Measuring Knowledge Gaps in Student Responses by Mining Networked Representations of Texts

Chen Qiao
The University of Hong Kong
Pokfulam Road
Hong Kong SAR, China
qiaochen@outlook.com

Xiao Hu
The University of Hong Kong
Pokfulam Road
Hong Kong SAR, China
xiaoxhu@hku.hk

ABSTRACT

Gaps between knowledge sources are interesting to various stakeholders: they might indicate potential misconceptions awaiting correction, complex or novel knowledge that requires careful delivery or studying. Motivated by these underlying values, this study explores the knowledge gap phenomenon in the context of student textual responses. In the method proposed in this study, discourses are first mapped into structured knowledge spaces where gaps between correct/incorrect responses and assessed knowledge are measured by network-based metrics. Empirical results demonstrate the effectiveness of the proposed method in measuring gaps in student responses. The networked representation of texts proposed in this study is novel in quantitatively framing gaps of knowledge. It also offers a set of validated metrics for analyzing student responses in research and practice.

CCS CONCEPTS

• **Computing methodologies~Natural-language processing** • **Applied computing**~Semantic networks

KEYWORDS

knowledge gap measurement, student responses, educational data mining, network analysis, text mining.

ACM Reference Format:

C. Qiao and X. Hu, 2019. Measuring Knowledge Gaps in Student Responses by Mining Networked Representations of Texts. In *The 9th International Learning Analytics and Knowledge Conference (LAK19), March, 2019, Tempe, AZ, USA.* ACM, New York, NY, USA. 5 pages. https://doi.org/10.1145/3303772.3303822

LAK19, March, 2019, Tempe, AZ, USA

ACM 978-1-4503-6256-6/19/03…$15.00
https://doi.org/10.1145/3303772.3303822

1 INTRODUCTION

Gaps between various knowledge sources occur when the sources cannot be bridged properly. In text comprehension, readers consciously or unconsciously construct connections between the knowledge mediated by the text and the knowledge existing in their minds. Should every effort fail in bridging the two sources of knowledge to a sound degree, significant gaps would occur, which would in turn undermine the comprehension process. This study is motivated by the underlying values of knowledge gap analysis.

This study narrows the scope of exploration to the context of educational question answering (QA). In QA, the primary goal is to assess whether students have mastered certain knowledge. In other words, it examines the gaps between the two sources: students' knowledge and the assessed knowledge. Rather than abstract mental construcs, this study focuses on the concretized text representations of both souces. The task of measuring the gaps between the two knowledge sources is henceforce interpreted as collecting evidence from textual student responses and assessment knowledge descriptions to make judgements. Naturally, the correct responses, in contrast to the incorrect ones, should connect better to the underlying knowledge of the questions. However, how to measure the goodness of connections, that is, how much more a gap is filled by a correct response than by an incorrect one, remains unresolved and calls for a quantitative measurement scheme. This study aims to provide such a scheme based on network analysis of the QA discourses in the networked representation space. A preliminary empirical experiment on student responses is conducted to prove the effectiveness of the proposed approach.

This study will contribute to the literature of automated assessment with the proposed method and metrics for measuring gaps of knowledge in educational discourses. The method and metrics can also enrich toolsets in education practices regarding computer-assisted assessment and feedback generation, which could potentially improve efficiency in both teaching and learning.

2 RELATED WORKS

2.1 Networked Representation of Knowledge

In the networked scheme of knowledge representation, each node represents a knowledge unit and each edge denotes a relation between two nodes. The notion of a networked scheme has been adopted in many learning and cognitive theories for describing cognitive processes. For example, Siemens' [1] connectivism theory describes learning as a process of establishing connections among knowledge components in an information network. The Construction-Integration Model [2,3] explains the cognitive process of reading comprehension as mental operations on the concept networks produced from the texts. Similarly, many other theories of human cognitive processes such as [4-7] share the same underlying assumption of networked knowledge representation.

These classical cognitive theories demonstrate the powerful representational capacity of the networked approach, especially in explaining complex cognitive processes related to knowledge, and provide theoretical support for adopting a networked scheme in this study. Henceforth, in this study, the question of measuring knowledge gaps boils down to computations in the networks that represent knowledge of the targeted sources.

