ParaBank: Monolingual Bitext Generation and Sentential Paraphrasing via Lexically-constrained Neural Machine Translation

J. Edward Hu Rachel Rudinger Matt Post Benjamin Van Durme
3400 North Charles Street
Johns Hopkins University
Baltimore, MD, USA

Abstract

We present ParaBank, a large-scale English paraphrase dataset that surpasses prior work in both quantity and quality. Following the approach of ParaNMT (Wieting and Gimpel, 2018), we train a Czech-English neural machine translation (NMT) system to generate novel paraphrases of English reference sentences. By adding lexical constraints to the NMT decoding procedure, however, we are able to produce *multiple* high-quality sentential paraphrases per source sentence, yielding an English paraphrase resource with more than 4 billion generated tokens and exhibiting greater lexical diversity. Using human judgments, we also demonstrate that ParaBank's paraphrases improve over ParaNMT on both semantic similarity and fluency. Finally, we use ParaBank to train a monolingual NMT model with the same support for lexically-constrained decoding for sentence rewriting tasks.

1 Introduction

In natural languages, mappings between meaning and utterance may be many-to-many. Just as ambiguity allows for multiple semantic interpretations of a single sentence, a single meaning can be realized by different sentences. The ability to identify and generate *paraphrases* has been pursued in the context of many natural language processing (NLP) tasks, e.g., semantic similarity, plagiarism detection, translation evaluation, monolingual transduction tasks such as text simplification and style transfer, textual entailment, and short-answer grading.

Paraphrastic resources exist at many levels of granularity, e.g., WordNet (Miller, 1995) for word-level and the Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch, 2013) for phrase-level. We are interested in building a large resource for *sentence*-level paraphrases in English. In this work, we introduce ParaBank, the largest publicly-available collection of English paraphrases we are aware of to date. We follow and extend the approach of Wieting and Gimpel (2018), who generate a set of 50 million English paraphrases, under the name ParaNMT-50M, via neural machine translation (NMT).

A part of ParaBank is trained and decoded on the same Czech-English parallel corpus (Bojar et al., 2016) as ParaNMT. However, ParaBank not only contains more reference sentences but also more paraphrases per reference than ParaNMT with overall improvements in lexical diversity, semantic similarity, and fluency. We are able to obtain a larger number of paraphrases per sentence by applying explicit lexical constraints to the NMT decoder, requiring specific words to appear (*positive constraint*) or not to appear (*negative constraint*) in the decoded sentence. Using ParaBank, we train, and release to the public, a monolingual sentence re-writing system, which may be used to paraphrase unseen English sentences with lexical constraints.

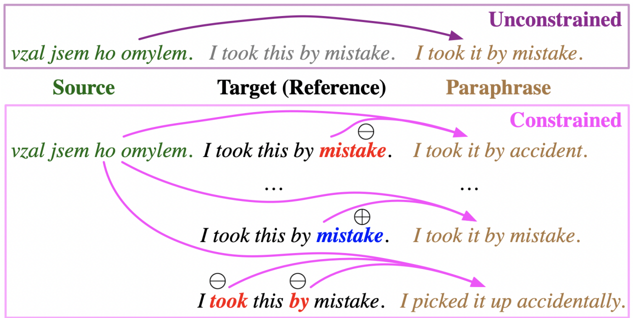

Figure 1: Contrasting prior work (e.g., ParaNMT) on building sentential paraphrase collections via translation, ParaBank conditions on both the source and *target* side of a translation pair, employing positive and negative lexical constraints derived from the reference to result in *multiple*, *diverse* paraphrases.

The main contributions of this work are:

- A novel approach to translation-based generation of monolingual bitext;
- The largest and highest quality English-to-English bitext resource to date with 79.5 million references and predicted human-judgment scores;
- Manual assessment of candidate paraphrases on semantic similarity, across a variety of generation strategies;
- A trained and freely released model for English-to-English rewriting, supporting positive and negative lexical constraints.

ParaBank is available for download at: `http://nlp.jhu.edu/parabank`.

2 Background

The following summarizes key background in approaches to monolingual paraphrasing with regard to PARABANK, along with the essential prior efforts that enable PARABANK to improve on related work. We first discuss work on sub-sentential resources that may be: hand-curated, automatically expanded from hand-curated, or fully automatically created. We then describe efforts at gathering or creating monolingual sentential bitexts, or otherwise sentence-to-sentence paraphrastic rewriting. For additional background, we refer readers to Madnani and Dorr (2010).

