2020). On the second, more challenging benchmark, our method outperforms all rule-based baselines and also allows for a slight improvement over the top-performing system. Perhaps more importantly, the high precision of our method makes it possible to combine it with a neural NLI system based on

RoBERTa (Zhuang *et al.* 2021), improving its overall performance. As our grounds for establishing entailment between a pair of lexical entries is always some overlap between two concept graphs, predictions made by our system are all directly explainable as common subgraphs of premise and hypothesis definitions, providing an example of transparent, trustworthy, and explainable AI (xAI).

The article is structured as follows. Section 2.1 reviews the formulations of the lexical entailment task and corresponding datasets, with special emphasis on the two datasets used in our work. Then in Section 2.2, we survey recent approaches to entailment and inference tasks. We then provide a short overview of common approaches to semantic parsing in Section 2.3 and review approaches to modeling entailment with semantic graphs in Section 2.4. Section 3 describes our pipeline for building `4lang` semantic graphs from Wiktionary entries and for obtaining additional synonyms both from Wiktionary and WordNet. The section also provides an overview of our method of using graph grammars to transform dependency trees to `4lang` graphs, and finally describes our method for establishing entailment over pairs of `4lang` graphs. Section 4 evaluates our method on two recent benchmarks and compares their performance to previous systems, also experimenting with simple strategies for combining them with the output of state-of-the art NLI systems for improved performance. Finally, Section 5 presents the results of manual error analysis on both datasets, providing insight about the differences between the two formulations of the entailment task and identifying current shortcomings of our approach, along with possible solutions. All software described in this article is open-source, released under an MIT license.[a,b,c]

## 2. Related work

### *2.1 Tasks and datasets*

Entailment between pairs of words has been studied extensively both as one of the fundamental relationships in the lexicon and as an essential building block of models of natural language inference. Several recent task formulations equate lexical entailment with *hypernymy/hyponymy* or the `IS_A` relationship (Vulić *et al.* 2017; Vulić, Ponzetto, and Glavaš 2019), treating it as a relationship between two entries in a lexicon and creating datasets of labeled pairs of words such as Hyperlex (Vulić *et al.* 2017). Other works are concerned with the entailment relationship between two words in their respective contexts. Pointing out that *eliminate* entails *treat* in *Aspirin eliminates headaches* but not in *Aspirin eliminates patients*, Levy and Dagan (2016) introduce a dataset of annotated relation pairs. This dataset uses question–answer pairs as context for lexical entailment, other approaches involve providing context as pairs of arguments (Zeichner, Berant, and Dagan 2012) or pairs of argument types (Berant, Dagan, and Goldberger 2011; Schmitt and Schütze 2019). Lexical entailment can also be viewed as a special case of natural language inference (NLI), modern systems for this task are commonly trained and evaluated on the Stanford Natural Language Inference (SNLI) (Bowman *et al.* 2015) dataset and the Multi-Genre NLI Corpus (MultiNLI) (Williams, Nangia, and Bowman 2018) dataset.

The approach taken in this article, outlined in Section 3.2, is based on the `4lang` formalism for representing (lexical) semantics and is capable of inspecting the relationship between the meaning of any two utterance fragments. We will evaluate our system on two recent benchmarks. The datasets used in the 2020 Semeval task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment" (Glavaš *et al.* 2020) are derived from HyperLex (Vulić *et al.* 2017), a dataset of monolingual and cross-lingual–graded lexical entailment. Candidate word pairs for human annotation were gathered from the USF (Nelson, McEvoy, and Schreiber 2004) and WordNet (Miller 1995) databases. For our experiments, we will use the binary labels of the monolingual subsets for English, German, and Italian. Some examples for each language are shown in Table 1. On

[a] https://github.com/adaamko/wikt2def/tree/nle.
[b] https://github.com/adaamko/wikt2def/tree/nle_semeval.
[c] https://github.com/recski/tuw-nlp.

**Table 1.** Example entries of the monolingual binary portion of the Semeval lexical entailment dataset (Glavaš *et al*. 2020)

| Premise | | Hypothesis | | Label |
|---|---|---|---|---|
| *sandwich* | | *food* | | True |
| *morning* | | *intelligence* | | False |
| *Sohn* | 'son' | *Kind* | 'child' | True |
| *Küche* | 'kitchen' | *Schlafzimmer* | 'bedroom' | False |
| *ciliegia* | 'cherry' | *frutto* | 'fruit' | True |
| *formaggio* | 'cheese' | *burro* | 'butter' | False |

**Table 2.** Example entries of the SherLlic dataset (Schmitt and Schütze 2019). Argument labels indicate entity types: PER – person, LOC – location, ORGF – organization_founder, EMPL – employer, AUTH – book_author

| Premise | Hypothesis | Label |
|---|---|---|
| ORGF[A] is granting to EMPL[B] | ORGF[A] is giving to EMPL[B] | True |
| PER[A] is interviewing AUTH[B] | PER[A] is asking AUTH[B] | True |
| LOC[A] is fighting with ORGF[B] | LOC[A] is allied with ORGF[B] | False |
| LOC[A] is city of LOC[B] | LOC[A] is capital of LOC[B] | False |

this dataset, we will compare our method to the GLEN system for measuring multilingual and cross-lingual lexical entailment using specialized word embeddings (Vulić *et al*. 2019), which outperforms previous baselines in Upadhyay *et al*. (2018), and also to the other systems participating in the shared task, including our own earlier system presented in Kovács *et al*. (2020).

A more challenging task formulation is provided by the SherLlic dataset of lexical inference in context (Schmitt and Schütze 2019), which was built by extracting inference candidates from an entity-linked portion of the ClueWeb corpus (Gabrilovich, Ringgaard, and Subramanya 2013) and using them as input to human annotation. Because annotation candidates were chosen based on distributional evidence, many entailment pairs in the final dataset are completely novel, missing from existing knowledge bases such as WordNet. Argument types for event pairs are necessary to disambiguate between word senses. For example, *run* entails *lead* if its arguments are of type PERSON and COMPANY (e.g., *Bezos runs Amazon*) but not if they are COMPUTER and SOFTWARE, as in *my mac runs macOS*. Table 2 shows further examples of entries in the SherLlic dataset. An extensive evaluation of various LE systems on this dataset presented in Schmitt and Schütze (2019) will serve as the starting point for our evaluations in Section 4.

The need for such novel datasets has been made clear by several recent experiments that point out the biases of deep learning based models of NLI. Glockner, Shwartz, and Goldberg (2018) constructed a new NLI test set from SNLI by replacing a single word in sentences from the training set, and used this new benchmark to expose top NLI systems' inability to perform true lexical inference. The only model in their evaluation not showing a major drop in performance was the one incorporating lexical knowledge (Chen *et al*. 2018). The inability of deep learning based NLI models to generalize across datasets was also shown in a more recent study across six systems and three datasets (Talman and Chatzikyriakidis 2019). These findings call into question whether black box models trained for high performance on any single NLI benchmark can be regarded as true models of inference. We believe that rule-based models such as the one proposed in this paper

can facilitate further qualitative analysis of deep learning models by providing strong explainable baselines on multiple tasks and datasets.

### 2.2 Approaches to entailment and inference

When seen as a task of detecting hypernymy, lexical entailment is most often addressed using distributional methods. Hypernymy candidates are encoded using word embeddings and classified by either neural networks (Nguyen *et al*. 2017; Shwartz, Goldberg, and Dagan 2016; Yu *et al*. 2015) or non-neural classifiers such as Support Vector Machines (SVMs) and logistic regression (Baroni *et al*. 2012; Levy *et al*. 2015; Roller, Erk, and Boleda 2014). Yu *et al*. (2015) proposes a neural model for supervised learning of hypernymy-specific embeddings. Nguyen *et al*. (2017) argues that standard distributional models cannot account for the asymmetric property of hypernymy, and introduces *HyperVec*, a hierarchical approach to learning hypernymy embeddings that allowed for significant improvement over the state-of-the-art on the HyperLex dataset. Glavaš and Ponzetto (2017) proposes Dual Tensor, an approach based on neural models to explicitly model the asymmetric nature of the hypernymy relation. Dual Tensor transforms generic embeddings into specialized vectors for scoring concept pairs based on whether the asymmetric relation holds. A different approach is taken by HypeNET (Shwartz *et al*. 2016), a method based on extracting paths between premise and hypothesis from dependency trees and using them as inputs to Long Short-term Memory Networks (LSTMs). Fine-tuning generic word vectors using external knowledge such as WordNet (Miller 1995) has improved performance on a range of language understanding tasks (Glavaš and Vulić 2018). To extend this method to unseen words, Kamath *et al*. (2019) introduced `POSTLE` (post-specialization for LE), a model that learns an explicit global specialization function captured with feed forward neural networks.

