BMEAUT at SemEval-2020 Task 2: Lexical entailment with semantic graphs

Ádám Kovács[*], **Kinga Gémes**
Dept. of Automation and Applied Informatics
Budapest University of Technology and Economics
kovacs.adam@aut.bme.hu
gemes.kingaandrea@aut.bme.hu

Andras Kornai
SZTAKI Institute of Computer Science
kornai@sztaki.hu

Gábor Recski
TU Wien
gabor.recski@tuwien.ac.at

Abstract

In this paper we present a novel rule-based, language independent method for determining lexical entailment relations using semantic representations built from Wiktionary definitions. Combined with a simple WordNet-based method our system achieves top scores on the English and Italian datasets of the Semeval-2020 task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment" (Glavaš et al., 2020). A detailed error analysis of our output uncovers future directions for improving both the semantic parsing method and the inference process on semantic graphs. reported in this document.

1 Introduction

We present a rule-based, multilingual system for detecting monolingual binary entailment between pairs of words using Wiktionary definitions, dependency parsing, and semantic graphs. We define entailment over pairs of semantic graphs built from dictionary definitions and achieve near-perfect precision on the Semeval-2020 task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment" (Glavaš et al., 2020), where we participate in the *ANY* track of the monolingual task that allows for the use of external lexico-semantic resources. Our system improves the performance of strong WordNet-based baselines on three languages, achieving top results on English and Italian and second-best on German. Our pipeline can be easily extended to support any language given a monolingual dictionary and a Universal Dependency (UD) parser. A detailed error analysis shows multiple directions for further improvement, most notably the refinement of the mechanism responsible for recursively extending semantic graphs based on the definition of its nodes. Section 2 briefly describes the lexical entailment task, the `4lang` semantic representation (Kornai et al., 2015), and the `dict_to_4lang` tool (Recski, 2016) for generating graphs from dictionary definitions (Recski, 2016; Recski, 2018). Section 3 outlines the architecture of our current system and presents our method for detecting entailment over pairs of `4lang` graphs. Our results on the shared task and a detailed error analysis is presented in Section 4. Our system is available for download under an MIT license from GitHub under `https://github.com/adaamko/wikt2def/tree/semeval`.

2 Background

2.1 Lexical Entailment

A common definition of lexical entailment, used also for the current shared task, is that of recognizing `IS_A` relationships between pairs of words (e.g. *lettuce* entails *food*). Datasets used in this shared task are derived from the HyperLex dataset (Vulić et al., 2017), methods for measuring multi-lingual and cross-lingual lexical entailment using specialized word embeddings are presented in (Vulić et al., 2019) and outperform previous baselines in (Upadhyay et al., 2018). Another common lexical inference task

[*] corresponding author

is defined between pairs of predicates in context, where context can be defined as pairs of arguments (Zeichner et al., 2012), pairs of argument types (Berant et al., 2011; Schmitt and Schütze, 2019), or question-answer pairs (Levy and Dagan, 2016).

Dictionary definitions have recently been used for explainable modeling of lexical entailment: (Silva et al., 2018) build semantic graphs from WordNet definitions (glosses) using a recurrent neural network trained on thousands of annotated examples, then search for paths of distributionally similar words between premise and hypothesis. Our method differs from their approach in its lack of training, which makes its applicable to any language for which a monolingual dictionary and a dependency parser is available. While the semantic formalism used in this paper treats lexical inference as a broader term subsuming not just hypernymy but also attribution and predication (such that *dog* entails not only *mammal* but also *four-legged* and *bark*), the context-free detection of `IS_A` relations between pairs of largely unambiguous words proves challenging enough to provide insight about the current shortcomings of our semantic representations.

2.2 4lang

The `4lang` formalism (Kornai et al., 2015) represents the meaning of linguistic units (both words and phrases) as directed graphs of language- and syntax-independent concepts. Nodes roughly correspond to content words, edges connecting them can have one of three labels: 0-edges simultaneously represent attribution (`dog` $\xrightarrow{0}$ `four-legged`), hypernymy (`dog` $\xrightarrow{0}$ `mammal`) and unary predication (`dog` $\xrightarrow{0}$ `bark`). Predicates are connected to their arguments via edges labeled 1 and 2, e.g. `cat` $\xleftarrow{1}$ `catch` $\xrightarrow{2}$ `mouse`. Concepts have no grammatical attributes and no event structure, e.g. the phrases *water freezes* and *frozen water* would both be represented as `water` $\xrightarrow{0}$ `freeze`.

