Data Collection vs. Knowledge Graph Completion: What is Needed to Improve Coverage?

Kenneth Church
Baidu USA
kennethchurch@baidu.com

Yuchen Bian
Baidu USA
yuchenbian@baidu.com

Abstract

This survey/position paper discusses ways to improve coverage of resources such as WordNet. Rapp estimated correlations, ρ, between corpus statistics and psycholinguistic norms. ρ improves with quantity (corpus size) and quality (balance). 1M words are enough for simple estimates (unigram frequencies), but at least 100M are required for pairs of words (word associations, edges). Knowledge Graph Completion (KGC) attempts to learn missing links in WN18. Unfortunately, WN18 is flawed with information leaking from train to test. More seriously, WN18 is based on SemCor (just 200k words) and dated (collected in 1960s). KGC cannot learn anything that happened since the 1960s, or associations requiring 100M words.

1 Quantity (Size) and Quality (Balance)

How large do the corpora have to be to learn what? In the early 1980s, corpora were about 1M words. The Brown Corpus (Kučera and Francis, 1967; Francis and Kučera, 1979, 1982) was large enough for first order statistics (counts of words), but not for second order statistics (word associations and counts of pairs of words).

The Brown Corpus was a *balanced* corpus. That is, the corpus was intended to be a representative sample of text that the system will see at inference time. The 1M word Brown Corpus consists of 500 samples[1] of 2000 words, representative of contemporary American English (from 1960s).

Over time, balanced corpora became larger. When the community decided to increase the size of balanced corpora from 1M words for the Brown Corpus to 100M for the British National Corpus (BNC) (Aston and Burnard, 1998; Burnard, 2002), it was known that 1M was too small for second order statistics (collocations and word associations), but it was hoped that 100M would be sufficient.

Around this time, Church and Hanks (1990) used an unbalanced sample of 44M words from the AP (Associated Press) to make the case for PMI (pointwise mutual information). Given the estimates in Table 1, it appears in retrospect that 44M words were just barely enough to make the case for PMI.

It was also believed that quality (balance) mattered, but there were few, if any, empirical studies to justify such beliefs. It was extremely controversial when engineers such as Mercer questioned these deeply held beliefs in 1985[2] with: "there is no data like more data." Most people working on corpus-based methods in lexicography were deeply committed to balance as a matter of faith, and were deeply troubled by Mercer's heresy.

More recently, Rapp (2014a,b) provided some empirical evidence that bears on this debate. He used 5 corpora to study quantity (sample size) and quality (balance). In addition to the two balanced corpora mentioned above, Brown and BNC, Rapp looked at 3 unbalanced corpora:

1. 300M words of Wikipedia (Wiki)
2. 2B words of web pages (ukWaC)
3. 4B words of newswire (Gigaword)

This study used correlations, ρ, to compare statistical summaries with psycholinguistic norms: familiarity (Coltheart, 1981), association (Kiss et al., 1973) and relatedness (Fernald, 1896). We will refer to unigram statistics and familiarity norms as *first order*; statistics on pairs of words (such as PMI) and the other norms will be referred to as *second order*. In Table 1, ρ_1 refers to correlations of first order quantities and ρ_2 refers to correlations of second order quantities.

[1] 500 samples span 15 categories: Press Reportage (44 texts), Press Editorial (27), Press Reviews (17), Religion (17), Skills and Hobbies (36), Popular Lore (48), Miscellaneous US Government & House Organs (30), Learned (80), General Fiction (29), Mystery and Detective Fiction (24), Science Fiction (6), Adventure and Western (29), Romance (29), Humor (9).

[2] http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6210–6215
November 7–11, 2021.

	First Order: ρ_1			Second Order: ρ_2		
N words	1M	10M	100M	10M	100M	1B
Brown	0.67	NA	NA	NA	NA	NA
BNC	0.69	0.74	**0.75**	0.35	0.53	NA
ukWaC	0.64	0.71	0.73	0.30	0.48	**0.56**
Wiki	0.60	0.66	0.67	0.27	0.43	NA
Giga	0.55	0.62	0.66	0.14	0.25	0.36

Table 1: ρ_1 and ρ_2 increase with quantity (N) and quality (balance: top 2 rows). Results from (Rapp, 2014b)

Rapp (2014a,b) showed that both ρ_1 and ρ_2 increase with quantity and quality, as shown in Table 1. We suggest two simple rules of thumb:

1. Balance Trade-off: ρ_1 over N balanced words $\approx \rho_1$ over $100N$ unbalanced words
2. First order is 100x easier than second order: ρ_1 on N words $> \rho_2$ on $100N$ words

Among the unbalanced corpora, web pages (ukWaC) have relatively large ρ, better than Wiki and Giga, though not as good as BNC. Note that 1B words of web pages has a better ρ_2 than 100M words of BNC.