Inference systems trained on the SNLI and MultiNLI datasets mostly use neural language models based on the Transformer architecture (Vaswani *et al*. 2017), in particular BERT (Devlin *et al*. 2019). SemBERT (Zhang *et al*. 2020) uses a BERT backbone enhanced with a Semantic Role Labeler (SRL), MT-DNN (Liu *et al*. 2019) enhances the system presented in Liu *et al*. (2015) with BERT. Top results on the MultiNLI benchmark were achieved by optimized, pretrained, and fine-tuned versions of BERT, RoBERTa (Zhuang *et al*. 2021), and ALBERT (Lan *et al*. 2020). Using rule-based models in NLI and combining them with deep learning based language models (BERT, ALBERT, RoBERTa) has recently also led to competitive results (Haruta, Mineshima, and Bekki 2020; Kalouli, Crouch, and de Paiva 2020). In this article, we will also present an improvement over such a model using our fully rule-based method for detecting entailment between pairs of words.

### 2.3 Semantic parsing

Semantic parsing is the task of mapping natural language text to some model of its meaning, and as a language processing step it can only be defined with respect to a particular system of semantic representation. As interest in symbolic representations has increased, so has the variety of such systems. Most practical frameworks link to lexical databases, including the graph of synonym sets in WordNet (Miller 1995), the semantic role inventory in VerbNet (Kipper *et al*. 2008), or the ontology of semantic frames in FrameNet (Ruppenhofer *et al*. 2006). Abstract meaning representations (AMR) (Banarescu *et al*. 2013), which became one of the most widely used representations in semantic parsing, is based on the PropBank database (Palmer, Gildea, and Kingsbury 2005). AMR handles a range of phenomena in the semantics of English but does not provide any treatment of word meaning and is also not intended as a universal representation of natural language semantics, despite some early efforts in Xue *et al*. (2014) and the release of a Chinese AMR bank (Li *et al*. 2016). Automatic construction of AMR graphs from text is usually performed using deep neural models (Konstas *et al*. 2017; Lyu and Titov 2018; Zhang *et al*. 2019) trained on AMR banks. A language-agnostic approach to meaning representation is taken by Universal Conceptual

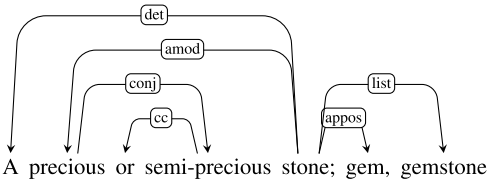


**Figure 1.** Dependency parse of the definition of `jewel`.

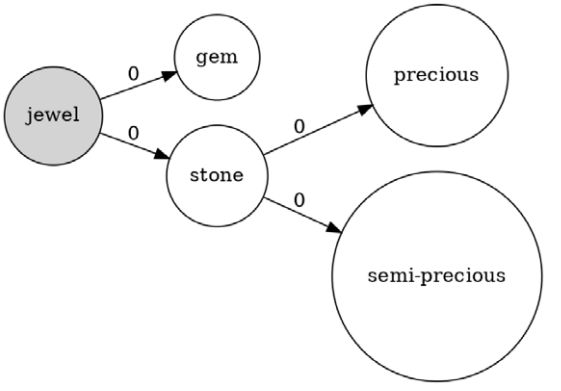


**Figure 2.** 4lang definition graph of `jewel`.

Cognitive Annotation (UCCA) (Abend and Rappoport 2013), which abstracts away from syntax by representing words of a sentence as leaf nodes of directed acyclic graphs (DAGs) representing *scenes* evoked by predicates. As in the case of AMRs, top-performing parsers for UCCA all employ neural networks trained on manually annotated sentences (Hershcovich, Abend, and Rappoport 2017, 2018; Ozaki *et al.* 2020; Samuel and Straka 2020). Not only are these formalisms often dependent on large manually built databases, existing parsing systems also rely on large datasets of hand-crafted representations (sembanks) for training. Transferring such systems across domains and languages therefore requires a considerable amount of expert human labor. In this article, we will use a meaning representation formalism designed to enable robust rule-based parsing and relying only on language-agnostic resources. Here we provide an overview only, our extensions and modifications shall be presented in Section 3.

`4lang` is a theory and formalism for representing the semantics of natural language, developed in Kornai (2012, 2019), Kornai *et al.* (2015), and partially implemented in Recski (2016, 2018). `4lang` concept graphs represent meaning as directed graphs of language-independent concepts. Edges connecting concepts have one of three labels. Predicates are connected to their arguments via edges labeled 1 and 2, for example, `cat` $\overset{1}{\leftarrow}$ `catch` $\overset{2}{\rightarrow}$ `mouse`, while 0-edges represent all relationships involving inheritance, including not only hypernymy (`dog` $\overset{0}{\rightarrow}$ `mammal`) but also attribution (`dog` $\overset{0}{\rightarrow}$ `four-legged`), and unary predication (`dog` $\overset{0}{\rightarrow}$ `bark`). This implies a broad definition of lexical entailment: unless explicitly overridden, *dog* entails not only *mammal*, but also *bark* and *four-legged*. In its current implementations, `4lang` concepts have no grammatical attributes and no event structure, for example, the phrases *water freezes* and *frozen water* would both be represented as *water* $\overset{0}{\rightarrow}$ *freeze*. Figure 2 shows the `4lang` definition of the concept `jewel` obtained by processing the dependency parse in Figure 1 of the Wiktionary definition *A precious or semi-precious stone; gem, gemstone.*

Optionally, the `4lang` system allows us to *expand* graphs, a process which unifies one graph with the definition graphs of the concepts within that graph. For example, a graph containing the

node `jewel` will be expanded to include as a subgraph the entire definition graph in Figure 2. This step will be essential to our method presented in Section 3.2. `4lang` graphs can be built automatically from Universal Dependencies (Nivre *et al.* 2018) using the rule-based `dep_to_4lang` module, which we extend to improve performance across languages. Section 3 will describe these changes, as well as our reimplementation of the parsing algorithm using Interpreted Regular Tree Grammars (IRTGs) (Koller 2015). While in the present work this method is used to map natural language definitions to concept graphs representing the meaning of individual words, the system is capable of processing any UD graph and can be used to construct the `4lang` semantic representation of any text.

### *2.4 Entailment in semantic graphs*

Lexical entailment is explicitly encoded by several semantic formalisms. The hyponymy/ hypernymy relation, the narrow interpretation of lexical entailment used for example in the Semeval 2020 shared task, is represented directly by WordNet. The 0-edge in `4lang` graphs is a more generic relation that subsumes the hypernymy relation along with all other types of predication, and accessibility of one concept from another in a `4lang` graph can be seen as a broad definition of lexical entailment. Recently, we have shown (Kovács *et al.* 2020) the direct applicability of such semantic graphs to hypernymy detection task by using them in a competitive system at the Semeval entailment task (Glavaš *et al.* 2020). When using `4lang` definition graphs, we defined entailment to hold between a pair of premise and hypothesis words if and only if in the twice-expanded definition graph of the premise there is a directed path of 0-edges leading from the premise word to the hypothesis word (the maximum number of expansions was chosen arbitrarily). Because entailment does not flow through locative and negative modifier clauses, inference had to be blocked explicitly where the path would go through prepositions (e.g. English *in, of, on*, German *in, auf*, Italian *di, su, il*) or words conveying negation (English *not*, German *keine*, etc.). For example, where *nose* is defined as "a protuberance on the face", 4lang graphs would contain a path of 0-edges from *nose* to *face*, falsely representing entailment. One of the modifications of the parsing algorithm to be presented in this article will ensure that subgraphs representing such relations do not contain 0-paths.

On the Semeval dataset of word-level entailment, the above method detected only about one-third of all true entailments in the `dev` dataset but achieved nearly perfect precision. On well-resourced languages such as English, WordNet was shown to be a very strong baseline both in terms of precision and recall, and the main contribution of `4lang` was its ability to increase recall without hurting precision, increasing the performance of strong WordNet-based baselines on three languages, ranking first in English and Italian and second-best on German, see Table 3. For English and Italian official WordNet releases were accessed via the `nltk`[d] package. In Kovács *et al.* (2020) for German, we did not have access to a high-coverage WordNet release, word pairs were therefore translated from German to English using the wikt2dict system (ács, Pajkossy, and Kornai 2013) to enable the use of English WordNet on the German task. In Section 4, when evaluating the methods presented in this article and comparing them to previous methods, we include a variant of this system using a recent German WordNet release, GermaNet (Hamp and Feldweg 1997; Henrich and Hinrichs 2010).

## 3. Semantic parsing and representation

Our method for establishing entailment between pairs of words requires that we create `4lang` semantic graphs for each word. Using a modified version of the pipeline described in Recski

[d] https://www.nltk.org/howto/wordwet.html.

**Table 3.** Official monolingual LE results on the `ANY` track (F-scores)

| System | en | de | it |
|---|---|---|---|
| GLEN baseline (Glavaš and Vulić 2019) | 79.87 | 59.88 | 66.27 |
| **BMEAUT** (Kovács *et al.* 2020) | **91.77** | 67.00 | **81.41** |
| SHIKEBLCU (Wang and Kuo 2020) | 87.90 | **71.43** | 75.94 |

Bold values represents best scores.

```
NOUN -> NOUN_AMOD_ADJ(NOUN, ADJ)
[ud] f_dep1(merge(merge(?1,"(r<root> :AMOD (d1<dep1>))"),
    r_dep1(?2)))
[4lang] f_dep1(merge(merge(?1,"(r<root> :0 (d1<dep1>))"),
     r_dep1(?2)))
```

**Figure 3.** Example of an IRTG rule.

(2016), we process dictionary definitions with Universal Dependency (UD) (Nivre *et al.* 2018) using the `stanza` library (Qi *et al.* 2020) and transform the resulting dependency trees using rules from the `dep_to_4lang` (Recski 2018) module. In this section, we present our modified version of this pipeline, for which we reimplement the transformation of dependency trees into `4lang` graphs using Interpreted Regular Tree Grammars (IRTGs) (Koller 2015) generated dynamically using a lexicon of rule templates. Then we present our method for detecting entailment over pairs of semantic graphs corresponding to premise and hypothesis.