2.2 Knowledge Gaps in the Networked Representation

Cognitive models on text comprehension interpret comprehension as a complex process of interactions between knowledge in the texts and readers' background. The product of a successful comprehension process is a coherent representation integrating both knowledge sources in the reader's mind (e.g., [2,5,8-14]). However, if a coherent representation cannot be established, that is, if there are significant gaps between the two sources, the comprehension process would be unsuccessful. In the theory of *inference* [14], integration of knowledge sources happens through establishing associations among knowledge components in the two sources, which results in a new knowledge structure that merges those of the two sources. The extent to which the resultant, merged knowledge structure is coherent indicates the success of the inference process [14].

Inspired by the concept association and structure integration processes in the *inference* procedure [14], this study views knowledge gaps as specific phenomena that can be observed from the merging operation between targeted knowledge sources. Based on the networked knowledge representation, the merging process is operationalized as associations between concept nodes and integration of network structures. The results are new knowledge networks with different landscapes, differing based on the degree of gaps between the knowledge sources. Therefore, knowledge gaps between sources can be captured by tracing the changes of the networks in the merging process.

3 THE MEASUREMENT SCHEME

As a network view is adopted in representing knowledge, quantitative measurement can be based on the various metrics in network analysis [16]. Classical network indices grant us powerful tools to capture network structures, and we propose a set of metrics built on these indices to capture network changes. The basic idea is introduced below.

Whether or not a knowledge integration process is successful can be revealed by comparing the network structures before and after the merging of two knowledge sources. A lower degree of node and edge integration after merging may indicate greater knowledge gaps between the sources. Leveraging methods in network analysis, knowledge gaps can be measured with the changes of node-level (micro) and global (macro) network indices before and after the merging operation. The task of quantifying knowledge gap phenomenon has thus been transformed to the computation of value changes in network indices at different levels. Consequently, texts with larger or smaller gaps with respect to certain background knowledge can be hypothesized to have different value changes in specific network indices.

In the context of educational QA where student responses are assessed against certain given knowledge (noted as "assessment knowledge" thereafter), two dimensions of the network dynamics are considered (shown in Fig. 1). One dimension is student response vs. assessment knowledge represented in the network and captured with network metrics; the other is network index changes in the macro (global) vs. micro (local) levels. The two dimensions result in four combinations of analysis: node-level and global network-level index changes with respect to student responses; and node-level and global network-level index changes with respect to assessment knowledge.

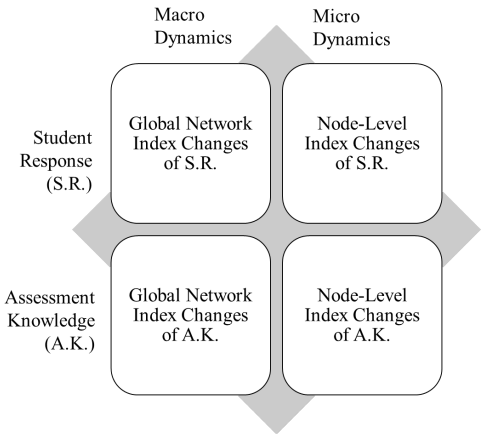

Figure 1: Focused Aspects of Network Dynamics Before and After Merging Operation

In the macro view, network indices characterizing the whole network structure are used to measure the global changes. Our metrics are computed on the differences of the indices after merging: average shortest path length, assortative coefficient, Pearson correlation coefficient,

diameter, radius, node connectivity, Estrada index, global efficiency, and the number of label propagation communities. These indices reflect the overall shape, connectivity and sparsity of the networks.

In the micro view, indices characterizing individual nodes are used to capture the changes of local structures. Our metrics are computed on the dynamics of the following indices: clustering coefficient, node degree, neighbor degree, betweenness centrality, closeness centrality, degree centrality, eccentricity, effective size, efficiency, eigenvector centrality, harmonic centrality, Newman betweenness centrality, PageRank, subgraph centrality, average neighbor degree. These indices can reveal relative importance, position and neighbor conditions of nodes in a network.

