Cross-Lingual Definition Generation from an mT5

Brandon Wilde
Montclair State University

Anna Feldman
Montclair State University

Jing Peng
Montclair State University

Abstract

Recent progress in definition modeling calls for an expansion into new territory. In this paper, we explore the little-studied prospect of cross-lingual definition generation from a resource-scarce perspective. We show that a small pretrained multilingual Text-to-Text Transfer Transformer (mT5) model can be transformed into a language-agnostic zero-shot definition generator, producing rudimentary English definitions for terms in multiple foreign languages. Throughout the project, several task-specific modifications to the model are devised and tested. We further recommend research paths that may progress the field of cross-lingual definition generation.

1 Introduction

The present study undertakes the task of generating definitions for words supplied (in context) in a different language. Although marginally similar to translation, this process bears the distinct advantage of not depending on previous translation data, further enabling the explanation of terms in less commonly spoken languages.

Researchers have rightly highlighted the prospects of cross-lingual definition generation benefiting language learners (Kong et al., 2020) and those documenting endangered languages (Bear and Cook, 2021), but the list of potential beneficiaries is longer still. This type of system could also serve as a translation aid, clarifying terms otherwise not easily translatable. It could serve as a means of censorship evasion, selectively reformulating terms while preserving semantic content. Such a system could also be a boon in automatic text simplification, which would benefit from the ability to not only shorten phrases, but also expand them with simpler terms when needed.

2 Previous Work

2.1 Monolingual Definition Modeling

The history of automatic definition procurement has largely developed in synchrony with other NLP tasks dominated by machine learning methods. In 2010, Navigli and Velardi began a template-driven approach to identify and extract definitions from raw text. As distributional language models became popular, attempts were made to map word embeddings to natural language glosses and vice versa by crafting embeddings directly from dictionary definitions (Hill et al., 2016; Bahdanau et al., 2017). The formulation of the task was then radically redefined by the seminal work of Noraset et al. (2017), as they pursued a purely generative means of defining words. Soon thereafter, researchers began to consider how polysemy and ambiguity could be accounted for. Some (Zhu et al., 2019) experimented with multi-sense embeddings. Others incorporated local (example sentence) context (Ni and Wang, 2017) as well as global context from outside sources (Ishiwatari et al., 2019). Latent variable modeling was also investigated as a means of allowing multiple embeddings per word (Gadetsky et al., 2018; Kabiri, 2020). While all of the previously mentioned studies, except for Navigli and Velardi's, were accomplished with Recurrent Neural Networks (RNN) or Long-Short Term Memory (LSTM) models, later work (Bevilacqua et al., 2020; Chang and Chen, 2019; Fan et al., 2020; Kong et al., 2020; Mickus et al., 2019; Reid et al., 2020) has experimented with transformer architectures (Vaswani et al., 2017).

2.2 Cross-lingual Definition Modeling

The majority of research in definition modeling has been conducted with English-language data and models, with some notable exceptions exploring similar methods in other languages (Fan et al., 2020; Kabiri, 2020; Mickus et al., 2020; Yang et al.,

2020; Zheng et al., 2021). Bear and Cook (2021) recently demonstrated the value of an effective cross-lingual definition modeling system for the preservation of an endangered low-resource language. Their model was devised and trained much like a machine translation (MT) task. Their language of interest, Wolastoqey, is characterized by a polysynthetic concatenative morphology, meaning that single Wolastoqey words often translate to full English phrases or sentences, blurring the line between translation and definition. A small cross-lingual dictionary of Wolastoqey words and their English definitions was used to train a LSTM model, which was then utilized to generate more definitions, expanding the dictionary. Kong et al. (2020) explored the prospects for transfer learning to impact this domain by adapting large pretrained multilingual language models to the task of definition generation. This was done by training variations of mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) on an English definition data set and then testing its ability to define Chinese words in context. The model they created is most similar to that devised for the present study.

