

ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis

Qingyun Wang¹, Qi Zeng¹, Lifu Huang¹,
Kevin Knight², Heng Ji¹ and Nazneen Fatema Rajani³

¹ University of Illinois at Urbana-Champaign

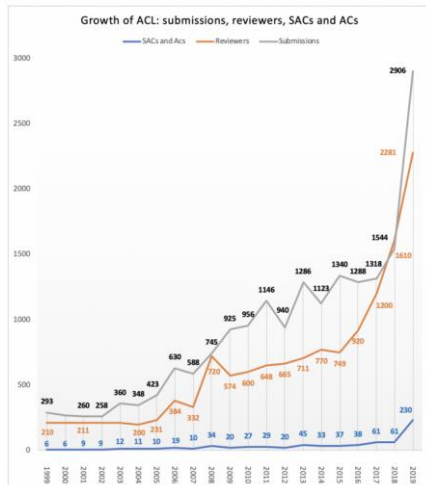
²DiDi Labs and University of Southern California

³Salesforce Research



Background

- Rapid growth of paper submission
 - The number of papers submitted for ACL 2020 reached a record for 3,429
 - ICLR 2021 has 3,000+ submitted papers, which is doubled compared to ICLR 2019
- Bad review quality
 - Due to the rapid growth of submission, some paper submissions lack qualified review comments



Gui Lohmann
November 25 at 8:06 PM · 🌐

When you had enough of Reviewer 2 attitude and decide to mention the "Reviewer 2 Must Be Stopped!" Facebook group in your response letter to the editor/reviewers... #hadenough

bibliography from different indexed journals, reports and local press. Additional reference on policy fragmentation in transport infrastructures can be ref. 1-2-3 here below; ref. 4 then further rebates the aspect of the crescent importance of cruise tourism worldwide. Finally, authors can add some more bibliography related to the topic of involving communities and stakeholders from the earliest phases of new project and development to ease the process and its effects on reducing demonstration and opposition from the citizenship. Ref. n. 3 and its related bibliography can be helpful with regard to this topic."

Thank you for your comments. We understand the suggestion made here in terms of references. However, if the point is to show evidence of the "importance of cruise tourism worldwide", we believe there are better references than the Ref. n. 4 proposed here. The request to include Ref. n. 4 might be an opportunity to increase citations for the authors? — on an anecdotal way, there is a brilliant Facebook group "Reviewer 2 Must Be Stopped!" (https://www.facebook.com/groups/reviewer2) which I fully encourage the reviewer to join — yhe might raise that various colleagues actually mock academic behaviours like this I hope you take this suggestion in high spirits...

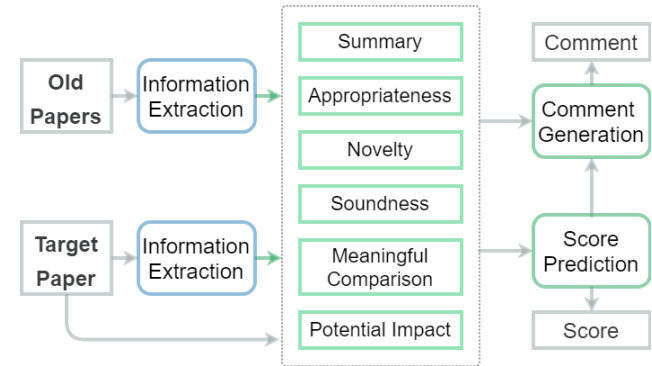
1. Isaksson K, Antonson H, Eriksson L (2017) Layering and parallel policy making - complementary concepts for understanding implementation challenges to sustainable mobility. Transp Policy 53:50-57
2. Bulckaen J, Keseru I, Macharis C (2016) Sustainability versus stakeholder preferences: searching for synergies in

216
23 Comments 4 Shares



Motivation

- How do humans review a paper?
 - Have sufficient background knowledge about the topic
 - Understand knowledge of the current paper
 - Compare the current paper with previous paper
 - Give score and review comments for each review category
- Can machine repeat such a process?
 - Build large-scale background KG from previous papers
 - Construct KG for related work section and other sections
 - Extract evidence based on the difference of KGs and corresponding paper sentences
 - Predict category scores and corresponding review comments based on the extracted evidence



Dataset Construction

- Based on PeerRead^[1], we further collect additional paper-review pairs from Openreview and NeurIPS

Conference	Year							
	2013	2014	2015	2016	2017	2018	2019	2020
ICLR	-	-	-	-	404	874	1,342	2,067
NeurIPS	342	399	389	545	655	963	-	-
ACL	-	-	-	-	130	-	-	-

- We construct the background KG from 174,165 papers from the open research corpus^[2]

Years (1965~)	2011	2012	2013	2014	2015	2016	2017	2018	2019
# of Entities	535,075	585,321	628,713	683,686	737,878	801,740	870,992	950,457	1,008,955
# of Relations	160,123	175,780	188,876	205,898	222,592	242,312	263,827	288,805	307,636

[1] Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1647–1661).

