In Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003), May 27 – June 1, 2003, Edmonton, Canada

Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics

Chin-Yew Lin and Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292
{cyl,hovy}@isi.edu

Abstract

Following the recent adoption by the machine translation community of automatic evaluation using the BLEU/NIST scoring process, we conduct an in-depth study of a similar idea for evaluating summaries. The results show that automatic evaluation using unigram co-occurrences between summary pairs correlates surprising well with human evaluations, based on various statistical metrics; while direct application of the BLEU evaluation procedure does not always give good results.

1 Introduction

Automated text summarization has drawn a lot of interest in the natural language processing and information retrieval communities in the recent years. A series of workshops on automatic text summarization (WAS 2000, 2001, 2002), special topic sessions in ACL, COLING, and SIGIR, and government sponsored evaluation efforts in the United States (DUC 2002) and Japan (Fukusima and Okumura 2001) have advanced the technology and produced a couple of experimental online systems (Radev et al. 2001, McKeown et al. 2002). Despite these efforts, however, there are no common, convenient, and repeatable evaluation methods that can be easily applied to support system development and just-in-time comparison among different summarization methods.

The Document Understanding Conference (DUC 2002) run by the National Institute of Standards and Technology (NIST) sets out to address this problem by providing annual large scale common evaluations in text summarization. However, these evaluations involve human judges and hence are subject to variability (Rath et al. 1961). For example, Lin and Hovy (2002) pointed out that 18% of the data contained multiple judgments in the DUC 2001 single document evaluation[1].

To further progress in automatic summarization, in this paper we conduct an in-depth study of automatic evaluation methods based on n-gram co-occurrence in the context of DUC. Due to the setup in DUC, the evaluations we discussed here are intrinsic evaluations (Spärck Jones and Galliers 1996). Section 2 gives an overview of the evaluation procedure used in DUC. Section 3 discusses the IBM BLEU (Papineni et al. 2001) and NIST (2002) n-gram co-occurrence scoring procedures and the application of a similar idea in evaluating summaries. Section 4 compares n-gram co-occurrence scoring procedures in terms of their correlation to human results and on the recall and precision of statistical significance prediction. Section 5 concludes this paper and discusses future directions.

2 Document Understanding Conference

The 2002 Document Understanding Conference[2] included the follow two main tasks:

- Fully automatic single-document summarization: given a document, participants were required to create a generic 100-word summary. The training set comprised 30 sets of approximately 10 documents each, together with their 100-word human written summaries. The test set comprised 30 unseen documents.
- Fully automatic multi-document summarization: given a set of documents about a single subject, participants were required to create 4 generic summaries of the entire set, containing 50, 100, 200, and 400 words respectively. The document sets were of four types: a single natural disaster event; a

[1] Multiple judgments occur when more than one performance score is given to the same system (or human) and human summary pairs by the same human judge.

[2] DUC 2001 and DUC 2002 have similar tasks, but summaries of 10, 50, 100, and 200 words are requested in the multi-document task in DUC 2002.