2.1 Paraphrasing Resources

Lexical Resources WordNet (Miller, 1995) includes manually-curated sub-sentential paraphrases. It groups single words or short-phrases with similar meanings into synonym sets (*synsets*). Each synset is related to other synsets through semantic relations (e.g., hypernym, entailment) to allow the construction of hierarchies and entailment relations.

VerbNet (Schuler, 2006) is a manually-constructed lexicon of English verbs. It is augmented with syntactic and semantic usage of its verb sense members. As a paraphrase resource, VerbNet groups verb senses into classes like "say-37.7-1" or "run-51.3.2". Members under the same class share a general meaning as noted by the class name.

FrameNet (Baker, Fillmore, and Lowe, 1998) contains manually annotated sentences classified into semantic frames. Though designed as a readable reference and as training data for semantic role labeling, FrameNet can be used to construct sub-sentential paraphrases, owing to rich semantic contents like semantic types and frame-to-frame relations.

Many efforts have aimed to automatically expand gold resources, for example: Snow, Jurafsky, and Ng (2006) augmented WordNet by combining existing semantic taxonomies in WordNet with hypernym predictions and coordinate term classifications; and Pavlick et al. (2015b) tripled the lexical coverage of FrameNet by substituting words through PPDB, with verification of quality via crowdsourcing.

Larger Paraphrases There is a rich body of work in automatically inducing phrasal, syntactic or otherwise structural paraphrastic resources. Some examples include: DIRT (Lin and Pantel, 2001), which extracts paraphrastic expressions over paths in dependency trees, based on an extension of the distributional hypothesis, which states that words that occur in the same contexts tend to have similar meanings (Harris, 1954); Weisman et al. (2012) explored learning inference relations between verbs through broader scopes (document or corpus level), resulting in a richer set of cues for verb entailment detections; and PPDB (Ganitkevitch, Van Durme, and Callison-Burch, 2013), a multi-lingual effort to construct paraphrase pairs by linking words or phrases that share the same translation in another language. Related to the human scoring pursued here for ParaBank evaluation, in PPDB 2.0 (Pavlick et al., 2015a), the authors collected annotations via Mechanical Turk to measure the quality of the induced paraphrases, in order to train a model for scoring all the entries in PPDB for semantic adequacy.

2.2 Monolingual Bitexts

Manually Created Some monolingual bitexts are created for research in text generation (Pang, Knight, and Marcu, 2003; Robin, 1995) where models benefit from exposure to multiple elicitations of the same concept. The utility of these resources is largely limited by their scale, as the cost of creation is high. Other sources include different translations of the same foreign text, which exist for many classic readings. Work has been done on identifying and collecting sub-sentential paraphrases from such sources (Barzilay and McKeown, 2001). However, the artistic nature of literature could result in various interpretations, often rendering this type of resource unreliable.

PARANMT Wieting and Gimpel (2018) leverage the relative abundance of bilingual bitext to generate sentence-level paraphrases through machine translation. This approach trains a neural machine translation (NMT) model from a non-English source language to English over the entire bitext (Czech-English) (Bojar et al., 2016), and decodes the source to obtain outputs that are semantically close to the training target. Decoding in PARANMT solely depends on the trained model and the source text with no inputs derived from the target English sentences. The approach in PARANMT exhibits little control over the diversity and adequacy of its paraphrastic sentence pairs: the application of lexical constraints during decoding is a key distinction between PARANMT and the approach described herein.

2.3 Lexically Constrained Decoding

Lexically constrained decoding (Hokamp and Liu, 2017) is a modification to beam search for neural machine translation that allows the user to specify tokens and token sequences that must (or must not) appear in the decoder output. A lexical constraint can be either *positive* or *negative*. A positive constraint requires the model to *include* the constrained token or tokens in the output. Negative constraints, on the other hand, require the model to *avoid* certain token or tokens in the output. The effect of constraints is to cause the system to generate the best decoding (translation or paraphrase) under those constraints.

Recently, Post and Vilar (2018) proposed a variant of lexically constrained decoding that reduced complexity from linear to constant-time (in the number of constraints). This allows us to decode hundreds of millions of sentences with constraints in a reasonable amount of time, and forms a key enabling technology for PARABANK. An implementation of it is included in Sockeye (Hieber et al., 2017), which we use for this work.

