Blindness to Modality Helps Entailment Graph Mining

Liane Guillou[*†], **Sander Bijl de Vroe**[*†], **Mark Johnson**[‡], **Mark Steedman**[†]
[†]University of Edinburgh, [‡]Oracle Digital Assistant
{sbdv, liane.guillou}@ed.ac.uk
mark.mj.johnson@oracle.com, steedman@inf.ed.ac.uk

Abstract

Understanding linguistic modality is widely seen as important for downstream tasks such as Question Answering and Knowledge Graph Population. Entailment Graph learning might also be expected to benefit from attention to modality. We build Entailment Graphs using a news corpus filtered with a modality parser, and show that stripping modal modifiers from predicates in fact increases performance. This suggests that for some tasks, the pragmatics of modal modification of predicates allows them to contribute as evidence of entailment.

1 Introduction

The ability to recognise textual entailment and paraphrase is crucial in many downstream tasks, including Open Domain Question Answering from text. For example, if we pose the question "Did Joe Biden run for President?" and the text states that "Joe Biden was elected President", producing the correct answer (Yes) necessitates understanding that *being elected President* entails *running for President*[1].

Entailment Graphs, constructed via unsupervised learning techniques over large text corpora, provide a solution to this problem. They consist of nodes representing predicates and directed edges representing entailment relations between them. Given the importance of detecting uncertainty for other downstream NLP tasks such as Information Extraction (Karttunen and Zaenen, 2005; Farkas et al., 2010), Information Retrieval (Vincze, 2014), machine reading (Morante and Daelemans, 2012a), and Question Answering (Jean et al., 2016) one might expect that it would also be useful in learning Entailment Graphs. That is, they would be more reliable if learned from data in which predications are *asserted* as actually happening, rather than data with *uncertain* predications under scope of various types of modality. We investigate whether this is the case.

The Entailment Graph-learning algorithm depends on descriptions of eventualities in the news, observing directional co-occurrences of typed predicates and their arguments. For example, we expect to observe all the arguments of *being president*, such as *Biden* and *Obama*, also to be encountered in a sufficiently large multiply-sourced body of text as arguments of *running for president*, but not the other way around (*Hillary Clinton* will *run* but not *be president*). However, if all the reports of *Clinton **might** be president* are extracted as *be_president(Clinton)*, one might expect the learning signal to be confusing to the algorithm.

We use the method of Hosseini et al. (2018) combined with a modality parser (Bijl de Vroe et al., 2021) to construct typed Entailment Graphs from raw text corpora under two different settings. Modality-aware: modal predications are removed from the data entirely, and modality-unaware: the model learns from both asserted and modal predications. Our contributions are 1) a comparison of Entailment Graphs learned from modal and non-modal data, showing (counterintuitively) that ignoring modal distinctions in fact improves Entailment Graph-learning, and 2) insights as to whether this effect applies uniformly across different subdomains.

2 Background

Entailment rules specify directional inferences between linguistic predicates (Szpektor and Dagan, 2008), and can be stored in an Entailment Graph, whose global structural properties can be used to learn more accurately (Berant et al., 2011, 2015). They are defined as a directed graph $\mathcal{G} = \{N, E\}$, in which the nodes N are typed predicates and edges E represent the entailment relation. The lex-

[*]Equal contribution

[1]Assuming a democratic election. We use the typical definition of the premise *most likely* entailing the hypothesis (Dagan et al., 2006)

Category	Example
∅	Protesters attacked the police
Modal operator	Protesters **may** have attacked the police
Conditional	**If** protesters attack the police...
Counterfactual	**Had** protesters attacked the police...
Propositional attitude	Journalists **said** that protesters attacked the police

Table 1: Modality categories

ical entailment knowledge stored within them is useful for Question Answering (McKenna et al., 2021), as well as other tasks such as email categorisation (Eichler et al., 2014), relation extraction (Eichler et al., 2017) and link prediction (Hosseini et al., 2019).

