Large-scale, Diverse, Paraphrastic Bitexts via Sampling and Clustering

J. Edward Hu **Abhinav Singh** **Nils Holzenberger**
Matt Post **Benjamin Van Durme**
Johns Hopkins University
{edward.hu,asingh78,nholzen1}@jhu.edu; {post,vandurme}@cs.jhu.edu

Abstract

Producing diverse paraphrases of a sentence is a challenging task. Natural paraphrase corpora are scarce and limited, while existing large-scale resources are automatically generated via back-translation and rely on beam search, which tends to lack diversity. We describe PARABANK 2, a new resource that contains multiple *diverse* sentential paraphrases, produced from a bilingual corpus using negative constraints, inference sampling, and clustering. We show that PARABANK 2 significantly surpasses prior work in both lexical and syntactic diversity while being meaning-preserving, as measured by human judgments and standardized metrics. Further, we illustrate how such paraphrastic resources may be used to refine contextualized encoders, leading to improvements in downstream tasks.

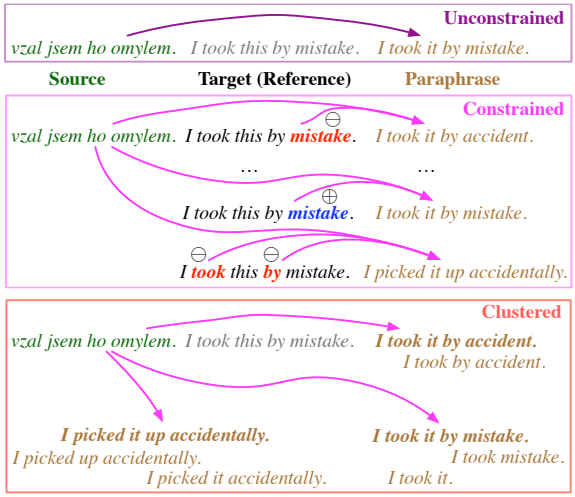

Figure 1: Contrived example paraphrases from previous work (unconstrained and constrained—used with permission) and ours (clustered).

1 Introduction

The ability to understand and produce paraphrases is a basic competency task, one that is often used as a teaching aid to validate if a student *understands* a statement or a concept. Current deep learning systems struggle with this task, exhibiting brittleness to both understanding and producing paraphrastic expressions (Iyyer et al., 2018).

One crucial factor behind this incompetence is the dearth of sentential paraphrastic data. Many works have sought to leverage the relative abundance of sub-sentential paraphrastic resources in paraphrase detection or generation (Napoles et al., 2016). Yet, they fail to capture contextualized word choices or syntactical variations, as word- or phrase-level resources cannot incorporate information from the whole input sentence.

Recent works have focused on leveraging bilingual resources to create large sentence-level paraphrastic collections using translation-based methods (Wieting and Gimpel, 2018; Hu et al., 2019). However, these works are confined to using beam search in decoding, which tend not to produce diverse candidates. One approach to force diverse translations is the use of hard lexical constraints at inference time (Hu et al., 2019). While effective in some cases, current approaches to automatic selection of such constraints is based on heuristics and task-oriented trial-and-error.

We present a novel resource with accurate and *collectively diverse* paraphrases, generated using stochastic decoding and clustering. By *collectively diverse*, we mean that the paraphrases of a given sentence cover a wide lexical and syntactic spectrum. Given a bilingual input pair, our core idea is to *sample* a large space of outputs from a translation system, cluster the results according to a notion of token-sequence similarity, score them with two translation models (one in each direction), and then select the best item from each cluster. We believe that sampling from the word distribution at each decoder time-step bet-

Proceedings of the 23rd Conference on Computational Natural Language Learning, pages 44–54
Hong Kong, China, November 3-4, 2019.

ter preserves the decoder's level of uncertainty, which is intrinsic to the goals of paraphrasing. We also sample *ancillary* lexical constraints to discourage, instead of explicitly prohibiting (Hu et al., 2019), certain words from being used by the decoder. While our experiment produces a large-scale English resource, our approach is dependent only on the availability of large bitexts and so is language-agnostic. We chose to build an English resource from CzEng to enable a direct comparison with Wieting and Gimpel (2018) and Hu et al. (2019).