[2] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., ... Etzioni, O. (2018). Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)* (pp. 84–91).



Review Category

- We adopt the following most common categories from NeurIPS2019^[1] and PeerRead^[2]:
 - **Summary:** What is this paper about?
 - **Appropriateness:** Does the paper fit in the venue?
 - **Clarity:** Is it clear what was done and why? Is the paper well-written and well-structured?
 - **Novelty:** Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
 - **Soundness:** Can one trust the empirical claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?
 - **Meaningful Comparison:** Do the authors make clear where the problems and methods sit with respect to existing literature? Are the references adequate?
 - **Potential Impact:** How significant is the work described? If the ideas are novel, will they also be useful or inspirational? Does the paper bring any new insights into the nature of the problem?

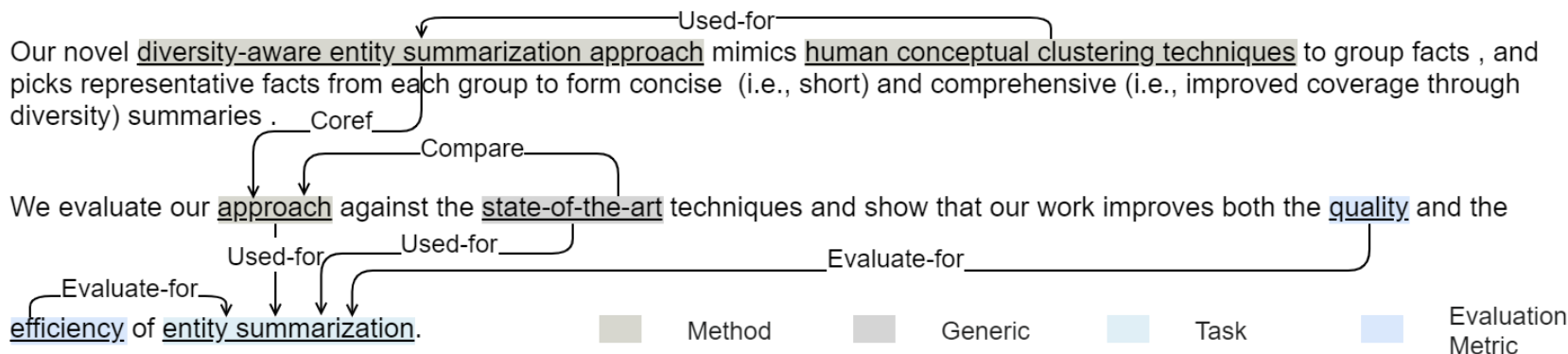
[1] <https://nips.cc/Conferences/2019/PaperInformation/ReviewerGuidelines>

[2] Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1647–1661).



