JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

Bayesian analysis of ConceptNet relations on PubMed dataset

Rajeswaran Viswanathan

Sathya Priya. S

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

ConceptNet is a popular knowledge graph built using crowd sourcing. To construct a knowledge graph from plain text relationship identification between words is a critical task. Downstream tasks like finding similarity is sensitive to these relationships. From PubMed abstracts, words are extracted and stop words removed. Using Glove (word vector) “Nearest neighbor” words are identified as candidate words to this root PubMed word. Relationship between these words is identified via numberbatch vectors of ConceptNet. Similarity for each word pair is calculated. Bayesian Random Effects Model (REM) is used to study this relationship strata. Analysis shows that there is heterogeneity among the relationships.

Core business problem

Introduction

Relationship extraction is critical in Natural Language Processing (NLP) when the objective is to construct knowledge graphs. There is considerable research on automated extraction from unstructured text into knowledge graph. ConceptNet is a commonsense knowledge base, composed mainly from the Open Mind Project from Media Laboratory, Massachusetts Institute of Technology. It contains 1.6 million edges connecting more than 300 000 nodes. Combining Word2vec , GloVe , PPDB , and ConceptNet, using retrofitting and Linked Open Data, a space of multilingual term embeddings called ConceptNet Numberbatch has been created. ConceptNet Numberbatch is a set of semantic vectors: it associates words and phrases in a variety of languages with lists of 600 numbers, representing the gist of what they mean. Using this we identify similarity of root word and candidate word. To truly understand concepts that appear in natural language text, it is important to recognize the informal relations between these concepts that are part of everyday knowledge, which are often under-represented in other lexical resources, which is the reason ConceptNet was selected for this study. Active research is being done on the application of various similarity metrics. Cosine similarity is used for identification of similarity in this research.

SUMPUBMED is a dataset for abstractive summarization over scientific article. From this corpus, stop words were removed and unique words identified. These unique words are taken as seed words. Word2Vec, Glove and FastText is used to identity "nearest" word, which is taken as candidate word. This study has been motivated by the need to understand the similarity and differences in relationship quality extracted from ConceptNet for a pair of words.

Quality of relationships are most critical in Natural Language Processing (NLP) tasks. There is considerable research on automated relation extraction from unstructured text into knowledge graph. The quality of relationship is not taken into account which establishing this relationship. Similarity can be used as a quality metric to find out if a pair of words are "related" or not. Using retrofitting and Linked Open Data, ConceptNet team has created ConceptNet Numerbatch which is a multilingual term embeddings. ConceptNet Numberbatch is a set of semantic vectors: it associates words and phrases in a variety of languages with lists of 600 numbers, representing the gist of what they mean. Using this we identify similarity of root word and candidate word. Cosine similarity is used for identification of similarity in this research.

To analyze this a Bayesian Random Effects Model (REM) is used. This is a popular modeling technique used in clinical trials analysis among other areas. The concern is explaining and not prediction, which the domain of REM - core idea being synthesis of multiple studies. Results show that the quality of relationship varies considerably when compared with Glove and there is considerable heterogeneity within the relationships.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

The paper is organized as follows: details of the data set have been presented in Section 3. Core ideas for the new statistical model are outlined in Section 4. Section 5 details the process of data preparation. Analysis of the data with results are discussed Section 6. Section 7 provides the concluding remark and scope for future research.

Data

The focus of this paper is well supported by the freely available pre-trained Word Vector Glove . They can be downloaded and experiments can be done in local or cloud environment. ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use. It is available[1] for use as API as well as local setup. This paper uses ConceptNet 5. The SUMPUBMED is a dataset is a standard benchmark data for summarization tasks openly available from Github[2].

Given two words, the ConceptNet Similarity score (CNS) can be found. These are numeric values in the range of 0 to 1. A criteria of 0.5 as a cutoff is applied and convert this into binary measure. This criterion is mainly related to the planned random effects model for dichotomized data that are further stratified by a variable.