Our contributions include:

- A large, high quality paraphrase collection[1] with up to 5 paraphrases per reference, close to 100 million pairs in total, which are more diverse than prior work in two distinct ways, as measured by standardized metrics;
- An evaluation of semantic similarity, lexical and syntactic diversity, compared against prior works, along with results on Sentence Textual Similarity (STS) Benchmark;
- Experiments on how our resource can be leveraged to improve performance on a set of language tasks.

2 Paraphrase generation pipeline

Prior works in constructing sentential paraphrastic resources have worked from large collections of bitext, producing translations of the foreign language sentence which, when paired with the target-language reference, constitute a set of paraphrases. Working from the very large CzEng parallel corpus, Wieting and Gimpel (2018) produced a single paraphrase for each English sentence by translating from the Czech source. Hu et al. (2019) expanded on this by translating the Czech sentence several times, using positive or negative constraints obtained from the English reference.

In terms of producing diverse paraphrases, both approaches are limited because they rely on beam search. There are potentially billions of paraphrases of a sentence (Dreyer and Marcu, 2012), yet beam search with recurrent models can only search a constant subset of them (in the beam size). There are techniques for producing more diverse paraphrases, such as the use of positive and negative constraints (Hu et al., 2019) or syntactic fragments (Iyyer et al., 2018), but these require the user to manually specify them, which can be cumbersome and unreliable.

We follow these prior works in working with the CzEng, a Czech–English dataset (Bojar et al., 2016b), due to its size, diverse domain coverage, and rich syntactic variations (Wieting and Gimpel, 2018), and to allow for a direct comparison in methodologies. However, we propose a new approach to paraphrase generation designed to increase paraphrastic diversity, using a multi-step process: the first part of the pipeline generates a large number of candidate paraphrases through a random process, and the second part whittles them down to a much shorter list. For each {source, target} input pair, we run the following pipeline:

1. *Constrained sampling*. We sample translations using a source→target translation model with lexical constraints. We obtain negative constraints by randomly selecting a set of tokens from the "source", so that they are not allowed to appear in the translations. Then, we decode each translation by sampling from only the top-k most probable tokens at each time step, after excluding constrained tokens (§2.1).
2. *Dual scoring*. The set of samples is then scored against the original source input using a target→source translation model. The scores from the forward and backward models are summed (§2.2).
3. *Clustering*. The samples are then clustered. The best item from each cluster (according to the summed score) is then returned (§2.3).

2.1 Constrained sampling

Sampling is a more effective way to explore model search space than beam search, particularly in auto-regressive models that do not permit dynamic programming. We introduce two means by which we can expand the hypothesis space, and produce a more diverse set of paraphrases, relative to straightforward beam search.

Top-k sampling In auto-regressive neural MT, the standard sampling approach would be to choose a word w_t at each decoder timestep t by sampling from the distribution $P(w_t \mid w_{1\ldots t-1})$. This approach has been found effective over 1-best beam search in generating source sentences in

[1] Available at http://nlp.jhu.edu/parabank2

back-translation (Edunov et al., 2018). However, for paraphrasing, this is not ideal, since words that are not semantically licensed by the source may be selected. Instead, we propose top-k sampling, in which we choose w_t from the top k most-probable tokens at each time step. This way, we allow the model to sample flexibly, vastly opening up the hypothesis space, without creating a large risk of producing nonsensical translations.

Randomized negative constraints Negative constraints are tokens that are not permitted in the decoder output. They are not formally described in the literature, but an implementation was provided with the associated positive constraints (Post and Vilar, 2018). Negative constraints can be provided as tokens or phrases; the decoder tracks the progress of generation through each constraint and adds an infinite cost to the final word of any constraints, precluding its selection in both sampling and beam search.

In order to further increase sample diversity when generating the hypotheses (§2.1), we obtain negative constraints from the source by randomly choosing a subset of tokens. We do this independently multiple times for each input sentence. This provides new sets of constraints for the inputs, independent of the decoding.

Note that we use subword regularization (Kudo, 2018) during training, causing different subword segmentations to be applied to training data types each time they are encountered and helping to build more robust models. We only constrain on the Viterbi segmentation, effectively discouraging negatively constrained words from appearing in the output, instead of prohibiting them, since there are often ways for the model to produce a word by generating a different decomposition.