2.4 Efficient Annotation of Scalar Labels (EASL)

Sakaguchi and Van Durme (2018) propose an efficient and accurate method of collecting scalar-valued scores from human annotators, called EASL, by combining pairwise rank-

System	Reference	Constraints	Paraphrase
⊖2ndIDF	How often do earthquakes **occur**?	⊖occur	How often are earthquakes happening?
⊖2nd3rdIDF	How **often** do earthquakes **occur**?	⊖occur, often	What frequency do earthquakes happen?
⊖⟨BOS⟩ 1sttoken	**How** often do earthquakes occur?	⊖⟨BOS⟩ How	What frequency do earthquakes occur?
PPDB equ	How **often** do earthquakes occur?	⊖often ⊕frequently	How **frequently** are earthquakes happening?
PPDB rev	This **myth** involves three misconceptions.	⊖myth ⊕mythology	There are three misconceptions in this **mythology**.
⊖1stIDF	This myth involves three **misconceptions**.	⊖misconceptions	This myth has three false ideas.
⊖3rdIDF	This myth **involves** three misconceptions.	⊖involves	The myth has three misconceptions.
⊖2ndIDF	It didn't mean **anything**, okay?	⊖anything	It didn't mean a thing, okay?
⊖1stIDF	It didn't mean anything, **okay**?	⊖okay	It didn't mean anything, all right?
⊖1st, 3rdIDF w/ lexical variants	It didn't **mean** anything, **okay**?	⊖okay,mean, means,meaning, meant	It was nothing, all right?
⊖1st, 2ndIDF	It didn't mean **anything**, **okay**?	⊖okay,anything	It meant nothing, all right?

Table 1: Examples of different constraint selection methods in PARABANK leading to multiple different paraphrases per reference. System names are defined in Tab. 3 with full descriptions in §3.2. Negative constraints are labeled in red and positive constraints are labeled in blue.

ing aggregation and direct assessment. In manually evaluating the quality of our system's paraphrases, we adopt an annotator interface based on EASL. Human annotators are asked to assess paraphrases' semantic similarity to the reference sentence through a combination of direct numerical assessment and pairwise comparison. This mode of evaluation is akin to the method employed by the Workshop on Statistical Machine Translation (WMT) evaluation through an adaptation of TrueSkill™ (Sakaguchi, Post, and Van Durme, 2014).

3 Approaches

3.1 Training the model

We use Sockeye (Hieber et al., 2017) to train the machine translation model, with which we generate paraphrases under different constraint systems. The training data, CzEng 1.7 (Bojar et al., 2016), is tokenized[1] and processed through Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016). To reduce the vocabulary size, we tokenized all numbers to digit-level.

The model's encoder and decoder are both 6-layer LSTMs with a hidden size of 1024 and an embedding size of 512. Additionally, the model has one dot-attention layer. We trained the model on 2 Nvidia GTX 1080Ti for two weeks.

3.2 Selection of Lexical Constraints

Lexical constraints (§2.3) can directly influence the diversity and sufficiency of the NMT decoder output (i.e., the translation). We generate paraphrases of English translations of Czech sentences using different sets of constraints obtained from the English side of the bitext. These constraints may be positive or negative, and multiple constraints of either type may be combined simultaneously (provided they are consistent).

The tokens on which we base these constraints are the tokens that appear in the reference sentence (or are morphological variants thereof), though in principle any token could be used as a constraint. To select the constraints from this pool, we experiment with different ways of selecting these constraints from the reference, resulting in 37 experimental system configurations (Tab. 3): one baseline system with no constraints, three with tokens selected positionally, 30 with positive and negative constraints selected via inverse document frequency (IDF), and three additional systems based on PPDB lookups. Here we describe these selection criteria in detail.

IDF Criteria We compute each token's inverse document frequency (IDF) from the training data. To avoid constraints with misspelled or overly-specialized words, we exclude tokens with an IDF above 17.0 from consideration as lexical constraints. We also avoid constraints based on the most frequent English words by setting a minimum IDF threshold of 7.0. These thresholds are heuristics and we leave optimizations to future works. Among the remaining candidates, the constraint token may be selected by the highest IDF, lowest IDF, or randomly.

Prepositions We make one exception to the minimum IDF threshold in the case of prepositions, which we found fruitful as diversity-promoting constraints (see Fig. 1). The allowed

[1] We used spaCy (Honnibal and Montani, 2017) to tokenize English text, and MorphoDiTa (Straková, Straka, and Hajič, 2014) to tokenize Czech text.

Token	**IDF**	**Token**	**IDF**
proud	11.1	told	7.9
work	7.4	them	6.2
her	5.8	was	4.3
for	3.6	to	2.3

Table 2: Tokens assessed, along with their IDFs, of the sentence *"I told her I was proud to work for them."*

prepositions are: *about*, *as*, *at*, *by*, *for*, *from*, *in*, *into*, *of*, *on*, *onto*, *over*, *to*, and *with*.