3 Model

The present study aims to further investigate the suitability of pretrained language models, especially for use in resource-light settings. A multilingual variant of the T5 model (Xue et al., 2021) built by Raffel et al. (2020), the mT5, similarly reframes all tasks as text-to-text transformations, theoretically enabling its natural adaptation to any generative NLP task. The model was pretrained on raw text from the multilingual Colossal Clean Crawled Corpus (mC4), incorporating text from over 100 languages (Xue et al., 2021). Although not all languages were equally represented in the training data, their inclusion suggests at least some capacity for the system to properly handle text in those languages given sufficient downstream fine-tuning. The mT5 model thus seems well-suited for adaptation to a relatively novel task such as definition modeling, and its inclusion of several languages and publication in several sizes suggests itself as an optimal tool for use in low-resource settings. The smallest published version, mT5-small, is accessed via the Hugging Face model hub and used for this study (Wolf et al., 2020). Modest adjustments to the model were tested, as will be described in section 5.

4 Data

The mT5 model did not include definition generation among its pretraining tasks, so task-specific fine-tuning is critical in fitting the model for this role. Due to the immaturity of the field and the diversity within the cross-lingual definition generation task, there are not yet any standardized evaluation regimes in place. The similarly designed Kong et al. (2020) research was dedicated to generating simple definitions, which is not a priority in the present case.

4.1 Definition Data

No suitable cross-lingual definition data set was found, so a high-quality publicly available English dataset was targeted. The fine-tuning data for this study were obtained from the SEMEVAL shared task CODWOE (COmparing Dictionaries and WOrd Embeddings) (Mickus et al., 2022). The data consist of approximately 64,000 training examples, where each example includes an English word and its gloss as well as its part of speech, an example sentence and several pretrained word embeddings for the given word. We make use of the word, gloss, and example sentence for each training example and use these English data as the basis for model fine-tuning.

In order to test the model's performance generating definitions in a different language than was input, an 1,800-example subset of the CODWOE data set was translated into German. The translation was obtained via DeepL Translator, an MT software highly regarded for its English-German translations [1]. Following translation of the words to be defined, hereafter termed *definienda* (singular: *definiendum*), and their example sentences, the German data set was filtered to remove any instances in which the example sentence did not contain the German definiendum or a similar enough form of the word (minimum edit distance ≤ 1). The data set decreased to approximately 900 examples and was then split evenly into a validation set and a test set. The remaining data (about 63,000 English examples) were used as the model training set.

4.2 Translation Data

German-English translation samples from the WMT-16 data set (Bojar et al., 2016) were also accessed and trimmed to 10,000 sentence pairs in order to simulate a low-resource setting. These data

[1] https://www.deepl.com/translator

were used experimentally as a training supplement while fine-tuning the definition generation model.

5 Fine-Tuning

Several fine-tuning strategies were tested in order to ascertain the optimal approach for producing a cross-lingual definition generator from a pretrained mT5 model. Variations included fine-tuning on different data types, differing input formats, selectively adapting the model's cross attention mechanism, and constraining the outputs.

5.1 Data Types

Due to the scarcity of cross-lingual dictionary data, we operate under the assumption that such data will typically not be available to developers of cross-lingual definition generation systems. In lieu of data matched perfectly to the task, monolingual dictionaries and translation data are posited as potential substitutes.

The model in this study is, in one experiment, first fine-tuned on German-to-English translation data and then further tuned with English definition data. We compare this against a model fine-tuned solely on English definitions. After each of these fine-tuning schemas were completed, we further test the effect of priming the model with a few handcrafted cross-lingual definition examples.

5.2 Input Formats

Previous research using definienda in context indicated the definiendum in several ways. Two methods are assessed here: (1) prepending the definiendum to the example sentence as demonstrated by Chang and Chen (2019) and Kong et al. (2020); and (2) demarcating the definiendum with special-purpose sentinel tokens. We additionally test the model with both of these methods employed concurrently.

5.3 Adapted Cross-Attention

Inspired by Mickus et al. (2019), we devise a cross-attention mask to restrict the information passed to the model decoder stack. A given example sentence is first passed through the model encoder stack with the definiendum demarcated by sentinel tokens. The encoders output a sequence of contextualized embeddings. A cross-attention mask is then applied such that all embeddings outside of the sentinel token span (i.e. embeddings not associated with the definiendum) are multiplied by 0 prior to the model's decoding phase, thus disabling them from passing on any information. We hypothesize that the encoding phase of the model will imbue sufficient contextual information to the definiendum to disambiguate the word sense from its contextualized embedding(s) alone.