Knowledge Graph Construction

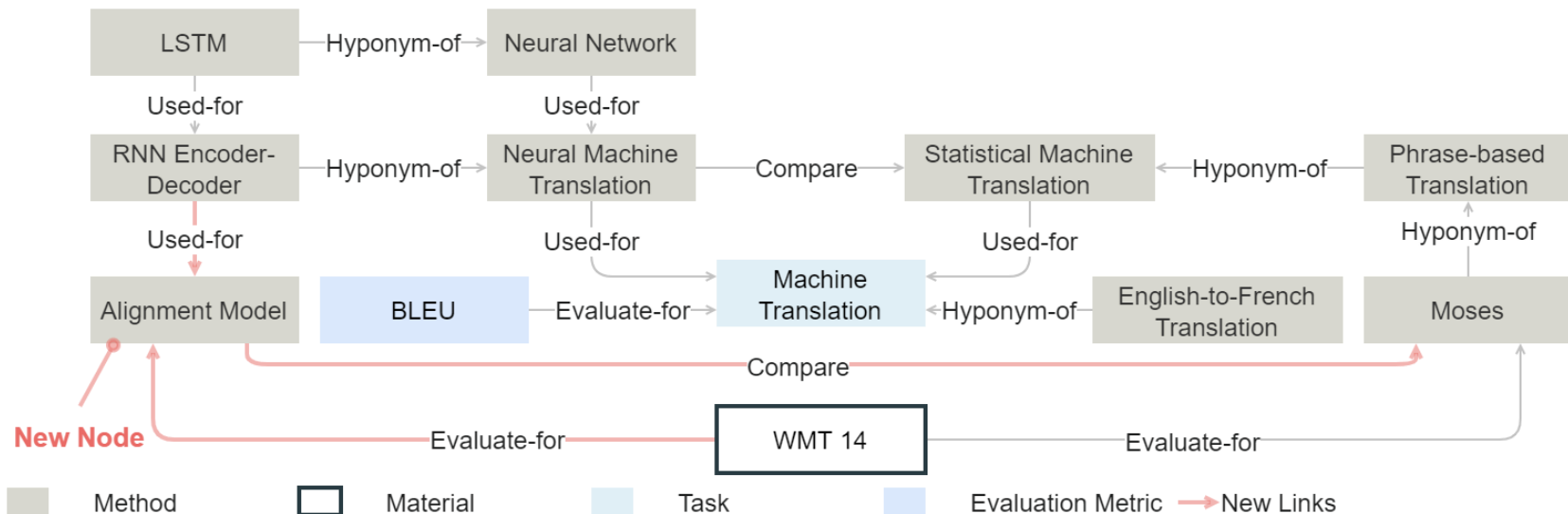
- Given the target paper under review P_τ , we first construct the following knowledge graphs using a state-of-the-art Information Extraction (IE) system for Natural Language Processing (NLP) and Machine Learning (ML) domains:
 - G_{P_τ} : A KG constructed from the abstract and conclusion sections of a target paper P_τ , which describes the main techniques
 - $\overline{G_{P_\tau}}$: A KG constructed from the related work section of P_τ , which describes related techniques.
 - G_B : A background KG constructed from all of the old NLP/ML papers published before the publication year of P_τ , in order to teach ReviewRobot what's happening in the field.



[1] Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings the 2018 Conference on Empirical Methods in Natural Language Processing, Volume 1 (Long Papers)* (pp. 3219–3232).



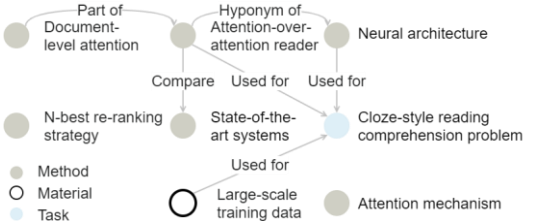

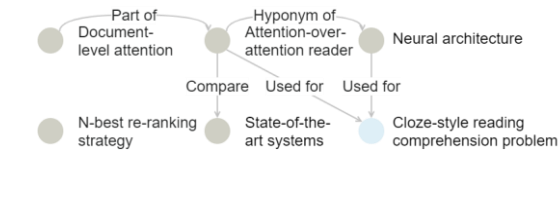
Knowledge Graph Construction



[1] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 5th International Conference on Learning Representations*.



Evidence Extraction for an Example Paper^[1]

Category	Evidence	Example
Summary	<ul style="list-style-type: none"> • G_{P_T} 	 <p>The diagram shows a network of entities and their relationships. Entities are represented by circles: grey for Method, white for Material, and light blue for Task. Relationships are shown as arrows with labels: 'Part of', 'Hyponym of', 'Compare', 'Used for', and 'Used for'. For example, 'Document-level attention' is a part of 'Attention-over-attention reader', which is a hyponym of 'Neural architecture'. 'Attention-over-attention reader' is compared to 'State-of-the-art systems' and used for 'Cloze-style reading comprehension problem'. 'Neural architecture' is used for 'Cloze-style reading comprehension problem'. 'State-of-the-art systems' is used for 'Large-scale training data' and 'Attention mechanism'. 'Large-scale training data' is used for 'Attention mechanism'.</p>
Appropriateness	<ul style="list-style-type: none"> • The number of entities overlapped between the target paper and the domain's background KG: $\nu \nu \in G_{P_T} \cap G_B$ • Abstract 	 <p>Legend for the knowledge graph diagram:</p> <ul style="list-style-type: none"> Grey circle: Neural architecture Light blue circle: Cloze-style reading comprehension problem White circle: Large-scale training data Grey circle: Attention mechanism
Novelty	<ul style="list-style-type: none"> • New knowledge elements that appear in the target paper but not in the background KG: $G_{P_T} - G_B$ • Paper sentences that contain new knowledge elements 	 <p>The diagram shows a network of entities and their relationships. Entities are represented by circles: grey for Method, white for Material, and light blue for Task. Relationships are shown as arrows with labels: 'Part of', 'Hyponym of', 'Compare', 'Used for', and 'Used for'. For example, 'Document-level attention' is a part of 'Attention-over-attention reader', which is a hyponym of 'Neural architecture'. 'Attention-over-attention reader' is compared to 'State-of-the-art systems' and used for 'Cloze-style reading comprehension problem'. 'Neural architecture' is used for 'Cloze-style reading comprehension problem'. 'State-of-the-art systems' is used for 'Large-scale training data' and 'Attention mechanism'. 'Large-scale training data' is used for 'Attention mechanism'.</p>

[1] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-Attention Neural Networks for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 593–602).