Morphological Variants To discourage trivial paraphrases, some negative constraint systems include morphological variants of the word, and all negative constraint systems exclude capitalization. For positive constraints, we only consider morphological variants for verbs[2], and only one variant of the selected token is used. For all constraints, only lowercased alphabetical tokens are considered.

Positional Constraints It has been observed that RNN decoders in dialogue systems can be nudged toward producing more diverse outputs by modifying decoding for only the first few tokens (Li et al., 2016). Motivated by this observation, we include **positional constraints**, which require that a given constraint apply only at the *beginning* of the sentence (denoted as ⟨BOS⟩ in Table 3). In particular, we require the first one, two, or three tokens *not* to match the reference translation (i.e., a *negative* constraint).

PPDB Constraints We also use PPDB 2.0 (Pavlick et al., 2015a) as a source for introducing *positive* lexical constraints. For each token in the original English sentence that passes the IDF filter (above), we look up its paraphrases in PPDB[3]. We randomly select up to three lexical paraphrases, one each of the type `Equivalence`, `ForwardEntailment`, and `ReverseEntailment`, if present. We further require the selected lexical paraphrases to coarsely match the original token's POS tag (e.g., any form of verb, etc.) A negative constraint is then added for the original token, and a positive constraint is added for the lexical paraphrase from PPDB. These negative-positive constraint pairs are applied one at a time (i.e., one pair per decoding).

Example of constraint selection Here we work through the process of selecting lexical constraints to produce a new paraphrase of the sentence *"I told her I was proud to work for them."*. We follow the rules of system number 18 (as designated in Tab. 3).

First, tokens with only lower-cased alphabetical letters are assessed; they are listed in Tab. 2 along with their IDF values. After applying the IDF thresholds and exception for prepositions, the following tokens are in the candidate pool, from which we choose tokens to constrain on (listed in descending order of IDF values): *proud*, *told*, *work*, *for*, and *to*.

No.	$\oplus/\ominus$	Token(s) Selected	Lex.
1,2,3	$\ominus$	1st, 2nd, 3rd highest IDF[4]	None
4,5,6	$\ominus$	(1st, 2nd), (2nd, 3rd), (1st, 3rd) IDF	None
7	$\ominus$	(1st, 2nd, 3rd) highest IDF	None
8,9,10	$\ominus$	1st, 2nd, 3rd highest IDF	All
11,12,13	$\ominus$	(1st, 2nd), (2nd, 3rd), (1st, 3rd) IDF	All
14	$\ominus$	(1st, 2nd, 3rd) highest IDF	All
15,16,**17**	$\ominus$	1st, 2nd, **3rd lowest IDF**	None
18,19,20	$\ominus$	(1st, 2nd), (2nd, 3rd), (1st, 3rd) low	None
21	$\ominus$	**(1st, 2nd, 3rd) lowest IDF**	None
22,23,24	$\ominus$	1, 2, 3 random tokens	None
25,26,27	$\ominus$	1, 2, 3 random tokens	All
28		**no constraints**	
29,30,**31**	$\oplus$	1st, 2nd, **3rd highest IDF**	Verb[5]
32,33,**34**	$\ominus$	positional ⟨BOS⟩: 1, 2, **3 tokens**	None
35,36,37	$\ominus\oplus$	**PPDB equ, fwd, rev entailment**	None

Table 3: Different system configurations to generate paraphrases. $\oplus$/$\ominus$ designates the type of constraint we impose on the model: *negative* constraints ($\ominus$) are sets of tokens or ngrams which the decoder must *not* include in its output, while *positive* constraints ($\oplus$) are sets of tokens or ngrams *required* in the output. Additional constraints may be included for lexical variations of the selected token(s), as indicated by the Lex. column. Systems in bold fonts are presented here for evaluation. Evaluations of all systems will be made available with the resource.

Under the configuration of PARABANK System 18 (Tab. 3), which avoids tokens with the lowest and the second lowest IDF, a negative constraint is generated for the tokens *for*, *For*, *to*, and *To*. With these constraints applied, the resulting decoded paraphrase is: *"I told her I was really proud of working with them."*.

More examples are shown in Tab. 1.

4 Extension to other bilingual corpora

Our methods are independent of the source bilingual corpora. We apply the same pipeline to the 10^9 word French-English parallel corpus (Giga) (Callison-Burch et al., 2009), which has different domain coverage than CzEng. This adds an additional 22.4 million English references, resulting in a total of 79.5 million reference sentences for PARABANK. We conducted manual evaluation on paraphrases generated from both CzEng and Giga, and included the additional PPDB constraint systems for Giga.

