Putting WordNet's Dictionary Examples in the Context of Definition Modelling: An Empirical Analysis

Fatemah Almeman* **Luis Espinosa-Anke*◇**
*CardiffNLP, School of Computer Science and Informatics, Cardiff University, UK
◇AMPLYFI, UK
{almemanf, espinosa-ankel}@cardiff.ac.uk

Abstract

Definition modeling is the task to generate a valid definition for a given input term. This relatively novel task has been approached either with no context (i.e., given a word embedding alone) and, more recently, as word-in-context modeling. Despite their success, most works make little to no distinction between resources and their specific features (e.g., type and style of definitions, or quality of examples) when used for training. Given the high diversity lexicographic resources exhibit in terms of topic coverage, style and formal structure, it is desirable for downstream definition modeling to better understand which of them are better suited for the task. In this paper, we propose an empirical evaluation of the well-known lexical database WordNet, and specifically, its dictionary examples. We evaluate them both directly, by matching them against criteria for *good* dictionary writing, and indirectly, in the task of definition modeling. Our results suggest that WordNet's dictionary examples could be improved by extending them in length, and incorporating prototypicality.

1 Introduction

Definition modeling (DM), as introduced by Noraset et al. (2017), is the task of generating a dictionary definition for a given word. This task was made possible by the adoption in NLP of sequence-to-sequence architectures based on RNNs (Gardner et al., 2022). Recently, DM systems have shown impressive performance in several intrinsic and downstream tasks, mostly thanks to being able to go from context-less (Noraset et al. only used the definiendum[1] as a conditioning token at all timesteps) to a contextually richer setting, e.g., by conditioning the generated definition to an example of usage of the target word (Ni and Wang, 2017; Gadetsky et al., 2018; Chang et al., 2018; Zhu et al., 2019; Mickus et al., 2019; Ishiwatari et al., 2019).

Recently, a notable leap in DM was achieved in Bevilacqua et al. (2020), who fine-tuned BART (Lewis et al., 2019) on example-definition pairs, and reported high results in intrinsic benchmarks and, more importantly, used their DM system for downstream NLP, specifically word sense disambiguation (WSD) and word-in-context classification. DM has also been explored from other perspectives, e.g., generating definitions with appropriate specificity using re-ranking mechanisms (Huang et al., 2021), or extending the generation cross-entropy loss with a reconstruction objective (Kong et al., 2022) (reminiscent of works that used dictionary definitions for improving word embeddings via autoencoders (Bosc and Vincent, 2018) or LSTMs (Hill et al., 2016)). Moreover, Barba et al. (2021) explore a BART-based model for performing the reverse task to DM, i.e., *exemplification modeling*, or generating a dictionary example given a term and its definition. Other applications of DM range from the aforementioned lexical semantics tasks to reverse dictionary (predict a word given a definition), interpretability, or for clarifying technical and medical terminology (Chen and Zhao, 2022; August et al., 2022), whereas recent applications of BART to tasks not originally designed to be solved generatively are semantic role labeling (Bevilacqua et al., 2020), relation extraction (Cabot and Navigli, 2021) or entity linking (De Cao et al., 2020).

Despite the above successes, little attention has been paid so far to the quality of the dictionary examples (or *contexts*) used for fine-tuning these models. In fact, most existing DM systems train on WordNet (**WN**) (Miller, 1995), which is the de-facto lexical database for English. However, we are not aware of previous work that has explored the quality (and hence, suitability for DM) of WN examples. Therefore, in this paper, we first inves-

[1]The *genus-et-differentia* Aristotelian definitions follows an *A is a B which Z* structure, with *A* being the *definiendum*, *B* the *genus* and *Z* the *definiens* or *differentia specifica*.

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 42–48

https://doi.org/10.26615/978-954-452-056-4_006

tigate the quality of WN examples by evaluating against the GDEX (Good Dictionary Examples) set of criteria (Kilgarriff et al., 2008), and use as a point of comparison a widely adopted open dataset used in DM, which is primarily based on the Oxford Dictionary (Chang and Chen, 2019) (**CHA**). It is worth noting, however, that these two resources were built for different objectives, as the initial purpose behind creating WN was to explain how lexical meaning is stored in the mind (Broda et al., 2009), and its primary use may be as a sense inventory (Agirre and Edmonds, 2007). However, with this caveat in mind, and given how lexicographic resources are currently converging into useful pretraining and fine-tuning datasets for lexical semantics, we also propose to extrinsically test these two resources in the DM task. Specifically, in our second set of experiments we fine-tune a BART-based model on WN and CHA, and show that generally speaking, results of models fine-tuned on WN perform slightly worse than if fine-tuned on CHA. Our preliminary results suggest that WN's examples sometimes do not provide enough context, making it difficult to learn a good representation for the word being contextualized. We also report an experiment comparing DM modeling results on WN nouns vs. WN verbs; which suggests that a DM model trained on WN nouns performs slightly better.