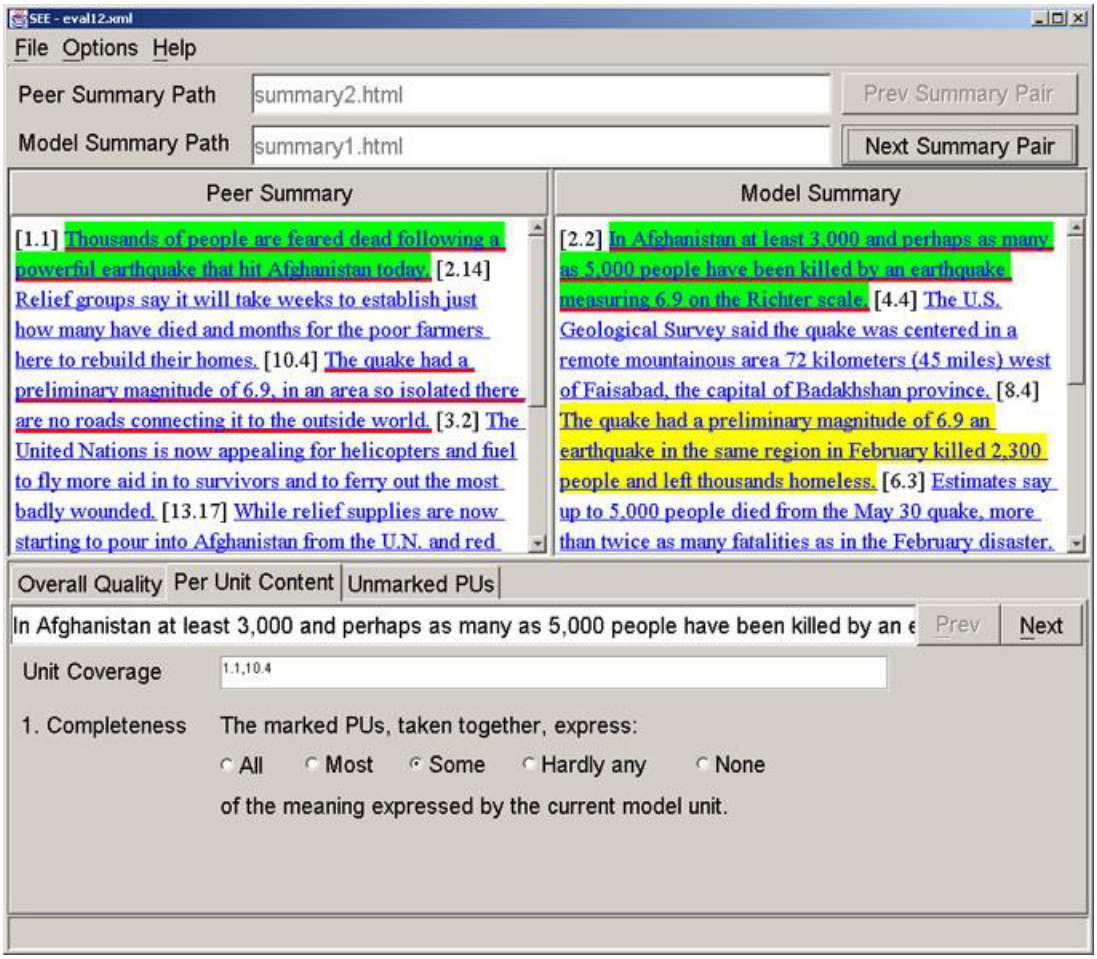

Figure 1. SEE in an evaluation session.

single event; multiple instances of a type of event; and information about an individual. The training set comprised 30 sets of approximately 10 documents, each provided with their 50, 100, 200, and 400-word human written summaries. The test set comprised 30 unseen sets.

A total of 11 systems participated in the single-document summarization task and 12 systems participated in the multi-document task.

2.1 Evaluation Materials

For each document or document set, one human summary was created as the 'ideal' model summary at each specified length. Two other human summaries were also created at each length. In addition, baseline summaries were created automatically for each length as reference points. For the multi-document summarization task, one baseline, *lead baseline*, took the first 50, 100, 200, and 400 words in the last document in the collection. A second baseline, *coverage baseline*, took the first sentence in the first document, the first sentence in the second document and so on until it had a summary of 50, 100, 200, or 400 words. Only one baseline (baseline1) was created for the single document summarization task.

2.2 Summary Evaluation Environment

To evaluate system performance NIST assessors who created the 'ideal' written summaries did pairwise comparisons of their summaries to the system-generated summaries, other assessors' summaries, and baseline summaries. They used the Summary Evaluation Environment (SEE) 2.0 developed by (Lin 2001) to support the process. Using SEE, the assessors compared the system's text (the *peer* text) to the ideal (the *model* text). As shown in Figure 1, each text was decomposed into a list of units and displayed in separate windows.

SEE 2.0 provides interfaces for assessors to judge both the content and the quality of summaries. To measure content, assessors step through each model unit, mark all system units sharing content with the current model unit (green/dark gray highlight in the model summary window), and specify that the marked system units express *all, most, some*, or *hardly any* of the content of the

current model unit. To measure quality, assessors rate grammaticality[3], cohesion[4], and coherence[5] at five different levels: *all, most, some, hardly any*, or *none*[6]. For example, as shown in Figure 1, an assessor marked system units 1.1 and 10.4 (red/dark underlines in the left pane) as sharing *some* content with the current model unit 2.2 (highlighted green/dark gray in the right).

