Gathering and Generating Paraphrases from Twitter with Application to Normalization

Wei Xu[+] Alan Ritter[ˆ] Ralph Grishman[+]
[+]New York University, New York, NY, USA
{xuwei, grishman}@cs.nyu.edu
[ˆ]University of Washington, Seattle, WA, USA
aritter@cs.washington.edu

Abstract

We present a new and unique paraphrase resource, which contains meaning-preserving transformations between informal user-generated text. Sentential paraphrases are extracted from a comparable corpus of temporally and topically related messages on Twitter which often express semantically identical information through distinct surface forms. We demonstrate the utility of this new resource on the task of paraphrasing and normalizing noisy text, showing improvement over several state-of-the-art paraphrase and normalization systems [1].

1 Introduction

Social media services provide a massive amount of valuable information and demand NLP tools specifically developed to accommodate their noisy style. So far not much success has been reported on a key NLP technology on social media data: paraphrasing. Paraphrases are alternative ways to express the same meaning in the same language and commonly employed to improve the performance of many other NLP applications (Madnani and Dorr, 2010). In the case of Twitter, Petrović et al. (2012) showed improvements on first story detection by using paraphrases extracted from WordNet.

Learning paraphrases from tweets could be especially beneficial. First, the high level of information redundancy in Twitter provides a good opportunity to collect many different expressions. Second, tweets contain many kinds of paraphrases not available elsewhere including typos, abbreviations, ungrammatical expressions and slang, which can be particularly valuable for many applications, such as phrase-based text normalization (Kaufmann and Kalita, 2010) and correction of writing mistakes (Gamon et al., 2008), given the difficulty of acquiring annotated data. Paraphrase models that are derived from microblog data could be useful to improve other NLP tasks on noisy user-generated text and help users to interpret a large range of up-to-date abbreviations (e.g. dlt → Doritos Locos Taco) and native expressions (e.g. oh my god → {oh my goodness | oh my gosh | oh my gawd | oh my jesus}) etc.

This paper presents the first investigation into automatically collecting a large paraphrase corpus of tweets, which can be used for building paraphrase systems adapted to Twitter using techniques from statistical machine translation (SMT). We show experimental results demonstrating the benefits of an in-domain parallel corpus when paraphrasing tweets. In addition, our paraphrase models can be applied to the task of normalizing noisy text where we show improvements over the state-of-the-art.

Relevant previous work has extracted sentence-level paraphrases from news corpora (Dolan et al., 2004; Barzilay and Lee, 2003; Quirk et al., 2004). Paraphrases gathered from noisy user-generated text on Twitter have unique characteristics which make this comparable corpus a valuable new resource for mining sentence-level paraphrases. Twitter also has much less context than news articles and much more diverse content, thus posing new challenges to control the noise in mining paraphrases while retaining the desired superficial dissimilarity.

2 Related Work

There are several key strands of related work, including previous work on gathering parallel monolingual text from topically clustered news articles, normalizing noisy Twitter text using word-based

[1] Our Twitter paraphrase models are available online at https://github.com/cocoxu/twitterparaphrase/

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 121–128,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

models, and applying out-of-domain paraphrase systems to improve NLP tasks in Twitter.

On the observation of the lack of a large paraphrase corpus, Chen and Dolan (2011) have resorted to crowdsourcing to collect paraphrases by asking multiple independent users for descriptions of the same short video. As we show in §5, however, this data is very different from Twitter, so paraphrase systems trained on in-domain Twitter paraphrases tend to perform much better.

The task of paraphrasing tweets is also related to previous work on normalizing noisy Twitter text (Han and Baldwin, 2011; Han et al., 2012; Liu et al., 2012). Most previous work on normalization has applied word-based models. While there are challenges in applying Twitter paraphrase systems to the task of normalization, access to parallel text allows us to make phrase-based transformations to the input string rather than relying on word-to-word mappings (for more details see §4).

Also relevant is recent work on collecting bilingual parallel data from Twitter (Jehl et al., 2012; Ling et al., 2013). In contrast, we focus on monolingual paraphrases rather than multilingual translations.