It is hard to know what will happen for much (1000x) larger corpora, but one might expect diminishing returns. Of course, extrapolating estimates like these by 10x or more is known to be risky (Efron and Thisted, 1976). Figures 1a-b of (Rapp, 2014b) suggest that while ρ is increasing almost everywhere, there may be some deceleration (negative second derivative), especially for large N.

Although Rapp's estimates predate much of the work on embeddings, we expect these estimates of quantity and quality to hold for static embeddings (Mikolov et al., 2013; Pennington et al., 2014) and contextual embeddings (Devlin et al., 2019; Sun et al., 2020), assuming the connection between PMI and Word2vec in Levy and Goldberg (2014).

In addition to size and balance, there are many other factors to consider. Different languages are different. Languages are constantly evolving. Variations are to be expected over time[3,4] (Hamilton et al., 2016; Szymanski, 2017)[5] and space, as well as sociolinguistic factors, demographics, gender bias (Pearce, 2008; Drozd et al., 2016; Sheng et al., 2019; Nissim et al., 2020; Kumar et al., 2020), etc.

In addition to language change, topics and domains are also constantly evolving. Obviously, news, Wikipedia and web pages are very different from social media (Twitter) and academic writing (ACL Anthology (Radev et al., 2013), ArXiv,[6] PubMed[7]). The Brown Corpus predates social media, and most publications in repositories such as: PubMed, ACL Anthology and ArXiv (Church, 2017). The Brown Corpus also predates huge changes in technology (computers and cell phones), the news media (cable TV and the Internet), and modern medicine (e.g., COVID-19, SARS, HIV, affordable DNA sequencing). Nevertheless, many resources in our field are still based on the Brown Corpus, including the Penn TreeBank (Marcus et al., 1993) and SemCor[8].

2 WordNet Coverage and SemCor

WordNet[9] (Miller et al., 1990; Miller, 1995; Feldbaum, 1998; Miller and Fellbaum, 2007; Vossen and Fellbaum, 2021) is widely cited because of accessibility[10] as well as coverage. Why is the coverage as good as it is, and how can it be improved? Unlike other methods for constructing lexical resources (Lenat, 1995; Sinclair, 1989; Hanks, 2008), WordNet was developed in tandem with SemCor, a small subset of the Brown Corpus, tagged with pointers into WordNet. The team constantly tracked coverage as indicated by the reference to 96% below:

> [SemCor] *starts with the corpus and proceeds through it word by word... This procedure has the advantage of immediately revealing deficiencies in the lexicon: not only missing words (which could be found more directly), but also missing senses and indistinguishable definitions–deficiencies that would not surface so quickly with* [alternatives]... *we ... adopted the* [SemCor] *approach for the bulk of our semantic tagging... over several months ... estimates of ... coverage have been slowly improving... it is currently averaging a little better than 96%.* (Miller et al., 1993)

The SemCor process helped manage growth. In 1993, they were adding almost 1k concepts per

[3] https://books.google.com/ngrams

[4] https://github.com/dimazest/google-ngram-downloader

[5] https://nlp.stanford.edu/projects/histwords/

[6] https://arxiv.org/help/bulk_data

[7] https://pubmed.ncbi.nlm.nih.gov/

[8] https://web.eecs.umich.edu/~mihalcea/downloads.html

[9] https://wordnet.princeton.edu/

[10] www.nltk.org/howto/wordnet.html

month. The number of synsets (word senses) nearly doubled from 63k in 1993 to 118k today. In addition, the process led to the creation of SemCor 3.0, a subset of about 20% of the Brown Corpus tagged with WordNet senses.

While SemCor has much to recommend it, there are also some obvious concerns. SemCor is only 200k words, probably not enough given Rapp's estimates above. Coverage of WordNet could be improved by building something like SemCor, but based on a larger corpus of more modern material. Alternatively, it might be possible to combine small annotated corpora with larger unannotated corpora.

3 Knowledge Graph Completion (KGC)

An alternative suggestion for improving WordNet coverage is: Knowledge Graph Completion (KGC)[11] (Nguyen, 2017; Wang et al., 2017; Yu et al., 2019). A standard KGC benchmark is WN18.[12] WN18 is a graph $G = (V, E)$. There are 41k vertices, V. Each vertex is a WordNet synset, a pointer to a set of synonymous lemmas in WordNet. There are 118k such synsets in WordNet.