2.3 Evaluation Metrics

Recall at different compression ratios has been used in summarization research to measure how well an automatic system retains important content of original documents (Mani et al. 1998). However, the simple sentence recall measure cannot differentiate system performance appropriately, as is pointed out by Donaway et al. (2000). Therefore, instead of pure sentence recall score, we use coverage score *C*. We define it as follows[7]:

$$C = \frac{(\text{Number of MUs marked}) \bullet E}{\text{Total number of MUs in the model summary}} \quad (1)$$

E, the ratio of completeness, ranges from 1 to 0: 1 for *all*, 3/4 for *most*, 1/2 for *some*, 1/4 for *hardly any*, and 0 for *none*. If we ignore *E* (set it to 1), we obtain simple sentence recall score. We use average coverage scores derived from human judgments as the references to evaluate various automatic scoring methods in the following sections.

3 BLEU and N-gram Co-Occurrence

To automatically evaluate machine translations the machine translation community recently adopted an n-gram co-occurrence scoring procedure BLEU (Papineni et al. 2001). The NIST (NIST 2002) scoring metric is based on BLEU. The main idea of BLEU is to measure the translation closeness between a candidate translation and a set of reference translations with a numerical metric. To achieve this goal, they used a weighted average of variable length n-gram matches between system translations and a set of human reference translations and showed that a weighted average metric, i.e. BLEU, correlating highly with human assessments.

Similarly, following the BLEU idea, we assume that the closer an automatic summary to a professional human summary, the better it is. The question is: "Can we apply BLEU directly without any modifications to evaluate summaries as well?". We first ran IBM's BLEU evaluation script unmodified over the DUC 2001 model and peer summary set. The resulting Spearman rank order correlation coefficient (ρ) between BLEU and the human assessment for the single document task is 0.66 using one reference summary and 0.82 using three reference summaries; while Spearman ρ for the multi-document task is 0.67 using one reference and 0.70 using three. These numbers indicate that they positively correlate at $\alpha = 0.01$[8]. Therefore, BLEU seems a promising automatic scoring metric for summary evaluation.

According to Papineni et al. (2001), BLEU is essentially a precision metric. It measures how well a machine translation overlaps with multiple human translations using n-gram co-occurrence statistics. N-gram precision in BLEU is computed as follows:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \quad (2)$$

Where $Count_{clip}$(*n-gram*) is the maximum number of *n-gram*s co-occurring in a candidate translation and a reference translation, and *Count*(*n-gram*) is the number of *n-gram*s in the candidate translation. To prevent very short translations that try to maximize their precision scores, BLEU adds a brevity penalty, *BP*, to the formula:

$$BP = \begin{cases} 1 & if\ |c| > |r| \\ e^{(1-|r|/|c|)} & if\ |c| \le |r| \end{cases} \quad (3)$$

Where |*c*| is the length of the candidate translation and |*r*| is the length of the reference translation. The BLEU formula is then written as follows:

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (4)$$

N is set at 4 and w_n, the weighting factor, is set at 1/*N*. For summaries by analogy, we can express equation (1) in terms of n-gram matches following equation (2):

$$C_n = \frac{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count_{match}(n-gram)}{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count(n-gram)} \quad (5)$$

Where $Count_{match}$(*n-gram*) is the maximum number of *n-gram*s co-occurring in a peer summary and a model unit and *Count*(*n-gram*) is the number of *n-gram*s in the model unit. Notice that the average n-gram coverage score, C_n, as shown in equation 5 is a recall metric

[3] Does the summary observe English grammatical rules independent of its content?

[4] Do sentences in the summary fit in with their surrounding sentences?

[5] Is the content of the summary expressed and organized in an effective way?

[6] These category labels are changed to numerical values of 100%, 80%, 60%, 40%, 20%, and 0% in DUC 2002.

[7] DUC 2002 uses a length adjusted version of coverage metric *C'*, where $C' = \alpha * C + (1-\alpha) * B$. B is the brevity and α is a parameter reflecting relative importance (DUC 2002).

[8] The number of instances is 14 (11 systems, 2 humans, and 1 baseline) for the single document task and is 16 (12 systems, 2 humans, and 2 baselines) for the multi-document task.

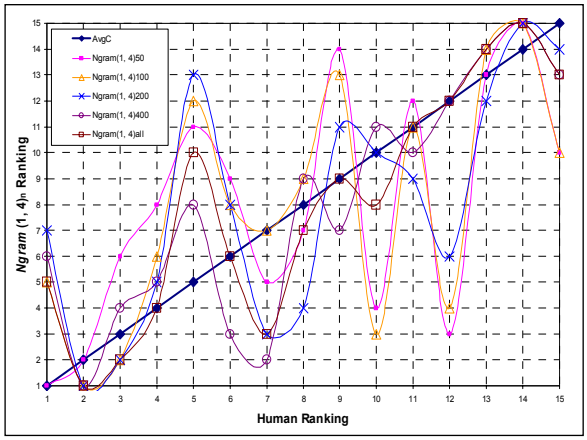

Figure 2. Scatter plot of $Ngram(1,4)_n$ score rankings versus human ranking for the multi-document task data from DUC 2001. The same system is at each vertical line with ranking given by different $Ngram(1,4)_n$ scores. The straight line (AvgC) is the human ranking and n marks summaries of different sizes. $Ngram(1,4)_{all}$ combines results from all sizes.

