Constructing Corpora for the Development and Evaluation of Paraphrase Systems

Trevor Cohn*
University of Edinburgh

Chris Callison-Burch**
Johns Hopkins University

Mirella Lapata†
University of Edinburgh

Automatic paraphrasing is an important component in many natural language processing tasks. In this article we present a new parallel corpus with paraphrase annotations. We adopt a definition of paraphrase based on word alignments and show that it yields high inter-annotator agreement. As Kappa is suited to nominal data, we employ an alternative agreement statistic which is appropriate for structured alignment tasks. We discuss how the corpus can be usefully employed in evaluating paraphrase systems automatically (e.g., by measuring precision, recall, and F1) and also in developing linguistically rich paraphrase models based on syntactic structure.

1. Introduction

The ability to paraphrase text automatically carries much practical import for many NLP applications ranging from summarization (Barzilay 2003; Zhou et al. 2006) to question answering (Lin and Pantel 2001; Duboue and Chu-Carroll 2006) and machine translation (Callison-Burch, Koehn, and Osborne 2006). It is therefore not surprising that recent years have witnessed increasing interest in the acquisition of paraphrases from real world corpora. These are most often *monolingual* corpora containing parallel translations of the same source text (Barzilay and McKeown 2001; Pang, Knight, and Marcu 2003). Truly *bilingual* corpora consisting of documents and their translations have also been used to acquire paraphrases (Bannard and Callison-Burch 2005; Callison-Burch 2007) as well as *comparable* corpora such as collections of articles produced by two different newswire agencies about the same events (Barzilay and Elhadad 2003).

Although paraphrase induction algorithms differ in many respects—for example, the acquired paraphrases often vary in granularity as they can be lexical (*fighting, battle*) or structural (*last week's fighting, the battle last week*), and are represented as words or

* School of Informatics, University of Edinburgh, EH8 9AB, Edinburgh, UK. E-mail: tcohn@inf.ed.ac.uk.
** Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, 21218. E-mail: ccb@cs.jhu.edu.
† School of Informatics, University of Edinburgh, EH8 9AB, Edinburgh, UK. E-mail: mlap@inf.ed.ac.uk.

Submission received: 10 September 2007; revised submission received: 8 February 2008; accepted for publication: 26 March 2008.

syntax trees—they all rely on some form of alignment for extracting paraphrase pairs. In its simplest form, the alignment can range over individual words, as is often done in machine translation (Quirk, Brockett, and Dolan 2004). In other cases, the alignments range over entire trees (Pang, Knight, and Marcu 2003) or sentence clusters (Barzilay and Lee 2003).

The obtained paraphrases are typically evaluated via human judgments. Paraphrase pairs are presented to judges who are asked to decide whether they are semantically equivalent, that is, whether they can be generally substituted for one another in the same context without great information loss (Barzilay and Lee 2003; Barzilay and McKeown 2001; Pang, Knight, and Marcu 2003; Bannard and Callison-Burch 2005). In some cases the automatically acquired paraphrases are compared against manually generated ones (Lin and Pantel 2001) or evaluated indirectly by demonstrating performance increase for a specific application, such as machine translation (Callison-Burch, Koehn, and Osborne 2006).

Unfortunately, manually evaluating paraphrases in this way has at least three drawbacks. First, it is infeasible to perform frequent evaluations when assessing incremental system changes or tuning system parameters. Second, it is difficult to replicate results presented in previous work because there is no standard corpus, and no standard evaluation methodology. Consequently comparisons across systems are few and far between. The third drawback concerns the evaluation studies themselves, which primarily focus on precision. Recall is almost never evaluated directly in the literature. And this is for a good reason: There is no guarantee that participants will identify the same set of paraphrases as each other or with a computational model. The problem relates to the nature of the paraphrasing task, which has so far eluded formal definition (see the discussion in Barzilay [2003]). Such a definition is not so crucial when assessing precision, because subjects are asked to rate the paraphrases without actually having to identify them. However, recall might be measured with respect to some set of "gold-standard" paraphrases which will have to be collected according to some concrete definition.

In this article we present a resource that could potentially be used to address these problems. Specifically, we create a monolingual parallel corpus with human paraphrase annotations. Our working definition of paraphrase is based on word and phrase[1] alignments between semantically equivalent sentences. Other definitions are possible, for instance we could have asked our annotators to identify all constituents that are more or less meaning preserving in our parallel corpus. We chose to work with alignments for two reasons. First, the notion of alignment appears to be central in paraphrasing—most existing paraphrase induction algorithms rely on alignments either implicitly or explicitly for identifying paraphrase units. Secondly, research in machine translation, where several gold-standard alignment corpora have been created, shows that word alignments can be identified reliably by annotators (Melamed 1998; Och and Ney 2000b; Mihalcea and Pedersen 2003; Martin, Mihalcea, and Pedersen 2005). We therefore create word alignments similar to those observed in machine translation, namely, featuring one-to-one, one-to-many, many-to-one, and many-to-many links between words. Alignment blocks larger than one-to-one are used to specify phrase correspondences.

1 Our definition of the term *phrase* follows the SMT literature. It refers to any contiguous sequence of words, whether it is a syntactic constituent or not. See Section 2 for details.

In the following section we explain how our corpus was created and summarize our annotation guidelines. Section 3 gives the details of an agreement study, demonstrating that our annotators can identify and align paraphrases reliably. We measure agreement using alignment overlap measures from the SMT literature, and also introduce a novel agreement statistic for non-enumerable labeling spaces. Section 4 illustrates how the corpus can be used in paraphrase research, for example, as a test set for evaluating the output of automatic systems or as a training set for the development of paraphrase systems. Discussion of our results concludes the article.

2. Corpus Creation and Annotation

Our corpus was compiled from three data sources that have been previously used for paraphrase induction (Barzilay and McKeown 2001; Pang, Knight, and Marcu 2003; Dolan, Quirk, and Brockett 2004): the Multiple-Translation Chinese (MTC) corpus, Jules Verne's *Twenty Thousand Leagues Under the Sea* novel (Leagues), and the Microsoft Research (MSR) paraphrase corpus. These are monolingual parallel corpora, aligned at the sentence level. Both source and target sentences are in English, and express the same content using different surface forms.

The MTC corpus contains news stories from three sources of journalistic Mandarin Chinese text.[2] These stories were translated into English by 11 translation agencies. Because the majority of the translators were non-native English speakers, occasionally translations contain syntactic or grammatical errors and are not entirely fluent. After inspection, we identified four translators with consistently fluent English, and used their sentences for our corpus. The Leagues corpus contains two English translations of the French novel *Twenty Thousand Leagues Under the Sea*. The corpus was created by Tagyoung Chung and manually aligned at the *paragraph* level.[3] In order to obtain *sentence* level paraphrase pairs, we sampled from the subset of one-to-one sentence alignments. The MSR corpus was harvested automatically from online news sources.[4] The obtained sentence pairs were further submitted to judges who rated them as being semantically equivalent or not (Dolan, Quirk, and Brockett 2004). We only used semantically equivalent pairs. The sentence pairs were filtered for length (≤ 50) and length ratio ($\leq 1:9$ between the shorter and longer sentence). This was necessary to prune out incorrectly aligned sentences.

