Generative Dictionary: Improving Language Learner Understanding with Contextual Definitions

Kevin Tuan[1], **Hai-Lun Tu**[2*] and **Jason S. Chang**[1]
[1]Department of Computer Science
National Tsing Hua University, Taiwan
[2]Department of Library and Information Science Fu Jen Catholic University, Taiwan
{kevintuan, helen.tu, jason}@nlplab.cc

Abstract

We introduce *GenerativeDictionary*, a novel dictionary system that generates word sense interpretations based on the given context. Our approach involves transforming context sentences to highlight the meaning of target words within their specific context. The method involves automatically transforming context sentences into sequences of low-dimensional vector token representations, automatically processing the input embeddings through multiple layers of transformers, and automatically generate the word senses based on the latent representations derived from the context. At runtime, context sentences with target words are processed through a transformer model that outputs the relevant word senses. Blind evaluations on a combined set of dictionary example sentences and generated sentences based on given word senses demonstrate that our method is comparable to traditional word sense disambiguation (WSD) methods. By framing WSD as a generative problem, *GenerativeDictionary* delivers more precise and contextually appropriate word senses, enhancing the effectiveness of language learning tools.

1 Introduction

The need for effective language mastery grows more critical as the world becomes increasingly interconnected. Reference resources like dictionaries are pivotal in language learning and vocabulary acquisition. Traditional dictionaries such as WordNet, Cambridge Learner Dictionary, and Macmillan Dictionary, curated by professional lexicographers, provide a solid foundation for understanding language. They organize word senses into related synsets or lists of meanings that help learners grasp the nuances of vocabulary, enhancing their communication skills. Traditional lexical resources typically list the most common senses of a word, sometimes overlooking figurative or less frequent usages (e.g., Figure 1 shows the list of word senses for "baton" as presented by WordNet). Users might struggle to identify the intended word sense in specific contexts, especially when the context extends beyond the provided sense inventory. This limitation underscores the need for more context-sensitive tools.

Consider the word "baton" in the sentence: "The successful passing of the [baton] demonstrated the team's ability to work collaboratively and manage responsibilities efficiently." In this context, "baton" does not refer to "a hollow cylinder passed from runner to runner in a relay race" but rather symbolizes "the responsibility for a person or a position." Traditional dictionaries like WordNet may not capture this symbolic meaning, making it difficult for learners to grasp the metaphorical connection.

We present a new system, *GenerativeDictionary*[1], that addresses this gap by interpreting words within their specific contexts. For example, in the sentence "The successful passing of the [baton] demonstrated the team's ability to work collaboratively and manage responsibilities efficiently," GenerativeDictionary identifies "baton" as "short staff symbolizing authority" By leveraging advanced Generative AI technology, this system fine-tunes pre-trained transformer models to analyze context and generate accurate word sense descriptions.

The GenerativeDictionary system enhances the usefulness of traditional dictionaries by providing concise, context-sensitive meanings. It bridges the gap between conventional word sense inventories and the nuanced interpretations required for effective communication. This innovative tool offers a more dynamic and practical approach to understanding language, making it an invaluable resource for learners and professionals alike.

čorresponding author

[1]http://joker.nlplab.cc:3000/

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 390–396

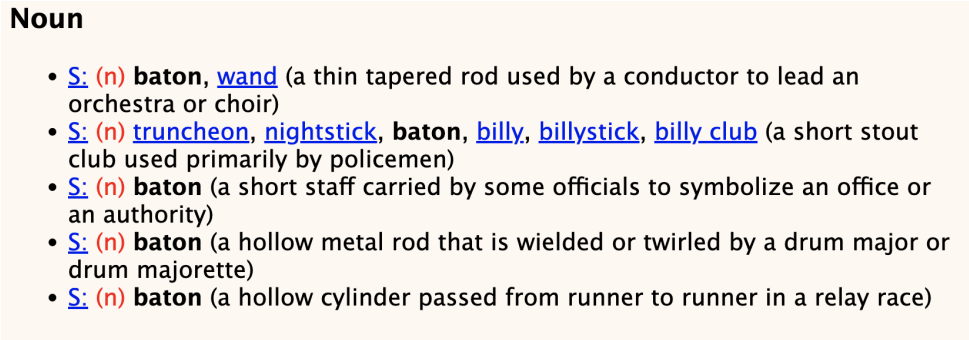

Figure 1: WordNet's sense inventory for the word "baton"