The required format for the proposed analysis is a two-fold contingency table that classifies two dichotomous variables (Table 1). In this case, two levels of X_1 are ConceptNet similarity (CNS) ($>= 0.5$) or ConceptNet similarity (CNS) (< 0.5) for the 40 relationship types identified by ConceptNet with regard to a word pair and that of X_2 are Glove (WV) ($>= 0.5$) and Glove (WV)(< 0.5) for the pair of words. Each cell count is the number of words that accounts for the respective combination of X_1 and X_2. The stratifying variables are the different relationship types. The format of transformed (numeric to count) data is illustrated in Table [tab:MA-data] for the ConceptNet relationships in scope.

Data format for ConceptNet similarity (CNS) < 0.5)/*ConceptNet similarity (CNS)* $>= 0.5$) *counts of root word with candidate words*

	ConceptNet relations	
X_1/X_2	CNS $>= 0.5$	CNS< 0.5
Glove (WV) $>= 0.5$	n_1	n_2
Glove (WV) < 0.5	n_3	n_4

Relationship	n_1	n_2	n_3	n_4
Antonym	721	541	178	498
AtLocation	313	272	79	324
CapableOf	23	35	4	32
Causes	32	50	15	32
CausesDesire	5	12	2	8
CreatedBy	18	24	2	6
DefinedAs	3	1	0	0
DerivedFrom	3897	143	3904	508
Desires	4	8	0	5
DistinctFrom	295	280	26	178
EtymologicallyDerivedFrom	13	11	8	13

[1] https://Conceptnet.io

[2] https://github.com/vgupta123/sumpubmed

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

EtymologicallyRelatedTo	351	53	138	141
FormOf	8180	53	1487	144
HasA	61	41	28	28
HasContext	317	178	216	223
HasFirstSubevent	1	4	0	10
HasLastSubevent	3	9	2	7
HasPrerequisite	50	42	5	47
HasProperty	47	74	15	44
HasSubevent	21	38	11	37
IsA	2930	1025	1193	1423
LocatedNear	4	4	0	3
MadeOf	25	23	4	18
MannerOf	194	169	137	314
MotivatedByGoal	10	19	2	8
PartOf	476	173	94	166
ReceivesAction	22	38	3	28
SimilarTo	819	136	424	280
Synonym	3830	570	1854	1238
UsedFor	214	202	46	158

Two step method

Bayesian Random Effects Model (REM), also called variance components model is used to study the variability within strata. REM is applied when there are multiple strata being analyzed and they are assumed to be drawn from a hierarchy of different populations whose differences relate to that hierarchy. Fixed Effects Model (FEM) is not the right choice to summarize the results for this type of data, since there is inherent variability within each of the strata. In NLP it is important to understand the sources of variability, within-strata and between strata when making inferences about the population.

In the Normal–Normal Hierarchical Model (NNHM) of REM, there are 2 parts - a sampling and a parameter model. The sampling model assumes approximately normally distributed estimates $Y_1, \dots, Y_k$ for the trial-specific parameters $\theta_1, \dots, \theta_k$

$$Y_j \mid \theta_j \sim N\left(\theta_j, \sigma_j^2\right), j = 1, \dots, k$$

The similarity model assumes the parameters as random effects

$$\theta_j \mid \mu \sim N(\mu, \tau^2), j = 1, \dots, k$$

The variance component for between-strata variability, $\theta_j = \mu + \epsilon_j$, with $\epsilon_j \sim N(\theta, \tau^2)$. The between-strata standard deviation τ determines the degree of similarity across parameters. For the mean parameter μ inference can be simplified by considering only the marginal model given by

$$Y_j \mid \mu, \tau \sim N\left(\mu, \sigma_j^2 + \tau^2\right), j = 1, \dots, k$$

Suitable prior will be used to estimate μ and τ^2 In our study Y_j is the observed OR for each table. The corresponding within variance as shown in equation (3). Usually delta method is applied to calculate Asymptotic Standard Error ASE. The choice of the OR is encouraged due to its property. The ASE in delta method the ASE is independent of the design in OR as can be seen below.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