Finally we highlight recent work on applying out-of-domain paraphrase systems to improve performance at first story detection in Twitter (Petrović et al., 2012). By building better paraphrase models adapted to Twitter, it should be possible to improve performance at such tasks, which benefit from paraphrasing Tweets.

3 Gathering A Parallel Tweet Corpus

There is a huge amount of redundant information on Twitter. When significant events take place in the world, many people go to Twitter to share, comment and discuss them. Among tweets on the same topic, many will convey similar meaning using widely divergent expressions. Whereas researchers have exploited multiple news reports about the same event for paraphrase acquisition (Dolan et al., 2004), Twitter contains more variety in terms of both language forms and types of events, and requires different treatment due to its unique characteristics.

As described in §3.1, our approach first identifies tweets which refer to the same popular event as those which mention a unique named entity and date, then aligns tweets within each event to construct a parallel corpus. To generate paraphrases, we apply a typical phrase-based statistical MT pipeline, performing word alignment on the parallel data using GIZA++ (Och and Ney, 2003), then extracting phrase pairs and performing decoding uses Moses (Koehn et al., 2007).

3.1 Extracting Events from Tweets

As a first step towards extracting paraphrases from popular events discussed on Twitter, we need a way to identify Tweets which mention the same event. To do this we follow previous work by Ritter et al. (2012), extracting named entities and resolving temporal expressions (for example "tomorrow" or "on Wednesday"). Because tweets are compact and self-contained, those which mention the same named entity and date are likely to reference the same event. We also employ a statistical significance test to measure strength of association between each named entity and date, and thereby identify important events discussed widely among users with a specific focus, such as the release of a new iPhone as opposed to individual users discussing everyday events involving their phones. By gathering tweets based on popular real-world events, we can efficiently extract pairwise paraphrases within a small group of closely related tweets, rather than exploring every pair of tweets in a large corpus. By discarding frequent but insignificant events, such as "I like my iPhone" and "I like broke my iPhone", we can reduce noise and encourage diversity of paraphrases by requiring less lexical overlap. Example events identified using this procedure are presented in Table 1.

3.2 Extracting Paraphrases Within Events

Twitter users are likely to express the same meaning in relation to an important event, however not every pair of tweets mentioning the same event will have the same meaning. People may have opposite opinions and complicated events such as presidential elections can have many aspects. To build a useful monolingual paraphrase corpus, we need some additional filtering to prevent unrelated sentence pairs.

If two tweets mention the same event and also share many words in common, they are very likely to be paraphrases. We use the Jaccard distance metric (Jaccard, 1912) to identify pairs of sentences within an event that are similar at the lexical level. Since tweets are extremely short with little context and include a broad range of topics, using only surface similarity is prone to unrelated sen-

Entity/Date	Example Tweets
Obama 11/6/2012	Vote for Obama on November 6th!
	OBAMA is #winning his 2nd term on November 6th 2012.
	November 6th we will re-elect Obama!!
James Bond 11/9/2012	Bought movie tickets to see James Bond tomorrow. I'm a big #007 fan!
	Who wants to go with me and see that new James Bond movie tomorrow?
	I wanna go see James Bond tomorrow
North Korea 12/29/2012	North Korea Announces December 29 Launch Date for Rocket
	Pyongyang reschedules launch to December 29 due to 'technical deficiency'
	North Korea to extend rocket launch period to December 29

Table 1: Example sentences taken from automatically identified significant events extracted from Twitter. Because many users express similar information when mentioning these events, there are many opportunities for paraphrase.

tence pairs. The average sentence length is only 11.9 words in our Twitter corpus, compared to 18.6 words in newswire (Dolan et al., 2004) which also contains additional document-level information. Even after filtering tweets with both their event cluster and lexical overlap, some unrelated sentence pairs remain in the parallel corpus. For example, names of two separate music venues in the same city might be mismatched together if they happen to have concerts on the same night that people tweeted using a canonical phrasing like "I am going to a concert at ______ in Austin tonight".