2 Related Works

Dictionaries play a crucial role in language learning, particularly in the area of vocabulary expansion. Studies had emphasize that a well-developed vocabulary is fundamental to language proficiency and effective communication, and dictionaries provide the necessary support for learners to access this vocabulary. (Nation and Nation, 2001; Laufer and Hulstijn, 2001) Furthermore, Schmitt and Schmitt (2020) suggests that engaging with dictionaries encourages autonomous learning and helps learners to become more effective at decoding unfamiliar words independently. Additionally, research by Knight (1994) illustrates that dictionaries aid in vocabulary learning by enabling learners to confirm their understanding of words and explore various meanings and contexts. These studies collectively highlight the pivotal role of dictionaries in enriching a learner's vocabulary and enhancing their overall language competence.

In our research, we represent word senses as concise, simple English descriptions (e.g., "steal – move quietly and secretly"), rather than an entry id in the sense inventory (e.g., "steal.v.2" in WordNet). More specifically, we focus on generating simple glosses for a target word in a given sentence for the purpose of assisted reading in language learning. The body of the WSD research most closely related to our work focuses on automatically classifying the target word in a given sentence into one of sense in pre-determined inventory, using information in the given sentence.

The advent of word embeddings has revolutionized WSD by providing low-dimensional dense vector representations that capture semantic relationships between words. Notably, word2vec (Mikolov et al., 2013) utilize skip-gram and continuous bag-of-words (CBOW) to enable models to capture syntactic and semantic properties of words from large corpora.

The lack of annotated data for WSD became increasingly evident as the capability of embedding and model architecture became increasingly sophisticated, prompting researchers to explore innovative solutions to enhance performance. One significant approach involved incorporating additional contextual data, exemplified by models like glossBERT (Huang et al., 2019), EWISER (Bevilacqua et al., 2020), and ARES (Scarlini et al., 2020). GlossBERT leveraged gloss definitions to enrich the context used by BERT, thus providing more comprehensive information about word senses. EWISER integrated synset embeddings from WordNet into its model, allowing it to utilize the rich semantic information encoded in these synsets. ARES utilized large-scale multilingual data to train sense embeddings, enhancing the model's ability to disambiguate words in various languages and contexts. Another approach aimed at reducing the complexity of sense inventories by compressing them, such as combining WordNet hypernyms. The ESR (Extended Synset Representation) (Song et al., 2021) model effectively merged similar senses into broader categories, thereby simplifying the inventory space. Additionally, researchers sought to generate annotated data from unannotated corpora. Techniques like distant supervision and semi-supervised learning enabled the automatic labeling of large text corpora, providing a substantial increase in training data without the need for extensive manual annotation.

Recent advances in pre-trained large language models (LLMs) have opened new avenues for reformatting the Word Sense Disambiguation problem into a generative problem, where the context is provided as input and the sense is generated as

Source	Words	Senses
Cambridge	49,521	80,666
Collins	177,238	322,199
Longman	41,015	80,796
Macmillan	40,766	75,091
Merriam-Webster	217,865	327,926
WordNet	148,730	206,978
Total	428,255	1,093,656

Table 1: Dictionary Data

Part of Speech	Words	Senses
NOUN	344,265	721,657
VERB	30,673	156,529
ADJ	66730	193,083
ADV	9,465	22,387

Table 2: Part of Speech Distribution

Dataset	Words	Senses	Sentences
DSD	56,784	190,833	181,369
Ex-DSD	71,302	224,504	662,640

Table 3: Definition-Sentence Dataset

output. The emergence of pre-trained transformer-based architectures (Vaswani et al., 2017), such as T5 (Raffel et al., 2020), has revolutionized Natural Language Processing by enabling models to understand and generate human-like text through extensive pre-training on large corpora. Specifically, T5 (Text-to-Text Transfer Transformer) has demonstrated the potential of framing various NLP tasks, such as translation, and summarization, as a text-to-text problem, thereby simplifying the modeling process. By leveraging these pre-trained models, we can alleviate the issue of limited annotated data, as these models possess a rich understanding of language nuances. This paradigm shift not only enhances the accuracy of WSD but also offers a scalable and adaptable solution, paving the way for more robust applications across different languages and domains.