### *3.1 Semantic parsing with Interpreted Regular Tree Grammars*

Many NLP tasks involve constructing or transforming graphs that represent syntax and/or semantics. Interpreted Regular Tree Grammars (IRTGs) (Koller 2015) can be used to encode the correspondence between sets of such structures and have in recent years been used to perform syntactic parsing (Koller and Kuhlmann 2012), generation (Koller and Engonopoulos 2017), semantic parsing (Groschwitz, Koller, and Teichmann 2015; Groschwitz *et al.* 2018), and surface realization (Kovács *et al.* 2019; Recski *et al.* 2020). The system presented in this article uses an IRTG for transforming UD trees into `4lang` graphs.

An IRTG rule is a (possibly weighted) rewrite rule of a Regular Tree Grammar that is mapped to an arbitrary number of named *interpretations*, each of which are operations of an algebra with the same arity as the RTG rule. Thereby any particular derivation of an IRTG grammar deterministically maps to a sequence of operations in each of the interpretation algebras. Parsing an object of one algebra involves finding the IRTG derivation(s) of the highest likelihood that would generate this object, while *decoding* is the subsequent construction of an object in another algebra using this sequence. The grammars used by our system establish a mapping between operations of two algebras of directed graphs, one for constructing UD representations and another for constructing `4lang` graphs. The overall structure of an IRTG rule with two interpretations is shown in Figure 3. The graph operations used in this example will now be introduced.

Following the practice of Koller and Kuhlmann (2011), we use *s-graph* algebras. We give an informal overview of its operations, see Courcelle and Engelfriet (2012) for a more formal explanation. S-graphs are graphs whose vertices may be labeled by one of a countable set of *sources*, which are essentially special node labels accessible by operations of the algebra. The binary `merge` operation creates an s-graph by taking the union of its argument graphs and merging nodes with identical sources. In other words, when two s-graphs $G_1$ and $G_2$ are `merged`, the resulting s-graph G' will contain all nodes of $G_1$ and $G_2$, and when a pair of nodes $(v_1, v_2) \in V(G_1) \times V(G_2)$ have

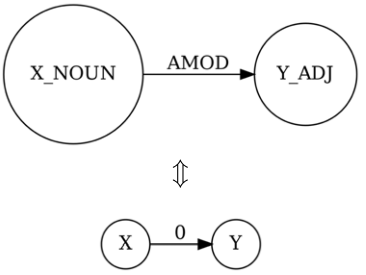


**Figure 4.** Graph representation of sample rule in Figure 3.

the same source name, they will be mapped to a single node *v*' in *G*' that has all adjacent edges of $v_1$ and $v_2$. Sources can be manipulated by the *rename* and *forget* operations for changing or deleting a given source label from all nodes of an s-graph.

We illustrate the algebra operations and the structure of an IRTG using a simple example. Figure 3 shows an IRTG rule encoding the correspondence between two operations: one that adds a directed `amod` edge between the root nodes of two UD graphs, and another that adds a directed `0` edge between the root nodes of two `4lang` graphs. The first line in the example is the IRTG rule, specifying that an operation called `NOUN_AMOD_ADJ` takes as its arguments two objects of type `NOUN` and `ADJ`. The second and third lines are the interpretations, each specifying the same sequence of operations: the first argument is merged with a graph consisting of a single directed edge, the second argument is merged with the resulting graph after its `root` source has been changed to `dep1` using the `rename` operation. Finally, the `dep1` source is deleted, using the `forget` operation. This sequence of operations is equivalent to adding a single directed edge between the root nodes of the two argument graphs, ensuring that the root of the first becomes the root of the new graph. The syntax used in this example is specified by the Algebraic Language Toolkit, or `alto`[e] (Gontrum *et al.* 2017), an open-source parser for IRTGs that implements a variety of algebras, including the s-graph algebras used in this article. Graph literals in each interpretation are given using the PENMAN notation,[f] in this simple case the only difference between the two strings is the edge label. The correspondence expressed by the rule in Figure 3 can also be represented by the pair of graph templates in Figure 4, we will use this simplified format in future examples.

Parsing UD graphs and transforming them into `4lang` graphs on a large scale would be possible using a single grammar with terminal rules corresponding to each word of the input graph. But since building and continuously extending such a large set of rules would be inefficient, we instead chose to dynamically generate individual grammars for each input UD graph, a process that makes it possible to use sets of rule templates for generating similar rules, and to organize them into configurable, application-specific rule lexica. We then construct IRTG grammars for individual UD trees by looking up their edges in such lexica, for example, the UD edge `NOUN` $\xrightarrow{amod}$ `ADJ` will always warrant an IRTG rule that maps this edge to the `4lang` edge `NOUN` $\xrightarrow{0}$ `ADJ`. Other patterns require the lexicon to reference additional nodes in the input graph, we discuss some examples below. Terminal rules that map POS-tags to words in both interpretations are added to each grammar in a trivial step. The grammar generation framework described here is available as open-source software as part of the `tuw-nlp`[g] Python package, and is a core dependency of the system presented in this article, which is also available as part of an open-source library.[h]

[e] https://github.com/coli-saar/alto.
[f] https://penman.readthedocs.io/en/latest/notation.html.
[g] https://github.com/recski/tuw-nlp.
[h] https://github.com/adaamko/wikt2def/tree/nle.

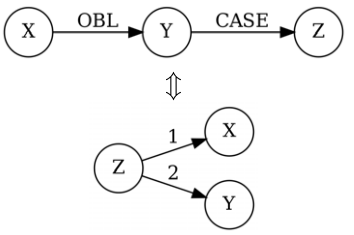


**Figure 5.** Obliques in UD and `4lang`.

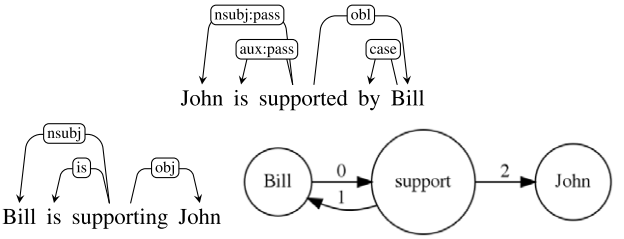


**Figure 6.** 4lang definition graph of *John is supported by Bill* and *Bill is supporting John*.

The starting point for creating rule lexica was the mapping of the original `dep_to_4lang` paper (Recski 2018), which establishes all trivial 1-to-1 mappings between pairs of UD and `4lang` relations. For example, all UD relations representing modification (`amod, advmod, nummod`) are mapped to 0-edges, while relations between predicates and their objects (`obj, nsubj:pass`) become 2-edges. Subjects relations (`nsubj, csubj`) are mapped to a pair of 0- and 1-edges; in the sentence *John is supporting Bill*, the UD relation *support* $\xrightarrow{nsubj}$ *John* becomes `support` $\overset{1}{\underset{0}{\rightleftharpoons}}$ `John`. Clausal modifiers (`acl, advcl`) are generally also mapped to 0-edges, some newly introduced exceptions will be discussed below. Additionally, we introduce a new mechanism for non-core (oblique) arguments marked by the `obl` relation.

Consider the Wiktionary definition of *teacher: someone who teaches, especially in a school*. The UD edge `teach` $\xrightarrow{obl}$ `school` does not in itself reveal the semantic relationship between the two concepts, thus we introduce a pattern that will also take into account the `case` relation marking the argument: the full UD analysis of this sentence contains the subgraph `teach` $\xrightarrow{obl}$ `school` $\xrightarrow{case}$ `in`, allowing us to build the `4lang` graph `teach` $\xleftarrow{1}$ `in` $\xrightarrow{2}$ `school` using an IRTG rule represented in Figure 5. We also implemented an English-specific exception to this rule: the preposition by will trigger the configuration for the predicate–subject relation, so that the UD analyses of the sentences *John is supported by Bill* and *Bill is supporting John* shall both be mapped to the same `4lang` graph (see Figure 6).

Another shortcoming of the original algorithm for mapping dependency graphs to `4lang` representations is its treatment of coordination. The strategy introduced in Section 3.4.1 of Recski (2018) simply copied all semantic relations between all elements of coordinating constructions, which has proved practical for downstream applications despite introducing some erroneous edges. Our system replicates this behavior in its mapping from UD to `4lang`. Some simple patterns over specific conjunctions (*and*, *or*, etc.) could be used to differentiate between occurrences of the `conj` dependency, similar to the approach of Enhanced Universal Dependencies (Schuster and Manning 2016), but modeling the semantics of coordinating conjunctions would nevertheless require considerable language-specific effort (see Gerdes and Kahane 2015; Kanayama *et al.* 2018).