instead of a precision one as p_n. Since the denominator of equation 5 is the total sum of the number of *n-gram*s occurring at the model summary side instead of the peer side and only one model summary is used for each evaluation; while there could be multiple references used in BLEU and $Count_{clip}(n\text{-}gram)$ could come from matching different reference translations. Furthermore, instead of a brevity penalty that punishes overly short translations, a brevity bonus, *BB*, should be awarded to shorter summaries that contain equivalent content. In fact, a length adjusted average coverage score was used as an alternative performance metric in DUC 2002. However, we set the brevity bonus (or penalty) to 1 for all our experiments in this paper. In summary, the n-gram co-occurrence statistics we use in the following sections are based on the following formula:

$$Ngram(i,j) = BB \bullet \exp\left(\sum_{n=i}^{j} w_n \log C_n\right) \quad (6)$$

Where $j \geq i$, i and j range from 1 to 4, and w_n is $1/(j-i+1)$. $Ngram(1, 4)$ is a weighted variable length n-gram match score similar to the IBM BLEU score; while $Ngram(k, k)$, i.e. $i = j = k$, is simply the average k-gram coverage score C_k.

With these formulas, we describe how to evaluate them in the next section.

4 Evaluations of N-gram Co-Occurrence Metrics

In order to evaluate the effectiveness of automatic evaluation metrics, we propose two criteria:

	SD-100	MD-All	MD-50	MD-100	MD-200	MD-400
SX	0.604	0.875	0.546	0.575	0.775	0.861
S	0.615	0.832	0.646	0.529	0.814	0.843

Table 1. Spearman rank order correlation coefficients of different DUC 2001 data between $Ngram(1, 4)_n$ rankings and human rankings including (S) and excluding (SX) stopwords. SD-100 is for single document summaries of 100 words and MD-50, 100, 200, and 400 are for multi-document summaries of 50, 100, 200, and 400 words. MD-All averages results from summaries of all sizes.

1. Automatic evaluations should correlate highly, positively, and consistently with human assessments.
2. The statistical significance of automatic evaluations should be a good predictor of the statistical significance of human assessments with high reliability.

The first criterion ensures whenever a human recognizes a good summary/translation/system, an automatic evaluation will do the same with high probability. This enables us to use an automatic evaluation procedure in place of human assessments to compare system performance, as in the NIST MT evaluations (NIST 2002). The second criterion is critical in interpreting the significance of automatic evaluation results. For example, if an automatic evaluation shows there is a significant difference between run A and run B at $\alpha = 0.05$ using the z-test (t-test or bootstrap resampling), how does this translate to "real" significance, i.e. the statistical significance in a human assessment of run A and run B? Ideally, we would like there to be a positive correlation between them. If this can be asserted with strong reliability (high recall and precision), then we can use the automatic evaluation to assist system development and to be reasonably sure that we have made progress.

4.1 Correlation with Human Assessments

As stated in Section 3, direct application of BLEU on the DUC 2001 data showed promising results. However, BLEU is a precision-based metric while the human evaluation protocol in DUC is essentially recall-based. We therefore prefer the metric given by equation 6 and use it in all our experiments. Using DUC 2001 data, we compute average $Ngram(1,4)$ scores for each peer system at different summary sizes and rank systems according to their scores. We then compare the $Ngram(1,4)$ ranking with the human ranking. Figure 2 shows the result of DUC 2001 multi-document data. Stopwords are ignored during the computation of $Ngram(1,4)$ scores and words are stemmed using a Porter stemmer (Porter 1980). The x-axis is the human ranking and the y-axis gives the corresponding $Ngram(1,4)$ rankings for summaries of difference sizes. The straight line marked by AvgC is the ranking given by human assessment. For example, a system at (5,8)

means that human ranks its performance at the 5th rank while $Ngram(1,4)_{400}$ ranks it at the 8th. If an automatic ranking fully matches the human ranking, its plot will coincide with the heavy diagonal. A line with less deviation from the heavy diagonal line indicates better correlation with the human assessment.