A subgraph containing predicates of a type-pair (e.g. PERSON-LOCATION) can be learned in an unsupervised way from collections of multiply-sourced text. A vector of argument-pair counts for every predicate is first machine read from the corpus. Typically, relation extraction systems used for reading these corpora ignore modal modifiers, possibly introducing noise in the graph. Next, a (directed) similarity score (e.g. DIRT (Lin and Pantel, 2001), Weed's score (Weeds and Weir, 2003) or BInc (Szpektor and Dagan, 2008)) is computed between the vectors, producing a local entailment score between each predicate pair. Then a globalisation process such as the soft constraints algorithm of Hosseini et al. (2018), which transfers information both within and between type-pair subgraphs, can be used to refine these local scores. When using the graph in practice, all edges with a score above a chosen threshold can be considered an entailment.

There are various semantic phenomena a speaker can use to mark veridicality (see Table 1). *Modal operators*, e.g. *probably, might, should, need to*, allow the user to indicate their attitude beyond the propositional content of a phrase, and often don't entail that the eventuality occurs (Kratzer, 2012). The same holds for predications under scope of *conditionals* and *counterfactuals* (Dancygier, 1998; Lewis, 1973). *Propositional attitude*, indicated by verbs such as *say, imagine or want*, allows the speaker to attribute thoughts regarding some possible eventuality to a source (Nelson, 2019).

These phenomena have been investigated for various NLP tasks, including uncertainty detection (Vincze, 2014), hedge detection (Medlock and Briscoe, 2007) and modality annotation (Saurı et al., 2006). Capturing this information is valuable to tasks such as Information Extraction, Question Answering and Knowledge Base Population (Karttunen and Zaenen, 2005; Morante and Daelemans, 2012b).

Early approaches to detecting modality focused on lexicon design (Szarvas, 2008; Kilicoglu and Bergler, 2008; Baker et al., 2010), with later approaches using machine learning over annotated corpora (Morante and Daelemans, 2009; Rei and Briscoe, 2010; Jean et al., 2016; Adel and Schütze, 2017). Recently, Bijl de Vroe et al. (2021) designed a parser similar to that by Baker et al. (2010), to cover a wider range of phenomena, including conditionality and propositional attitude. While modality annotation is clearly useful for recognising entailment from a given text (Snow et al., 2006; De Marneffe et al., 2006), to our knowledge no research has been conducted on its effect on learning Entailment Graphs.

3 Methods

We extend relation extraction to pay attention to modality, so that we can distinguish modal and non-modal relations in the Entailment Graph mining algorithm. This allows us to investigate the impact of modalised predicate data on the accuracy of learned entailment edges.

We extract *binary relations* of the form arg1-predicate-arg2 using MONTEE, an open-domain modality-aware relation extraction system (Bijl de Vroe et al., 2021). MONTEE uses the RotatingCCG parser (Stanojević and Steedman, 2019) as the basis for extracting binary relations and a modality lexicon to identify modality triggers. A relation is tagged as modal (MOD), propositional attitude (ATT_SAY, ATT_THINK) or conditional (COND) if the CCG dependency graph contains a path between a relation node and a node matching an entry in the MONTEE lexicon. Counterfactuals (COUNT) are tagged according to hand-crafted rules. Since we focus on uncertainty and not negation, lexical negation (LNEG) tagging is ignored.

In the modality-aware setting, we remove relations tagged by MONTEE as any kind of modal ({MOD, ATT_SAY, ATT_THINK, COUNT, COND}). In local learning, learned entailment edges then have access only to non-modal evidence: eventualities that were asserted as actually happening. For example, the edge between *win* and *lose* should now be learned only from non-modal descriptions such as *A won today against B* or *A has been defeated by B*, leaving out modal descriptions

(*A could beat B*). The local and globalisation parts of the algorithm are otherwise unchanged.

4 Experimental Setup

Using MONTEE[2], we extract 40,669,812 binary relation triples from the NewsSpike corpus (Zhang and Weld, 2013). Of these, 14.57% are tagged; 10.04% MOD, 3.51% REP_SAY, 0.38% REP_THINK, 0.61% COND, and 0.03% COUNT. We then construct three different datasets and build an Entailment Graph with each. The modality-unaware baseline, **BaselineLarge**, is trained on the complete set of relations with modality tags removed. This corresponds to the data and model in Hosseini et al. (2018). For the modality-aware **Asserted** graph, we extract only the set of 34,744,216 asserted relations (∼85% of the relations), i.e. all modal relations are excluded. To rule out effects of data size, we construct **BaselineSmall**, which is trained on a random sample of 85% relations from the total set. Comparing Asserted to BaselineLarge shows us whether it is worth filtering out modal data, and comparing Asserted to BaselineSmall shows whether asserted data or mixed data (i.e. asserted and modal) is more effective for learning entailment relations.

