Collecting Highly Parallel Data for Paraphrase Evaluation

David L. Chen
Department of Computer Science
The University of Texas at Austin
Austin, TX 78712, USA
dlcc@cs.utexas.edu

William B. Dolan
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
billdol@microsoft.com

Abstract

A lack of standard datasets and evaluation metrics has prevented the field of *paraphrasing* from making the kind of rapid progress enjoyed by the machine translation community over the last 15 years. We address both problems by presenting a novel data collection framework that produces highly parallel text data relatively inexpensively and on a large scale. The highly parallel nature of this data allows us to use simple n-gram comparisons to measure both the semantic adequacy and lexical dissimilarity of paraphrase candidates. In addition to being simple and efficient to compute, experiments show that these metrics correlate highly with human judgments.

1 Introduction

Machine paraphrasing has many applications for natural language processing tasks, including machine translation (MT), MT evaluation, summary evaluation, question answering, and natural language generation. However, a lack of standard datasets and automatic evaluation metrics has impeded progress in the field. Without these resources, researchers have resorted to developing their own small, ad hoc datasets (Barzilay and McKeown, 2001; Shinyama et al., 2002; Barzilay and Lee, 2003; Quirk et al., 2004; Dolan et al., 2004), and have often relied on human judgments to evaluate their results (Barzilay and McKeown, 2001; Ibrahim et al., 2003; Bannard and Callison-Burch, 2005). Consequently, it is difficult to compare different systems and assess the progress of the field as a whole.

Despite the similarities between paraphrasing and translation, several major differences have prevented researchers from simply following standards that have been established for machine translation. Professional translators produce large volumes of bilingual data according to a more or less consistent specification, indirectly fueling work on machine translation algorithms. In contrast, there are no "professional paraphrasers", with the result that there are no readily available large corpora and no consistent standards for what constitutes a high-quality paraphrase. In addition to the lack of standard datasets for training and testing, there are also no standard metrics like BLEU (Papineni et al., 2002) for evaluating paraphrase systems. Paraphrase evaluation is inherently difficult because the range of potential paraphrases for a given input is both large and unpredictable; in addition to being meaning-preserving, an ideal paraphrase must also diverge as sharply as possible in form from the original while still sounding natural and fluent.

Our work introduces two novel contributions which combine to address the challenges posed by paraphrase evaluation. First, we describe a framework for easily and inexpensively crowdsourcing arbitrarily large training and test sets of independent, redundant linguistic descriptions of the same semantic content. Second, we define a new evaluation metric, PINC (Paraphrase In N-gram Changes), that relies on simple BLEU-like n-gram comparisons to measure the degree of novelty of automatically generated paraphrases. We believe that this metric, along with the sentence-level paraphrases provided by our data collection approach, will make it possi-

ble for researchers working on paraphrasing to compare system performance and exploit the kind of automated, rapid training-test cycle that has driven work on Statistical Machine Translation.

In addition to describing a mechanism for collecting large-scale sentence-level paraphrases, we are also making available to the research community 85K parallel English sentences as part of the Microsoft Research Video Description Corpus [1].

The rest of the paper is organized as follows. We first review relevant work in Section 2. Section 3 then describes our data collection framework and the resulting data. Section 4 discusses automatic evaluations of paraphrases and introduces the novel metric PINC. Section 5 presents experimental results establishing a correlation between our automatic metric and human judgments. Sections 6 and 7 discuss possible directions for future research and conclude.

2 Related Work

Since paraphrase data are not readily available, various methods have been used to extract parallel text from other sources. One popular approach exploits multiple translations of the same data (Barzilay and McKeown, 2001; Pang et al., 2003). Examples of this kind of data include the Multiple-Translation Chinese (MTC) Corpus [2] which consists of Chinese news stories translated into English by 11 translation agencies, and literary works with multiple translations into English (e.g. Flaubert's *Madame Bovary*.) Another method for collecting monolingual paraphrase data involves aligning semantically parallel sentences from different news articles describing the same event (Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004).

