Wordnet for Definition Augmentation with Encoder-Decoder Architecture

Konrad Wojtasik, Arkadiusz Janz, Bartłomiej Alberski, Maciej Piasecki
Wrocław University of Science and Technology
{konrad.wojtasik|arkadiusz.janz}@pwr.edu.pl

Abstract

Data augmentation is a difficult task in Natural Language Processing. Simple methods that can be relatively easily applied in other domains like insertion, deletion or substitution, mostly result in changing the sentence meaning significantly and obtaining an incorrect example. Wordnets are potentially a perfect source of rich and high quality data that when integrated with the powerful capacity of generative models can help to solve this complex task. In this work, we use plWordNet, which is a wordnet of the Polish language, to explore the capability of encoder-decoder architectures in data augmentation of sense glosses. We discuss the limitations of generative methods and perform qualitative review of generated data samples.

1 Introduction

Transformer models have appeared to be very successful in solving a large variety of Natural Language Processing tasks and applications. The research on neural language modeling has been intensified in recent years and has yielded many new developments, such as pre-trained autoregressive language models for text generation. Text generation models such as BART (Lewis et al., 2020), GPT (Brown et al., 2020) or T5 (Raffel et al., 2020) have increased the performance even further, due to their few-shot abilities (Radford et al., 2019).

The knowledge resources such as wordnets (Miller et al., 1990) are often incomplete and still require constant development, especially for low-resourced languages. In Słowosieć (Dziob et al., 2019) (also called plWordNet) – a wordnet of the Polish language, one of the largest wordnets in the world – over 40% senses still lack a definition, and over 60% of senses do not have any sense use example. This area might be addressed by utilising large language models pre-trained on text generation tasks. Adding missing definitions and sense use examples is a crucial task for further wordnet development.

The definition generation problem is tightly interconnected with Word Sense Disambiguation (WSD) problem, as the words have different meanings in different contexts. The modern language models have significantly improved WSD performance in recent years. Transformer-based models such as BERT (Devlin et al., 2019) have proved to be very effective in contextual word sense recognition (Bevilacqua et al., 2021). While very effective, large language models require at least a small data sample to effectively fine-tune them for the WSD task. Nevertheless, large pre-trained language models with billions of parameters have been shown to require less training data to effectively tune them for downstream tasks (Chowdhery et al., 2022).

In this paper, we investigate generation abilities of large pre-trained language models in the task of wordnet gloss generation for the Polish language. We treat this problem as a data augmentation problem, as some senses in under-resourced wordnets are missing their definitions. We evaluate gloss generation performance on the example of Polish wordnet – Słowosieć (Dziob et al., 2019) – in the version 4.2.[1]

2 Related Work

The acquisition and completion of missing sense glosses has been addressed in the literature in many different ways. Enrichment of synset glosses in wordnets can be partially achieved by utilising machine translation models (Chakravarthi et al., 2019). However, these approaches do not take into account the discrepancy between sense inventories in different languages, as some senses do not exist in the source or target languages. Thus, an automated translation of Princeton WordNet glosses (Miller et al., 1990) to other language might not be able

[1] The code and the training data, as well as the generated sense definitions, are available at https://gitlab.clarin-pl.eu/knowledge-extraction/prototypes/gwc-t5-wordnet.

to completely solve the task of gloss completion. The other approaches rely on interlinking the wordnets with external resources and semantic networks such as multilingual thesauri in linked open data, Wikipedia[2], Wikidata[3], BabelNet (Navigli et al., 2021), or with Open Multilingual WordNet grid (Bond and Foster, 2013). Some solutions solve the problem as a joint task in which translations and potential glosses available in large semantic networks are analysed with WSD algorithms to increase the accuracy of gloss acquisition (Camacho-Collados et al., 2019). Still, an overall coverage of senses is strongly dependent on the target domain of application, and for specific domains the WSD models are biased towards more frequent senses. The closest to our work are generative approaches in which the encoder–decoder architectures are used to generate definitions in an autoregressive manner and treating the language models as knowledge bases (Huang et al., 2021; Mickus et al., 2021; Bevilacqua et al., 2020; Zhang et al., 2022). The approaches such as (Huang et al., 2021) utilise large pre-trained transformers, mainly T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) models, to generate definitions. The solution proposed in (Huang et al., 2021) is the closest to our work since it's based on the same pre-trained T5 transformer architecture, but the authors have added reranking models to control the specificity of generated sense definitions. In our work we expand the research on generative definition acquisition and investigate the performance of raw generative language models for the Polish language. The Japanese corpus for definition generation (Huang et al., 2022) also provides words with usage and definition, but it was generated via linking Wikidata items with sentences in Wikipedia articles.

