Evaluating a Multi-sense Definition Generation Model for Multiple Languages

Arman Kabiri and Paul Cook

Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3 Canada
{arman.kabiri,paul.cook}@unb.ca

Abstract. Most prior work on definition modelling has not accounted for polysemy, or has done so by considering definition modelling for a target word in a given context. In contrast, in this study, we propose a context-agnostic approach to definition modelling, based on multi-sense word embeddings, that is capable of generating multiple definitions for a target word. In further contrast to most prior work, which has primarily focused on English, we evaluate our proposed approach on fifteen different datasets covering nine languages from several language families. To evaluate our approach we consider several variations of BLEU. Our results demonstrate that our proposed multi-sense model outperforms a single-sense model on all fifteen datasets.

Keywords: Definition modelling · Multi-sense embeddings · Polysemy

1 Introduction

The advent of pre-trained distributed word representations, such as [12], led to improvements in a wide range of natural language processing (NLP) tasks. One limitation of such word embeddings, however, is that they conflate all of a word's senses into a single vector. Subsequent work has considered approaches to learn multi-sense embeddings, in which a word is represented by multiple vectors, each corresponding to a sense [3,10]. More recent work has considered contextualized word embeddings, such as [5], which provide a representation of the meaning of a word in a given context.

Definition modelling, recently introduced by [16], is a specific type of language modelling which aims to generate dictionary-style definitions for a given word. Definition modelling can provide a transparent interpretation of the information represented in word embeddings, and has the potential to be applied to generate definitions for newly-emerged words that are not yet recorded in dictionaries.

The approach to definition modelling of [16] is based on a recurrent neural network (RNN) language model, which is conditioned on a word embedding for the target word to be defined, specifically pre-trained word2vec [12] embeddings. As such, this model does not account for polysemy. To address this limitation, a number of studies have proposed context-aware definition generation models

[15,7,9,11,4]. In all of these approaches, the models generate a definition corresponding to the usage of a given target word in a given context.

In contrast, in this paper we propose a context-agnostic multi-sense definition generation model. Given a target word type (i.e., without its usage in a specific context) the proposed model generates multiple definitions corresponding to different senses of that word. Our proposed model is an extension of [16] that incorporates pre-trained multi-sense embeddings. As such, the definitions that are generated are based on the senses learned by the embedding model on a background corpus, and reflect the usage of words in that corpus. Under this setup — i.e., generating multiple definitions for each word corresponding to senses present in a corpus — the proposed definition generation model has the potential to generate partial dictionary entries. In order to train the proposed model, pre-trained sense vectors for a word need to be matched to reference definitions for that word. We consider two approaches to this matching based on cosine similarity between sense vectors and reference definitions.

Recently, [19] propose a multi-sense model for generating definitions for the various senses of a target word. This model utilizes word embeddings and coarse-grained atom embeddings to represent senses [1], in which atoms are shared across words. In contrast, we only rely on fine-grained multi-sense embeddings. To match sense vectors to reference definitions during training, [19] propose a neural approach, and also consider a heuristic-based approach that incorporates cosine similarity between senses and definitions. Our proposed approach to this matching is similar to their heuristic-based approach, although we explore two variations of this method. Furthermore, [19] only consider English for evaluation, whereas we consider fifteen datasets covering nine languages.

Following [19] we evaluate our proposed model using variations of BLEU [17]. We evaluate our model on fifteen datasets covering nine languages from several families. Our experimental results show that, for every language and dataset considered, our proposed approach outperforms the benchmark approach of [16] which does not model polysemy.

2   Proposed Model

Here we briefly describe the model of [16], referred to as the base model, and then present our proposed multi-sense model which builds on the base model.

The base model is an RNN-based language model which, given a target word to be defined (w^*), predicts the target word's definition ($D = [w_1, ..., w_T]$). The probability of the tth word of the definition sequence, w_t, is calculated based on the previous words in the definition as well as the word being defined, as shown in Equation 1.

$$P(D|w^*) = \prod_{t=1}^{T} p(w_t|w_1, ..., w_t - 1, w^*)$$ (1)

The probability distribution is estimated by a softmax function. The model further incorporates a character-level CNN to capture knowledge of affixes. A full explanation of this model is in [16].