2 Data

WordNet (WN) is an electronic lexical dictionary for English that describes words (11,7097 nouns, 11,488 verbs, 22,141 adjectives, and 4,601 adverbs) organized in groups of synonyms called "synsets" (Miller, 1995; Fellbaum, 2013). Each synset is described by its definition, lemmas, examples of usage (for some but not all words), and the relations between synsets, e.g., hypernymy (is-a), meronymy (is-part) or troponymy (manner-of). WN has typically been used in lexicographic and language learning settings (Morato et al., 2004), but more importantly, also in NLP, e.g., as a natural language interface for optimizing the precision of search engines, WSD or query expansion (Moldovan and Mihalcea, 2000; Banerjee and Pedersen, 2002). Moreover, relations in WN have been used extensively, for example for improving word embeddings via retrofitting (Faruqui et al., 2014; Espinosa-Anke et al., 2016; Vulić and Mrkšić, 2017; Mrkšić et al., 2017).

CHA (Chang and Chen, 2019), the other resource we consider in this paper, is based on Oxford Dictionaries. It was released with two splits, namely *seen*, where definitions in the training set also exist in the test set, and *unseen*, which contains a set of words not available in the training set (Bevilacqua et al., 2020). This is similar to the lexical splits (as opposed to random splits) present in other analogous tasks such as graded lexical entailment (Shwartz et al., 2016; Vulić et al., 2017). In this paper, we are concerned with the quality of examples in WN (and how they compare with CHA), i.e., sentences where a target word appears, and which should be informative enough to convey the necessary contextual information to clarify fully or partially the word's meaning (encoded in a natural language definition or gloss, instead of e.g., a word embedding).

We show in Table 1 examples from WN and CHA, where it becomes apparent that WN examples have a different pattern, e.g., they are much shorter, and are crucially limited in the contextual information they provide, as opposed to the examples in CHA, which features, first, full-fledged grammatical examples, and second, associated vocabularies that help position the target word in the mental lexicon, which is crucial for word access (Zock et al., 2010).

Data	Lemma	Definition	Example
WN	**people**	(plural) any group of human beings (men or women or children) collectively	old people
CHA	**people**	human beings in general or considered collectively	each day he has looked at a key issue facing us as a nation as a people as frail human beings
WN	**sheet**	any broad thin expanse or surface	a sheet of ice
CHA	**sheet**	a large rectangular piece of cotton or other fabric used on a bed to cover the mattress and as a layer beneath blankets when these are used	Mary quietly got off the bed and covered him with the sheet and blanket
WN	**tall**	great in vertical dimension; high in stature	tall people
CHA	**tall**	of great or more than average height especially with reference to an object relative to width	the elevator came to a stop and the doors slid open revealing the sixth floor of the tall building

Table 1: WN vs CHA definitions and examples for a given lemma (in bold).

3 Experiments

In this section, we introduce the two sets of experiments we perform. First, the descriptive comparison between WN and CHA examples using GDEX as a proxy (Section 3.1). Second, we describe the setting for the DM experiment, where we test WN as supervision signal (Section 3.2).

3.1 GDEX-based comparison

As a proxy for determining the quality of dictionary examples in WN, and given that there is no manually annotated dataset for this purpose, we used GDEX (Good Dictionary Examples) criteria. GDEX is a system that added around 8,000 new example sentences to Macmillan English Dictionary by automatically finding good examples in corpora using a set of rules of thumb (Kilgarriff et al., 2008; Bejoint, 2014).

In our work, we used some of the features that are introduced in GDEX, specifically:

- **sentence length**: according to Kilgarriff et al. (2008), good dictionary examples should range between 10 and 25 words, and thus we penalize shorter or longer dictionary examples proportionally (the more an example deviates from the acceptable minimum or maximum, the more it is penalized).

- **word frequency**: a sentence is penalized for each non-frequent word that is not in the list of the top 20,000 most frequent words in English Wikipedia.

