Evaluating an Evaluation Method: The Pyramid Method Applied to 2003 Document Understanding Conference (DUC) Data

Rebecca Passonneau
Columbia University
December 16, 2005

Abstract

A pyramid evaluation dataset was created for DUC 2003 in order to compare results with DUC 2005, and to provide an independent test of the evaluation metric. The main differences between DUC 2003 and 2005 datasets pertain to the document length, cluster sizes, and model summary length. For five of the DUC 2003 document sets, two pyramids each were constructed by annotators working independently. Scores of the same peer using different pyramids were highly correlated. Sixteen systems were evaluated on eight document sets. Analysis of variance using Tukey's Honest Significant Difference method showed significant differences among all eight document sets, and more significant differences among the sixteen systems than for DUC 2005.

1 Introduction

The 2005 Document Understanding Conference used the pyramid method (Nenkova and Passonneau, 2004) for manual evaluation of content selection in automated summarization systems. In prior years, DUC has used other manual methods that made use of automated tools for pre-processing (Lin, 2001) (Marcu, 1999). In parallel with manual evaluation, DUC has been using the automated ROUGE system (Lin and Hovy, 2003). DUC's use of the pyramid method, and the advantages of automated scoring, has inspired automated approaches to the scoring phase of the pyramid method (Fuentes et al., 2005) (Harnly et al., 2005). Given the large amount of pyramid data from DUC 2005, and the fact that the pyramid method will be used in 2006, further investigation of the evaluation method itself is merited in order to better understand what it can reveal in system comparisons.

The main differences between the 2003 and 2005 datasets pertains to the document length, cluster sizes, and model summary length. For DUC 2005, average cluster size was 30.4 documents, with an average document length of 720 words; each of twenty pyramids was constructed from seven model summaries of 250 words each. For DUC 2003, average cluster size was ten documents, with average document length of 500 words. To re-examine the DUC 2003 data, pyramids were constructed from model summaries of 100 words. Pyramids with ten models were used for measuring interannotator agreement on peers and for computing score correlations, when different peer annotators using the same pyramid. Pyramids with seven models were used for measuring interannotator agremeent on pyramids, and for computing score correlations when different peer annotators used different pyramids. This last comparison is unique to the DUC 2003 dataset.

The questions posed are:

- What is the peer interannotator reliability for DUC 2003, and how do scores from different peer annotations correlate?
- How do the two metrics, interannotator reliability on peer annotation, and score correlations, compare across the two datasets?
- What is the pyramid interannotator reliability

for DUC 2003, and how do scores from different pyramids correlate?

- What is the result of system evaluation for DUC 2003?

- How does system evaluation compare across the two datasets?

The major results are as follows. Interannotator agreement for peer annotation is somewhat better for the 2003 dataset, while score correlations are about the same. Interannotator agreement for pyramids is high, and score correlations are as high as when independent annotators use the same pyramid.[1] System scores in 2003 are higher than in 2005, and the difference between original and modified scores is less great, which is most likely due to differences in the sizes of the document sets, and the difference in model summary lengths. There are relatively more significant differences among systems in 2003. Finally, in both years, there were significant differences between document sets, but details are not given for the 2005 data. Here, all document sets are significantly different from each other, with respect to the modified pyramid score.

2 The Two Pyramid Datasets

Briefly, in the pyramid method, multiple human model summaries of the target document clusters are manually annotated to create pyramid models for each cluster. A pyramid consists of clusters of semantically similar phrases, with no more than one phrase from each model summary. Human annotators match peer summaries against the pyramid models, attempting to determine which Summary Content Units (SCUs) from the pyramid models appear in the peer. The matching process is semantic and subjective, rather than based on the ngram similarity approach underlying ROUGE and BE (Hovy et al., 2005). In sum, there are two phases of human annotation, each of which should be evaluated.