In the base model, the target word being defined (w^*) is represented by its word2vec word embedding. This reliance on single-sense embeddings limits the model's ability to generate definitions for different senses of polysemous target words. To address this limitation, we propose to extend the base model by incorporating multi-sense embeddings, in which each word is represented by multiple vectors which correspond to different meanings or senses for that word. Specifically, we replace w^* in Equation 1 by a sense of the target word, represented as a sense vector.

Most prior work on definition modelling has considered polysemy through context-aware approaches [15,7,9,11,4] that require an example of the target word in context for definition generation. In contrast, the model we propose is context agnostic (as is the base model) and is able to generate multiple definitions for a target word without requiring that specific contexts of the target word be given in order to generate definitions.

The base model is trained on instances consisting of pairs of a word — represented by a word2vec embedding — and one of its definitions, i.e., from a dictionary. Our proposed approach is trained on pairs of a word sense — represented as a sense vector — and one of the corresponding word's definitions. In order to train our proposed approach, we require a way to associate pre-trained sense vectors with dictionary definitions, where the number of sense vectors and definitions is often different for a given word.

We consider two approaches to associating sense vectors with definitions: definition-to-sense and sense-to-definition. For both approaches we require a representation of definitions. We represent a definition as the average of its word embeddings, after removing stopwords. For each word in the training data, we then calculate the pairwise cosine similarity between its sense vectors and definitions. For definition-to-sense, each definition is associated with the most similar sense vector for the corresponding word. For sense-to-definition, on the other hand, each sense is associated with the most similar definition. For both approaches, the selected sense–definition pairs form the training data.

These approaches to pairing senses and definitions are only used to create training instances. At test time, to generate definitions for a given target word, each sense vector for the target word is fed to the definition generation model, which then generates one definition for each of the target word's sense vectors.

3   Materials and Methods

In this section, we describe the datasets, word and sense embeddings, and evaluation metrics used in our experiments.

3.1   Datasets

In this work, we conduct a multi-lingual study of definition modelling. We extract monolingual dictionaries for nine languages covering several language families, from three different sources: Wiktionary,[1] OmegaWiki,[2] and WordNet [13].

Wiktionary is a free collaboratively-constructed online dictionary for many languages. The structure of Wiktionary pages is not consistent across languages. Extracting word–definitions pairs from Wiktionary pages for a given language requires a carefully-designed language-specific parser, which moreover requires some knowledge of that language to build. We therefore use publicly-available Wiktionary parsers. We use WikiParsec for English, French, and German,[3] and Wikokit for Russian,[4] to extract word–definition pairs for these languages.

OmegaWiki, like Wiktionary, is a free collaborative multilingual dictionary. In OmegaWiki data is stored in a relational database, and so language-specific parsers are not required to automatically extract words and definitions. We extract the word–definition pairs from OmegaWiki for English, Dutch, French, German, Italian, and Spanish — the six languages with the largest vocabulary size in OmegaWiki — using the BabelNet Java API [14].

Finally, we consider WordNets. We only use WordNets for which the words and definitions are in the same language. We again use the BabelNet Java API to extract the word–definition entries from English [13], Italian [2], and Spanish [6] WordNets. We separately extract word–definition pairs from Greek [18] and Japanese [8] WordNets.

Properties of the extracted datasets are shown in Table 1. Each dataset is partitioned into train (80%), dev (10%), and test (10%) sets. We ensure that, for each word in each dataset, all of its definitions are included in only one of the train, dev, or test sets, so that models are only evaluated on words that were not seen during training.

3.2   Word and Sense Embeddings

Following [16], we use word2vec embeddings in the singe-sense definition generation model (i.e., the base model). For the proposed multi-sense models, we utilize AdaGram embeddings [3]. AdaGram is a non-parametric Bayesian extension of Skip-gram which learns a variable number of sense vectors for each word, unlike many multi-sense embedding models which learn a fixed number of senses for every word. Note that although here we use AdaGram, any multi-sense embedding method could potentially be used.[5]

For each language, word2vec and AdaGram embeddings are trained on the most recent Wikipedia dumps as of January 2020.[6] We extract plain text from

---

[1] https://en.wiktionary.org

[2] http://www.omegawiki.org

[3] https://github.com/LuminosoInsight/wikiparsec

[4] https://github.com/componavt/wikokit

[5] In preliminary experiments with MUSE embeddings [10] we found MUSE to perform poorly compared to AdaGram, and so only report results for AdaGram here.