2.2 Back-translation likelihoods

Some semantic changes during paraphrasing, especially omission, are not well-reflected by the (forward) probability $p_{generate}$ from the generating model. However, a model running in the other direction can penalize this omission, as found by Goto and Tanaka (2017). Thus, we obtain the back-translation probability p_{back} of each sampled candidate paraphrase, and define the final score for each candidate paraphrase as the joint probability $p^* = p_{generate} * p_{back}$, which is the sum of negative log-likelihood.

2.3 Edit-distance-based clustering

The above process produces a large set of translations of the source sentence. Many of them will be minor variants of one another, but we expect that there will be a lot of variety in the large pool. The task now is to reduce this pool to a small set of *collectively diverse* paraphrastic candidates.

We address this problem with k-means clustering via Levenshtein (or edit) distance (Miller et al., 2009). We compute this on lowercased, segmented candidates, after striping punctuation. Clusters are initialized with the k furthest candidates measured by edit-distance. We also add the reference sentence as the centroid of an additional cluster and skip the re-centering for that cluster. This improves the chance of the k clusters congregating candidates different from the reference in different ways. When the clustering has converged, we take the candidate with the best score from each cluster (except for the one with the reference sentence), rank them by score, and take the best n as the final output.

3 Evaluations

3.1 Data

All of our experiments are based on the CzEng 1.7 corpus, a subset of CzEng 1.6 (Bojar et al., 2016b) that has been chosen for higher quality. Based on experience with data quality issues in neural MT (Ott et al., 2018; Junczys-Dowmunt, 2018), we decided to further clean the corpus. First, we normalize Unicode punctuation, and keep only bilingual pairs whose English side can be encoded with `latin-1` and Czech side with `latin-2`. We then filter the data with dual cross-entropy filtering (Junczys-Dowmunt, 2018). We use Sockeye (Hieber et al., 2017) to train two NMT models, CS–EN and EN-CS, on a relatively clean subset of the data provided for WMT 2018 (Bojar et al., 2016a): Europarl, Wiki titles, and news commentary. We use 4 layer Transformer models (Vaswani et al., 2017) trained to convergence, with held-out likelihood evaluated on a random 500-sentence subset of the WMT16 and WMT17 news test data. These models are then used to score all the remaining CzEng data after deduplication. We kept all sentences with a model score (negative log-likelihood) of less than 3.5. After applying the above two filters, we keep $19,723,003$ out of the $57,065,358$ pairs in CzEng 1.7.

3.2 Translation models

We train two new translation models on the filtered data, the CS–EN *generation model* (for generating English candidates via sampling) and the EN–CS *scoring model* (for providing backwards scores of the candidates). Both are Transformer models built with AWS SOCKEYE. The generation model is a 12 layer Transformer with a model and embedding size of 768, 12 attention heads, a feed-forward layer size of 3072. The scoring model has 6 layers, model and embedding size of 512, 8 attention heads, and a feed-forward layer size of 2048.

All training data is pre-processed with subword sampling using SentencePiece[2] (Kudo, 2018) with a vocabulary size of 20k and character coverage of 0.9999. We used separate models for Czech and English. At inference time, we use the Viterbi segmentation of each input sentence, for both the generation and scoring models.

3.3 Parameters

There are a few parameters involved in the sample-score-cluster pipeline. For each Czech input sentence, we generate 5 sets of random constraints (§2.1), creating 5 variants of the input. From each of these inputs, we generate 30 samples using top-k sampling with $k = 10$ (i.e., at each timestep, the model randomly chooses from the top 10 most probable words, according to their scaled distribution, and excluding negatively constrained words). The resulting 150 sentences are scored, and anything with a combined score greater than 3.5 is thrown out. The remaining sentences are clustered into 8 clusters, one of them centered on the English reference. The reference cluster is thrown out, and a list of the best-scoring translation from the remaining 7 clusters is constructed. From this list, the top 5 translations are returned as hypotheses.

3.4 Setup

We follow the evaluation framework of Hu et al. (2019), which judged semantic similarity between paraphrases and their reference through human evaluation, and lexical diversity via automatic metrics. We use the evaluation result made public by Hu et al. (2019) to enable a direct comparison. Rather than focusing on improving semantic similarity, which is limited by the quality of the bilingual resource, we seek to build a resource that contains both more lexical and syntactical diversity.

