A Continuously Growing Dataset of Sentential Paraphrases from Twitter

Wuwei Lan[1], Siyu Qiu[2], Hua He[3], Wei Xu[1]
[1] Department of Computer Science and Engineering
Ohio State University
`{lan.105@osu.edu, weixu@cse.ohio-state.edu}`
[2] Computer and Information Science Department
University of Pennsylvania
`siqiu@seas.upenn.edu`
[3] Department of Computer Science
University of Maryland, College Park
`huah@umd.edu`

Abstract

A major challenge in paraphrase research is the lack of parallel corpora. In this paper, we present a new method to collect large-scale sentential paraphrases from Twitter by linking tweets through shared URLs. The main advantage of our method is its simplicity, as it gets rid of the classifier or human in the loop needed to select data before annotation and subsequent application of the paraphrase identification algorithms in the previous work. We present the largest human-labeled paraphrase corpus to date of 51,524 sentence pairs and the first cross-domain benchmarking for automatic paraphrase identification. In addition, we show that more than 30,000 new sentential paraphrases can be easily and continuously captured every month at ∼70% precision, and demonstrate their utility for downstream NLP tasks through phrasal paraphrase extraction. We will make our code and data publicly available.[1]

1 Introduction

A paraphrase is a restatement of meaning using different expressions (Bhagat and Hovy, 2013). Researchers have shown benefits of using paraphrases in a wide range of applications (Madnani and Dorr, 2010), including question answering (Fader et al., 2013), semantic parsing (Berant and Liang, 2014), information extraction (Sekine, 2006; Zhang et al., 2015), machine translation (Mehdizadeh Seraj et al., 2015), textual entailment (Dagan et al., 2006; Bjerva et al., 2014; Marelli

a 15-year-old girl, a 15yr old, a 15 y/o girl
fetuses, fetal tissue, miscarried fetuses
responsible for, guilty of, held liable for, liable for
UVA administrator, UVa official, U-Va. dean, University of Virginia dean
FBI Director backs CIA finding, FBI agrees with CIA, FBI backs CIA view, FBI finally backs CIA view, FBI now backs CIA view, FBI supports CIA assertion, FBI-Clapper back CIA's view, The FBI backs the CIA's assessment, FBI Backs CIA,
Donald Trump, DJT, Drumpf, Mr Trump, Idiot Trump, Chump, Evil Donald, #OrangeHitler, Donald @realTrump, D*nald Tr*mp, Comrade #Trump, Crooked #Trump, CryBaby Trump, Daffy Trump, Donald KKKrump, Dumb Trump, GOPTrump, Incompetent Trump, He-Who-Must-Not-Be-Named, Pres-elect Trump, President-Elect Trump, President-elect Donald J . Trump, PEOTUS Trump, Emperor Trump

Table 1: Up-to-date phrasal paraphrases automatically extracted from Twitter with our new method.

et al., 2014; Izadinia et al., 2015), vector semantics (Faruqui et al., 2015; Wieting et al., 2015), and semantic textural similarity (Agirre et al., 2015). Studying paraphrases in Twitter can also help track unfolding events (Vosoughi and Roy, 2016) or the spread of information (Bakshy et al., 2011) on social networks.

In this paper, we address a major challenge in paraphrase research — the lack of parallel corpora. There are **only two** publicly available datasets of naturally occurring sentential paraphrases and non-paraphrases:[2] the MSRP corpus derived from clustered news articles (Dolan and Brockett, 2005) and the PIT-2015 corpus from Twitter trending topics (Xu et al., 2014, 2015). Our goal is not only to create a new annotated paraphrase corpus, but to identify a new data source and method that can narrow down the

---

[2]Meaningful non-paraphrases (pairs of sentences that have similar wordings or topics but different meanings, and that are not randomly or artificially generated) have been very difficult to obtain but are very important, because they serve as necessary distractors in training and evaluation.

Name	Genre	Size	Sentence Length	Multi-Ref.	Non-Para.
MSR Paraphrase Corpus (MSRP)	news	5801 pairs	18.9 words	no	yes
Twitter Paraphrase Corpus (PIT-2015)	Twitter	18,762 pairs	11.9 words	some	yes
Twitter News URL Corpus (this work)	Twitter	44,365 pairs	14.8 words	yes	yes
MSR Video Description Corpus	YouTube	123,626 sentences	7.03 words	yes	no

Table 2: Summary of publicly available large sentential paraphrase corpora with manual quality assurance. Our Twitter News URL Corpus has the advantages of including both meaningful non-paraphrases (Non-Para.) and multiple references (Multi-Ref.), which are important for training paraphrase identification and evaluating paraphrase generation, respectively.

search space of paraphrases without using the classifier-biased or human-in-the-loop data selection as in MSRP and PIT-2015. This is so that sentential paraphrases can be conveniently and continuously harvested in large quantities to benefit downstream applications.

