ParaEval: Using Paraphrases to Evaluate Summaries Automatically

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{liangz, cyl, dragos, hovy} @isi.edu

Abstract

ParaEval is an automated evaluation method for comparing reference and peer summaries. It facilitates a tiered-comparison strategy where recall-oriented global optimal and local greedy searches for paraphrase matching are enabled in the top tiers. We utilize a domain-independent paraphrase table extracted from a large bilingual parallel corpus using methods from Machine Translation (MT). We show that the quality of ParaEval's evaluations, measured by correlating with human judgments, closely resembles that of ROUGE's.

1 Introduction

Content coverage is commonly measured in summary comparison to assess how much information from the reference summary is included in a peer summary. Both manual and automatic methodologies have been used. Naturally, there is a great amount of confidence in manual evaluation since humans can infer, paraphrase, and use world knowledge to relate text units with similar meanings, but which are worded differently. Human efforts are preferred if the evaluation task is easily conducted and managed, and does not need to be performed repeatedly. However, when resources are limited, automated evaluation methods become more desirable.

For years, the summarization community has been actively seeking an automatic evaluation methodology that can be readily applied to various summarization tasks. ROUGE (Lin and Hovy, 2003) has gained popularity due to its simplicity and high correlation with human judgments. Even though validated by high correlations with human judgments gathered from previous Document Understanding Conference (DUC) experiments, current automatic procedures (Lin and Hovy, 2003; Hovy et al., 2005) only employ lexical *n-gram* matching. The lack of support for word or phrase matching that stretches beyond strict lexical matches has limited the expressiveness and utility of these methods. We need a mechanism that supplements literal matching—i.e. paraphrase and synonym—and approximates semantic closeness.

In this paper we present ParaEval, an automatic summarization evaluation method, which facilitates paraphrase matching in an overall three-level comparison strategy. At the top level, favoring higher coverage in reference, we perform an optimal search via dynamic programming to find multi-word to multi-word paraphrase matches between phrases in the reference summary (usually human-written) and those in the peer summary (system-generated). The non-matching fragments from the previous level are then searched by a greedy algorithm to find single-word paraphrase/synonym matches. At the third and the lowest level, we perform literal lexical unigram matching on the remaining texts. This tiered design for summary comparison guarantees at least a ROUGE-1 level of summary content matching if no paraphrases are found.

The first two levels employ a paraphrase table. Since manually created multi-word paraphrases—phrases determined by humans to be paraphrases of one another—are not available in sufficient quantities, we automatically build a paraphrase

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 447–454,
New York, June 2006.

table using methods from the Machine Translation (MT) field. The assumption made in creating this table is that if two English phrases are translated into the same foreign phrase with high probability (shown in the alignment results from a statistically trained alignment algorithm), then the two English phrases are paraphrases of each other.

This paper is organized in the following way: Section 2 introduces previous work in summarization evaluation; Section 3 describes the motivation behind this work; paraphrase acquisition is discussed in Section 4; Section 5 explains in detail our summary comparison mechanism; Section 6 validates ParaEval with human summary judgments; and we conclude and discuss future work in Section 7.

2 Previous Work

There has been considerable work in both manual and automatic summarization evaluations. Three most noticeable efforts in manual evaluation are SEE (Lin and Hovy, 2001), Factoid (Van Halteren and Teufel, 2003), and the Pyramid method (Nenkova and Passonneau, 2004).

SEE provides a user-friendly environment in which human assessors evaluate the quality of system-produced peer summary by comparing it to a reference summary. Summaries are represented by a list of summary units (sentences, clauses, etc.). Assessors can assign full or partial content coverage score to peer summary units in comparison to the corresponding reference summary units. Grammaticality can also be graded unit-wise.

The goal of the Factoid work is to compare the information content of different summaries of the same text and determine the minimum number of summaries, which was shown through experimentation to be 20-30, needed to achieve stable consensus among 50 human-written summaries.

The Pyramid method uses identified consensus—a pyramid of phrases created by annotators—from multiple reference summaries as the gold-standard reference summary. Summary comparisons are performed on Summarization Content Units (SCUs) that are approximately of clause length.