The edges, E, connect two vertices with one of 18 relations. The relations also come from WordNet. Some relations are more frequent than others.

Many of the relations come in pairs, as shown in Table 2. By construction, if x is-a y, then there will be a hypernym link from x to y, as well as a hyponym link from y to x. We will refer to the backward links as *inverses*.

The KGC task is to infer subsets of these graphs from other subsets of these graphs. That is, KGC splits E randomly into three sets: train, validation and test. WN18 consists of 141k edges in train, 5k in validation and 5k in test.

For each set, we have a set of input features, X, and a set of output labels, Y. The standard procedure uses X_{train} and Y_{train} to fit a model. This model is used to predict $\hat{Y}_{test}$ from X_{test}. The predicted values, $\hat{Y}_{test}$, are compared with the gold labels, Y_{test}, to compute a score.

There is a considerable literature on KGC methods, e.g., Trans[DEHRM], KG2E, ConvE, Complex, DistMult (Bordes et al., 2013; Wang et al., 2014; Yang et al., 2014; Lin et al., 2015; Nickel et al., 2016; Trouillon et al., 2016; Nguyen et al., 2017; Sun et al., 2019).

Relation	Edges	Inverse	Edges
hypernyms	37,221	hyponyms	37,221
derivationally related forms	31,867		
member meronym	7928	member holonum	7928
has part	5142	part of	5148
synset domain topic of	3335	member of domain topic	3341
instance hypernym	3150	instance hyponym	3150
also see	1396		
verb group	1220		
member of domain region	983	synset domain region of	982
member of domain usage	675	synset domain usage of	669
similar to	86		

Table 2: 18 Relations in WN18. By construction, many of these relations have inverses (with similar counts).

Cum %	Freq	Relation
40%	1251	hypernym
74%	1074	derivationally related form
82%	253	member meronym
88%	172	has part
92%	122	instance hypernym
95%	114	synset domain topic of

Table 3: Six relations cover 95% of WN18RR test set.

The next two subsections address two concerns with the KGC literature and the WN18 benchmark:

1. information leakage and
2. size: WN18 is based on SemCor (20% of Brown Corpus), too small for ρ_2 given Table 1

3.1 Information Leakage in KGC

Some of the leakage in the WordNet benchmark, WN18, is well-known and some is not. WN18RR is a reduced subset of WN18 that corrects for the known leakage (Dettmers et al., 2018).[13] The correction removes the 7 inverse relations on the right hand side of Table 2, resulting in the test set shown in Table 3. Before the correction, there are 5000 edges over 18 relations in the WN18 test set. After the correction, there are 3134 edges over 11 relations in the WN18RR test set.

Unfortunately, there is even more leakage in WN18RR that has not been previously reported. Note that "derivationally related forms" also come in pairs. By construction, derivationally related links are symmetric: $xRy \Rightarrow yRx$. That is, if there

[11] https://github.com/Sujit-O/pykg2vec

[12] https://paperswithcode.com/dataset/wn18

[13] See Elangovan et al. (2021) for discussion of leakage in other benchmarks.

Forward	**Inverse Links:** yRx			
Links	**test**	**train**	**valid**	**Totals**
xRy **test**	24	1011	39	1074
xRy **train**	1011	27,701	1003	29,715
xRy **valid**	39	1003	36	1078
Totals	1074	29,715	1078	31,867

Table 4: Information Leakage in WN18RR: Derivationally related links are symmetric ($xRy \Rightarrow yRx$).

is an edge in one direction, then there will also be an inverse edge in the reverse direction. This symmetry will leak information between train and test because it is likely that one member of the pair appears in train and the other appears in test.

Table 4 shows that many of these pairs are indeed leaking information in this way. The table shows how these "derivationally related" edges, xRy, and their inverses, yRx, are distributed across the WN18RR test, train and validation splits.

In particular, of the 1074 derivationally related edges in the WN18RR test set, all of them are also in one of the other sets, but in the reverse direction. The 1074 reversed edges are split across test (24), train (1011) and valid (39).

Because of this leakage, a system can do very well on this benchmark without learning anything useful about WordNet. Simply reverse edges in the training set and predict that those reversed edges will appear in the test set (unless they have already been seen in the training or validation sets). Such a system will correctly predict $1 - 24/1074 = 98\%$ of the derviationally related edges ($1074/3134 = 34\%$ of the test set).

One could correct for this leakage by removing the redundant edges, just as we removed redundant edges to reduce WN18 to WN18RR.