We present an effective method to collect sentential paraphrases from tweets that refer to the same URL and contribute a new gold-standard annotated corpus of 51,524 sentence pairs, which is the largest to date (Table 2). We show the different characteristics of this new dataset contrasting the two existing corpora through the first systematic study of paraphrase identification across multiple datasets. Our new corpus is complementary to previous work, as the corpus contains multiple references of both formal well-edited and informal user-generated texts. This is also the first work that provides a continuously growing collection, with more than 30,000 new sentential paraphrases per month automatically labeled at ~70% precision. We demonstrate that up-to-date phrasal paraphrases can then be extracted via word alignment (see examples in Table 1). We plan to continue collecting paraphrases using our method and release a constantly updating paraphrase resource.

2 Existing Paraphrase Corpora and Their Limitations

To date, there exists only two publicly available corpora of both sentential paraphrases and non-paraphrases:

MSR Paraphrase Corpus [MSRP] (Dolan et al., 2004; Dolan and Brockett, 2005) This corpus contains 5,801 pairs of sentences from news articles, with 4,076 for training and the remaining 1,725 for testing. It was created from clustered news articles by using a SVM classifier (using features including string similarity and Word-Net synonyms) to gather likely paraphrases, then annotated by human on semantic equivalence. The

MSRP corpus has a known deficiency skewed toward over-identification (Das and Smith, 2009), because the "purpose was not to evaluate the potential effectiveness of the classifier itself, but to identify a reasonably large set of both positive and plausible 'near-miss' negative examples" (Dolan and Brockett, 2005). It contains a large portion of sentence pairs with many ngrams shared in common.

Twitter Paraphrase Corpus [PIT-2015] (Xu et al., 2014) This corpus was derived from Twitter's trending topic data. The training set contains 13,063 sentence pairs on 400 distinct topics, and the test set contains 972 sentence pairs on 20 topics. As numerous Twitter users spontaneously talk about varied topics, this dataset contains many lexically divergent paraphrases. However, this method requires a manual step of selecting topics to ensure the quality of collected paraphrases, because many topics detected automatically are either incorrect or too broad. For example, the topic "New York" relates to tweets with a wide range of information and cannot narrow the search space down enough for human annotation and the subsequent application of the classification algorithms.

3 Constructing the Twitter URL Paraphrase Corpus

For paraphrase acquisition, it has been crucial to find a simple and effective way to locate paraphrase candidates (see related work in Section 6). We show the efficacy of tracking URLs in Twitter. This method does not rely on automatic news clustering as in MSRP or topic detection as in PIT-2015, but it keeps collecting good candidate paraphrase pairs in large quantities.

3.1 Data Source: News vs. Streaming

We extracted the embedded URL in each tweet and used Twitter's Search API to retrieve all tweets

Twitter News URL Corpus	
Original Tweet	Samsung halts production of its Galaxy Note 7 as battery problems linger
Paraphrase	#Samsung temporarily suspended production of its Galaxy #Note7 devices following reports
	News hit that @Samsung is temporarily halting production of the #GalaxyNote7.
	Samsung still having problems with their Note 7 battery overheating. Completely halt production.
	SAMSUNG HALTS PRODUCTS OF GALAXY NOTE 7 . THE BATTERIES ARE * STILL * EXPLODING .
Non-Paraphrase	in which a phone bonfire in 1995–a real one–is a metaphor for samsung's current note 7 problems
	samsung decides, "if we don't build it, it won't explode."
	Samsung's Galaxy Note 7 Phones AND replacement phones have been going up in flames due to the defective batteries

Table 3: A representative set of tweets linked by a URL originated from news agencies (this work).

Twitter Streaming URL Data	1 dasviness louistomlinson overhears harrystyles on the phone
	when she likes tall guys ??? ??? vine by justjamiie
	shineeasvines jonghyun when he wears shoe lifts
	idaliaorellana kimmvanny ladyfea 21 hahaha if he does it he needs heels

Table 4: A representative set of tweets linked by a URL in streaming data (generally poor readability).

that contain the same URL. Some tweets use shortened URLs, which we resolve as full URLs. We tracked 22 English news accounts in Twitter to create the paraphrase corpus in this paper (see examples in Table 3). We will extend the corpus to include other languages and domains in future work.