3 Methods

3.1 Text Generation Models

Text generation task is formally defined as conditional sequence generation $\mathcal{Y} = (y_1, y_2, \ldots, y_M)$, where a model should predict sequence $\mathcal{Y}$ conditioned on the sequential input data $\mathcal{X} = (x_1, x_2, \ldots, x_P)$, with $p(\mathcal{Y}|\mathcal{X}) = p(y1, y2, \ldots, y_M|\mathcal{X})$. The models for text generation task usually descend from *sequence-to-sequence* architectures with sequential *encoders* and sequential *decoders*. Modern text generators such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), or GPT (Radford et al., 2018, 2019; Brown et al., 2020) utilise transformer networks and autoregressive decoders. In this work, we investigate text generation abilities of pre-trained T5 language models for Polish language, more specifically the `plT5` language models (Chrabrowa et al., 2022) pre-trained on Polish corpora.

[2] https://www.wikipedia.org/
[3] https://www.wikidata.org

3.2 Sense Definitions and Sense Examples

Following (Huang et al., 2021), we prepared a dataset of sense definitions and sense use examples for target words selected for the task of definition generation. Princeton WordNet has a great collection of glosses and sense examples, which have been frequently used in various natural language processing tasks, including word sense disambiguation (Huang et al., 2019; Bevilacqua and Navigli, 2020). Polish sense inventories, such as plWordNet, do not provide complete description of senses in terms of their glosses and sense use examples. Thus, we decided to incorporate sense annotated corpora from (Janz et al., 2022) and (Hajnicz and Bartosiak, 2019) to obtain a larger and diversified collection of sense definitions and their usage examples.

3.3 T5 for Definition Generation

Let $\mathcal{D} = \{(w, D, E)\}_{i=1}^{N}$ will be a dataset with instances representing a sense use example E and sense definitions D of a target word w and its sense $s \in \mathcal{S}_w$. Glosses D and a sense use examples E are defined as sequences of tokens $D = (d_1, d_2, \ldots, d_T)$ and $E = (e_1, e_2, \ldots, e_M)$. The senses and their textual descriptions are obtained from the sense inventory $s \in \mathcal{S}$. We use the data from plWordNet and additional sense-annotated corpora (see Section 3.2).

To fine-tune a model to the definition generation task for target words and their sense use contexts, we prepare the training data according to the methodology presented in (Raffel et al., 2020; Zhang et al., 2022) for the T5 model. A single training example consists of a word and its sense use example concatenated with a colon, e.g. „*cat: the cat was jumping on the bed in the middle of the night*". The target for T5 model represents the definition of the sense expressed by the given sense use example („*feline mammal usually having thick soft fur and no ability to roar, domestic cats*").

We split the dataset into two parts $(\mathcal{D}_L, \mathcal{D}_T)$, where $\mathcal{D}_L$ is a labeled training corpus for text generation model, and $\mathcal{D}_T$ is the held-out testing sample with lemmas outside the training set – lexical data split. The generation task is defined as follows.

