KEViN: A Knowledge Enhanced Validity and Novelty Classifier for Arguments

Ameer Saadat-Yazdi[1] and **Xue Li**[1] and **Sandrine Chausson**[1]
Vaishak Belle[1] and **Björn Ross**[1] and **Jeff Z. Pan**[1,2] and **Nadin Kökciyan**[1]
[1]School of Informatics, University of Edinburgh
[2] Huawei Edinburgh Centre, CSI, Huawei
(ameer.saadat, xue.shirley.li, sandrine.chausson,
vbelle, b.ross, j.z.pan, nadin.kokciyan)@ed.ac.uk

Abstract

The ArgMining 2022 Shared Task is concerned with predicting the validity and novelty of an inference for a given premise and conclusion pair. We propose two feed-forward network based models ($\mathcal{KEV}i\mathcal{N}_1$ and $\mathcal{KEV}i\mathcal{N}_2$), which combine features generated from several pre-trained transformers and the WikiData knowledge graph. The transformers are used to predict entailment and semantic similarity, while WikiData is used to provide a semantic measure between concepts in the premise-conclusion pair. Our proposed models show significant improvement over RoBERTa, with $\mathcal{KEV}i\mathcal{N}_1$ outperforming $\mathcal{KEV}i\mathcal{N}_2$ and obtaining second rank on both subtasks (A and B) of the ArgMining 2022 Shared Task.

1 Introduction

A number of frameworks have been proposed to evaluate the quality of natural language arguments. Many of these frameworks consider some notion of logical soundness (validity), (Wachsmuth et al., 2017). The ArgMining 2022 shared task also highlights the importance of novelty in measuring the usefulness of a conclusion in order to avoid redundant or non-informative conclusions. These metrics were more formally introduced in (Opitz et al., 2021) to assess the quality of arguments.

In our work, we combine the power of pre-trained language models with external knowledge sources to provide additional information for predictions (Wang et al., 2019; Pan et al., 2019). For this, we extract paths from WikiData (Vrandečić and Krötzsch, 2014) that link the premise to the conclusion, and generate numerical features from these paths.

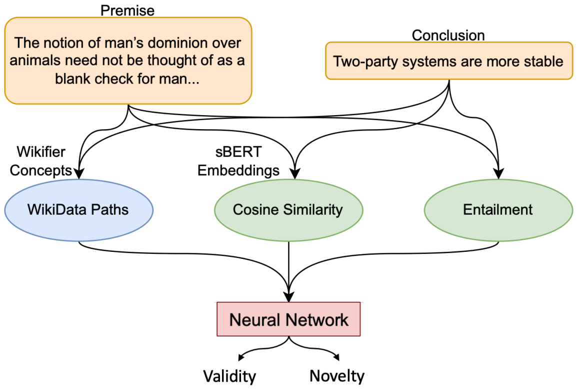
Premise
The notion of man's dominion over animals need not be thought of as a blank check for man...
Conclusion
Two-party systems are more stable
Wikifier Concepts
sBERT Embeddings
WikiData Paths
Cosine Similarity
Entailment
Neural Network
Validity
Novelty

Figure 1: $\mathcal{KEV}i\mathcal{N}_1$ uses WikiData knowledge, and pre-trained transformers to predict similarity and entailment prior to feeding the data into the neural network.

Having generated several sets of features using WikiData and pre-trained models, we proceed to use these features as inputs to a small feed-forward network trained to predict validity and novelty. Our results show a significant improvement from simply fine-tuning a pre-trained model on the task, and we identify that textual entailment serves as a strong indicator of argument validity, while a combination of textual entailment and knowledge graph distance serves to improve the model's ability to detect novelty[1].

We trained and tested two versions of our model on Task A (binary classification). The first model, $\mathcal{KEV}i\mathcal{N}_1$, predicts both validity and novelty using the same network (Figure 1). The second model, $\mathcal{KEV}i\mathcal{N}_2$, uses two separate networks, which were trained separately for each label, and then combines their predictions. Both models show significant improvement over RoBERTa with $\mathcal{KEV}i\mathcal{N}_1$ significantly outperforming $\mathcal{KEV}i\mathcal{N}_2$. We additionally evaluated the $\mathcal{KEV}i\mathcal{N}_1$ model trained on Task A on the testing set of Task B, the corresponding details and results are given in the Appendix.