In contrast to previous researches on word sense disambiguation, we present a system that automatically learns to generate a short gloss for a target word in a given sentence, by curating a collection of data to fine-tune a pretrained text-to-text model. We exploit the inherent regularity in dictionary definitions and examples to build a model for effective word sense interpretation.

3 Methodology

Our method can be summarized in a series of streamlined steps. First, we collect a comprehensive sense inventory from various lexicographic resources. Next, we simplify definitions that are excessively long or cumbersome, making them more accessible and easier to understand. To further enhance our dataset, we use generative AI to create additional example sentences based on the definitions. Finally, we train our model using this enriched dataset, which now includes both the simplified definitions and the newly generated example sentences. This systematic approach ensures a comprehensive and effective method for improving our context-based dictionary system.

3.1 Data Collection

We compiled a Comprehensive English Sense Inventory (CEI) from six lexicographic resources: Cambridge Dictionary[2], Collins Online Dictionary[3], Longman Dictionary of Contemporary English[4], Macmillan English Dictionary for Advanced Learners (Rundell, 2007), Merriam-Webster: America's Most Trusted Dictionary[5], and WordNet (Fellbaum, 2010).

From these resources, we gathered a total of 428,255 unique words associated with 1,093,656 senses. Table 1 and Table 2 summarize the information regarding CEI. From this dataset, we extracted 181,369 example sentences and format them into definition-sentence pairs dataset formatted as $\langle sent, defi \rangle$ pairs to form the Definition-Sentence Dataset (DSD). For each $\langle sent, defi \rangle$ pair, we surround the target word with square brackets to mark it. One example is listed below:

> Input sequence: The [dwindling] attendance at the meetings suggests a loss of interest among members.
> Output sequence: becoming gradually less

3.2 Definition Simplification

We identified that many dictionary definitions are excessively long or cumbersome. Simplifying

[2]https://dictionary.cambridge.org/
[3]https://www.collinsdictionary.com/
[4]https://www.ldoceonline.com/
[5]https://www.merriam-webster.com/

these definitions serve two primary purpose: simplicity and improved model performance.

For example, the Cambridge Dictionary defines one sense of "mortgage" as "a legal arrangement where you borrow money from a financial institution in order to buy land or a house, and you pay back the money over a period of years. If you do not make your regular payments, the lender normally has the right to take the property and sell it in order to get back their money." This could be simplified to "a loan secured by property."

Moreover, transformer models are known to degrade in coherency as the output grows longer and longer. Since the encoder packages all the information of the input sequence into a contextualized embedding, the model might forget earlier parts of the output sequence as the output grows longer. For example, without controlling the definition length, our model generates "a cabochon is shaped like a ring and is shaped like a ring, and is shaped like a ring with a ring on it" for the word "cabochon." By simplifying these definitions, we can improve the model's performance and obtain coherent outputs.

To achieve our goal, we tasked GPT-4-turbo with summarizing 236,275 senses into short definitions of seven words or less aimed at high school students level. We denote the modified dataset as Simplified-CEI (S-CEI).

3.3 Data Expansion

While dictionary editors typically provide examples for commonly used words or specific usages, this results in a limited coverage. Only 13.3% of the words and 17.4% of the senses in CEI has example sentences in the Definition-Sentence Dataset (DSD). Consequently, only a small fraction of the words and senses in our sense inventory are represented in the training dataset. We identified 99,342 senses in WordNet for which we intended to generate new sentences. For each word sense, we instructed gpt-4-turbo model to generate five sentences. To address this issue, we leveraged ChatGPT to generate additional sentences for given words and their specific definitions. In total, we obtained 495,914 additional $\langle sent, defi \rangle$ pairs and produced the Extended Definition-Sentence Dataset (Ex-DSD) (some of the generated sentences failed to include the target word.)

By incorporating these generated sentences into our training data, we improve the robustness and accuracy of our context-based dictionary system.