$$p(D|E, w) = \prod_{t=1}^{T} p(D_t|w, D_{t-1}, \ldots, D_1, E)$$

4 Evaluation

Output of generative models was a definition for a given word in relation to the particular context and the evaluation of such an output is a nontrivial task. In language generation different evaluation metrics are used. We chose BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) metrics which are widely applied in many benchmarks. This automatic evaluation gave us information, if a model is overfitting to provided data or not. We could also estimate the difference between basic and large models performance on the test set. But to evaluate definitions properly, syntactic-level metrics are not sufficient. That is why we also performed manual validation of the generated definitions together with doing error analysis of the model's predictions. The manual validation was performed by professional lexicographers specialising in wordnets. We used a subset of error tags from (Huang et al., 2021) as a basis for our manual evaluation, namely:

- `self-reference` – error is assigned when a word being defined is described by using the word itself,
- `completely-wrong` – the word being defined has been assigned a definition representing as wrong sense,
- `partially-wrong` – some part of the generated definition is incorrect or refers to a different sense,
- `incoherent` – the definition contains contradictory parts.

To decrease memorisation impact on our evaluation, we evaluated the predictions by ensuring both the lemmas and the definitions in our test data were not included in the training dataset. We also provide the results with respect to part-of-speech of analysed lemmas.

Hard evaluation In this setting, a lexicographer accepts a generated definition if and only if any of the defined errors has not occurred in it.

Soft evaluation A generated definition is considered to be correct, even if the `self-reference` or `partially-wrong` errors have been spotted, but other errors are not observed.

4.1 Experimental Setting

We fine-tuned a pre-trained plT5 (Chrabrowa et al., 2022) generative language model for the task of definition generation. We trained `plT5-base` and `plT5-large` models available on HuggingFace[4] model repository. They have correspondingly 220 millions a parameters and 770 millions parameters. We trained them on single Nvidia RTX3090 GPU. The batch size for `plT5-base` was set to 16 and the model was trained for 40 epochs. In case of `plT5-large`, the batch size was set to 4 and the model was trained for 15 epochs, due to increased computational complexity of the model. We applied batch gradient accumulation steps for every 8 the batches and set a learning rate to 1e-4. The prompts of pre-selected T5 language models were set to `'[generate definition]'`.

4.2 Datasets

Training Data To train the models we used the following sense annotated corpora. The main dataset used for training was created from plWordNet's sense definitions and sense use examples.

- Verb's Valency Dictionary – Składnica (SK) is a sense-annotated treebank (Hajnicz, 2014) used as a benchmark dataset for knowledge-based WSD solutions for Polish language (Kędzia et al., 2015). The dataset was updated at *PolEval's WSD competition Task 3* (Janz et al.).
- The Corpus of Wroclaw University of Science and Technology (KPWr) (Broda et al., 2012) – contains the documents from various sources and represents different genres and domains. The manual sense annotation was based on a lexical sampling approach – the occurrences of words pre-selected by experts were manually annotated with senses in relation to their contexts (Broda et al., 2012; Kędzia et al., 2015). In (Janz et al.) the corpus

[4] `https://huggingface.co`

was extended with full-text sense annotation – 100 documents were manually tagged with plWordNet senses.

- Sherlock Holmes: The Adventure of The Speckled Band (SPEC) by Sir Arthur Conan Doyle, translated to Polish by a team of experts as a part of The NTU Multilingual Corpus (Tan and Bond, 2011). The corpus was manually tagged both with morphological information and sense tags (Janz et al.).

All of the aforementioned datasets are fully compatible with sense inventory of plWordNet 4.2, as they were described in (Janz et al., 2022). To improve the coverage of senses, we incorporated additional silver dataset built upon plWordNet Corpus 10.0 (Kocoń and Gawor, 2019), in short KGR10.

- Data Sample for Monosemous Lemmas – the KGR10 corpus is a corpus built from web-based data sources, covering a broad range of styles, genres and topics. It contains over 4 billion tokens with over 18 million distinct words. We synthesized a collection of additional sense use examples by extracting context windows from KGR10 corpus for senses representing potentially monosemous lemmas. To select monosemous lemmas we used plWordNet's sense inventory, mainly its multi-word expressions and lemmas with single sense and lower occurrence frequency in the corpus.