5 Evaluation

We evaluate the quality of paraphrases by both semantic similarity and lexical diversity. A good paraphrase should

[2] We POS-tagged the reference sentence using SpaCy.

[3] We use the `ppdb-2.0-lexical-xl` packet downloaded from `paraphrase.org`.

[4] Highest within the token pool. Same for lowest.

[5] If the token is a verb, we pick a random lexical variation to include. E.g., we might constrain on one of "taken", "taking", "takes", or "take" if the original token is "took".

	#	System	Czech-English					French-English				
			len=5	len=10	len=20	len=40	Avg.	len=5	len=10	len=20	len=40	Avg.
Semantic Similarity	-	PARANMT	73.47	73.49	75.49	71.09	73.39	-	-	-	-	-
	17	$\ominus$3rdlow IDF	71.59	**75.71**	**79.67**	**77.31**	**76.07**	73.64	72.69	83.60	80.15	77.52
	21	$\ominus$3 low IDF	59.86	66.56	70.76	69.77	66.74	62.84	69.56	81.73	77.08	72.80
	28	no con.	**76.63**	**78.35***	**83.19***	**80.35**	**79.63***	79.09	74.28	84.86	83.02*	80.31*
	31	$\oplus$3rdtop IDF	**77.12***	**75.22**	**82.98**	**79.91**	**78.81**	80.22*	73.78	85.01	80.02	79.76
	34	$\ominus$first 3 tks	**74.03**	**77.74**	**81.88**	**81.04***	**78.67**	78.12	74.37*	84.14	82.74	79.84
	35	PPDB Equ	-	-	-	-	-	79.22	67.30	82.09	79.98	77.15
	36	PPDB Fwd	-	-	-	-	-	65.40	69.39	85.10*	82.75	75.66
	37	PPDB Rev	-	-	-	-	-	64.25	66.28	76.35	72.49	69.84
Lexical Diversity	35	PARANMT	18.32	25.49	32.25	33.84	27.48	-	-	-	-	-
	17	$\ominus$3rd low IDF	**9.25**	**20.62**	35.76	41.88	**26.88**	6.59	15.57	26.78	34.54	20.87
	21	$\ominus$3 low IDF	**0.00***	**7.84***	**22.04***	**28.90***	**14.70***	1.21*	6.90*	18.28	22.02	12.10
	28	no con.	19.32	28.41	42.02	46.63	34.10	14.07	21.12	30.57	38.66	26.11
	31	$\oplus$3rd top IDF	19.09	29.16	41.42	46.96	34.16	15.28	21.69	31.49	37.56	26.51
	34	$\ominus$first 3 tks	**13.22**	**24.90**	39.11	44.67	30.47	10.87	18.15	27.70	36.87	23.40
	35	PPDB equ	-	-	-	-	-	4.46	13.33	25.88	29.54	18.30
	36	PPDB fwd	-	-	-	-	-	1.51	8.87	14.14*	19.03*	10.89*
	37	PPDB bkw	-	-	-	-	-	3.94	9.69	18.94	26.24	14.70

Table 4: **Top**: Semantic similarity between paraphrases and reference sentences, as scored by human annotators on a 0-100 scale (least to most similar). Results are grouped by length of reference sentences $\{5, 10, 20, 40\}$. System names and numbers correspond to Tab. 3. Improvements over PARANMT (Czech-English only) in bold. Asterisk (*) indicates best in column. **Bottom**: Lexical diversity between generated paraphrases and reference sentences, as computed by a modified BLEU score with no length penalty. Results are grouped by length of reference sentences, and BLEU is computed over concatenated references and concatenated paraphrases. Lower BLEU scores indicate greater lexical divergence; the lowest per column (bottom half) is indicated by (*).

strive to preserve as much meaning as possible while using lexically diverse expression. Otherwise, the result is either a trivial re-write or fails to convey the original meaning. We understand these two metrics as interdependent and sometimes conflicting – a high lexical diversity likely sacrifices semantic similarity. The goal of PARABANK is to offer not only a balance between the two, but also options across the spectrum for different applications. Of course, good paraphrases should also be fluent in their expression, so we also evaluate paraphrases for grammaticality and meaningfulness, independent of their reference.

For brevity, we picked 5 PARABANK systems (bold in Tab. 3) to cover negative, positive, positional, and no constraint. We also include system 21 (3 lowest IDF tokens) to show that too many constraints might significantly hurt semantic similarity. Full evaluations on all proposed systems will be included with the release of the resource.