The delta method calculation give you ASE. In the Binomial case, Assuming two row totals are fixed, $n_1 \sim Bin(r_1, p_1)$ and $n_2 \sim Bin(r_2, p_2)$ where n_1, n_2 are cell counts, r_1, r_2 are row totals and p_1, p_2 are the respective parameters. Then we have $f(p_1, p_2) = \ln(p_1) + \ln(1 - p_2)$. Further,

$$\nabla f = \left(\frac{1}{p_1(1-p_1)} \frac{-1}{p_2(1-p_2)}\right)$$

$$\Sigma = \begin{pmatrix} \frac{p_1(1-p_1)}{r_1} & 0 \\ 0 & \frac{p_2(1-p_2)}{r_2} \end{pmatrix}$$

The expression $\nabla f \sum (\nabla f)^T$ simplifies to

$$\left(\frac{1}{r_1} \frac{-1}{r_2}\right) \begin{pmatrix} \frac{1}{p_1(1-p_1)} \\ \frac{-1}{p_2(1-p_2)} \end{pmatrix} = \frac{1}{r_1 - n_2} + \frac{1}{n_1} + \frac{1}{r_2 - n_2} + \frac{1}{n_2}$$

In terms of cell counts,

$$\nabla f \sum (\nabla f)^T = \frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}$$

The Asymptotic Standard Error (ASE) is thus given by

$$ASE(\hat{\phi}) = ASE(\widehat{lnOR}) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}$$

When we consider the multinomial case, assuming cell counts $n; n_1, n_2 n_3, n_4$ with respective parameters $p_1, p_2, p_3, p_4 \sum p_i = 1$; $n = \sum n_i$. The log of odds ratio can be expressed as, $\log OR = \log(n_1) + \log(n_4) - \log(n_2) - \log(n_3)$. Also, $f(\theta) = \log(p_1) + \log(p_4) - \log(p_2) - \log(p_3)$. Now,

$$\nabla f = \left(\frac{1}{p_1} \frac{-1}{p_2} \frac{-1}{p_3} \frac{1}{p_4}\right)_{1\times 4}$$

Since, n_i follows multinomial distribution, $V(\widehat{p_i}) = \frac{p_i(1-p_i)}{n}$ and the covariance, $cov(n_i, n_j) = np_ip_j$ will simplify to $\frac{1}{n} p_i p_j$.

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & -p_1p_4 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & -p_2p_4 \\ -p_3p_1 & -p_3p_2 & -p_3(1-p_3) & -p_3p_4 \\ -p_4p_1 & p_4p_2 & -p_4p_3 & -p_4(1-p_4) \end{bmatrix}_{4\times 4}$$

Here also $\nabla f \sum (\nabla f)^T$ simplifies to

$$\nabla f \sum (\nabla f)^T = \sum \left[\frac{1}{n_i}\right] and$$

$$ASE(\widehat{lnOR}) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

As can be seen above for odds ratio, there is no difference between binomial or multinomial design.

Authors have compared different approaches to compute standard error where bootstrap and delta methods have been well discussed. Bootstrap has been evaluated for Deep Learning and performance of bootstrap in terms of computation cost has been the main concern. Most software solutions R (metafor, meta, bayesmeta), Strata implement the asymptotic method only. Efron published bootstrap method, which is widely used extensively to estimate standard errors and confidence intervals in NLP . Attempt is made here to use bootstrap to directly calculate Y_i and the variance instead of the asymptotic method.

The statistical inference aims to provide following summaries to understand the association between the variables in the individual and overall levels together with the amount of heterogeneity.

1. Point estimate and confidence interval for the true θ_i
2. Point and interval estimates of μ to understand the presence or absence of an overall effect and its statistical significance.
3. Estimates of variability measures indicating the variation between strata.

Data pre-processing

SUMPUBMED is a dataset for abstractive summarization over scientific article. PubMed comprises of more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. SUMPUBMED was created by downloading around 33,772 documents identified as BMC literature. BMC (BIO MED CENTRAL) literature incorporates BMC health services research papers related to medicine, pharmacy, nursing, dentistry, health care, and so on. After downloading this dataset, stop words are removed as part of pre-processing. Our interest is to find relationship for meaningful words and not stop words.

