Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

Rebecca Passonneau

Columbia University
New York, New York, USA
becky@cs.columbia.edu

Abstract

Annotation projects dealing with complex semantic or pragmatic phenomena face the dilemma of creating annotation schemes that oversimplify the phenomena, or that capture distinctions conventional reliability metrics cannot measure adequately. The solution to the dilemma is to develop metrics that quantify the decisions that annotators are asked to make. This paper discusses MASI, distance metric for comparing sets, and illustrates its use in quantifying the reliability of a specific dataset. Annotations of Summary Content Units (SCUs) generate models referred to as pyramids which can be used to evaluate unseen human summaries or machine summaries. The paper presents reliability results for five pairs of pyramids created for document sets from the 2003 Document Understanding Conference (DUC). The annotators worked independently of each other. Differences between application of MASI to pyramid annotation and its previous application to co-reference annotation are discussed. In addition, it is argued that a paradigmatic reliability study should relate measures of inter-annotator agreement to independent assessments, such as significance tests of the annotated variables with respect to other phenomena. In effect, what counts as sufficiently reliable intera-annotator agreement depends on the use the annotated data will be put to.

1. Introduction

To capture gradations in meaning or function, semantic and pragmatic annotation projects have taken various approaches. The project on Interlingual Annotation of Multilingual Text Corpora (IAMTC; Farwell et al., 2004) explicitly directed annotators to make multiple selections when no single selection seems sufficient. A related approach was taken in an email domain in which annotators were allowed to make multiple selections, but were asked to designate one as primary (Rosenberg & Binkowski, 2004). A contrasting implicit method relies on frequency of a category across multiple annotators to represent stronger or weaker presence of pragmatic units (Passonneau & Litman, 1997) or semantic ones (Nenkova & Passonneau, 2004).

Annotations with multiple choices or graded categories require new approaches to measuring agreement. Rosenberg & Binkowski (2004), for example, developed an augmented version of Cohen's kappa (1960) to assess inter-annotator agreement for the email domain. It yields different kappa scores, depending on the weight assigned to the primary selection. If the weight is 1, the secondary selection is ignored; if it is .5, both are considered equally. (Passonneau, 2004) presented a weighted metric for measuring agreement on set-valued items (referred to here as MASI) and compared it with other measures of agreement on co-reference annotation. MASI has also been applied to the IAMTC data (Passonneau et al., 2006). It will be demonstrated here for a semantic annotation task pertaining to the evaluation of automatic summarization: the creation of pyramids from human model summaries. A pyramid is a weighted model of the semantic content in a set of human model summaries (Nenkova & Passonneau, 2004), and can be used to score machine-generated summaries. It was used in the 2005 Document Understanding Conference (DUC) (Passonneau et al., 2005) and will be used in DUC 2006. This paper will address the inter-annotator reliability of pyramid construction for a DUC 2003 dataset.

Section two describes the annotation task, and gives an example of a representative pair of SCUs from pyramids created by different annotators. Section three gives an overview of a standard framework for assessing reliability, a definition and simple illustration of MASI, and a brief discussion of related work. Section four compares inter-annotator agreement results for the five pairs of pyramids using three metrics, including MASI. Together, the three metrics indicate a very high degree of overlap in pyramid annotations.

2. Pyramid Annotation Task

Summaries written by different humans will share information, but will also have information that does not appear in any other summary. This long observed fact was dramatically quantified by (van Halteren & Teufel, 2003) for a set of fifty summaries of a single source text.

Pyramids represent shared content in summaries by

having annotators select spans of words, or *contributors,*[1] from different summaries such that each expresses more or less the same information. We refer to a set of contributors as a Summary Content Unit (SCU). An SCU will have at most the same number of contributors as there are model summaries. The cardinality of an SCU, its *weight*, indicates how many of the model summaries expresses the given content. The set of all SCUs found in the models constitutes a pyramid. Annotators assign a label that serves as a mnemonic for the meaning.