While utilizing multiple translations of literary work or multiple news stories of the same event can yield significant numbers of parallel sentences, this data tend to be noisy, and reliably identifying good paraphrases among all possible sentence pairs remains an open problem. On the other hand, multiple translations on the sentence level such as the MTC Corpus provide good, natural paraphrases, but rela-

tively little data of this type exists. Finally, some approaches avoid the need for monolingual paraphrase data altogether by using a second language as the pivot language (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Kok and Brockett, 2010). Phrases that are aligned to the same phrase in the pivot language are treated as potential paraphrases. One limitation of this approach is that only words and phrases are identified, not whole sentences.

While most work on evaluating paraphrase systems has relied on human judges (Barzilay and McKeown, 2001; Ibrahim et al., 2003; Bannard and Callison-Burch, 2005) or indirect, task-based methods (Lin and Pantel, 2001; Callison-Burch et al., 2006), there have also been a few attempts at creating automatic metrics that can be more easily replicated and used to compare different systems. Para-Metric (Callison-Burch et al., 2008) compares the paraphrases discovered by an automatic system with ones annotated by humans, measuring precision and recall. This approach requires additional human annotations to identify the paraphrases within parallel texts (Cohn et al., 2008) and does not evaluate the systems at the sentence level. The more recently proposed metric PEM (Paraphrase Evaluation Metric) (Liu et al., 2010) produces a single score that captures the semantic adequacy, fluency, and lexical dissimilarity of candidate paraphrases, relying on bilingual data to learn semantic equivalences without using n-gram similarity between candidate and reference sentences. In addition, the metric was shown to correlate well with human judgments. However, a significant drawback of this approach is that PEM requires substantial in-domain bilingual data to train the semantic adequacy evaluator, as well as sample human judgments to train the overall metric.

We designed our data collection framework for use on crowdsourcing platforms such as Amazon's Mechanical Turk. Crowdsourcing can allow inexpensive and rapid data collection for various NLP tasks (Ambati and Vogel, 2010; Bloodgood and Callison-Burch, 2010a; Bloodgood and Callison-Burch, 2010b; Irvine and Klementiev, 2010), including human evaluations of NLP systems (Callison-Burch, 2009; Denkowski and Lavie, 2010; Zaidan and Callison-Burch, 2009). Of particular relevance are the paraphrasing work by Buzek et al. (2010)

---

[1]Available for download at `http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/`

[2]Linguistic Data Consortium (LDC) Catalog Number LDC2002T01, ISBN 1-58563-217-1.

and Denkowski et al. (2010). Buzek et al. automatically identified problem regions in a translation task and had workers attempt to paraphrase them, while Denkowski et al. asked workers to assess the validity of automatically extracted paraphrases. Our work is distinct from these earlier efforts both in terms of the task – attempting to collect linguistic descriptions using a visual stimulus – and the dramatically larger scale of the data collected.

3 Data Collection

Since our goal was to collect large numbers of paraphrases quickly and inexpensively using a crowd, our framework was designed to make the tasks short, simple, easy, accessible and somewhat fun. For each task, we asked the annotators to watch a very short video clip (usually less than 10 seconds long) and describe in one sentence the main action or event that occurred in the video clip

We deployed the task on Amazon's Mechanical Turk, with video segments selected from YouTube. A screenshot of our annotation task is shown in Figure 1. On average, annotators completed each task within 80 seconds, including the time required to watch the video. Experienced annotators were even faster, completing the task in only 20 to 25 seconds.

One interesting aspect of this framework is that each annotator approaches the task from a linguistically independent perspective, unbiased by the lexical or word order choices in a pre-existing description. The data thus has some similarities to parallel news descriptions of the same event, while avoiding much of the noise inherent in news. It is also similar in spirit to the 'Pear Stories' film used by Chafe (1997). Crucially, our approach allows us to gather arbitrarily many of these independent descriptions for each video, capturing nearly-exhaustive coverage of how native speakers are likely to summarize a small action. It might be possible to achieve similar effects using images or panels of images as the stimulus (von Ahn and Dabbish, 2004; Fei-Fei et al., 2007; Rashtchian et al., 2010), but we believed that videos would be more engaging and less ambiguous in their focus. In addition, videos have been shown to be more effective in prompting descriptions of motion and contact verbs, as well as verbs that are generally not imageable (Ma and Cook, 2009).