We randomly sampled 300 sentence pairs from each corpus (900 in total). Of these, 300 pairs (100 per corpus) were first annotated by two coders to assess inter-annotator agreement. The remaining 600 sentence pairs were split into two distinct sets, each consisting of 300 sentences (100 per corpus), and were annotated by a single coder. Each coder annotated the same amount of data. In addition, we obtained a trial set of 50 sentences from the MTC corpus which was used for familiarizing our annotators with the paraphrase alignment task (this set does not form part of the corpus). In sum, we obtained paraphrase annotations for 900 sentence pairs, 300 of which are doubly annotated.

To speed up the annotation process, the data sources were first aligned automatically and then hand-corrected. We used Giza++ (Och and Ney 2003), a publicly available

2 The corpus is made available by the LDC, Catalog Number LDC2002T01, ISBN 1-58563-217-1.

3 The corpus can be downloaded from `http://www.isi.edu/~knight/`.

4 The corpus is available at `http://research.microsoft.com/research/downloads/Details/607D14D9-20CD-47E3-85BC-A2F65CD28042/Details.aspx`.

implementation of the IBM word alignment models (Brown et al. 1993). Giza++ was trained on the full 993-sentence MTC part1 corpus[5] using all 11 translators and all pairings of English translations as training instances. This resulted in $55 = \frac{11 \cdot (11-1)}{2}$ training pairs per sentence and a total of 54,615 training pairs. In addition, we augmented the training data with a word-identity lexicon, as proposed by Quirk, Brockett, and Dolan (2004). This follows standard practice in SMT where entries from a bilingual dictionary are added to the training set (Och and Ney 2000a), except in our case the "dictionary" is monolingual and specifies that each word type can be paraphrased as itself. This is necessary in order to inform Giza++ about word identity.

A common problem with automatic word alignments is that they are asymmetric: one source word can only be aligned to one target word, whereas one target word can be aligned to multiple source words. In SMT, word alignments are typically predicted in both directions: source-to-target and target-to-source. These two alignments are then merged (symmetrized) to produce the final alignment (Koehn, Och, and Marcu 2003). Symmetrization improves the alignment quality compared to that of a single directional model, while also allowing a greater range of alignment types (i.e., some many-to-one, one-to-many, and many-to-many alignments can be produced). Analogously, we obtained word alignments in both directions[6] which we subsequently merged by taking their intersection. This resulted in a high precision and low recall alignment.

Our annotators (two linguistics graduates) were given pairs of sentences and asked to show which parts of these were in *correspondence* by aligning them on a word-by-word basis.[7] Our definition of alignment was fairly general (Och and Ney 2003): Given a source string $X = x_1, \ldots, x_N$ and a target string $Y = y_1, \ldots, y_M$, an alignment $\mathcal{A}$ between two word strings is the subset of the Cartesian product of the word positions:

$$\mathcal{A} \subseteq \{(n, m) : n = 1, \ldots, N; m = 1, \ldots, M\} \quad (1)$$

We did not provide a formal definition of what constitutes a correspondence. As a rule of thumb, annotators were told to align words or phrases $x \leftrightarrow y$ in two sentences (X, Y) whenever the words x could be substituted for y in Y, or vice versa. This relationship should hold within the context of the sentence pair in question: the relation $x \leftrightarrow y$ need not hold in general contexts. Trivially this definition allowed for identical word pairs.

Following common practice (Och, Tillmann, and Ney 1999; Och and Ney 2003; Daumé III and Marcu 2004), we distinguished between *sure* (S) and *possible* (P) alignments, where $S \subseteq P$. The intuition here is that sure alignments are clear-cut decisions and typical of genuinely substitutable words or phrases, whereas possible alignments flag a correspondence that has slightly divergent syntax or semantics. Annotators were encouraged to produce sure alignments. They were also instructed to prefer smaller alignments whenever possible, but were allowed to create larger block alignments. Smaller alignments were generally used to indicate lexical correspondences, whereas block alignments were reserved for non-compositional phrase pairs (e.g., idiomatic expressions) or simply expressions with radically different syntax or vocabulary. In

5 The IBM alignment models require a large amount of parallel data to yield reliable alignments. We therefore selected the MTC for training purposes as it was the largest of our parallel corpora.

6 We used five iterations for each of Model 1, Model 2, and the HMM model.

7 The annotation was conducted using a Web-based alignment tool available at `http://demo.linearb.co.uk/paraphrases/`.

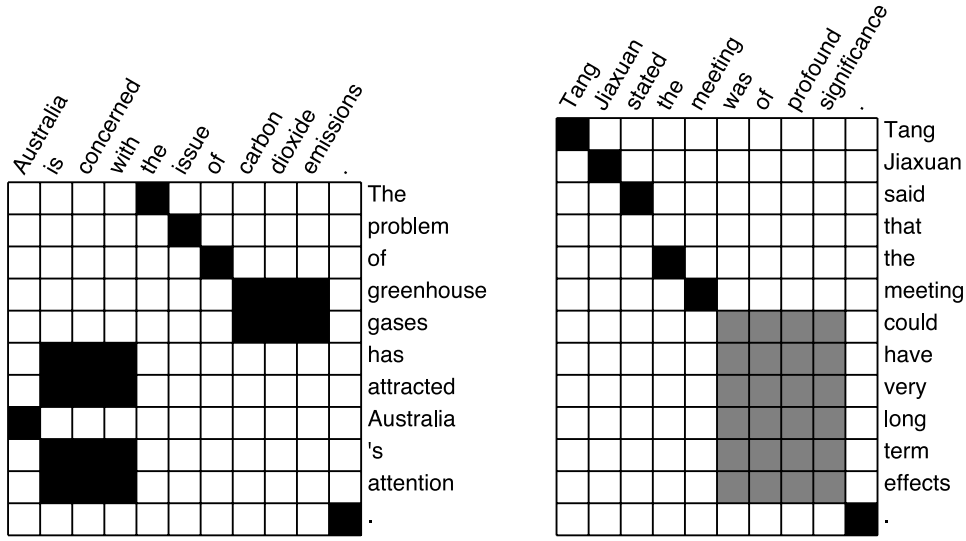

Figure 1
Manual alignment between two sentence pairs from the MTC corpus, displayed as a grid. Black squares represent sure alignment, gray squares represent possible alignment.

cases where information in one sentence was not present in the other, the annotators were asked to leave this information unaligned.

