Creating a Gold Standard for Sentence Clustering in Multi-Document Summarization

Johanna Geiss
University of Cambridge
Computer Laboratory
15 JJ Thomson Avenue
Cambridge, CB3 0FD, UK
johanna.geiss@cl.cam.ac.uk

Abstract

Sentence Clustering is often used as a first step in Multi-Document Summarization (MDS) to find redundant information. All the same there is no gold standard available. This paper describes the creation of a gold standard for sentence clustering from DUC document sets. The procedure of building the gold standard and the guidelines which were given to six human judges are described. The most widely used and promising evaluation measures are presented and discussed.

1 Introduction

The increasing amount of (online) information and the growing number of news websites lead to a debilitating amount of redundant information. Different newswires publish different reports about the same event resulting in information overlap. Multi-Document Summarization (MDS) can help to reduce the amount of documents a user has to read to keep informed. In contrast to single document summarization information overlap is one of the biggest challenges to MDS systems. While repeated information is a good evidence of importance, this information should be included in a summary only once in order to avoid a repetitive summary. Sentence clustering has therefore often been used as an early step in MDS (Hatzivassiloglou et al., 2001; Marcu and Gerber, 2001; Radev et al., 2000). In sentence clustering semantically similar sentences are grouped together. Sentences within a cluster overlap in information, but they do not have to be identical in meaning. In contrast to paraphrases sentences in a cluster do not have to cover the same amount of information. One sentence represents one cluster in the summary. Either a sentences from the cluster is selected (Aliguliyev, 2006) or a new sentence is regenerated from all/some sentences in a cluster (Barzilay and McKeown, 2005). Usually the quality of the sentence clusters are only evaluated indirectly by judging the quality of the generated summary. There is still no standard evaluation method for summarization and no consensus in the summarization community how to evaluate a summary. The methods at hand are either superficial or time and resource consuming and not easily repeatable. Another argument against indirect evaluation of clustering is that troubleshooting becomes more difficult. If a poor summary was created it is not clear which component e.g. information extraction through clustering or summary generation (using for example language regeneration) is responsible for the lack of quality.

However there is no gold standard for sentence clustering available to which the output of a clustering systems can be compared. Another challenge is the evaluation of sentence clusters. There are a lot of evaluation methods available. Each of them focus on different properties of a set of clusters. We will discuss and evaluate the most widely used and most promising measures. In this paper the main focus is on the development of a gold standard for sentence clustering using DUC clusters. The guidelines and rules that were given to the human annotators are described and the interjudge agreement is evaluated.

2 Related Work

Sentence Clustering is used for different application in NLP. Radev et al. (2000) use it in their MDS system MEAD. The centroids of the clusters are used to create a summary. Only the summary is evaluated, not the sentence clusters. The same applies to Wang et al. (2008). They use symmetric matrix factorisation to group similar sentences together and test their system on DUC2005 and DUC2006 data set, but do not evaluate the clusterings. However Zha (2002) created a gold stan-

Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, pages 96–104,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

dard relying on the section structure of web pages and news articles. In this gold standard the section numbers are assumed to give the true cluster label for a sentence. In this approach only sentences within the same document and even within the same paragraph are clustered together whereas our approach is to find similar information between documents.

A gold standard for event identification was built by Naughton (2007). Ten annotators tagged events in a sentence. Each sentence could be assigned more than one event number. In our approach a sentence can only belong to one cluster.

For the evaluation of SIMFINDER Hatzivassiloglou et al. (2001) created a set of 10.535 manually marked pairs of paragraphs. Two human annotator were asked to judge if the paragraphs contained 'common information'. They were given the guideline that only paragraphs that described the same object in the same way or in which the same object was acting the same are to be considered similar. They found significant disagreement between the judges but the annotators were able to resolve their differences. Here the problem is that only pairs of paragraphs are annotated whereas we focus on whole sentences and create not pairs but clusters of similar sentences.