[1] Our code is available on GitLab: https://git.ecdf.ed.ac.uk/xli3310/KEViN_2022.

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 104–110
October 17, 2022.

Split	Val	¬Val	Nov	¬Nov	Total
train	401	320	123	595	750
dev	125	74	82	118	202
test	314	206	226	294	520

Table 1: Data split from Task A. Note that the training and development sets also contain ambiguous examples that are neither valid/novel nor non-valid/-novel.

2 Task and Data Description

The ArgMining 2022 Shared Task consists of two subtasks: for a given textual premise, 1) classifying a conclusion as being valid/novel (Task A) and 2) comparing conclusions in terms of validity/novelty (Task B). For both tasks, the premise consists of multiple sentences while the conclusion is a single statement. *Validity* requires that there exists a sound logical inference linking the premise to the conclusion; *Novelty*, on the other hand, requires the conclusion to contain new information compared to the premise, and as such to be more than a simple paraphrase[2]. Table 1 provides the class counts per data split for Task A, while Table 2 provides examples of premise/conclusions pairs to illustrate the concepts of validity and novelty.

3 Feature Extraction

In this section, we explain the process for extracting the features that were used as input to our feed-forward network.

3.1 Neural Features

A subset of the features we used as input for our classifier were extracted using large pre-trained neural networks.

3.1.1 Textual Entailment

Given a text *t* and hypothesis *h*, the task of Recognising Textual Entailment (RTE), also called Natural Language Inference (NLI), consists of determining whether *t* entails *h* ("Entailment" class), contradicts *h* ("Contradiction" class), or neither ("Neutral" class) (Zeng et al., 2021).

We used a BART model (Lewis et al., 2020) fine-tuned on the Multi-genre Natural Language Inference dataset (Williams et al., 2018) to predict the textual entailment between each premise/conclusion input pair. We did this first with the premise as the text *t* and the conclusion as the hypothesis *h* (**TE_P2C**), and then the other way around (**TE_C2P**). In each case, the model returns the probability distribution for the three entailment classes; i.e., a vector of three real numbers adding up to 1. We chose the BART_MNLI model[3] to extract the entailment features because of its state-of-the-art performance on the RTE/NLI task (Yin et al., 2019).

3.1.2 Cosine Similarity

For each premise/conclusion input pair, we used the SBERT package (Reimers and Gurevych, 2019) to obtain $\vec{p}$ and $\vec{c}$, the vector representations of the premise and conclusion respectively. To measure the similarity between these two vectors, we calculated their cosine similarity (**CoSim**), as defined by the following equation:

$$cos(\vec{p}, \vec{c}) = \frac{\vec{p} \cdot \vec{c}}{||\vec{p}||||\vec{c}||} \tag{1}$$

3.1.3 BERT Predictions

We trained two separate BERT models to predict Validity and Novelty on the training set. The probabilities of validity (**BERT_pred_val**) and novelty (**BERT_pred_nov**) were used as additional neural input features.

3.2 Knowledge Graph Features

Knowledge Graphs (KGs) represent knowledge in a graph-based structure, in which nodes represent entities and edges represent relations connecting them. Within the KG formalism, the connection between two entities is denoted as the triple $\langle s, r, o \rangle$, where s, r and o represent the subject, relation and object, respectively.

KGs have many applications, including query answering (Huang et al., 2019; Yasunaga et al., 2021) and modelling 5G networks (Zhu et al., 2022; Wang et al., 2021). In this paper, we chose to work with WikiData, one of the biggest KGs in the literature (Vrandečić and Krötzsch, 2014), to extract KG features that can assist the validity/novelty classification of a conclusion c for a given premise p.