We follow the example of Hosseini et al. (2018) and construct typed graphs for all possible type pairs (e.g. PERSON-LOCATION). Relation arguments are typed by linking to a Named Entity Freebase identifier (Bollacker et al., 2008) using the AIDA-light linker (Nguyen et al., 2014), and mapping these identifiers to a type in the FIGER hierarchy (Ling and Weld, 2012). The typed relations become the input to the graph learning step of the Entailment Graph mining algorithm. Following previous research, we use the BInc similarity score (Szpektor and Dagan, 2008) to compute entailment scores. We first construct local typed Entailment Graphs and then globalise the scores across graphs as in Hosseini et al. (2018).

We evaluate the Entailment Graphs on two datasets. The first is the *Levy/Holt Entailment Dataset*, a set of 18,407 entailment pairs for the general domain (Levy and Dagan, 2016; Holt, 2018). As our training method is unsupervised and we do not tune hyperparameters, we evaluate on the complete Levy/Holt dataset rather than the dev/test split. We also evaluate on the *Sports Entailment Dataset* (Guillou et al., 2020), focusing on the subset of 718 examples comprising entailments and pairs of match outcome predicates (e.g. *win*, *lose*, *tie*, and their paraphrases) which are always non-entailments. This subset evaluates whether Entailment Graphs can recognise, for example, that win/lose → play but win ↮ lose (with similar patterns for other paraphrases of *win, play* and *lose*). We focus on the subgraph of ORGANISATIONs as all predicates are assumed to apply to sports teams. Both datasets use binary labels for each premise/hypothesis pair: entailment (1) and non-entailment (0).

We used the entGraph[3] code developed by Hosseini et al. (2018) to construct each of the Entailment Graphs, and the corresponding evaluation scripts[4] to evaluate performance on the Levy/Holt dataset. Performance on the Sports Entailment Dataset[5] is evaluated using scripts[6] developed for this paper. For details on hyperparameters and computational costs see Appendix A.

	Levy/Holt all	Levy/Holt directional	Sports
BaselineLarge	**0.190**	**0.163**	0.453
BaselineSmall	0.184	0.157	0.422
Asserted	0.171	0.136	**0.468**

Table 2: AUC scores

	Nodes	Edges	% Levy preds found	
			all ex.	directional
BaselineLarge	334K	72,7M	63.06	70.29
BaselineSmall	277K	58,4M	61.13	69.29
Asserted	254K	46,3M	58.51	67.92

Table 3: Graph size comparison and predicate coverage for Levy/Holt dataset (all examples) and its directional portion

	Nodes	Edges	% Sports preds found
BaselineLarge	4,514	1.65M	92.86
BaselineSmall	3,823	1.29M	90.48
Asserted	3,682	1.09M	88.10

Table 4: ORGANISATIONs subgraph size comparison and predicate coverage for the Sports Entailment Dataset