3 Data Set for Clustering

The data used for the creation of the gold standard was taken from the Document Understanding Conference (DUC)[1] document sets. These document clusters were designed for the DUC tasks which range from single-/multi-document summarization to update summaries, where it is assumed that the reader has already read earlier articles about an event and requires only an update of the newer development. Since DUC has moved to TAC in 2008 they focus on the update task. In this paper only clusters designed for the general multi-document summarization task are used.

Our clustering data set consists of four sentence sets. They were created from the document sets d073b (DUC 2002), D0712C (DUC 2007), D0617H (DUC 2006) and d102a (DUC 2003). Especially the newer document clusters e.g. from DUC 2006 and 2007 contain a lot of documents. In order to build good sentence clusters the judges have to compare each sentence to each other sentence and maintain an overview of the topics within the documents. Because of human cognitive limitations the number of documents and sentences have to be reduced. We defined a set of constraints for a sentence set: (i) from one set, (ii) a sentence set should consist of 150 – 200 sentences[2]. To obtain sentence sets that comply with these requirements we designed an algorithm that takes the number of documents in a DUC set, the date of publishing, the number of documents published on the same day and the number of sentences in a document into account. If a document set includes articles published on the same day they were given preference. Furthermore shorter documents (in terms of number of sentences) were favoured. The properties of the resulting sentence sets are listed in table 1. The documents in a set were ordered by date and split into sentences using the sentence boundary detector from RASP (Briscoe et al., 2006).

name	DUC	DUC id	docs	sen
Volcano	2002	D073b	5	162
Rushdie	2007	D0712C	15	103
EgyptAir	2006	D0617H	9	191
Schulz	2003	d102a	5	248

Table 1: Properties of sentence sets

4 Creation of the Gold Standard

Each sentence set was manually clustered by at least three judges. In total there were six judges which were all volunteers. They are all second-language speakers of English and hold at least a Master's degree. Three of them (Judge_A, Judge_J and Judge_O) have a background in computational linguistics. The judges were given a task description and a list of guidelines. They were only using the guidelines given and worked independently. They did not confer with each other or the author. Table 2 gives details about the set of clusters each judge created.

4.1 Guidelines

The following guidelines were given to the judges:

1. Each cluster should contain only one topic.
2. In an ideal cluster the sentences are very similar.

[1] DUC has now moved to the Text Analysis Conference (TAC)

[2] If a DUC set contains only 5 documents all of them are used to create the sentence set, even if that results in more than 200 sentences. If the DUC set contains more than 15 documents, only 15 documents are used for clustering even if the number of 150 sentences is not reached.

judge	Rushdie			Volcano			EgyptAir			Schulz		
	s	c	s/c	s	c	s/c	s	c	s/c	s	c	s/c
Judge_A	70	15	4.6	92	30	3	85	28	3	54	16	3.4
Judge_B	41	10	4.1	57	21	2.7	44	15	2.9	38	11	3.5
Judge_D				46	16	2.9						
Judge_H	74	14	5.3				75	19	3.9			
Judge_J										120	7	17.1
Judge_O							53	20	2.6			

Table 2: Details of manual clusterings: *s* number of sentences in a set, *c* number of clusters, *s/c* average number of sentences in a cluster

3. The information in one cluster should come from as many different documents as possible. The more different sources the better. Clusters of sentences from only one document are not allowed.
4. There must be at least two sentences in a cluster, and more than two if possible.
5. Differences in numbers in the same cluster are allowed (e.g. vagueness in numbers (300,000 - 350,000), update (two killed - four dead))
6. Break off very similar sentences from one cluster into their own subcluster, if you feel the cluster is not homogeneous.
7. Do not use too much inference.
8. Partial overlap – If a sentence has parts that fit in two clusters, put the sentence in the more important cluster.
9. Generalisation is allowed, as long as the sentences are about the same person, fact or event.