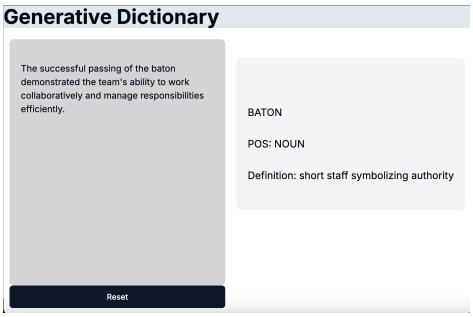

Figure 2: The interface of *GenerativeDictionary*

This data augmentation approach enhances both the breadth and depth of our training data, ensuring a more comprehensive representation of the words and senses in our inventory.

3.4 Model Training

In this work, we fine-tune the pre-trained T5 model (Text-to-Text Transfer Transformer) (Raffel et al., 2020) on the Definition-Sentence Dataset. The dataset is split into 70% training, 15% validation, and 15% testing. Due to computational constraints, we utilize the T5-base variant from the Hugging Face transformers library (Wolf et al., 2019). The T5-base model's architecture includes 12 encoder and decoder layers, with each block having 768 hidden sizes. In total, the model has 220 million parameters.

To evaluate the effectiveness of our data processing methods, we trained four different models using the same T5-base model but with variations in the training datasets. Our baseline model, T5-DSD, was trained on the original DSD without any modifications, maintaining the original sense inventory and the original $\langle defi, sense \rangle$ pairs. The second model, T5-Ex-DSD, was trained on the extended version of the DSD (Ex-DSD) but retained the original sense inventory. The third model, T5-S-DSD, was trained on the original DSD with Simplified definitions. The fourth model, T5-S-Ex-DSD, utilized the Ex-DSD with the Simplified Comprehensive English Sense Inventory, aiming to test the impact of sense length on model performance. These variations allow us to systematically analyze the contributions of data augmentation and sense inventory simplification to overall performance.

4 System

GenerativeDictionary features a simple and user-friendly interface. Users can write or copy-paste

the desired text into the textbox on the left. By pressing the "Enter" key, the text is submitted to the system. Users can then hover over any word in the text and click on it to see the word's part of speech and definition, which will be displayed in the text box on the right. Behind the scenes, our system identifies the target word and places brackets around it to identify it. The augmented sentence is then sent to the model, which generates the definition based on the context. This process allows *GenerativeDictionary* to provide precise and context-sensitive definitions, enhancing the user's understanding of the text.

5 Evaluations

To ensure our system produces results that are applicable in real life, we asked five English teachers to rate the generated definitions of a sample of DSD and Ex-DSD. Additionally, to benchmark our model's performance against the latest advancements in natural language processing, we compared our results with those generated by GPT-4-turbo. This comparison allowed us to gauge the relative effectiveness of our approach.

The test set consists of a total of 500 sentences, each containing a marked target word. Of these, 300 sentences are drawn from dictionary example sentences (DSD), while the remaining 200 sentences are generated using GPT-4 (Ex-DSD), based on the definitions of the target words.

We asked evaluators to assess the quality of the generated definitions for each sentence, using a grading scale from 0 to 2. This scale measures the degree of correctness, taking into account the precision of the definition and its suitability for the given context. A score of 0 is given when the generated definition is entirely incorrect or fails to capture the intended meaning of the target word.

A score of 1 is assigned when the definition is partially correct. In these cases, the definition may still convey the general sense of the target word, but could include issues such as incorrect part of speech, definitions that are overly broad or narrow. A score of 2 reflects an accurate and appropriate definition that aligns well with both the meaning of the target word and its usage in the sentence. Definitions awarded this score are not only correct but also contextually appropriate, effectively conveying the intended meaning without significant omissions or errors. The results are shown in Table 4.

Model	2	1	0
T5-DSD	.581	.222	.197
T5-Ex-DSD	.541	.286	.173
T5-S-DSD	.563	.289	.148
T5-S-Ex-DSD	.542	.322	.136
GPT-4	.643	.225	.132

Table 4: Expert Evaluations

Our evaluation yielded several key insights into the performance of our system and highlighted areas for further improvement. Firstly, expanding the datasets had a modest impact on improving scores of 1 or higher. While increasing the size of the datasets generally offers more training data and contextual information for the models, the observed gains in performance were incremental. This suggests that beyond a certain threshold, simply adding more data does not lead to substantial improvements in the model's accuracy or quality.