As shown in Table 5, nearly all the tweets posted by news agencies have embedded URLs. About 51.17% posts contain two URLs, usually one pointing to a news article and the other to media such as a photo or video. In contrast, although close to half of the tweets in the streaming data contain at least one URL, most of them are very hard to read (see examples in Table 4).

Data Source	tweets w/o url	avg #url per tweet	avg # url (news) per tweet
Streaming Data	55.8%	0.52	
@nytimes	1.2%	1.31	0.988
@cnnbrk	0.0%	1.17	1
@BBCBreaking	1.0%	1.32	0.99
@CNN	0.0%	1.85	1
@ABC	1.7%	1.26	0.983
@NBCNews	1.1%	1.63	0.989

Table 5: Statistics of tweets in Twitter's streaming data and news account data. Many tweets contain more than one URL because media such as photo or video is also represented by URLs.

3.2 Filtering of Retweets

Retweeting is an important feature in Twitter and consists two types: automatic and manual retweets. An automatic retweet is done by clicking the retweet button on Twitter and is easy to remove using the Twitter API. A manual retweet

occurs when the user creates a new tweet by copying and pasting the original tweet and possibly adding some extras, such as hashtags, usernames or comments. We preprocessed the tweets using a tokenizer[3] (Gimpel et al., 2011) and an in-house sentence splitter. We then filtered out the manual retweets using a set of rules, checking if one tweet was a sub- or super- string of the original tweet, or if it only differed in punctuation, or the contents of the "twitter:title" or "twitter:description" tag in the linked HTML file of the news article.

This filtering step is crucial for removing the redundant tweets with slight rephrases, as shown in Table 6. We used PINC, a standard paraphrase metric, to measure ngram-based dissimilarity (Chen and Dolan, 2011). Jaccard is a token-based string similarity measure (Jaccard, 1912).

	avg #tweets (STD)	avg PINC	avg Jaccard
before filtering	205.51 (219.66)	0.6153	0.3635
after filtering	74.75 (94.39)	0.7603	0.2515

Table 6: Impact of filtering of manual retweets. Higher PINC and lower Jaccard scores indicate more significant rephrasing.

3.3 Gold Standard Corpus

To get the gold-standard paraphrase corpus, we obtained human labels on Amazon Mechanical Turk. We showed annotators an original sentence, and asked them to select sentences with the same meaning from 10 candidate sentences. For each question, we recruited 6 annotators and paid $0.03 to each worker. On average, each question

---

[3]http://www.cs.cmu.edu/~ark/TweetNLP/

took about 53 seconds to finish.[4] For each sentence pair, we aggregated the paraphrase and non-paraphrase labels using the majority vote.

We constructed the largest gold standard paraphrase corpus to date, with 42,200 tweets of 4,272 distinct URLs annotated in the training set and 9,324 tweets of 915 distinct URLs in the test set. The training data was collected between 10/10/2016 and 11/22/2016, and testing data between 01/09/2017 and 01/19/2017. In Section 4, we contrast the characteristics of our data against existing paraphrase corpora.

Quality Control We evaluated the annotation quality of each worker using Cohen's kappa agreement (Artstein and Poesio, 2008) against the majority vote of other workers. We askd the best workers (the top 528 out of 876) to label more data by republishing the questions done by workers with low reliability (Cohen's kappa <0.4).

Inter-Annotator Agreement In addition, we also had 300 sampled sentence pairs annotated by an expert. The annotated agreement is 0.739 by Cohen's kappa between the expert and the majority vote of 6 crowdsourcing workers. If we assume the expert annotation is gold, the precision of worker vote is 0.871, the recall is 0.787, and F1 is 0.827, similar to those of PIT-2015.

3.4 Continuous Harvesting of Sentential Paraphrases

Since our method directly applies to raw tweets, we could train automatic paraphrase identifiers using the human-labeled data to continuously extract sentential paraphrases from Twitter. It turns out that we could construct a silver-standard paraphrase corpus at about 70% precision that grows monthly with more than 30,000 new sentential paraphrases. In Section 4, we discuss automatic paraphrase identification in detail. In Section 5, we demonstrate the usefulness of these automatically identified sentential paraphrases.

4 Comparison of Paraphrase Corpora

Though paraphrasing has been widely studied, supporting analyses and experiments have thus far often only been conducted on a single dataset. In this section, we present a comparative analysis of

our newly constructed gold-standard corpus with two existing corpora by 1) individually examining the instances of paraphrase phenomena and 2) benchmarking a range of automatic paraphrase identification approaches.