To facilitate fast summarization system design-evaluation cycles, ROUGE was created (Lin and Hovy, 2003). It is an automatic evaluation package that measures a number of *n-gram* co-occurrence statistics between peer and reference summary pairs. ROUGE was inspired by BLEU (Papineni et al., 2001) which was adopted by the machine translation (MT) community for automatic MT evaluation. A problem with ROUGE is that the summary units used in automatic comparison are of fixed length. A more desirable design is to have summary units of variable size. This idea was implemented in the Basic Elements (BE) framework (Hovy et al., 2005) which has not been completed due to its lack of support for paraphrase matching. Both ROUGE and BE have been shown to correlate well with past DUC human summary judgments, despite incorporating only lexical matching on summary units (Lin and Hovy, 2003; Hovy et al., 2005).

SCU1: the crime in question was the Lockerbie {Scotland} bombing
1 [for the Lockerbie bombing]
2 [for blowing up] [over Lockerbie, Scotland]
3 [of bombing] [over Lockerbie, Scotland]
4 [was blown up over Lockerbie, Scotland,]
5 [the bombing of Pan Am Flight 103]
6 [bombing over Lockerbie, Scotland,]
7 [for Lockerbie bombing]
8 [bombing of Pan Am flight 103 over Lockerbie.]
9 [linked to the Lockerbie bombing]
10 [in the Lockerbie bombing case.]

Figure 1. Paraphrases created by Pyramid annotation.

3 Motivation

3.1 Paraphrase Matching

An important difference that separates current manual evaluation methods from their automatic counterparts is that semantic matching of content units is performed by human summary assessors. An essential part of the semantic matching involves paraphrase matching—determining whether phrases worded differently carry the same semantic information. This paraphrase matching process is observed in the Pyramid annotation procedure shown in (Nenkova and Passonneau, 2004) over three summary sets (10 summaries each). In the example shown in Figure 1 (reproduced from Pyramid results), each of the 10 phrases (numbered 1 to 10) extracted from summary sentences carries the same *semantic content* as the overall summary content unit labeled SCU1 does. Each extracted phrase is identified as a summary content unit (SCU). In our work in building an automatic evaluation procedure that enables paraphrase

matching, we aim to automatically identify these 10 phrases as paraphrases of one another.

3.2 Synonymy Relations

Synonym matching and paraphrase matching are often mentioned in the same context in discussions of extending current automated summarization evaluation methods to incorporate the matching of semantic units. While evaluating automatically extracted paraphrases via WordNet (Miller et al., 1990), Barzilay and McKeown (2001) quantitatively validated that synonymy is not the only source of paraphrasing. We envisage that this claim is also valid for summary comparisons.

From an in-depth analysis on the manually created SCUs of the DUC2003 summary set D30042 (Nenkova and Passonneau, 2004), we find that 54.48% of 1746 cases where a non-stop word from one SCU did not match with its supposedly human-aligned pairing SCUs are in need of some level of paraphrase matching support. For example, in the first two extracted SCUs (labeled as 1 and 2) in Figure 1—"for the Lockerbie bombing" and "for blowing up … over Lockerbie, Scotland"—no non-stop word other than the word "Lockerbie" occurs in both phrases. But these two phrases were judged to carry the same semantic meaning because human annotators think the word "bombing" and the phrase "blowing up" refer to the same action, namely the one associated with "explosion." However, "bombing" and "blowing up" cannot be matched through synonymy relations by using WordNet, since one is a noun and the other is a verb phrase (if tagged within context). Even when the search is extended to finding synonyms and hypernyms for their categorical variants and/or using other parts of speech (verb for "bombing" and noun phrase for "blowing up"), a match still cannot be found.

To include paraphrase matching in summary evaluation, a collection of less-strict paraphrases must be created and a matching strategy needs to be investigated.

4 Paraphrase Acquisition

Paraphrases are alternative verbalizations for conveying the same information and are required by many Natural Language Processing (NLP) applications. In particular, summary creation and evaluation methods need to recognize paraphrases and their semantic equivalence. Unfortunately, we have yet to incorporate into the evaluation framework previous findings in paraphrase identification and extraction (Barzilay and McKeown, 2001; Pang et al., 2003; Bannard and Callison-Burch, 2005).