Moreover, the results indicated a slight decline in the number of definitions receiving a perfect score of 2 when sentences were generated rather than taken directly from dictionary examples. This suggests that the additional variables involved in generating sentences—such as variability in sentence structure, word usage, and contextual nuances—may introduce complexities that make it more challenging for the model to produce fully accurate and contextually appropriate definitions.

In our comparisons with the latest large language models (LLMs), specifically GPT-4-turbo, we found that our models are not yet on par with the performance of chat-GPT. However, this disparity is understandable given the differences in model architecture and data size. GPT-4-turbo benefits from extensive training on vast datasets, which is not feasible for smaller models like ours. Despite this, our models performed satisfactorily within their operational constraints.

Additionally, we observed that chat-GPT rarely utilized the same wording as the reference definitions. As a result, incorporating sequence similarity metrics could provide a more accurate assessment of semantic similarity between the generated definitions and the reference texts. This shift would allow us to better capture the essence of the definitions, even when the exact phrasing differs.

6 Conclusion and Futureworks

In this study, we introduced *Generative Dictionary*, a novel context-based dictionary system that enhances the user experience by providing context-sensitive definitions. By reframing Word Sense Disambiguation (WSD) from a traditional classification problem to a text-to-text generation task, we harnessed the power of pre-trained transformer models. These models, embedded with extensive lexical knowledge, generate definitions for ambiguous words, simplifying the WSD task and leveraging advancements in natural language processing for improved accuracy and applicability.

Our evaluation results demonstrate that our models perform well, especially when generating concise definitions. This improvement underscores the effectiveness of short text generation in addressing the challenges associated with longer outputs for T5-based models. However, a performance gap remains between our models and state-of-the-art large language models like GPT-4, primarily due to differences in model size and training data volume.

Future work should aim to bridge this gap by refining our evaluation metrics and exploring more sophisticated methods for sequence similarity, which could provide a more accurate measure of the semantic quality of the generated definitions. Enhancing our models' ability to produce precise and contextually appropriate definitions will be crucial for advancing WSD.

Overall, *Generative Dictionary* marks a significant step forward in WSD research, offering a user-centric approach that improves the dictionary experience by delivering relevant and context-aware definitions. By focusing on improved evaluation techniques and advanced similarity measures, future research can build on this foundation to achieve even greater performance and applicability in real-world scenarios.

Limitations

Following are some of the limitations we faced in this project:

1. We are limited by the computational resources available to us. One straightforward way to improve our model performance is to increase the model size, which we do not have the resources for. We believe this is the most important factor contributing to the performance gap between our system and GPT-4.

2. Another limitation is the inherent stability and degradation problems of transformer models. While we tried to alleviate this issue by simplifying the definitions, this process might introduce new errors. The final system is still enough that changing one word in the input sentence might drastically change the output.

3. Current evaluation relies solely on human evaluation of a relatively small set of test data. To measure performance on a large scale, we need a method that can automatically rate the performance of our models.

4. Our fine-tuning process is tailored specifically for the task of word sense disambiguation, which might limit the model's generalizability to other natural language processing tasks without further adjustments

References

Michele Bevilacqua, Roberto Navigli, et al. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the conference-Association for Computational Linguistics. Meeting*, pages 2854–2864. Association for Computational Linguistics.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Susan Knight. 1994. Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The modern language journal*, 78(3):285–299.

Batia Laufer and Jan Hulstijn. 2001. Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied linguistics*, 22(1):1–26.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ian SP Nation and ISP Nation. 2001. *Learning vocabulary in another language*, volume 10. Cambridge university press Cambridge.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *Journal of machine learning research*, 21(140):1–67.

Michael Rundell. 2007. *Macmillan English Dictionary for Advanced Learners*, 2nd edition.

Bianca Scarlini, Tommaso Pasini, Roberto Navigli, et al. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539. The Association for Computational Linguistics.

Norbert Schmitt and Diane Schmitt. 2020. *Vocabulary in language teaching*. Cambridge university press.

Yang Song, Xin Cai Ong, Hwee Tou Ng, and Qian Lin. 2021. Improved word sense disambiguation with enhanced sense representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4311–4320.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.