Figure 1: A screenshot of our annotation task as it was deployed on Mechanical Turk.

3.1 Quality Control

One of the main problems with collecting data using a crowd is quality control. While the cost is very low compared to traditional annotation methods, workers recruited over the Internet are often unqualified for the tasks or are incentivized to cheat in order to maximize their rewards.

To encourage native and fluent contributions, we asked annotators to write the descriptions in the language of their choice. The result was a significant amount of translation data, unique in its multilingual parallelism. While included in our data release, we leave aside a full discussion of this multilingual data for future work.

To ensure the quality of the annotations being produced, we used a two-tiered payment system. The idea was to reward workers who had shown the ability to write quality descriptions and the willingness to work on our tasks consistently. While everyone had access to the Tier-1 tasks, only workers who had been manually qualified could work on the Tier-2 tasks. The tasks were identical in the two tiers but each Tier-1 task only paid 1 cent while each Tier-2 task paid 5 cents, giving the workers a strong incentive to earn the qualification.

The qualification process was done manually by the authors. We periodically evaluated the workers who had submitted the most Tier-1 tasks (usually on the order of few hundred submissions) and granted them access to the Tier-2 tasks if they had performed well. We assessed their work mainly on the grammaticality and spelling accuracy of the submitted descriptions. Since we had hundreds of submissions to base our decisions on, it was fairly easy to identify the cheaters and people with poor English skills [3]. Workers who were rejected during this process were still allowed to work on the Tier-1 tasks.

While this approach requires significantly more manual effort initially than other approaches such as using a qualification test or automatic post-annotation filtering, it creates a much higher quality workforce. Moreover, the initial effort is amortized over time as these quality workers are retained over the entire duration of the data collection. Many of them annotated all the available videos we had.

3.2 Video Collection

To find suitable videos to annotate, we deployed a separate task. Workers were asked to submit short (generally 4-10 seconds) video segments depicting single, unambiguous events by specifying links to YouTube videos, along with the start and end times. We again used a tiered payment system to reward and retain workers who performed well.

Since the scope of this data collection effort extended beyond gathering English data alone, we

- Someone is coating a pork chop in a glass bowl of flour.
- A person breads a pork chop.
- Someone is breading a piece of meat with a white powdery substance.
- A chef seasons a slice of meat.
- Someone is putting flour on a piece of meat.
- A woman is adding flour to meat.
- A woman is coating a piece of pork with breadcrumbs.
- A man dredges meat in bread crumbs.
- A person breads a piece of meat.
- A woman is breading some meat.
- Someone is breading meat.
- A woman coats a meat cutlet in a dish.
- A woman is coating a pork loin in bread crumbs.
- The laldy coated the meat in bread crumbs.
- The woman is breading pork chop.
- A woman adds a mixture to some meat.
- The lady put the batter on the meat.

Figure 2: Examples of English descriptions collected for a particular video segment.

tried to collect videos that could be understood regardless of the annotator's linguistic or cultural background. In order to avoid biasing lexical choices in the descriptions, we muted the audio and excluded videos that contained either subtitles or overlaid text. Finally, we manually filtered the submitted videos to ensure that each met our criteria and was free of inappropriate content.

3.3 Data

We deployed our data collection framework on Mechanical Turk over a two-month period from July to September in 2010, collecting 2,089 video segments and 85,550 English descriptions. The rate of data collection accelerated as we built up our workforce, topping 10K descriptions a day when we ended our data collection. Of the descriptions, 33,855 were from Tier-2 tasks, meaning they were provided by workers who had been manually identified as good performers. Examples of some of the descriptions collected are shown in Figure 2.