To get close vectors for these root words in the corpus - The technique of nearest neighbors is used. For each root work, using nearest neighbors 100 candidate words from Word Vectors (WV) Word2Vec, Glove and FastText are identified. WV can take a seed word and "nearest neighbors" 100 words can be identified using simple distance metric and this is used as candidate words, after removing duplicates. The scope of investigation is the similarity of these words to the root word in WV compared with ConceptNet similarity. The size of the word pairs is 93196 at this stage of analysis.

ConceptNet has 33 relationships[3]. In this dataset only 30 relationships exists. The 3 relationships - ExternalURL, ObstructedBy and SymbolOf does not exist in this dataset. Using Python package, the relationships between the root word and the candidate word is identified. Then a simple count of each of the relationship type is made to get overall totals for each relationship. If there is no relationship, these are dropped from the analysis.

For each word pair (root word and candidate word) cosine similarity is calculated using ConceptNet numberbatch. This metric data (WV similarity and CNS) is then converted into binary data by having 0.5 as cutoff. This allows us to do count modeling on the data to study the random effects. After transformation into binary, a simple count for arriving at the 2x2 tables for each ConceptNet relationship as shown in Table [tab:MA-data].

It can be seen from the Table [tab:MA-data] that there are many cells with zero values (for example, HasFirstSubevent has 0 in n_3. Odds ratio cannot be directly computed for such tables. Researchers have strongly argued that zero counts should not be dropped. Since this is sampling zero as opposed to structural zero, 1 is added as pseudo-count when there are zeros in the table.

[3] https://github.com/commonsense/ConceptNet5/wiki/Relations

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

Analysis

Python programming language (Python Software Foundation)[4] was used to extract CNS. Libraries like pandas, numpy and API access to ConceptNet which are freely available for install as libraries to base Python language, which were used for data preparation. Python was also used to calculate the similarity metrics for extraction of candidate words from Glove. The REM modeling after data extraction has been carried out in the computational tool R (R Core Team, 2016) especially with Bayesmeta package in R.

If we have a collection K effects sizes, then Y_k is the kth effect-size estimate (with sample variance v_k) of a parameter, k. For notational convivence we group all effect-size estimates as **T**, sample variances as **V**, and effect-size parameters as . We are interested in some overall mean effect size (μ) and the between-studies standard deviation (τ). Thus, given our observed data (**T** and **V**) and our parameters ($\theta, \mu,$ and τ), using Bayes' theorem our meta-analytic model is

$$P(\mu, \tau, \boldsymbol{\theta} \mid \mathrm{T,V}) = \frac{P(\mathrm{T,V} \mid \mu, \tau, \boldsymbol{\theta}) P(\mu, \tau, \boldsymbol{\theta})}{P(\mathrm{T,V})}$$

Because $P(\mathrm{T,V})$ is independent of any parameters, meta-analysts will typically use a proportional model as shown below,

$$P(\mu, \tau, \boldsymbol{\theta} \mid \mathrm{T,V}) \propto P(\mathrm{T,V} \mid \mu, \tau, \boldsymbol{\theta}) P(\mu, \tau, \boldsymbol{\theta})$$

Expanding this equation for the terms to be specified,

$$P(\mu, \tau, \boldsymbol{\theta} \mid \mathrm{T,V}) \propto \prod_{k=1}^{K} [P(Y_k \mid \theta_k, v_k) P(\theta_k \mid \mu, \tau) P(\mu) P(\tau)]$$

Four terms must be specified for this equation: $P(Y_k \mid \theta_k, v_k)$, $P(\theta_k \mid \mu, \tau)$, $P(\mu)$, and $P(\tau)$. In a hierarchical model format this can be written as

$$\begin{aligned} Y_i &\sim N(\theta_k, v_k) \\ \theta_k &\sim N(\mu, \tau) \\ \mu &\sim U(\bullet) \\ \tau &\sim \pi(\bullet) \end{aligned}$$