We build `4lang` graphs using a reimplementation of the `dict_to_4lang` tool (Recski, 2016), essentially a pipeline of dependency parsing and a set of simple, hand-written rules mapping UD substructures (Nivre et al., 2018) to `4lang` subgraphs. For example, dependency relations such as `amod` and `advmod` are mapped to 0-edges but `obj` and `nsubjpass` are mapped to 2-edges (see (Recski, 2016) for the full mapping). Figure 1 shows an example, the `4lang` definition and corresponding UD parse of the concept `jewel`, obtained by processing the Wiktionary definition *A precious or semi-precious stone; gem, gemstone*. Optionally, the `4lang` system allows us to *expand* graphs, a process which unifies the graph with the definition graphs of each concept within the graph. Figure 1 is an example of applying the *expand* method on the concept `jewel`. This operation will be essential to our method presented in Section 3. Our system currently supports three languages, but extending `dict_to_4lang` to further languages only requires a trained UD parser and a language-specific extractor for Wiktionary[1].

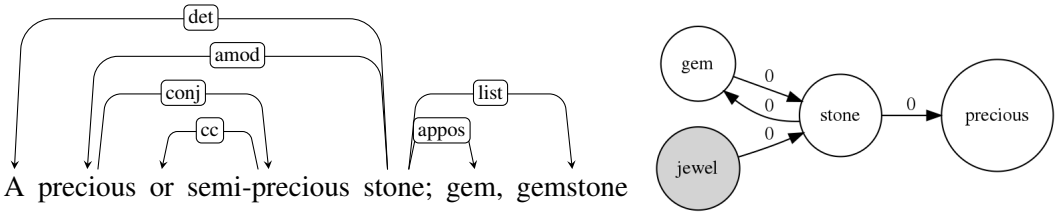

Figure 1: 4lang graph and UD parse of the definition of `jewel`

3 Method

In this section we describe the pipeline used for creating pairs of 4lang graphs from each word pair in the SemEval dataset and our method for determining, based on these graphs, whether entailment between the two words can be established. The only language-specific components of our pipeline are the UD parser, which is available for 50+ languages from the `stanfordnlp` module alone, and the templates used to

[1]The mapping from UD representations to 4lang graphs can be extended to incorporate morphological tags for languages where UD relations convey insufficient information, an example is presented in (Recski et al., 2016)

extract definitions from Wiktionary dumps, which are currently implemented for English, German, and Italian.

3.1 Parser

We use the `dep_to_4lang` module (see Section 2) for mapping universal dependency parses of Wiktionary definitions to 4lang graphs. To obtain these definitions we process publicly available dumps of Wiktionary for each language, extract markdown text from the XML format, find definitions using language-specific templates, and finally strip all markdown to obtain raw text phrases or sentences. By default our system uses the first definition of each word present on its page. Some definitions are explicitly marked as `obsolete`, `archaic`, `historical`, or `rare`, if these appear in the first position we skip them and use the second definition, if available. Based on manual error analysis this method appears to choose in over 98% of cases a definition corresponding to the word sense used in the entailment dataset. (Technically the entailment task may be considered ill-defined for highly polysemous words such as *letter*, since the question of whether *letter* entails *mail* hinges on the choice between the definitions 'a written or printed communication' and 'a symbol in an alphabet', only one of which entails any sense of *mail*. Similarly, the entailment pair *mole-animal* in the dev dataset goes undetected by our system because it chooses the definition 'pigmented skin' instead of 'any of several small, burrowing insectivores of the family Talpidae'. See Section 4 for details.)

3.2 Method on graphs

Given a Wiktionary dataset and a UD parser for some language we can generate the `4lang` definition graph for any word in the dataset using the system described in the previous sections. We therefore develop a method for detecting entailment for a pair of graphs corresponding to premise and hypothesis words. All relationships between concepts marked by 0-edges in `4lang` graphs (attribution, predication, hypernymy) constitute entailment, although the Semeval dataset is limited to hypernymy. We shall extend premise graphs by recursively expanding nodes accessible from the root via a path of 0-edges. We then define entailment to hold iff in this extended graph there is a directed path of 0-edges leading from the premise word to the hypothesis word. The single tunable parameter of our system is the number of times we perform the `expand` operation recursively, which we set to 2, as further expansion yields false positive matches such as when *four* is found to entail *two* after the third expansion. The main source of false positives generated by our method are word pairs where the hypothesis word is part of a locative phrase accessible from the premise word via a dependency path that is mapped to a path of 0-edges in the `4lang`-representation. For example, in the Wiktionary definition of `nose` *A protuberance on the face*, the dependency relation `nmod(nose, face)` is established, and in the resulting `4lang` graph the concept *face* becomes accessible via a 0-path from *nose*. We overcome this issue by deleting nodes that connect to any of a short language-specific list of function words such as certain prepositions (e.g. English *in, of, on*, German *in, auf*, Italian *di, su, il*) and words conveying negation (English *not*, German *keine*, etc.).