Test Data We prepared two distinct test sets for the evaluation. The first test set was prepared for manual evaluation, and the second test set was created to perform automated evaluation using BLEU and Rouge-L scores.

To create the test set for automated evaluation, we have split the data from plWordNet and sense-annotated corpora into training part and test part. We acquired almost 237k examples with words, usage examples and definitions. From those examples around 213k were acquired from plWordNet, 6.2k from The Corpus of Wroclaw University of Science and Technology (KPWr), 16k from Verb's Valency Dictionary, and 1.5k Sherlock Holmes. To create the test set, we randomly sampled 10k examples.

The test set for manual evaluation contained 146 examples with words and representative usage examples. We sampled these examples from the test set prepared for automated evaluation. All usage examples were new and were not seen by the model before. We split the data by words according to the following criteria. There were 102 instances that were already provided with expected sense definition in plWordNet. We denoted this subset as *WordNet+*. The subset of 44 words that had no definition in plWordNet was denoted as *WordNet-*. The examples were given to experts to measure defining capabilities of language models.

5 Results and Discussion

The results indicate that there is a significant difference between *base* and *large* model sizes. Our automatic evaluation results on 10k test set containing definitions from plWordNet, showed that BLEU score (see figure 1) and Rouge-L score (see figure 2) were getting better over time at higher pace for the *large* model than for the *base* model. The highest scores achieved after 13k iterations were (*0.31*, *0.44*) and (*0.44*, *0.54*) for BLEU score and Rouge-L score, respectively. The final difference in scores was greater than *0.1* for both metrics.

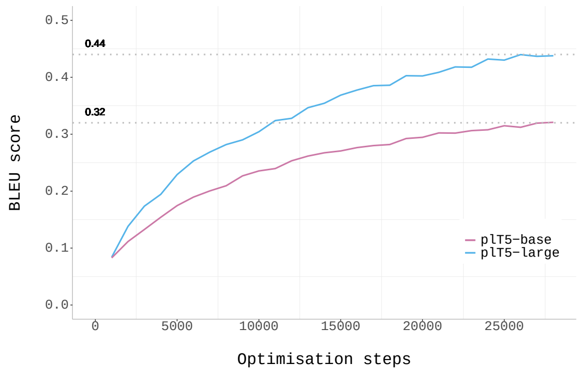

Figure 1: Evaluation of text generation models in the task of definition generation. We plot the performance of fine-tuned language models measured by BLEU score with respect to optimisation steps during fine-tuning. One iteration is equal to 256 shown examples.

The examples of generated definitions for provided contexts (see Table 1) showed different definition patterns. The first example represents the word *to devastate*. The model generated a correct definition explaining the meaning of analysed word. The second example, the word *to solve*, was explained using the word itself and passed the soft evaluation. However, the generated definition did not pass the hard evaluation test (*definiendum* case). The third example, the word *covered by*, had its meaning correctly explained by the generated definition in

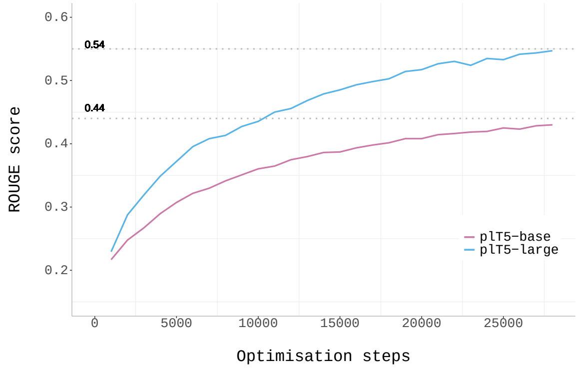

Figure 2: Evaluation of text generation models in the task of definition generation. We plot the performance of fine-tuned language models measured by ROUGE score with respect to optimisation steps during fine-tuning. One iteration is equal to 256 shown examples.

the given context, and the model did not repeat the existing definition from plWordNet. The fourth example, the word *tapir*, shows that the model was able to use previously acquired knowledge from Wikipedia pages or other knowledge bases (available at pre-training time) and created a new definition for that word, even though it was not present in plWordNet.