3.2 Corpus Sizes

Size is perhaps more serious than leakage. Leaning edges in WN18 is a second order task. As shown in Table 1, second order tasks typically require a corpus of 100M words or more. Unfortunately, WN18 is based on SemCor (indirectly via WordNet). SemCor is a 200k word sample, too small for second order tasks. Inferences on downstream graphs (such as WN18) are unlikely to capture associations on pairs of words.

KGC is learning subsets of WordNet from other subsets of WordNet. But given Table 1, to improve WordNet, we need more data, not less. Modern corpora are 1000x larger than SemCor, and more representative of text from this century. We believe it is more profitable to collect more data (and more representative data) than to infer information that is not in the WordNet graph (or the underlying SemCor corpus).

KGC can be viewed as similar to downsampling in speech, where there is a well-known difference between upsampling and downsampling. In speech, it is relatively easy to downsample a waveform from 16 kHz down to telephone bandwidth (8 kHz), but harder to invert the process (upsampling). That is, we can always throw away information by low pass filtering and decimating. But it is harder to recover the high frequency information after it has been thrown away.

SemCor can be viewed as a small sample of contemporary language, downsampled with a strong bias favoring American English from the 1960s. Rapp's estimates suggest there is more information in larger corpora than in smaller corpora. Thus, the downsampling process is throwing away information that cannot be recovered. Obviously, KGC cannot recover information that cannot be recovered, but it is also unlikely to learn anything since 1960, let alone other dialects/languages.

4 Conclusions

What is needed to improve WordNet coverage? We started with Rapp's estimates of ρ, correlations of corpus statistics and psycholinguistic norms. ρ improves with quantity (corpus size) and quality (balance). Unbalanced corpora need to be larger (100x) than balanced. Estimates of second order quantities (word associations and edges in WordNet) require at least 100x more data than first order quantities (frequency/familiarity). Rapp's estimates suggest there is more information in larger samples than in smaller samples.

WordNet is based on SemCor. It is remarkable that WordNet works as well as it does, given Rapp's estimates. One approach to improving coverage is Knowledge Graph Completion (KGC). KGC attempts to learn missing links from subsets. The KGC Benchmarks, WN18 and WN18RR, are deeply flawed. Information is leaking between training and test sets. Some of this leakage has been previously reported, and some has not. But more seriously, if SemCor is already too small and dated, data collection is more likely to succeed than attempts to infer information that is not there.

References

Guy Aston and Lou Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Capstone.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Lou Burnard. 2002. *Where did we go wrong? A retrospective look at the British National Corpus*. Brill Rodopi.

Kenneth Church. 2017. Emerging trends: Inflation. *Natural Language Engineering*, 23(5):807–812.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530, Osaka, Japan. The COLING 2016 Organizing Committee.

Bradley Efron and Ronald Thisted. 1976. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Christine Feldbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

James Champlin Fernald. 1896. *English Synonyms and Antonyms*. Funk & Wagnalls Company.

W Nelson Francis and Henry Kučera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Winthrop Nelson Francis and Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Patrick Hanks. 2008. The lexicographical legacy of john sinclair. *International Journal of Lexicography*, 21(3):219–229.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI, USA.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A Miller. 1995. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38(11):39.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller and Christiane Fellbaum. 2007. Wordnet then and now. *Language Resources and Evaluation*, 41(2):209–214.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2017. A novel embedding model for knowledge base completion based on convolutional neural network. *arXiv preprint arXiv:1712.02121*.

Dat Quoc Nguyen. 2017. An overview of embedding models of entities and relationships for knowledge base completion. *arXiv preprint arXiv:1703.08098*.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Thirtieth Aaai conference on artificial intelligence*.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Michael Pearce. 2008. Investigating the collocational behaviour of man and woman in the bnc using sketch engine. *Corpora*, 3(1):1–29.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Reinhard Rapp. 2014a. Using collections of human language intuitions to measure corpus representativeness. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2117–2128, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Reinhard Rapp. 2014b. Using word familiarities and word associations to measure corpus representativeness. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2029–2036, Reykjavik, Iceland. European Language Resources Association (ELRA).

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

John Sinclair. 1989. *Collins COBUILD English language dictionary*. BOOK. Collins Publishers.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *AAAI*.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. Association for Computational Linguistics.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning (ICML)*.

Piek Vossen and Christiane Fellbaum, editors. 2021. *Proceedings of the 11th Global Wordnet Conference*. Global Wordnet Association, University of South Africa (UNISA).

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Twenty-Eighth AAAI conference on artificial intelligence*.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Shih Yuan Yu, Sujit Rokka Chhetri, Arquimedes Canedo, Palash Goyal, and Mohammad Abdullah Al Faruque. 2019. Pykg2vec: A python library for knowledge graph embedding. *arXiv preprint arXiv:1906.04239*.