In an overview paper of the results of applying the pyramid method in DUC 2005 (Passonneau et al., 2005), the authors point to several factors affecting the metric, such as the size of document clusters and model summaries. They present results of interannotator agreement on annotation of peer summaries in conjunction with an analysis of the correlation of scores from distinct peers. No results on evaluating the pyramid creation phase are presented for DUC 2005. (Nenkova and Passonneau, 2004) report an interannotator reliability result on pyramid creation, but provide few details, in contrast to (Teufel and van Halteren, 2004) which reports good results on a two-phase approach to interannotator agreement on factoid annotation, a related type of content unit for summarization evaluation.

For the present study, the three original DUC 2003 pyramids reported on in (Nenkova and Passonneau, 2004) were used. In addition, pairs of annotators working independently constructed pyramids for five additional document sets, giving a total of eight document sets across which to compare peer performance. Five annotators performed peer annotation of summaries produced by sixteen systems, distributed so that every peer had two distinct annotations. For three docsets, annotators used the same pyramid; for the remaining five, annotators used different pyramids constructed from the same model summaries. The annotators who performed the peer annotations were a mix of unpaid, experienced volunteers and paid annotators who were given some initial training.

In the DUC 2005 dataset, five annotators constructed pyramids for all twenty document sets, working in pairs and adjudicating differences. Six of the twenty-five system peers were annotated twice, by different annotators. Twenty-seven sites participated in DUC 2005, and each site performed peer annotation. Though the DUC 2003 dataset presented here is smaller, due to more limited resources for annotation, it is sufficient for making meaningful system comparisons.

(Passonneau et al., 2005) gives the full formulas for the original pyramid score, that is analogous to precision, and a modified score, that is analgous to recall. As discussed there, both scores are a ratio of the sum of the weights of the SCUs found in the peer (OBServed) to the sum for an ideal summary (MAXimum). OBS is a sum of the weights of the SCUs that occur in a peer annotation. MAX is the maximum sum that can be computed by sampling SCUs without replacement from the pyramid, first

[1] We currently have the results from half the annotations; the remaining annotations are still being done as of this writing.

peers	better than
12	15, 17, 21, 19, 23, 11, 22, 18 (n=8)
.53	.32
13, 16	15, 17, 21, 19, 23, 11, 22 (n=7)
.50	.31
6, 10, 20, 26	15, 17, 21, 19, 23 (n=5)
.46	.29
14, 18	15, 17 (n=2)
.42	.23
19, 11, 22, 23	15
.35	.19

Table 1: System differences in mean modified scores, Tukey's HSD

taking those of the maximum weight, then each next weight, until the total number of the sample equals the observed number of SCUs in the peer (original score), or until it equals the average number of SCUs per human model summary in the associated pyramid.

docset	dice	wdice	alphaN	alphaD
30042	0.84	0.96	0.75	0.81
31041	0.76	0.92	0.68	0.76
31050	0.79	0.95	0.73	0.77

Table 2: Association and reliability measures on sixteen peers for three document sets

3 Peer Annotation

3.1 Interannotator Reliability Results

Following the example of (Passonneau et al., 2005), interannotator reliability on peers was computed using three metrics: Dice (a measure of term association), and Krippendorff's α (Krippendorff, 1980). αN is α with a nominal distance metric that treats all differences the same. αD incorporates a distance metric.[2] Because α measures disagreements rather than agreements, each unit comparison between annotators is weighted by (1-Dice), so that greater values reflect greater disagreement. In addition, a weighted version of Dice was also used.

To compute Dice, counts are collected of the number of times both annotators find the same SCU (a), the number of times one annotator finds an SCU that the other does not (b), the converse (c); it is the ratio of 2a to (2a + b + c). It is thus closer to 1 when there are fewer items the annotators disagree on. In the weighted Dice, SCU weights are factored in so that a is the sum of the weights of the SCUs that both annotators find in a peer, and so forth.

Analyis of variance tests were conducted on linear models with each metric in turn as the dependent variable, and docset, peer and annotator as factors. This provides a two-way analysis of variance test, testing for row and column effects on each metric. With Dice as the dependent variable, docset had a significant affect at the .05 level (p=.02). There were no significant effects of the other factors on any metric; p values ranged from .15 to .85.