Finally, annotators were given a list of heuristics to help them decide how to make alignments in cases of ambiguity. These heuristics handled the alignment of named entities (e.g., *George Bush*) and definite descriptions (e.g., *the president*), tenses (e.g., *had been* and *shall be*), noun phrases with mismatching determiners (e.g., *a man* and *the man*), verb complexes (e.g., *was developed* and *had been developed*), phrasal verbs (e.g., *take up* and *accept*), genitives (e.g., *Bush's infrequent speeches* and *the infrequent speeches by Bush*), pronouns, repetitions, typographic errors, and approximate correspondences. For more details, we refer the interested reader to our annotation guidelines.[8]

Figure 1 shows the alignment for two sentence pairs from the MTC corpus. The first pair (*Australia is concerned with the issue of carbon dioxide emissions.* ↔ *The problem of greenhouse gases has attracted Australia's attention.*) contains examples of word-to-word (*the* ↔ *The*; *issue* ↔ *problem*; *of* ↔ *of*; *Australia* ↔ *Australia*) and many-to-many alignments (*carbon dioxide emissions* ↔ *greenhouse gases*). Importantly, we do not use a large many-to-many block for *Australia is concerned with* and *has attracted Australia's attention* because it is possible to decompose the two phrases into smaller alignments. The second sentence pair illustrates a possible alignment (*could have very long term effects* ↔ *was of profound significance*) indicated by the gray squares. Possible alignments are used here because the two phrases only loosely correspond to each other. Possible alignments are also used to mark significant changes in syntax where the words denote a similar concept: for example, in cases where two words have the same stem but are

8 Both the corpus and the annotation guidelines can be found at: `http://homepages.inf.ed.ac.uk/tcohn/paraphrase_corpus.html`.

expressed with different parts of speech, (e.g., *co-operative* ↔ *cooperation*) or when two verbs are used that are not synonyms (e.g., *this is also* ↔ *this also marks*).

3. Human Agreement

As mentioned in the previous section, 300 sentence pairs (100 pairs from each subcorpus) were doubly annotated, in order to measure inter-annotator agreement. Here, we treat one annotator as gold-standard (reference) and measure the extent to which the other annotator deviates from this reference.

Word-Based Measures. The standard technique for evaluating word alignments is to represent them as a set of links (i.e., pairs of words) and compare them against gold-standard alignments. The quality of an alignment $\mathcal{A}$ (defined in Equation (1)) compared to reference alignment $\mathcal{B}$ can be then computed using standard recall, precision, and F1 measures (Och and Ney 2003):

$$Precision = \frac{|\mathcal{A}_S \cap \mathcal{B}_P|}{|\mathcal{A}_S|} \qquad Recall = \frac{|\mathcal{A}_P \cap \mathcal{B}_S|}{|\mathcal{B}_S|} \qquad F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2}$$

where the subscripts S and P denote sure and possible word alignments, respectively. Note that both precision and recall are asymmetric in that they compare sets of possible and sure alignments. This is designed to be maximally generous: sure predictions which are present in the reference as possibles are not penalized in precision, and the converse applies for recall. We adopt Fraser and Marcu (2007)'s definition of F1, an F-measure between precision and recall over the sure and possibles. They argue that it is a better alternative to the commonly used Alignment Error Rate (AER), which does not sufficiently penalize unbalanced precision and recall.[9] As our corpus is monolingual, in order to avoid artificial score inflation, we limit the precision and recall calculations to consider only pairs of non-identical words (and phrases, as discussed subsequently).

To give an example, consider the sentence pairs in Figure 2, whose alignments have been produced by the two annotators A (left) and B (right). Table 1 shows the individual word alignments for each annotator and their type (sure or possible). In order to measure F1, we must first estimate *Precision* and *Recall* (see Equation (2)). Treating annotator B as the gold standard, $|\mathcal{A}_S| = 4$, $|\mathcal{B}_S| = 5$, $|\mathcal{A}_S \cap \mathcal{B}_P| = 4$, and $|\mathcal{A}_P \cap \mathcal{B}_S| = 4$. This results in a precision of $\frac{4}{4} = 1$, a recall of $\frac{4}{5}$, and F1 of $\frac{2 \times 1 \times 0.8}{1+0.8} = 0.89$. Note that we ignore alignments over identical words (i.e., *discussed* ↔ *discussed*, *the* ↔ *the*, *and* ↔ *and*, . ↔ .).

Phrase-Based Measures. The given definitions are all word-based; however, our annotators, and several paraphrasing models, create correspondences not only between words but also between phrases. To take this into account, we also evaluate these measures over larger blocks (similar to Ayan and Dorr [2006]). Specifically, we extract phrase pairs from the alignments produced by our annotators using a modified version of the standard SMT phrase extraction heuristic (Och, Tillmann, and Ney 1999). The heuristic

9 Fraser and Marcu (2007) also argue for an unbalanced F-measure to bias towards recall. This is shown to correlate better with translation quality. For paraphrasing it is not clear if such a bias would be beneficial.

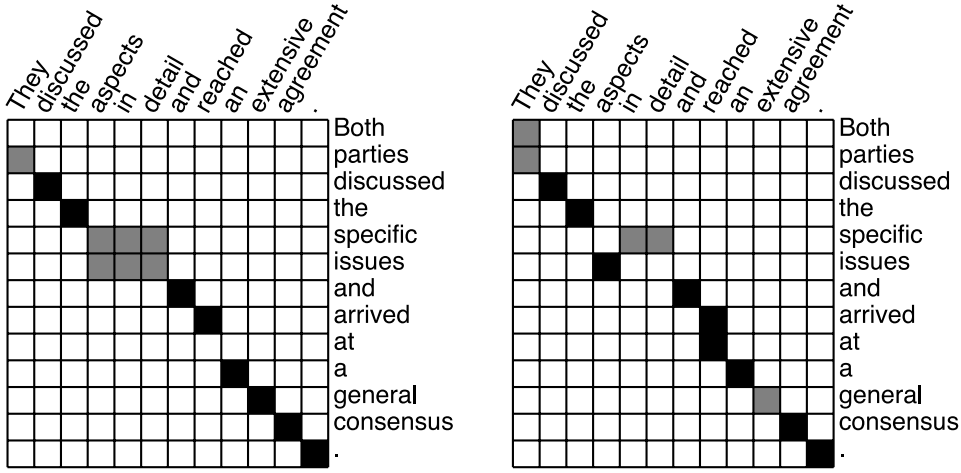

Figure 2
Sample sentence pair showing the word alignments from two annotators.

extracts all phrase pairs *consistent* with the word alignment. These include phrase pairs whose words are aligned to each other or nothing, but not to words outside the phrase boundaries.[10] The phrase extraction heuristic creates masses of phrase pairs, many of which are of dubious quality. This is often due to the inclusion of unaligned words or simply to the extraction of overly-large phrase pairs which might be better decomposed into smaller units. For our purposes we wish to be maximally conservative in how we process the data, and therefore we do not extract phrase pairs with unaligned words on their boundaries.