- **anaphoric references**: we penalize the number of pronouns in the dictionary example, normalized by sentence length.

- **sentence probability**: we use the GPT-2 (Radford et al., 2019) language model to score the probability of dictionary examples. Intuitively, this can be a useful metric for semantic coherence and fluency.

3.2 Definition Modeling

The general formulation of DM is as follows. To generate a gloss g that defines a target lemma t in a context c, the standard sequence-to-sequence conditional generation probability is computed by factorising it auto-regressively (Bevilacqua et al., 2020):

$$P(g|c,t) = \prod_{k=1}^{|g|} P(g_k|g_{0:k-1}, c, t) \quad (1)$$

where g_{k} is the k^{th} token of g and g_0 is a special start token (Bevilacqua et al., 2020). We fine-tune BART, a pre-trained encoder-decoder system, to perform the definition generation task by taking the pair (context, target lemma) as an input to produce the corresponding definition. The dataset includes (c, t, g) triples where t is the target word (lemma) in a context c (example) and g is the gold gloss which defines t in c (definition). We encode the input as (t, c) pairs and special tokens are used to identify the target lemma in each context such as *The cherry tree <target> bloomed </target>.*, with the lemma "bloom" as the target word in this context.

Exp. 1 (WN vs CHA) Since we are concerned with using WN in definition modeling, we trained and tested the definition generation model (BART) on WN lemmas that have examples (44,351 lemmas) using an 80/20 split for training and testing. Additionally, we trained the same model using a CHA-derived training set of the same size as our WN training set, and tested it on the same WN test set. We ensured that no duplicates/leakage occured between sets in both experiments. We train both models with a maximum of 50 epochs with early stopping[2].

Exp. 2 (WN Nouns vs WN Verbs) We trained and tested the same BART model with same hyperparameters as in the WN vs CHA experiment on random 10k noun lemmas and 10k verb lemmas from WN separately (using again an 80/20 ratio for training and testing) to evaluate whether there are noticeable differences between these two grammatical categories.

4 Analysis

In this section, we discuss the results of our two experiments, namely GDEX-wise comparison between WN and CHA, and WN's stress test in the DM task.

[2]We implemented our experiments using the `simpletransformers` (`http://simpletransformers.ai/`) library, a wrapper on top of `transformers` (Wolf et al., 2020).

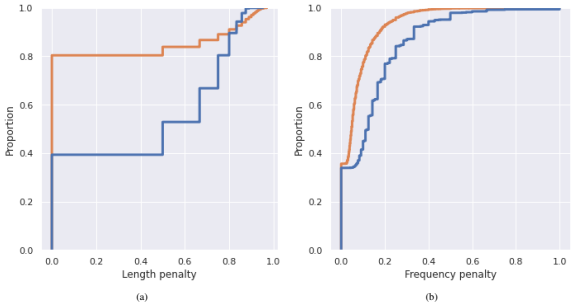

Figure 1: Empirical distribution functions between WN (blue) and CHA (orange) for length (a) and frequency (b) penalties.

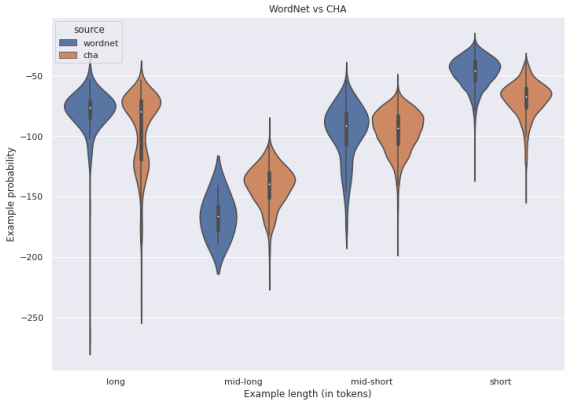

Figure 2: Violin plot showing the difference in log-likelihood assigned by GPT-2 to WN vs CHA examples (higher is better).

4.1 GDEX score

Since Kilgarriff et al. did not specify an optimal weighting for the different factors they took into account in the GDEX metric, we look individually at each of the four factors discussed in Section 3.1. We leave for future work investigating optimal weighting for these and other metrics, for example, by tuning them on downstream applications. When comparing these scores for both WN and CHA examples, Figure 1 (lower is better in both metrics) shows that WN has generally higher penalties both for example length and for usage of infrequent words. Specifically, for instance, we found that 80% of CHA's examples have a length penalty of .6 or less, whereas for the same proportion, the length penalty reaches more than .8 in WN. In a subsequent analysis, we found that these differences, if studied between WN's nouns and verbs, clearly favour nouns, that is, WN's nouns are in general accompanied by better examples. Specifically, we found that, on average, the length penalty is .49 for nouns, and .62 for verbs, and that the frequency penalty is .10 for nouns and .15 for verbs.