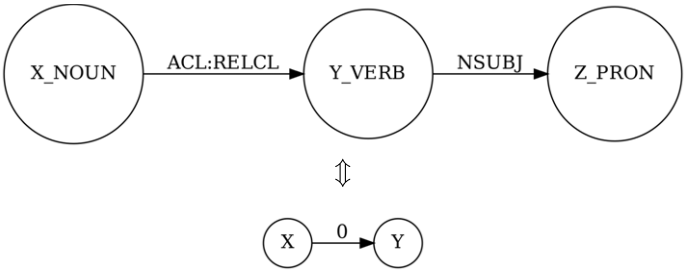


**Figure 7.** Relative clause modifier of a noun in UD and 4lang.

Perhaps the most significant limitation of the original system was its lack of treatment for relationships between clauses. This is partly due to the fact that the type of the UD relation connecting the heads of two clauses often reveals very little about the semantic function of the dependent clause. The UD relation `acl` is defined[i] as *clausal modifier of noun (adjectival clause)*, and indeed the general case can be handled by 0-edges, but this rule will currently create erroneous edges for sentences such as *I have a parakeet named Cookie*, since there is no mechanism to detect that in this case *parakeet* is the object of *named*. Another issue is posed by the UD relation `acl:relcl` (relative clause modifier of a noun), which is used both in *an animal that moves* and *the man you love*, constructions that would warrant the edges `animal` $\xrightarrow{0}$ `move` and `man` $\xleftarrow{2}$ `love`, respectively. The first of these two examples can receive the correct treatment based on the presence of the `nsubj` edge between *move* and *that*, thus we introduce a rule that implements the correspondence in Figure 7. Other occurrences of `acl:relcl` are currently not processed. For the full mapping between UD and 4lang structures see Table 4.

While some of these newly introduced mechanisms are still rudimentary, improving the general mechanisms of building task-independent semantic representations is a central goal of our work—we present our error analysis in Section 5. The pipeline for building `4lang` definition graphs from Wiktionary is currently implemented for three languages: English, German, and Italian. Extending it to additional languages requires the existence of an accurate UD parser and a machine-readable monolingual dictionary of sufficient coverage. Once the UD parses of definitions are available, the remainder of the pipeline is language-independent, although for some languages it might be necessary to extend the `dep_to_4lang` mapping to include reference to morphological features, as done in Recski, Borbély, and Bolevácz (2016). We apply this pipeline to dictionary definitions extracted from data dumps of Wiktionary, a large crowd-sourced dictionary containing more than 100,000 entries for 40+ languages (and more than 10,000 entries for about twice as many). Section 3.2 will describe our method for detecting entailment between pairs of words or predicates using their corresponding `4lang` graphs. Entailment relations extracted using this method may be of lower quality than those encoded by manually built databases such as WordNet, but cover larger vocabularies and thus improve performance even for languages with large WordNets (see Section 4 for details). While in our current experiments we chose to build pipelines for three relatively well-resourced languages (English, German, Italian), both a large Wiktionary and an UD parser model are available for many more.

### *3.2 Modeling lexical entailment*

A Wiktionary page for a given word form typically contains several definitions corresponding to multiple word senses and/or parts-of-speech, but, given the crowd-sourced nature of the dataset, without adhering to any particular lexicographic principles. After identifying individual entries

[i] https://universaldependencies.org/u/dep/all.html.

**Table 4.** Mapping from UD relations to 4lang subgraphs

| Dependency | Edge |
|---|---|
| advcl<br>advmod<br>amod<br>nmod<br>nummod<br>obl:npmod | $w_1 \xrightarrow{0} w_2$ |
| nsubj<br>csubj | $w_1 \underset{0}{\overset{1}{\rightleftharpoons}} w_2$ |
| obj<br>ccomp<br>xcomp | $w_1 \xrightarrow{2} w_2$ |
| appos | $w_1 \underset{0}{\overset{0}{\rightleftharpoons}} w_2$ |
| nmod:poss | $w_2 \xleftarrow{1} \texttt{HAS} \xrightarrow{2} w_1$ |
| nmod:npmod | $w_1 \xleftarrow{1} \texttt{NPMOD} \xrightarrow{2} w_2$ |
| nmod:tmod<br>obl:tmod | $w_1 \xleftarrow{1} \texttt{AT} \xrightarrow{2} w_2$ |
| $w_1 \xrightarrow{\text{obl}} w_2 \xrightarrow{\text{case}} w_3$ | $w_1 \xleftarrow{1} w_3 \xrightarrow{2} w_2$ |
| $w_1 \xrightarrow{\text{acl:relcl}} w_2 \xrightarrow{\text{nsubj}} w_3$ | $w_1 \xrightarrow{0} w_2$ |

on each Wiktionary page using simple language-specific templates, our approach is to pick the first definition of each word unless it is explicitly marked by editors as `obsolete`, `archaic`, `historical`, or `rare`. Based on manual analysis of a small sample, we estimate that for over 98% of words in the Semeval dataset this method chooses the sense that is apparently intended in the dataset. For example, for the word pair *letter–mail* in the Semeval dataset, we assume that the intended sense is that defined as *a written or printed communication* and not *a symbol in an alphabet*.

The entailment candidates in the SherLliC dataset pose a greater challenge. Predicates are implicitly disambiguated by their type signatures, yet we do not attempt to select some subset of the available definitions based on this information, instead we establish entailment between a pair of predicates iff there is any pair of definitions for which the conditions of entailment, to be defined in this section, are fulfilled. In Section 4, we shall see that this does not lead to a drastic decrease of precision as it would on the Semeval dataset, likely because of the relatively higher bar of matching an argument-predicate structure as opposed to a single word. Meanwhile this enables detecting entailment based on any listed definition of predicates; for example, we correctly detect that the premise *"A is releasing update for B"* entails the hypothesis *"A is releasing version of B"* based on the third Wiktionary definition of *update*: *"A modification of something to a more recent,*

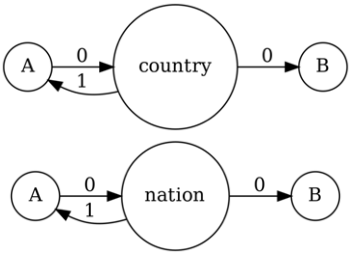


**Figure 8.** 4lang representations of *A is nation in B* and *A is country in B*.

*up-to-date version; (in software) a minor upgrade"*. We shall discuss the issue of polysemy in more detail in Section 5.

Our core method for modeling entailment is based on the intuition that all semantic relationships marked by 0-edges in `4lang` graphs constitute entailment, that is *dog* entails not only *mammal* but also *four-legged* and *bark*, and that this relation is transitive, that is if *dog* entails *bark* and *bark* entails *sound* then *dog* must also entail *sound*. In other words, we equate entailment with inheritance and consequently with accessibility via a directed path of 0-edges in a `4lang` concept graph. Here, we shall not discuss whether the relationship between for example the concepts *dog* and *bark* are an example of entailment, causation, correlation, or something else; this comprehensive definition is rooted in our view that modeling lexical entailment should be an enabler of the modeling of natural language inference (NLI). We note that this view on entailment may be incompatible with task formulations involving statements about specific events or scenes, such as the SNLI dataset that is largely based on image captions, since in such descriptions *A dog is there* should not entail *A dog is barking*. It can be argued, however, that it is precisely the omission of such type-theoretic details that gives the `4lang` system its flexibility. Where there is a dog, there can be a bark(ing). Kornai (2010) further argues that the episodic readings that have occupied Montague Grammarians ever since the inception of MG are practically nonexistent in natural language.

Implementing the above definition as a function over pairs of semantic graphs involves the recursive expansion of the premise graph based on the definitions of its defining words. This, after only a small number of iterations, leads to the proliferation of errors and ambiguities, caused primarily by imperfections of the `dep_to_4lang` mapping and the inherent ambiguity of definitions (see Section 5 for more details). We found it practical in our experiments to limit the depth of expansion to 2 (although in certain cases 3 would lead to a higher F-score, see Section 4 for details). On the Semeval dataset of word pairs, we then establish entailment iff the hypothesis word is present in the expanded premise graph. The depth of recursive expansion is the major tunable parameter of our method, set to 2 for both tasks based on early experimental results, but recall can be increased considerably by allowing our system to establish entailment without full coverage, requiring only some percentage of edges in the hypothesis graph to be covered by the premise graph. Optimal values for this threshold were obtained by optimizing on the development portion of the dataset. For the high-precision configurations, we set the threshold to 0.8 while for the highest F-score required a value of 0.2. Such tweaking admittedly weakens full explainability, since one has to justify a partial overlap of premise and hypothesis graphs. Still, the interpretability of our representation remains, enabling deep error analysis (see Section 5) and further development of our semantic parsing methods.