Table 2 presents the means of each metric broken down by docset; standard deviations were relatively small, ranging from .03 to .16 and are not shown here. Column two shows where the significance arises in the anova for dice: for 30042, the dice values were on average higher than for the other two doc sets. This may be a training effect. The annotator who did most of 31050 and 30042, who had no prior experience with pyramid annotation, did 31050 first. Note that that all metrics are higher on 30042 than 31050. In addition, Dice is more sensitive to the actual number of agreements, as opposed to the likelihood of agreement.

A comparison of the dice and wdice measures in 2 shows that taking the weights of the SCUs into account yields much higher measures of association, meaning that while the annotations have high pro-

[2] For general discussion of weighted reliability metrics, see (Artstein and Poesio, 2005).

docset	cor	p	95% conf inter for delta
original scores			
30042	.96	0	(0.8936, 0.9873)
31041	.76	0	(0.4234, 0.9120)
31050	.74	.0011	(0.3856, 0.9040)
modified scores			
30042	.97	0	(0.9081, 0.9891)
31041	.83	0	(0.5722, 0.9400)
31050	.73	.0014	(0.3612, 0.8987)

Table 3: Pearsons correlation of scores two annotators, same pyramids

portions of agreement on SCUs, they have even better agreement on a per SCU weight basis.

The unweighted α values indicate that even without counting partial matches, agreement is good. (Passonneau et al., 2005) argue that weighted α values, however, are the best stand-alone measure of interannotator reliability. On this measure, the interannotator peer agreement is roughly the same for 2003 and 2005. An interesting difference from DUC 2005 is that there is much less difference in the 2003 data between the unweighted and Dice-weighed α measures.

3.2 Score Correlation Results

The scores and modified scores from annotator pairs using the same pyramid are quite close. On average, the difference in scores for all forty-eight peers was .02 for the original score and .0003 for the modified score. There are a variety of ways to test the degree to which the scores differ. Since system performance is measured by average performance across document sets, a paired t-test of the means could be used, and in fact, shows no significant difference. However, Pearson's correlation gives a more conservative test of the degree to which two series of individual scores provide the same ranking. Table 3 gives Pearson's correlations between the scores from different annotations. Separate correlation tests are done on each document set, in part because document set has significant effect on score variance, and in part, to permit direct comparison with DUC 2005 results reported in (Passonneau et al., 2005) . Correlations are generally high, and more so for the modified scores.

Ann1	(W=4) Poland is affected worse
Ann2	(W=3) Poland was hit hardest by the brutal cold
Sum1	Throughout Poland, Europe's most affected country$_{Ann1}$
Sum2	and$_{Ann2}$ **Poland suffered most.**$_{Ann1,2}$
Sum3	**Hardest hit were Poland,**$_{Ann1,2}$
Sum4	Many deaths were attributed to the cold$_{Ann1}$ **with**$_{Ann1,2}$ **Poland apparently the hardest hit.**$_{Ann1,2}$

Figure 1: SCU overlap for one SCU from independently constructed pyramids

Set relation	$\mathcal{P}$
Set A equals set B	0
Set A subsumes set B	$\frac{1}{3}$
Set A does not subsume B and the intersection is non-empty	$\frac{2}{3}$
Sets A and B are disjoint	1

Figure 2: $\mathcal{P}$, as used in MASI

4 Pyramid Annotation

The pyramid annotation method has been described in (Nenkova and Passonneau, 2004) and in the DUC 2005 annotation guidelines. Briefly, the goal of the annotation is to determine semantic content that recurs across different humans' summaries of the same source documents, independent of the words used, and to let frequency of occurrence of the same content across summaries represent how highly to weight the same content in a new summary. The annotation units are referred to as Summary Content Units (SCUs).

Figure 1 illustrates a simple case where two annotators independently created very similar SCUs for pyramids from the same model summaries. SCU labels assigned by each annotator are shown, along with the distinct weights. Bold face words were selected by both annnotators. Impressionistically, the annotations are similar. The questions addressed in this section are how to quantify the similarity of independently created pyramids, and whether the use of pyramids constructed by different annotators has an impact on scores.