To quantify the correlation, we compute the Spearman rank order correlation coefficient (ρ) for each $N\text{-}gram(1,4)_n$ run at different summary sizes (n). We also test the effect of inclusion or exclusion of stopwords. The results are summarized in Table 1.

Although these results are statistically significant (α = 0.025) and are comparable to IBM BLEU's correlation figures shown in Section 3, they are not consistent across summary sizes and tasks. For example, the correlations of the single document task are at the 60% level; while they range from 50% to 80% for the multi-document task. The inclusion or exclusion of stopwords also shows mixed results. In order to meet the requirement of the first criterion stated in Section 3, we need better results.

The $Ngram(1,4)_n$ score is a weighted average of variable length n-gram matches. By taking a log sum of the n-gram matches, the $Ngram(1,4)_n$ favors match of longer n-grams. For example, if "*United States of America*" occurs in a reference summary, while one peer summary, *A*, uses "*United States*" and another summary, *B*, uses the full phrase "*United States of America*", summary *B* gets more contribution to its overall score simply due to the longer version of the name. However, intuitively one should prefer a short version of the name in summarization. Therefore, we need to change the weighting scheme to not penalize or even reward shorter equivalents. We conduct experiments to understand the effect of individual n-gram co-occurrence scores in approximating human assessments. Tables 2 and 3 show the results of these runs without and with stopwords respectively.

For each set of DUC 2001 data, single document 100-word summarization task, multi-document 50, 100, 200, and 400 -word summarization tasks, we compute 4 different correlation statistics: Spearman rank order correlation coefficient (Spearman ρ), linear regression *t*-test (LR_t, 11 degree of freedom for single document task and 13 degree of freedom for multi-document task), Pearson product moment coefficient of correlation (Pearson ρ), and coefficient of determination (CD) for each $Ngram(i,j)$ evaluation metric. Among them Spearman ρ is a nonparametric test, a higher number indicates higher correlation; while the other three tests are parametric tests. Higher LR_t, Pearson ρ, and CD also suggests higher linear correlation.

Analyzing all runs according to Tables 2 and 3, we make the following observations:

(1) Simple unigram, $Ngram(1,1)$, and bi-gram, $Ngram(2,2)$, co-occurrence statistics consistently outperform ($0.99 \geq$ Spearman $\rho \geq 0.75$) the weighted average of n-gram of variable length $Ngram(1, 4)$ ($0.88 \geq$ Spearman $\rho \geq 0.55$) in single and multiple document tasks when stopwords are ignored. Importantly, unigram performs especially well with Spearman ρ ranging from 0.88 to 0.99 that is better than the best case in which weighted average of variable length n-gram matches is used and is consistent across different data sets.

(2) The performance of weighted average n-gram scores is in the range between bi-gram and tri-gram co-occurrence scores. This might suggest some summaries are over-penalized by the weighted average metric due to the lack of longer n-gram matches. For example, given a model string "*United States, Japan, and Taiwan*", a candidate

		Ngram (1,4)	*Ngram* (1,1)	*Ngram* (2,2)	*Ngram* (3,3)	*Ngram* (4,4)
Single Doc 100	Spearman ρ	0.604	0.989	0.868	0.527	0.505
	LR_t	1.025	7.130	2.444	0.704	0.053
	Pearson ρ	0.295	0.907	0.593	0.208	0.016
	CD	0.087	0.822	0.352	0.043	0.000
Multi-Doc All	Spearman ρ	0.875	0.993	0.950	0.782	0.736
	LR_t	3.910	13.230	5.830	3.356	2.480
	Pearson ρ	0.735	0.965	0.851	0.681	0.567
	CD	0.540	0.931	0.723	0.464	0.321
Multi-Doc 50	Spearman ρ	0.546	0.879	0.746	0.496	0.343
	LR_t	2.142	5.681	3.350	2.846	2.664
	Pearson ρ	0.511	0.844	0.681	0.620	0.594
	CD	0.261	0.713	0.463	0.384	0.353
Multi-Doc 100	Spearman ρ	0.575	0.896	0.761	0.543	0.468
	LR_t	2.369	7.873	3.641	1.828	1.385
	Pearson ρ	0.549	0.909	0.711	0.452	0.359
	CD	0.301	0.827	0.505	0.204	0.129
Multi-Doc 200	Spearman ρ	0.775	0.979	0.904	0.782	0.754
	LR_t	3.243	15.648	4.929	2.772	2.126
	Pearson ρ	0.669	0.974	0.807	0.609	0.508
	CD	0.447	0.950	0.651	0.371	0.258
Multi-Doc 400	Spearman ρ	0.861	0.982	0.961	0.854	0.661
	LR_t	4.390	10.569	6.409	3.907	2.755
	Pearson ρ	0.773	0.946	0.872	0.735	0.607
	CD	0.597	0.896	0.760	0.540	0.369