We obtained the evaluation set from Hu et al. (2019), which contains 400 English sentences from CzEng. Due to additional filtering, 24 out of 400 (6%) reference sentences aren't in PARABANK 2 and therefore excluded in this evaluation.

We set the output size $n = 5$. After sorting the candidates by negative log-likelihood for each reference, we treat candidates at each rank as an individual system to investigate the expected quality of paraphrases under our approach. For references that produce fewer than 5 paraphrases, the paraphrase with the highest negative log-likelihood is duplicated to fill in ranks that otherwise would be empty. We also artificially pick the paraphrase with the maximum, minimum, and median human semantic similarity judgment under each reference as three additional oracle systems.

3.5 Semantic similarity via human judgments

For a fair comparison, we used the evaluation setup released by Hu et al. (2019), which uses the interface from EASL (Sakaguchi and Van Durme, 2018) to collect semantic similarity and gammaticality judgments. Each human annotator is presented with a reference sentence and five paraphrases from different sources. Annotators use a slider bar under each paraphrase to rate the semantic similarity from 0 (Opposite/Irrelevant) to 100 (Identical Meaning). Annotators are also asked to comment on whether the paraphrase is ungrammatical or nonsensical. The reference sentence is repeated next to the paraphrase for easier visual comparison.

Each paraphrase receives at least 3 independent judgments. Following Hu et al. (2019), we randomly add in the reference sentence as a paraphrase and filter out annotators who fail to score them 100 more than 10% of such encounters. The result includes only annotators who contributed at least 25 judgments and is shown in Tab. 1.

3.6 Paraphrastic diversity

BLEU has been a successful metric in evaluating MT systems. However, as noted earlier, monolingual paraphrasing has inherently different objectives than cross-lingual translation. BLEU, in tandem with human evaluation in semantic similarity, makes a good metric for paraphrastic diversity.

[2] https://github.com/google/sentencepiece

System	Semantics↑	Grammar↑	1-BLEU↑	∩/∪ ↓	Tree ED↑	Len. Ratio
ParaNMT	83.2	89.2	66.29	48.76	6.62	1.00
ParaBank$_{17}$	84.5	92.1	62.85	46.21	6.21	1.01
ParaBank$_{34}$	**85.7**	**92.7**	58.16	51.01	6.51	1.02
Our work$_1$	84.4±.0	90.2±.2	75.83±.10	37.75±.02	7.16±.05	1.04±.00
Our work$_2$	83.8±.0	88.3±.4	76.98±.07	36.19±.36	7.15±.17	1.05±.00
Our work$_3$	83.5±.0	87.3±.1	78.29±.69	35.22±.43	7.47±.11	1.05±.00
Our work$_4$	83.2±.2	86.6±.8	78.92±.19	34.49±.06	7.51±.11	1.06±.01
Our work$_5$	81.7±.1	87.3±.8	**81.55±.35**	**32.50±.32**	**7.80±.19**	1.09±.00
Our work$_{max}$	*91.2±.2**	*93.1±.8**	76.71±.11	37.15±.33	7.38±.06	1.05±.01
Our work$_{med.}$	84.1±.1	88.2±.1	78.34±.10	35.34±.25	7.52±.08	1.06±.00
Our work$_{min}$	72.5±.2	81.5±.2	*79.29±.21**	*33.13±.55**	*7.65±.10**	1.05±.00

Table 1: Paraphrastic diversity measured by (1-BLEU)×100, bag-of-word intersection/union score×100, and Tree edit-distance. Systems from this work that receive the best human judgments, worst human judgments, and the median, are included in the table. A higher 1-BLEU suggests higher paraphrastic diversity; a higher Intersection/Union score suggests a higher lexical diversity; a higher Tree edit-distance suggests a higher syntactic diversity. Best in each column, excluding oracle systems, is in bold. * denotes best oracle systems.

Here, we use 1-BLEU to measure how different the paraphrases are to the references.

We generate 5 paraphrases for each reference sentence using the approach outlined in this work. To account for randomness, we average over two independent runs in the result, shown in Tab. 1.