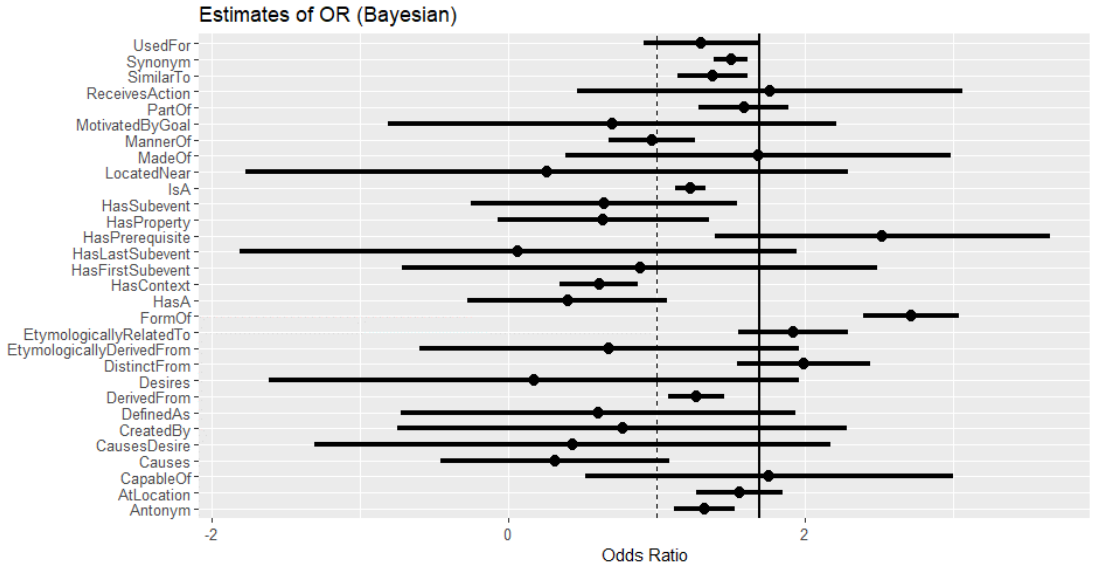
where $\pi(\bullet)$ distributions attempted here are - Half Normal, Log Normal and Gamma. For the μ parameter, the *bayesmeta* package has two main prior distribution options: Normal (with mean and variance) or uniform distribution. For τ six priors have been considered - Halfnormal05(0.5), Halfnormal10(1.0), lognormal05($\mu = -1.7413, \sigma = 1.0464$), lognormal10($\mu = -1.0482, \sigma = 1.0464$), Gamma05($\alpha = 0.5, \beta = 0.001889$), Gamma10($\alpha = 0.5, \beta = 0.007553$). The log-normal parameters and the square root of inverse Gamma parameters have been matched to half-normal distributions via 5% and 95% quantiles. These are well known priors from published meta-analysis , ,. The output when applying these different priors is shown in Table [tab:Final.Prior] along with the Credible intervals.

Numerical and graphical summaries are quite straightforward with R. Point and credible interval estimates (CrI) for overall odds ratio (μ) for CNS $>= 0.5$ or CNS < 0.5 in context of Glove (WV) $>= 0.5$ and WV being < 0.5 are presented in Table [tab:OR-estimate].

The forest plot shown in Figure [fig:Forest] helps in visualizing the heterogeneity between the 30 relationships and also shows the credible intervals for each of the point estimates.

[4] https://www.python.org/

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452



Title	Estimated OR	LL	UL
Antonym	1.319	1.115	1.523
AtLocation	1.558	1.264	1.853
CapableOf	1.757	0.516	2.999
Causes	0.315	-0.456	1.087
CausesDesire	0.433	-1.304	2.170
CreatedBy	0.770	-0.746	2.285
DefinedAs	0.604	-0.726	1.933
DerivedFrom	1.268	1.078	1.459
Desires	0.176	-1.609	1.960
DistinctFrom	1.993	1.545	2.440
EtymologicallyDerivedFrom	0.681	-0.594	1.955
EtymologicallyRelatedTo	1.921	1.550	2.292
FormOf	2.712	2.390	3.033
HasA	0.401	-0.271	1.074
HasFirstSubevent	0.886	-0.717	2.488
HasLastSubevent	0.067	-1.810	1.944
HasPrerequisite	2.520	1.393	3.646
HasProperty	0.638	-0.073	1.350
HasSubevent	0.648	-0.250	1.545
IsA	1.227	1.123	1.331
LocatedNear	0.259	-1.770	2.288
MadeOf	1.686	0.388	2.984
MannerOf	0.970	0.681	1.260
MotivatedByGoal	0.699	-0.813	2.211
PartOf	1.586	1.281	1.891
ReceivesAction	1.765	0.467	3.062
SimilarTo	1.380	1.145	1.616
Synonym	1.501	1.388	1.614