Table 2. Various $Ngram(i,j)$ rank/score correlations for 4 different statistics (without stopwords): Spearman rank order coefficient correlation (Spearman ρ), linear regression *t*-test (LR_t), Pearson product moment coefficient of correlation (Pearson ρ), and coefficient of determination (CD).

		Ngram (1,4)	*Ngram* (1,1)	*Ngram* (2,2)	*Ngram* (3,3)	*Ngram* (4,4)
Single Doc 100	Spearman ρ	0.615	0.951	0.863	0.615	0.533
	LR_t	1.076	4.873	2.228	0.942	0.246
	Pearson ρ	0.309	0.827	0.558	0.273	0.074
	CD	0.095	0.683	0.311	0.075	0.005
Multi-Doc All	Spearman ρ	0.832	0.918	0.936	0.832	0.732
	LR_t	3.752	6.489	5.451	3.745	2.640
	Pearson ρ	0.721	0.874	0.834	0.720	0.591
	CD	0.520	0.764	0.696	0.519	0.349
Multi-Doc 50	Spearman ρ	0.646	0.586	0.650	0.589	0.600
	LR_t	2.611	2.527	2.805	2.314	1.691
	Pearson ρ	0.587	0.574	0.614	0.540	0.425
	CD	0.344	0.329	0.377	0.292	0.180
Multi-Doc 100	Spearman ρ	0.529	0.636	0.625	0.571	0.468
	LR_t	2.015	3.338	2.890	2.039	1.310
	Pearson ρ	0.488	0.679	0.625	0.492	0.342
	CD	0.238	0.462	0.391	0.242	0.117
Multi-Doc 200	Spearman ρ	0.814	0.964	0.879	0.814	0.746
	LR_t	3.204	10.134	4.926	3.328	2.173
	Pearson ρ	0.664	0.942	0.807	0.678	0.516
	CD	0.441	0.888	0.651	0.460	0.266
Multi-Doc 400	Spearman ρ	0.843	0.914	0.946	0.857	0.721
	LR_t	4.344	5.358	6.344	4.328	3.066
	Pearson ρ	0.769	0.830	0.869	0.768	0.648
	CD	0.592	0.688	0.756	0.590	0.420

Table 3. Various $Ngram(i, j)$ rank/score correlations for 4 different statistics (with stopwords).

string "*United States, Taiwan, and Japan*" has a unigram score of 1, bi-gram score of 0.5, and tri-gram and 4-gram scores of 0 when the stopword "*and*" is ignored. The weighted average n-gram score for the candidate string is 0.

(3) Excluding stopwords in computing n-gram co-occurrence statistics generally achieves better correlation than including stopwords.

4.2 Statistical Significance of N-gram Co-Occurrence Scores versus Human Assessments

We have shown that simple unigram, *Ngram*(1,1), or bi-gram, *Ngram*(2,2), co-occurrence statistics based on equation 6 outperform the weighted average of n-gram matches, *Ngram*(1,4), in the previous section. To examine how well the statistical significance in the automatic *Ngram*(*i*,*j*) metrics translates to real significance when human assessments are involved, we set up the following test procedures:

(1) Compute pairwise statistical significance test such as *z*-test or *t*-test for a system pair (X,Y) at certain α level, for example α = 0.05, using automatic metrics and human assigned scores.

(2) Count the number of cases a *z*-test indicates there is a significant difference between X and Y based on the automatic metric. Call this number N_{As}.

(3) Count the number of cases a *z*-test indicates there is a significant difference between X and Y based on the human assessment. Call this number N_{Hs}.

(4) Count the cases when an automatic metric predicts a significant difference and the human assessment also does. Call this N_{hit}. For example, if a *z*-test indicates system X is significantly different from Y with α = 0.05 based on the automatic metric scores and the corresponding *z*-test also suggests the same based on the human agreement, then we have a hit.