[2]https://gitlab.com/lianeg/montee

[3]https://github.com/mjhosseini/entGraph

[4]https://github.com/mjhosseini/entgraph_eval

[5]https://gitlab.com/lianeg/temporal-entailment-sports-dataset

[6]https://gitlab.com/lianeg/sports-entailment-evaluation

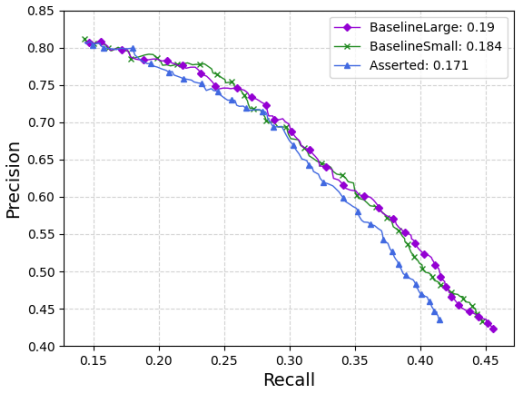

Figure 1: Precision/recall on Levy/Holt dataset

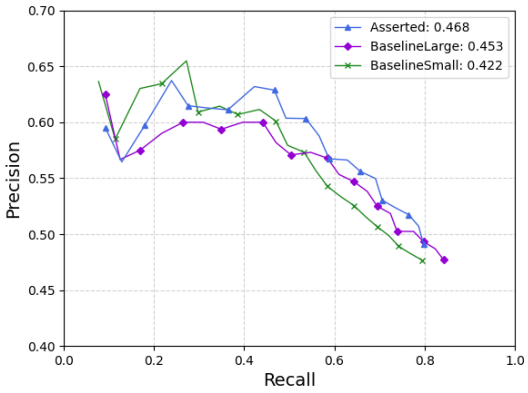

Figure 2: Precision/recall on Sports Entailment Dataset

5 Results

Table 2 contains area under the precision/recall curve (AUC) scores for Asserted, BaselineSmall, and BaselineLarge on the Levy/Holt dataset (**all** examples), the directional portion of the Levy/Holt dataset (2,414 examples), and the Sports Entailment dataset. The precision-recall curves for the Levy/Holt (all examples) and Sports Entailment datasets are displayed in Figures 1 and 2 respectively. Every point on the curve represents a different entailment score threshold (higher thresholds correspond to lower recall and vice versa). We follow the example of Hosseini et al. (2018) and compute AUC for precision in the range [0.5, 1]. All three Entailment Graphs cover this range and predictions with precision higher than random are important for downstream applications.

On the Levy/Holt dataset (all examples), BaselineLarge performs best overall. The strong performance of BaselineLarge compared to Asserted is in itself surprising, and indicates that it is usually not beneficial to distinguish modality when building Entailment Graphs. This can be understood as a data size issue: filtering out data is harmful as it introduces sparsity, and modal data is useful enough to provide a learning signal.

More counterintuitive, however, is that even BaselineSmall, which controls for training dataset size, outperforms Asserted. To understand why, we measured the size of each graph in terms of the number of nodes (predicates) and edges (entailment relations) it contained, and the percentage of predicates in the Levy/Holt dataset that were present in the graph (see Table 3). This revealed that BaselineSmall contained more of the predicates present in the Levy/Holt dataset, while also being larger in terms of both nodes and edges than Asserted. Thus, Asserted learns with more relations per predicate, while BaselineSmall has more predicate nodes overall. This may lead to the increase in recall that we see for the BaselineSmall graph.

Another explanation might be that this richer predicate coverage allows BaselineSmall to accurately correlate more of the common paraphrase examples in the Levy/Holt dataset. To this end we investigated the directional portion of the Levy/Holt dataset, which contains 2,414 examples of both the entailment pair and its reverse, where the entailment is true in one direction and false in the other. As noted by Hosseini et al. (2018) this task is much harder than that represented by the original dataset. However, the baselines both outperform the Asserted graph on the directional entailment task. We also observe a similar pattern in the percentage of predicates covered (see last column in Table 3). In general, we conclude that modal data is useful even for learning directional entailments.

Performance on the Sports Entailment dataset (Figure 2) reveals a different pattern. BaselineLarge outperforms BaselineSmall as expected, but Asserted performs best, despite lower coverage of the predicates in the Sports Entailment Dataset (see Table 4 for a size comparison of the ORGANISATIONs subgraph). This supports the suggestion by Guillou et al. (2020) that excluding modal data may help to avoid learning entailments between disjunctive outcomes, i.e. that winning entails losing, which is not measured by the Levy/Holt dataset.

6 Discussion & Future Work

Another appealing intuition for the usefulness of modal relations is that they might generally be expressed in text when the prior probability of the main predicate is already high. This would lead the distributions for the main predicates to be improved in spite of the uncertainty of the evidence. Additionally, if the probability of a premise is high enough to be worth mentioning, then in general that of its entailments will be too. However, this may not hold for the sports scenario because the outcomes are widely speculated upon despite being highly uncertain.