Figure 3 illustrates the types of phrase pairs our extraction heuristic permits. Here, the pair *and reached* $\leftrightarrow$ *and arrived at* is consistent with the word alignment. In contrast, the pair *and reached* $\leftrightarrow$ *and arrived* isn't; there is an alignment outside the hypothetical phrase boundary which is not accounted for (*reached* is also aligned to *at*). The phrase pair *and reached an* $\leftrightarrow$ *and arrived at* is consistent with the word alignment; however it has an unaligned word (i.e., *an*) on the phrase boundary, which we disallow.

Our phrase extraction procedure distinguishes between two types of phrase pairs: **atomic**, that is, the smallest possible phrase pairs, and **composite**, which can be created by combining smaller phrase pairs. For example, the phrase pair *and reached* $\leftrightarrow$ *and arrived at* in Figure 3 is composite, as it can be decomposed into *and* $\leftrightarrow$ *and* and *reached* $\leftrightarrow$ *arrived at*. Table 2 shows the atomic and composite phrase pairs extracted from the possible alignments produced by annotators A and B for the sentence pair in Figure 2.

We compute recall, precision, and F1 over the phrase pairs extracted from the word alignments as follows:

$$Precision = \frac{|\mathcal{A}^p_{atom} \cap \mathcal{B}^p|}{|\mathcal{A}^p_{atom}|} \quad Recall = \frac{|\mathcal{A}^p \cap \mathcal{B}^p_{atom}|}{|\mathcal{B}^p_{atom}|} \quad F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

10 The term *phrase* is not used here in the linguistic sense; many extracted phrases will not be constituents.

Table 1
Single word pairs specified by the word alignments from Figure 2, for two annotators, A and B. The column entries specify the alignment type for each annotator, either sure (S) or possible (P). Dashes indicate that the word pair was not predicted by the annotator. *Italics* denote lexically identical word pairs.

Word alignments			A	B
they	↔	both	–	P
they	↔	parties	P	P
discussed	↔	*discussed*	*S*	*S*
the	↔	*the*	*S*	*S*
aspects	↔	specific	P	–
in	↔	specific	P	P
detail	↔	specific	P	P
aspects	↔	issues	P	S
in	↔	issues	P	–
detail	↔	issues	P	–
and	↔	*and*	*S*	*S*
reached	↔	arrived	S	S
reached	↔	at	–	S
an	↔	a	S	S
extensive	↔	general	S	P
agreement	↔	consensus	S	S
.	↔	.	*S*	*S*

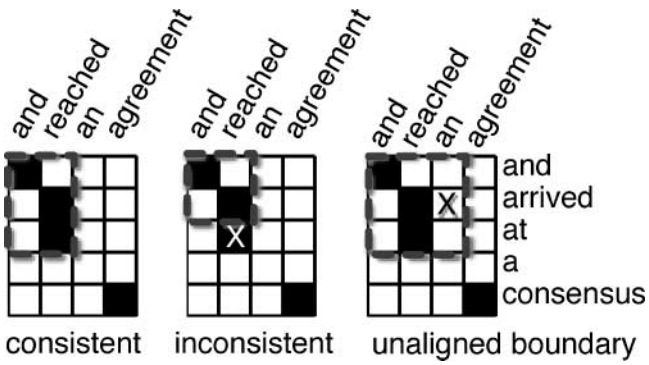

Figure 3
Validity of phrase pairs according to the phrase extraction heuristic. Only the leftmost phrase pair is valid. The others are inconsistent with the alignment or have an unaligned word on a boundary, respectively, indicated by a cross.

where $\mathcal{A}^p$ and $\mathcal{B}^p$ are the predicted and reference phrase pairs, respectively, and the *atom* subscript denotes the subset of atomic phrase pairs, $\mathcal{A}^p_{atom} \subseteq \mathcal{A}^p$. As shown in Equation (3) we measure precision and recall between atomic phrase pairs and the full space of atomic and composite phrase pairs. This ensures that we do not multiply reward composite phrase pair combinations,[11] while also not unduly penalizing non-matching phrase pairs which are composed of atomic phrase pairs in

11 This contrasts with Ayan and Dorr (2006), who use all phrase pairs up to a given size, and therefore might multiply count phrase pairs.

Table 2
Phrase pairs are specified by the word alignments from Figure 2, using the possible alignments. The entire set of atomic phrase pairs for either annotator (labeled A or B) and a selection of the remaining 57 composite phrase pairs are shown. The *italics* denote lexically identical phrase pairs. *This phrase pair is atomic in A but composite in B.

Atomic phrase pairs			A	B
they	↔	parties	P	–
they	↔	both parties	–	P
discussed	↔	*discussed*	*S*	*S*
the	↔	*the*	*S*	*S*
aspects in detail	↔	specific issues	P	–
in detail	↔	specific	–	P
aspects	↔	issues	–	S
and	↔	*and*	*S*	*S*
reached	↔	arrived	S	S
reached	↔	arrived at	–	S
reached an	↔	arrived at a	S	P*
an	↔	a	S	S
extensive	↔	general	S	P
agreement	↔	consensus	S	S
.	↔	.	*S*	*S*
Composite phrase pairs			**A**	**B**
they discussed	↔	both parties discussed	–	P
they discussed	↔	parties discussed	P	–
they discussed the	↔	both parties discussed the	–	P
they discussed the	↔	parties discussed the	P	–
they ... reached an	↔	both parties ... arrived at a	P	–
the aspects in detail	↔	the specific issues	P	P
reached an extensive	↔	arrived at a general	S	S
extensive agreement .	↔	general consensus .	S	S
		...		

the reference. Returning to the example in Table 2, with annotator B as the gold standard, $|\mathcal{A}^p_{atom}| = 7$, $|\mathcal{B}^p_{atom}| = 8$, $|\mathcal{A}^p_{atom} \cap \mathcal{B}^p| = 5$, and $|\mathcal{A}^p \cap \mathcal{B}^p_{atom}| = 4$. Consequently, precision= $\frac{5}{7} = 0.71$, recall= $\frac{4}{8} = 0.50$, and F1= $\frac{2 \times 0.71 \times 0.50}{0.71 + 0.50} = 0.59$. Again we ignore identical phrase pairs.

A potential caveat here concerns the quality of the atomic phrase pairs, which are automatically induced and may not correspond to linguistic intuition. To evaluate this, we had two annotators review a random sample of 166 atomic phrase pairs drawn from the MTC corpus (sure), classifying each phrase pair as correct, incorrect, or uncertain given the sentence pair as context. From this set, 73% were deemed correct, 22% uncertain, and 5% incorrect.[12] Annotators agreed in their decisions 75% of the time (using the Kappa[13] statistic, their agreement is 0.61). This confirms that the phrase-extraction process produces reliable phrase pairs from our word-aligned data (although we cannot claim that it is exhaustive).