We also provided some examples of errors in the generated definitions (see Table 2). For the word *anesthetized*, the model resolved the first part of the definition correctly, but the second part was contradictory, because a person who is under anesthesia is out of touch with reality. The second example, the word *to guide*, was defined using the word itself, and was classified by the expert as incorrect. The third example represents the word *get involved*. It was defined in an unspecific way, and semantically the definition is only partially correct. In the fourth example, the word *snarky* not only defines itself, but the definition is wrong and the word is used in an incorrect sense.

The overall results are presented in Table 4. We measured the average accuracy of the model's predictions according to experts. There was a substantial difference between plT5-base and plT5-large models, where the larger model was better by more than 10 percent points in the overall evaluation. The words that existed already in plWordNet were easier to be defined and the unseen words seemed to be more challenging for the model. The main reason for that is that the model was able to memorize well seen texts and generated definitions accordingly, but for the unseen examples, we expected the model to generate definitions for meanings that have not been seen before.gw There were cases, where the model was able to access knowledge from pre-training corpora and use it to generate definition for new words. The accuracy difference between word examples that were not defined in plWordNet for hard evaluation and soft evaluation is almost doubled for both models. This may be the result of the model's understanding of the main concept of the word it defines, but at the same time hallucinating the definition, resulting in a formally incorrect definition. Another reason why the model performs poorly on unseen examples is that the sense-annotated resources for Polish language may be insufficient for learning to define senses completely unrelated to the training data and existing in completely different areas of plWordNet's lexico-semantic structure. This issue will be investigated in our future work.

6 Conclusion

The plWordNet definitions and usage examples are perfect data source for fine-tuning T5 generative models. Fine-tuned models are able to understand the new task and partially to generate definitions of unseen word senses. There is a significant difference in performance for seen and unseen words. Models are able to memorize definitions for examples in training set and modify them. Large model performs better than base, not only in automatic evaluation but also when evaluated by trained linguist. It is able to generalize better and recall more from it's memory.

As our results are promising, we believe that further research in this area will lead to models that are able to generalize better and define polysemic words properly in different contexts. Those models can be applied to WSD, Entity Linking and other knowledge intensive tasks.

Acknowledgements

The work was partially supported by (1) the Polish Ministry of Education and Science, CLARIN-PL; (2) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, (3) project CLARIN-Q (agreement no. 2022/WK/09), and (4) the Department of Artificial Intelligence at Wroclaw University of Science and Technology.

Word	Context	Generated Definition
`wyniszczać, verb`	*„Wreszcie świadczenia na rzecz gospodarki okazały się* ***wyniszczające****. Stopa akumulacji kapitału osiągnęła bezprecedensowy poziom.*"	`powodować, że coś nie nadaje się do użytku albo straciło jakieś swoje funkcje;`
(to devastate)	(*Finally, the benefits to the economy have proved* ***devastating*** *in the economic field: the rate of capital accumulation has reached unprecedented levels*)	`cause something to become unusable or lose some of its functions;`
`rozwiązywać, verb`	*„ Ich zdaniem obiekt tego rodzaju zostanie zrealizowany dopiero wówczas, gdy maszyna będzie w stanie* ***rozwiązywać*** *zadania, których dotąd nie udało się rozwiązać człowiekowi, i to nie na skutek braku szybkości i dokładność"*	`zapoznawać się z jakimś problemem, rozwiązywać go, odgadywać na podstawie jakichś przesłanek`
(to solve)	(*In their opinion, an object of this kind will be realized only when the machine will be able to solve tasks that humans have so far failed* ***to solve****, and not due to a lack of speed and accuracy.*)	`get acquainted with some problem, solve it, make a guess based on some premises`
`opieka, noun`	*„Klient Rekuperatory od pierwszego dnia funkcjonowania systemu objęty jest* ***opieką*** *Działu Serwisowego, który jest do dyspozycji od poniedziałku do piątku"*	`czynność, zabezpieczenie kogoś lub czegoś przed czymś, strzeżenie czegoś, chronienie przed zagrożeniami`
(covered by)	(*The customer of Rekuperatory from the first day of operation of the system is* ***covered by*** *the Service Department, which is available from Monday to Friday*)	`action, protecting someone or something from something, guarding something, protecting from threats`
`tapirek, noun`	*„****Tapirek*** *podszedł i otarł pysk o nogę Hala. Wielki wąż wpatrywał się w niego głodnym wzrokiem, a potem cofnął głowę i z potworną siłą uderzył nią o drzwi klatki."*	`tapir górski, Tapirus pinchaque - ssak z rodziny tapirowatych, najmniejszy spośród czterech gatunków i jedyny, który może żyć na wolności poza regionem lasów tropikalnych;`
(Tapir)	(***Tapir*** *approached and rubbed his snout against Hal's leg. The big snake stared at him with hungry eyes.*)	`mountain tapir, Tapirus pinchaque - a mammal of the tapir family, the smallest of the four species and the only one that can live in the wild outside the rainforest region;`