The guidelines were designed by the author and her supervisor – Dr Simone Teufel. The starting point was a single DUC document set which was clustered by the author and her supervisor with the task in mind to find clusters of sentences that represent the main topics in the documents. The minimal constraint was that each cluster is specific and general enough to be described in one sentence (see rule 1 and 2). By looking at the differences between the two manual clustering and reviewing the reasons for the differences the other rules were generated and tested on another sentence set.

One rule that emerged early says that a topic can only be included in the summary of a document set if it appears in more than one document (rule 3). From our understanding of MDS and our definition of importance only sentences that depict a topic which is present in more than one source document can be summary worthy. From this it follows that clusters must contain at least two sentences which come from different documents. Sentences that are not in any cluster of at least two are considered irrelevant for the MDS task (rule 4). We defined a spectrum of similarity. In an ideal cluster the sentences would be very similar, almost paraphrases. For our task sentences that are not paraphrases can be in the same cluster (see rule 5, 8, 9). In general there are several constraints that pull against each other. The judges have to find the best compromise.

We also gave the judges a recommended procedure:

1. Read all documents. Start clustering from the first sentence in the list. Put every sentence that you think will attract other sentences into an initial cluster. If you feel, that you will not find any similar sentences to a sentence, put it immediately aside. Continue clustering and build up the clusters while you go through the list of sentences.
2. You can rearrange your clusters at any point.
3. When you are finished with clustering check that all important information from the documents is covered by your clusters. If you feel that a very important topic is not expressed in your clusters, look for evidence for that information in the text, even in secondary parts of a sentence.
4. Go through your sentences which do not belong to any cluster and check if you can find a suitable cluster.
5. Do a quality check and make sure that you wrote down a sentence for each cluster and that the sentences in a cluster are from more than one document.
6. Rank the clusters by importance.

4.2 Differences in manual clusterings

Each judge clustered the sentence sets differently. No two judges came up with the same separation into clusters or the same amount of irrelevant sentences. When analysing the differences between the judges we found three main categories:

Generalisation One judge creates a cluster that from his point of view is homogeneous:

1. Since then, the Rushdie issue has turned into a big controversial problem that hinders the relations between Iran and European countries.
2. The Rushdie affair has been the main hurdle in Iran's efforts to improve ties with the European Union.

3. In a statement issued here, the EU said the Iranian decision opens the way for closer cooperation between Europe and the Tehran government.
4. "These assurances should make possible a much more constructive relationship between the United Kingdom, and I believe the European Union, with Iran, and the opening of a new chapter in our relations," Cook said after the meeting.

Another judge however puts these sentences into two separate cluster (1,2) and (3,4).The first judge chooses a more general approach and created a cluster about the relationship between Iran and the EU, whereas the other judge distinguishes between the improvement of the relationship and the reason for the problems in the relationship.

Emphasise Two judges can emphasise on different parts of a sentence. For example the sentence "All 217 people aboard the Boeing 767-300 died when it plunged into the Atlantic off the Massachusetts coast on Oct. 31, about 30 minutes out of New York's Kennedy Airport on a night flight to Cairo." was clustered together with other sentence about the number of casualties by one judge. Another judge emphasised on the course of events and put it into a different cluster.

Inference Humans use different level of interference. One judge clustered the sentence "Schulz, who hated to travel, said he would have been happy living his whole life in Minneapolis." together with other sentences which said that Schulz is from Minnesota although this sentence does not clearly state this. This judge interfered from "he would have been happy living his whole life in Minneapolis" that he actually is from Minnesota.