[6] https://dumps.wikimedia.org

Table 1. The number of words, and proportion of polysemous words (PPW) in each dataset.

Language	Omega		Wiktionary		WordNet	
	#Words	PPW	#Words	PPW	#Words	PPW
Dutch	13093	0.18	–	–	–	–
English	17000	0.20	17000	0.27	20000	0.18
French	15869	0.17	20000	0.26	–	–
German	13338	0.12	16000	0.26	–	–
Greek	–	–	–	–	11517	0.26
Italian	18351	0.21	–	–	16290	0.22
Japanese	–	–	–	–	20000	0.30
Russian	–	–	15000	0.17	–	–
Spanish	17000	0.19	–	–	18934	0.12

these dumps, and then pre-process and tokenize the corpora using tools from AdaGram,[7] modified for multilingual support, except in the case of Japanese where we use the Mecab tokenizer.[8] The resulting corpora range in size from roughly 86 million tokens for Greek to 3.7 billion tokens for English. The same pre-processing and tokenization is also applied to the datasets of words and definitions extracted from dictionaries.

We train word2vec embeddings using Gensim with its default parameters.[9] We also use the default parameter settings for AdaGram. To obtain representations for words, as opposed to senses, from AdaGram sense embeddings, as required to form representations for definitions (Section 2), we take the most frequent sense vector of each word (as indicated by Adagram) as the representation of the word itself.

3.3   Evaluation Metrics

BLEU [17] has been widely used for evaluation in prior work on definition modelling [16,9,15]. BLEU is a precision-based metric that measures the overlap of a generated sequence (here a definition) with respect to one or more references. For multi-sense models, we calculate BLEU as the average BLEU score over each generated definition.

While BLEU is appropriate for evaluation of single-sense definition generation models, it does not capture the ability of a model to produce multiple definitions corresponding to different senses of a polysemous word. We therefore also consider a recall-based variation of BLEU, known as rBLEU, in which the generated and reference definitions are swapped [19], i.e., the overlap of a reference definition is measured with respect to the generated definition(s). For

---

[7] https://github.com/sbos/AdaGram.jl/blob/master/utils/tokenize.sh

[8] https://github.com/jordwest/mecab-docs-en

[9] https://radimrehurek.com/gensim/

each target word, we calculate rBLEU as the average rBLEU score for each of its reference definitions (for both single and multi-sense models).

In addition to precision-based BLEU, and recall-based rBLEU, we report the harmonic mean of BLEU and rBLEU, referred to as fBLEU.

4   Results

In this section, we present experimental results comparing the proposed multi-sense definition generation models against the single-sense base model [16]. All models are trained using parameter settings from [16], i.e., a two-layer LSTM as the RNN component with 300 units in each level; a character-level CNN with kernels of length 2–6 and size $\{10, 30, 40, 40, 40\}$ with a stride of 1; and Adam optimization with a learning rate of 0.001.

To generate definitions at test time, for each word and sense for the single-sense and multi-sense models, respectively, we sample tokens at each time step from the predicted probability distribution with a temperature of 0.1. We compute BLEU, rBLEU, and fBLEU for each word, and then the average of these measures over all words in a dataset. We repeat this process 10 times, and report the average scores over these 10 runs.

Results are shown in Table 2. Focusing on fBLEU, for every dataset, the best results are obtained using a multi-sense model — i.e., sense-to-definition (S2D), or definition-to-sense (D2S). Moreover, for every dataset, D2S improves over the base model. These results show that definition modelling can be improved by accounting for polysemy through the incorporation of multi-sense embeddings.