4 Paraphrasing Tweets for Normalization

Paraphrase models built from grammatical text are not appropriate for the task of normalizing noisy text. However, the unique characteristics of the Twitter data allow our paraphrase models to include both normal and noisy language and consequently translate between them. Our models have a tendency to normalize because correct spellings and grammar are most frequently used,[2] but there is still danger of introducing noise. For the purposes of normalization, we therefore biased our models using a language model built using text taken from the New York Times which is used to represent grammatical English.

Previous work on microblog normalization is mostly limited to word-level adaptation or out-of-domain annotated data. Our phrase-based models fill the gap left by previous studies by exploiting a large, automatically curated, in-domain paraphrase corpus.

Lexical normalization (Han and Baldwin, 2011) only considers transforming an out-of-vocabulary (OOV) word to its standard form, i.e. in-vocabulary (IV) word. Beyond word-to-word conversions, our phrase-based model is also able to handle the following types of errors without requiring any annotated data:

Error type	Ill form	Standard form
1-to-many	everytime	every time
incorrect IVs	can't want for	can't wait for
grammar	I'm going a movie	I'm going to a movie
ambiguities	4	4 / 4th / for / four

Kaufmann and Kalita (2010) explored machine translation techniques for the normalization task using an SMS corpus which was manually annotated with grammatical paraphrases. Microblogs, however, contain a much broader range of content than SMS and have no in-domain annotated data available. In addition, the ability to gather paraphrases *automatically* opens up the possibility to build normalization models from orders of magnitude more data, and also to produce up-to-date normalization models which capture new abbreviations and slang as they are invented.

5 Experiments

We evaluate our system and several baselines at the task of paraphrasing Tweets using previously developed automatic evaluation metrics which have been shown to have high correlation with human judgments (Chen and Dolan, 2011).

[2] Even though misspellings and grammatical errors are quite common, there is much more variety and less agreement.

In addition, because no previous work has evaluated these metrics in the context of noisy Twitter data, we perform a human evaluation in which annotators are asked to choose which system generates the best paraphrase. Finally we evaluate our phrase-based normalization system against a state-of-the-art word-based normalizer developed for Twitter (Han et al., 2012).

5.1 Paraphrasing Tweets

5.1.1 Data

Our paraphrase dataset is distilled from a large corpus of tweets gathered over a one-year period spanning November 2011 to October 2012 using the Twitter Streaming API. Following Ritter et al. (2012), we grouped together all tweets which mention the same named entity (recognized using a Twitter specific name entity tagger[3]) and a reference to the same unique calendar date (resolved using a temporal expression processor (Mani and Wilson, 2000)). Then we applied a statistical significance test (the G test) to rank the events, which considers the corpus frequency of the named entity, the number of times the date has been mentioned, and the number of tweets which mention both together. Altogether we collected more than 3 million tweets from the 50 top events of each day according to the p-value from the statistical test, with an average of 229 tweets per event cluster.

Each of these tweets was passed through a Twitter tokenizer[4] and a simple sentence splitter, which also removes emoticons, URLs and most of the hashtags and usernames. Hashtags and usernames that were in the middle of sentences and might be part of the text were kept. Within each event cluster, redundant and short sentences (less than 3 words) were filtered out, and the remaining sentences were paired together if their Jaccard similarity was no less than 0.5. This resulted in a parallel corpus consisting of 4,008,946 sentence pairs with 800,728 unique sentences.

We then trained paraphrase models by applying a typical phrase-based statistical MT pipeline on the parallel data, which uses GIZA++ for word alignment and Moses for extracting phrase pairs, training and decoding. We use a language model trained on the 3 million collected tweets in the decoding process. The parameters are tuned over development data and the exact configuration are released together with the phrase table for system replication.

Sentence alignment in comparable corpora is more difficult than between direct translations (Moore, 2002), and Twitter's noisy style, short context and broad range of content present additional complications. Our automatically constructed parallel corpus contains some proportion of unrelated sentence pairs and therefore does result in some unreasonable paraphrases. We prune out unlikely phrase pairs using a technique proposed by Johnson et al. (2007) with their recommended setting, which is based on the significance testing of phrase pair co-occurrence in the parallel corpus (Moore, 2004). We further prevent unreasonable translations by adding additional entries to the phrase table to ensure every phrase has an option to remain unchanged during paraphrasing and normalization. Without these noise reduction steps, our system will produce paraphrases with serious errors (e.g. change a person's last name) for 100 out of 200 test tweets in the evaluation in §5.1.5.