We consider two sources of paraphrastic diversity: 1) lexical diversity, the use of different words; and 2) syntactic diversity, the change of sentence or phrasal structure. We separately measure them using bag-of-word Intersection/Union scores and parse-tree edit-distances, respectively.

Lexical diversity A sentence is lexically different from the reference when it uses lexical paraphrases (e.g., synonyms) to convey similar meanings. We calculate the case-insensitive piece Intersection/Union score after striping punctuation and the SentencePiece white space symbol. All pieces are put to lowercase and into a set. The more pieces the two sentences share, the higher the score will be. The Intersection/Union scores between the reference and the paraphrases are shown in Tab. 1.

Syntactic diversity We consider the edit-distance between the parse trees of the reference and the paraphrase as a metric of syntactic diversity. Parse tree edit-distance is considered a useful feature in NLP tasks (Yao et al., 2013). The more syntactic variations there are between two sentences, the larger the tree edit-distance will be. We consider only the top 3 levels of the parse trees, excluding any terminals. Sentences are parsed with Stanford CoreNLP (Manning et al., 2014); the tree edit-distance is calculated with the APTED (Pawlik and Augsten, 2015a,b) algorithm. The average tree edit-distance for each system is shown in Tab. 1.

Diversity among paraphrases Hu et al. (2019) produced multiple paraphrases for each reference. While shown to be diverse compared to the reference, the authors did not investigate whether these paraphrases are trivial rewrites of one another, as it is likely the case with beam search under a few lexical constraints. Our clustering step is specifically designed to retrieve collectively diverse paraphrases.

We use the same metrics to evaluate pairs of systems from our work and compare them against ParaBank (Hu et al., 2019), as shown in Tab. 2. The max/min/median systems are oracle systems derived from human semantic similarity judgment scores. The human judgments from Tab. 1 show our paraphrases are of comparable quality to ParaBank, while maintaining a much higher degree of diversity among paraphrases of the same reference, as shown by automatic metrics.

3.7 Semantic similarity on STS Benchmark

In addition to evaluating via human judgments, we consider the same evaluation mechanism as ParaNMT (Wieting and Gimpel, 2018): the use

Systems Compared	1-BLEU↑	∩/∪ ↓	Tree ED↑
ParaBank$_{17}$/ParaBank$_{34}$	20.58	80.93	2.26
Our work$_1$/Our work$_3$	64.16±.21	52.77±.48	5.51±.01
Our work$_3$/Our work$_5$	**71.05±.22**	**45.00±.51**	**6.40±.19**
Our work$_1$/Our work$_5$	69.46±.27	46.79±.12	6.25±.18
Our work$_{max}$/Our work$_{min}$	66.03±.86	49.10±.16	5.84±.33

Table 2: Collective diversity within our work compared to ParaBank, as measured by (1-BLEU)×100, intersection/union score×100, and parse tree edit-distance.

of paraphrase corpora as training data for the Semantic Textual Similarity (STS) task. STS aims to measure the degree of equivalence in meaning or semantics between a pair of sentences. Notably, Agirre et al. (2016) having been a part of the SemEval workshop (2012 -2017). The evaluation consists of human annotated English sentence pairs, scored on a scale of 0 to 5 to quantify similarity of meaning, with 0 being the least, and 5 the most similar.

Wieting and Gimpel (Wieting and Gimpel, 2018) compared three encoding mechanisms: WORD, TRIGRAM and LSTM. The WORD model (Wieting et al., 2016) averages the embedding for each word in the sentence into a fixed length vector embedding for the sentence; the TRIGRAM model (Huang et al., 2013) averages over character trigrams; and the LSTM (Hochreiter and Schmidhuber, 1997) approach averages over the final hidden states to obtain the sentence embedding.