This method detects about a third of all true entailments in the `dev` dataset (see Section 4 for details), and achieves nearly perfect precision (only two false positives on both the English and Italian development datasets). We combine this system with a simple method based on WordNet: we also establish entailment between a pair of words if the hypothesis word is present in the set of hypernyms for any synset containing the premise word in the WordNet of the given language. For English and Italian official WordNet releases are available in the `nltk`[2] package. For German we did not have access to a high-coverage WordNet release, therefore we translated word pairs from German to English using the wikt2dict system (Ács et al., 2013) and used the union of English WordNet synsets corresponding to each of the translations. These hybrid systems proved superior in terms of F-score to both individual systems on all three languages.

[2] https://www.nltk.org/howto/wordnet.html

Lang	Method	Precision	Recall	F-score
EN	wordnet	**95.75**	88.76	92.12
	4lang	95.74	25.28	40.00
	both	94.70	**90.44**	**92.52**
IT	wordnet	88.96	75.88	81.90
	4lang	**95.55**	25.29	40.00
	both	88.81	**79.41**	**83.85**
DE	wordnet (en)	61.61	79.22	69.31
	4lang	**90.38**	30.51	45.63
	both	61.97	**85.71**	**71.93**

Table 1: Performance of our methods on the development dataset

System	en	de	it
GLEN baseline	79.87	59.88	66.27
BMEAUT	**91.77**	67.00	**81.41**
FERRYMAN	72.13	53.12	62.71
SHIKEBLCU	87.90	**71.43**	75.94

Table 2: Official monolingual LE results on the `ANY` track (F-scores)

4 Evaluation

We participated in the `ANY` track of the Semeval-2020 task "Predicting Multilingual and Cross-lingual (graded) Lexical Entailment", a detailed description of which is available in (Glavaš et al., 2020). We did not experiment with detecting cross-lingual or graded entailment, our systems produce binary output for pairs of words of the same language only, which we submitted to both the binary and graded subtasks of the monolingual task. We implemented Wiktionary extractor modules for three languages: English, German, and Italian. For each of these we measured on the development set not only the performance of our best-performing hybrid system but also that of the stand-alone WordNet and 4lang-based systems. Figures are shown in Table 1. Additionally, the official Semeval evaluation compared our system's performance on the test set to those of other participants and the GLEN baseline (Glavaš and Vulić, 2019), a hybrid system that specializes distributional vectors for lexical entailment using English synonymy, antonymy, and hypernymy constraints from WordNet and then transfers the specialization to other languages via cross-lingual word embedding spaces. Results from this evaluation are shown in Table 2.

While WordNet baselines outperform our method in terms of F-score due to our low recall, the high precision of the 4lang-based system allows us to improve overall performance on each language by increasing recall by 2 (4, 6) percentage points, corresponding to 3 (6, 9) additional true positives for English, Italian, and German, respectively. In Table 3 we list some examples of entailment pairs that have been detected as such by our method but not by WordNet, along with their Wiktionary definition of the premise that was used for building `4lang` representations. Results from the official evaluation shows that on all three languages our system outperforms a competitive baseline by a wide margin and scores higher than any other system on English and Italian data. Since our method yields very high precision at the cost of low recall, for the English dataset we conducted a detailed error analysis of false negative pairs to better understand the shortcomings of our method and representation.

The most common case of our method failing to detect entailment between two concepts based on their definition graphs is when the recursive extension of the premise graph contains most of the semantic content of the hypothesis without actually making the connection with the right concept. An example is the entailment pair *lettuce* $\rightarrow$ *food*. The graph built from the definition of the premise (*An edible plant, Lactuca sativa and its close relatives, having a head of green and/or purple leaves.*), then extended using the definition of `edible` (*can be eaten without harm*) and finally with that of `eat` (*to ingest*) will still fail to contain the concept `food`. Such mismatches highlight the need for reducing all such semantic representations to a small common set of defining concepts, a step which could then be performed for both premise and hypothesis words. Our future plans include implementing such a reduction along the