A1's SCU: Weight=4
[Label: *Americans asked Saudi officials for help*]

Sum1	**<Saudi Arabian officials, under American pressure>1**
Sum2	**<sought help from Saudi officials>2**
Sum3	**<Through the Saudis, the United States asked>3**
Sum4	**<U.S. and Saudi Arabian requests>4**

A2's SCU: Weight=5
[Label: *Through the Saudis, the U.S. tried to get cooperation from the Taliban*]

Sum1	**<Saudi Arabian officials, under American pressure,>1 <asked Afghan leaders>5**
Sum2	**<U.S. and Saudi officials then attempted>6**
Sum3	**<sought help from Saudi officials>2, <who tried to convince Taliban leaders>7**
Sum4	**<Through the Saudis, the United States asked>3**
Sum5	**<U.S. and Saudi Arabian requests>4**

Equivalence classes:
A1 { 1, 2, 3, 4 } {5, 7} {6}
A2 { 1, 2, 3, 4, 5, 6, 7}

Figure 1. Semantically similar SCUs from two annotators, A1 and A2.

In the original annotation method (Passonneau & Nenkova, 2003), the contributors constituted equivalence classes over the words in the model summaries, and SCUs were equivalence classes over the contributors. This is the annotation style whose results are reported on here. For the annotations in DUC 2005 (Passonneau et al., 2005), the annotation constraints were relaxed to allow a word or phrase to be part of multiple contributors, and a contributor could be part of multiple SCUs.

For the 2003 Document Understanding Conference, NIST assembled thirty clusters of documents to use in the evaluation of automatic summarizers. In addition, four 100-word human summaries per document cluster (Docset) were collected. We recruited journalism majors, English majors and others in the Columbia University community who demonstrated high verbal skills (such as high verbal GREs scores) to write additional summaries. For five of the document clusters from DUC 2003, we had two annotators work independently to create pyramids, each using seven model summaries per Docset.

Figure 1 shows a pair of SCUs from the two independently annotated pyramids for one of the five Docsets (31038). It is typical of what we see from different annotators across the five pairs of pyramids investigated here. The two SCUs are very similar, but not identical. They differ in the weight (four versus five), and in the constituency of the contributors. By giving each span of words a unique identifier, we can see that there are two contributors that are the same for both annotators (spans 3 and 4), two that partly overlap (spans 1 and 2 for A1, versus A2's combination of spans 1 with 5, and of spans 2 with 7), and one that is unique to annotator A2 (span 6). Annotator A1 placed spans 5 and 7 in a distinct SCU, and span 6 was a singleton SCU.

The two annotators' equivalence classes (of spans rather than words) are shown at the bottom of Figure 1.

3. Inter-annotator Agreement

3.1. Standard Approach

Different types of tables have been used as a basis for computing inter-annotator agreement, including contingency tables, and simple agreement tables having rows for each unit and columns for each category and where cells record how often each unit was assigned each category. For two coders, Di Eugenio & Glass (2004) prefer contingency tables. They note that in comparison to contingency tables, simple agreement tables lose information about what choices individual coders make.

Contingency tables also lose information: they don't represent the coding units, and are inconvenient for more than two coders. An agreement matrix in Krippendorff's (1986) canonical form has one row per coder, and one column per coding unit. Cell values indicate the category **k** assigned by the ***i*th** coder to the ***j*th** unit. Such matrices lose no information, and can be used to tabulate counts of the number of coders who assigned the ***k*th** category to the ***j*th** unit, the number of categories assigned by the ***i*th** coder to the ***j*th** unit, and so on.

Apart from the assumptions used for computing the probability of the ***k*th**, most reliability metrics use the same or equivalent general formula to factor out chance agreement (Passonneau, 1997) (Arstein & Poesio, 2005). Where $p(A_O)$ and $p(A_E)$ are the probabilities of observed and expected agreement, the general formula is:

$$\frac{p(A_O) - p(A_E)}{1 - p(A_E)}$$

The metrics all have the same range: one for perfect agreement, to zero for no difference from chance, to values that approach minus one for ever greater than chance disagreement. The devil is in the details, namely how to estimate $p(A_E)$.