To obtain our KG features, we first extracted WikiData entities from p and c, respectively. We tested two entity extraction tools, Wikifier (Brank et al., 2017) and Falcon2.0 (Sakor et al., 2020). Both performed similarly for our task, however we

[2] https://phhei.github.io/ArgsValidNovel/

[3] https://huggingface.co/facebook/bart-large-mnli

Premise	The notion of man’s dominion over animals need not be thought of as a blank check for man to exploit animals. Indeed, it may be appropriate to connect the notion of “dominion” to stewardship” over animals. Yet, humans can be good stewards of animals while continuing to eat them. It is merely necessary that humans maintain balance, order, and sustainability in the animal kingdom. But, again, this does not require the abandonment of meat-eating.		
		Valid?	**Novel?**
Conclusion	Two-party systems are more stable	no	no
	Man’s “dominion” over animals does not imply abandoning meat.	yes	no
	The idea of “domiminism” is unnecessary.	no	yes
	Dominion over animals can and should be used responsibly.	yes	yes

Table 2: Example from Task A on the topic of *Vegetarianism*.

chose Wikifier for its convenient interface. Our entity extractions from p and c are written as $p \mapsto \mathbb{E}_p$ and $c \mapsto \mathbb{E}_c$, where $\mathbb{E}_p$ and $\mathbb{E}_c$ are sets of WikiData entity IDs, respectively. Having done this, we then identified the Knowledge Graph Paths connecting entities from the premise to entities from the conclusion.

Definition 3.1 (Knowledge Graph Path (KGP)) *Given a pair (e_h, e_t) in the KG, their KGP, $\mathbb{K}(e_h, e_t)$, is defined as:*

- *$\emptyset$, if e_h and e_t are disconnected;*
- *$\{\langle e_h \rangle\}$, if $e_h = e_t$;*
- *$\{\langle e_h, r_1, x_1\rangle, \ \langle x_1, r_2, x_2\rangle, ..., \langle x_n, r_n, e_t\rangle\}$, a set of n triples where the object of the former triple is the subject of the following triple, otherwise.*

Multiple KGPs can exist for a single pair of entities. Moreover, there is no guarantee for a KGP to be finite. For our task, we aimed to find the shortest KGPs with a limit. Our search of KGP over WikiData is based on SPARQL queries (Pérez et al., 2009), for which breadth-first search (BFS) was the easiest to implement. To reduce the search space of KGP, we applied an interactive depth limit to the BFS algorithm with a termination depth limit D equal to 3.[4] As a result, the search terminates with the shortest KGPs whose length is less or equal to 3, or with failure if no such path is found.

Some relations denote extremely close proximity, e.g. ‘same as’, while others the opposite, e.g. ‘different from’. These two kinds of extreme relations are summarised $\mathbb{L}_1$ and $\mathbb{L}_2$, respectively. Both sets are given in the Appendix. Based on our test, the extreme relations in $\mathbb{L}_1$ and $\mathbb{L}_2$ make KGPs less representative in our tasks. We compute the semantic length between two entities e_1 and e_2, ($\mathcal{L}_s(e_1, e_2)$) as defined in Equation 2.

$$\mathcal{L}_s = \begin{cases} 0, \ \mathbb{K}(e_1, e_2) = \{e_1\} \bigvee \forall \langle s, r, o \rangle \in \mathbb{K}(e_1, e_2), r \in \mathbb{L}_1 \\ D + 1, \ \mathbb{K}(e_1, e_2) = \emptyset \bigvee \exists \langle s, r, o \rangle \in \mathbb{K}(e_1, e_2) \wedge r \in \mathbb{L}_2 \\ |\{\langle s, r, o \rangle \mid \langle s, r, o \rangle \in \mathbb{K}(e_1, e_2) \wedge r \notin \mathbb{L}_1\}|, \ otherwise \end{cases} \quad (2)$$

Finally, we compute the final KG features, i.e. **Irrelevancy** and **Avg_Dist**, as shown below, where $p \mapsto \mathbb{E}_p, \ c \mapsto \mathbb{E}_c, \ e_p \in \mathbb{E}_p$ and $\ e_c \in \mathbb{E}_c$.