12 Taking a more conservative position by limiting the proportion of unaligned words within the phrase pair improves these figures monotonically to 90% correct and 0% incorrect (fully aligned phrase pairs).

13 This Kappa is computed over three nominal categories (correct, incorrect, and uncertain) and should not be confused with the agreement measure we develop in the following section for phrase pairs.

Chance-Corrected Agreement. Besides precision and recall, inter-annotator agreement is commonly measured using the Kappa statistic (Cohen 1960). Thus is a desirable measure because it is adjusted for agreement due purely to chance:

$$\kappa = \frac{\Pr(A) - \Pr(E)}{1 - \Pr(E)} \tag{4}$$

where $\Pr(A)$ is the proportion of times two coders[14] agree, corrected by $\Pr(E)$, the proportion of times we would expect them to agree by chance.

Kappa is a suitable agreement measure for nominal data. An example would be a classification task, where two coders must assign n linguistic instances (e.g., sentences or words) into one of m categories. Given this situation, it would be possible for each coder to assign each instance to the same category. Kappa allows us to quantify whether the coders agree with each other about the category membership of each instance. It is relatively straightforward to estimate $\Pr(A)$—it is the proportion of instances on which the two coders agree. $\Pr(E)$ requires a model of what would happen if the coders were to assign categories randomly. Under the assumption that coders r_1 and r_2 are independent, the chance of them agreeing on the jth category is the product of each of them assigning an instance to that category: $\Pr(C_j|r_1)\Pr(C_j|r_2)$. Chance agreement is then the sum of this product across all categories: $\Pr(E) = \sum_{j=1}^{m} \Pr(C_j|r_1)\Pr(C_j|r_2)$. The literature describes two different methods for estimating $\Pr(C_j|r_i)$. Either a separate distribution is estimated for each coder (Cohen 1960) or the same distribution for all coders (Scott 1955; Fleiss 1971; Siegel and Castellan 1988). We refer the interested reader to Di Eugenio and Glass (2004) and Artstein and Poesio (2008) for a more detailed discussion.

Unfortunately, Kappa is not universally suited to every categorization task. A prime example is structured labeling problems that allow a wide variety of output categories. Importantly, the number and type of categories is not fixed in advance and can vary from instance to instance. In parsing, annotators are given a sentence for which they must specify a tree, of which there is an exponential number in the sentence length. Similarly, in our case the space of possible alignments for a sentence pair is also exponential in the input sentence lengths. Considering these annotations as nominal variables is inappropriate.

Besides, alignments are only an intermediate representation that we have used to facilitate the annotation of paraphrases. Ideally, we would like to measure agreement over the set of phrase pairs which are specified by our annotators (via the word alignments), not the alignment matrices themselves.

Kupper and Hafner (1989) present an alternative measure similar to Kappa that is especially designed for sets of variables:

$$\hat{C} = \frac{\hat{\pi} - \pi_0}{1 - \pi_0}, \tag{5}$$

$$\text{where } \hat{\pi} = \sum_{i=1}^{I} \frac{|A_i \cap B_i|}{\min(|A_i|, |B_i|)}, \text{ and } \pi_0 = \frac{1}{Ik} \sum_{i} \min(|A_i|, |B_i|)$$

14 Kappa has been extended to more than two coders (Fleiss 1971; Bartko and Carpenter 1976). For simplicity's sake our discussion and subsequent examples involve two coders. Also note that we use the term *coder* instead of the more common *rater*. This is because in our task the annotators must identify (a.k.a. code) the paraphrases rather than rate them.

Here, A_i and B_i are the coders' predictions on sentence pair i from our corpus of I sentence pairs. Each prediction is a subset of the full space of k items. Expression (5) measures the agreement (or concordance) between coders A and B and follows the general form of Kappa from Equation (4), which is defined analogously with $\Pr(A)$ and $\Pr(E)$ taking the roles of $\hat{\pi}$ and π_0, but with different definitions.

Kupper and Hafner (1989) developed their agreement measure with medical diagnostic tasks in mind. For example, two physicians classify subjects into $k = 3$ diagnostic categories and wish to find out whether they agree in their diagnoses. Here, each coder must decide which (possibly empty) subset from k categories best describes each subject. The size of k is thus invariant with the instance under consideration. This is not true in our case, where k will vary across sentence pairs as sentences of different lengths license different numbers of phrase pairs. More critically, the formulation in Equation (5) assumes that items in the set are independent: All subsets of the same cardinality as k are equally likely, and no combination is impossible. This independence assumption is inappropriate for the paraphrase annotation task. The phrase extraction heuristic allows each contiguous span in a sentence to be aligned to either zero or one span in the other sentence; that is, nominating a phrase pair precludes the choice of many other possible phrase pairs. Consequently relatively few of the subsets of the full set of possible phrase pairs are valid. Formally, an alignment can specify only $O(N^2)$ phrase pairs from a total set of $k = O(N^4)$ possible phrase pairs. This disparity in magnitudes leads to increasingly underestimated $\hat{\pi}$ for larger N, namely, $\lim_{N\to\infty} \pi_0 = \lim_{N\to\infty} O(N^2)/O(N^4) = 0$. The end result is an overestimate of $\hat{C}$ on longer sentences.

For these reasons, we adapt the method of Kupper and Hafner (1989) to account for our highly interdependent item sets. We use $\hat{C}$ from Equation (5) as our agreement statistic defined over sets of atomic phrase pairs, that is, $A = \mathcal{A}^p_{atom}$, $B = \mathcal{B}^p_{atom}$. We redefine π_0 as follows:

$$\pi_0 = \frac{1}{I} \sum_{i=1}^{I} \sum_{\mathcal{A}^p_{atom}} \sum_{\mathcal{B}^p_{atom}} \Pr(\mathcal{A}^p_{atom}) \Pr(\mathcal{B}^p_{atom}) \frac{|\mathcal{A}^p_{atom} \cap \mathcal{B}^p_{atom}|}{\min(|\mathcal{A}^p_{atom}|, |\mathcal{B}^p_{atom}|)} \tag{6}$$

where $\mathcal{A}^p_{atom}$ and $\mathcal{B}^p_{atom}$ range over the sets of atomic phrase pairs licensed by sentence pair i, and $\Pr(\mathcal{A}^p_{atom})$ and $\Pr(\mathcal{B}^p_{atom})$ are priors over these sets for each annotator. A consequence of dropping the independence assumptions is that calculating π_0 is considerably more difficult.