3.1 An Illustrative Case

To help illustrate the proposed measurement scheme, the analysis of a sample case (replicated from [15]) is presented here. The knowledge to be assessed is described in the following discourse:

"*Water and other materials necessary for biological activity in trees are* transported *throughout the stem and branches in thin, hollow tubes in the xylem, or wood tissue*".

The following are *examples* of incorrect (A) and correct (B) student responses in recapturing the knowledge:

A. stems transport water to other parts of the plant by converting water to food.

B. stems transport water to other parts of the plant through a system of tubes.

The left and right subgraphs of Fig. 2 illustrate the networked representations of the results after merging the two responses with the assessment knowledge respectively. In each subgraph, circle, rectangle and triangle nodes denote concepts in assessment knowledge, those in student response and overlapped concepts respectively.

It could be noted from Fig.2 that *response A* has a larger gap with the assessment knowledge compared to *response B*. Specifically, compared to the right network (*response B*), the left network (*response A*) has fewer shared concepts (nodes) with the assessment knowledge; the shared concepts are less central in the resultant network; the none-overlapped concepts have fewer connections to those in the assessment knowledge network. All the observations indicate a looser connection and thus a lower degree of integration between *response A* and the assessment knowledge.

Such differences can be quantified by tracing the dynamics/changes of the network indices. Fig. 3 shows a subset of the proposed metrics capturing the differences of network indices of the two response networks before and after the merging operations. The first two metrics measure the changes of the *global* network structure. The larger values of the two metrics on *response A* indicate that the merged network of *response A* has a more elongated structure than that of *response B* (as also shown in Fig. 2). It is also

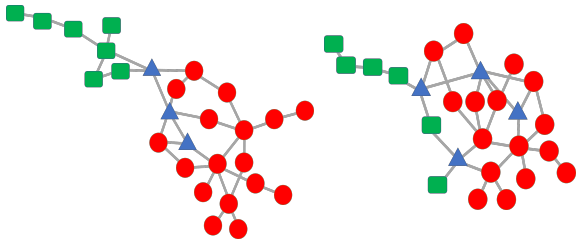

Figure 2: Example Results of Knowledge Merging. (The left and right subgraphs represent the merging results of incorrect and correct responses.)

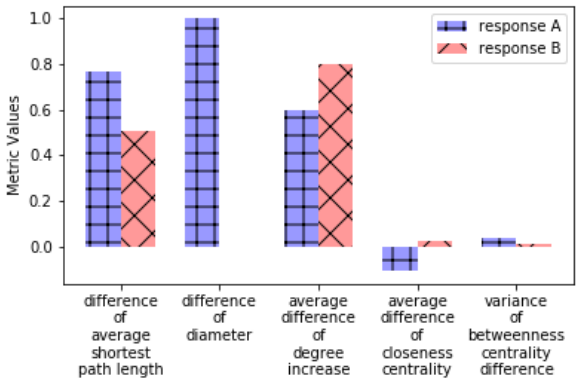

Figure 3: Changes of network indices after merging assessment knowledge with two responses

noteworthy that the *diameter* of the merged network of *response B* did not change from that of the assessment knowledge: the connections between the assessment knowledge and *response B* were established at more central positions and hence are stronger. The connections between *response A* and the assessment knowledge are, however, quite biased to the top left part of the assessment knowledge network, which implies a greater gap between the assessment knowledge and the response.

The last three metrics in Fig. 3 measure the changes of node-level (local) structures. *Average difference of degree increase* measures the average changes of node connectivity after merging. It could be noted that, after merging with the assessment knowledge, the concept nodes in the network of *response B* gained more connectivity (with higher values) than that of *response A,* indicating a better compatibility between *response B* and the assessment knowledge. *Average difference of closeness centrality* measures the changes of closeness centrality of the nodes of the response networks. The situation in this example is that the concepts in *response A* on average lost closeness centrality after merging (the change value is negative), whereas the concepts in *response B* obtained slightly increased centrality. This metric, at the micro level, indicates a better integration between *response B* and the assessment knowledge. The last metric, *variance of betweenness centrality difference* captures the disperse degrees of changes in the nodes' betweenness centrality values. By comparison, we can see that the changes of

betweenness centrality are more even across the concept nodes in *response B*. This metric at the local level corroborates the fact that the connection of the networks of *response B* and the assessment knowledge is better integrated, and hence there is a smaller knowledge gap between *response B* and the assessment knowledge.