At the same time, it is also important to promote lexical dissimilarity in the paraphrase task. Following Ritter et. al. (2011) we add a lexical similarity penalty to each phrase pair in our system, in addition to the four basic components (translation model, distortion model, language model and word penalty) in SMT.

5.1.2 Evaluation Details

The beauty of lexical similarity penalty is that it gives control over the degree of paraphrasing by adjusting its weight versus the other components. Thus we can plot a BLEU-PINC curve to express the tradeoff between semantic adequacy and lexical dissimilarity with the input, where BLUE (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) are previously proposed automatic evaluation metrics to measure respectively the two criteria of paraphrase quality.

To compute these automatic evaluation metrics, we manually prepared a dataset of gold paraphrases by tracking the trending topics on Twitter[5] and gathering groups of paraphrases in November 2012. In total 20 sets of sentences were collected and each set contains 5 different sentences that express the same meaning. Each sentence is used

[3] https://github.com/aritter/twitter_nlp

[4] https://github.com/brendano/tweetmotif

[5] https://support.twitter.com/articles/101125-faqs-about-twitter-s-trends

Input	Output
Hostess is going outta biz	hostess is going out of business
REPUBLICAN IMMIGRATION REFORM IS A THING NOW	gop imigration law is a thing now
Freedom Writers will always be one of my fav movies	freedom writers will forever be one of my favorite movies
sources confirm that Phil Jackson has cancelled all weekend plans and upcoming guest appearances, will meet with LAL front office	source confirms that phil jackson has canceled all weekend plans , upcomin guest appearances and will meet with lakers front office

Table 2: Example paraphrases generated by our system on the test data.

once as input while other 4 sentences in the same set serve as reference translation for automatic evaluation of semantic adequacy using BLEU.

5.1.3 Baselines

We consider two state-of-the-art paraphrase systems as baselines, both of which are trained on parallel corpora of aligned sentences. The first one is trained on a large-scale corpus gathered by asking users of Amazon's Mechanical Turk Service (Snow et al., 2008) to write a one-sentence description of a short video clip (Chen and Dolan, 2011). We combined a phrase table and distortion table extracted from this parallel corpus with the same Twitter language model, applying the Moses decoder to generate paraphrases. The additional noise removal steps described in §5.1.1 were found helpful for this model during development and were therefore applied. The second baseline uses the Microsoft Research paraphrase tables that are automatically extracted from news articles in combination with the Twitter language model.[6]

5.1.4 Results

Figure 1 compares our system against both baselines, varying the lexical similarity penalty for each system to generate BLEU-PINC curves. Our system trained on automatically gathered in-domain Twitter paraphrases achieves higher BLEU at equivalent PINC for the entire length of the curves. Table 2 shows some sample outputs of our system on real Twitter data.

One novel feature of our approach, compared to previous work on paraphrasing, is that it captures many slang terms, acronyms, abbreviations and misspellings that are otherwise hard to learn. Several examples are shown in table 3. The rich semantic redundancy in Twitter helps generate a large variety of typical paraphrases as well (see an example in table 4).

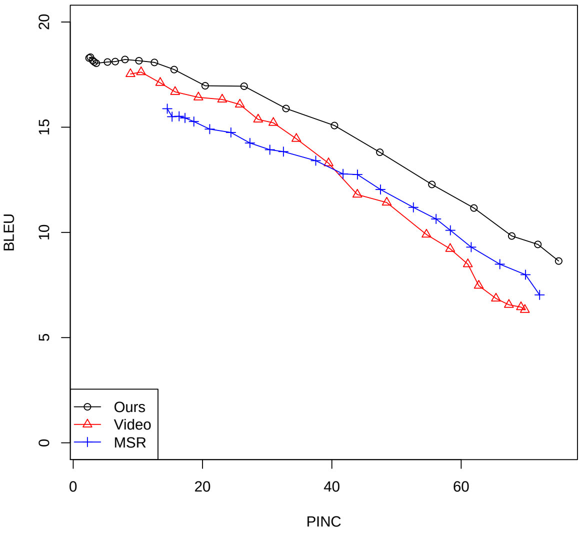

Figure 1: Results from automatic paraphrase evaluation. PINC measures n-gram dissimilarity from the source sentence, whereas BLEU roughly measures n-gram similarity to the reference paraphrases.