4.1 Related Work on Paraphrasing

Three major approaches in paraphrase collection are manual collection (domain-specific), collection utilizing existing lexical resources (i.e. WordNet), and derivation from corpora. Hermjakob et al. (2002) view paraphrase recognition as reformulation by pattern recognition. Pang et al. (2003) use word lattices as paraphrase representations from semantically equivalent translations sets. Using parallel corpora, Barzilay and McKeown (2001) identify paraphrases from multiple translations of classical novels, where as Bannard and Callison-Burch (2005) develop a probabilistic representation for paraphrases extracted from large Machine Translation (MT) data sets.

4.2 Extracting Paraphrases

Our method to automatically construct a large domain-independent paraphrase collection is based on the assumption that two different English phrases of the same meaning may have the same translation in a foreign language.

Phrase-based Statistical Machine Translation (SMT) systems analyze large quantities of bilingual parallel texts in order to learn translational alignments between pairs of words and phrases in two languages (Och and Ney, 2004). The sentence-based translation model makes word/phrase alignment decisions probabilistically by computing the optimal model parameters with application of the statistical estimation theory. This alignment process results in a corpus of word/phrase-aligned parallel sentences from which we can extract phrase pairs that are translations of each other. We ran the alignment algorithm from (Och and Ney, 2003) on a Chinese-English parallel corpus of 218 million English words. Phrase pairs are extracted by following the method described in (Och and Ney, 2004) where all contiguous phrase pairs having consistent alignments are extraction candidates. The resulting phrase table is of high quality; both the alignment models and phrase extraction meth-

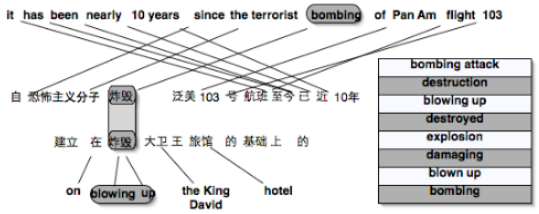


Figure 2. An example of paraphrase extraction.

ods have been shown to produce very good results for SMT. Using these pairs we build paraphrase sets by joining together all English phrases with the same Chinese translation. Figure 2 shows an example word/phrase alignment for two parallel sentence pairs from our corpus where the phrases "blowing up" and "bombing" have the same Chinese translation. On the right side of the figure we show the paraphrase set which contains these two phrases, which is typical in our collection of extracted paraphrases.

5 Summary Comparison in ParaEval

This section describes the process of comparing a peer summary against a reference summary and the summary grading mechanism.

5.1 Description

We adopt a three-tier matching strategy for summary comparison. The score received by a peer summary is the ratio of the number of reference words matched to the total number of words in the reference summary. The total number of matched reference words is the sum of matched words in reference throughout all three tiers. At the top level, favoring high recall coverage, we perform an optimal search to find multi-word paraphrase matches between phrases in the reference summary and those in the peer. Then a greedy search is performed to find single-word paraphrase/synonym matches among the remaining text. Operations conducted in these two top levels are marked as linked rounded rectangles in Figure 3. At the bottom level, we find lexical identity matches, as marked in rectangles in the example. If no paraphrases are found, this last level provides a guarantee of lexical comparison that is equivalent to what other automated systems give. In our system, the bottom level currently performs unigram matching. Thus, we are ensured with at least a ROUGE-1 type of summary comparison. Alternatively, equivalence of other ROUGE configurations can replace the ROUGE-1 implementation.

There is no theoretical reason why the first two levels should not merge. But due to high computational cost in modeling an optimal search, the separation is needed. We explain this in detail below.

5.2 Multi-Word Paraphrase Matching

In this section we describe the algorithm that performs the multi-word paraphrase matching between phrases from reference and peer summaries. Using the example in Figure 3, this algorithm creates the phrases shown in the rounded rectangles and establishes the appropriate links indicating corresponding paraphrase matches.