Indeed it is easy to find examples in the news corpus to support these intuitions. In the general domain we observe examples of eventualities initially being discussed with uncertainty, and later mentioned as asserted. An example of this is the acquisition of Dell by Michael Dell: on February 5th, 2013 we observe *"... founder and CEO Michael Dell and investment firm Silver Lake Partners will buy Dell."*, and subsequently, on February 6th, 2013 we read *"So Michael Dell and a private equity group have bought Dell and taken it private."*. We also observe the reverse scenario in the sports domain. For example, on January 10th, 2013 we observe *"The popular opinion on this game seems to be Seattle beating Atlanta because..."*, while shortly afterwards we are informed that *"Falcons come back to beat Seahawks"*. The latter is likely rather domain-specific, and we may expect to find a similar effect for other domains that share the disjunctive outcome property, for example elections, court cases and battles, where modals are used when speculating about potential and counterfactual outcomes.

We will explore ways to leverage this information and consider other sub-domains for which it is useful to retain or remove modal data. This may involve creating more domain-specific datasets. It is also worth investigating the effects of negation, which shares similar properties to modality, on learning Entailment Graphs.

Relatedly, we could retain predicates under specific modal modifiers, as these correspond to different prior probabilities of eventualities, carrying a different epistemic commitment from the writer. Eventualities that happen "undoubtedly" might be preferred over those that are "unlikely", for instance, and the modality parser can output specific categories of modality, allowing us to choose the subsets that should be kept.

Finally, we will experiment with learning Entailment Graphs with modal predicate nodes, by retaining modal relations with tags attached as input. Many of these entailments are trivial, because any entailment of a consequence can be reproduced under modal scope (if *buy* $\rightarrow$ *own*, then also *MOD_buy* $\rightarrow$ *MOD_own*). More notably, we might recover that following an entailment in the reverse direction can produce a modal entailment (e.g. if *beat* $\rightarrow$ *play*, then we know *play* $\rightarrow$ *MOD_beat*), and many preconditions will behave interestingly (e.g. *beat* $\rightarrow$ *play*, but also *MOD_beat* $\rightarrow$ *play*). To evaluate this idea, we will design a dataset of modal entailments, drawing inspiration from previous research on veridicality in entailment datasets (Staliūnaitė, 2018).

7 Conclusion

We have investigated the role of modally modified relations in Entailment Graph mining, and shown that, contrary to results from other tasks, uncertain predications actually constitute a valuable learning signal overall. Further analysis shows that there are specific predicate domains in which removing modal data is beneficial.

Acknowledgements

This work was funded by the ERC H2020 Advanced Fellowship GA 742137 SEMANTAX and a grant from The University of Edinburgh and Huawei Technologies.

The authors would like to thank Mohammad Javad Hosseini and Nick McKenna for helpful discussions, and the reviewers for their valuable feedback.

References

Heike Adel and Hinrich Schütze. 2017. Exploring different dimensions of attention for uncertainty detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 22–34, Valencia, Spain. Association for Computational Linguistics.

Kathryn Baker, Michael Bloodgood, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin, and Christine Piatko. 2010. A modality lexicon and its use in automatic tagging. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jonathan Berant, Noga Alon, Ido Dagan, and Jacob Goldberger. 2015. Efficient global learning of entailment graphs. *Computational Linguistics*, 41(2):221–263.

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.

Sander Bijl de Vroe, Liane Guillou, Miloš Stanojević, Nick McKenna, and Mark Steedman. 2021. Modality and negation in event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 31–42, Online. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.

Barbara Dancygier. 1998. *Conditionals and Prediction: Time, Knowledge and Causation in Conditional Constructions*, volume 87 of *Cambridge Studies in Linguistics*. Cambridge University Press.

Marie-Catherine De Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D Manning. 2006. Learning to distinguish valid textual entailments. In *Second Pascal RTE Challenge Workshop*, volume 62.

Kathrin Eichler, Aleksandra Gabryszak, and Günter Neumann. 2014. An analysis of textual inference in german customer emails. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 69–74.

Kathrin Eichler, Feiyu Xu, Hans Uszkoreit, and Sebastian Krause. 2017. Generating pattern-based entailment graphs for relation extraction. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 220–229, Vancouver, Canada. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task*, pages 1–12.