The interannotator metric use here, referred to as MASI (Measuring Agreement on Set-valued Items), is a distance measure for use with interannotator agreement metrics like Krippendorff's α or the β^3 metric proposed in (Artstein and Poesio, 2005). It

Docset	MASI	1-Dice
D30016	.79	.66
D30040	.80	.68
D31001	.68	.54
D31010	.69	.52
D31038	.71	.53

Table 4: Interannotator agreement on five pyramids using Krippendorff's α, and two different distance metrics

docset	cor	p	95% conf int
Original score			
30016	.90	0	(.7304, .9652)
30040	.84	0	(.5333, .9436)
Modified score			
30016	.91	0	(.7591, .9693)
30040	.88	0	(.7029, .9611)

Table 5: Pearson's correlations of scores from two annotators, different pyramids

is specifically intended for set-valued semantic and pragmatic annotations, and has been used for coreference annotation ((Passonneau, 2004)). The formula for MASI is $(1 - Dice) \times \mathcal{P}$, where $\mathcal{P}$ is a penalty factor that takes into account the monotonicity of the relations between two sets, as given in Figure 2. As argued in (Popescu-Belis et al., 2004), it is useful to compare results using different ways of measuring the same thing. Thus we present results for α using two distance metrics: MASI, and (1-Dice).

Assessing interannotator reliability typically depends on representing the data in an i by j matrix, where i is the number of units being coded, j is the number of coders, and each cell i, j contains the value that coder j assigned to unit i. As noted in ((Passonneau, 2004)), certain semantic and pragmatic annotations require annotators to group units into categories, as is the case with coreference annotation where the units are discourse referential noun phrases, and the categories are equivalence classes of coreferring expressions. Pyramid annotation requires annotators to group words from different summaries into equivalence classes of expressions that express the same content.

The coding units are the words in a model summary. For every word in a model, annotators either assign the word to the same SCU or not, thus each coding value is the set of tokens in the SCU the word was assigned to, excluding the current token.

Table 4 shows the results of interannotator agreement on the five pyramids. 1-Dice as a distance metric rewards the proportion of words from one annotator's SCU that overlap the other annotator's SCU, without consideration of the overall semantic compatibility of the SCUs. For the same amount of word overlap, MASI gives a greater reward when one SCU subsumes the other, and penalizes intersection. In general, the MASI values are much higher than for (1-Dice), meaning that when SCUs from annotator one and annotator two are not identical, one subsumes the other more often than not. This would occur when the associated concepts are in a part whole relation, or other semantically monotonic relation. Intersection would reflect mixing of concepts.

4.1 Score Correlation Results

Correlations of the scores from two annotators using different pyramids are as high as those when two annotators use the same pyramid, as shown in Table 5. We expected high correlations, but we expected score correlation to be higher with the same pyramid. The high correlations for the different pyramid condition may again be a training effect. For the two sets reported on in Table 5, the first annotator in both cases had extensive prior experience with pyramid annotation, and the second annotator had some prior experience, or had already annotated multiple sets for the same pyramid condition.[3]

5 Peer Scores

Analysis of variance with original pyramid score as the dependent variable and peer and docset as factors indicates a significant effect of both factors at the .1% level, with a p-value for peer of p=.000334, and for docset of 0. Similar results obtain for the modified score (peer, p=.0004637; docset, p=0).

Table 1 shows the system differences that result from using Tukey's Honest Significant Difference

[3] Data on the remaining three sets will be available in January 2006.

(HSD) method to examine peer differences using the modified score. Similar results obtain for the original scores but are not shown due to space limitations. Tukey's HSD is more conservative than the Least Significant Difference (LSD), and is used to guard against finding spurious differences in testing all pairwise comparisons when doing analysis of variance. Confidence intervals are created for each pairwise system comparison, and those that are significant are the ones greater than HSD. The first of each pair of rows in Table 1 indicates the sets of peers for which a significant difference is found. For expository purposes, the next row in each pair shows the mean modified score for the set. This conveys impressionistically the size of the difference, but again, the means are compared peer by peer, not set by set.