4 DATASET AND EXPERIMENT

The empirical exploration consists of 1) statistical tests on the significance of the proposed metrics in differentiating incorrect and correct responses, and 2) a prediction experiment taking the metrics as features. A baseline prediction method with common lexicon-based features is also included in 2), for comparison purposes. The experiment was conducted on the dataset of SemEval-2013 Task 7 which consists of questions typically seen in school exercises, assessments or instructional conversations, all with reference answers and student responses. The dataset provides training and testing sets with 135 and 196 instances respectively. Each instance contains one question, one reference answer and a number of correct and incorrect student responses. We randomly sampled the student responses to include one correct and one incorrect response for each instance in the experiment. The testing set contains questions from the training set with unseen responses, new questions unseen in the training set, as well as new questions from unseen domains. In this experiment, the assessment knowledge network of each sample was constructed on the questioned knowledge and the reference answer, while the response networks were built on the correct and incorrect student responses respectively.

All texts were processed through tokenization, stop words filtering and lemmatization of words to their root forms. Each lemmatized token is taken as a node in the network. In constructing and merging concept networks, a 5-word sliding window scheme was used within which the lemmatized tokens (nodes) are connected to one another in the network.

For statistical testing, paired t-test was conducted with Bonferroni corrections for each metric. To observe the performance of the metrics in classifying correct and incorrect responses, three predictive models with distinctive natures were chosen: 1) *Logistic Regression* (RL) - predicting on linear combinations of the features; 2) *Random Forest* (RF) - predicting on non-linear splits of the feature space, with a mechanism of avoiding overfitting (by reducing variance); 3) *Gradient Boosting* (GBT) - predicting on non-linear splits of feature space, with an ensembled mechanism of minimizing assumptions about the target function (reducing bias).

As baselines for comparison, word features based on lexicons were also extracted and used for prediction. Two widely applied types of lexical features were adopted: *Word2Vec* [17] and *Latent Semantic Index* (LSI) [18] each of which was specified with a vector size of 32 to constrain the feature size to prevent overfitting. Moreover, six additional features deemed helpful in this case were also included: size of response vocabulary gain after merging with assessment knowledge; size of assessment knowledge vocabulary gain after merging with response; *BLEU* (BiLingual Evaluation Understudy) [19] and *Rouge* (Recall-Oriented Understudy for Gisting Evaluation)-1, -2 and -l scores [20] between the reference answer and student response. While the first two additional features could indicate how much extra information is obtained after the merge of student response and assessment knowledge, *BLEU* and *Rouge* scores are used to measure the closeness of the student response to the reference answer in respect of vocabulary agreement.

5 RESULTS AND ANALYSIS

5.1 Significant Metrics

The metrics that significantly differed between incorrect and correct responses are listed in Table 1. From the angle of assessment knowledge, the first section of Table 1 shows that significant changes are primarily detected by local metrics. In general, local structures of the assessment knowledge network (AKN) after merging with correct response network (CRN) changed more evenly across all nodes (indicated by lower values in metrics 9-14) than those after merging with incorrect response network (IRN). The local metrics also indicate increased path lengths among nodes after merging, and merging with CRN makes AKN less dispersed than merging with IRN, as shown by distance- and efficiency-related metrics 1-3. Likewise, more node connections were established in the merged network of AKN + CRN than that of AKN + IRN (metrics 4 and 8). In addition, metrics 5-7 show that the centrality decreased less in the merge of AKN + CRN than that of AKN + IRN, indicating more central connections were formed in the AKN + CRN case.