On the SherLlic task, we use `4lang` graphs built from the example sentences associated with each predicate in the dataset. For example, the pair of premise and hypothesis predicates *"A is nation in B"* and *"A is country in B"* will be represented by the graphs in Figure 8. We define entailment to hold iff all edges of the hypothesis graph are found in the expanded premise graph, which requires us to refine the expansion process to also allows nodes to inherit relations along

```
For each pair of nodes (v, w) in graph G:
        IF there is a directed path of 0-edges from v to w:
                copy all edges of v to w
```

**Figure 9.** Appending zero edges to premise.

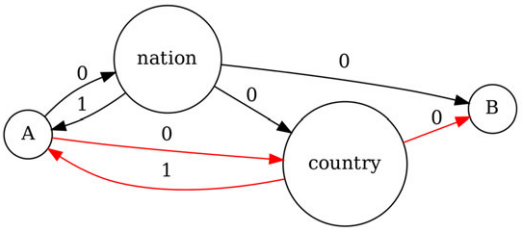


**Figure 10.** Expanded `4lang` representations of *A is nation in B*.

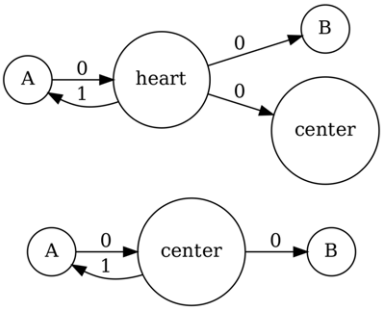


**Figure 11.** 4lang definition graphs of expanding and reducing *A is heart of B*.

paths of 0-edges. After a graph has been extended with the definition graphs of its nodes, we allow relations to be inherited along directed paths of 0-edges using the algorithm in Figure 9 and exemplified in Figure 10, representing the *modus ponens* reasoning that if A is a nation in B and nation is a type of country, then A is a country in B.

Let us now consider the premise *"A is center of B"* and the hypothesis *"A is heart of B"*. The concept `heart` is not accessible from the premise, but one of its definitions is *The centre*. In this case, we can detect entailment via the *reduction* mechanism, which substitutes a concept with its definition graph (while still allowing relations to be inherited by defining concepts, as in the case of expansion). In other words, reduction is equivalent to expansion followed by removal of the original concept, as shown in Figure 11. This mechanism represents the intuition that if all defining properties of some concept can be inferred then the concept itself is also entailed.

On the simpler word-level task, we increase our overall performance by incorporating additional lexical resources: we extract for all premise words lists of synonyms from their Wiktionary pages as well as a list of synonyms and all their hypernyms from WordNet (Miller 1995). We use such additional resources in two ways. First, we can choose to establish entailment if the hypothesis word is present in the set of hypernyms for any WordNet synset containing the premise word. Second, we may use synonyms of the premise word from both WordNet synsets and Wiktionary to extend the `4lang` definition graph of the premise with additional concepts before performing expansion. In Section 4, we shall quantify the contribution of these additions. On the SherLlic dataset, we decided not to use these extensions as they appeared to introduce too much noise.

Our method for detecting entailment described in this section produces decisions that are directly explainable due to the presence of at least one path of 0-edges leading from the premise word to the hypothesis word, and these paths can be further explained by the dictionary definitions on which they were based. For example, the decision in the above example that *"A is center of B"* entails "A is heart of B" could be explained by citing the original Wiktionary definition *heart: 12. (figuratively) The centre, essence, or core*. Although we have not observed it during our manual error analysis (see Section 5), it is possible for our method to make a correct decision based on edges that were introduced in error, which would ultimately result in an incorrect explanation.

Our approach treats lexical entailment detection as a binary classification task and in this work we shall not attempt to give an account of entailment as a graded phenomenon, even though the Semeval dataset (see Section 2.1) provides evaluation data also for graded lexical entailment. While this dataset is validated by high agreement among annotators answering the question *To what degree is X a type of Y* and represents common intuitions such as that *chess* is a *sport* to a lesser degree than *basketball* (Vulić *et al*. 2017), we consider such distinctions more an issue of prototypicality than of implicature as such. Basketball may be uniformly viewed as more of a sport than chess, but the implications we want to employ, for example that it is a contest, a fight between participants with a winner and a loser, and that the winner is proven stronger by this outcome in the sport at hand, are the same, and they operate in a 0-1, rather than a gradient, fashion.

## 4. Evaluation

In this section, we present the results of experiments on both the Semeval and SherLlic datasets. On the Semeval dataset, we compare our system to the previous state-of-the-art and some strong baselines, while on the SherLlic task we use for comparison both the set of baselines published by Schmitt and Schütze (2019) and a recent SOTA NLI system. We also evaluate the effect of some of the processing steps discussed in Section 3 by comparing various configurations of our method on both benchmarks.

For the word-level entailment task, we evaluate on the English, German, and Italian portions of the Semeval 2020 dataset. We used the development portion of each dataset for experimentation and for determining the optimal value of the single tunable parameter of our system, the depth of recursion when expanding `4lang` graphs. We set this value to 2 in all configurations and on both datasets. It was also based on these experiments that we made the determination to limit expansion to nodes connected to each word by 0-paths, as opposed to expanding all nodes in the definition graphs, as we have done for the SherLlic task.

Performance of various configurations on the Semeval development data is presented in Table 5. Our core method, using expanded `4lang` definition graphs built from Wiktionary definitions, achieves high precision on all three languages and recall values in the 0.25–0.35 range. Extending our definition graphs with additional synonyms from WordNet and Wiktionary improves recall at the cost of some precision, on all languages. On the German data, we also compare the contribution of the translation-based method `EN_WordNet` that uses English Wordnet to the high-quality German WordNet release GermaNet (Hamp and Feldweg 1997; Henrich and Hinrichs 2010). Since WordNet graphs explicitly encode the hypernymy–hyponymy relationship between synsets, they can be evaluated as standalone baselines and achieve strong precision and recall scores on all languages. Our method, however, can be used to improve their recall further, thus the top-performing system for all three languages is that which labels word pairs as entailment if either our system or the `WordNet` baseline labeled it as such. On the English dataset, we also illustrate the negative effect of including all Wiktionary definitions. On the SherLlic task the effect is reversed, using multiple definitions increases performance (`4lang_multidef`). In Table 6, we list some examples of entailment pairs that have been detected by our method but not by WordNet, along with their Wiktionary definition of the premise that was used for building `4lang` representations.

**Table 5.** Performance on the Semeval development set. `4lang` and `4lang_syn` is our method without and with additional synonym nodes from WordNet and Wiktionary. `WordNet` is the baseline using WordNet hypernyms, `4lang_syn+WordNet` is the union of `4lang+syn` and `WordNet`

| Lang | Method | P | R | F |
|---|---|---|---|---|
| EN | always yes | 56.33 | 100.0 | 72.07 |
| | WordNet | 95.75 | 88.76 | 92.12 |
| | `4lang` | **96.30** | 29.21 | 44.83 |
| | `4lang_multidef` | 81.16 | 62.92 | 70.89 |
| | `4lang_syn` | 92.85 | 36.51 | 52.41 |
| | `4lang_syn+WordNet` | 93.22 | 92.69 | **92.95** |
| | RoBERTa | 83.25 | **94.94** | 88.71 |
| DE | always yes | 37.11 | 100.0 | 54.13 |
| | EN_WordNet | 61.61 | 79.22 | 69.31 |
| | GermaNet | 90.83 | 70.77 | 79.56 |
| | `4lang` | **88.88** | 36.36 | 51.61 |
| | `4lang_syn` | 87.87 | 37.66 | 52.72 |
| | `4lang_syn+EN_WordNet` | 61.86 | **86.36** | 72.08 |
| | `4lang_syn+WordNet` | 87.23 | 79.87 | **83.33** |
| IT | always yes | 41.67 | 100.0 | 58.82 |
| | WordNet | 88.96 | 75.88 | 81.90 |
| | `4lang` | **93.47** | 25.29 | 39.81 |
| | `4lang_syn` | 81.17 | 40.58 | 54.11 |
| | `4lang_syn+WordNet` | 83.92 | **82.94** | **83.43** |

Bold values represents best scores.

**Table 6.** Examples of entailment pairs not in WordNet but detected by our system

| Premise | Hypothesis | Premise definition |
|---|---|---|
| *graph* | *chart* | *a data chart (graphical representation of data) intended to illustrate the relationship between a set (or sets) of numbers* |
| *Saturn* | *Planet* | *sechster und zweitgröter Planet unseres Sonnensystem* |
| | | 'sixth and second-largest planet of our solar system' |
| *test* | *esame* | *esame per verificare qualcosa* |
| | | 'exam to check something' |

**Table 7.** Performance on the Semeval test set. `4lang` and `4lang_syn` is our method without and with additional synonym nodes from WordNet and Wiktionary. `WordNet` is the baseline using WordNet hypernyms, `all` is the union of `4lang+syn` and `WordNet`. Previous top-scoring systems on each task are BMEAUT (Kovács *et al.* 2020) and SHIKEBLCU (Wang *et al.* 2020)

| Lang | Method | P | R | F |
|---|---|---|---|---|
| EN | always yes | 56.33 | 100.0 | 72.07 |
| | WordNet | 94.40 | 89.29 | 91.77 |
| | `4lang_syn` | 94.40 | 38.74 | 54.94 |
| | BMEAUT | | | 91.77 |
| | all | 93.02 | **94.70** | **93.85** |
| | RoBERTa | 85.66 | 91.33 | 88.41 |
| DE | GermaNet | 89.57 | 67.25 | 76.82 |
| | `4lang_syn` | 81.02 | 28.11 | 41.74 |
| | SHIKEBLCU | | | 71.43 |
| | all | 84.12 | **72.59** | **77.93** |
| IT | always yes | 41.67 | 100.0 | 58.82 |
| | WordNet | 87.08 | 76.43 | 81.41 |
| | `4lang_syn` | 88.58 | 28.57 | 43.20 |
| | BMEAUT | | | 81.41 |
| | all | 84.88 | **79.38** | **82.03** |

Bold values represents best scores.