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 zweitgrößter Planet unseres Sonnensystem* 'sixth and second-largest planet of our solar system'
test	*esame*	*esame per verificare qualcosa* 'exam to check something'

Table 3: Examples of entailment pairs detected by our system

principles outlined in (Kornai et al., 2015) and (Kornai, 2019). The second class of errors is caused by definitions where the necessary pieces of information are expressed by prepositional phrases. As discussed in Section 3, we block inference across nodes such as `in`, `on`, etc. to avoid false positive entailments such as *nose* $\rightarrow$ *face*. This filter also reduces our knowledge of `husband`, defined as *A man in a marriage or marital relationship, especially in relation to his spouse* to `husband` $\overset{0}{\rightarrow}$ `man` $\overset{0}{\rightarrow}$ `human` $\overset{0}{\rightarrow}$ `male` and missing the entailment `husband` $\rightarrow$ `spouse`

A further error class is caused by 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*. Choosing the first sense defined nevertheless remains a strong heuristic, but see Section 4.4.3 of (Recski, 2018) for a discussion on how multiple definitions of a word might be incorporated in a single semantic representation. Our current approach of choosing the first and usually most common sense of a word also fails when there is no clear "main sense" of the word and it is only the entailment candidate that allows us to disambiguate between multiple senses. An example in the dev dataset is *letter* $\rightarrow$ *mail* which is labelled as entailment but simply isn't if we choose the definition *"A symbol in an alphabet."*. A possible remedy for this issue might be to establish entailment if *any* of the multiple definitions of a word warrants it, but such a modification of our method would cause many false positives due to the exponential growth in the number of nodes involved in the expansion process.

5 Conclusion

We presented a system of entailment detection that relies on a considerable amount of manual work for its data sources: both Wikitionary and WordNet were crafted by many years of human labor, and the UD parser trainsets are generally hand-corrected silver or hand-parsed gold sets. But the adaptive layer between these two, consisting mostly in trivial scripts that convert the formats, and a simple rule-based parser to extract the pivot representations from UD parses, is relatively thin, and quite easy to extend to further languages. Perhaps the most important takeaway from our work is that the classical resources of computational linguistics are exactly the kind of structures a system needs to learn. As the old miners' adage goes, *gold is where you find it*. Knowledge, in distilled and higly leverageable form, is in the dictionaries. Even if our ultimate goal is, as it should be, to extract the knowledge from raw data, experimentation with hybrid systems is warranted by the fact that symbolic systems, and so far only these, can be meaningfully debugged on the kind of relatively small but well-crafted datasets our shared tasks provide.

Acknowledgements

Kovács was partly supported by the ÚNKP-19-3 New National Excellence Program of the Ministry for Innovation and Technology. Kovács and Gémes were partly supported by project FIEK16-1-2016-0007 of the National Research, Development and Innovation Fund of Hungary, financed under the FIEK16 scheme of the Centre for Higher Education and Industrial Cooperation. Kovács and Kornai were supported by 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence, and in part by

the Hungarian Scientific Research Found (OTKA), contract number 120145. Recski was partly supported by BRISE-Vienna (UIA04-081), a European Union Urban Innovative Actions project.

References

Judit Ács, Katalin Pajkossy, and András Kornai. 2013. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria. Association for Computational Linguistics.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 610–619, Portland, Oregon, USA. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 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*, pages 4824–4830, Florence, Italy. Association for Computational Linguistics.

Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Ponzetto. 2020. SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 165–175, Denver, Colorado. Association for Computational Linguistics.

András Kornai. 2019. *Semantics*. Springer Verlag.

Omer Levy and Ido Dagan. 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)*, pages 249–255, Berlin, Germany. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, 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.

Gábor Recski, Gábor Borbély, and Attila Bolevácz. 2016. Building definition graphs using monolingual dictionaries of Hungarian. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *XI. Magyar Számitógépes Nyelvészeti Konferencia [11th Hungarian Conference on Computational Linguistics*.

Gábor Recski. 2016. Building concept graphs from monolingual dictionary entries. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Gábor Recski. 2018. Building concept definitions from explanatory dictionaries. *International Journal of Lexicography*, 31:274–311.

Martin Schmitt and Hinrich Schütze. 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*, pages 902–914, Florence, Italy. Association for Computational Linguistics.

Vivian S Silva, Siegfried Handschuh, and André Freitas. 2018. Recognizing and justifying text entailment through distributional navigation on definition graphs. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 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)*, pages 607–618, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2019. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4963–4974, Florence, Italy. Association for Computational Linguistics.

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