While it may be possible to calculate π_0 analytically, this gets increasingly complicated for larger phrase pairs or with an expressive prior. For the sake of flexibility we estimate π_0 using Monte Carlo sampling. Specifically, we approximate the full sum by drawing samples from a prior distribution over sets of phrase pairs for each of our annotators ($\Pr(\mathcal{A}^p_{atom})$ and $\Pr(\mathcal{B}^p_{atom})$ in Equation (6)). These samples are then compared using the intersection metric. This is repeated many times and the results are then averaged. More formally:

$$\hat{\pi}_0 = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{J} \sum_{j=1}^{J} \frac{|\mathcal{A}^{p\ (j)}_{atom} \cap \mathcal{B}^{p\ (j)}_{atom}|}{\min(|\mathcal{A}^{p\ (j)}_{atom}|, |\mathcal{B}^{p\ (j)}_{atom}|)} \tag{7}$$

where for each sentence pair, i, we draw J samples of pairs of sets of phrase pairs, $(\mathcal{A}^p_{atom}, \mathcal{B}^p_{atom})$. We use $J = 1,000$, which is ample to give reliable estimates. So far we have

not defined how we sample valid sets of phrase pairs. This is done via the word alignments. Recall that the annotators start out with alignments from an automatic word-aligner. Firstly, we develop a distribution to predict how often an annotator changes a cell from the initial alignment matrix. We model the number of changes made with a binomial distribution, that is, each local change is assumed independent and has a fixed probability, $\Pr(edit|r, N_i, M_i)$ where r is the coder and N_i and M_i are the sentence lengths. This distribution is fit to each annotator's predictions using a linear function over the combined length of two sentences. Next we sample word alignments. Each sample starts with the automatic alignment, and each cell is changed with probability $\Pr(edit)$. These changes are binary, swapping alignments for non-alignments and vice versa. Finally, the phrase-extraction heuristic is run over the alignment matrix to produce a set of phrase pairs. This is done for each annotator, A and B, after which we have a sample, $(\mathcal{A}^p_{atom}, \mathcal{B}^p_{atom})$. Each sample is then fed into Equation (7). Admittedly, this is not the most accurate prior, as annotators are not just randomly changing the alignment, but instead are influenced by the content expressed by the sentence pair and other factors such as syntactic complexity. However, this prior produces estimates for $\hat{\pi}_0$ which are several orders of magnitude larger than those using Kupper and Hafner's model of π_0 in Equation (5).

We now illustrate the process of measuring chance-corrected agreement, $\hat{C}$, with respect to the example in Figure 2. Here, $|\mathcal{A}^p_{atom}| = 7$, $|\mathcal{B}^p_{atom}| = 8$, $|\mathcal{A}^p_{atom} \cap \mathcal{B}^p_{atom}| = 4$, and therefore $\hat{\pi} = \frac{4}{7} = 0.571$. For this sentence our annotators edited eight and nine alignment cells, respectively, of the initial alignment matrix. This translates into $\Pr(edit|r = A) = \frac{8}{12\times13} = 5.13\%$ and $\Pr(edit|r = B) = 5.77\%$. Given these priors, we run the Monte Carlo sampling process from Equation (7), which results in $\hat{\pi}_0 = 0.147$. Combining the agreement estimate, $\hat{\pi}$, and chance correction estimate, $\hat{\pi}_0$, using Equation (6) results in $\hat{C} = \frac{0.571-0.147}{1-0.147} = 0.497$.

Now, imagine a hypothetical case where $\hat{\pi} = \frac{4}{7} = 0.571$ (i.e., the agreement is the same as before), annotator B edits nine alignment cells, but annotator A chooses not to make any edits. This leads to an increased estimate of $\hat{\pi}_0 = 0.259$ and a decreased $\hat{C} = 0.442$. If both annotators were not to make any edits, $\hat{\pi}_0 = 1$ and $\hat{C} = -\infty$. Interestingly, at the other extreme when $\Pr(edit|r = A) = \Pr(edit|r = B) = 1$, agreement is also perfect, $\hat{\pi}_0 = 1$ and $\hat{C} = -\infty$. This is because only one phrase pair can be extracted which consists of the two full sentences.

Results. Tables 3 and 4 display agreement statistics on our three corpora using precision, recall, F1, and $\hat{C}$. Specifically, we estimate $\hat{C}$ by aggregating $\hat{\pi}$ and $\hat{\pi}_0$ into corpus-level estimates. Table 3 shows agreement scores for individual words, whereas Table 4 shows agreement for phrase pairs. In both cases the agreement is computed over non-identical word and phrase pairs which are more likely to correspond to paraphrases. The agreement figures are broken down into possible (Poss) and sure alignments (Sure) for precision and recall.

When agreement is measured over words, our annotators obtain high F1 on all three corpora (MTC, Leagues, and News). Recall on Possibles seems worse on the News corpus when compared to MTC or Leagues. This is to be expected because this corpus was automatically harvested from the Web, and some of its instances may not be representative examples of paraphrases. For example, it is common for one sentence to provide considerably more details than the other, despite the fact that both describe the same event. The annotators in turn have difficulty deciding whether such instances are valid paraphrases. The $\hat{C}$ scores for the three corpora are in the same ballpark.

Table 3
Inter-annotator agreement using precision, recall, F1, and $\hat{C}$; the agreement is measured over *words*.

MTC			Leagues			News		
Measure	Poss	Sure	Measure	Poss	Sure	Measure	Poss	Sure
Prec	0.79	0.59	Prec	0.85	0.73	Prec	0.78	0.55
Rec	0.77	0.73	Rec	0.74	0.75	Rec	0.57	0.70
F1	0.76		F1	0.79		F1	0.74	
$\hat{C}$	0.85		$\hat{C}$	0.87		$\hat{C}$	0.89	

Table 4
Inter-annotator agreement using precision, recall, F1, and $\hat{C}$; the agreement is measured over *atomic phrase pairs*.

MTC			Leagues			News		
Measure	Poss	Sure	Measure	Poss	Sure	Measure	Poss	Sure
Prec	0.77	0.67	Prec	0.74	0.72	Prec	0.72	0.68
Rec	0.77	0.66	Rec	0.77	0.73	Rec	0.69	0.81
F1	0.71		F1	0.74		F1	0.76	
$\hat{C}$	0.63		$\hat{C}$	0.62		$\hat{C}$	0.53	

Interestingly, $\hat{C}$ is highest on the News corpus, whereas F1 is lowest. Whereas precision and recall are normalized by the number of predictions from annotators A and B, respectively, $\hat{C}$ is normalized by the minimum number of predictions between the two. Therefore, when the predictions are highly divergent, $\hat{C}$ will paint a rosier picture than F1 (which is the combination of precision and recall). This indeed seems to be the case for the News corpus, where precision and recall have a higher spread in comparison to the other two corpora (see the Poss column in Table 3).

Agreement scores tend to be lower when taking phrases into account (see Table 4). This is expected because annotators are faced with a more complex task; they must generally make more decisions: for example, determining the phrase boundaries and how to align their constituent words. An exception to this trend is the News corpus where the F1 is higher for phrase pairs than for individual word pairs. This is due to the fact that there are many similar sentence pairs in this data. These have many identical words and a few different words. The differences are often in a clump (e.g., person names, verb phrases), rather than distributed throughout the sentence. The annotators tend to block align these and there is a large scope for disagreement. Whereas estimating agreement over words heavily penalizes block differences, when phrases are taken into account in the F1 measure, these are treated more leniently. Note that $\hat{C}$ is not so lenient, as it measures agreement over the sets of atomic phrase pairs rather than between atomic and composite phrase pairs in the F1 measure. This means that under $\hat{C}$, choosing different granularities of phrases will be penalized, but would not have been under the F1 measure.