4.1 Paraphrase Phenomena

In order to show the paraphrase differences across these three datasets, we sampled 100 sentential paraphrases from each training set and counted the occurrences of each phenomenon in the following categories: Elaboration (textual pairs can differ in total information content, such as *Trump's ex-wife Ivana* and *Ivana Trump*), Phrasal (alternates of phrases, such as *taking over* and *replaces*), Spelling (spelling variants, such as *Trump* and *Trumpf*), Synonym (such as *said* and *told*), Anaphora (a full noun phrase in one sentence that corresponds to the counterpart, such as *@MarkKirk* and *Kirk*) and Reordering (when a word, phrase or the whole sentence reorders, or even logically reordered, such as *Matthew Fishbein questioned him* and *under questioning by Matthew Fishbein*). We report the average number of occurrences of each paraphrase type per sentence pair for each corpus in Table 7. As sentences tend to be longer in MSRP and shorter in PIT-2015, we also normalized the numbers by the length of sentences to be more comparable to the URL dataset.

These three datasets exhibit distinct and complementary compositions of paraphrase phenomena. MSRP has more synonyms, because authors of different news articles may use different and rather sophisticated words. PIT-2015 contains a lot of phrasal paraphrases, probably due to the fact that most tweets under the same trend-

per sentence	MSRP	PIT-2015	URL
Elaboration	0.60	0.23	**0.79**
Spelling	0.17	0.13	**0.35**
Synonym	**0.26**	0.10	0.13
Phrasal	0.42	**0.56**	0.35
Anaphora	0.27	0.08	**0.33**
Reordering	**0.53**	0.33	0.49
adjusted by sent length*	MSRP*	PIT-2015*	URL
Elaboration	0.42	0.36	**0.79**
Spelling	0.12	0.21	**0.35**
Synonym	**0.18**	0.16	0.13
Phrasal	0.29	**0.89**	0.35
Anaphora	0.19	0.13	**0.33**
Reordering	0.37	**0.52**	0.49

Table 7: Mean number of instances of paraphrase phenomena per sentence pair across three corpora.

ing topic are written spontaneously and independently. Our URL dataset shows more elaboration, spelling and anaphora paraphrase phenomena, showing that many URL-embedded tweets are created by users with a conscious intention to rephrase the original news headline.

4.2 Automatic Paraphrase Identification

In the following we provide a benchmark on paraphrase identification to better understand various models, as well as the characteristics of our new corpus compared to the existing ones. We focus on the binary classification of the paraphrase/non-paraphrase, and report the maximum F1 measure of any point on the precision-recall curve.

4.2.1 Models

We chose several representative technical approaches for automatic paraphrase identification:

GloVe (Pennington et al., 2014) This is a word representation model trained on aggregated global word-word co-occurrence statistics from a corpus. We used 300-dimensional word vectors trained on Common Crawl and Twitter, summed the vectors for each sentence, and computed the cosine similarity.

WMF/OrMF Weighted Matrix Factorization (WMF) (Guo and Diab, 2012) is an unsupervised latent space model. The unobserved words are carefully handled, which results in more robust embeddings for short texts. Orthogonal Matrix Factorization (OrMF) (Guo et al., 2014) is the extension of WMF, with an additional objective to obtain nearly orthogonal dimensions in matrix factorization to discount redundant information.

LR The logistic regression (LR) model incorporates 18 features based on 1-4 gram overlaps between two sentences (s_1 and s_2) (Das and Smith, 2009). The features are of the form $precision_n$ (number of n-gram matches divided by the number of n-grams in s_1), $recall_n$ (number of n-gram matches divided by the number of n-grams in s_2), and F_n (harmonic mean of recall and precision).

LEX-WMF/LEX-OrMF This is an open-sourced adaptation (Xu et al., 2014) of LEXDIS-CRIM (Ji and Eisenstein, 2013) with comparable performance. Specifically, for the **(vec)** version, vectors of a pair of sentences $\vec{v_1}$ and $\vec{v_2}$ are converted into one feature vector, $[\vec{v_1} + \vec{v_2}, |\vec{v_1} - \vec{v_2}|]$, by concatenating the element-wise sum $\vec{v_1} + \vec{v_2}$

Method	F1	Precision	Recall
Random	0.799	0.665	1.0
Edit Distance	0.799	0.666	1.0
GloVe	0.812	0.707	0.952
LR	0.829	0.741	0.941
WMF (vec)	0.817	0.713	0.956
LEX-WMF (vec)	**0.836**	**0.751**	**0.943**
OrMF (vec)	0.820	0.733	0.930
LEX-OrMF (vec)	**0.833**	**0.741**	**0.950**
WMF (sim)	0.812	0.728	0.918
LEX-WMF (sim)	0.831	0.732	0.962
OrMF (sim)	0.815	0.699	0.976
LEX-OrMF (sim)	0.832	0.735	0.958
MultiP	0.800	0.667	0.998
DeepPairwiseWord	**0.834**	**0.763**	**0.919**

Table 8: Paraphrase models in the MSR Paraphrase Corpus (MSRP). The bold font in the table represents top three models in the dataset.