Evidence Extraction for an Example Paper^[1]

Category	Evidence	Example
Soundness	<ul style="list-style-type: none"> The number of knowledge elements that appear in the contribution claims in the introduction section and that are verified in the experiment section Abstract 	<ul style="list-style-type: none"> attention-over-attention reader, n-best re-ranking, strategy is verified in the related work section
Meaningful Comparison	<ul style="list-style-type: none"> The number of papers about relevant knowledge elements which are missed in the related work section: $G_{P_\tau} \cap G_B - \overline{G_{P_\tau}}$ The number of papers about relevant knowledge elements which are claimed new in the related work section: $G_{P_\tau} \cap G_B \cap \overline{G_{P_\tau}}$ The description sentences about comparison with related work If the related work section is not available, we use the difference between G_{P_τ} and G_B instead 	<pre> graph TD A((Large-scale training data)) -- Used for --> B((Cloze-style reading comprehension problem)) A -- Used for --> C((Attention mechanism)) D((Neural architecture)) -- Used for --> B </pre> <p>[2],[3]</p>

[1] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-Attention Neural Networks for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 593–602).

[2] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 5th International Conference on Learning Representations*.

[3] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems28* (pp. 1693–1701).



Evidence Extraction for an Example Paper^[1]

Category	Evidence	Example
Potential Impact	<ul style="list-style-type: none">• The number of new knowledge elements in the future work section• The number of new software, systems, data sets, and other resources	<ul style="list-style-type: none">• 5 new knowledge elements• 1 new architecture
Overall Recommendations	<ul style="list-style-type: none">• All features mentioned in the above categories• Abstract	

[1] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-Attention Neural Networks for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 593–602).



Score Prediction

- We first encode its category related sentences with an attentional Gated Recurrent Unit (GRU)^[1] to obtain attentional contextual sentence embedding.
- We also encode the extracted evidence for each review category with an embedding layer.
- Then we concatenate the context embedding and evidence embedding to predict the quality score r in the range of 1 to 5 with a linear output layer. We use the prediction probability as the confidence score.

[1] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 5th International Conference on Learning Representations*.



Category Annotation

- We manually annotate 200 paper-review pairs from ACL2017 and ICIR2017 datasets to select those constructive and informative human review sentences that are supported by certain evidence in the papers.

Category	# of Pairs	Evidence Sentence in Paper	Corresponding Review Sentence
Summary	236	In this paper, we present a simple but novel model called attention-over-attention reader for better solving cloze-style reading comprehension task.	The paper describes a new method called attention-over-attention for reading comprehension .
Novelty	33	The paper presents a new framework to solve the SR problem - amortized MAP inference and adopts a pre-learned affine projection layer to ensure the output is consistent with LR.	It introduces a novel neural network architecture that performs a projection to the affine subspace of valid SR solutions ensuring that the high resolution output of the network is always consistent with the low resolution input.
Soundness	174	In high dimensions we empirically found that the GAN based approach, AffGAN produced the most visually appealing results.	Combined with GAN , this framework can obtain plausible and good results.
Meaningful Comparison	16	As a concrete instantiation, we show in this paper that we can enable recursive neural programs in the NPI model, and thus enable perfectly generalizable neural programs for tasks such as sorting where the original, non-recursive NPI program fails.	This paper improves significantly upon the original NPI work, showing that the model generalizes far better when trained on traces in recursive form.
Potential Impact	14	Since there may be several rounds of questioning and reasoning, these requirements bring the problem closer to task-oriented dialog and represent a significant increase in the difficulty of the challenge over the original bAbI (supporting fact) problems.	I am a bit worried that the tasks may be too easy (as the bAbI tasks have been), but still, I think locally these will be useful.