In Figure 4 we show how $\hat{C}$ varies with sentence length for our three corpora. Specifically, we plot observed agreement $\hat{\pi}$, chance agreement π_0, and $\hat{C}$ against sentence pairs

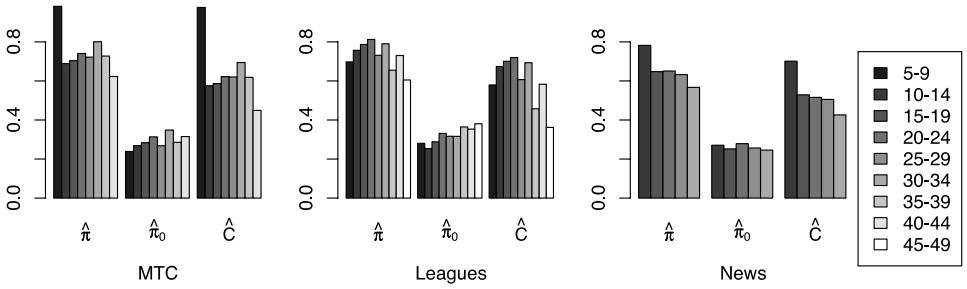

Figure 4
Agreement statistics plotted against sentence length for the three sub-corpora. Each group of three columns correspond to $\hat{\pi}$, $\hat{\pi}_0$, and $\hat{C}$, respectively. The statistics were measured over non-identical phrase pairs using all phrase pairs, atomic and composite.

Table 5
Agreement between automatic Giza++ predicted word alignments and our manually corrected alignments, measured over *atomic phrase pairs*.

MTC			Leagues			News		
Measure	Poss	Sure	Measure	Poss	Sure	Measure	Poss	Sure
Prec	0.58	0.55	Prec	0.63	0.60	Prec	0.63	0.65
Rec	0.42	0.49	Rec	0.39	0.47	Rec	0.50	0.64
F1	0.53		F1	0.54		F1	0.63	

binned by (the shorter) sentence length. In all cases we observe that chance agreement is substantially lower than observed agreement for all sentence lengths. We also see that $\hat{C}$ tends to be higher for shorter sentences. Differences in $\hat{C}$ across sentence lengths are mostly of small magnitude across all three corpora. This indicates that disagreements may be due to other factors, besides sentence length.

Unfortunately, there are no comparable annotation studies that would allow us to gauge the quality of the obtained agreements. The use of precision, recall, and F1 is widespread in SMT, but these measures evaluate automatic alignments against a gold standard, rather than the agreement between two or more annotators (but see Melamed [1998] for an exception). Nevertheless, we would expect the humans to agree more with each other than with Giza++, given that the latter produces many erroneous word alignments and is not specifically tuned to the paraphrasing task. Table 5 shows agreement between one annotator and Giza++ for atomic phrase pairs.[15] We obtained similar results for the other annotator and with the word-based measures. As can be seen, human–Giza++ agreement is much lower than human–human agreement on all three corpora (compare Tables 5 and 4). Taken together the results in Tables 3–5 show a substantial level of agreement, thus indicating that our definition of paraphrases via word alignments can yield reliable annotations. In the following section we discuss how our corpus can be usefully employed in the study of paraphrasing.

15 Note that we cannot meaningfully measure $\hat{C}$ for this data because the Giza++ predictions are already being used to estimate π_0 in our formulation. Consequently, $P(A) = P(B)$ and $\hat{C}$ is zero.

4. Experiments

Our annotated corpus can be used in a number of ways to help paraphrase research: for example, to inform the linguistic analysis of paraphrases, as a training set for the development of discriminative paraphrase systems, and as a test set for the automatic evaluation of computational models. Here, we briefly demonstrate some of these uses.

Paraphrase Modeling. Much previous research has focused on *lexical* paraphrases (but see Lin and Pantel [2001] and Pang, Knight, and Marcu [2003] for exceptions). We argue that our corpus should support a richer range of *structural* (syntactic) paraphrases. To demonstrate this we have extracted paraphrase rules from our annotations using the grammar induction algorithm from Cohn and Lapata (2007). Briefly, the algorithm extracts tree pairs from word-aligned text by choosing aligned constituents in a pair of equivalent sentences. These pairs are then generalized by factoring out aligned subtrees, thereby resulting in synchronous grammar rules (Aho and Ullman 1969) with variable nodes.

We parsed the MTC corpus with Bikel's (2002) parser and extracted synchronous rules from the gold-standard alignments. A sample of these rules are shown in Figure 5. Here we see three lexical paraphrases, followed by five structural paraphrases. In example 4, *also* is replaced with *moreover* and is moved to the start of the sentence from the pre-verbal position. Examples 5–8 show various reordering operations, where the boxed numbers indicate correspondences between non-terminals in the two sides of the rules.

The synchronous rules in Figure 5 provide insight into the process of paraphrasing at the syntactic level, and also a practical means for developing algorithms for paraphrase *generation*—a task which has received little attention to date. For instance, we could envisage a paraphrase model that transforms parse trees of an input sentence into parse trees that represent a sentential paraphrase of that sentence. Our corpus can be used to learn this mapping using discriminative methods (Cowan, Kučerová, and Collins 2006; Cohn and Lapata 2007).

Evaluation Set. As mentioned in Section 1, it is currently difficult to compare competing approaches due to the effort involved in eliciting manual judgments of paraphrase output. Our corpus could fill the role of a gold-standard test set, allowing for automatic evaluation techniques.

Developing measures for automatic paraphrase evaluation is outside the scope of this article. Nevertheless, we illustrate how the corpus can be used for this purpose. For example we could easily measure the precision and recall of an automatic system