The family of metrics including Scott's pi (1955) and Siegel & Castellan's K (1988) use a single probability distribution for all coders, based on the observed rate of each category **k** across all coders. Cohen's kappa (1960) uses a distinct probability distribution for each coder

[1] A contributor can have discontinuities in the word string, e.g., for discontinuous constituents.

based on the rate at which the ***k*th** category appears in the ***i*th** coder's annotation. Cohen's (1960) kappa makes fewer assumptions, so in principle it provides stronger support for inferences about reliability. In practice, kappa may not always be the best choice.

Di Eugenio & Glass (2004) argue that kappa suffers from coder bias. The size of kappa will be relatively higher than Siegel & Castellan's K if two coders assign the categories **k** at different rates. Whether one views bias as an obstacle depends on one's goals. If the probability distributions over the values **k** are very different for two coders, then the probability that they will agree will necessarily be lower, and kappa accounts for this. Whether the difference in distribution arises from the inherent subjectivity of the task, insufficient specification in the annotation guidelines of when to use each category, or differences in the skill and attention of the annotators, cannot be answered by one metric in one comparison.

Artstein and Poesio (2005) review several families of reliability metrics, the associated assumptions, and differences in the resulting values that arise given the same data. The quantitative differences tend to be small. In order to illustrate the impact of different distance metrics, results are reported here using a single method of computing $p(A_E)$, Krippendorff's Alpha (1980).

The formula for Alpha, given m coders and r units, is:

$$\alpha = 1 - \frac{rm-1}{m} \frac{\sum_i \sum_b \sum_{c>b} n_{bi} n_{ci} \delta_{bc}}{\sum_b \sum_c n_b n_c \delta_{bc}}$$

The numerator is a summation over the product of counts of all pairs of values b and c, times the distance metric δ, across rows. The denominator is a summation of agreements and disagreements within columns. For categorical scales, because Alpha measures disagreements, δ is 0 when b=c, and 1 when $b \neq c$. For very large samples, Alpha is equivalent to Scott's (1955) pi; it corrects for small sample sizes, applies to multiple coders, and generalizes to many scales of annotation data.

Interpreting inter-annotator reliability raises two questions: what value of reliability is good enough, and how does one decide. Krippendorff (1980) is often cited as recommending a threshold of 0.67 to support cautious conclusions. The comment he made that introduced his discussion should be quoted more often. For the question of how reliable is reliable enough, he said: "there is no set answer" (p. 146). He offered the 0.67 threshold in the context of reliability studies in which the same variables also played a role in independent significance tests. In his data, variables below the 0.67 threshold happened never to be significant. He noted that in contrast, "some content analyses are very robust in the sense that unreliabilities become hardly noticeable in the result" (p. 147).

I will refer to the simultaneous investigation of reliability values of annotated data, and significance tests of the annotated variables with respect to independent measures, as a paradigmatic reliability study. (Passonneau et al., 2005) includes an analysis of the reliability of peer annotations for pyramid evaluation, and of the significance of correlations of pyramid scores using peer annotations from different annotators. It is a paradigmatic reliability study of *peer* annotation. The average Kappa across six document sets was .57, the average Alpha with Dice (1945) as a distance metric was 0.62, and Pearson's correlations were highly significant. A distance metric was used to count partial agreement for annotators who agreed that a given SCU occurred in a peer summary, but disagreed as to how often. MASI was not relevant here, because the counts of SCUs per summary did not constitute a unit of representation.

In concurrent work (Passonneau, 2005), we present results of a study in which the five pyramids discussed here were used to score summaries. Thus the present paper in combination with (Passonneau, 2005) constitutes a paradigmatic reliability study of *pyramid* annotation.

3.3. MASI

MASI is a distance metric for comparing two sets, much like an association measure such as Jaccard (1908) or Dice (1945). In fact, it incorporates Jaccard, as explained below. When used to weight the computation of inter-annotator agreement, it is independent of the method in which probability is computed, thus of the expected agreement. It can be used in any weighted agreement metric, such as Krippendorff's Alpha (Passonneau, 2004) or Artstein & Poesio's (2005) Beta[3] .