5 Evaluation measures

The evaluation measures will compare a set of clusters to a set of classes. An ideal evaluation measure should reward a set of clusters if the clusters are pure or homogeneous, so that it only contains sentences from one class. On the other hand it should also reward the set if all/most of the sentences of a class are in one cluster (completeness). If sentences that in the gold standard make up one class are grouped into two clusters, the measure should penalise the clustering less than if a lot of irrelevant sentences were in the same cluster. Homogeneity is more important to us.
D is a set of N sentences d_a so that $D = \{d_a|a = 1, ..., N\}$. A set of clusters $L = \{l_j|j = 1, ..., |L|\}$ is a partition of a data set D into disjoint subsets called clusters, so that $l_j \cap l_m = \emptyset$. $|L|$ is the number of clusters in L. A set of clusters that contains only one cluster with all the sentences of D will be called L_{one}. A cluster that contains only one object is called a singleton and a set of clusters that only consists of singletons is called L_{single}.
A set of classes $C = \{c_i|i = 1, ..., |C|\}$ is a partition of a data set D into disjoint subsets called classes, so that $c_i \cap c_m = \emptyset$. $|C|$ is the number of classes in C. C is also called a gold standard of a clustering of data set D because this set contains the "ideal" solution to a clustering task and other clusterings are compared to it.

5.1 V-measure and V_{beta}

The V-measure (Rosenberg and Hirschberg, 2007) is an external evaluation measure based on conditional entropy:

$$V(L, C) = \frac{(1 + \beta)hc}{\beta h + c} \quad (1)$$

It measures homogeneity (h) and completeness (c) of a clustering solution (see equation 2 where n^i_j is the number of sentences l_j and c_i share, n_i the number of sentences in c_i and n_j the number of sentences in l_j)

$$h = 1 - \frac{H(C|L)}{H(C)} \qquad c = 1 - \frac{H(L|C)}{H(L)}$$

$$H(C|L) = -\sum_{j=1}^{|L|}\sum_{i=1}^{|C|} \frac{n^i_j}{N} log \frac{n^i_j}{n_j}$$

$$H(C) = -\sum_{i=1}^{|C|} \frac{n^i}{N} log \frac{n^i}{N} \quad (2)$$

$$H(L) = -\sum_{j=1}^{|L|} \frac{n^j}{N} log \frac{n^j}{N}$$

$$H(L|C) = -\sum_{i=1}^{|C|}\sum_{j=1}^{|L|} \frac{n^i_j}{N} log \frac{n^i_j}{n_i}$$

A cluster set is homogeneous if only objects from a single class are assigned to a single cluster. By calculating the conditional entropy of the class distribution given the proposed clustering it can be measured how close the clustering is to complete homogeneity which would result in zero entropy. Because conditional entropy is constrained by the size of the data set and the distribution of the class sizes it is normalized by $H(C)$ (see equation 2). Completeness on the other hand is achieved if all data points from a single class are assigned to a single cluster which results in $H(L|C) = 0$.

The V-measure can be weighted. If $\beta > 1$ the completeness is favoured over homogeneity whereas the weight of homogeneity is increased if $\beta < 1$.

Vlachos et al. (2009) proposes V_{beta} where β is set to $\frac{|L|}{|C|}$. This way the shortcoming of the V-measure to favour cluster sets with many more clusters than classes can be avoided. If $|L| > |C|$ the weight of homogeneity is reduced, since clusterings with large $|L|$ can reach high homogeneity quite easily, whereas $|C| > |L|$ decreases the weight of completeness. V-measure and V_{beta} can range between 0 and 1, they reach 1 if the set of clusters is identical to the set of classes.

5.2 Normalized Mutual Information

Mutual Information (I) measures the information that C and L share and can be expressed by using entropy and conditional entropy:

$$I = H(C) + H(L) - H(C, L) \tag{3}$$

There are different ways to normalise I. Manning et al. (2008) uses

$$NMI = \frac{I(L,C)}{\frac{H(L)+H(C)}{2}} = 2\frac{I(L,C)}{H(L) + H(C)} \tag{4}$$

which represents the average of the two uncertainty coefficients as described in Press et al. (1988).