1. **Irrelevancy**: the number of conclusion entities e_c that are disconnected from all premise entities e_p:

$$\mathcal{I} = |\{e_c | \forall e_p, \ \mathbb{K}(e_p, e_c) = \emptyset\}| \quad (3)$$

2. **Avg_Dist**: the average minimal distance between premise entities $\mathbb{E}_p$ to conclusion entities $\mathbb{E}_c$, based on the semantic length ($\mathcal{L}_s$) of all possible pairs of entities from the premise and the conclusion.

$$\mathcal{A} = \frac{\sum_{e_p, e_c} \min |\mathcal{L}_s(e_p, e_c)|}{|\mathbb{E}_p| \times |\mathbb{E}_c|} \quad (4)$$

These two KG features were shown to be significant for our task. Other KG features that we experimented with but that were not as useful are given in the Appendix.

4 Preprocessing and Training

Once the features were computed, we applied several preprocessing steps to improve results. Given the distribution shift between the training and development data of Task A, as shown in Table 1, we used a simple upsampling strategy to ensure that all classes ($Val\&Nov$, $Val\&\neg Nov$, $\neg Val\&Nov$, $\neg Val\&\neg Nov$) were relatively balanced. To do this, we duplicated 200 $Val\&Nov$ examples and 250 $\neg Val\&\neg Nov$ examples so that we would have roughly 300 samples for each class. We also duplicated ambiguous examples, such that if a sample

[4] We choose 3 as the depth limit, because helpful KG features can be found under that limit and the program terminates within reasonable time.

Model	Precision	Recall	F1
RoBERTa	0.21	0.26	0.21
$\mathcal{KEV}i\mathcal{N}_1$	0.44	0.43	0.43
$\mathcal{KEV}i\mathcal{N}_2$	0.41	0.40	0.40

Table 3: Performance of the $\mathcal{KEV}i\mathcal{N}_1$, $\mathcal{KEV}i\mathcal{N}_2$ and a fine-tuned RoBERTa model on the test set for the combined task of validity and novelty prediction.

Model	Validity	Novelty
$\mathcal{KEV}i\mathcal{N}_1$	0.70	0.62
$\mathcal{KEV}i\mathcal{N}_2$	0.67	0.62

Table 4: F1 scores of models on task A, broken down by validity and novelty.

has ambiguous validity, it would appear once as valid and once as invalid, and likewise for novelty. Finally, MinMax scaling was applied to each feature across the training, development, and test sets to ensure that all values were between 0 and 1.

The features were then concatenated and input to a small neural network with two hidden layers of widths five and two respectively and a softmax output layer. We used the Adam optimizer (Kingma and Ba, 2014) with a constant learning rate of 0.001 with L2 regularization. To optimize the regularization term and find the best combination of features for the task we performed an exhaustive grid search using L2 parameters $\alpha \in \{0.001, 0.01, 0.1, 1.0\}$ and all possible combinations of features with a limit of five features at most.

5 Results and Analysis

Table 3 shows the performance of $\mathcal{KEV}i\mathcal{N}_1$, $\mathcal{KEV}i\mathcal{N}_2$ and a RoBERTa baseline on the test set. All models were trained on the same upsampled version of the train set. The RoBERTa baseline was fine-tuned over 3 epochs: the checkpoint that minimised loss on the dev set, obtained after the first epoch, was used for the evaluation. We see clearly that $\mathcal{KEV}i\mathcal{N}_1$ outperforms the RoBERTa baseline and the model with two independent classifiers both in precision and recall. The increase in performance mostly affects the predicition of validity as shown in Table 4.

For $\mathcal{KEV}i\mathcal{N}_1$, our validation run identified **irrelevancy**, **average KGP distance**, **TE_P2C** and **TE_C2P** as the set of features leading to the best performance. We performed an ablation study to identify the relative contribution of each feature. Table 5 shows that removing **TE_P2C** or **Irrelevancy** has the most significant impact overall. We also see that the neural features play a more important role than KG features, especially for validity classification. The results suggest that neural features are crucial to improve the model performance for the combined task of validity and novelty prediction. While KG features are also useful for detecting validity, removing them particularly harms the novelty detection. We expect this result since the existence of KGPs and their lengths reflect the semantic relatedness between the premise and the conclusion, which is relevant for novelty detection.