In conjunction with the cross-attention mask application, we also manipulate the residual connection weights after each self-attention layer in the encoder stack. Whereas the mT5 model (and presumably all other typical transformer models) customarily adds each self-attention vector to its attending token's vector, we re-weight this sum in order to decrease the effect of self-attention and the contextualization it confers. Multiple weight coefficients are tested. The intent here is to ensure that the "semantic" content of the definiendum's original token embedding(s) is largely preserved and only slightly augmented by contextualization.

5.4 Constrained Outputs

The final system adjustments deal with the constraints placed on the decoder stack during text generation. Beam search was used to generate the most probable definition for each example. The model was then further limited so as not to produce any bigram more than once within a single output nor to reproduce any definiendum within its own definition.

6 Evaluation

6.1 Metrics

Bilingual Evaluation Understudy (BLEU) scores and perplexity are selected to assess our definition modeling system. The BLEU method measures the proportion of n-grams that two texts have in common, and is a common metric for sequence-to-sequence NLP tasks, owing to its frequent correlation with grammaticality and meaning preservation (Papineni et al., 2002; Štajner et al., 2014). Unfortunately, the n-gram comparison method also causes these scores to become less reliable when dealing with very short or highly subjective outputs. Perplexity is simply the exponential of the loss for the validation/test set. Its validity is not dependent on the length of output, per se, but rather reflects a proxy of the model's uncertainty given some input.

7 Results

Early models were allowed to fine-tune for up to 17 epochs, but peak performance was usually reached

at about 3 epochs, so subsequent models were only fine-tuned for 3 epochs, with occasional 6-epoch comparisons. Initial model evaluations including a cross-attention mask took longer to converge, but reached comparable BLEU scores after twice the fine-tuning period (6 epochs in total). A qualitative comparison of the results, however, suggested that the definitions generated with a cross-attention mask were more generic than the others. Rather than homing in on the definiendum, the generated definitions were often overgeneralized (e.g., "Of or pertaining to a certain manner"). We hypothesized that the encoder stack redistributed semantic content among the tokens in a non-intuitive way, such that the contextualized embeddings no longer mapped well to the original token embeddings. This impression led to the introduction of the residual connection reweighting, an attempt to scale back the extent of semantic redistribution. Unfortunately, reweighting the residual connections did not improve the overgeneralized outputs, and the BLEU and perplexity scores only worsened. These configurations were not tested for fine-tuning periods longer than 6 epochs. Further testing investigated how best to indicate the definiendum without using cross-attention masks or residual connection reweighting.

The model configurations which produced the best results with the validation set were subsequently evaluated with the reserved German-English test set. Results are summarized in Table 1, where the top-performing model is referred to as Model X. Hyperparameters and example outputs for this model are included in Appendix Table A1 and Table A2, respectively. Because the models were fine-tuned on solely English data, their evaluation on other languages constitutes a zero-shot inference task. We further test 1-, 5-, and 10-shot inference by exposing Model X to 1-10 German-English examples prior to evaluation, updating model parameters after each example with a linearly decreasing learning rate starting at 0.0003. The 3-run average for each scenario is included in Table 2. Priming the model with 10 cross-lingual examples in one's language of interest may increase the average BLEU score of subsequent model outputs by up to 5%. A dual fine-tuning strategy entailing preliminary fine-tuning on a translation task prior to the cross-lingual definition task did not improve model performance. No model in this scenario was fine-tuned on definition data for longer than 3 epochs.

Model	BLEU ↑	PPL ↓
Model X	**2.323** (0%)	**12.944** (0%)
Model Y	2.124 (-9%)	14.037 (+8%)
Model Z	2.086 (-10%)	13.870 (+7%)

Table 1: BLEU and perplexity scores for the top-performing models, as well as percentage difference from Model X. Model Y differs from Model X by a lower learning rate (0.0003). Model Z differs from Model X by a lower learning rate (0.0002) and not prepending the definiendum to the example sentence.