(5) Compute the recall and precision using the following formulas:

$$\text{recall} = \frac{N_{hit}}{N_{Hs}}$$

$$\text{precision} = \frac{N_{hit}}{N_{As}}$$

A good automatic metric should have high recall and precision. This implies that if a statistical test indicates a significant difference between two runs using the automatic metric then very probably there is also a significant difference in the manual evaluation. This would be very useful during the system development cycle to gauge if an improvement is really significant or not.

Figure 3 shows the recall and precision curves for the DUC 2001 single document task at different α levels and Figure 4 is for the multi-document task with different summary sizes. Both of them exclude stopwords. We use *z*-test in all the significance tests with α level at 0.10, 0.05, 0.25, 0.01, and 0.005.

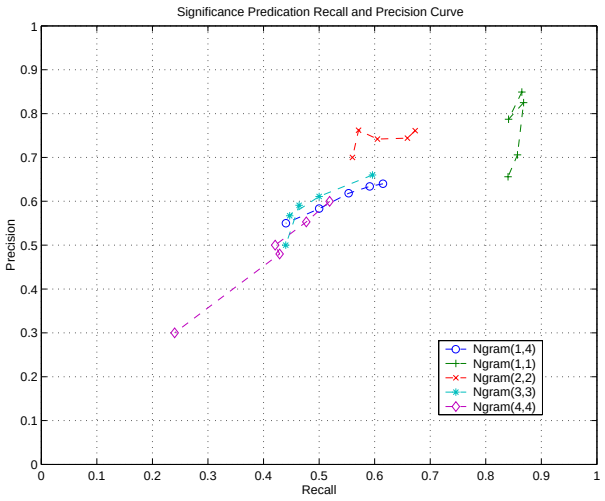

Figure 3. Recall and precision curves of N-gram co-occurrence statistics versus human assessment for DUC 2001 single document task. The 5 points on each curve represent values for the 5 α levels.

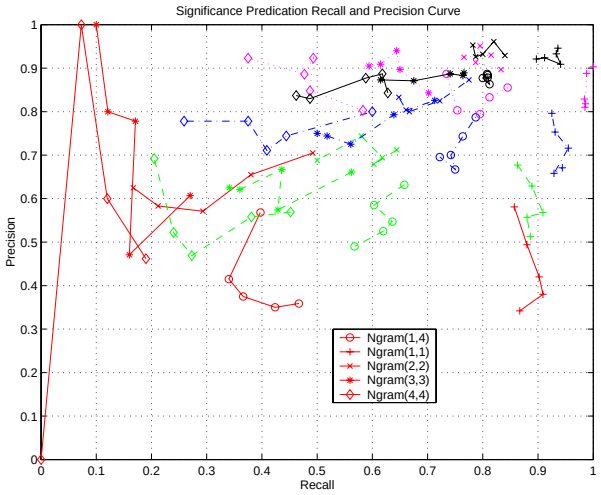

Figure 4. Recall and precision curves of N-gram co-occurrence statistics versus human assessment for DUC 2001 multi-document task. Dark (black) solid lines are for average of all summary sizes, light (red) solid lines are for 50-word summaries, dashed (green) lines are for 100-word summaries, dash-dot lines (blue) are for 200-word summaries, and dotted (magenta) lines are for 400-word summaries.

From Figures 3 and 4, we can see *Ngram*(1,1) and *Ngram*(2,2) reside on the upper right corner of the recall and precision graphs. *Ngram*(1,1) has the best overall behavior. These graphs confirm *Ngram*(1,1) (simple

unigram) is a good automatic scoring metric with good statistical significance prediction power.

5 Conclusions

In this paper, we gave a brief introduction of the manual summary evaluation protocol used in the Document Understanding Conference. We then discussed the IBM BLEU MT evaluation metric, its application to summary evaluation, and the difference between precision-based BLEU translation evaluation and recall-based DUC summary evaluation. The discrepancy led us to examine the effectiveness of individual n-gram co-occurrence statistics as a substitute for expensive and error-prone manual evaluation of summaries. To evaluate the performance of automatic scoring metrics, we proposed two test criteria. One was to make sure system rankings produced by automatic scoring metrics were similar to human rankings. This was quantified by Spearman's rank order correlation coefficient and three other parametric correlation coefficients. Another was to compare the statistical significance test results between automatic scoring metrics and human assessments. We used recall and precision of the agreement between the test statistics results to identify good automatic scoring metrics.