Removed Feature	Val	Nov	Both
TE_P2C	0.65	0.45	0.31
TE_C2P	0.67	0.61	0.39
Avg_Dist	0.59	0.59	0.40
Irrelevancy	0.67	0.57	0.35
Neural features	0.54	0.54	0.26
KG features	0.68	0.59	0.37

Table 5: Ablation study on test set showing the F1 score of $\mathcal{KEV}i\mathcal{N}_1$ when a given feature is removed. The F1 scores of $\mathcal{KEV}i\mathcal{N}_1$ for validity, novelty and the combined task are 0.70, 0.62 and 0.43, respectively. Colours represent the relative performance decrease with respect to the original $\mathcal{KEV}i\mathcal{N}_1$ model (as a percentage).

6 Related Work

This work identifies a strong similarity between argument validity and textual entailment which has been previously explored (Cabrio and Villata, 2012; Bosc et al., 2016) with mixed success for argument mining. Likewise, the introduction of external KGs into the argument mining pipeline has been studied by Fromm et al. (2019); Paul et al. (2020); Li et al. (2021). In the textual entailment literature, a substantial amount of work has shown the importance of external KGs in making accurate inferences over new domains (Wang et al., 2019, 2020).

7 Conclusion

In this paper, we have shown how features can be extracted from knowledge graphs and pre-trained neural networks that are both relevant and complementary for the task of argument novelty and validity detection. We did this by demonstrating how a small neural network trained on these features outperforms fine-tuning with large transformers.

We defined KG paths in terms of their semantic length and the corresponding KG distance features, which gave promising results and provides a basis for future work. For example, we would also like to consider semantic representations of paths, such as natural language representations, vector-based KG embedding approaches, and other KGs to improve the performance of the proposed model. In addition, it would be interesting to see if learning weights to predict the semantic length based on the relations in KG paths or if extending the graph containing the premise and conclusion concepts with semantic dependency relations would boost the performance.

Acknowledgements

This work was supported by the ELIAI (Edinburgh Laboratory for Integrated Artificial Intelligence) EPSRC (grant no EP/W002876/1); the UKRI (grant EP/S022481/1) and the University of Edinburgh; the Edinburgh-Huawei Joint Lab and Huawei's grant CIENG4721/LSC. We appreciate the time given by the shared task reviewers for their useful feedback, which contributed to improvements in our paper.

References

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media. *Computational Models of Argument*, pages 21–32. Publisher: IOS Press.

Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*, 472.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *ECAI*.

Michael Fromm, Evgeniy Faerman, and Thomas Seidl. 2019. TACAM: Topic And Context Aware Argument Mining. *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 99–106. ArXiv: 1906.00923.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 105–113.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Weichen Li, Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Topic-guided knowledge graph construction for argument mining. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 315–322.

Juri Opitz, Philipp Heinisch, Philipp Wiesenbach, Philipp Cimiano, and Anette Frank. 2021. Explainable unsupervised argument similarity rating with Abstract Meaning Representation and conclusion generation. In *Proceedings of the 8th Workshop on Argument Mining*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie, and Dong Yu. 2019. Improving question answering with external knowledge. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 27–37, Hong Kong, China. Association for Computational Linguistics.

Debjit Paul, Juri Opitz, Maria Becker, Jonathan Kobbe, Graeme Hirst, and Anette Frank. 2020. Argumentative Relation Classification with Background Knowledge. *Computational Models of Argument*, pages 319–330. Publisher: IOS Press.

Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. 2009. Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, 34(3):1–45.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3141–3148.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Fangrong Wang, Alan Bundy, Xue Li, Ruiqi Zhu, Kwabena Nuamah, Lei Xu, Stefano Mauceri, and Jeff Z Pan. 2021. Lekg: A system for constructing knowledge graphs from log extraction. In *The 10th International Joint Conference on Knowledge Graphs*, pages 181–185.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7208–7215. Number: 01.