1 (NP (CD six) (NNS months)) ↔ (NP (PDT half) (DT a) (NN year))
2 (NP (JJ United) (NNS Nations)) ↔ (NN UN)
3 (ADVP (RB all) (RB over)) ↔ (ADVP (RB completely))
4 (S NP$_{\boxed{0}}$ (ADVP (RB also)) VP$_{\boxed{1}}$.$_{\boxed{2}}$) ↔ (S (ADVP (RB moreover)) (, ,) NP$_{\boxed{0}}$ VP$_{\boxed{1}}$.$_{\boxed{2}}$)
5 (PP IN$_{\boxed{0}}$ (NP DT$_{\boxed{1}}$ NN$_{\boxed{2}}$ (NN issue))) ↔ (NP (NP DT$_{\boxed{1}}$ (NN issue)) (PP IN$_{\boxed{0}}$ NP$_{\boxed{2}}$))
6 (NP NP$_{\boxed{0}}$ (PP IN$_{\boxed{1}}$ (NP DT$_{\boxed{2}}$ NN$_{\boxed{3}}$))) ↔ (PP IN$_{\boxed{1}}$ (NP (NP DT$_{\boxed{2}}$ NN$_{\boxed{3}}$ (POS 's)) NNS$_{\boxed{0}}$))
7 (NP NP$_{\boxed{0}}$ (PP (IN of) NP$_{\boxed{1}}$)) ↔ (NP JJ$_{\boxed{1}}$ NNP$_{\boxed{0}}$)
8 (NP (NP (NP NN$_{\boxed{0}}$ (POS 's)) JJ$_{\boxed{1}}$ NN$_{\boxed{2}}$) S$_{\boxed{3}}$) ↔ (NP (NP (DT the) JJ$_{\boxed{1}}$ NN$_{\boxed{2}}$) (PP (IN by) NP$_{\boxed{0}}$) S$_{\boxed{3}}$)

Figure 5
Synchronous grammar rules extracted from the MTC corpus.

against our annotations. Computing precision and recall for an individual system is not perhaps the most meaningful test, considering the large potential for paraphrasing in a given sentence pair. A better evaluation strategy would include a comparison across many systems on the same corpus. We could then rank these systems without, however, paying so much attention to the absolute precision and recall values. We expect these comparisons to yield relatively low numbers for many reasons. First and foremost the task is hard, as shown by our inter-annotator agreement figures in Tables 3 and 4. Secondly, there may be valid paraphrases that the systems identify but are not listed in our gold standard. Thirdly, systems may have different biases, for example, towards producing more lexical or syntactic paraphrases, but our comparison would not take this into account. Despite all these considerations, we believe that comparison against our corpus would treat these systems on an equal footing against the same materials while factoring out nonessential degrees of freedom inherent in human elicitation studies (e.g., attention span, task familiarity, background).

We evaluated the performance of two systems against our corpus. Our first system is simply Giza++ trained on the 55,615 sentence pairs described in Section 4. The second system uses a co-training-based paraphrase extraction algorithm (Barzilay and McKeown 2001). It was also trained on the MTC part 1 corpus, on the same data set used for Giza++, with its default parameters. For each system, we filtered the predicted paraphrases to just those which match part of a sentence pair in the test set. These paraphrases were then compared to the sure phrase pairs extracted from our manually aligned corpus. Giza++'s precision is 55% and recall 49% (see Table 5). The co-training system obtained a precision of 30% and recall of 16%. To confirm the accuracy of the precision estimate, we performed a human evaluation on a sample of 48 of the predicted paraphrases which were treated as errors. Of these, 63% were confirmed as being incorrect and only 20% were acceptable (the remaining were uncertain). The inter-annotator agreement in Table 4 can be used as an upper bound for precision and recall (precision for Sure phrase pairs is 67% and recall 66%). These results seem to suggest that a hypothetical paraphrase extractor based on automatic word alignments would obtain performance superior to the co-training approach. However, we must bear in mind that the co-training system is highly parametrized and was not specifically tuned to our data set.

5. Conclusions

In this article we have presented a human-annotated paraphrase corpus and argued that it can be usefully employed for the evaluation and modeling of paraphrases. We have defined paraphrases as word alignments in a corpus containing pairs of equivalent sentences and shown that these can be reliably identified by annotators. In measuring agreement, we used the standard measures of precision, recall, and F1, but also proposed a novel formulation of chance-corrected agreement for word (and phrase) alignments. Beyond alignment, our formulation could be applied to other structured tasks including parsing and sequence labeling.

The uses of the corpus are many and varied. It can serve as a test set for evaluating the precision and recall of paraphrase induction systems trained on parallel monolingual corpora. The corpus could be further used to develop new evaluation metrics for paraphrase acquisition or novel paraphrasing models. An exciting avenue for future research concerns paraphrase *prediction*, that is, determining when and how to paraphrase single sentence input. Because our corpus contains paraphrase annotations at the sentence level, it could provide a natural test-bed for prediction algorithms.

Acknowledgments

The authors acknowledge the support of the EPSRC (Cohn, grant GR/T04557/01; Lapata, grant GR/T04540/01), the National Science Foundation (Callison-Burch, grant IIS-071344), and the EuroMatrix project (Callison-Burch) funded by the European Commission (6th Framework Programme). We are grateful to our annotators Vasilis Karaiskos and Tom Segler. Thanks to Regina Barzilay for providing us the output of her system on our data and to the anonymous referees whose feedback helped to substantially improve the present article.

References

Aho, A. V. and J. D. Ullman. 1969. Syntax directed translations and the pushdown assembler. *Journal of Compute System Sciences*, 3(1):37–56.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics*.

Ayan, Necip Fazil and Bonnie J. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Sydney.

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604, Ann Arbor, MI.

Bartko, John J. and William T. Carpenter. 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163(5):307–317.

Barzilay, Regina. 2003. *Information Fusion for Multi-Document Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York, NY.

Barzilay, Regina and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Sapporo.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 16–23, Edmonton.

Barzilay, Regina and Kathy McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse.

Bikel, Daniel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the Human Language Technology Conference*, pages 24–27, San Diego, CA.

Brown, Peter F., Stephen A. Della-Pietra, Vincent J. Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, Chris. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 17–24, New York, NY.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cohn, Trevor and Mirella Lapata. 2007. Large margin synchronous generation and its application to sentence compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*, pages 73–82, Prague.

Cowan, Brooke, Ivona Kučerová, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 232–241, Sydney.

Daumé III, Hal and Daniel Marcu. 2004. A phrase-based HMM approach to document/abstract alignment. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 119–126, Barcelona.

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Dolan, William, 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*, pages 350–356, Geneva.

Duboue, Pablo and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 33–36, New York, NY.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fraser, Alexander and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton.

Kupper, Lawrence L. and Kerry B. Hafner. 1989. On assessing interrater agreement for multiple attribute responses. *Biometrics*, 45(3):957–967.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7(4):342–360.

Martin, Joel, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 67–74, Ann Arbor, MI.

Melamed, I. Dan. 1998. Manual annotation of translational equivalence: The Blinker project. IRCS Technical Report #98-07, University of Pennsylvania, Philadelphia, PA.

Mihalcea, Rada and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–6, Edmonton.

Och, Franz Josef and Hermann Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 1086–1090, Saarbrücken.

Och, Franz Josef and Hermann Ney. 2000b. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Och, Franz Josef, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park, MD.

Pang, Bo, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 181–188, Edmonton.

Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona.

Scott, William A. 1955. Reliability of content analysis: The case of nominal scale. *Public Opinion Quarterly*, 19:127–141.

Siegel, Sidney and N. John Castellan. 1988. *Non Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Zhou, Liang, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference*, pages 447–454, New York, NY.