Overall, 688 workers submitted at least one English description. Of these workers, 113 submitted at least 100 descriptions and 51 submitted at least 500. The largest number of descriptions submitted by a single worker was 3496 [4]. Out of the 688 workers, 50 were granted access to the Tier-2 tasks. The

---

[3] Everyone who submitted descriptions in a foreign language was granted access to the Tier-2 tasks. This was done to encourage more submissions in different languages and also because we could not verify the quality of those descriptions other than using online translation services (and some of the languages were not available to be translated).

---

[4] This number exceeds the total number of videos because the worker completed both Tier-1 and Tier-2 tasks for the same videos

	Tier 1	Tier 2
pay	$0.01	$0.05
# workers (English)	683	50
# workers (total)	835	94
# submitted (English)	51510	33829
# submitted (total)	68578	55682
# accepted (English)	51052	33825
# accepted (total)	67968	55658

Table 1: Statistics for the two video description tasks

success of our data collection effort was in part due to our ability to retain these good workers, building a reliable and efficient workforce. Table 1 shows some statistics for the Tier-1 and Tier-2 tasks [5]. Overall, we spent under $5,000 including Amazon's service fees, some pilot experiments and surveys.

On average, 41 descriptions were produced for each video, with at least 27 for over 95% of the videos. Even limiting the set to descriptions produced from the Tier-2 tasks, there are still 16 descriptions on average for each video, with at least 12 descriptions for over 95% of the videos. For most clusters, then, we have a dozen or more high-quality parallel descriptions that can be paired with one another to create monolingual parallel training data.

4 Paraphrase Evaluation Metrics

One of the limitations to the development of machine paraphrasing is the lack of standard metrics like BLEU, which has played a crucial role in driving progress in MT. Part of the issue is that a good paraphrase has the additional constraint that it should be lexically dissimilar to the source sentence while preserving the meaning. These can become competing goals when using n-gram overlaps to establish semantic equivalence. Thus, researchers have been unable to rely on BLEU or some derivative: the optimal paraphrasing engine under these terms would be one that simply returns the input.

To combat such problems, Liu et al. (2010) have proposed PEM, which uses a second language as pivot to establish semantic equivalence. Thus, no n-gram overlaps are required to determine the semantic adequacy of the paraphrase candidates. PEM

also separately measures lexical dissimilarity and fluency. Finally, all three scores are combined using a support vector machine (SVM) trained on human ratings of paraphrase pairs. While PEM was shown to correlate well with human judgments, it has some limitations. It only models paraphrasing at the phrase level and not at the sentence level. Further, while it does not need reference sentences for the evaluation dataset, PEM does require suitable bilingual data to train the metric. The result is that training a successful PEM becomes almost as challenging as the original paraphrasing problem, since paraphrases need to be learned from bilingual data.

The highly parallel nature of our data suggests a simpler solution to this problem. To measure semantic equivalence, we simply use BLEU with multiple references. The large number of reference paraphrases capture a wide space of sentences with equivalent meanings. While the set of reference sentences can of course never be exhaustive, our data collection method provides a natural distribution of common phrases that might be used to describe an action or event. A tight cluster with many similar parallel descriptions suggests there are only few common ways to express that concept.

In addition to measuring semantic adequacy and fluency using BLEU, we also need to measure lexical dissimilarity with the source sentence. We introduce a new scoring metric PINC that measures how many n-grams differ between the two sentences. In essence, it is the inverse of BLEU since we want to minimize the number of n-gram overlaps between the two sentences. Specifically, for source sentence s and candidate sentence c:

$$PINC(s,c) = \frac{1}{N}\sum_{n=1}^{N} \ 1 - \frac{|\text{ n-gram}_s \cap \text{n-gram}_c \ |}{|\text{ n-gram}_c \ |}$$

where N is the maximum n-gram considered and n-gram$_s$ and n-gram$_c$ are the lists of n-grams in the source and candidate sentences, respectively. We use $N = 4$ in our evaluations.