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

Title	Estimated OR	LL	UL
UsedFor	1.299	0.916	1.683

It can be seen that the individual odds ratio for the various relationships vary in the REM model. This heterogeneity is clearly captured in the REM model. FormOf and HasPrerequisite have the highest OR estimate (2.712 and 2.520 respectively) followed by DistinctFrom, EtymologicallyRelatedTo, PartOf, AtLocation and Synonym in the range of 1.5 to 2. In the next group, SimilarTo Antonym, DerivedFrom and IsA all have OR greater than 1. Although the OR estimates for ReceivesAction, CapableOf, MadeOf and UsedFor is greater than 1, the Credible Interval covers 1 and so it is not statistically significant. MannerOf has OR 0.970 which is not statistically significant since the Credible Interval covers 1. For items with OR less than 1, HasFirstSubevent, CreatedBy, MotivatedByGoal, EtymologicallyDerivedFrom, HasSubevent, HasProperty, DefinedAs, CausesDesire, HasA, Causes, LocatedNear, Desires and HasLastSubevent do not have statically significant OR because the Credible Interval covers 1. This wide range of OR estimations make it difficult to trust the quality of relationships identified.

The overall results for the different priors are presented in Table [tab:Final.Prior]

Prior	τ	LL	UL	μ	LL	UL
Half Normal (1.0)	0.543	0.359	0.770	1.255	0.995	1.503
Half Normal (0.5)	0.526	0.353	0.735	1.258	1.005	1.500
log Normal (1.0)	0.522	0.347	0.738	1.259	1.007	1.500
log Normal (0.5)	0.511	0.340	0.720	1.262	1.013	1.498
Gamma (1.0)	0.511	0.338	0.725	1.262	1.013	1.498
Gamma (0.5)	0.510	0.337	0.724	1.262	1.013	1.498

Conclusion

When importing text into knowledge graphs the quality of relationships is critical. If different relationships within the same corpus have different strength and there is disagreement on the notion of closeness then downstream automation is risky. Data scientist does not know if the relationship when extracted can be trusted for the task, the odds of it being in agreement is favorable for only 9 out of 30 relations studied. In most cases the odds are not statistically significant (18/30). The initial candidate words were identified using "closeness" concept of Word2Vec, Glove and FastText. But, this notion of closeness is not shared by CNS all the time. Although ConceptNet is trained on Word2Vec and Glove the overall OR estimate for Glove is weak. This is one aspect of variability - wide disagreement between the different relations within ConceptNet on the "similarity" of the words identified as "nearest neighbor" by Word2Vec, FastText and Glove.

A different aspect studied by this Bayesian REM model is about how much these relationships agree with Web Vector models. Since Word2Vec, Glove and FastText are older methods, they have not taken the context of the words into account for generating the vector. ConceptNet is hand curated (crowd sourced). Difference in OR estimate and wide difference in CI length within each REM and disagreement between the REM models on the strength of relationship gives low confidence on the core concept of "similarity" being addressed.

This presents a challenge for data scientists while automating knowledge creation - what to trust and in which context when there is such high heterogeneity among these relations. Unfortunately, this core concept of "closeness" is not addressed by researchers, while new metrics for similarity are being developed. It means that currently this problem cannot be solved without manual intervention for knowledge creation.

Further research can be done on a wider corpus to see if this heterogeneity is prevalent. Authors plan to extend this research to other languages beyond English. Handling of zero's in REM is a active topic . Zero cell problem can be further probed by other techniques like Bayesian model.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2016-2025
https://publishoa.com
ISSN: 1309-3452

The key objective is to automate various downstream NLP tasks and so care must be taken to generate quality relationships. When this itself is questionable, the value of these identified relations for automation decreases dramatically.