Method	F1	Precision	Recall
Random	0.346	0.209	1.0
Edit Distance	0.363	0.236	0.789
GloVe	0.484	0.396	0.617
LR	0.645	0.669	0.623
WMF (vec)	0.594	0.681	0.526
LEX-WMF (vec)	0.635	0.655	0.617
OrMF (vec)	0.594	0.681	0.526
LEX-OrMF (vec)	0.638	0.579	0.709
WMF (sim)	0.553	0.570	0.537
LEX-WMF (sim)	**0.651**	**0.657**	**0.646**
OrMF (sim)	0.563	0.591	0.537
LEX-OrMF (sim)	0.644	0.632	0.657
MultiP	**0.721**	**0.705**	**0.737**
DeepPairwiseWord	**0.667**	**0.725**	**0.617**

Table 9: Paraphrase models in the Twitter Paraphrase Corpus (PIT-2015).

Method	F1	Precision	Recall
Random	0.327	0.195	1.000
Edit Distance	0.526	0.650	0.442
GloVe	0.583	0.607	0.560
LR	0.683	0.669	0.698
WMF (vec)	0.660	0.640	0.680
LEX-WMF (vec)	**0.693**	**0.687**	**0.698**
OrMF (vec)	0.662	0.625	0.703
LEX-OrMF (vec)	**0.691**	**0.709**	**0.674**
WMF (sim)	0.659	0.595	0.738
LEX-WMF (sim)	0.688	0.632	0.754
OrMF (sim)	0.660	0.690	0.632
LEX-OrMF (sim)	0.688	0.630	0.758
MultiP	0.536	0.386	0.875
DeepPairwiseWord	**0.749**	**0.803**	**0.702**

Table 10: Paraphrase models in Twitter URL Corpus (this work).
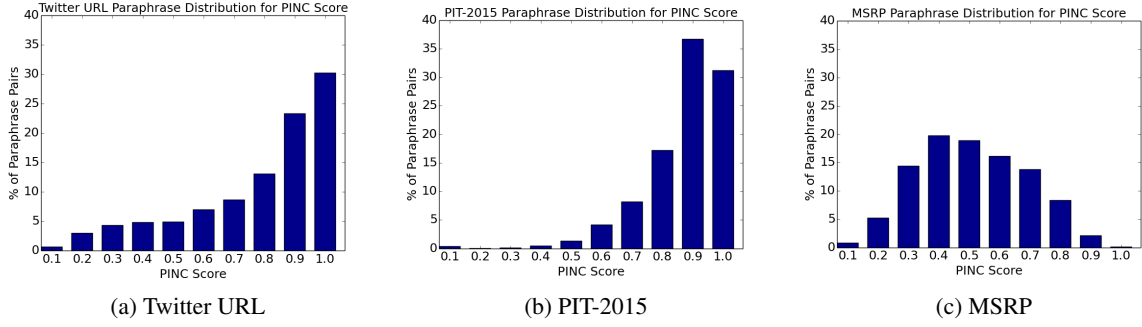
| (a) Twitter URL | (b) PIT-2015 | (c) MSRP |

Figure 1: Comparison of ngram dissimilarity (PINC score) in sentential paraphrases across three corpora. The MSRP contains sentential paraphrases with more ngram overlaps (low PINC). Our URL corpus and PIT-2015 contain more lexically divergent paraphrases (high PINC).

and absolute difference $|\vec{v_1} - \vec{v_2}|$, then combining it with n-gram overlapping features to train a LR classifier. We also provide the **(sim)** variation, which directly uses the single cosine similarity score between two sentence vectors.

MultiP MultiP (Xu et al., 2014) is a multi-instance learning model suited for short messages on Twitter. The at-least-one-anchor assumption in this model looks for two sentences that have a topical phrase in common, plus at least one pair of anchor words that carry a similar key meaning. This model achieves the best performance in PIT-2015 (Xu et al., 2014) dataset.

DeepPairwiseWord He et al. (2016) developed a deep neural network model that focuses on important pairwise word interactions across input sentences. This model innovates in proposing a similarity focus layer and a 19-layer very deep convolutional neural network to guide model attention to important word pairs. It has shown state-of-the-art performance on several textual similarity measurement datasets.