The PINC score computes the percentage of n-grams that appear in the candidate sentence but not in the source sentence. This score is similar to the Jaccard distance, except that it excludes n-grams that only appear in the source sentence and not in the candidate sentence. In other words, it rewards candi-

---

[5]The numbers for the English data are slightly underestimated since the workers sometimes incorrectly filled out the form when reporting what language they were using.

dates for introducing new n-grams but not for omitting n-grams from the original sentence. The results for each n are averaged arithmetically. PINC evaluates single sentences instead of entire documents because we can reliably measure lexical dissimilarity at the sentence level. Also notice that we do not put additional constraints on sentence length: while extremely short and extremely long sentences are likely to score high on PINC, they still must maintain semantic adequacy as measured by BLEU.

We use BLEU and PINC together as a 2-dimensional scoring metric. A good paraphrase, according to our evaluation metric, has few n-gram overlaps with the source sentence but many n-gram overlaps with the reference sentences. This is consistent with our requirement that a good paraphrase should be lexically dissimilar from the source sentence while preserving its semantics.

Unlike Liu et al. (2010), we treat these two criteria separately, since different applications might have different preferences for each. For example, a paraphrase suggestion tool for a word processing software might be more concerned with semantic adequacy, since presenting a paraphrase that does not preserve the meaning would likely result in a negative user experience. On the other hand, a query expansion algorithm might be less concerned with preserving the precise meaning so long as additional relevant terms are added to improve search recall.

5 Experiments

To verify the usefulness of our paraphrase corpus and the BLEU/PINC metric, we built and evaluated several paraphrase systems and compared the automatic scores to human ratings of the generated paraphrases. We also investigated the pros and cons of collecting paraphrases using video annotation rather than directly eliciting them.

5.1 Building paraphrase models

We built 4 paraphrase systems by training English to English translation models using Moses (Koehn et al., 2007) with the default settings. Using our paraphrase corpus to train and to test, we divided the sentence clusters associated with each video into 90% for training and 10% for testing. We restricted our attention to sentences produced from the Tier-2 tasks
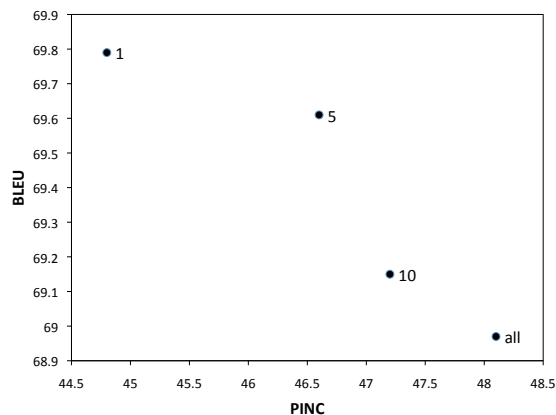

Figure 3: Evaluation of paraphrase systems trained on different numbers of parallel sentences. As more training pairs are used, the model produces more varied sentences (PINC) but preserves the meaning less well (BLEU)

in order to avoid excessive noise in the datasets, resulting in 28,785 training sentences and 3,367 test sentences. To construct the training examples, we randomly paired each sentence with 1, 5, 10, or all parallel descriptions of the same video segment. This corresponds to 28K, 143K, 287K, and 449K training pairs respectively. For the test set, we used each sentence once as the source sentence with all parallel descriptions as references (there were 16 references on average, with a minimum of 10 and a maximum of 31.) We also included the source sentence as a reference for itself.

Overall, all the trained models produce reasonable paraphrase systems, even the model trained on just 28K single parallel sentences. Examples of the outputs produced by the models trained on single parallel sentences and on all parallel sentences are shown in Table 2. Some of the changes are simple word substitutions, e.g. *rabbit* for *bunny* or *gun* for *revolver*, while others are phrasal, e.g. *frying meat* for *browning pork* or *made a basket* for *scores in a basketball game*. One interesting result of using videos as the stimulus to collect training data is that sometimes the learned paraphrases are not based on linguistic closeness, but rather on visual similarity, e.g. substituting *cricket* for *baseball*.