To qualitatively compare the base model and the proposed model, we consider the definitions generated for the word *state*. The following three definitions are generated for this word by the base model: (1) *a state of a government*, (2) *to make a certain or permanent power*, and (3) *to make a certain or administrative power*. In contrast, the proposed multi-sense model using D2S generates the following three definitions, which appear to capture a wider range of the usages of the word *state*: (1) *a place of government*, (2) *a particular region of a country*, and (3) *a particular place of time*.

Comparing S2D and D2S in terms of fBLEU, we observe that D2S often performs better. The number of sense vectors learned by Adagram for a given word is on average higher than the number of reference definitions available for that word, for every dataset. We hypothesize that the poor performance of S2D relative to D2S could therefore be due to sense vectors being associated with inappropriate definitions.

rBLEU is a recall-based evaluation metric that indicates the extent to which the reference definitions are covered by the generated definitions. A multi-sense definition generation model — which produces multiple definitions for a target word — is therefore particularly advantaged compared to a single-sense model — such as the base model — which produces only one, with respect to this metric. Indeed, we see that for every dataset, both S2D and D2S, outperform the base model in terms of rBLEU. BLEU, on the other hand, is a precision-based

Table 2. BLEU, rBLEU, and fBLEU for the single-sense definition generation model (base) and the proposed multi-sense models using sense-to-definition (S2D) and definition-to-sense (D2S) for each dataset. The best result for each evaluation metric and dataset is shown in boldface.

Lang.	Model	OmegaWiki			Wiktionary			WordNet		
		BLEU	rBLEU	fBLEU	BLEU	rBLEU	fBLEU	BLEU	rBLEU	fBLEU
DE	base	12.12	11.55	11.83	11.35	08.80	09.91	–	–	–
	S2D	12.43	16.26	14.09	**15.00**	15.82	**15.40**	–	–	–
	D2S	**12.44**	**16.83**	**14.31**	14.07	**16.54**	15.21	–	–	–
EL	base	–	–	–	–	–	–	**13.21**	12.06	12.61
	S2D	–	–	–	–	–	–	12.44	12.85	12.64
	D2S	–	–	–	–	–	–	13.08	**13.63**	**13.35**
EN	base	14.74	14.32	14.53	20.21	16.88	18.40	13.78	12.77	13.26
	S2D	14.23	16.02	15.07	18.88	16.99	17.89	12.85	13.09	12.97
	D2S	**15.22**	**17.80**	**16.41**	**21.49**	**19.78**	**20.60**	**13.84**	**14.84**	**14.32**
ES	base	**17.68**	17.70	17.69	–	–	–	**26.46**	24.69	25.54
	S2D	16.52	19.00	17.67	–	–	–	25.80	**28.14**	**26.92**
	D2S	17.54	**20.28**	**18.81**	–	–	–	25.68	27.97	26.78
FR	base	**12.58**	12.66	12.62	63.48	59.87	61.62	–	–	–
	S2D	11.70	14.26	12.85	63.56	60.00	61.73	–	–	–
	D2S	11.94	**14.82**	**13.23**	**64.12**	**60.41**	**62.21**	–	–	–
IT	base	**12.29**	11.93	12.11	–	–	–	21.33	20.65	20.98
	S2D	11.43	13.61	12.43	–	–	–	20.35	23.67	21.88
	D2S	11.74	**13.95**	**12.75**	–	–	–	**21.96**	**25.10**	**23.43**
JA	base	–	–	–	–	–	–	10.13	08.50	09.24
	S2D	–	–	–	–	–	–	**11.53**	**11.96**	**11.74**
	D2S	–	–	–	–	–	–	09.42	09.37	09.39
NL	base	14.37	14.04	14.20	–	–	–	–	–	–
	S2D	13.49	15.88	14.59	–	–	–	–	–	–
	D2S	**14.46**	**17.07**	**15.66**	–	–	–	–	–	–
RU	base	–	–	–	47.04	46.04	46.53	–	–	–
	S2D	–	–	–	46.24	46.69	46.46	–	–	–
	D2S	–	–	–	**47.52**	**48.09**	**47.80**	–	–	–

metric that indicates whether a generated definition contains material present in the reference definitions. The improvements of the multi-sense models over the base model with respect to rBLEU do not substantially impact BLEU — as observed by the overall higher fBLEU obtained by the multi-sense models. Overall, these results indicate that a multi-sense model is able to generate definitions that better reflect the various senses of polysemous words than a single-sense model, without substantially impacting the quality of the individual generated definitions.