Problem Description

Measuring content coverage of a peer summary using a single reference summary requires computing the recall score of how much information from the reference summary is included in the peer. A summary unit, either from reference or peer, cannot be matched for more than once. For

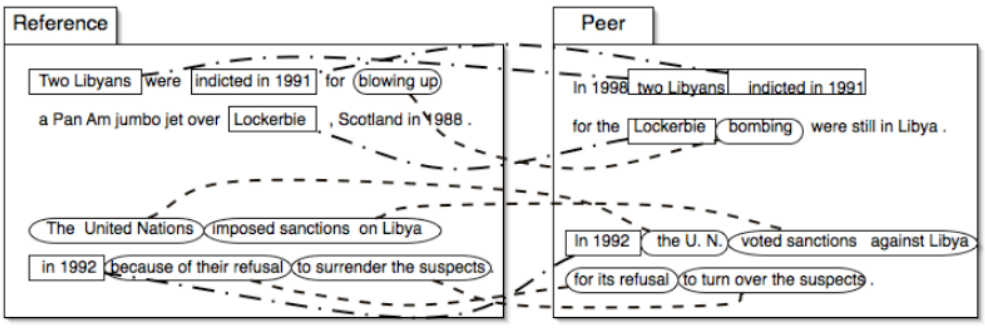


Figure 3. Comparison of summaries.

example, the phrase "imposed sanctions on Libya" (r_1) in Figure 3's reference summary was matched with the peer summary's "voted sanctions against Libya" (p_1). If later in the peer summary there is another phrase p_2 that is also a paraphrase of r_1, the match of r_1 cannot be counted twice. Conversely, double counting is not permissible for phrase/words in the peer summary, either.

We conceptualize the comparison of peer against reference as a task that is to complete over several time intervals. If the reference summary contains n sentences, there will be n time intervals, where at time t_i, phrases from a particular sentence i of the reference summary are being considered with all possible phrases from the peer summary for paraphrase matches. A decision needs to be made at each time interval:

- Do we employ a local greedy match algorithm that is recall generous (preferring more matched words from reference) towards only the reference sentence currently being analyzed,
- Or do we need to explore globally, inspecting all reference sentences and find the best overall matching combinations?

Consider the scenario in Figure 4:

1) at t_0: $L(p_1 = r_2) > L(p_2 = r_1)$ and r_2 contains r_1. A local search algorithm leads to $match(p_1, r_2)$. $L()$ indicates the number of words in reference matched by the peer phrase through paraphrase matching and $match()$ indicates a paraphrase match has occurred (more in the figure).

2) at t_1: $L(p_1 = r_3) > L(p_1 = r_2)$. A global algorithm reverses the decision $match(p_1, r_2)$ made at t_0 and concludes $match(p_1, r_3)$ and $match(p_2, r_1)$. A local search algorithm would have returned no match.

Clearly, the global search algorithm achieves higher overall recall (in words). The matching of paraphrases between a reference and its peer becomes a global optimization problem, maximizing the content coverage of the peer compared in reference.

Solution Model

We use dynamic programming to derive the solution of finding the best paraphrase-matching combinations. The optimization problem is as follows: Sentences from a reference summary and a peer summary can be broken into phrases of various lengths. A paraphrase lookup table is used to find whether a reference phrase and a peer phrase are paraphrases of each other. What is the optimal paraphrase matching combination of phrases from reference and peer that gives the highest recall score (in number of matched reference words) for this given peer? The solution should be recall oriented (favoring a peer phrase that matches more reference words than those match less).

Following (Trick, 1997), the solution can be characterized as:

1) This problem can be divided into n stages corresponding to the n sentences of the reference summary. At each stage, a decision is required to determine the best combination of matched paraphrases between the reference sentence and the entire peer summary that results in no double counting of phrases on the peer side. There is no double counting of reference phrases across stages since we are processing one reference sentence at a time and are finding the best paraphrase matches using the entire peer summary. As long as there is no double counting in peers, we are guaranteed to have none in reference, either.

2) At each stage, we define a number of possible states as follows. If, out of all possible phrases of any length extracted from the reference sentence, m phrases were found to have matching paraphrases in the peer summary, then a state is any subset of the m phrases.