To evaluate the results quantitatively, we used the BLEU/PINC metric. The performance of all the trained models is shown in Figure 3. Unsurprisingly, there is a tradeoff between preserving the meaning

Original sentence	Trained on 1 parallel sentence	Trained on all parallel sentences
a bunny is cleaning its paw	a rabbit is licking its paw	a rabbit is cleaning itself
a man fires a revolver	a man is shooting targets	a man is shooting a gun
a big turtle is walking	a huge turtle is walking	a large tortoise is walking
a guy is doing a flip over a park bench	a man does a flip over a bench	a man is doing stunts on a bench
milk is being poured into a mixer	a man is pouring milk into a mixer	a man is pouring milk into a bowl
children are practicing baseball	children are doing a cricket	children are playing cricket
a boy is doing karate	a man is doing karate	a boy is doing martial arts
a woman is browning pork in a pan	a woman is browning pork in a pan	a woman is frying meat in a pan
a player scores in a basketball game	a player made a basketball game	a player made a basket

Table 2: Examples of paraphrases generated by the trained models.

and producing more varied paraphrases. Systems trained on fewer parallel sentences are more conservative and make fewer mistakes. On the other hand, systems trained on more parallel sentences often produce very good paraphrases but are also more likely to diverge from the original meaning. As a comparison, evaluating each human description as a paraphrase for the other descriptions in the same cluster resulted in a BLEU score of 52.9 and a PINC score of 77.2. Thus, all the systems performed very well in terms of retaining semantic content, although not as well in producing novel sentences.

To validate the results suggested by the automatic metrics, we asked two fluent English speakers to rate the generated paraphrases on the following categories: semantic, dissimilarity, and overall. Semantic measures how well the paraphrase preserves the original meaning while dissimilarity measures how much the paraphrase differs from the source sentence. Each category is rated from 1 to 4, with 4 being the best. A paraphrase identical to the source sentence would receive a score of 4 for meaning and 1 for dissimilarity and overall. We randomly selected 200 source sentences and generated 2 paraphrases for each, representing the two extremes: one paraphrase produced by the model trained with single parallel sentences, and the other by the model trained with all parallel sentences. The average scores of the two human judges are shown in Table 3. The results confirm our finding that the system trained with single parallel sentences preserves the meaning better but is also more conservative.

5.2 Correlation with human judgments

Having established rough correspondences between BLEU/PINC scores and human judgments of se-

	Semantic	Dissimilarity	Overall
1	3.09	2.65	2.51
All	2.91	2.89	2.43

Table 3: Average human ratings of the systems trained on single parallel sentences and on all parallel sentences.

mantic equivalence and lexical dissimilarity, we quantified the correlation between these automatic metrics and human ratings using Pearson's correlation coefficient, a measure of linear dependence between two random variables. We computed the inter-annotator agreement as well as the correlation between BLEU, PINC, PEM (Liu et al., 2010) and the average human ratings on the sentence level. Results are shown in Table 4.

In order to measure correlation, we need to score each paraphrase individually. Thus, we recomputed BLEU on the sentence level and left the PINC scores unchanged. While BLEU is typically not reliable at the single sentence level, our large number of reference sentences makes BLEU more stable even at this granularity. Empirically, BLEU correlates fairly well with human judgments of semantic equivalence, although still not as well as the inter-annotator agreement. On the other hand, PINC correlates as well as humans agree with each other in assessing lexical dissimilarity. We also computed each metric's correlation with the overall ratings, although neither should be used alone to assess the overall quality of paraphrases.

PEM had the worst correlation with human judgments of all the metrics. Since PEM was trained on newswire data, its poor adaptation to this domain is expected. However, given the large amount of training data needed (PEM was trained on 250K Chinese-

	Semantic	Dissimilarity	Overall
Judge A vs. B	0.7135	0.6319	0.4920
BLEU vs. Human	0.5095	N/A	0.2127
PINC vs. Human	N/A	0.6672	0.0775
PEM vs. Human	N/A	N/A	0.0654
PINC vs. Human (BLEU > threshold)			
threshold = 0	N/A	0.6541	0.1817
threshold = 30	N/A	0.6493	0.1984
threshold = 60	N/A	0.6815	0.3986
threshold = 90	N/A	0.7922	0.4350
Combined BLEU and PINC vs. Human			
Arithmetic Mean	N/A	N/A	0.3173
Geometric Mean	N/A	N/A	0.3003
Harmonic Mean	N/A	N/A	0.3036
PINC × Sigmoid(BLEU)	N/A	N/A	0.3532

Table 4: Correlation between the human judges as well as between the automatic metrics and the human judges.