The results indicate that out of sixteen systems, there are four peers sets of peers that do significantly better than at least one other peer. As shown, there are two systems that do better than seven peers, or half the remaining systems. A second system performs significantly better than four of the seven. Eleven systems do better than the two bottom ranked systems, one of which does better than the other.

The findings here show twice as many sets of system differences than reported in (Passonneau et al., 2005), for a smaller number of systems overall. Using Tukey's method, they found two sets of orderings: ten peers out of twenty five systems did better than the lowest performing system, and two more did better than the two lowest performing systems. The two lowest performing systems had original scores of .12 and .14, and modified scores of .06 and .09. The original and modified scores for the highest performing system were .25 and .20. Thus in general, absolute score values were lower for DUC 2005 than they are for the DUC 2003 evaluation.

6 Score variance due to document set

There is a greater differentiation of scores due to document set than due to peer. Table 6 shows the results of using Tukey's HSD to compare mean modified scores for each pair of document sets. Again, the mean for each set is shown for illustrative purposes only. Using the modified score gives a complete ranking among the document sets; that is, every document set is significantly different from every other document set with respect to average system performance. Somewhat fewer differences were found using the original score.

7 Mean SCU Weight

Mean SCU weight (MSW) has been suggested as a metric for comparing pyramids (Passonneau et al., 2005). A higher MSW reflects a greater number of SCUs at higher weights. A lower MSW indicates more low weighted SCUs.

For the three DUC 2003 pyramids constructed from ten models, the mean MSW was 2.88. Mean MSW for the five pyramids with the seven model summaries is 2.58, computed by taking the average of the ten MSWs from both sets of pyramid annotators. Interestingly, MSW does not increase by much given ten models instead of seven. However, MSW is higher for this data than for DUC 2005: (Passonneau et al., 2005) report that the twenty DUC 2005 pyramids have an average MSW of 1.9. The number of model summaries was also seven, but they were two and a half times longer (100 vers 250 words), and the document sets were larger. Anyone of these factors might correlate with variations in MSW.

Table 7 gives the MSWs from each annotator for the pairs of pyramids that were constructed independently from the same models. Note that in addition to the measures reported above (interannotator reliability and score correlations), the similarity in MSW within each pair argues for the similarity of the pyramids for scoring purposes.

8 Discussion

It is unlikely that systems have gotten worse between 2003 and 2005. Instead, it is likely that the use of larger document clusters and longer articles made the difference in task difficulty exceed the improvements in system design. Also, the different results for each year suggest that whether the pyramid method distinguishes systems may depend on characteristics of the test design, such as how many documents are in the clusters, how long the documents are, and how long the model summaries are. The large effect of document set in both years presumably points to deeper differences that affect task difficulty, independent of cluster and document size,

docsets	higher scores than
30016	31038, 31041, 31001, 30042, 30040, 31050, 31010 (n=7)
.55	.38
31010	31038, 31041, 31001, 30042, 30040, 31050 (n=6)
.51	.35
31050	31038, 31041, 31001, 30042, 30040 (n=5)
.46	.33
30040	31038, 31041, 31001, 30042 (n=4)
.42	.31
30042	31038, 31041, 31001 (n=3) (n=3)
.36	.29
31001	31038, 31041
.34	.27
31041	31038
.32	.23

Table 6: Docset differences in mean original and modified scores, Tukey's HSD

such as topic dependent difficulties.

docset	annotator	mean scu wt
30016	ann1	2.31
	ann2	2.34
30040	ann1	2.33
	ann2	2.25
31001	ann1	2.69
	ann2	2.71
31010	ann1	3.43
	ann2	3.43
31038	ann1	2.14
	ann2	2.20

Table 7: Similarity of SCU weights across independently constructed pyramids using the same model summaries