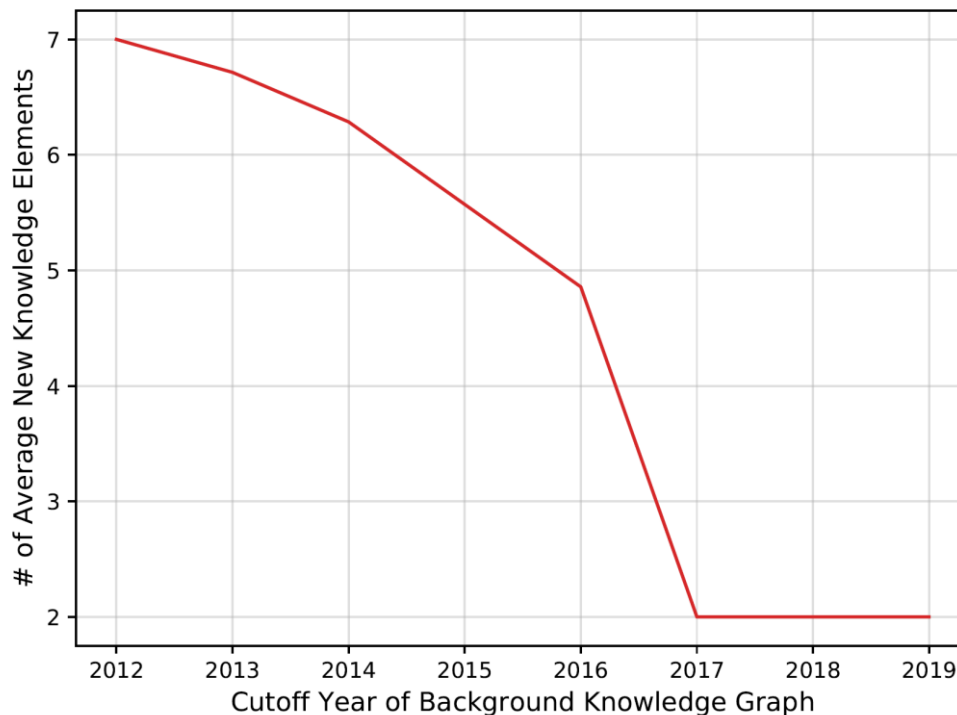
Comment Generation

- For appropriateness, soundness, and potential impact categories, we generate generic positive or negative comments based on the predicted scores.
- For summary, novelty, and meaningful comparison categories, we consider review generation as a template-based graph-to-text generation task.
 - Specifically, for summary and novelty, we generate reviews by describing the Used-for, Feature-of, Compare and Evaluate-for relations in evidence graphs. We choose positive or negative templates depending on whether the predicted scores are above 3. We use the predicted overall recommendation score to control summary generation. For related work, we keep the knowledge elements in the evidence graph with a TF-IDF score^[1] higher than 0.5. For each knowledge element, we recommend the most recent 5 papers that are not cited as related papers.

[1] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.