In (Passonneau, 2004), MASI was used for measuring agreement on co-reference annotations. Earlier work on assessing co-reference annotations did not use reliability measures of canonical agreement matrices, in part because of the data representation problem of determining what the coding values should be. The annotation task in co-reference does not involve selecting categories from a predefined set, but instead requires annotators to group expressions together into sets of those that co-refer.

(Passonneau, 2004) proposed a means for casting co-reference annotation into a conventional agreement matrix by treating the equivalence classes that annotators grouped NPs into as the coding values. Application of MASI for comparing the equivalence classes that annotators assign an NP to made it possible to quantify the degree of similarity across annotations. Since it is typically the case that annotators assign the same NP to very similar, but rarely identical, equivalence classes, applying an unweighted metric to the agreement matrices yields misleadingly low values.

The annotation task in creating pyramids has similar properties to the NP co-coreference annotation task. Neither the number of distinct referents, nor the number of distinct SCUs, is given in advance: both are the outcome of the annotation. The annotations both yield equivalence classes in which every NP token, or every word token, belongs to exactly one class (corresponding to a referent, or an SCU). NPs that are not grouped with other NPs (e.g., NPs annotated as non-referential), and words that are not grouped with other words (e.g., closed-class lexical items like "and" that contribute little or nothing to the semantics of an SCU, form singleton sets.

Figure 2 and Figure 3 schematically represent agreement matrices using set-based annotations. A3 and A4 stand for two annotators; **x**, **y** and **z** are the units from

which to create sets, and the coding values are the sets shown in the cells of the matrices.

	Units		
Annotator	**x**	**y**	**z**
A3	**{x, y}**	**{x, y}**	**{x}**
A4	**{x, y, z}**	**{x, y, z}**	**{x, y, z}**

Figure 2. Annotation with set subsumption.

	Units		
Annotator	**x**	**y**	**z**
A3	**{x, y }**	**{x, y}**	**{z}**
A4	**{x }**	**{ y, z}**	**{ y, z}**

Figure 3. Annotation with symmetric difference in column "y".

Figure 2 is like the SCU example in Figure 1 in that there is a monotonic relationship among all the sets in the matrix. Within columns, A3's sets always share properties with A4's sets, and there are no conflicting properties. This is not the case in Figure 3, where A3 has a set **{x,y}**, and A4 has a set **{y,z}.** The two sets have a non-null intersection (**{y}**), and non-null set-differences (**{x},{z}**). Figure 3 represents a case where A3 thinks **x** and **y** have the same set of properties, not shared by **z**; A3 thinks **y** and **z** have the same set of properties, not shared by **x**, thus the semantic or pragmatic elements being represented are in conflict.

MASI ranges from 1, when two sets are identical, to 0, when they are disjoint. It has two terms which weight different aspects of set comparison: MASI = J*M. The Jaccard (1908) metric (the J term) is used to weight the differences in size of two sets, independent of whether sets are monotonic. The M term is for monotonicity, and penalizes a case like Figure 3 more heavily than Figure 2. Their role in computing MASI will now be illustrated.

Spans	A1	A2
1	**{ 1, 2, 3, 4 }**	**{ 1, 2, 3, 4, 5, 6, 7}**
2	**{ 1, 2, 3, 4 }**	**{ 1, 2, 3, 4, 5, 6, 7}**
3	**{ 1, 2, 3, 4 }**	**{ 1, 2, 3, 4, 5, 6, 7}**
4	**{ 1, 2, 3, 4 }**	**{ 1, 2, 3, 4, 5, 6, 7}**
5	**{5, 7}**	**{ 1, 2, 3, 4, 5, 6, 7}**
6	**{6}**	**{ 1, 2, 3, 4, 5, 6, 7}**
7	**{5, 7}**	**{ 1, 2, 3, 4, 5, 6, 7}**

Figure 4. Agreement matrix for Figure 1, using spans (instead of words) as the coding units.