5.1 Scoring PARABANK paraphrases

Following the approach of PPDB 2.0 (Pavlick et al., 2015a), we trained a supervised model on the human annotations we collected. We extracted several features from reference-paraphrase pairs to predict human judgments of semantic similarity on all paraphrases, with the exception of those whose reference contains more than 100 tokens post-BPE. The regression model achieves reasonable correlation with human judgment on the test data with a Spearman's ρ of 0.53 on CzEng and 0.63 on Giga.

5.2 Baseline comparison

Our baseline system with no lexical constraints applied shows substantial improvement compared to PARANMT. This could be a combination of improved training data and NMT framework. PARANMT is trained on a 51.4M subset of CzEng1.6, while PARABANK used CzEng1.7, a 57.0M subset of CzEng1.6. We also switched to SOCKEYE as our training framework. This baseline improvement gives us more flexibility to pursue explicit lexical diversity with reasonable compromise in semantic similarity.

5.3 Semantic similarity

Human judgment remains the gold standard of semantic similarity. We randomly sampled 100 Czech-English sentence pairs from each of the four English token lengths: 5, 10, 20, and 40. We translate 400 sentences from CzEng under 34 PARABANK systems (without PPDB constraints) and 400 sentences from Giga under 37 PARABANK systems. Then, we merge identical outputs and add in the corresponding PARANMT entries to the CzEng evaluation.

We randomize the paraphrase pool and formulate them into Mechanical Turk Human Intelligence Tasks. Inspired by the interface of EASL (Sakaguchi and Van Durme, 2018), we ask workers to assign each paraphrase a score between 0 and 100 by adjusting a slider bar. Each worker is presented with one reference sentence and five attempted paraphrases at the same time. Occasionally, we present workers the reference sentence itself as a candidate paraphrase and expect it to receive a perfect score. Workers who fail to do so more than 10% of all times are disqualified for inattentiveness. In total, we incorporated the annotations of 44 workers who contributed at least 25 judgments. Each paraphrase receives independent judgments from at least 3 different workers.

We then calculate the average score for each sentence pair, before averaging over all pairs from each ParaBank system (or ParaNMT). The final score for each system is a number between 0 and 100.

The top half of Tab. 4 shows the average human judgment over 100 sentences per reference length for ParaBank systems and ParaNMT, grouped by sentence length.

Best performing ParaBank systems from each reference length outperform ParaNMT relatively by 5.0%, 6.6%, 10.2%, and 14.0% in terms of semantic similarity (corresponding to 5, 10, 20, 40 tokens per reference sentence).

5.4 Lexical diversity

We used a modified BLEU score to evaluate lexical diversity and **a lower score suggests a higher lexical diversity**. Specifically, we concatenate[6] multiple paraphrastic sentence pairs into one reference paragraph and one paraphrase paragraph, and calculate the associated BLEU score **without brevity penalty**. This modification ensures that we don't reward shorter paraphrases. We compared the result with a naive unigram precision metric and they show a strong correlation with a Spearman's ρ of 0.98.

The bottom half of Tab. 4 shows this modified BLEU score for each ParaBank system and ParaNMT, grouped by reference length. For every length, there is at least one ParaBank system that exhibits higher lexical diversity than ParaNMT; unsurprisingly, the ParaBank systems that apply the greatest number of lexical constraints tend to yield the greatest lexical diversity (e.g., system 21).

5.5 Meaningfulness and grammaticality

We ask annotators to comment on each paraphrase's fluency by flagging sentences that are completely nonsensical or indisputably ungrammatical. We consider a sentence nonsensical or ungrammatical when at least one independent annotator flags it as so.

We then calculate the percentage of sentences that are deemed both meaningful and grammatical for each ParaBank system and ParaNMT.

The result is shown in Tab. 5. System 34 (avoid first 3 tokens) shows a 12.6% improvement over ParaNMT. In all, 21 out of 34 proposed ParaBank systems contain a smaller proportion of nonsensical or ungrammatical sentences than ParaNMT. The full set of annotations are available with the resource.

#	System	Cz-En	Fr-En
-	ParaNMT	73.25	-
17	$\ominus$3rdlow IDF	**76.50**	73.75
21	$\ominus$3 low IDF	65.50	65.00
28	no con.	**81.75**	80.00*
31	$\oplus$3rdtop IDF	**74.00**	72.75
34	$\ominus$first 3 tks	**82.50***	77.50
35	PPDB equ	-	69.97
36	PPDB fwd	-	71.19
37	PPDB rev	-	52.46

Table 5: Percentage of paraphrases for each system that are rated by human annotators as both grammatical *and* meaningful, independent of similarity to the reference sentence. Improvements over ParaNMT (Czech-English only) in bold. Asterisk (*) indicates best in column.