We also evaluate our system on the Semeval test set, figures are presented in Table 7. We compare our current configuration with the top-scoring system from the 2020 Semeval competition (Glavaš *et al.* 2020). On English and Italian, the previous top system is our own BMEAUT submission (Kovács *et al.* 2020) and on German it is the SHIKEBLCU (Wang *et al.* 2020) system, which specializes distributional word vectors for lexical relations. The results on the test dataset shows that after improving `4lang` with additional methods and synonyms we achieve state-of-the-art results in all three languages. For German, the greatest improvement is clearly brought about by the new, high-quality German WordNet release GermaNet. Finally, we also evaluate a state-of-the-art neural NLI system on the Semeval datasets. RoBERTa (Zhuang *et al.* 2021) was trained on a dataset with three labels: entailment, neutral, and contradiction. We interpret its predictions by merging the latter two labels into a single label "not entailment". In absence of sufficient training data, we tuned the model to the dataset by setting the threshold of the output weights for optimal performance on the development set. The resulting model was evaluated on both the development and test portions of the dataset.

Next we performed evaluation on the SherLlic dataset, comparing several configurations of our method to both the RoBERTa system and a wide variety of baselines published alongside the dataset (Schmitt and Schütze 2019). Figures are presented in Table 8. We experimented with two configurations of our current system, tuned to high F-score and to high precision respectively. The trivial "always yes" baseline yields $F = 49.9$ on this dataset, and tellingly, the best earlier rule-based systems, `Berant II` (Berant 2012) and PPDB (Pavlick *et al.* 2015), achieved only $F = 30.0$ and $F = 34.7$. In the high F-score configuration (using the method described in Section 3.2 complete

**Table 8.** Performance on the SherLlic test set. `WordNet` is the baseline using WordNet hypernyms, `ESIM` (Chen *et al*. 2017) is the strongest system evaluated that wasn't tuned on SherLlic's held-out portion and `w2v+tsg_rel_emb` is the overall strongest system of Schmitt and Schütze (2019). `4lang_high_prec` and `4lang_high_fscore` are the configurations of our system tuned for high precision and high F-score, respectively

| Method | P | R | F |
|---|---|---|---|
| always yes | 33.3 | 100 | 49.9 |
| WordNet | 38.8 | 35.7 | 37.2 |
| `4lang_high_prec` | **89.06** | 11.46 | 20.32 |
| `4lang_high_fscore` | 44.65 | 52.51 | 48.26 |
| ESIM | 39.0 | **83.3** | 53.1 |
| `w2v + tsg_rel_emb` | 51.8 | 72.7 | 60.5 |
| RoBERTa | 70.0 | 76.35 | **73.0** |
| + `4lang_high_prec` | 69.79 | 76.96 | **73.20** |
| + `4lang_high_fscore` | 51.00 | 85.00 | 64.00 |

Bold values represents best scores.

with all postprocessing steps (expansion, reduction) and using all Wiktionary definitions), 4lang, though clearly better than the earlier systems and WordNet, is still below the trivial baseline. The high-precision configuration, obtained by blocking the inheritence of relations in the premise graph after expansion and also the reduction of the hypothesis graphs, is worse than the earlier rule-based systems, but proves its usefulness in the hybrid configuration with RoBERTa, where it improves not just over the high-F variant but also over RoBERTa used alone, which defined the state-of-the-art before this work.

## 5. Error analysis

The graph-based method presented in previous sections is an example of eXplainable AI (XAI). Decisions taken by any variant of our algorithm can be represented as paths of concepts between premise and hypothesis in a concept graph, or lack thereof, making qualitative error analysis straightforward. Our method is high-precision by design, and on the simpler word-level task its false positives are limited to a small number of unique accidents, such as Wiktionary defining *video* as *television*. Therefore, in this section, we focus on tracing the reasons behind false negatives, a task that simply cannot be performed outside the XAI context.

We begin with the simpler, word-level task, where our standalone method achieves considerably lower recall and F-score than what is already possible by including another lexical resource with an explicit model of hypernymy. We manually inspected 60 positive entailment pairs in the English Semeval dev dataset that were missed by our core method, that is the system without additional synonyms.

By far the most common, and in our opinion, the most interesting, source of false negatives is when the expanded premise graph correctly contains most or even all of the semantic content of the hypothesis word, yet there is no direct match, the hypothesis word cannot be accessed. An example is the word pair *lettuce* → *food*. *Lettuce* is defined in Wiktionary as "an edible plant, Lactuca sativa and its close relatives, having a head of green and/or purple leaves", `edible` means "can be eaten without harm" and finally `eat` is simply defined as "to ingest". Since this is as far as our iterative expansion goes, we are missing the word *food* altogether.

The main issue here is that the default object of eating is *food*, a fact well represented in the definition of *eat* in dictionaries such as LDOCE (Bullon 2003) "to put food in your mouth and chew and swallow it"; Webster's New World (Guralnik 1958) "to chew and swallow (food)"; The Concise Oxford (McIntosh 1951) "masticate and swallow (solid food)"; Webster's 3rd (Gove 1961) "to take in through the mouth as food". The connection with the object is so strong that it is also regularly present in the definition of *edible*: "fit to eat, food" (Collins); "suitable by nature for use as food" (Webster's 3rd); etc. Even *ingest*, a word that can be appropriately used for drinks, poison, etc. that are clearly not fit to be eaten, will generally include the default: "to take food or other substances into your body" (LDOCE); "to take (food) to the stomach" (Concise Oxford); "consume food or drink" (FrameNet) etc.

Since the association between eating and food is direct, we simply need to enhance the definition base further. A more interesting case is presented by *mussel* and *seafood*, where we do not believe that any reasonable lexical resource can help. This is because dictionaries are not at all good sources for plants, animals, minerals, and common encyclopedic knowledge about them: basically every dictionary we consulted defines *mussel* as a kind of mollusc (often using the technical term "bivalvate"), but gives up on defining *mollusc* by shunting the reader to the encyclopedia "animal belonging to *Mollusca*, a subkingdom of soft-bodied and usually hard-shelled animals". As we learn for example from the Concise Oxford, this subkingdom includes limpets, snails, cuttlefish, etc, and most of these (with the exception of oysters) are actually not considered as seafood. All of this points to the conclusion that for these cases a different, encyclopedic knowledge source should be consulted and here Wikipedia works well: the article en.wikipedia.org/wiki/Mussel actually has a heading As food.

A further error class comprises words for which the first definition in Wiktionary does not correspond to the sense intended in the entailment pair, most often because it is in fact not the most common sense of the word. An example is *submarine*, whose first sense in Wiktionary is defined as "underwater". Quite often, there is no clear "main sense" of the word and it is only the entailment candidate that allows us to disambiguate between multiple senses. Examples are *letter* → *mail* which is labeled as entailment but simply is not if we choose the definition "symbol in an alphabet", or *mole* → *animal*, which fails for the sense "pigmented skin" the same way. When constructing the Hyperlex dataset (Vulić *et al*. 2017), which later became the basis of the Semeval dataset used in this paper, annotators were instructed that *"two words stand in a type-of relation if any of their senses stand in a type-of relation."*(Vulić *et al*. 2017, p. 797). This might suggest that we consider the union of all definitions of a word for our method, but our early experiments showed that (because of the crowd-sourced nature of Wiktionary entries) such an approach would very often lead to the proliferation of erroneous representations built from low-quality and/or unwarranted definitions. Alternative solutions might include disambiguating among definitions based on context and/or building meaning representations from groups of definitions that describe multiple uses of the same abstract word sense. For a discussion of the difficulties of such approaches, the reader is referred to Section 4.4.3 of Recski (2018).

In Table 9, we highlight some cases where the predictions of `4lang` and RoBERTa differ. In the first case, `4lang` misses the meaning "support, protect" for the verb *back* (also seen in the multiword expression *have the back of*), but we simply do not know why RoBERTa is getting this right, and have no performance guarantees that further training or other improvements to F will preserve this particular instance. The second example is more subtle: ambassadors always represent their country but presidents do this much more rarely. Under a strict logical reading the implication fails, but as a practical matter, we should accept it, since representing their country is a *typical* activity of their presidents. While `4lang` has the means of expressing typicality (defaults), and some dictionaries such as Guralnik (1958) make reference to "formal head", it is again the more encyclopedic sources that must be consulted to find *formal* or *ceremonial* from which eventually represent(ative) can be inferred.