Taking the two MASI terms in turn, J is the ratio of the cardinality of the intersection to the cardinality of the union of the two sets. For two sets P and Q, it is one if P=Q, and grows closer to one the more members P and Q have in common. J is zero if P and Q are disjoint, and is closer to zero the larger P and Q are, and the fewer members they have in common.

The value of Jaccard is 2/3 for the **x** and **y** columns of Figure 2, and 1/3 for the **z** column. Similarly, it is 2/3 for the **y** column of Figure 3, and 1/3 for the **x** and **z** columns. The mean Jaccard for Figure 2 is 5/9, and for Figure 3 it is 4/9. Thus Figure 2 is appropriately closer to one than Figure 3, but the quantitative difference is small.

The second term of MASI (M, for monotonicity) penalizes the case in Figure 3 more heavily than that in Figure 2. If two sets Q and P are identical, M is 1. If one set is a subset of the other, M is 2/3. If the intersection and the two set differences are all non-null, then M is 1/3. If the sets are disjoint, M is 0.

Before comparing the sets assigned by A3 and A4 to a coding unit **y**, the coding unit itself must be removed. Otherwise, the coding values will necessarily intersect. For column y in Figure 2, {**x**} would be compared with {**x, z**}. For column **y** in Figure 3, {**x**} would be compared with {**z**}. The mean MASI for Figure 2 is 10/27 (.37) and for Figure 3 it is 6/27 (.22).

A1's SCU-101: Weight=2
[Label: *Worker's agree to Estrada's terms*]

Sum1	**<61% voted yes>1**
Sum2	**<Unions agreed to some employee cuts with separation benefits>2**

A1's SCU-102: Weight=1
[Label: *Ground crew accepts 2 weeks after initial rejection*]

Sum3	**<which it accepted two weeks later>3**

A2's SCU-201: Weight=2
[Label: The settlement was finally accepted]

Sum1	**<61% voted yes>1**
Sum2	**<which it accepted two weeks later>3**

A2's SCU-202: Weight=1
[Label: Unions agree to some employee cuts]

Sum3	**< Unions agreed to some employee cuts with separation benefits >2**

Figure 5. Pairs of SCUs from two annotators illustrating non-monotonicity.

Figure 4 shows a canonical agreement matrix for the example from Figure 1; to save space it is presented with the coding units in rows instead of columns. For the sake of illustration, the coding units are spans instead of words. The set of coding categories consists of the equivalence classes from both annotations. Annotator A2 placed all the spans shown in a single SCU labeled ***[Through the Saudis, the U.S. tried to get cooperation from the Taliban***]. Annotator A1 created a similar SCU labeled, [***Americans asked Saudi officials for help*]**, but did not include spans 5 ("*asked Afghan leaders")* and 7 ("who tried to convince Taliban leaders"). A1 placed 5 and 7 in a distinct SCU (with other contributing spans, omitted from discussion), labeled **[Saudi officials asked Afghan leaders to release Bin Laden]**.

The labels assigned by A1 and A2 in Figure 1 reflect the difference in content. A2 chose a more comprehensive label expressing a 3-way relation in which

the Saudis would mediate between the U.S. and the Taliban. In comparison, A1's labels describe two binary relations, one relating the U.S. and the Saudis, and one relating the Saudis and the Taliban. The labels would suggest that A2's annotation subsumes A1's, and the SCU representation confirms this.

In contrast to the SCU example illustrated in Figure 1, we occasionally find groups of SCUs across annotators that are semantically more distinct, corresponding to cases like Figure 3. Figure 5 gives an example from a pyramid whose reliability was reported on in (Nenkova & Passonneau, 2004).[2]

Table 1 shows the reliability values for the data from Figure 1 using Krippendorff's Alpha with three different distance metrics. Because Krippendorff's Alpha measures disagreements, one minus Jaccard, and one minus MASI, are used in computing Alpha. The "Nominal" column shows the results treating all non-identical sets as categorically distinct (see section 3.1). For illustrative purposes, the top portion of the table uses spans as the coding units, i.e., computing Alpha from the agreement matrix given in Figure 4. Since spans were not given in advance, but were decided on by coders, this underestimates the number of decisions that annotators were required to make. The very low value in the Jaccard column is due to the disparity in size between the two annotations for rows five through seven of Figure 4.