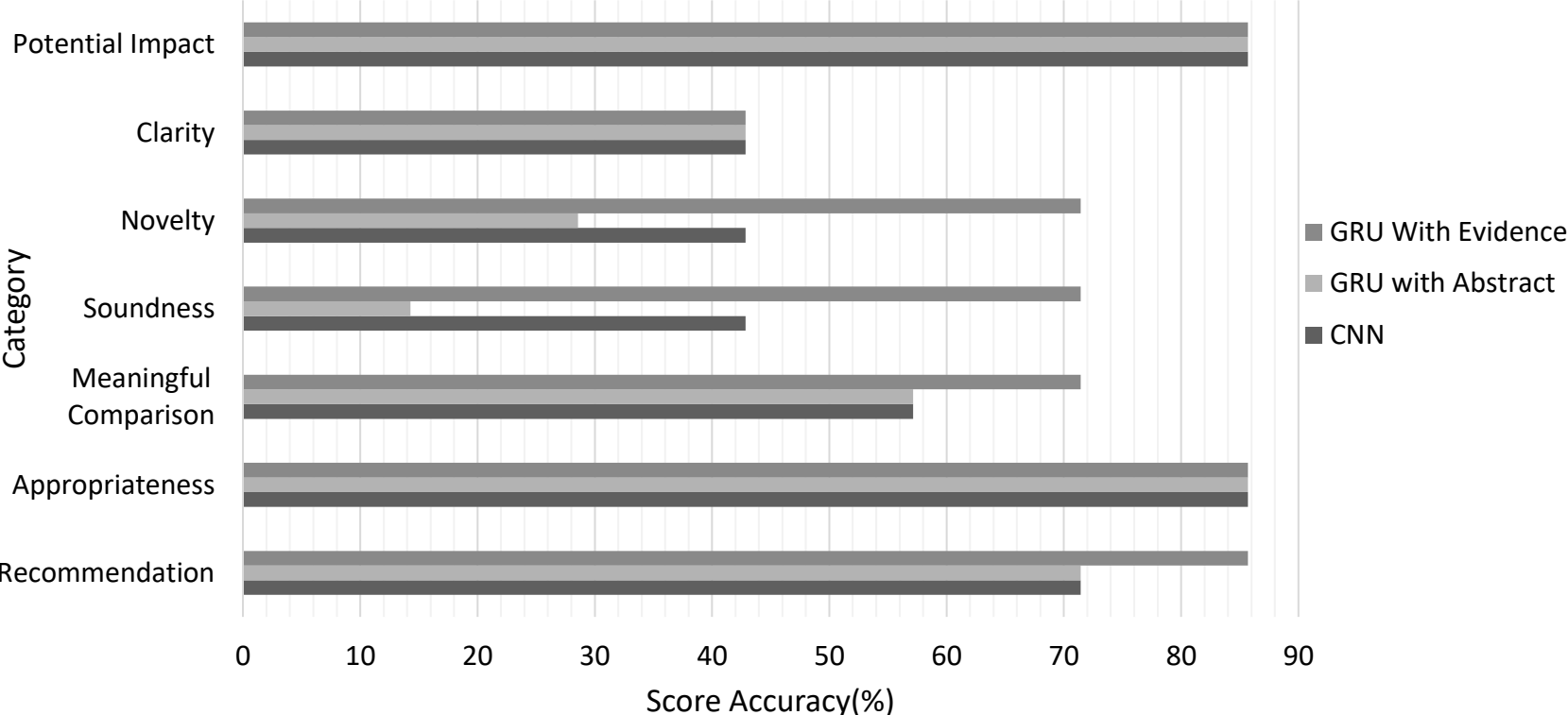
Score Prediction Performance



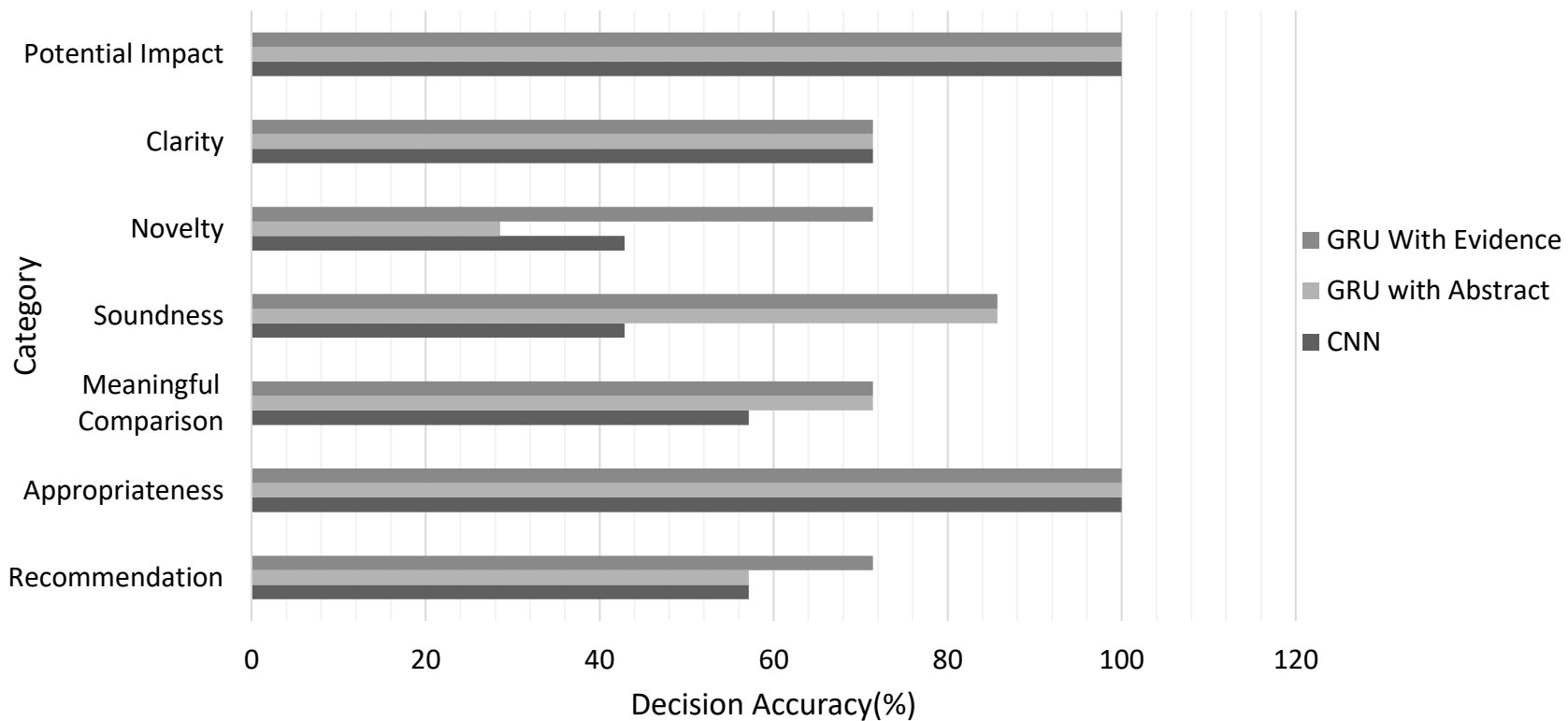
The average number of new knowledge elements in ACL2017 test papers given the background KG constructed from (1965~cutoff year)