Generalise NMI to $NMI_\beta = \frac{(1+\beta)I}{\beta H(L)+H(C)}$. Then NMI_β is actually the same as V_β:

$$\begin{gathered} h = 1 - \frac{H(C|L)}{H(C)} \\ \Rightarrow H(C)h = H(C) - H(C|L) \\ = H(C) - H(C,L) + H(L) = I \\ \\ c = 1 - \frac{H(L|C)}{H(L)} \\ \Rightarrow H(L)c = H(L) - H(L|C) \\ = H(L) - H(L,C) + H(C) = I \\ V = \frac{(1+\beta)hc}{\beta h + c} \\ = \frac{(1+\beta)H(L)H(C)hc}{\beta H(L)H(C)h + H(L)H(C)c} \end{gathered} \tag{5}$$

$H(C)h$ and $H(L)c$ are substituted by I:

$$\begin{gathered} \frac{(1+\beta)I^2}{\beta H(L)I + H(C)I} \\ = \frac{(1+\beta)I}{\beta H(L) + H(C)} = NMI_\beta \\ V_1 = 2\frac{I}{H(L) + H(C)} = NMI \end{gathered} \tag{6}$$

5.3 Variation of Information (VI) and Normalized VI

The VI-measure (Meila, 2007) also measures completeness and homogeneity using conditional entropy. It measure the distance between two clusterings and thereby the amount of information gained in changing from C to L. For this measure the conditional entropies are added up:

$$VI(L, C) = H(C|L) + H(L|C) \tag{7}$$

Remember small conditional entropies mean that the clustering is near to complete homogeneity/ completeness, so the smaller VI the better ($VI = 0$ if $L = C$). The maximum of VI is $log\ N$ e.g. for $VI(L_{single}, C_{one})$. VI can be normalized, then it can range from 0 (identical clusters) to 1.

$$NVI(L, C) = \frac{1}{log\ \mathrm{N}} VI(L, C) \tag{8}$$

V-measure, V_{beta} and VI measure both completeness and homogeneity, no mapping between classes and clusters is needed (Rosenberg and Hirschberg, 2007) and they are only dependent on the relative size of the clusters (Vlachos et al., 2009).

5.4 Rand Index (RI)

The Rand Index (Rand, 1971) compares two clusterings with a combinatorial approach. Each pair of objects can fall into one of four categories:

- TP (true positives) = objects belong to one class and one cluster
- FP (false positives) = objects belong to different classes but to the same cluster
- FN (false negatives) = objects belong to the same class but to different clusters
- TN (true negatives) = objects belong to different classes and to different cluster

By dividing the total number of correctly clustered pairs by the number of all pairs, RI gives the percentage of correct decisions.

$$RI = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

RI can range between 0 and 1 where 1 corresponds to identical clusterings. Meila (2007) mentions that in practise RI concentrates in a small interval near 1 (for more detail see section 5.7). Another shortcoming is that RI gives equal weight to FPs and FNs.

5.5 Entropy and Purity

Entropy and Purity are widely used evaluation measures (Zhao and Karypis, 2001). They both can be used to measure homogeneity of a cluster. Both measures give better values when the number of clusters increase, with the best result for L_{single}. Entropy ranges from 0 for identical clusterings or L_{single} to $log\ N$ e.g. for C_{single} and L_{one}. The values of P can range between 0 and 1, where a value close to 0 represents a bad clustering solution and a perfect clustering solution gets a value of 1.

$$\begin{aligned} Entropy &= \sum_{j=1}^{|L|} \frac{n_j}{N} \left(-\frac{1}{\log |C|} \sum_{i=1}^{|C|} \frac{n_j^i}{n_j} \log \frac{n_j^i}{n_j} \right) \\ Purity &= \frac{1}{N} \sum_{j=1}^{|L|} \max_i \left(n_j^i \right) \end{aligned} \tag{10}$$

5.6 F-measure

The F-measure is a well known metric from IR, which is based on Recall and Precision. The version of the F-score (Hess and Kushmerick, 2003) described here measures the overall Precision and Recall. This way a mapping between a cluster and a class is omitted which may cause problems if $|L|$ is considerably different to $|C|$ or if a cluster could be mapped to more than one class. Precision and Recall here are based on pairs of objects and not on individual objects.

$$\begin{aligned} P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \\ F(L, C) = \frac{2PR}{P + R} \end{aligned} \tag{11}$$

5.7 Discussion of the Evaluation measures

We used one cluster set to analyse the behaviour and quality of the evaluation measures. Variations of that cluster set were created by randomly splitting and merging the clusters. These modified sets were then compared to the original set. This experiment will help to identify the advantages and disadvantages of the measures, what the values reveal about the quality of a set of clusters and how the measures react to changes in the cluster set.