According to our experiments, we found that unigram co-occurrence statistics is a good automatic scoring metric. It consistently correlated highly with human assessments and had high recall and precision in significance test with manual evaluation results. In contrast, the weighted average of variable length n-gram matches derived from IBM BLEU did not always give good correlation and high recall and precision. We surmise that a reason for the difference between summarization and machine translation might be that extraction-based summaries do not really suffer from grammar problems, while translations do. Longer n-grams tend to score for grammaticality rather than content.

It is encouraging to know that the simple unigram co-occurrence metric works in the DUC 2001 setup. The reason for this might be that most of the systems participating in DUC generate summaries by sentence extraction. We plan to run similar experiments on DUC 2002 data to see if unigram does as well. If it does, we will make available our code available via a website to the summarization community.

Although this study shows that unigram co-occurrence statistics exhibit some good properties in summary evaluation, it still does not correlate to human assessment 100% of the time. There is more to be desired in the recall and precision of significance test agreement with manual evaluation. We are starting to explore various metrics suggested in Donaway et al. (2000). For example, weight n-gram matches differently according to their information content measured by tf, tfidf, or SVD. In fact, NIST MT automatic scoring metric (NIST 2002) already integrates such modifications.

One future direction includes using an automatic question answer test as demonstrated in the pilot study in SUMMAC (Mani et al. 1998). In that study, an automatic scoring script developed by Chris Buckley showed high correlation with human evaluations, although the experiment was only tested on a small set of 3 topics.

According to Over (2003), NIST spent about 3,000 man hours each in DUC 2001 and 2002 for topic and document selection, summary creation, and manual evaluation. Therefore, it would be wise to use these valuable resources, i.e. manual summaries and evaluation results, not only in the formal evaluation every year but also in developing systems and designing automatic evaluation metrics. We would like to propose an annual automatic evaluation track in DUC that encourages participants to invent new automated evaluation metrics. Each year the human evaluation results can be used to evaluate the effectiveness of the various automatic evaluation metrics. The best automatic metric will be posted at the DUC website and used as an alternative in-house and repeatable evaluation mechanism during the next year. In this way the evaluation technologies can advance at the same pace as the summarization technologies improve.

References

Donaway, R.L., Drummey, K.W., and Mather, L.A. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceeding of the Workshop on Automatic Summarization*, post-conference workshop of ANLP-NAACL-2000, pp. 69-78, Seattle, WA, 2000.

DUC. 2002. *The Document Understanding Conference*. http://duc.nist.gov.

Fukusima, T. and Okumura, M. 2001. Text Summarization Challenge: Text Summarization Evaluation at NTCIR Workshop2. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, NII, Tokyo, Japan, 2001.

Lin, C.-Y. 2001. *Summary Evaluation Environment*. http://www.isi.edu/~cyl/SEE.

Lin, C.-Y. and E. Hovy. 2002. Manual and Automatic Evaluations of Summaries. In *Proceedings of the Workshop on Automatic Summarization,* post-conference workshop of ACL-2002, pp. 45-51, Philadelphia, PA, 2002.

McKeown, K., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, S. Sigelman. Tracking and Summarizing

News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of Human Language Technology Conference 2002* (HLT 2002). San Diego, CA, 2002.

Mani, I., D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. *The TIPSTER SUMMAC Text Summarization Evaluation: Final Report*. MITRE Corp. Tech. Report.

NIST. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics.

Over, P. 2003. Personal Communication.

Papineni, K., S. Roukos, T. Ward, W.-J. Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. *IBM Research Report RC22176 (W0109-022)*.

Porter, M. F. 1980. An Algorithm for Suffix Stripping. *Program*, 14, pp. 130-137.

Radev, D. R., S. Blair-Goldensohn, Z. Zhang, and R. S. Raghavan. Newsinessence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization. In *Proceedings of human Language Technology Conference* (HLT 2001), San Diego, CA, 2001.

Spärck Jones, K. and J. R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. New York: Springer.

Rath, G.J., Resnick, A., and Savage, T.R. 1961. The Formation of Abstracts by the Selection of Sentences. *American Documentation*, 12(2), pp. 139-143. Reprinted in Mani, I., and Maybury, M., eds, *Advances in Automatic Text Summarization*, MIT Press, pp. 287-292.

WAS. 2000. *Workshop on Automatic Summarization*, post-conference workshop of ANLP-NAACL-2000, Seattle, WA, 2000.

WAS. 2001. *Workshop on Automatic Summarization*, pre-conference workshop of NAACL-2001, Pittsburgh, PA, 2001.

WAS. 2002. *Workshop on Automatic Summarization*, post-conference workshop of ACL-2002, Philadelphia, PA, 2002.