Finally, while the **sentence probability** is a valid metric, we observe that it is more likely that shorter sentences exhibit lower perplexity, and therefore will be scored higher by a language model. To further investigate this, we conduct an analysis where we split WN's and CHA's examples into 4 bins, namely *short*, *mid-short*, *mid-long* and *long*, with *short* examples containing between 1 and 15 tokens, *mid-short* up to 30 tokens, *mid-long* up to 45 tokens, and *long* above 45 tokens. Then, we compared the probability assigned by the language model to these examples, and verified that, indeed, WN has better short and mid-short examples, but worse mid-long examples. It also important to note that among the *long* examples, most of them were close to 45 tokens for WN, while for CHA they are much longer. To (perhaps anecdotally) illustrate this point, the longest dictionary example in WN is only 46 tokens long, while the longest in CHA is 141. Finally, in terms of usage of anaphoric references, we did not find significantly different results between WN and CHA.

4.2 Definition Modeling

Evaluating the quality of the generated definitions is a subjective matter, as delivering the meaning of words can take many forms. Table 2 shows examples of the predicted definitions generated by a WN-trained model and a CHA-trained model. When analysing these definitions and annotating the error types (following the typification proposed in Noraset et al. (2017)), it seems that the predicted definitions generated by the WN-trained model show evidence of under-specificity (first and second rows), since in each case the definition represents the general idea, but where part of the meaning of the target lemma in context is lost. In the third row, the generated definition falls into the self-reference type of error, since it refers to the same lemma in a circular way.

We also noticed that, generally speaking, the CHA-trained model learned to explicitly mention the prototypical concept or the idea to which a definition applies, and this is interesting from a commonsense learning point of view, which has recently received considerable attention (Gajbhiye et al., 2022; Nguyen et al., 2022). Therefore, given that CHA has many definitions that start with the prototypical concept/entity that embodies that prop-

No.	Lemma	Example	Gold definition	PD_WN	PD_CHA
(1)	accelerate	The car accelerated	move faster	become more powerful or efficient	of a vehicle or aircraft move forward at a high rate of speed
(2)	appear	Did your latest book appear yet?	be issued or published	have a physical form or appearance	of a book or other product reach the shelves of a bookstore or other store
(3)	immigrate	Many people immigrated at the beginning of the 20th century	come into a new country and change residency	become immigratory	of a person move to a foreign country to settle permanently

Table 2: Sample of predicted definitions generated by WN-trained model and CHA-trained model. PD_WN: predicted definition by WN-trained model, PD_CHA: predicted definition by CHA-trained model

	WN	CHA
BLEU	0.18	00.16
METEOR	12.28	14.89
ROUGE-L	16.49	17.37

Table 3: DM evaluation results for WN and CHA

	Nouns	Verbs
BLEU	3.67	0.47
METEOR	20.66	14.13
ROUGE-L	26.85	18.72
Average	1 7.06	11.12

Table 4: DM evaluation results for WN Nouns vs WN Verbs

erty (e.g., "*accelerate*" having a definition starting with "*of a vehicle*"), for the future, this resource could be helpful to map prototypical features to concepts, using dictionary examples as additional contexts.

We evaluated the definitions intrinsically using automatic string matching measures, specifically BLEU, ROUGE-L and METEOR. BLEU is a metric used for machine translation evaluation and compares n-grams matches of the candidate sentence with the reference sentence (Papineni et al., 2002) (we used the default BLEU-4). Rouge-L measures the longest common sub-sequence between the candidate sentence with the reference sentence (Lin, 2004). METEOR is another improved machine translation evaluation metric that matches uni-grams based on their surface forms, stemmed forms, and meanings (Lavie and Agarwal, 2007).

Exp. 1 (WN vs CHA) Table 3 shows the average BLEU, METEOR and ROUGE-L scores for the definitions generated by WN-trained model and CHA-trained model. The results show that the overall scores for evaluating the definition generation model that uses WN examples are low in general, even when comparing it with the model that uses CHA examples for training.