4.2.2 Model Performance and Dataset Difference

The results on three benchmark paraphrase corpora are shown in Table 8, 9 and 10. The random baseline reflects that close to 80% sentence pairs are paraphrases in the MSPR corpus. This is untypical in the real-world text data and may cause falsely positive predictions.

Both the edit distance and the LR models exploit surface word features. In particular, the LR model that uses lemmatization and ngram overlap features achieves very competitive performance on all datasets. Figure 1 shows a closer look at

ngram differences across datasets measured by the PINC metric (Chen and Dolan, 2011), which is the opposite of BLEU. MSRP consists of paraphrases with more ngram overlap (lower PINC), while PIT-2015 contains shorter and more lexically dissimilar sentences. Our new URL corpus is in between the two, and is more similar to PIT-2015. It includes user's intentional rephrasing of an original tweet from a news agency with some words untouched, as well as some dramatic paraphrases that are challenging for any automatic identification methods, such as *CO2 levels mark 'new era' in the world's changing climate* and *CO2 levels haven't been this high for 3 to 5 million years*.

MultiP model exploits a restrictive constraint that the candidate sentence pairs share a same topical phrase. It achieves the best performance on PIT-2015, which naturally contains such phrases. However, for MSRP and URL datasets, we had to use the name entity tagged with the longest span as an approximation of a shared topic phrase in MultiP and thus suffered a performance drop.

Both Glove and WMT/OrMF utilize the underlying co-occurrence statistics of the text corpus. WMT/OrMF use global matrix factorization to project sentences into lower dimension and show great advantages on measuring sentence-level semantic similarities over Glove, which focuses on word representations. Figure 2 shows the fine-grained distribution of the OrMF-based cosine similarities and that the URL-linked Twitter data works well with OrMF to yield sentential paraphrases. Once combined with ngram overlap features, LEX-WMF and LEX-OrMF show consistently high performance across different datasets, close to the more complicated DeepPairwiseWord. The similarity focus mechanism on important pair-
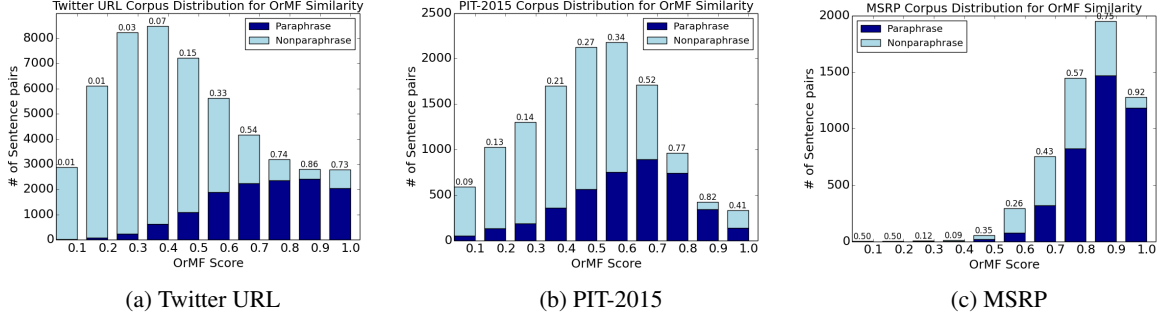
| (a) Twitter URL | (b) PIT-2015 | (c) MSRP |

Figure 2: Comparison of OrMF-based distributional semantic similarity across three paraphrase corpora.

wise word interactions in DeepPairwiseWord is more helpful for the two Twitter datasets, due to the fact that they contain lexically divergent paraphrases while MSRP has an artificial bias toward ngram overlapping sentences.

5 Extracting Phrasal Paraphrases

We can apply paraphrase identification models trained on our gold standard corpus to unlabeled Twitter data and continuously harvest sentential paraphrases in large quantities. We used the opensourced LEX-OrMF model and obtained 114,025 sentential paraphrases (system predicted probability ≥ 0.5 and average precision $= 69.08\%$) from raw 1% free Twitter data between 10/10/2016 and 01/10/2017. To demonstrate the utility, we show that we can extract up-to-date lexical and phrasal paraphrases from this data.

5.1 Phrase Extraction and Ranking

One of the most successful ideas to obtain lexical and phrasal paraphrases in large quantities is through word alignment, then ranking for better quality. It is originated from (Bannard and Callison-Burch, 2005) and previously applied to bilingual parallel data to create PPDB (Ganitkevitch et al., 2013; Pavlick et al., 2015). There was little previous work utilizing monolingual parallel data since it is not as naturally available as bitexts.