English sentence pairs and 2400 human ratings of paraphrase pairs), it is difficult to use PEM as a general metric. Adapting PEM to a new domain would require sufficient in-domain bilingual data to support paraphrase extraction. In contrast, our approach only requires monolingual data, and evaluation can be performed using arbitrarily small, highly-parallel datasets. Moreover, PEM requires sample human ratings in training, thereby lessening the advantage of having automatic metrics.

Since lexical dissimilarity is only desirable when the semantics of the original sentence is unchanged, we also computed correlation between PINC and the human ratings when BLEU is above certain thresholds. As we restrict our attention to the set of paraphrases with higher BLEU scores, we see an increase in correlation between PINC and the human assessments. This confirms our intuition that PINC is a more useful measure when semantic content has been preserved.

Finally, while we do not believe any single score could adequately describe the quality of a paraphrase outside of a specific application, we experimented with different ways of combining BLEU and PINC into a single score. Almost any simple combination, such as taking the average of the two, yielded decent correlation with the human ratings. The best correlation was achieved by taking the product of PINC and a sigmoid function of BLEU. This follows the intuition that semantic preservation is closer to a
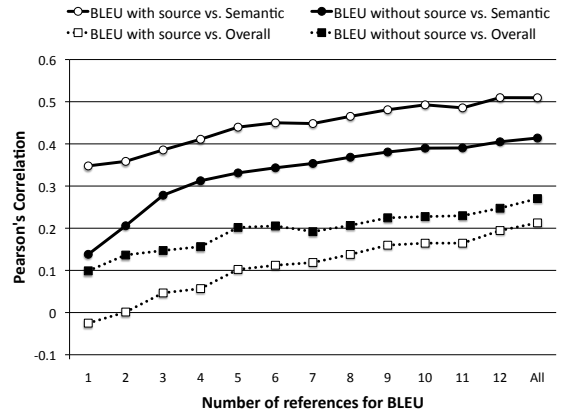

Figure 4: Correlation between BLEU and human judgments as we vary the number of reference sentences.

binary decision (i.e. a paraphrase either preserves the meaning or it does not, in which case PINC does not matter at all) than a linear function. We used an oracle to pick the best logistic function in our experiment. In practice, some sample human ratings would be required to tune this function. Other more complicated methods for combining BLEU and PINC are also possible with sample human ratings, such as using a SVM as was done in PEM.

We quantified the utility of our highly parallel data by computing the correlation between BLEU and human ratings when different numbers of references were available. The results are shown in Figure 4. As the number of references increases, the correlation with human ratings also increases. The graph also shows the effect of adding the source sentence as a reference. If our goal is to assess semantic equivalence only, then it is better to include the source sentence. If we are trying to assess the overall quality of the paraphrase, it is better to exclude the source sentence, since otherwise the metric will tend to favor paraphrases that introduce fewer changes.

5.3 Direct paraphrasing versus video annotation

In addition to collecting paraphrases through video annotations, we also experimented with the more traditional task of presenting a sentence to an annotator and explicitly asking for a paraphrase. We randomly selected a thousand sentences from our data and collected two paraphrases of each using Mechanical Turk. We conducted a post-annotation sur-

vey of workers who had completed both the video description and the direct paraphrasing tasks, and found that paraphrasing was considered more difficult and less enjoyable than describing videos. Of those surveyed, 92% found video annotations more enjoyable, and 75% found them easier. Based on the comments, the only drawback of the video annotation task is the time required to load and watch the videos. Overall, half of the workers preferred the video annotation task while only 16% of the workers preferred the paraphrasing task.

The data produced by the direct paraphrasing task also diverged less, since the annotators were inevitably biased by lexical choices and word order in the original sentences. On average, a direct paraphrase had a PINC score of 70.08, while a parallel description of the same video had a score of 78.75.