Exp. 2 (WN Nouns vs WN Verbs) Finally, with regards to the WN nouns vs WN verbs experiment, Table 4 shows the results of the three metrics used for evaluating the generated definitions. When comparing these results and the average of the scores, we can see that the quality of generated definitions of nouns is generally better than that of verbs. We leave for future work to further explore the differences between WN's noun vs verb examples, and why nouns seem to be easier to learn.

5 Conclusion

Definition modeling is the task to generate a dictionary definition given an input word and, optionally, some context. While different lexicographic resources are used as supervision for DM systems, there is little work analyzing their intrinsic quality. Our evaluation is focused on the examples available in WordNet and the Oxford Dictionary, where we train a sequence-to-sequence definition modeling architecture based on BART using these two dictionaries. We found that WN's dictionary examples are written in a style that may make them hard to learn (especially verbs), and that they are, generally, (perhaps too) short. For the future, we would like to explore extrinsic evaluations and perform additional experiments with other datasets and language models.

6 Acknowledgements

We thank the anonymous reviewers for their insightful comments. The experiments were executed using the computational facilities of the Advanced Research Computing @Cardiff (ARCCA) at Cardiff University.

References

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Tal August, Catharina Reinecke, and Noah Smith. 2022. generating scientific definitions with controllable commplexity.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. volume 2276, pages 136–145.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3779–3785. International Joint Conferences on Artificial Intelligence Organization.

Henri Bejoint. 2014. The bloomsbury companion to lexicography edited by howard jackson. *Dictionaries: Journal of the Dictionary Society of North America*, 35:374–381.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *EMNLP*, pages 1522–1532.

Bartosz Broda, Maciej Piasecki, and Stan Szpakowicz. 2009. A wordnet from the ground up. *Oficyna Wydawnicza Politechniki Wrocławskiej*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. 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 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Ting-Yun Chang, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xsense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks.

Pinzhen Chen and Zheng Zhao. 2022. A unified model for reverse dictionary and definition modelling. *arXiv preprint arXiv:2205.04602*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Luis Espinosa-Anke, Jose Camacho-Collados, Sara Rodríguez Fernández, Horacio Saggion, and Leo Wanner. 2016. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of COLING 2016: Technical Papers. The 26th International Conference on Computational Linguistics; 2016 Dec. 11-16; Osaka (Japan).[place unknown]: COLING; 2016. p. 900-10.* COLING.

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Eduard Hovy, and Noah Smith. 2014. Retrofitting word vectors to semantic lexicons.

Christiane Fellbaum. 2013. Wordnet. In Carol Chapelle, editor, *The encyclopedia of applied linguistics*, pages 6739–6746. Blackwell Publishing Ltd.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Amit Gajbhiye, Luis Espinosa-Anke, and Steven Schockaert. 2022. Modelling commonsense properties using pre-trained bi-encoders.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: literature review and dataset analysis. *Applied Computing and Intelligence*, 2(1):83–98.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. 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 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Kilgarriff, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus.

In *Proceedings of the 13th EURALEX International Congress*, pages 425–432, Barcelona, Spain. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Cunliang Kong, Yujie Wang, Ruining Chong, Liner Yang, Hengyuan Zhang, Erhong Yang, and Yaping Huang. 2022. Blcu-icall at semeval-2022 task 1: Cross-attention multitasking framework for definition modeling. *arXiv preprint arXiv:2204.07701*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

George Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–.

D.I. Moldovan and Rada Mihalcea. 2000. Using wordnet and lexical operators to improve internet searches. *Internet Computing, IEEE*, 4:34 – 43.

Jorge Morato, Miguel Marzal, Juan Llorens, and Jos Moreiro. 2004. Wordnet applications. *Proceedings of the 2nd Global Wordnet Conference*, 2004.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Tuan-Phong Nguyen, Simon Razniewski, Julien Romero, and Gerhard Weikum. 2022. Refined commonsense knowledge from large-scale web contents. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–16.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. pages 3259–3266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. *arXiv preprint arXiv:1603.06076*.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Ivan Vulić and Nikola Mrkšić. 2017. Specialising word vectors for lexical entailment. *arXiv preprint arXiv:1710.06371*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ruimin Zhu, Thanapon Noraset, Alisa Liu, Wenxin Jiang, and Doug Downey. 2019. Multi-sense definition modeling using word sense decompositions.

Michael Zock, Olivier Ferret, and Didier Schwab. 2010. Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, 13(4):201–218.