Table 1: A sample of correctly predicted definitions for test words and their contexts. The T5 model was provided with words and their sense use examples as an input for definition generation.

Word	Context	Generated Definition	Error Tag
znieczulony, adj	*„Nic nie czułem! — Był pan **znieczulony**, sir. Zresztą podobno to nie boli.”*	o człowieku: taki, który jest pod wpływem znieczulenia, ma kontakt z rzeczywistością;	**incoherent**
(anesthetized)	(*I didn't feel anything! - You were **anesthetized**, sir. Anyway, supposedly it doesn't hurt.*)	about a person: one who is under anesthesia, is in touch with reality;	
przyświecać, verb	*„W rzeczywistości zapewne **przyświecała** mi zupełnie bezsensowna myśl, żeby spróbować ułożyć z kamieni coś w rodzaju muru.”*	przyświecać, przyświecać komuś;	**definiendum**
(to guide)	(*In fact, I was probably **guided** by a completely nonsensical thought to try to arrange something like a wall out of the stones.*)	to guide, to guide someone;	
wikłać, verb	*„Niestety, dziewczyna pada ofiarą zakazanej miłości i **wikła się** w niebezpieczny romans.”*	przenośnie: pakować się w coś, dawać się wciągnąć czemuś lub komuś;	**partially-wrong**
(get involved)	(*Unfortunately, the girl falls prey to forbidden love and **get involved** in a dangerous affair.*)	metaphorically: get sucked into something or someone;	
warkliwy, adj	*„Gdy spytałem go o tego człowieka, przyznał **warkliwym** tonem, że jest najlepszy z całej zgrai.”*	taki, który jest pełen warczenia, warczący;	**completely-wrong**
(snarky)	(*When I asked him about the man, he admitted in a **snarky** tone that he was the best of the bunch.*)	One that is full of growling, growling	

Table 2: A sample of incorrectly predicted definitions for test words and their contexts. The predictions were tagged with error tags explaining possible error types in definition generation.

Polysemic Word	Context	Generated Definition
`zamek, noun`	*„Na górze widniał wielki, piękny* ***zamek*** *otoczony fosą.*"	`budowla mieszkalna, rezydencja pana, króla, księcia lub magnata.`
(castle)	(*On the mountain was a large, beautiful* ***castle*** *surrounded by a moat.*)	`A residential building, the residence of a lord, king, prince or magnate.`
`zamek, noun`	„ *Masz rozpięty* ***zamek*** *w spodniach.*"	`wyrób tekstylny, który ma zamek, jest skonstruowany w taki sposób, że można go zamknąć.`
(zipper)	(*You have the* ***zipper*** *of your pants undone.*)	`a textile product that has a lock is constructed in such a way that it can be closed.`
`zamek, noun`	*„Dorobił sobie klucz do* ***zamka****."*	`urządzenie do zamykania np. drzwi, szuflad, walizek.`
(lock)	(*He made up a key for the* ***lock****.*)	`A device for locking, for example, doors, drawers, suitcases.`
`zamek, noun`	*„Po raz któryś z kolei odciągnął* ***zamek*** *i zajrzał do komory nabojowej swego kałasznikowa."*	`mechanizm broni palnej, wyposażony w ruchomy zamek.`
(bolt)	(*For the umpteenth time, he pulled back the* ***bolt*** *and looked into the cartridge chamber of his kalashnikov.*)	`firearms mechanism, equipped with a movable bolt.`