We used the GIZA++ word aligner in the Moses machine translation toolkit (Koehn et al., 2007) and extracted 245,686 phrasal paraphrases. Some examples are shown in Table 1. We additionally explored two supervised monolingual aligners: Jacana aligner (Yao et al., 2013) and Md Sultan's aligner (Sultan et al., 2014). We ranked the phrase pairs using four different scores:

- **Language Model Score** Let $w_{-2}w_{-1}pw_1w_2$

be the context of the phrase p. We considered a phrase p' to be a good substitute for p if $w_{-2}w_{-1}p'w_1w_2$ is a likely sequence according to a language model (Heafield, 2011) trained on Twitter data.

- **Translation Score** Moses provides translation probabilities $\varphi(p|p')$.
- **Glove Score** We used Glove (Pennington et al., 2014) pretrained 100-dimensional Twitter word vectors and cosine similarity.
- **Our Score** We trained a supervised SVM regression model using 500 phrase pairs with human ratings. We used the language model, translation, and glove scores as features, and additionally used the inverse phrase translation probability $\varphi(p'|p)$, lexical weighting $lex(p|p')$, and $lex(p'|p)$ from Moses.

Figure 3 compares the different ranking methods against the human judgments on 200 phrase pairs randomly sampled from GIZA++.

5.2 Paraphrase Quality Evaluation

We compared the quality of paraphrases extracted by our method with the closest previous work (BUCC-2013) (Xu et al., 2013), in which a similar phrase table was created using Moses from monolingual parallel tweets that contain the same named entity and calendar date. We randomly sampled 500 phrase pairs from each phrase table and collected human judgements on a 5-point Likert scale, as described in Callison-Burch (Callison-Burch, 2008). Table 11 shows the evaluation results. We focused on the highest-quality paraphrases that rated as 5 ("all of the meaning of the original phrase is retained, and nothing is added") and their presence among all extracted paraphrases sorted by ranking scores.

We were also interested in how these phrasal paraphrases compared with those in PPDB. We

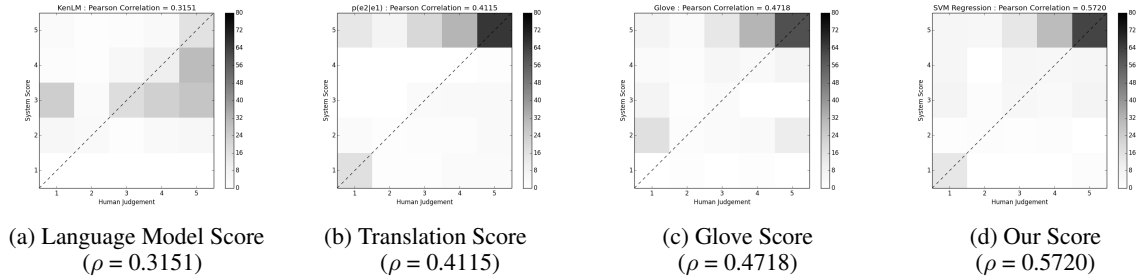| (a) Language Model Score | (b) Translation Score | (c) Glove Score | (d) Our Score |
| $(\rho = 0.3151)$ | $(\rho = 0.4115)$ | $(\rho = 0.4718)$ | $(\rho = 0.5720)$ |

Figure 3: Correlation between automatic scores (vertical axis) and 5-point human scores (horizontal axis) for ranking phrasal paraphrases. The darker squares along the diagonal line indicate a higher ranking.

sampled an equal amount of 420 paraphrase pairs from our phrase tables and PPDB, and then checked what percentage out of the total 840 could be found in our phrase tables and PPDB, respectively. As shown in Table 12, there is little overlap between URL data and PPDB, only 1.3% (51.3-50%) plus 0.8% (50.8-50%). Our Twitter URL data complements well with the existing paraphrase resources, such as PPDB, which are primarily derived from well-edited texts.

Top Rankings	BUCC 2013	GIZA++	Jacana	Sultan
10%	76.0	85.5	90.0	90.0
20%	65.6	86.5	91.0	91.0
30%	62.7	79.2	86.0	88.0
40%	56.6	73.2	85.5	84.5
50%	52.1	68.1	83.4	84.8
100% (all)	36.3	49.8	75.8	77.2

Table 11: Percentage of high-quality phrasal paraphrases extracted from Twitter URL data (this work) by GIZA++, Jacana, Sultan aligners, comparing to the previous work (BUCC-2013).