5.1.5 Human Evaluation

In addition to automatic evaluation, we also performed a human evaluation in which annotators were asked to pick which system generated the best paraphrase. We used the same dataset of 200 tweets gathered for the automatic evaluation and generated paraphrases using the 3 systems in Figure 1 with the highest BLEU which achieve a PINC of at least 40. The human annotators were then asked to pick which of the 3 systems generated the best paraphrase using the criteria that it should be both different from the original and also

[6] No distortion table or noisy removal process is applied because the parallel corpus is not available.

Input	Top-ranked Outputs
amped	pumped
lemme kno	let me know
bb	bigbang, big brother
snl	nbcsnl, saturday night live
apply 4 tix	apply for tickets, ask for tickets, applying for tickets
the boys	one direction (a band, whose members are often referred as "the boys"), they, the boy, the gys, the lads, my boys, the direction (can be used to refer to the band "one direction"), the onedirection, our boys, our guys
oh my god	oh my gosh, omfg, thank the lord, omg, oh my lord, thank you god, oh my jesus, oh god
can't wait	cant wait, cant wait, cannot wait, i cannot wait, so excited, cnt wait, i have to wait, i can'wait, ready, so ready, so pumped, seriously can'wait, really can't wait

Table 3: Example paraphrases of noisy phrases and slang commonly found on Twitter

Input	Top-ranked Outputs
who want to get a beer	wants to get a beer, so who wants to get a beer, who wants to go get a beer, who wants to get beer, who want to get a beer, trying to get a beer, who wants to buy a beer, who wants to get a drink, who wants to get a rootbeer, who trying to get a beer, who wants to have a beer, who wants to order a beer, i want to get a beer, who wants to get me a beer, who else wants to get a beer, who wants to win a beer, anyone wants to get a beer, who wanted to get a beer, who wants to a beer, someone to get a beer, who wants to receive a beer, someone wants to get a beer

Table 4: Example paraphrases of a given sentence "who want to get a beer"

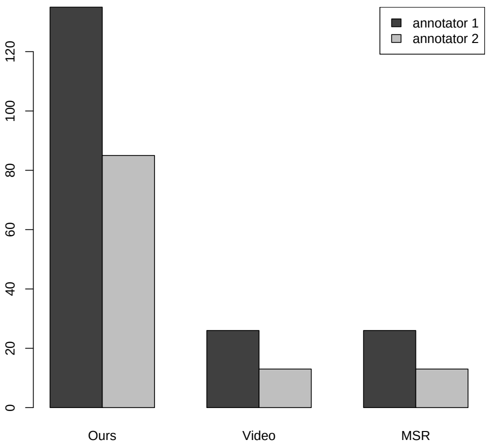

Figure 2: Number of paraphrases (200 in total) preferred by the annotators for each system

capture as much of the original meaning as possible. The annotators were asked to abstain from picking one as the best in cases where there were no changes to the input, or where the resulting paraphrases totally lost the meaning.

Figure 2 displays the number of times each annotator picked each system's output as the best. Annotator 2 was somewhat more conservative than annotator 1, choosing to abstain more frequently and leading to lower overall frequencies, however in both cases we see a clear advantage from paraphrasing using in-domain models. As a measure of inter-rater agreement, we computed Cohen's Kappa between the annotators judgment as to whether the Twitter-trained system's output best. The value of Cohen's Kappa in this case was 0.525.

5.2 Phrase-Based Normalization

Because Twitter contains both normal and noisy language, with appropriate tuning, our models have the capability to translate between these two styles, e.g. paraphrasing into noisy style or normalizing into standard language. Here we demonstrate its capability to normalize tweets at the sentence-level.

