# Applying the Pyramid Method in the 2006 Document Understanding Conference

**Rebecca J. Passonneau** and **Kathleen McKeown** and **Sergey Sigelman** and **Adam Goodkind**
Columbia University
Computer Science Department
New York, NY 10027
{becky,kathy,ss1792}@cs.columbia.edu, ang2108@columbia.edu

## Abstract

The pyramid evaluation effort for the 2006 Document Understanding Conference involved twenty-two sites on twenty document sets. Each pyramid content model (one per document set) was constructed from four human summaries. Peer systems were scored using the modified pyramid score introduced in DUC 2005. ANOVAs with score as the independent variable and nine factors yielded three significant factors: document set, peer, and content responsiveness. There were many more significant differences among peer systems in 2006 than for DUC 2005. We speculate this is due to a combination of improved systems and improvements in our evaluation procedures.

## 1 Introduction

The 2005 Document Understanding Conference (DUC) administered by NIST included a voluntary evaluation phase to apply the pyramid evaluation method, an annotation procedure and accompanying set of metrics for assessing the content of automatic summaries. This supplemented other types of summarization evaluation, including automated methods such as ROUGE (Lin and Hovy, 2003). Columbia University administered the 2005 pyramid effort, and twenty five sites participated, as reported in (Passonneau et al., 2005). For a second time in 2006, NIST invited sites participating in the automatic summarizer evaluation to participate in a supplementary evaluation using the pyramid method. Columbia University again administered the effort, four sites consisting of Mitre, Microsoft, National University of Singapore and Columbia University participated in the advance preparation of the pyramid models, and twenty-two sites (peer systems) participated in the actual pyramid evaluation.

Summarization evaluation at DUC begins with NIST's preparation of materials and task definition. This includes creating document clusters of news articles of a given article length and cluster size. NIST's task definition includes specifying the length of peer summaries. The 2005 and 2006 summarization task prepared by NIST had similarly sized document clusters, and the same length requirement on peer summaries. Pyramid models for evaluating peer summaries are constructed from human summaries of the same length as the peers. The 2005 pyramids were constructed from seven model summaries; for 2006 we had four models per pyramid. In principle, pyramid scores are more consistent and robust with a larger number of models (Nenkova and Passonneau, 2004), but in practice, we found a greater ability to differentiate systems in 2006. In part, this is probably the result of improvements to the systems that participated. In addition, we made several improvements to the 2006 procedures for administering a large, group evaluation effort, and we believe this led to improved consistency among the individuals who constructed the pyramids, and among the sites who annotated the peer summaries.

This paper describes the 2006 pyramid evaluation, briefly compares results with previous years, and considers the semantic relationships within and across pyramids. In section 2, we review the key points of the pyramid approach, which has been presented in detail elsewhere (Nenkova and Passon-

neau, 2004) (Passonneau et al., 2005) (Harnly et al., 2005) (Passonneau, 2006). We present the specific methods and procedures used here in section 3, followed by a general description of the dataset, including characteristics of the 2006 pyramids in contrast to the 2005 pyramids (section 4). In section 5, we present the quantitative results, followed by a discussion of the results (section 6). In section 7, we present a brief qualitative analysis of the distributional differences of content units we see within across pyramids. We conclude in section 8.

## 2 Review of Pyramid Method

The pyramid content annotation and evaluation method is an approach designed to capitalize on an observation seen in human summaries that presents an obstacle to using a single summary as a model: summaries from different humans always have partly overlapping content. The pyramid method includes a manual annotation method to represent Summary Content Units (SCUs) and to quantify the proportion of model summaries that express this content. All SCUs have a weight representing the number of models they occur in, thus from 1 to $\max_n$, where $\max_n$ is the total number of models. A similar content evaluation method is presented in (Teufel and van Halteren, 2004); differences include their use of a quasi logical form representation for their *factoids*, no direct link to paraphrases, insistence on a much larger number of model summaries, no weighting of factoids, and a different evaluation metric.[1] We partition pyramids into tiers by the SCU weight, which gives $n$ tiers. There are very few SCUs expressed in all models (i.e., weight=$\max_n$), and increasingly many SCUs at each lower weight, with the most SCUs at weight=1. It is because of this *bottom-heavy* distribution of SCUs that we refer to the content model as a pyramid.