Zikang Wang, Linjing Li, and Daniel Zeng. 2020. Knowledge-Enhanced Natural Language Inference Based on Knowledge Graphs. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6498–6508, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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)*, pages 1112–1122. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qagnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Ricky Zhu, Xue Li, Sylvia Wang, Alan Bundy, Jeff Z Pan, Kwabena Nuamah, Stefano Mauceri, and Lei Xu. 2022. Treat: Automated construction and maintenance of probabilistic knowledge bases from logs. In *The 8th Annual Conference on Machine Learning, Optimization and Data Science*.

Appendix

The KG Features

Several KG features not described in the main body of this paper were also tested in our experiments. We provide their definition here for the record.

- **Max_Dist**: the length of the longest KGP from any e_p to any e_c.

$$\mathbb{K}_{max} = \max_{e_p,\, e_c} \mathcal{L}_s(e_p, e_c) \tag{5}$$

- **Total_Dist**: the sum of all KGPs from premise entities to conclusion entities.

$$S(t, c) = \sum_{e_p, e_c} \mathcal{L}_s(e_p, e_c) \tag{6}$$

- **Minimal KGP (MKGP):** the shortest KGPs from one entity e_c to a set of entities $\mathbb{T}$.

$$\mathbb{P}_{min}(h, \mathbb{T}) = \min_{e_t, e_c} \mathcal{L}_s(e_p, e_c) \tag{7}$$

The **Dist_max** feature was useful but was not selected as one of the optimal features during grid-search. The **Dist_max** and **MKGP** features, on the other hand, proved unhelpful for this task.

KG Relations

A set of common logical relations are summarised in Table 6, where the last two are used to invalidate a KGP. Relations not included in the list are omitted when calculating the semantic length of a KGP.

ID	Label	Type
P31	instance of	$\mathbb{L}_1$
P279	subclass of	$\mathbb{L}_1$
P527	has part(s)	$\mathbb{L}_1$
P361	part of	$\mathbb{L}_1$
P463	member of	$\mathbb{L}_1$
P1269	facet of	$\mathbb{L}_1$
P355	has subsidiary	$\mathbb{L}_1$
P460	said to be the same as	$\mathbb{L}_1$
P642	of	$\mathbb{L}_1$
P1889	different from	$\mathbb{L}_2$
P461	opposite of	$\mathbb{L}_2$

Table 6: A set of logical relations wither their WikiData IDs and type, where $\mathbb{L}_1$ represents semantic similarity and $\mathbb{L}_2$ represents semantic distance.

Task B (Comparitive Predictions)

In Task B, two conclusions are given for a single premise and the objective is to decide whether the first conclusion is more, less, or equally valid/novel as as the second. In both these tasks the model should output two labels, one for validity and one for novelty.

We approached this task by using the best model trained on Task A to predict the probability of novelty and validity of both conclusions. We then assigned the conclusion with the highest probability of validity/novelty as that which is more valid/novel. The results are given in Table 7.

	Validity			**Novelty**		
	$\mathbf{-1}$	**0**	**1**	$\mathbf{-1}$	**0**	**1**
Precision	0.39	0.00	0.38	0.26	0.00	0.32
Recall	0.67	0.00	0.66	0.64	0.00	0.57
F1	0.49	0.00	0.48	0.37	0.00	0.41
			Both (Macro)			
Precision			0.07			
Recall			0.19			
F1			0.09			

Table 7: Results of our best model on Task B.

The results show that this simple approach to Task B fails to identify cases where the two conclusions are equally valid/novel (classes 0). This can be explained by the fact that the classifier outputs continuous probabilities, which span the entire 0 to 1 range. As such, requiring both probabilities to be equal for the two conclusions to be considered equally valid/movel is an excessively stringent requirement. A better approach might require the difference between the two probabilities not to exceed a given threshold. This threshold could for instance be found using the training set provided for Task B.