5   Conclusions

Definition modelling is a recently-introduced language modelling task in which the aim is to generate dictionary-style definitions for a given word. In this paper, we proposed a multi-sense context-agnostic definition generation model which employed multi-sense embeddings to generate multiple senses for polysemous words. In contrast to most prior work on definition modelling which focuses on English, we conducted a multi-lingual study including nine languages from several language families. Our experimental results demonstrate that our proposed multi-sense model outperforms a single-sense baseline model. Code and datasets for these experiments is available.[10] In future work, we intend to consider incorporating alternative approaches to learning multi-sense embeddings into our model, as well as alternative approaches to associating sense vectors to definitions for constructing training instances.

References

1. Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: Linear algebraic structure of word senses, with applications to polysemy. TACL **6**, 483–495 (2018)
2. Artale, A., Magnini, B., Strapparava, C.: Wordnet for Italian and its use for lexical discrimination. In: Lenzerini, M. (ed.) AI*IA 97: Advances in Artificial Intelligence. pp. 346–356. Springer, Berlin, Heidelberg (1997)
3. Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D.: Breaking sticks and ambiguities with adaptive skip-gram. In: Proceedings of AISTATS 2016. pp. 130–138. Cadiz, Spain (2016)
4. Chang, T.Y., Chen, Y.N.: What does this word mean? Explaining contextualized embeddings with natural language definition. In: Proceedings EMNLP-IJCNLP 2019. pp. 6064–6070. Hong Kong, China (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL 2019. pp. 4171–4186. Minneapolis, Minnesota (2019)
6. Fernández-Montraveta, A., Vázquez, G., Fellbaum, C.: The Spanish version of WordNet 3.0. In: Text Resources and Lexical Knowledge. Selected Papers from KONENS 2008. pp. 175–182. Mouton de Gruyter (2008)
7. Gadetsky, A., Yakubovskiy, I., Vetrov, D.: Conditional generators of words definitions. In: Proceedings of ACL 2018. pp. 266–271. Melbourne, Australia (2018)

---

[10] `https://github.com/ArmanKabiri/Multi-sense-Multi-lingual-Definition-Modeling`

8. Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of the Japanese WordNet. In: Proceedings of LREC 2008. Marrakech, Morocco (2008)
9. Ishiwatari, S., Hayashi, H., Yoshinaga, N., Neubig, G., Sato, S., Toyoda, M., Kitsuregawa, M.: Learning to describe unknown phrases with local and global contexts. In: Proceedings of NAACL 2019. pp. 3467–3476. Minneapolis, Minnesota (2019)
10. Lee, G.H., Chen, Y.N.: MUSE: Modularizing unsupervised sense embeddings. In: Proceedings EMNLP 2017. pp. 327–337. Copenhagen, Denmark (2017)
11. Mickus, T., Paperno, D., Constant, M.: 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. pp. 1–11. Turku, Finland (2019)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. pp. 3111–3119 (2013)
13. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
14. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
15. Ni, K., Wang, W.Y.: Learning to explain non-standard English words and phrases. In: Proceedings of IJCNLP 2017. pp. 413–417. Taipei, Taiwan (2017)
16. Noraset, T., Liang, C., Birnbaum, L., Downey, D.: Definition modeling: Learning to define word embeddings in natural language. In: AAAI 2017. pp. 3259–3266 (2017)
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of ACL 2002. pp. 311–318. Philadelphia, Pennsylvania, USA (2002)
18. Stamou, S., Nenadic, G., Christodoulakis, D.: Exploring balkanet shared ontology for multilingual conceptual indexing. In: Proceedings of LREC 2004. Lisbon, Portugal (2004)
19. Zhu, R., Noraset, T., Liu, A., Jiang, W., Downey, D.: Multi-sense definition modeling using word sense decompositions. arXiv preprint arXiv:1909.09483 (2019)