Figure 1 illustrates an SCU from one of the DUC 2006 pyramids. At the top is a label assigned by the annotator during the annotation process that captures the annotators' view of the shared meaning found across model summaries. This SCU is of weight four because each of the four model summaries expressed this information. As illustrated, we avoid a formal representation of the semantic content of the SCU, and we link a range of surface forms to the same semantics. Here it happens that three of the four expressions, which we refer to as contributors, have no overt subject, however we take the fillers of the corresponding semantic arguments of *make* and *take* to be implicitly present, as indicated in the label: *the Concorde*.

The approach involves two phases of manual annotation: pyramid construction and annotation of unseen summaries against the pyramid to determine which SCUs in the pyramid have been expressed in the peer summary. The interannotator reliability of the annotation procedures are discussed in (Passonneau, 2006) and (Passonneau, 2005). We have found both types of annotation to be reliable. In (Passonneau, 2006) and (Passonneau et al., 2005) we argue that interannotator reliability is best investigated in the context of an independent application of the data, and we exemplify this approach by comparing the scores yielded by peer annotations from different annotators who have scored the same summaries against the same pyramids (Passonneau, 2006) and (Passonneau et al., 2005), and by comparing the scores yielded by peer annotations from different annotators scoring the same summaries using different pyramids ((Passonneau, 2005)). In both comparisons, we find that the parallel sets of scores from different annotators are significantly highly correlated.

Given a pyramid model for a document set, a variety of methods for scoring peer summaries are possible. In DUC 2005 we used two that are somewhat analogous to recall and precision metrics used in information retrieval. Both require a peer summary to be annotated against a pyramid so as to compute a sum of the weights of the SCUs that a given peer expresses. This sum of the observed weights in a peer is then normalized against an ideal sum. The original pyramid score normalizes the observed sum against the maximum sum the pyramid allows, given a count of all the SCUs in the peer. A pyramid generates multiple ways of assigning weights to $m$ SCUs, with the constraint that a given weight cannot be used more often than it appears in the pyramid. The original score requires that all sentences in a peer be annotated into SCUs, including ones that have not been represented in the pyramid model. While there

[1] For a critique of their approach to reliability of factoid annotation, see (Passonneau, 2006).

| Label | The Concorde crossed the Atlantic in less than 4 hours |
|---|---|
| Sum1 | making the transatlantic flight in three and one half hrs |
| Sum2 | The Concorde could make the flight in between New York and London or Paris in less than four hours |
| Sum3 | completing its journey from London to New York in about 3 hours, 30 minutes |
| Sum4 | took less than 4 hrs to cross the Atlantic |

Figure 1: An example SCU of weight 4 from a DUC 2006 pyramid

are guidelines to follow in this case, the more unseen SCUs there are, the less reliable is this aspect of the annotation. For DUC 2005, we used a modified pyramid score that does not require the annotation of unseen SCUs; instead, the observed sum is normalized against the maximum sum the pyramid can generate, given the average count of SCUs per summary in the pyramid.

## 3 Methods

For DUC 2006, new annotation guidelines for both the pyramid and peer phases of the process were written, informed by the experiences of 2005 and by pilot tests early in 2006. The primary focus of the changes to the pyramid annotation were to make the peer annotation process easier and more reliable by imposing further constraints on pyramid annotation. This led to the following changes:

- We developed an explicit set of guidelines for writing the SCU labels in a more uniform manner, making it easier to search the labels;
- we geared the instructions towards producing greater semantic uniformity among contributors to the same SCU, for example, encouraging annotators to prune unnecessary words, and to reuse words from the peer in multiple contributors in order to make implicit arguments explicit, so long as the full semantics of overlapping contributors remained distinct;
- we paid attention to the tradeoffs between annotating SCUs that are more atomic (semantically simpler, e.g., fewer arguments), and minimizing the total number of SCUs. We encouraged annotators:
  - to split SCUs of weight 2-4 into multiple SCUs under certain conditions; this facilitates the peer annotation process because there is less ambiguity about what to do if a new summary expresses only part of the content in the original (unsplit) SCU;
  - to find ways to merge SCUs of weight one with other SCUs; this facilitates peer annotation by minimizing the number of SCUs of weight one;
  - to make SCUs of weight one the same semantic granularity as a protoytpical clause, e.g., subj-verb-obj, which standardizes the semantics of SCUs of weight one, facilitating search;
- we created training samples from DUC 2005 data for annotators to try out and to compare with pre-annotated output.