6 Constrained monolingual rewriting

Napoles, Callison-Burch, and Post (2016) explored sentential rewriting with machine translation models. Inspired by their work, we use a subset of ParaBank, with more than 50 million English paraphrastic sentence pairs (English text from CzEng as source, ParaBank outputs as target), to train a monolingual NMT model, and decode with the same types of constraint systems. We present the following result as a proof of concept that highlights the potential for and problems with the most straightforward instantiation of the model. A thorough investigation of building such a monolingual model is outside the scope of this work.

We decide to use the same LSTM model instead of more advanced self-attention models to contrast between the bilingual and monolingual models. After training for one epoch, we decode the model with the same 5 constraint systems (no. 17, 21, 28, 31, 34) evaluated for the bilingual model, and ask human annotators[7] to compare their semantic similarity to the reference sentence in the same way. We sampled 100 sentences across the same 4 lengths (25 sentences per length); each sentence receives at least 3 independent judgments. The semantic similarity scores for this monolingual system are reported in Tab. 6 ("Monolingual") alongside lexical diversity scores (modified BLEU).

Outputs from the monolingual model show a significant boost in semantic similarity compared to bilingual counterparts, system 28 (no constraint) shows an improvement of 16.7%. This is accompanied by an increase in BLEU score, a sign of less lexical diversity. Example outputs from the monolingual model can be found in Tab. 7.

As evidenced in our examples, some monolingual systems may generate slightly more nonsensical or ungrammatical sentences than their bilingual counterparts: future work will pursue more extensive model training and data filtering for the monolingual model. Our intent here is to foremost illustrate the quality of ParaBank as a resource, while illustrating the feasibility of training and employing a monolingual

[6] After switching all tokens to lowercase and stripping punctuation to avoid rewarding trivial re-writes.

[7] Same setup as §5.3. The result includes 8 workers who contributed more than 25 judgments.

#	System	Semantic Similarity		Lexical Diversity	
		Bilingual (Cz-En)	Monolingual	Bilingual (Cz-En)	Monolingual
-	ParaNMT	73.89 (std. 4.60)	-	27.36 (4.89)	-
17	$\ominus$3rd low IDF	76.79 (std. 3.77)	74.56 (std. 9.17)	27.20 (std. 14.67)	33.04 (std. 18.58)
21	$\ominus$3 low IDF	66.48 (std. 8.25)	63.35 (std. 19.59)	14.28 (std. 13.35)	24.82 (std. 20.52)
28	no con.	81.91 (std. 3.11)	86.25 (std. 3.94)	33.90 (std. 11.58)	44.36 (std. 13.69)
31	$\oplus$3rd top IDF	80.64 (std. 3.60)	83.09 (std. 4.23)	33.48 (std. 12.66)	43.80 (std. 13.39)
34	$\ominus$first 3 tks	80.04 (std. 1.95)	84.47 (std. 4.10)	30.20 (std. 13.50)	39.81 (std. 10.52)
-	Reference	99.81 (std. 0.16)	-	100.00 (std. 0.00)	-

Table 6: Comparison of bilingual (Czech-English) and monolingual (English-English) paraphrasing systems in terms of (1) **semantic similarity** as rated by human annotators on a scale of 0-100, and (2) **lexical diversity** as measured by a modified BLEU score without length penalty, where lower BLEU scores are taken as evidence of greater lexical diversity. We observe that similarity and diversity scores for the monolingual rewriting systems exhibit higher variance than bilingual systems (sample standard deviations given in parentheses); however, the monolingual rewriter is able to generate English paraphrases in the absence of a Czech reference sentence.

Reference
Hey, it's nothing to be ashamed of.

Paraphrases from the monolingual model
Hey, it's nothing to be embarrassed.
Hey, it's nothing to be ashamed.
Hey, it's not like you're ashamed of.
Hey, you don't have to worry about that.
Hey, you don't have to be ashamed.
Hey, there's nothing you can be ashamed of.
You don't have to be ashamed of it.
Hey, there's nothing you can do about it.
Oh, hey, it's no big deal.

Table 7: Example paraphrases generated from the monolingual rewriting model, after applying the same set of lexical constraints described in Tab. 3 and merging duplicates.

sentence rewriting model built atop the ParaBank artifact.