Encoders are trained on paraphrase pairs (s, s') with a margin based loss function $l(s, s', t, t') =$

$$\max(0, \delta - \cos[g(s), g(s')] + \cos[g(s), g(t)]) + \max(0, \delta - \cos[g(s), g(s')] + \cos[g(s'), g(t')])$$

where g is one of (WORD, TRIGRAM, LSTM) and (t, t') is a negative sample selected from a *megabatch*, an aggregation of m mini-batches (Wieting and Gimpel, 2018).[3]

We evaluate the WORD model trained[4] on ParaNMT, ParaBank and ParaBank 2 (our work). We retrieved the paraphrases from ParaBank and our work that share the same references as ParaNMT-5m. Our work is evaluated as 5 systems, based on the rank in the output; the last available paraphrase is used when lower ranks are empty. We also include a system that uses a pair of paraphrases, instead of a reference and a paraphrase. We keep ParaBank paraphrases that have a bag-of-word intersection/union score of 0.7 or less, and use the 1-best based on regression scores. In Tab. 3, we report Pearson's r and Spearman's r on the STS'16 test set. Sentence embeddings trained on our work exhibit higher correlation with human judgments, which reflects the superior paraphrastic diversity of the corpus.

System	Pearson's r	Spearman's r
ParaNMT	75.378	76.322
ParaBank	76.006	76.961
Our work$_1$	**76.546**	77.528
Our work$_2$	76.143	77.240
Our work$_3$	76.397	77.500
Our work$_4$	76.414	**77.612**
Our work$_5$	75.882	77.075
Our work$_{1/5}$	75.680	76.882

Table 3: Pearson's r × 100 and Spearman's r × 100 computed on STS 2016 task. Our work$_{1/5}$ contains paraphrase pairs from system$_1$ paired with system$_5$, while all other systems are paired with the reference sentence.

3.8 Improving contextualized encoders with paraphrastic data

Paraphrastic data can be used to fine-tune contextualized encoders such as BERT (Devlin et al., 2018). We frame the fine-tuning task as paraphrase identification (Das and Smith, 2009), where given a pair of sentences, the task is to classify them as paraphrases or non-paraphrases. To generate the training data, we extract, for each

[3] We confirmed this loss with Wieting and Gimpel, that it captures their open implementation, which we employ. Wieting and Gimpel (2018) described their loss as: $\max(0, \delta - \cos(g(s), g(s')) + \cos(g(s), g(t)))$, which is equivalent under their assumption the paraphrases are equivalent.

[4] https://github.com/jwieting/para-nmt-50m

	QQP	MNLI	STS-B	MRPC
BERT	87.90	**83.86**	88.40	84.00
pBERT	**88.14**	82.64	**88.59**	**86.55**

Table 4: F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for MNLI. Numbers reported on Dev set

	Type	BERT	pBERT
F1	HasAns	**76.81**	74.21
	NoAns	71.44	**74.95**
	Total	74.12	**74.58**
Exact Match	HasAns	**70.34**	68.00
	NoAns	71.44	**74.95**
	Total	70.89	**71.48**

Table 5: SQuAD 2.0 results on dev set.

sentence in PARANMT-5M, the sentence embeddings generated by the WORD model trained in §3.7. For each sentence s, we then find the (approximate) nearest neighbour n which is not s', among all of the sentences. We thus obtain two pairs, where (s, s') is a paraphrase pair, and (s, n) is a non-paraphrase pair. We use these to train a binary classifier with cross-entropy loss.

We then use this BERT fine-tuned on paraphrases (henceforth pBERT) for fine-tuning on SQuAD 2.0 (Rajpurkar et al., 2018) and 4 NLP tasks present in the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019): Quora Question Pairs (QQP) (Chen et al., 2017), Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018), the Semantic Textual Similarity Benchmark (STS-B) (Agirre et al., 2016), and the Microsoft Research Paraphrase Corpus (MRPC) (Dolan et al., 2004). Following the model formulation, hyper-parameter selection and training procedure specified in Devlin et al. (2018), we add a single task-specific, randomly initialized output layer for the classifier.

We present our results in Tab. 4 and Tab. 5. We observe gains for STS-B, MRPC and QQP, tasks strongly related to paraphrase identification. Fine-tuning on our paraphrase corpus also improves performance on SQuAD, a question-answering task, while slightly degrading performance on MNLI. Overall, simple fine-tuning of BERT on our corpus leads to improvements on downstream tasks, in particular when the task is related to paraphrase detection.

4 Related works

4.1 Paraphrastic resources

Paraphrastic resources exist across different scopes (i.e., lexical, phrasal, sentential) and different creation strategies (i.e., manually curated, automatically generated). For a more comprehensive survey on data-driven approaches to paraphrasing, please refer to Madnani and Dorr (2010).