Priming	BLEU ↑	PPL ↓
0-shot	2.323 (0%)	12.944 (0%)
1-shot	2.228 (-4.1%)	12.935 (+0.2%)
5-shot	2.325 (+0.1%)	12.935 (-0.1%)
10-shot	**2.444** (+5.2%)	**12.876** (-0.5%)

Table 2: BLEU and perplexity scores for various priming strategies applied to Model X. Priming entailed exposing the model to 1-10 German-English examples prior to evaluation. Percentages indicate comparison to the base Model X.

8 Conclusion

This study has demonstrated the effectiveness of the mT5-small model as the basis for a cross-lingual definition generation system and explored means for improving system performance. The devised cross-attention mask and residual connection reweighting scheme did not improve model performance, but these may yet prove beneficial if applied throughout a model's pretraining. An alternative approach may be to first fine-tune the encoder layers with reweighted residuals and afterwards fine-tune the decoder layers on the definition generation task with a cross-attention mask applied. The model performed well regardless of whether the definiendum was demarcated within the example sentence or simply prepended to the front of the sentence. Indicating the definiendum in both ways concurrently seems an effective, if inelegant, method of up-weighting the most important tokens.

This study used minimal language resources, suggesting that a similar model could easily be developed for language pairs other than German-English. Considering that the mT5-small model is also the smallest in a family of related language models, it is likely that the results outlined here have substantial prospect for scalability.

Acknowledgements

This research has been funded by NSF grant 1704113 (A Linguistically-Informed Approach for Measuring and Circumventing Internet Censorship).

References

Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

Diego Bear and Paul Cook. 2021. Cross-lingual wolastoqey-English definition modelling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 138–146, Held Online. INCOMA Ltd.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7059–7069.

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.

Qinan Fan, Cunliang Kong, Liner Yang, and Erhong Yang. 2020. bert (chinese definition modeling based on bert and beam seach). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 336–348.

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.

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.

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.

Arman Kabiri. 2020. *A multi-sense context-agnostic definition generation model evaluated on multiple languages*. Ph.D. thesis, University of New Brunswick.

Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. 2020. Toward cross-lingual definition generation for language learners. *arXiv preprint arXiv:2010.05533*.

Timothee Mickus, Mathieu Constant, and Denis Paperno. 2020. Génération automatique de définitions pour le français (definition modeling in French). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 66–80, Nancy, France. ATALA et AFCP.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2022. SemEval-2022 Task 1: Codwoe – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. 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.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden. Association for Computational Linguistics.

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.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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.

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*, 21:1–67.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2020. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6331–6344.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10.

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, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021. Decompose, fuse and generate: A formation-informed method for Chinese definition generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online. Association for Computational Linguistics.

Ruimin Zhu, Thanapon Noraset, Alisa Liu, Wenxin Jiang, and Doug Downey. 2019. Multi-sense definition modeling using word sense decompositions. *arXiv preprint arXiv:1909.09483*.

A Supplemental Tables

Model Hyperparameter	Setting
Learning rate	0.0004
Learning rate scheduler type	Linear
Optimizer	AdamW
Mask context	False
Fine-tuning epochs	3
Residual weight	0.5 (default)
Definiendum indication method	Mark, Prepend
Number of beams	4
Nonrepeatable n-gram size	2 (bigram)

Table A1: Model X configuration.

Input sentence (and translation):	Generated definition:	Reference definition:
In diesem Jahr sind viele Arbeitstage durch **Krankheit** verloren gegangen. (Many working days this year have been lost through **illness**.)	The condition of being diseased.	A state of bad health or disease.
Häuser in New Mexico, Kalifornien und Florida zeigen einen starken **hispanischen** architektonischen Einfluss. (Houses in New Mexico , California and Florida exhibit a strong **Hispanic** architectural influence.)	Of or pertaining to the Spanish language.	Of or relating to a Spanish-speaking people or culture , as in Latin America.
Der **Änderungsantrag** wird nun zur Diskussion gestellt. (The **motion** to amend is now open for discussion.)	A formal announcement, especially a formal request.	A parliamentary action to propose something. A similar procedure in any official or business meeting.
die eigene Autorität zu **behaupten** (to **assert** one 's authority)	To make known; to recognize.	To use or exercise and thereby prove the existence of.

Table A2: Example definitions generated by Model X. Definienda are in red and reference data are in blue.