**Table 9.** Comparing our system to RoBERTa. Examples from Schmitt and Schütze (2019)

| Example | `41ang` | RoBERTa |
|---|---|---|
| ORGF[A] is supporter of ORGF[B] ⇒ ORGF[A] is backing ORGF[B] | False | True |
| AUTH[A] is president of LOC[B] ⇒ AUTH[A] is representing LOC[B] | False | True |
| PERSON[A] is REGION[B]'s ruler ⇒ PERSON[A] is dictator of REGION[B] | False | True |
| LOCATION[A] is winning war against LOCATION[B] ⇒ LOCATION[A] is declaring war on LOCATION[B] | False | True |

That such inference is fraught with difficulties is clear from the third and fourth examples, where RoBERTa has false positives, we think precisely because *dictator* often appears in text sufficiently close to *president*, and war-winning to declaring war. While in principle this hypothesis could be tested by retraining RoBERTa on its original training set minus these sentences, and rerunning the experiment with the RoBERTa' so obtained, devising a less CPU-intense experiment seems warranted.

False positives of our system are considerably easier to track: for example, we lack a rule for handling modally subordinated predicates. Therefore, we conclude *invade* from *planning to invade* and worse yet, *win* from *failing to win*. Short and too generic definitions in Wiktionary are another main source of our false positives: we conclude *A is selling to B* from *A is leaving to B* because *sell* and *leave* both contain *act* in their definition.

In some cases, valuable information is lost when we prune the contents of prepositional phrases from our premise graph (to avoid false positive entailments such as *nose* → *face*, as discussed in Section 3.2). For example, our system does not detect the entailment *husband* → *spouse* because Wiktionary defines *husband* as "a man in a marriage or marital relationship, especially in relation to his spouse". In dictionaries that are built on stronger lexicographic principles, this is avoided by reliance on a strict defining vocabulary, a limited set of concepts capable of defining all other concepts. For example in LDOCE, *spouse* is defined as "a husband or wife" which avoids any complication. Following Webster's 3rd, modern lexicographic practice avoids defining the simple by the more complex, and the idea of defining *eat* by means of "masticate" or "ingest" is seen as useless, as the language learner who does not know *eat* is quite unlikely to know these other verbs. Wiktionary, however, has not fully incorporated modern lexicographic principles.

Replacing Wiktionary with modern explanatory dictionaries would make it possible to extract higher quality representations, but only at the cost of broad applicability to the ever-increasing set of languages that already have significant Wiktionaries. That said, high-quality data sources can significantly reduce the inherent difficulty of polysemous words. For example, English does not mark causativization on the verb, so that *run* can express both "go fast" and "make go fast". If we consider the 50+ senses that Wiktionary provides for the verb *run*, there is little chance of finding the one appropriate for the type signature `(PERSON, COMPANY)`, which is meaning 22 "to control or manage, be in charge of" or the one appropriate for `(COMPUTER, SOFTWARE)` (meaning 26) "to execute or carry out a plan, procedure or program". Yet this level of disambiguation, between plain and causative forms, seems quite feasible, especially using dynamic embeddings earlier in the analysis process.

## 6. Conclusion

The value of a well-constructed dataset is that it leads to interesting problems. In this regard, `SherLIiC` is truly valuable, as it inspires us to think more deeply about synonymy, polysemy, disambiguation, definitional economy, prepositional linkers, modal subordination, causativization,

and a host of other questions that are traditionally considered central to natural language semantics.

We are fortunate to have WordNet, with its extensive hypernym links, tailor-made for entailment detection. Yet as we have seen, even WordNet can be profitably combined with other resources, dictionaries in particular. But for other relationals, such as causation, possession, mereological implications, spatiotemporal reasoning, etc, we would need similar datasets that highlight these very real problems. Since WordNet is less helpful there, we could expect considerably worse results, but it is unclear how progress could be made on grand semantic challenges in the absence of such new datasets.

In this article, we presented an explainable, multilingual method for detecting lexical entailment using a pipeline for the automatic construction of semantic graphs from dictionary definitions and a simple rule-based method for detecting entailment between pairs of such graphs. Our method presents a strong, high-precision baseline on the simpler task of detecting hypernymy and lets us improve over the performance of a manually created resource like WordNet. On the more complex task presented by the SherLlic dataset, our method outperforms all known rule-based baselines and is outperformed only by systems that have been adapted to the dataset. We also presented a baseline using a pretrained version of RoBERTa trained on MultiNLI, achieving even better results then the previously published distributional baselines, but combining it with our rule-based method further increases its performance eventually achieving state-of-the-art performance.

In future work, we hope to address several of the issues exemplified in our error analysis. One particularly promising avenue is to invoke a more explicit disambiguation process before, or in parallel to, the modeling of lexical entailment.

**Acknowledgements.** We are grateful to the three anonymous reviewers for their thoughtful comments, for questions leading to additional discussion in the manuscript, and for excellent references. Work partly supported by BRISE-Vienna (UIA04-081), a European Union Urban Innovative Actions project. Kovacs and Kornai were partially supported by MILAB, the Hungarian Artificial Intelligence National Laboratory. Competing interests: The authors declare none. The authors acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Program.

## References

**Abend O. and Rappoport A.** (2013). Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria. Association for Computational Linguistics, pp. 228–238.

**Ács J., Pajkossy K. and Kornai A.** (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, Sofia, Bulgaria. Association for Computational Linguistics, pp. 52–58.

**Banarescu L., Bonial C., Cai S., Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn P., Palmer M. and Schneider N.** (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria. Association for Computational Linguistics, pp. 178–186.

**Baroni M., Bernardi R., Do N.-Q. and Shan C.-c.** (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France. Association for Computational Linguistics, pp. 23–32.

**Berant J.** (2012). *Global Learning of Textual Entailment Graphs*. PhD Thesis, Tel Aviv University.

**Berant J., Dagan I. and Goldberger J.** (2011). Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA. Association for Computational Linguistics, pp. 610–619.

**Bowman S.R., Angeli G., Potts C. and Manning C.D.** (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics, pp. 632–642.

**Bullon S.** (2003). *Longman Dictionary of Contemporary English*, 4th Edn. Münich, Germany: Longman.

**Chen Q., Zhu X., Ling Z.-H., Wei S., Jiang H. and Inkpen D.** (2017). Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 1657–1668.

**Chen Z., Cui Y., Ma W., Wang S., Liu T. and Hu G.** (2018). Hfl-rc system at semeval-2018 task 11: Hybrid multi-aspects model for commonsense reading comprehension. arXiv preprint arXiv:1803.05655.

**Courcelle B. and Engelfriet J.** (2012). *Graph Structure and Monadic Second-Order Logic: A Language-Theoretic Approach*. *Encyclopedia of Mathematics and its Applications*. Cambridge: Cambridge University Press.

**Devlin J.**, **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp. 4171–4186.

**Gabrilovich E.**, **Ringgaard M. and Subramanya A.** (2013). Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).

**Gerdes K. and Kahane S.** (2015). Non-constituent coordination and other coordinative constructions as dependency graphs. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden. Uppsala University, Uppsala, Sweden, pp. 101–110.

**Glavaš G. and Ponzetto S.P.** (2017). Dual tensor model for detecting asymmetric lexico-semantic relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 1757–1767.

**Glavaš G. and Vulić I.** (2018). Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 34–45.

**Glavaš G. and Vulić I.** (2019). Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 4824–4830.

**Glavaš G.**, **Vulić I.**, **Korhonen A. and Ponzetto S.P.** (2020). SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics, pp. 24–35.

**Glockner M.**, **Shwartz V. and Goldberg Y.** (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 650–655.

**Gontrum J.**, **Groschwitz J.**, **Koller A. and Teichmann C.** (2017). Alto: Rapid prototyping for parsing and translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain. Association for Computational Linguistics, pp. 29–32.

**Gove P.B.** (ed.) (1961). *Webster's Third New International Dictionary of the English Language, Unabridged*. Ypsilanti, MI, USA: G. & C. Merriam.

**Groschwitz J.**, **Koller A. and Teichmann C.** (2015). Graph parsing with s-graph grammars. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China. Association for Computational Linguistics, pp. 1481–1490.

**Groschwitz J.**, **Lindemann M.**, **Fowlie M.**, **Johnson M. and Koller A.** (2018). AMR dependency parsing with a typed semantic algebra. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 1831–1841.

**Guralnik D.B.** (ed.) (1958). *Webster's New World Dictionary of the American Language*. Chicago, USA: The World Publishing Company.

**Hamp B. and Feldweg H.** (1997). GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

**Haruta I.**, **Mineshima K. and Bekki D.** (2020). Combining event semantics and degree semantics for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics, pp. 1758–1764.

**Henrich V. and Hinrichs E.** (2010). GernEdiT - the GermaNet editing tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association.

**Hershcovich D.**, **Abend O. and Rappoport**, **A.** (2017). A transition-based directed acyclic graph parser for UCCA. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 1127–1138.

**Hershcovich D.**, **Abend O. and Rappoport A.** (2018). Multitask parsing across semantic representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 373–385.

**Kalouli A.-L.**, **Crouch R. and de Paiva V.** (2020). Hy-NLI: A hybrid system for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics, pp. 5235–5249.