Sub-sentential resources WordNet (Miller, 1995), FrameNet (Baker et al., 1998), and VerbNet (Schuler, 2006) can be used to extract paraphrastic expressions at lexical levels. They contain the grouping of words or phrases that share similar semantics and sometimes entailment relations. While FrameNet and VerbNet do have example sentences or frames where lexical units are put into contexts, there is no explicit paraphrastic relations among these examples. Also, these datasets tend to be small, as they were curated manually. There have been efforts to augment such resources with automatic methods (Snow et al., 2006; Pavlick et al., 2015b), but they are still confined to lexical level and sometimes require the use of other paraphrastic resources (Pavlick et al., 2015b).

PPDB (Ganitkevitch et al., 2013; Pavlick et al., 2015a) automated the generation of lexical paraphrases via bilingual pivoting, taking advantage of the relative abundance of bilingual corpora. While significantly larger and more informative (e.g., ranking, entailment relations, etc.) than the above manually curated resources, PPDB suffers from ambiguity as words or phrases are removed from their sentential contexts.

Sentential resources There exists multiple human translations in the same language for some classic readings. Barzilay and McKeown (2001) sought to extract lexical paraphrastic expression from such sources. Unfortunately such resources – along with those manually constructed for text generation research (Robin, 1995; Pang et al., 2003) – are small and limited in domain.

PARANMT and PARABANK are two much larger sentential paraphrastic resources created through back-translation.

Reference:		
	Real life is sometimes thoughtless and mean.	Hey, stop right there!
PARANMT:		
	real life is sometimes reckless and cruel .	hey , stop .
PARABANK:		
	The real life is occasionally ruthless and cruel.	Stay where you are!
	The real world is occasionally ruthless and cruel.	
	The real life is sometimes reckless and cruel.	
Our work:		
	True life is sometimes ruthless and cruel.	Hold your position!
	Actual life is sometimes ruthless and cruel.	Stay where you are!
	Sometimes real life is ruthless and cruel.	Stay in position!
	Real life can be inconsiderate, cruel sometimes.	Remain where you are!
	Real living is a harsh and unscrupulous one, at times.	Stay put!

Table 6: Selected examples from our work, compared to paraphrastic resources with prior approaches. Our work has paraphrases that are not only different from the reference, but also diverse among themselves.

4.2 Translation-based Approaches

PARANMT is an automatically generated sentential paraphrastic resource through back-translating bilingual resources. It leveraged the imperfect ability of Neural Machine Translation (NMT) to recreate the translation target by conditioning on the source side of the bitext.

PARABANK took a similar approach but with the inclusion of lexical constraints from the target side of the bitext. This step allows for multiple translations from one bilingual sentence pair and promotes lexical diversity. Their work, despite being larger and shown to be less noisy than PARANMT, relies on heuristics to produce *hard* constraints on the decoder, which often causes unintended changes in semantics or grammar.

Both works largely follow standard approaches in NMT, generating 1-best hypotheses given a source text and a set of constraints using beam search. Sentential paraphrasing, nevertheless, has fundamentally different objectives than MT. The latter strives to find the best elicitation that is both fluent and semantically close to the **foreign** text to convey information across languages. The former, on the other hand, seeks syntactically and lexically diverse expressions that convey the same meaning, with the goal of capturing the **intrinsic flexibility and uncertainty** of human communications. This work attempts to adapt the methodology to these objectives of monolingual paraphrasing.

4.3 Leveraging paraphrases in NLP

In the context of semantic parsing, Berant and Liang (2014) use a paraphrase classification module to determine the match between a canonical utterance and a logical form, both using a phrase table and distributed representations. To improve question answering (QA), Duboue and Chu-Carroll (2006) generate paraphrases of a given question using back-translation, and optionally replace the original question with the most relevant paraphrase. Dong et al. (2017) tackle QA by marginalizing the probability of an answer over a set of paraphrases, generated using rule-based and NMT-based methods. Fader et al. (2013) use a corpus of questions with paraphrases, to construct a corpus of semantically equivalent queries.