1. Speer, Robyn, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An open multilingual graph of general knowledge." Thirty-first AAAI conference on artificial intelligence. 2017.
2. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781v3 [cs.CL].2013
3. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP). Pages 1532-1543. 2014.
4. Juri Ganitkevitch, Benjamin Van Durme, Chris Callison-Burch. PPDB: The Paraphrase Database. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. June 2013.
5. Yang, Jiaqi, et al. "Measuring the short text similarity based on semantic and syntactic information." Future Generation Computer Systems 114 (2021): 169-180.
6. Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Nils Y. Hammerla. Correlation Coefficients and Semantic Textual Similarity. arXiv:1905.07790v1 [cs.CL]. May 2019.
7. Wang, Haoyu, et al. "Joint constrained learning for event-event relation extraction." arXiv preprint arXiv:2010.06727 (2020).
8. Becker, Maria, Katharina Korfhage, and Anette Frank. "COCO-EX: A Tool for Linking Concepts from Texts to ConceptNet." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. (2021).
9. Faruqui, Manaal, et al. "Retrofitting word vectors to semantic lexicons." arXiv preprint arXiv:1411.4166 (2014).
10. Dowd, Bryan E et al. “Computation of standard errors.” Health services research vol. 49,2: 731-50. doi:10.1111/1475-6773.12122. (2014)..
11. Nilsen, Geir K., et al. "A Comparison of the Delta Method and the Bootstrap in Deep Learning Classification." arXiv preprint arXiv:2107.0160. (2021).
12. Hedges LV, Olkin I . Statistical Methods for Meta-Analysis. San Diego, CA, USA: Academic Press.(1995)
13. Efron B, Tibshirani R. An Introduction to the Bootstrap. Chapman and Hall: New York, 1993.
14. Solano, Quintin P., et al. "Natural Language Processing and Assessment of Resident Feedback Quality." Journal of Surgical Education (2021).
15. Zhu, Haotian, et al. "NLPStatTest: A Toolkit for Comparing NLP System Performance." arXiv preprint arXiv:2011.13231 (2020).
16. Cook, Thomas D. "Advanced statistics: up with odds ratios! A case for odds ratios when outcomes are common." Academic Emergency Medicine 9.12 (2002): 1430-1434.
17. Keus, F., Wetterslev, J., Gluud, C., Gooszen, H. G. & van Laarhoven, C. J. H. M. Robustness assessments are needed to reduce bias in meta-analyses that include zero-event randomized trials. Am. J. Gastroenterol. 104, 546–551 (2009).

18. Kuss, O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. Statist. Med. 34, 1097–1116 (2015).

19. Chang Xu, Luis Furuya-Kanamori, Liliane Zorzela, Lifeng Lin, Sunita Vohra, A proposed framework to guide evidence synthesis practice for meta-analysis with zero-events studies, Journal of Clinical Epidemiology, Volume 135, Pages 70-78, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2021.02.012.(2021)

20. Wei, Jiajin, et al. "Meta-analysis With Zero-event Studies: A Comparative Study With Application to COVID-19 Data." (2021).

21. Xu, Chang, et al. "Exclusion of studies with no events in both arms in meta-analysis impacted the conclusions." Journal of clinical epidemiology 123 (2020): 91-99.

22. Roever, Christian, Tim Friede, and Maintainer Christian Roever. "Package 'bayesmeta'." (2017).

23. Bernardo, José M., and Adrian FM Smith. Bayesian theory. Vol. 405. John Wiley & Sons, (2009).

24. Smith, T. C., Spiegelhalter, D. J., and Thomas, A. Bayesian approaches to random-effects meta-analysis: A comparative study. Statistics in Medicine, 14(24), 2685–2699. (1995).

25. Polson NG, Scott JG . On the half-Cauchy prior for a global scale parameter. Bayesian Analysis 7: 887–902.(2012)

26. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chichester, UK: Wiley & Sons.(2004)

27. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1: 515–534.(2006)

28. Zhang, Tianyi, et al. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675 (2019).

29. Rosen, Clifford J. "Revisiting the rosiglitazone story—lessons learned." New England Journal of Medicine 363.9 (2010): 803-806.