The lower portion of Table 1 shows the results using words as the coding units. The values across the three columns are similar to those for the full dataset as we will see in the discussion of Table 2.

	Alpha		
Coding units	Nominal	Jaccard	MASI
spans	0	-.44	0.14
words	0	.64	.81

Table 1. Reliability values for data from Figure 1, using spans versus words as coding units.

3.4. Related Work

As noted above, Teufel and van Halteren (2004) perform an annotation addressing a goal similar to the pyramid method. They create lists of factoids, atomic units of information. To compare sets of factoids that were independently created by two annotators, they first create a list of subsumption relations between factoids across annotations. Then they construct a table that lists all (subsumption-relation, summary) pairs, with counts of how often each subsumption relation occurs in each summary. Figure 6 reproduces their Figure 2. Every factoid is given an index, and in Figure 6, P30 represents a factoid created by one annotator that subsumes two created by the other annotator. Symbols **a** through **e** represent five summaries. They compute kappa from this type of agreement table.

[2] SCU-201 has been simplified for illustrative purposes; in the actual data, it had a third contributor.

	A1	A2		A1	A2
P30 ←F9.21 -a	1	1	P30 ←F9.22 -a	1	0
P30 ←F9.21 -b	0	0	P30 ←F9.22 -b	0	0
P30 ←F9.21 -c	1	0	P30 ←F9.22 -c	1	1
P30 ←F9.21 -d	0	0	P30 ←F9.22 -d	0	0
P30 ←F9.21 -e	1	0	P30 ←F9.22 -e	1	1

Figure 6. Agreement table representation used in Teufel and van Halteren (2004).

While this representation does not suffer from the loss of information De Eugenio & Glass (2004) fault Siegel & Castellan (1988) for, note that it differs from an agreement matrix or a contingency table in that it is not the case that each count represents an individual decision made by an annotator. We can see from the table that A1 is the annotator who created P30 and A2 is the one who created F9.21 and F9.22. Although there are two cells in A1's column for the two subsumption relations P30 ← F9.21 and P30←F9.22, it is unlikely that A1's original annotation involved decisions about F9.21 and F9.22. If the number of decisions is overestimated, $p(A_E)$ will be underestimated, leading to higher kappa values.

Another issue in using such an agreement table from two independently created factoid lists is that it requires the creation of a new level of representation that would itself be subject to reliability issues.

4. Results and Discussion

Canonical greement matrices of the form shown in Figure 4, but with words as the coding units, were computed for the five pairs of independently created pyramids for the Docsets listed in Table 2. The mean number of words per pyramid was 725; the mean number of distinct SCUs was 92. Results are shown for Alpha with the same three distance metrics used in Table 1.

Alpha			
Docset	Nominal	Jaccard	MASI
30016	.19	.55	0.79
30040	.24	.58	0.80
31001	.01	.40	0.68
31010	.03	.39	0.69
31038	.09	.40	0.71

Table 2. Inter-annotator agreement on 5 pyramids using unweighted Krippendorff's Alpha (nominal), and Alpha with Jaccard and MASI as δ.

The low values for the nominal distance metric are expected, given that there are few cases of word-for-word identity of SCUs across annotations. With Jaccard as the distance metric, the values increase manyfold, indicating that over all the comparisons of pairs of SCUs across annotators for a given pyramid, the size of the set intersection is closer to the size of the set union than not.

With MASI, values increase by approximately half of the difference between the Jaccard value and the maximum value of one. Since MASI rewards overlapping sets twice as much if one is a subset of the other than if they are not, this degree of increase indicates that most of the differences between SCUs are monotonic.