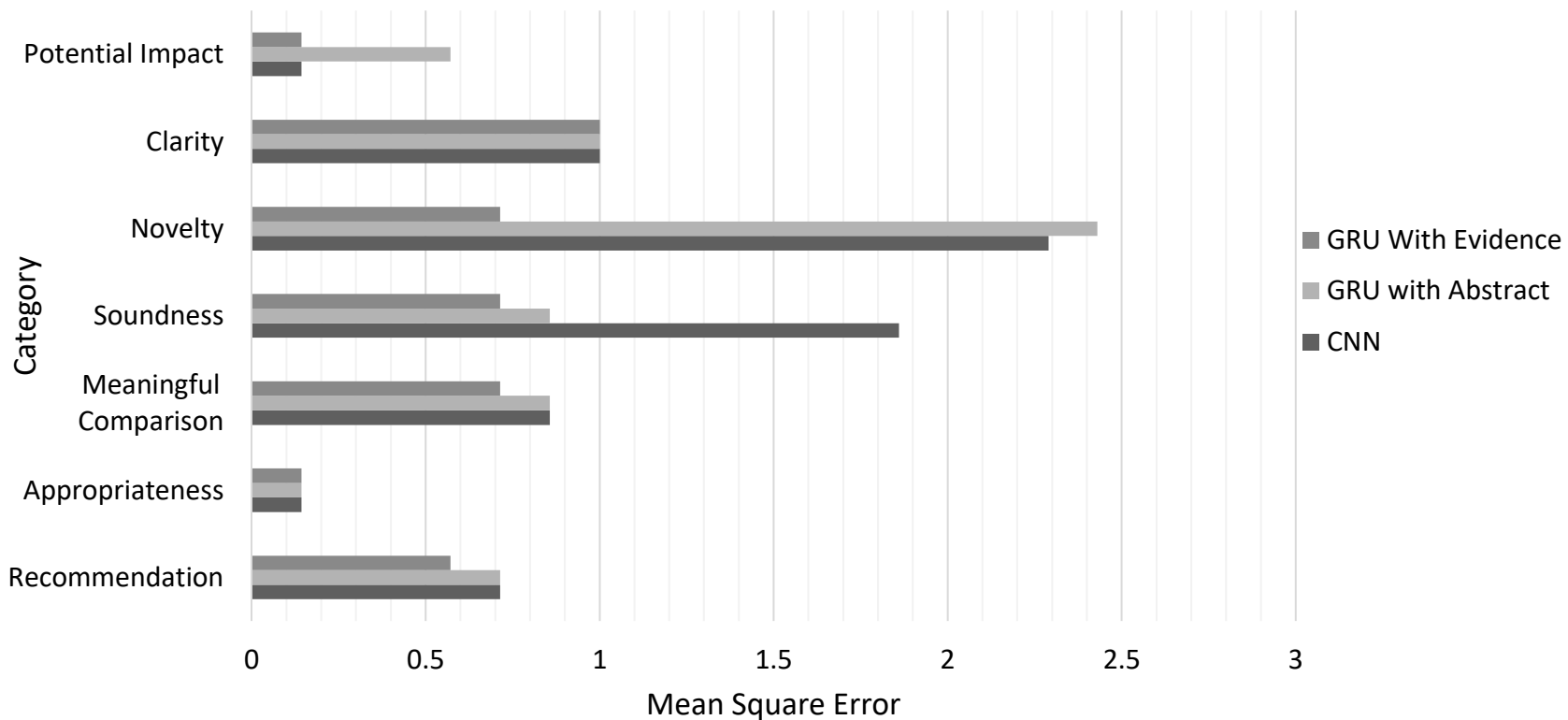
Score Prediction Accuracy



Decision Accuracy



Mean Square Error



Comment Generation Performance

- For the system generated review comments for 50 ACL2017 papers, we ask domain experts to check whether each comment is constructive and valid.
- Two researchers independently annotate the reviews and reach the inter-annotator agreement of 92%, 92%, and 82% for Novelty, Summary and Related Work, respectively. One expert annotator performs data adjudication.
- The percentages of constructive and valid comments are 70.5%, 44.6% and 41.7% for Summary, Novelty and Meaningful Comparison, respectively. Human assessors also find that for 20% of these papers, human reviewers do not suggest missing related work for Meaningful Comparison, while ReviewRobot generates constructive and informative comments.