We used the set of clusters created by Judge_A for the Rushdie sentence set. It contains 70 sentences in 15 clusters. This cluster set was modified by splitting and merging the clusters randomly until we got L_{single} with 70 clusters and L_{one} with one cluster. The original set of clusters (C_A) was compared to the modified versions of the set (see figure 1). The evaluation measures reach their best values if $C_A = 15$ clusters is compared to itself.

The F-measure is very sensitive to changes. It is the only measure which uses its full measurement range. $F = 0$ if C_A is compared to $L_{A-single}$, which means that the F-measure considers $L_{A-single}$ to be the opposite of C_A. Usually L_{one} and $L_{A-single}$ are considered to be observe and a measure should only reach its worst possible value if these sets are compared. In other words the F-measure might be too sensitive for our task.

The RI stays most of the time in an interval between 0.84 and 1. Even for the comparison between C_A and $L_{A-single}$ the RI is 0.91. This behaviour was also described in Meila (2007) who observed that the RI concentrates in a small interval near 1.

As described in section 5.5 Purity and Entropy both measure homogeneity. They both react to changes slowly. Splitting and merging have almost the same effect on Purity. It reaches ≈ 0.6 when the clusters of the set were randomly split or merged four times. As explained above our ideal evaluation measure should punish a set of clusters which puts sentences of the same class into two clusters less than if sentences are merged with irrelevant ones. Homogeneity decreases if unrelated clusters are merged whereas a decline in completeness follows from splitting clusters. In other words for our task a measure should decrease more if two clusters are merged than if a cluster is split.

Entropy for example is more sensitive to merging than splitting. But Entropy only measures homogeneity and an ideal evaluation measure should also consider completeness.

The remaining measures V_{beta}, $V_{0.5}$ and NVI/VI all fulfil our criteria of a good evaluation measure. All of them are more affected by merging than by splitting and use their measuring range appropriately. $V_{0.5}$ favours homogeneity over completeness, but it reacts to changes less than V_{beta}. The V-measure can also be inaccurate if the $|L|$ is considerably different to $|C|$. V_{beta} (Vlachos et al., 2009) tries to overcome this problem and the tendency of the V-measure to favour clusterings with a large number of clusters.

Since VI is measured in bits with an upper bound of $log\ N$, values for different sets are difficult to compare. NVI tries to overcome this problem by