**Kamath A.**, **Pfeiffer J.**, **Ponti E.M.**, **Glavaš G. and Vulić I.** (2019). Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Florence, Italy. Association for Computational Linguistics, pp. 72–83.

**Kanayama H.**, **Han N.-R.**, **Asahara M.**, **Hwang J.D.**, **Miyao Y.**, **Choi J.D. and Matsumoto Y.** (2018). Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Brussels, Belgium. Association for Computational Linguistics, pp. 75–84.

**Kipper K.**, **Korhonen A.**, **Ryant N. and Palmer M.** (2008). A large-scale classification of English verbs. *Language Resources and Evaluation* **42**(1), 21–40.

**Koller A.** (2015). Semantic construction with graph grammars. In *Proceedings of the 11th International Conference on Computational Semantics*, London, UK. Association for Computational Linguistics, pp. 228–238.

**Koller A. and Engonopoulos N.** (2017). Integrated sentence generation using charts. In *Proceedings of the 10th International Conference on Natural Language Generation*, Santiago de Compostela, Spain. Association for Computational Linguistics, pp. 139–143.

**Koller A. and Kuhlmann M.** (2011). A generalized view on parsing and translation. In *Proceedings of the 12th International Conference on Parsing Technologies (IWPT)*, Dublin.

**Koller A. and Kuhlmann M.** (2012). Decomposing TAG algorithms using simple algebraizations. In *Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11)*, Paris, France, pp. 135–143.

**Konstas I.**, **Iyer S.**, **Yatskar M.**, **Choi Y. and Zettlemoyer L.** (2017). Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, pp. 146–157.

**Kornai A.** (2010). The treatment of ordinary quantification in English proper. *Hungarian Review of Philosophy* **54**(4), 150–162.

**Kornai A.** (2012). Eliminating ditransitives. In de Groote P. and Nederhof M.-J. (eds), *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*, LNCS, vol. 7395. Berlin, Germany: Springer, pp. 243–261.

**Kornai A.** (2019). *Semantics*. Cham, Switzerland: Springer Verlag.

**Kornai A.**, **Ács J.**, **Makrai M.**, **Nemeskey D.M.**, **Pajkossy K. and Recski G.** (2015). Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, Denver, Colorado. Association for Computational Linguistics, pp. 165–175.

**Kovács Á.**, **Ács E.**, **Ács J.**, **Kornai A. and Recski G.** (2019). BME-UW at SRST-2019: Surface realization with interpreted regular tree grammars. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, Hong Kong, China. Association for Computational Linguistics, pp. 35–40.

**Kovács Á.**, **Gémes K.**, **Kornai A. and Recski G.** (2020). BMEAUT at SemEval-2020 task 2: Lexical entailment with semantic graphs. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics, pp. 135–141.

**Lan Z.**, **Chen M.**, **Goodman S.**, **Gimpel K.**, **Sharma P. and Soricut R.** (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net.

**Levy O. and Dagan I.** (2016). Annotating relation inference in context via question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 249–255.

**Levy O.**, **Remus S.**, **Biemann C. and Dagan I.** (2015). Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado. Association for Computational Linguistics, pp. 970–976.

**Li B.**, **Wen Y.**, **Qu W.**, **Bu L. and Xue N.** (2016). Annotating the little prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in Conjunction with ACL 2016 (LAW-X 2016)*, Berlin, Germany. Association for Computational Linguistics, pp. 7–15.

**Liu X.**, **Gao J.**, **He X.**, **Deng L.**, **Duh K. and Wang Y.-y.** (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado. Association for Computational Linguistics, pp. 912–921.

**Liu X.**, **He P.**, **Chen W. and Gao J.** (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 4487–4496.

**Lyu C. and Titov I.** (2018). AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 397–407.

**McIntosh E.** (ed.) (1951). *The Concise Oxford Dictionary of Current English*, 4th Edn. London, UK: Oxford University Press.

**Miller G.A.** (1995). Wordnet: A lexical database for English. *Communications of the ACM* **38**(11), 39–41.

**Nelson D.L.**, **McEvoy C.L. and Schreiber T.A.** (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* **36**(3), 402–407.

**Nguyen K.A.**, **Köper M.**, **Schulte im Walde S. and Vu N.T.** (2017). Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark. Association for Computational Linguistics, pp. 233–243.

**Nivre J.**, **Abrams M.**, **Agić Ž.** et al. (2018). Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

**Ozaki H.**, **Morio G.**, **Koreeda Y.**, **Morishita T. and Miyoshi T.** (2020). Hitachi at MRP 2020: Text-to-graph-notation transducer. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, Online. Association for Computational Linguistics, pp. 40–52.

**Palmer M.**, **Gildea D. and Kingsbury P.** (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* **31**(1), 71–106.

**Pavlick E.**, **Rastogi P.**, **Ganitkevitch J.**, **Van Durme B. and Callison-Burch C.** (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Beijing, China. Association for Computational Linguistics, pp. 425–430.

**Qi P.**, **Zhang Y.**, **Zhang Y.**, **Bolton J. and Manning C.D.** (2020). Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Online. Association for Computational Linguistics, pp. 101–108.

**Recski G.** (2016). Building concept graphs from monolingual dictionary entries. In Calzolari N., Choukri K., Declerck T., Grobelnik, M., Maegaard B., Mariani J., Moreno, A., Odijk J. and Piperidis S. (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

**Recski G.** (2018). Building concept definitions from explanatory dictionaries. *International Journal of Lexicography* **31**, 274–311.

**Recski G.**, **Borbély G. and Bolevácz A.** (2016). Building definition graphs using monolingual dictionaries of Hungarian. In Tanács A., Varga V. and Vincze V. (eds), *XI. Magyar Számitógépes Nyelvészeti Konferencia [11th Hungarian Conference on Computational Linguistics*, Szeged, Hungary.

**Recski G.**, **Kovács Á.**, **Gémes K.**, **Ács J. and Kornai A.** (2020). BME-TUW at SR'20: Lexical grammar induction for surface realization. In *Proceedings of the Third Workshop on Multilingual Surface Realisation*, Barcelona, Spain (Online). Association for Computational Linguistics, pp. 21–29.

**Roller S.**, **Erk K. and Boleda G.** (2014). Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland. Dublin City University and Association for Computational Linguistics, pp. 1025–1036.

**Ruppenhofer J.**, **Ellsworth M.**, **Petruck M.R.**, **Johnson C.R. and Scheffczyk J.** (2006). *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute. Distributed with the FrameNet data.

**Samuel D. and Straka M.** (2020). ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, Online. Association for Computational Linguistics, pp. 53–64.

**Schmitt**, **M. and Schütze**, **H.** (2019). SherLIiC: A typed event-focused lexical inference benchmark for evaluating natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 902–914.

**Schuster S. and Manning C.D.** (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA), pp. 2371–2378.

**Shwartz V.**, **Goldberg Y. and Dagan I.** (2016). Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, pp. 2389–2398.

**Talman A. and Chatzikyriakidis S.** (2019). Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy. Association for Computational Linguistics, pp. 85–94.

**Upadhyay S.**, **Vyas Y.**, **Carpuat M. and Roth D.** (2018). Robust cross-lingual hypernymy detection using dependency context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. ACL, pp. 607–618.

**Vaswani A.**, **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser L.U. and Polosukhin I.** (2017). Attention is all you need. In Guyon I., Luxburg U.V., Bengio S., Wallach H., Fergus R., Vishwanathan S. and Garnett R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 5998–6008.

**Vulić I.**, **Gerz D.**, **Kiela D.**, **Hill F. and Korhonen A.** (2017). Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics* **43**(4), 781–835.

**Vulić I.**, **Ponzetto S.P. and Glavaš G.** (2019). Multilingual and cross-lingual graded lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 4963–4974.

**Wang B. and Kuo C.-C.** (2020). Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 1–1.

**Wang S.**, **Fan Y.**, **Luo X. and Yu D.** (2020). SHIKEBLCU at SemEval-2020 task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics, pp. 255–262.

**Williams A.**, **Nangia N. and Bowman S.** (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, pp. 1112–1122.

**Xue N.**, **Bojar O.**, **Hajič J.**, **Palmer M.**, **Urešová Z. and Zhang X.** (2014). Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 1765–1772.

**Yu Z.**, **Wang H.**, **Lin X. and Wang M.** (2015). Learning term embeddings for hypernymy identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15. AAAI Press, pp. 1390–1397.

**Zeichner N.**, **Berant J. and Dagan I.** (2012). Crowdsourcing inference-rule evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Jeju Island, Korea, pp. 156–160.

**Zhang S.**, **Ma X.**, **Duh K. and Van Durme B.** (2019). AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 80–94.

**Zhang Z.**, **Wu Y.**, **Zhao H.**, **Li Z.**, **Zhang S.**, **Zhou X. and Zhou X.** (2020). Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 9628–9635.

**Zhuang L.**, **Wayne L.**, **Ya S. and Jun Z.** (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Huhhot, China. Chinese Information Processing Society of China, pp. 1218–1227.

**Cite this article:** Kovacs A, Gemes K, Kornai A and Recski G. Explainable lexical entailment with semantic graphs. *Natural Language Engineering* https://doi.org/10.1017/S1351324922000092