From the perspective of response network changes, the remaining two sections of Table 1 show that there are also significant structure differences between CRN and IRN after merging. In general, nodes in CRN were more likely to

Table1: Significant Metrics on SemEval Dataset

Source of interest	Metric	Mean diff (gapped)	Mean diff (None-gap)	P value	T
Background Knowledge (local metrics)	1. Avg. eccentricity	9.30E-02	5.35E-02	.027*	-2.24
	2. Local efficiency	1.08E-02	1.56E-02	.007**	2.76
	3. Avg. harmonic centrality	3.13E+00	3.68E+00	.009**	2.63
	4. Average neighbor degree	8.17E-01	1.26E+00	.000**	5.02
	5. Avg. closeness centrality	-3.11E-02	-2.26E-02	.009**	2.66
	6. Avg. current flow closeness centrality	-4.83E-02	-3.77E-02	.000**	3.71
	7. Avg. degree centrality	-1.17E-01	-8.52E-02	.005**	2.85
	8. Ratio of nodes with degree increase	1.35E-01	1.89E-01	.000**	5.49
	9. Var. average neighbor degree	5.89E-01	9.03E-01	.000**	4.24
	10. Var. current flow closeness centrality	2.26E-04	3.10E-04	.013*	2.52
	11. Var. degree centrality	1.35E-02	2.13E-02	.002**	3.09
	12. Var. effective size	3.33E+00	5.17E+00	.000**	3.65
	13. Var. eigenvector centrality	1.56E-04	3.41E-04	.002**	3.21
	14. Var. harmonic centrality	9.75E-01	1.84E+00	.000**	5.16
Answer (global metrics)	15. Average shortest path length	5.34E-01	4.74E-01	.000**	-5.29
	16. Diameter	1.13E+00	9.55E-01	.031*	-2.17
	17. Global efficiency	-2.61E-01	-2.32E-01	.000**	5.50
	18. Node connectivity	-2.26E+00	-1.39E+00	.000**	3.77
Answer (node-level metrics)	19. Local efficiency	-7.54E-02	-6.72E-02	.000**	3.77
	20. Avg. eccentricity	9.35E-01	7.48E-01	.001**	-3.32
	21. Avg. betweenness centrality	3.16E-02	2.47E-02	.000**	-4.83
	22. Avg. closeness centrality	-3.01E-01	-2.49E-01	.000**	5.10
	23. Avg. current flow closeness centrality	-3.64E-01	-3.04E-01	.000**	5.44
	24. Avg. degree centrality	-1.02E+00	-8.52E-01	.000**	5.23
	25. Avg. eigenvector centrality	-2.02E-01	-1.56E-01	.000**	4.42
	26. Avg. Newman betweenness centrality	3.16E-02	2.47E-02	.000**	-4.82
	27. Ratio of nodes with degree increase	4.87E-01	5.62E-01	.000**	3.74
	28. Var. betweenness centrality	3.56E-03	2.71E-03	.017*	-2.43
	29. Var. current flow betweenness centrality	3.76E-03	2.68E-03	.002**	-3.12
	30. Var. eigenvector centrality	6.07E-03	5.18E-03	.004**	-2.91
	31. Var. harmonic centrality	2.96E+01	2.66E+01	.020*	-2.35
	32. Var. Newman betweenness centrality	3.56E-03	2.70E-03	.016*	-2.44

establish connections with central nodes of AKN rather than peripheral nodes as in the IKN case (see Fig. 2 for an example). As a result, the path lengths of the nodes in CRN grew less than that of the IRN (metrics 15, 16, 20), while the efficiency and betweenness metrics of the former decreased less (metrics 17, 19, 21 and 26). In addition, when merging with the AKN, nodes in CRN tended to gain more connections while keeping decentralization at a lower degree than those in IRN (metrics 22-25 and 27). Likewise, the connectivity metric (metrics 18) decreased less for CRN than IRN. The lower values of the variance metrics (metrics 28-32) indicate that more nodes in CRN were merged in similar manners, while there were more extreme cases (either no merging or become bridging nodes) occurred in IRN merges.

5.2 Prediction Performances

The performances of different feature-model combinations in predicting correct and incorrect responses are listed in Table 2. The best performance is obtained by GBT classifier combined with the proposed metrics, with a classification accuracy reaching 0.9860, greatly surpassing the baselines with lexical features (accuracy values are around 0.60). The superiority remains true for LR and RF classifiers as well. The consistent advantages of the network-based metrics proposed in this study over traditional word-based features across multiple predication models evidence the effectiveness of the network-based measurement scheme.