We made the following modifications to the annotation software developed at Columbia and first used in DUC 2005:

- enforcement of the constraint that disallows SCUs whose weight is greater than the total number of models;
- enforcement of the constraint that an SCU not contain multiple contributors from the same summary;
- drag-and-drop capability within the SCU list;
- enhanced search of model text and SCU labels;
- enhancements to the appearance to clarify the different functions.

### 3.1 Pyramid Annotation

Pyramids were constructed for twenty document sets. NIST selected the twenty document sets using two criteria: high clarity ratings, and even distribution among the ten assessors who wrote the model summaries.

Six individuals at four sites collaborated on the pyramid annotation task, two of whom had prior experience. Two of the inexperienced annotators experimented with the training samples that were included with the guidelines, which led to more training prior to the pyramid construction process than occurred in 2005. The two remaining annotators had not previously created pyramids, but were familiar with the process from the previous year. Pyramid annotators were urged to communicate among each other to share questions, comments and problems. There was almost daily communication among annotators until the initial annotations were complete. Each individual was assigned a partner, and partners reviewed each other's pyramids. During this phase of the process, there was less discussion, and the types of questions that did arise suggested that most problems had been ironed out prior to this phase.

### 3.2 Peer Annotation

We decided in advance to use only the modified pyramid score. This simplified the annotation process in that annotators were not asked to identify how many content units were represented in the peer summary that did not match SCUs in the model pyramids. Once they identified all the SCUs from a given pyramid that were expressed in the peer, they were done with the peer annotation task.

Twenty-two sites participated in doing peer annotation of twenty-one peers plus the baseline. In DUC 2005, volunteers at Columbia reviewed the peer annotations. For DUC 2006, we partnered each site with another site, and peers reviewed each others' sets of peer annotations. This worked extremely well, both in terms of the care that annotators took with their annotations, and in terms of a more even distribution of work across sites.

### 3.3 Scoring

As noted above, we used only the modified pyramid score. In both the original and modified scores, the formula is a ratio of the sum of the observed SCU weights in a peer ($SW_{Obs}$) to a maximum sum of SCU weights ($SW_{Max}$), given some number of SCUs to normalize against:

$\mathcal{P} = \frac{SW_{Obs}}{SW_{Max}}$.

Where O is the number of SCUs in a peer,

$SW_{obs} = \sum_{i=1}^{n} i * O$.

$SW_{Max}$ varies depending on how many SCUs to use in constructing an ideal sum of weights to normalize against. In the original pyramid score, this was the observed number of SCUs in the peer. When a peer contains more than a few words that do not match the associated pyramid, it becomes difficult to estimate how many SCUs of weight zero the non-matching text represents. In the modified score, we use the average number of SCUs per model summary in the relevant pyramid, which eliminates this subjective estimate. We use $T_i$ to label each tier in the pyramid consisting of all the SCUs of a given weight $i$. Where $j = max_i$, the maximum sum a pyramid generates for an ideal summary with $j$ SCUs is:

$SW_{Max} = \frac{\sum_{i=1}^{n} |T_i|}{j}$.

The DUCView annotation tool can automatically assign scores once the peer annotation has been completed. The scores were computed at Columbia, and included in tables along with content responsiveness, overall responsiveness, and five linguistic quality ratings. The full table of results was sent to NIST who made it available to all participants.

## 4 Characteristics of Data and Contrast with Previous Years

The DUC 2006 pyramids were constructed from four model summaries each, in contrast to the seven model summaries used in DUC 2005. In other ways, the pyramids were similar. The document clusters being summarized had similar length and topic characteristics: clusters were comprised of twenty-five articles, compared with an average of thirty three in 2006. The model summaries for both years were the same length of 250 words. A pyramid's mean SCU weight is a good indicator of the overall distribution. Across the twenty 2006 pyramids, the mean of the mean SCU weight per pyramid was 1.56; for the twenty five 2005 pyramids, it was 1.9.