5.2.1 Baselines

Much effort has been devoted recently for developing normalization dictionaries for Microblogs. One of the most competitive dictionaries available today is HB-dict+GHM-dict+S-dict used by Han et al. (2012), which combines a manually-constructed Internet slang dictionary , a small (Gouws et al., 2011) and a large automatically-derived dictionary based on distributional and string similarity. We evaluate two baselines using this large dictionary consisting of 41181 words; following Han et. al. (2012), one is a simple dictionary look up. The other baseline uses the machinery of statistical machine translation using this dictionary as a phrase table in combination with Twitter and NYT language models.

5.2.2 System Details

Our base normalization system is the same as the paraphrase model described in §5.1.1, except that the distortion model is turned off to exclude reordering. We tuned the system towards correct spelling and grammar by adding a language model built from all New York Times articles written in 2008. We also filtered out the phrase pairs which map from in-vocabulary to out-of-vocabulary words. In addition, we integrated the dictionaries by linear combination to increase the coverage of phrase-based SMT model (Bisazza et al., 2011).

5.2.3 Evaluation Details

We adopt the normalization dataset of Han and Baldwin (2011), which was initially annotated for the token-level normalization task, and which we augmented with sentence-level annotations. It contains 549 English messages sampled from Twitter API from August to October, 2010.

5.2.4 Results

Normalization results are presented in figure 5. Using only our phrase table extracted from Twitter events we achieve poorer performance than the state-of-the-art dictionary baseline, however we find that by combining the normalization dictionary of Han et. al. (2012) with our automatically constructed phrase-table we are able to combine the high coverage of the normalization dictionary with the ability to perform phrase-level normalizations (e.g. “outta” → “out of” and examples in §4) achieving both higher PINC and BLEU than the systems which rely exclusively on word-level mappings. Our phrase table also contains many words that are not covered by the dictionary (e.g. “pts” → “points”, “noms” → “nominations”).

	BLEU	PINC
No-Change	60.00	0.0
SMT+TwitterLM	62.54	5.78
SMT+TwitterNYTLM	65.72	9.23
Dictionary	75.07	22.10
Dicionary+TwitterNYTLM	75.12	20.26
SMT+Dictionary+TwitterNYTLM	77.44	25.33

Table 5: Normalization performance

6 Conclusions

We have presented the first approach to gathering parallel monolingual text from Twitter, and built the first in-domain models for paraphrasing tweets. By paraphrasing using models trained on in-domain data we showed significant performance improvements over state-of-the-art out-of-domain paraphrase systems as demonstrated through automatic and human evaluations. We showed that because tweets include both normal and noisy language, paraphrase systems built from Twitter can be fruitfully applied to the task of normalizing noisy text, covering phrase-based normalizations not handled by previous dictionary-based normalization systems. We also make our Twitter-tuned paraphrase models publicly available. For future work, we consider developing additional methods to improve the accuracy of tweet clustering and paraphrase pair selection.

Acknowledgments

This research was supported in part by NSF grant IIS-0803481, ONR grant N00014-08-1-0431, and DARPA contract FA8750- 09-C-0179.

References

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 - Volume 1*, NAACL ’03.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus interpolation methods for phrase-based smt adaptation. In *International Workshop on Spoken Language Translation (IWSLT), San Francisco, CA.*

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-2011)*, Portland, OR, June.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of Coling 2004.*

Michael Gamon, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. 2008. Using contextual speller techniques and language modeling for esl error correction. *IJCNLP*.

S. Gouws, D. Hovy, and D. Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90. Association for Computational Linguistics.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 368–378.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432, Stroudsburg, PA, USA.

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

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421. Association for Computational Linguistics.

J.H. Johnson, J. Martin, G. Foster, and R. Kuhn. 2007. Improving translation quality by discarding most of the phrasetable.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International Conference on Natural Language Processing, Kharagpur, India*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broadcoverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Republic of Korea*.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.*

Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 69–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 333–340.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and twitter.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP 2004*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *KDD*, pages 1104–1112. ACM.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08.