7 Conclusions and Future Work

We created ParaBank by decoding a neural machine translation (NMT) model with lexical constraints. We applied our methods to CzEng 1.7 and Giga, leading to a large collection of paraphrases with 79.5 million references and on average 4 paraphrases per reference, which we make available for download at `http://nlp.jhu.edu/parabank`. Via large-scale crowdsourced annotations, we found the overall best performing ParaBank system exhibits an 8.5% relative improvement in terms of semantic similarity over prior work. Analysis on lexical diversity showed the potential of ParaBank as a more diverse and less noisy paraphrastic resource. In addition to releasing hundreds of millions of English sentential paraphrases, we also release a free, pre-trained, model for monolingual sentential rewriting, as trained on ParaBank.

With the existence of ParaBank and an initial monolingual rewriting model, future work can investigate how more advanced NMT models, such as those with self-attention structures, can lead to better rewriting. One may also investigate the automatic expansion of resources for a variety of NLP tasks. For example, in Machine Translation one might create sentential paraphrases from the English side of bitexts for low-resource languages: cases where only small numbers of gold translations exist in English, and are expensive or otherwise problematic to expand by hand. In Information Extraction, one may rewrite sentences with structured annotations such as for Named Entity Recognition (NER), with positive constraints that phrases representing known NER spans be preserved while some tokens of the remainder be negatively constrained, thereby providing additional novel sentential contexts for IE system training. In educational NLP technology, one might wish to rewrite a sentence that includes or excludes target vocabulary words a language learner does not understand or is trying to acquire. There are many other such examples in NLP where the ability to rewrite existing datasets with lexical constraints could lead to significantly larger and more diverse training sets, with no additional human labor. To pursue such work may require a large, high quality monolingual bitext to train a rewriting model, and an NMT decoder supporting both positive and negative constraints, such as we have introduced here.

Acknowledgments

We thank Chu-Cheng Lin and three anonymous reviewers for their thoughtful feedback. This research was supported in part by DARPA AIDA. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The berkeley framenet project. In *Proceedings of ACL/ICCL*, ACL '98, 86–90.

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

Bojar, O.; Dušek, O.; Kocmi, T.; Libovický, J.; Novák, M.; Popel, M.; Sudarikov, R.; and Variš, D. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing

Tools Dockered. In Sojka, P.; Horák, A.; Kopeček, I.; and Pala, K., eds., *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, 231–238. Cham / Heidelberg / New York / Dordrecht / London: Masaryk University.

Callison-Burch, C.; Koehn, P.; Monz, C.; and Schroeder, J. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 1–28. Athens, Greece: Association for Computational Linguistics.

Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2013. Ppdb: The paraphrase database. In *Proceedings NAACL-HLT 2013*, 758–764.

Ganitkevitch, J. 2018. *Large-Scale Paraphrasing for Text-to-Text Generation*. Ph.D. Dissertation, Johns Hopkins University.

Harris, Z. S. 1954. Distributional structure. *WORD* 10(2-3):146–162.

Hieber, F.; Domhan, T.; Denkowski, M.; Vilar, D.; Sokolov, A.; Clifton, A.; and Post, M. 2017. Sockeye: A toolkit for neural machine translation. *CoRR* abs/1712.05690.

Hokamp, C., and Liu, Q. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of ACL*, 1535–1546.

Honnibal, M., and Montani, I. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, 110–119. San Diego, California: ACL.

Lin, D., and Pantel, P. 2001. Dirt @sbt@discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, 323–328. New York, NY, USA: ACM.

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

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

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

Pang, B.; Knight, K.; and Marcu, D. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL*.

Pavlick, E.; Rastogi, P.; Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2015a. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of ACL/IJCNLP*, volume 2.

Pavlick, E.; Wolfe, T.; Rastogi, P.; Callison-Burch, C.; Dredze, M.; and Van Durme, B. 2015b. Framenet+: Fast paraphrastic tripling of framenet. In *Proceedings of the ACL/IJCNLP*, volume 2.

Post, M., and Vilar, D. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of NAACL*.

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

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

Sakaguchi, K.; Post, M.; and Van Durme, B. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of WMT*.

Schuler, K. K. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. Dissertation, University of Pennsylvania.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, 1715–1725. Berlin, Germany: ACL.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ICCL/ACL*.

Straková, J.; Straka, M.; and Hajič, J. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of ACL: System Demonstrations*.

Weisman, H.; Berant, J.; Szpektor, I.; and Dagan, I. 2012. Learning verb inference rules from linguistically-motivated evidence. In *Proceedings of EMNLP/CNLL*, EMNLP-CoNLL '12.

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