## 5 Quantitative Results

We did an analysis of variance of the data with the peer modified pyramid score as the dependent variable and with nine factors: document set (Setid), peer summarizer (Peerid), content responsiveness, overall responsiveness, and the five linguistic quality ratings. The results indicate a significant ef-

| Docsets | Mean modified pyramid score |
|---|---|
| 5 | .065 |
| 1, 3, 8, 15, 47 | .133 |
| 50 | .135 |
| 45, 30 | .158 |
| 28 | .164 |
| 16, 17, 20, 29 | .172 |
| 27 | .197 |
| 14 | .229 |
| 43 | .252 |
| 40 | .269 |
| 24 | .286 |
| 31 | .357 |

Table 1: Significantly distinct groups of docsets, using Tukey's HSD

fect for three of the factors. Setid and Peerid were highly significant predictors of score, with p values effectively zero. Content responsiveness was also a highly significant predictor of score, with p=.0001.

To investigate further the way in which the significant factors interacted with score means, we used Tukey's Honest Significant Difference (HSD) method to identify significant differences within each factor. Table 1 presents the results for Setid, Table 2 presents the results for Peerid, and Table 3 presents the results for content responsiveness.

In Table 1, the sets of document sets differentiated by Tukey's HSD are disjoint. Column one of each row shows a document set where the mean modified score is significantly different from rows above and below. To illustrate how the sets differ, column two shows the mean score for all peers on the relevant document set. For the twenty document sets, there are twelve significantly distinct groups, nine of which contain only one document set. The document set that is most difficult has a mean score of .065, which is less than one fifth that of the easiest document set (.0357).

Table 2 uses a different layout than Table 1 because the significant differences are not disjoint. The rows in plain type in the first column of Table 2 shows sets of peers that did significantly better than the peers listed on the same row in column two. Below each set of peers in column, the mean score for this set appears in italics; the means are presented for illustrative purposes, to indicate how the mean scores for the set in the left column increase as the set in the right column grows larger. As shown, there are eight distinct groups of peers. The baseline is in the lowest performing set (N=5); peers 10 and 23 are the highest performing, and do significantly better than all but one of the other peers.

| Content responsiveness significantly distinct from | content responsiveness |
|---|---|
| 3 (*.2015*) | 1 (*0.1253*) |
| 4 (*.2225*) | 1 (*0.1253*) |
| 5 (*.2241*) | 1 (*0.1253*) |

Table 3: Significantly distinct levels of content responsiveness, using Tukey's HSD

Table 3 shows the results of Tukey's HSD for content responsiveness. The scores associated with the five levels of content responsiveness serve to differentiate each of levels three, four and five from one. Thus there are really only two clearly distinct groups, given by the right and left columns in the table. As in Table 2, the mean modified score for each level of content responsiveness shown in the table is presented in italics.

## 6 Discussion of Quantitative Results

For three years for which we have pyramid analyses, 2003 (Passonneau, 2005), 2005 (Passonneau et al., 2005) and 2006, document set has always been a significant factor predicting mean pyramid score, and has always had relatively many distinct differences. In 2003, for example, eight document sets were evaluated and all were significantly differentiated based on the original or modified pyramid scores. This demonstrates that the DUC organizers at NIST have been consistently successful at creating document sets of wide-ranging difficulty. One thing that would be interesting to investigate further is to what degree the system differences shown in Table 2 are reflected within each of the distinct document sets shown in Table 1. We present a few observations about the relationship between score magnitude, document set difficulty, and peer ranking, illustrating why it is necessary to test on a large number of document sets.