By including several metrics whose relationship to each other is known, Table 2 indicates that the pyramid annotations do not have many cases of exact agreement (nominal), that the sets being compared have more members in common than not (Jaccard), and that the commonality is more often monotonic than not (MASI).

Whether these results are sufficiently reliable depends on the uses of the data. In a separate investigation (Passonneau, 2005), the pairs of pyramids for Docsets 30016 and 30014 have been used to produce parallel sets of scores for summaries from sixteen summarization systems that participated in DUC 2003. Pearson's correlations of two types of scores (original pyramid and modified) range from 0.84 to 0.91with p values always zero. This constitutes evidence that the pyramid annotations are more than reliable enough.

5. Conclusion

Measuring inter-annotator reliability involves more than a single number or a single study. Di Eugenio & Glass (2004) argue that using multiple reliability metrics with different methods for computing $p(A_E)$ can be more revealing of than a single metric. Passonneau et al. (2005) present a similar argument for the case of comparing different distance metrics. Here, inter-annotator reliability results have been presented using three metrics in order to more fully characterize the dataset.

This paper argues that full interpretation of a reliability measure is best carried out in a paradigmatic reliability study: a series of studies that link one or more measures of the reliability of a dataset to an independent assessment, such as a significance test. If the same dataset is used in different tasks, what is reliable for one task may not be for another.

Investigators faced with complex annotation data have shown ingenuity in proposing new data representations (Teufel & van Halteren, 2004), new reliability measures (Rosenberg & Binkowski, 2004), and techniques new to computational linguistics, as discussed in (Artstein & Poesio). While this paper argues for placing a greater burden on the interpretation of inter-annotator agreement, proposals such as these provide an expanding suite of tools for accomplishing this task.

Acknowledgments

This work was supported by DARPA NBCH105003 and NUU01-00-1-8919. The author thanks many annotators, especially Ani Nenkova and David Elson.

References

Artstein, R. and M. Poesio. 2005. Kappa^3 = Alpha (or Beta). University of Essex NLE Technote 2005-01.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26:297-302.

Farwell, D.; Helmreich, S.; Dorr, B. J.; Habash, N.; Reeder, F.; Miller, K.; Levin, L.; Mitamura, T.; Hovy, E.; Rambow, O.; Siddharthan, A. (2004). Interlingual annotation of multilingual text corpora. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Frontiers in Corpus Annotation*, Boston, MA, pp. 55-62, 2004.

van Halteren, H. and S. Teufel. 2003. Examining the consensus between human summaries. In *Proceedings of the Document Understanding Workshop*.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 44:223-270.

Krippendorff, K. 1980. *Content Analysis*. Newbury Park, CA: Sage Publications.

Nenkova, A. and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *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.

Passonneau, R.; Nenkova, A.; McKeown, K.; Sigelman, S. 2005. Applying the pyramid method in DUC 2005. *Document Understanding Conference Workshop*.

Passonneau, R.; Habash, N.; Rambow, O. 2006. Inter-annotator agreement on a multilingual semantic annotation task. *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). Genoa, Italy.

Passonneau, R. 2005. Evaluating an evaluation method: The pyramid method applied to 2003 Document Understanding Conference (DUC) Data. Technical Report CUCS-010-06, Department of Computer Science, Columbia University.

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

Passonneau, R. 1997. Applying reliability metrics to co-reference annotation. Technical Report CUCS-017-97, Department of Computer Science, Columbia University.

Passonneau, R. and D. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics* 23.1: 103-139.

Rosenberg, A. and Binkowski, E. (2004). Augmenting the kappa statistic to determine inter-annotator reliability for multiply labeled data points. In *Proceedings of the Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*.

Siegel, S. and N. John Castellan, Jr. (1988) *Non-parametric Statistics for the Behavioral Sciences*, 2nd edition. McGraw-Hill, New York.

Teufel, S. and H. van Halteren, 2004: Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of Empirical Methods in Natural Language Processing*. Barcelona.