Liane Guillou, Sander Bijl de Vroe, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2020. Incorporating temporal information in entailment graph mining. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 60–71.

Xavier Holt. 2018. Probabilistic models of relational implication. Master's thesis, Macquarie University.

Mohammad Javad Hosseini, Nathanael Chambers, Siva Reddy, Xavier R. Holt, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2018. Learning typed entailment graphs with global soft constraints. *Transactions of the Association for Computational Linguistics*, 6:703–717.

Mohammad Javad Hosseini, Shay B. Cohen, Mark Johnson, and Mark Steedman. 2019. Duality of link prediction and entailment graph induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4736–4746, Florence, Italy. Association for Computational Linguistics.

Pierre-Antoine Jean, Sébastien Harispe, Sylvie Ranwez, Patrice Bellot, and Jacky Montmain. 2016. Uncertainty detection in natural language: A probabilistic model. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Lauri Karttunen and Annie Zaenen. 2005. Veridicity. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC bioinformatics*, 9(11):1–10.

Angelika Kratzer. 2012. *Modals and conditionals: New and revised perspectives*, volume 36. Oxford University Press.

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.

David Lewis. 1973. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, pages 418–446.

Dekang Lin and Patrick Pantel. 2001. DIRT: Discovery of Inference Rules from Text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, pages 323–328, New York, NY, USA. ACM Press.

Xiao Ling and Daniel S. Weld. 2012. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 94–100. AAAI Press.

Nick McKenna, Liane Guillou, Mohammad Javad Hosseini, Sander Bijl de Vroe, and Mark Steedman. 2021. Multivalent entailment graphs for question answering. *arXiv preprint arXiv:2104.07846*.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 992–999.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *Proceedings of the BioNLP 2009 workshop*, pages 28–36.

Roser Morante and Walter Daelemans. 2012a. Annotating modality and negation for a machine reading evaluation. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Roser Morante and Walter Daelemans. 2012b. Annotating modality and negation for a machine reading evaluation. In *CLEF (Online Working Notes/Labs/Workshop)*, pages 17–20.

Michael Nelson. 2019. Propositional Attitude Reports. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. Metaphysics Research Lab, Stanford University.

Dat Ba Nguyen, Johannes Hoffart, Martin Theobald, and Gerhard Weikum. 2014. Aida-light: High-throughput named-entity disambiguation. *Workshop on Linked Data on the Web*, 1184:1–10.

Marek Rei and Ted Briscoe. 2010. Combining manual rules and supervised learning for hedge cue and scope detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task*, pages 56–63.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Rion Snow, Lucy Vanderwende, and Arul Menezes. 2006. Effectively using syntax for recognizing false entailment. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, page 33–40. Association for Computational Linguistics.

Ieva R. Staliūnaitė. 2018. Learning about non-veridicality in textual entailment. Master's thesis, Utrecht University.

Miloš Stanojević and Mark Steedman. 2019. CCG parsing algorithm with incremental tree rotation. 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)*, pages 228–239, Minneapolis, Minnesota. Association for Computational Linguistics.

György Szarvas. 2008. Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT*, pages 281–289.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 849–856, Manchester, UK. Coling 2008 Organizing Committee.

Veronika Vincze. 2014. *Uncertainty detection in natural language texts*. Ph.D. thesis, University of Szeged.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 81–88.

Congle Zhang and Daniel S Weld. 2013. Harvesting parallel news streams to generate paraphrases of event relations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786.

A Experimental Settings / Requirements

When using MoNTEE to extract relations we used the default settings, with the exception of disabling unary relation extraction (writeUnaryRels=False) and restricting binary relations to those that include at least one named entity (acceptGGBinary=False). When using entGraph to construct Entailment Graphs we raised the threshold values for infrequent predicates (minPredForArgPair=4) and argument pairs (minArgPairForPred=4), and used the default values for all other parameters.

All experiments were conducted on a single server with 330GB RAM, and two Intel Xeon E5-2697 v4 2.3GHz CPUs (each with 18 cores). The computational cost of training a single Entailment Graph is approximately one day for the local learning step, and eight hours for globalisation.