6 Discussions and Future Work

While our data collection framework yields useful parallel data, it also has some limitations. Finding appropriate videos is time-consuming and remains a bottleneck in the process. Also, more abstract actions such as *reducing the deficit* or *fighting for justice* cannot be easily captured by our method. One possible solution is to use longer video snippets or other visual stimuli such as graphs, schemas, or illustrated storybooks to convey more complicated information. However, the increased complexity is also likely to reduce the semantic closeness of the parallel descriptions.

Another limitation is that sentences produced by our framework tend to be short and follow similar syntactic structures. Asking annotators to write multiple descriptions or longer descriptions would result in more varied data but at the cost of more noise in the alignments. Other than descriptions, we could also ask the annotators for more complicated responses such as "fill in the blanks" in a dialogue (e.g. "If you were this person in the video, what would you say at this point?"), their opinion of the event shown, or the moral of the story. However, as with the difficulty of aligning news stories, finding paraphrases within these more complex responses could require additional annotation efforts.

In our experiments, we only used a subset of our corpus to avoid dealing with excessive noise. How-ever, a significant portion of the remaining data is useful. Thus, an automatic method for filtering those sentences could allow us to utilize even more of the data. For example, sentences from the Tier-2 tasks could be used as positive examples to train a string classifier to determine whether a noisy sentence belongs in the same cluster or not.

We have so far used BLEU to measure semantic adequacy since it is the most common MT metric. However, other more advanced MT metrics that have shown higher correlation with human judgments could also be used.

In addition to paraphrasing, our data collection framework could also be used to produces useful data for machine translation and computer vision. By pairing up descriptions of the same video in different languages, we obtain parallel data without requiring any bilingual skills. Another application for our data is to apply it to computer vision tasks such as video retrieval. The dataset can be readily used to train and evaluate systems that can automatically generate full descriptions of unseen videos. As far as we know, there are currently no datasets that contain whole-sentence descriptions of open-domain video segments.

7 Conclusion

We introduced a data collection framework that produces highly parallel data by asking different annotators to describe the same video segments. Deploying the framework on Mechanical Turk over a two-month period yielded 85K English descriptions for 2K videos, one of the largest paraphrase data resources publicly available. In addition, the highly parallel nature of the data allows us to use standard MT metrics such as BLEU to evaluate semantic adequacy reliably. Finally, we also introduced a new metric, PINC, to measure the lexical dissimilarity between the source sentence and the paraphrase.

References

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*.

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

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*.

Michael Bloodgood and Chris Callison-Burch. 2010a. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*.

Michael Bloodgood and Chris Callison-Burch. 2010b. Using Mechanical Turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Olivia Buzek, Philip Resnik, and Benjamin B. Bederson. 2010. Error driven paraphrase annotation using Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL-06)*.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*.

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

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*.

Wallace L. Chafe. 1997. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

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

Michael Denkowski and Alon Lavie. 2010. Exploring normalization techniques for human judgments of machine translation adequacy collected using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Michael Denkowski, Hassan Al-Haj, and Alon Lavie. 2010. Turker-assisted paraphrasing for English-Arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*.

Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):1–29.

Ali Ibrahim, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*.

Ann Irvine and Alexandre Klementiev. 2010. Using Mechanical Turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*.

Stanley Kok and Chris Brockett. 2010. Hitting the right paraphrases in good time. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-2010)*.

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-2001).

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. PEM: A paraphrase evaluation metric exploiting parallel texts. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).*

Xiaojuan Ma and Perry R. Cook. 2009. How well do visual verbs work in daily communication for young and old adults. In *Proceedings of ACM CHI 2009 Conference on Human Factors in Computing Systems.*

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of Human Language Technology Conference / North American Association for Computational Linguistics Annual Meeting (HLT-NAACL-2003).*

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 of the Association for Computational Linguistics (ACL-2002)*, pages 311–318, Philadelphia, PA, July.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004).*

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference.*

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems.*

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).*