Table 2: Prediction Performance of the Metrics

Classifier	Features	Train Set Accuracy	Test Set Accuracy
Gradient Boosting	proposed metrics	0.9924	**0.9860**
	w2v + lexical	0.6336	0.5839
	LSI + lexical	0.6183	0.5909
Logistic Regression	proposed metrics	0.9733	**0.9825**
	w2v + lexical	0.5992	0.5979
	LSI + lexical	0.6183	0.5874
Random Forest	proposed metrics	0.9885	**0.9790**
	w2v + lexical	0.6489	0.6084
	LSI + lexical	0.6450	0.6014

6 CONCLUSION AND FUTURE WORK

This study focuses on the measurement of knowledge gaps between student responses and assessment knowledge. Empirical results demonstrated the significant indicative and predictive power of the proposed metrics in knowledge gap detection tasks.

This study is our initial step of a series of works in progress towards automatic educational QA analysis. Future works include exploring these metrics on larger datasets; considering partially-correct responses and gaps in different text span scales; and embedding common-sense knowledge into the method to generalize it to scenarios where assessment knowledge is unavailable. It is also worthwhile to examine the effectiveness of the proposed method to educational practices such as computer-assisted grading and misconception diagnosis.

ACKNOWLEDGEMENTS

This study is partially supported by an Early Career Scheme grant from the Research Grants Council of the Hong Kong S.A.R., China (Project No.: HKU 27401114).

REFERENCES

[1] George Siemens. 2005. Connectivism: A learning theory for the digital age. In *International Journal of Instructional Technology and Distance Learning*. 3–10.

[2] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95, 2 (1988), 163–82

[3] Walter Kintsch and Teun A. Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* (1978), 363–394. Myroslava Dzikovska, Rodney Nielsen, Chris Brew, et al. 2013. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. *In 2nd Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation*. ACL, 263–274.

[4] Paul Van den Broek, Michael Young, Yuhtsuen Tzeng, Tracy Linderholm, et al. 1998. The landscape model of reading: Inferences and the online construction of a memory representation. In *The Construction of Mental Representations During Reading,* Herre van Oostendorp and Susan R. Goldman (Eds.). Erlbaum, Mahwah, NJ, USA, 71–98.

[5] Jerome L. Myers and Edward J. O'Brien. 1998. Accessing the discourse representation during reading. *Discourse Processes* 26, 2-3 (1998), 131–157.

[6] Anthony J Sanford and Simon C Garrod. 1981. *Understanding written language: Explorations of comprehension beyond the sentence*. John Wiley & Sons.

[7] David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group (Eds.). 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA.

[8] P.M.A. Gernsbacher and M.A. Gernsbacher. 1990. *Language Comprehension as Structure Building*. Taylor & Francis Group.

[9] Anne E. Cook, Jennifer G. Halleran, and Edward J. O'Brien. 1998. What is readily available during reading? A memory-based view of text processing. *Discourse Processes* 26, 2-3 (1998), 109–129.

[10] Rolf A Zwaan, Joseph P Magliano, and Arthur C Graesser. 1995. Dimensions of situation model construction in narrative comprehension. *Journal of experimental psychology: Learning, memory, and cognition* 21, 2 (1995), 386–397.

[11] Yuhtsuen Tzeng, Paul van den Broek, Panayiota Kendeou, and Chengyuan Lee. 2005. The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods* 37, 2 (2005), 277–286.

[12] Paul Van den Broek, Kirsten Risden, Charles R Fletcher, and Richard Thurlow. 1996. A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. *Models of understanding text* (1996), 165–187.

[13] A. M. Glenberg. 1997. What memory is for: Creating meaning in the service of action. *Behavioral and Brain Sciences* 20, 1 (1997), 41–55.

[14] Panayiota A Kendeou. 2015. *A general inference skill*. Cambridge University Press, 160–181.

[15] Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of The 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*. New Orleans, Louisiana, USA.

[16] John Scott. 2017. *Social Network Analysis*. SAGE.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems* 26, (2013), 3111–3119.

[18] S Deerwester, ST Dumais, GW Furnas, TK Landauer, and R Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information science* 41, 6 (1990), 391-407.

[19] K Papineni, S Roukos, T Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 311–318. Philadelphia, Pennsylvania, USA.

[20] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, 10.