3) Since no double counting in matched phrases/words is allowed in either the reference summary or the peer summary, the decision of which phrases (leftover text segments in reference

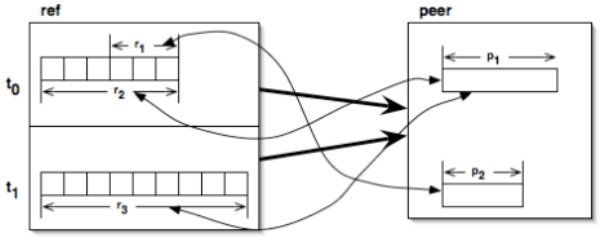


Figure 4. Local vs. global paraphrase matching.

P_j and r_i represent phrases chosen for paraphrase matching from peer and reference respectively.

$P_j = r_i$ indicates that the phrase P_j from peer is found to be a paraphrase to the phrase r_i from reference.

$L(P_j = r_i)$ indicates the number of words matched by P_j in r_i when they are found to be paraphrases of each other.

$L(P_j = r_i)$ and $L(P_j = r_{i+1})$ may not be equal if the number of words in r_i, indicated by $L(r_i)$, does not equal to the number of words in r_{i+1}, indicated by $L(r_{i+1})$.

and in peer) are allowed to match for the next stage is made in the current stage.

4) *Principle of optimality*: at a given state, it is not necessary to know what matches occurred at previous stages, only on the accumulated recall score (matched reference words) from previous stages and what text segments (phrases) in peer have not been taken/matched in previous stages.

5) There exists a recursive relationship that identifies the optimal decision for stage s (out of n total stages), given that stage $s+1$ has already been solved.

6) The final stage, n (last sentence in reference), is solved by choosing the state that has the highest accumulated recall score and yet resulted no double counting in any phrase/word in peer the summary.

Figure 5 demonstrates the optimal solution (12 reference words matched) for the example shown in Figure 4. We can express the calculations in the following formulas:

$$f_1(x_s) = \max_{x_b : c(x_b)} \{r(x_s)\}$$

and

$$f_y(x_b) = \max_{x_b : c(x_b)} \{r(x_b) + f_{y-1}(x_b - c(x_b))\}$$

where $f_y(x_b)$ denotes the optimal recall coverage (number of words in the reference summary matched by the phrases from the peer summary) at state x_b in stage y. $r(x_b)$ is the recall coverage given state x_b. And $c(x_b)$ records the phrases matched in peer with no double counting, given state x_b.

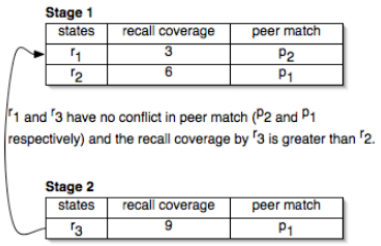


Figure 5. Solution for the example in Figure 4.

5.3 Synonym Matching

All paraphrases whose pairings do not involve multi-word to multi-word matching are called synonyms in our experiment. Since these phrases have either a *n-to-1* or *1-to-n* matching ratio (such as the phrases "blowing up" and "bombing"), a greedy algorithm favoring higher recall coverage reduces the state creation and stage comparison costs associated with the optimal procedure ($\mathbf{O}(m^6)$: $\mathbf{O}(m^3)$ for state creation, and for 2 stages at any time)). The paraphrase table described in Section 4 is used.

Synonym matching is performed only on parts of the reference and peer summaries that were not matched from the multi-word paraphrase-matching phase.

5.4 Lexical Matching

This matching phase performs straightforward lexical matching, as exemplified by the text fragments marked in rectangles in Figure 3. Unigrams are used as the units for counting matches in accordance with the previous two matching phases.

During all three matching phases, we employed a ROUGE-1 style of counting. Other alternatives, such as ROUGE-2, ROUGE-SU4, etc., can easily be adapted to each phase.