| Peers higher scoring than | peers |
|---|---|
| 1, 17, 18, 25, 25 (N=5) | NIL |
| *.113* | |
| 22, 29, 32 (N=3) | 1 |
| *.169* | |
| 19, 24, 33 (N=3) | 1, 35, 17, 18 (N=4) |
| *.176* | |
| 2, 3, 6, 14, 15 (N=5) | 1, 35, 17, 18, 25 (N=5) |
| *.199* | |
| 28 | 1, 35, 17, 18, 25, 29 (N=6) |
| *.205* | |
| 27 | 1, 35, 17, 18, 25, 29, 32, 22 (N=8) |
| *.210* | |
| 8 | 1, 35, 17, 18, 25, 29, 32, 22, 14 (N=9) |
| *.214* | |
| 10, 23 | 1, 35, 17, 18, 25, 29, 32, 22, 14, 19, 5, 33, 24, 3, 6, 2, 15 (N=17) |
| *.241* | |

Table 2: Significantly distinct groups of peers, using Tukey's HSD

If we look at the two highest scores in the dataset ($> 0.50$), both were on document set D0631. As shown in Table 1, D0631 was the document set with the highest mean modified score (.357). One of the systems that scored above 0.50 was peer 10, one of the top ranked peers as shown in Table 2. However, the other high score was from peer 15, which had only middling performance overall. Considering the data another way, let us look at the third most difficult document set: D0650. The four best performing systems on D0650 included the top two peers (10 and 23), a bottom ranked peer (25), and one close to the bottom (22). Thus on any one document set, a system can perform quite differently from its overall average.

A related question, presumably of interest both to assessors and to systems, is how many document sets, and which ones, would be minimally required to identify the peer differences shown in Table 2. However, this question interacts with whether peer performance is sufficiently consistent, and sufficiently distinct from other peers, to be differentiated at all (cf. (Nenkova, 2005)).

In the 2003 data, sixteen peer systems were analyzed. Peer and model summaries were significantly shorter in 2003 (100 words compared with 250), and clusters consisted of ten rather than thirty (DUC 2005) documents or twenty-five (DUC 2006) documents, with a document length of 500 words rather than 720 (DUC 2005). Pyramids were constructed from a seven to ten model summaries. Many more significant differences among peers were found for 2003 than in 2005, which we speculated resulted from the characteristics of the document sets, pyramids and peer annotation, not to system performance. With the shorter summaries in 2003, and a much higher mean SCU weight per pyramid (2.9), we found that the overall score range was about twice that of 2005. Also, annotators had more training, and as much time to perform the peer annotation as they wished. In 2006, we believe the greater differentiation of peers is due to a combination of genuine improvements in system performance and improved evaluation procedures.

DUC 2005 was the first large-scale application of the pyramid method, and our first experience with untrained annotators. The 2003 analysis reported in (Passonneau, 2005) was conducted concurrently with the 2005 study, on a much smaller scale, using a constraint on pyramid annotation that was not applied for DUC 2005 or 2006 (cf. (Passonneau, 2006)). The pyramids thus differ because the document set characteristics differ between 2003 versus 2005 and 2006, and also as a result of our increas-

ing experience, expertise in working with untrained annotators, enhancements to the annotation software between 2005 and 2006, and modifications to the annotation method itself.

To summarize, we list four characteristics of the pyramid evaluation task that seem to affect differences in how well the method differentiates systems:

- differences in difficulty of document clusters for summarizers;
- pyramid characteristics, such as overall size, clarity of labeling, and coherence within SCUs;
- score variability and overall score range;
- engineering improvements to peer systems.

## 7 Qualitative Observations: Semantics of SCUs and Pyramids

In preceding sections, we have primarily discussed the ability of SCUs and pyramids to sort peer systems into those that perform relatively better or worse with respect to content selection. Here we consider semantic characteristics of SCUs and of pyramids.

The topics for the document sets have semantic differences that might account for observed variations in difficulty for summarizers. Topics associated with the easier document sets tend to be about specific events over a narrow time period, and concrete cause-and-effect relations. The topics for the two easiest document sets are:

- D0631: Discuss the Concorde jet, its crash in 2000, and aftermaths of this crash.
- D0624: What is known about the murder of Stephen Lawrence, his killers, the actions of the government, and the reactions of the public?

In contrast, the topics associated with the more difficult document sets involve broader time spans and more abstraction:

- D0650: Describe former President Carter's international efforts including activities of the Carter Center.
- D0605: Describe what procedures for treatment of osteoarthritis have been attempted and the result of research on these treatments.