System Generated Review for an Example Paper^[1]

- Summary
 - **[SYSTEM]:** The paper proposes novel skip-gram, attention scheme, sememe-encoded models and word representation learning for NLP tasks. The authors uses linguistic common-sense knowledge bases.
 - **[HUMAN]** This paper proposes the use of HowNet to enrich embeddings.
- Novelty
 - **[SYSTEM]** The paper proposes novel attention scheme and sememe-encoded models.
 - **[HUMAN]** I see the main contribution of the work to be the results which show that we can learn better word representations (unsure about wsi) by modeling sememe information than other competitive baselines.

[1] Niu, Y., Xie, R., Liu, Z., & Sun, M. (2017). Improved Word Representation Learning with Sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2049–2058).



System Generated Review for an Example Paper [1]

● Meaningful Comparison

- **[SYSTEM]** The following related papers are missing:
 - About low-dimensional semantic space:
 - Unsupervised Approximate-semantic Vocabulary Learning for Human Action and Video Classification: *Qiong Zhao and Horace HS Ip. 2013. Unsupervised Approximate-semantic Vocabulary Learning for Human Action and Video Classification. Pattern Recognition Letters, 34(15):1870–1878.*
 - About sememes:
 - Chinese Word Sense Disambiguation with PageRank and HowNet: *Jinghua Wang, Jianyi Liu, and Ping Zhang. 2008. Chinese Word Sense Disambiguation with PageRank and HowNet. In Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing.*
 - A Maximum Entropy Approach to HowNet-based Chinese Word Sense Disambiguation: *Ping Wai Wong and Yongsheng Yang. 2002. A Maximum Entropy Approach to HowNet-based Chinese Word Sense Disambiguation. In COLING-02: SEMANET: Building and Using Semantic Networks.*
 - About word similarity and word analogy:
 - Open IE as an Intermediate Structure for Semantic Tasks: *Gabriel Stanovsky, Ido Dagan, et al. 2015. Open IE as an Intermediate Structure for Semantic Tasks. 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), pages 303–308.*
- **[HUMAN]** The paper would be stronger with the inclusion of more baselines based on related work.

[1] Niu, Y., Xie, R., Liu, Z., & Sun, M. (2017). Improved Word Representation Learning with Sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2049–2058).



Remaining challenges

- The quality of ReviewRobot is mainly limited by state-of-the-art Information Extraction performance for the scientific literature domain.
 - *“Japanese short-answer scoring and support machine”* cannot be recognized by the IE system
- Paper review generation requires background knowledge acquisition and comparison with the target paper content.
 - Given the following two sentences in a paper: *“Third, at least 93% of time expressions contain at least one time token.”*, and *“For the relaxed match on all three datasets, SynTime-I and SynTime-E achieve recalls above 92%.”*, the ReviewRobot cannot understand *“93%”* is the upper bound of the model performance.
- ReviewRobot cannot generalize knowledge elements into high-level comments such as *“deterministic”* as in *“The tasks 1-5 are also completely deterministic”*.
- ReviewRobot still lacks of deep knowledge reasoning ability to judge the soundness of algorithm design details, such as whether the split of data set makes sense, whether a model is able to generalize.
- ReviewRobot is not able to comment on missing hypotheses, the problems on experimental setting and future work.
- ReviewRobot currently focuses on text only and cannot comment on mathematical formulas, tables and figures.



Conclusion

- Propose a new research problem of generating paper reviews
- Present the first complete end-to-end knowledge-driven framework to generate scores and comments for each review category
- Create a new benchmark that includes 8K paper and review pairs, 473 manually selected pairs of paper sentences and constructive human review sentences, and a background KG constructed from 174K papers



Future Work

- Build a taxonomy on top of the background KG
- Incorporate multi-modal analysis of formulas, tables, figures, and citation networks



Thank you!

Code: <https://github.com/EagleW/ReviewRobot>

Dataset: <https://drive.google.com/file/d/1NclEwGEVcHCrSWk8s3lDjvEbMIWXQoXM/view?usp=sharing>