6 Evaluation of ParaEval

To evaluate and validate the effectiveness of an automatic evaluation metric, it is necessary to show that automatic evaluations correlate with human assessments highly, positively, and consistently (Lin and Hovy, 2003). In other words, an automatic evaluation procedure should be able to distinguish good and bad summarization systems by assigning scores with close resemblance to humans' assessments.

6.1 Document Understanding Conference

The Document Understanding Conference has provided large-scale evaluations on both human-created and system-generated summaries annually. Research teams are invited to participate in solving summarization problems with their systems. System-generated summaries are then assessed by humans and/or automatic evaluation procedures. The collection of human judgments on systems and their summaries has provided a test-bed for developing and validating automated summary grading methods (Lin and Hovy, 2003; Hovy et al., 2005).

The correlations reported by ROUGE and BE show that the evaluation correlations between these two systems and DUC human evaluations are much higher on single-document summarization tasks. One possible explanation is that when sum-

marizing from only one source (text), both human- and system-generated summaries are mostly extractive. The reason for humans to take phrases (or maybe even sentences) verbatim is that there is less motivation to abstract when the input is not highly redundant, in contrast to input for multi-document summarization tasks, which we speculate allows more abstracting. ROUGE and BE both facilitate lexical *n-gram* matching, hence, achieving amazing correlations. Since our baseline matching strategy is lexically based when paraphrase matching is not activated, validation on single-doc summarization results is not repeated in our experiment.

6.2 Validation and Discussion

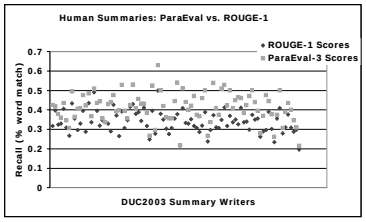
We use summary judgments from DUC2003's multi-document summarization (MDS) task to evaluate ParaEval. During DUC2003, participating systems created short summaries (~100 words) for 30 document sets. For each set, one assessor-written summary was used as the reference to compare peer summaries created by 18 automatic systems (including baselines) and 3 other human-written summaries. A system ranking was produced by taking the averaged performance on all summaries created by systems. This evaluation process is replicated in our validation setup for ParaEval. In all, 630 summary pairs were compared. Pearson's correlation coefficient is computed for the validation tests, using DUC2003 assessors' results as the gold standard.

Table 1 illustrates the correlation figures from the DUC2003 test set. ParaEval-para_only shows the correlation result when using only paraphrase and synonym matching, without the baseline unigram matching. ParaEval-2 uses multi-word paraphrase matching and unigram matching, omitting the greedy synonym-matching phrase. ParaEval-3 incorporates matching at all three granularity levels.

We see that the current implementation of ParaEval closely resembles the way ROUGE-1 differentiates system-generated summaries. We believe this is due to the identical calculations of recall scores. The score that a peer summary receives from ParaEval depends on the number of words matched in the reference summary from its paraphrase, synonym, and unigram matches. The counting of individual words in reference indicates a ROUGE-1 design in grading. However, a detailed examination on individual reference-peer comparisons shows that paraphrase and synonym comparisons and matches, in addition to lexical *n-gram* matching, do measure a higher level of content coverage. This is demonstrated in Figure 6a and b. Strict unigram matching reflects the content retained by a peer summary mostly in the 0.2-0.4 ranges in recall, shown as dark-colored dots in the graphs. Allowing paraphrase and synonym matching increases the detection of peer coverage to the range of 0.3-0.5, shown as light-colored dots.

DUC-2003	Pearson
ROUGE-1	0.622
ParaEval-para_only	0.41
ParaEval-2	0.651
ParaEval-3	0.657

Table 1. Correlation with DUC 2003 MDS results.



a). Human-written summaries.

b). System-generated summaries.

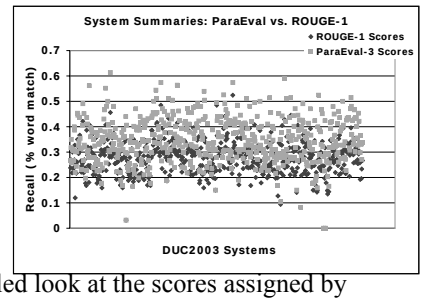


Figure 6. A detailed look at the scores assigned by lexical and paraphrase/synonym comparisons.