	PPDB	URL	GIZA++	Jacana	Sultan
Sample Size	50%	50%	16.7%	16.7%	16.7%
Coverage	51.3%	50.8%	18.7%	32.1%	34.4%

Table 12: Coverage comparison of phrasal paraphrases extracted from Twitter URL data (sampled 1:1:1 from GIZA++, Jacana and Sultan's aligner outputs) and the PPDB (Ganitkevitch et al., 2013).

6 Related Work

Sentential Paraphrase Data Researchers have found several data sources from which to collect sentential paraphrases: multiple news agencies reporting the same event (MSRP) (Dolan et al., 2004; Dolan and Brockett, 2005), multiple translated versions of a foreign novel (Barzilay and Elhadad, 2003; Barzilay and Lee, 2003) or other

texts (Cohn et al., 2008), multiple definitions of the same concept (Hashimoto et al., 2011), descriptions of the same video clip from multiple workers (Chen and Dolan, 2011) or rephrased sentences (Burrows et al., 2013; Toutanova et al., 2016). However, all these data collection methods are incapable of obtaining sentential paraphrases on a large scale (i.e. limited number of news agencies or books with multiple translated versions), and/or lack meaningful negative examples. Both of these properties are crucial for developing machine learning models that identify paraphrases and measure semantic similarities.

Non-sentential Paraphrase Data There are other phrasal and syntactic paraphrase data, such as DIRT (Lin and Pantel, 2001), POLY (Grycner et al., 2016), PATTY (Nakashole et al., 2012), DE-FIE (Bovi et al., 2015), and PPDB (Ganitkevitch et al., 2013; Pavlick et al., 2015). Most of these works focus on news or web data. Other earlier works on Twitter paraphrase extraction used unsupervised approaches (Xu et al., 2013; Wang et al., 2013) or small datasets (Zanzotto et al., 2011; Antoniak et al., 2015).

7 Conclusion and Future Work

In this paper, we show how a simple method can effectively and continuously collect large-scale sentential paraphrases from Twitter. We rigorously evaluated our data with automatic identification classification models and various measurements. We will share our new dataset with the research community; this dataset includes 51,524 sentence pairs manually labeled and a monthly growth of 30,000 sentential paraphrases automatically labeled. Future work could include expanding into many different languages present in social media and developing language-independent automatic paraphrase identification models.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval)*.

Maria Antoniak, Eric Bell, and Fei Xia. 2015. Leveraging paraphrase labels to extract synonyms from twitter. In *Proceedings of the 28th Florida Artificial Intelligence Research Society Conference (FLAIRS)*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4):555–596.

Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics* 39(3).

Johannes Bjerva, Johan Bos, Rob Van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval)*.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics (TACL)* pages 529–543.

Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4(3):43.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics* 34(4):597–614.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Proceedings of the First international conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment (MLCW)*.

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 Asian Federation of Natural Language Processing (ACL-IJCNLP)*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP)*.

William B. 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)*.

Anthony Fader, Luke S 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 (ACL)*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Adam Grycner, Saarland Informatics Campus, and Gerhard Weikum. 2016. POLY: Mining relational paraphrases from multilingual sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Weiwei Guo, Wei Liu, and Mona Diab. 2014. Fast tweet retrieval with compact binary codes. In *Proceedings of the 30th International Conference on Computational Linguistics (COLING)*.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi. 2011. Extracting paraphrases from definition sentences on the web. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*.

Hamid Izadinia, Fereshteh Sadeghi, Santosh K Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2015. Segment-phrase table for semantic segmentation, visual entailment and paraphrasing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

P. Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist* 11(2):37–50.

Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*.

Dekang Lin and Patrick Pantel. 2001. DIRT – Discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

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

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. PATTY: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, and Chris Callison-Burch Ben Van Durme. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Satoshi Sekine. 2006. On-demand information extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics (TACL)* 2:219–230.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and

short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Soroush Vosoughi and Deb Roy. 2016. A semi-automatic method for efficient detection of stories on social media. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*.

Ling Wang, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2013. Paraphrasing 4 microblog normalization. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

J. Wieting, M. Bansal, K. Gimpel, K. Livescu, and D. Roth. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics (TACL)* 3:345–358.

Wei Xu, Chris Callison-Burch, and William B. Dolan. 2015. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics (TACL)* 2:435–448.

Wei Xu, Alan Ritter, and Ralph Grishman. 2013. Gathering and generating paraphrases from twitter with application to normalization. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora (BUCC)*.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Kostas Tsioutsiouliklis. 2011. Linguistic redundancy in Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Congle Zhang, Stephen Soderland, and Daniel S Weld. 2015. Exploiting parallel news streams for unsupervised event extraction. *Transactions of the Association for Computational Linguistics (TACL)* 3:117–129.