Another factor to remember regarding the two evaluation sets compared here is that the pyramid method identifies subsentential content units, but the systems being evaluated create summaries by extracting full sentences. Inspection of the peer annotations on a sentence by sentence basis indicates that there is a wide variation in SCU weights within sentences. To illustrate this variation, the differences in weights between every pair of SCUs were summed on a sentence by sentence basis for a sample of 2003 peers. If all the SCUS in a sentence have the same weight, the sum will be 0. The sum increases if there are more pairs with widely distinct weights. Table 8 shows the results for three sentences produced by peer 13 for docset 30042. The columns indicate the docset, the original score, the sentence number, the number of SCUs per sentence, the sum of weight δs between all pairs of SCUs, and the average δ. This peer summary has a moderately high score, but seems to be penalized for the third sentence, which contains only two SCUS. The relatively high average δ indicates that one of the SCUs had a much higher weight than the other. Thus on a per sentence basis, this summary is very good, apart from the third sentence.

docset	score	s#	n SCUs	sum δ	avg δ
30042	.49	1	4	8	2
		2	8	14	1.75
		3	2	12	6

Table 8: Mean deltas among all pairs of SCU weights, by sentence for peer 13

9 Conclusion

The pyramid method has been applied to DUC 2003 by constructing new model pyramids, then compared with results for DUC 2005. Comparisons were made regarding the overall reliability of the method, and the types of observations it supports regarding the ability to compare systems across a reasonable number of document sets.

Three main conclusions are drawn. First, both phases of the annotation method appear to be reliable, thus it should be feasible to automate each of them. Results on 2003 data for interannotator agreement and correlation of scores indicate that the method is not sensitive to the use of distinct pyramids from the same models. Results on interannotator agreement of peer annotation were marginally higher for DUC 2003, possibly due to the higher mean SCU weights of the 2003 pyramids. Second, the method discriminates among systems better in 2003 than in 2005, possibly due to clearer differences across document sets, or possibly due to differences in the pyramids as reflected in the mean SCU weight. Third, there are highly significant differences across document sets that are greater in magnitude than system differences, which may make it difficult to detect system differences. Evaluation might be more revealing if we could better characterize differences across document sets.

There are many factors that differentiate corpora, and while methods for characterizing corpora quantitatively are difficult to come by (Kilgarriff, 2001), the large number of differences in performance across document sets, and the size of the differences, suggest that summarizaton systems would perform better on a large scale if they could adapt to the language differences associated with different document sets. All document sets are newswire, so genre is not a factor. Two likely dimensions to explore are the homogeneity of the document clusters, and differences associated with different topics.

Finally, it was suggested that the pyramid method may be more suited to related technologies, such as Q&A of the long answer variety, given the smaller granularity of SCUs relative to sentence extraction methods of summarization (cf. (Lin and Demner-Fushman, 2005).)

Ackowledgements. This work was supported by DARPA DI NBCH105003, and NUU01-00-1-8919. The author extends an enormous thanks to Ani Nenkova, who contributed to early discussions of the study design, and who annotated some of the new data collected for this study. Thanks also go to David Elson, who did a large part of the new annotations, and to the great job done by two paid annotators.

References

Ron Artstein and Massimo Poesio. 2005. Kappa cubed = alpha (or beta). Technical Report NLE Technote 2005-01, University of Essex, Essex.

M. Fuentes, E. Gonzalez, D. Ferres, and H. Rodriguez. 2005. Qasum-talp at duc 2005 automatically evaluated with a pyramid based metric. In *Proceedings of the 2005 Document Understanding Conference*.

Aaron Harnly, Ani Nenkova, Rebecca Passonneau, and Owen Rambow. 2005. Automation of summary evaluation by the pyramid method. In *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, September.

Ed Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating duc 2005 using basic elements. In *Proceedings of the 2005 Document Understanding Conference Workshop*.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, pages 1–37.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*.

Chin-Yew Lin. 2001. See: Summary evaluation evironment, user's guide.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. In Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–126. MIT Press.

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Joint Annual Meeting of Human Language Technology (HLT) and the North American chapter of the Association for Computational Linguistics (NACL)*, Boston, MA.

Rebecca Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in duc 2005. In *Proceedings of the 2005 Document Understanding Conference*.

Rebecca Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Portugal.

A. Popescu-Belis, L. Rigouste, S. Salmon-Alt, and L. Romary. 2004. Online evaluation of coreference resolution. In *Fourth International Conference on Language Resources and Evaluation*.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *EMNLP-04*.