Across pyramids, the SCUs of higher weight tend to be more general. For example, two SCUs of weight four (the maximum weight) from different pyramids are shown below. Note that both express general statements, e.g., about *wetlands* or *exercise*, rather than specific ones, e.g., about wetlands in a specific location, or a specific form of exercise, such as calisthenics or yoga.

- D0603 (W=4): Wetlands help control floods
- D0605 (W=4): Exercise helps arthritis

Compare the specificity of two low-weighted SCUs from the same pyramids:

- D0603 (W=1): In underdeveloped countries the increase of rice-planting has negative impacts on wetlands
- D0605 (W=1): Arthroscopic knee surgery appears to reduce pain, for unknown reasons

SCUs of higher weight also tend to be less dependent on the meaning of other SCUs. Here we see two SCUs of weight four that are very specific, referring to specific events involving specific entities; however, both of these highly weighted SCUs can be interpreted in isolation.

- D0640 (W=4): The Kursk sank in the Barents Sea
- D0617 (W=4): Egypt Air Flight 990 crashed

In contrast, the two SCUs of weight 1 shown below, from the same pyramids, refer to entities whose interpretation depends on entities mentioned in other SCUs:

- D0640 (W=1): The *escape hatch* was too badly damaged to dock in 7 attempts
- D0617 (W=1): *Tail elevators* were in an uneven position, indicating a possible malfunction

The *escape hatch* is a part of the Kursk; the *tail elevators* are a part of the airplane that crashed.

We hypothesize that the semantic differences among SCUs of different weights can be expressed as implicational relationships. For example, it seems

that if a pyramid contains SCUs that are very general, they will not have lower weights than specific SCUs. Also, pyramids whose highly weighted SCUs are neither general nor context independent tend to be more difficult. For example, document set D0647 is associated with lower mean pyramid scores (.133; cf. Table 1). It has nine SCUs of weight four that are all very specific: about the sea rescue of the Cuban child, Elian Gonzales. Of these, the interpretations of five depend on other SCUs.

## 8 Conclusion

We have presented a description of the pyramid evaluation effort at DUC 2006 and an analysis of score results. We believe the large number of system differences revealed by application of Tukey's HSD to the ANOVA results indicates several noteworthy achievements. The similarity of the 2005 and 2006 tasks permits relative comparison of the number of system differences; the greater score range in 2006, and larger number of significant differences among systems, suggest that summarization systems have improved in their ability to capture the content that humans find relevant. In 2005, two systems (14 and 15) performed significantly better than the baseline, based on modified scores; in 2006, this number rose to sixteen. The fact that the 2006 pyramids can differentiate systems well, despite fewer model summaries and the correlated lower mean SCU weights compared with 2005, indicates that the improvements to the annotation guidelines and annotation procedures specific to the DUC context may have paid off. Another possibility is that the proportion of difficult document clusters was smaller in 2006.

The pyramids differentiate document sets in addition to peer systems. This is evidenced by the significance of document set as a factor in predicting mean score, as well as by our qualitative observations on semantic differences within and across pyramids. Document sets that are more difficult for systems, as reflected by a lower average score across all peers, seem to have more complex topics, and fewer general, context-independent SCUs.

## 9 Acknowledgments

We thank Hoa Dang (NIST) and all the volunteers who made this evaluation possible, including all the sites who participated. Special thanks go to the volunteers who helped create the pyramids; Ben Gilbert (Microsoft), Qui Long (National University of Singapore), Inderjeet Mani (Mitre), and the co-authors at Columbia University; to participants who provided detailed critiques of the 2005 pyramid annotation or the 2006 pilot annotation: Lucy vanderwende (Microsoft), Guy LaPalme (University of Montreal).

## References

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.

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

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 (NAACL)*, Boston, MA.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, Pittsburgh, USA.

Rebecca Passonneau, Ani Nenkova, Kathleen McKeown, and Sergey Sigelman. 2005. Applying the pyramid method in DUC 2005. In *Proceedings of the 2005 DUC Workshop*, Vancouver, B.C.

Rebecca Passonneau. 2005. Evaluating an evaluation method: the pyramid method applied to 2003 document understanding conference (DUC) data. Technical Report CUCS-010-06, Columbia University Department of Computer Science.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.