Table 3: A sample of predicted definitions for polysemic word in polish language *zamek*.

Model	All samples		WordNet^{+}		WordNet^{-}	
	hard eval.	*soft eval.*	*hard eval.*	*soft eval.*	*hard eval.*	*soft eval.*
`plT5-base`	0.43	0.62	0.82	0.95	0.27	0.54
`plT5-large`	0.59	0.74	0.95	0.99	0.37	0.64

Table 4: Manual evaluation of T5-based definition generation models on test data sample of 200 words with examples. We provide the accuracy of text generation model for *hard evaluation* and *soft evaluation* settings. We split the evaluation into three distinct settings: i) WordNet^{+} – testing on senses with a proper definition in plWordNet, ii) WordNet^{-} – testing on senses which definitions are missing in plWordNet, iii) testing on all test samples.

References

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.

Michele Bevilacqua and Roberto Navigli. 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a free corpus of Polish. In *Proc. of the 8th International Conference on Language Resources and Evaluation*, pages 3218–3222, Istanbul, Turkey.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jose Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. 2019. Sensedefs: a multilingual corpus of semantically annotated textual definitions. *Language Resources and Evaluation*, 53(2):251–278.

Bharathi Raja Chakravarthi, Mihael Arcan, and John Philip McCrae. 2019. Wordnet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the second workshop on multilingualism at the intersection of knowledge bases and machine translation*, pages 1–7.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. 2022. Evaluation of transfer learning for polish with a text-to-text model. *arXiv preprint arXiv:2205.08808*.

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.

Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. 2019. plWordNet 4.1 – a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global Wordnet Conference*, pages 353–362.

Elżbieta Hajnicz. 2014. Lexico-semantic annotation of składnica treebank by means of PLWN lexical units. In *Proc. of the 7th Global Wordnet Conference*, pages 23–31, Tartu, Estonia.

Elżbieta Hajnicz and Tomasz Bartosiak. 2019. Connections between the semantic layer of walenty valency dictionary and plwordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 99–107.

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.

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2022. JADE: Corpus for Japanese definition modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6884–6888, Marseille, France. European Language Resources Association.

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

Arkadiusz Janz, Joanna Baran, Agnieszka Dziob, and Marcin Oleksy. 2022. A unified sense inventory for word sense disambiguation in polish. In *Proceedings of the International Conference on Computational Science: ICCS 2022)*, London, United Kingdom.

Arkadiusz Janz, Joanna Chlebus, Agnieszka Dziob, and Maciej Piasecki. Results of the poleval 2020 shared task 3: Word sense disambiguation. *Proc. of the PolEval 2020 Workshop*, page 65.

Paweł Kędzia, Maciej Piasecki, and Marlena Orlińska. 2015. Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies*, (15).

Jan Kocoń and Michal Gawor. 2019. Evaluating kgr10 polish word embeddings in the recognition of temporal expressions using bilstm-crf. *ArXiv*, abs/1904.04055.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Timothee Mickus, Mathieu Constant, and Denis Paperno. 2021. About neural networks and writing definitions. *Dictionaries: Journal of the Dictionary Society of North America*, 42(2):95–117.

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.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567.

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 of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Liling Tan and Francis Bond. 2011. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). In *Proc. of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 362–371, Singapore.

Hengyuan Zhang, Dawei Li, Shiping Yang, and Yanran Li. 2022. Fine-grained contrastive learning for definition generation. *arXiv preprint arXiv:2210.00543*.