We conducted a manual evaluation to further examine the paraphrases being matched. Using 10 summaries from the Pyramid data, we asked three human subjects to judge the validity of 128 (randomly selected) paraphrase pairs extracted and identified by ParaEval. Each pair of paraphrases was coupled with its respective sentences as contexts. All paraphrases judged were multi-word. ParaEval received an average precision of 68.0%. The complete agreement between judges is 0.582 according to the Kappa coefficient (Cohen, 1960). In Figure 7, we show two examples that the human judges consider to be good paraphrases produced and matched by ParaEval. Judges voiced difficul-

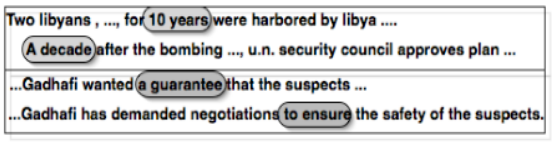


Figure 7. Paraphrases matched by ParaEval.

ties in determining "semantic equivalence." There were cases where paraphrases would be generally interchangeable but could not be matched because of non-semantic equivalence in their contexts. And there were paraphrases that were determined as matches, but if taken out of context, would not be direct replacements of each other. These two situations are where the judges mostly disagreed.

7 Conclusion and Future Work

In this paper, we have described an automatic summarization evaluation method, ParaEval, that facilitates paraphrase matching using a large domain-independent paraphrase table extracted from a bilingual parallel corpus. The three-layer matching strategy guarantees a ROUGE-like baseline comparison if paraphrase matching fails.

The paraphrase extraction module from the current implementation of ParaEval does not discriminate among the phrases that are found to be paraphrases of one another. We wish to incorporate the probabilistic paraphrase extraction model from (Bannard and Callison-Burch, 2005) to better approximate the relations between paraphrases. This adaptation will also lead to a stochastic model for the low-level lexical matching and scoring.

We chose English-Chinese MT parallel data because they are news-oriented which coincides with the task genre from DUC. However, it is unknown how large a parallel corpus is sufficient in providing a paraphrase collection good enough to help the evaluation process. The quality of the paraphrase table is also affected by changes in the domain and language pair of the MT parallel data. We plan to use ParaEval to investigate the impact of these changes on paraphrase quality under the assumption that better paraphrase collections lead to better summary evaluation results.

The immediate impact and continuation of the described work would be to incorporate paraphrase matching and extraction into the summary creation process. And with ParaEval, it is possible for us to evaluate systems that do incorporate some level of abstraction, especially paraphrasing.

References

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. *Proceedings of ACL-2005*.

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

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263—311, 1993.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Education and Psychological Measurement*, 43(6):37—46.

Diab, M. and P. Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. *Proceedings of ACL-2002*.

DUC. 2001—2005. Document Understanding Conferences.

Hermjakob, U., A. Echihabi, and D. Marcu. 2002. Natural language based reformulation resource and web exploitation for question answering. *Proceedings of TREC-2002*.

Hovy, E, C.Y. Lin, and L. Zhou. 2005. Evaluating DUC 2005 using basic elements. *Proceedings of DUC-2005*.

Hovy, E., C.Y. Lin, L. Zhou, and J. Fukumoto. 2005a. Basic Elements. http://www.isi.edu/~cyl/BE.

Lin, C.Y. 2001. http://www.isi.edu/~cyl/SEE.

Lin, C.Y. and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the HLT-2003*.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): 235—245.

Nenkova, A. and R. Passonneau. 2004. Evaluating content selection in summarization: the pyramid method. *Proceedings of the HLT-NAACL 2004*.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-—51, 2003.

Och, F. J. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), 2004.

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

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. IBM research report Bleu: a method for automatic evaluation of machine translation *IBM Research Division Technical Report*, RC22176, 2001.

Trick, M. A. 1997. A tutorial on dynamic programming.http://mat.gsia.cmu.edu/classes/dynamic/dynamic.html.

Van Halteren, H. and S. Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. *Proceedings of HLT-2003*.