The task of paraphrase identification, which we use as a fine-tuning objective, has been studied as a task in itself. Das and Smith (2009) use grammars to perform generative modeling of paraphrases. Madnani et al. (2012) identify paraphrases by relying only on MT metrics as features. Ferreira et al. (2018) feed sentence similarity measured with hand-crafted features to machine learning algorithms. Convolutional neural networks have been introduced by Yin and Schütze (2015) and Chen et al. (2018), and further augmented with LSTMs (Kubal and Nimkar, 2018) and attention mechanisms (Fan et al., 2018).

5 Conclusions and future work

A presumed goal for building a sentential paraphrase resource is to capture *different* ways of expressing the same thing: *diversity matters*. Previous work on paraphrastic resource creation relied on decoding techniques from NMT using bilingual corpora, with limited success in promoting diverse expressions. We have presented a new community resource produced by sampling and clustering. We evaluated our method against prior works (Wieting and Gimpel, 2018; Hu et al., 2019) and found significant gains in both lexical and syntactic diversity. Further, we've shown how straightforward fine-tuning of a state-of-the-art contextual encoder on our resource can improve performance on a variety of language tasks.

Acknowledgments

This work was supported in part by a National Science Foundation collaborative grant (BCS-1748969/BCS-1749025) The MegaAttitude Project: Investigating selection and polysemy at the scale of the lexicon.

References

Eneko Agirre, Carmen Banea, Daniel M. Cer, Mona T. Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval@NAACL-HLT*, pages 497–511. The Association for Computer Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL/ICCL*, ACL '98, pages 86–90, Stroudsburg, PA, USA.

Regina Barzilay and Kathleen R McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL*.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1415–1425.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 131–198.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Peixin Chen, Wu Guo, Zhi Chen, Jian Sun, and Lanhua You. 2018. Gated convolutional neural network for sentence matching. *memory*, 1:3.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. First quora dataset release: Question pairs.

Dipanjan Das and Noah A Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 468–476. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171. Association for Computational Linguistics.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1608–1618.

Miao Fan, Wutao Lin, Yue Feng, Mingming Sun, and Ping Li. 2018. A globalization-semantic matching neural network for paraphrase identification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2067–2075. ACM.

Rafael Ferreira, George DC Cavalcanti, Fred Freitas, Rafael Dueire Lins, Steven J Simske, and Marcelo Riss. 2018. Combining sentence similarities measures to identify paraphrases. *Computer Speech & Language*, 47:59–73.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings NAACL-HLT 2013*, pages 758–764.

Isao Goto and Hideki Tanaka. 2017. Detecting untranslated content for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 47–55, Vancouver. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. PARABANK: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI 2019*, Hawaii, USA. AAAI.

Po-Sen Huang, , Jianfeng Gao, , and and. 2013. Learning deep structured semantic models for web search using clickthrough data. ACM International Conference on Information and Knowledge Management (CIKM).

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895. Association for Computational Linguistics.

Divesh R Kubal and Anant V Nimkar. 2018. A hybrid deep learning architecture for paraphrase identification. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Courtney Napoles, Chris Callison-Burch, and Matt Post. 2016. Sentential paraphrasing as black-box machine translation. In *Proceedings of the NAACL 2016*, pages 62–66, San Diego, California. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965, Stockholmsmssan, Stockholm Sweden. PMLR.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015a. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL/IJCNLP*, volume 2.

Ellie Pavlick, Travis Wolfe, Pushpendre Rastogi, Chris Callison-Burch, Mark Dredze, and Benjamin Van Durme. 2015b. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the ACL/IJCNLP*, volume 2.

Mateusz Pawlik and Nikolaus Augsten. 2015a. Efficient computation of the tree edit distance. *ACM Trans. Database Syst.*, 40(1):3:1–3:40.

Mateusz Pawlik and Nikolaus Augsten. 2015b. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Jacques Pierre Robin. 1995. *Revision-based Generation of Natural Language Summaries Providing Historical Background: Corpus-based Analysis, Design, Implementation and Evaluation*. Ph.D. thesis, New York, NY, USA. UMI Order No. GAX95-33653.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of ACL*, pages 208–218, Melbourne, Australia. ACL.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ICCL/ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

John Wieting and Kevin Gimpel. 2018. PARANMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of ACL 2018*, pages 451–462. ACL.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Answer extraction as sequence tagging with tree edit distance. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 858–867, Atlanta, Georgia. Association for Computational Linguistics.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.