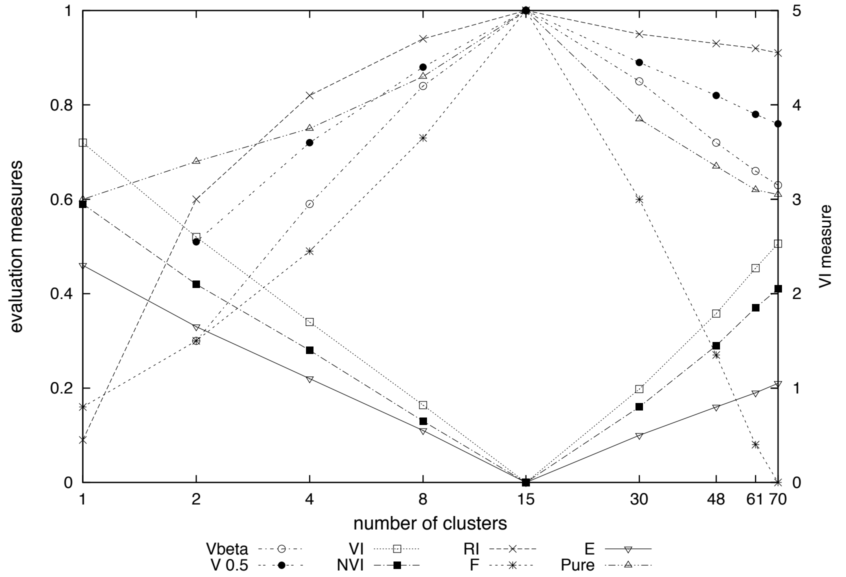

Figure 1: Behaviour of evaluation measure when randomly changed sets of clusters are compared to the original set.

normalising VI by dividing it by $log\ N$. As Meila (2007) pointed out, this is only convenient if the comparison is limited to one data set.

In this paper V_{beta}, $V_{0.5}$ and NVI will be used for evaluation purposes.

6 Comparability of Clusterings

Following our procedure and guidelines the judges have to filter out all irrelevant sentences that are not related to another sentence from a different document. The number of these irrelevant sentences are different for every sentence set and every judge (see table 2). The evaluation measures require the same number of sentences in each set of clusters to compare them. The easiest way to ensure that each cluster set for a sentence set has the same number of sentences is to add the sentences that were filtered out by the judges to the corresponding set of clusters. There are different ways to add these sentences:

1. singletons: Each irrelevant sentence is added to set of clusters as a cluster of its own
2. bucket cluster: All irrelevant sentences are put into one cluster which is added to the set of clusters.

Adding each irrelevant sentence as a singleton seems to be the most intuitive way to handle the problem with the sentences that were filtered out. However this approach has some disadvantages. The judges will be rewarded disproportionately high for any singleton they agreement on. Thereby the disagreement on the more important clustering will be less punished. With every singleton the judges agree on the completeness and homogeneity of the whole set of clusters increases.

On the other hand the sentences in a bucket cluster are not all semantically related to each other and the cluster is not homogeneous which is contradictory to our definition of a cluster. Since the irrelevant sentences are combined to only one cluster, the judges will not be rewarded disproportionately high for their agreement. However two bucket clusters from two different sets of clusters will never be exactly the same and therefore the judges will be punished more for the disagreement on the irrelevant sentences

We have to considers these factors when we interpret the results of the inter-judge agreement.

7 Inter-Judge Agreement

We added the irrelevant sentences to each set of clusters created by the judges as described in section 6. These modified sets were then compared to each other in order to evaluate the agreement between the judges. The results are shown in table 3. For each sentence set 100 random sets of clusters were created and compared to the modified sets (in total 1300 comparisons for each method of adding irrelevant sentences). The average values of these

set	judges	singleton clusters			bucket cluster		
		V$_{beta}$	$V_{0.5}$	NVI	V_{beta}	$V_{0.5}$	NVI
Volcano	A-B	0.92	0.93	0.13	0.52	0.54	0.39
	A-D	0.92	0.93	0.13	0.44	0.49	0.4
	B-D	0.95	0.95	0.08	0.48	0.48	0.31
Rushdie	A-B	0.87	0.88	0.19	0.3	0.31	0.59
	A-H	0.86	*0.86*	*0.2*	**0.69**	**0.69**	0.32
	B-H	*0.85*	0.87	*0.2*	*0.25*	*0.27*	*0.64*
EgyptAir	A-B	0.94	0.95	0.1	0.41	0.45	0.34
	A-H	0.93	0.93	0.12	0.57	0.58	0.31
	A-O	0.94	0.94	0.11	0.44	0.46	0.36
	B-H	0.93	0.94	0.11	0.44	0.46	0.3
	B-O	0.96	0.96	0.08	0.42	0.43	0.28
	H-O	0.93	0.94	0.12	0.44	0.44	0.34
Schulz	A-B	**0.98**	**0.98**	**0.04**	0.54	0.56	**0.15**
	A-J	0.89	0.9	0.17	0.39	0.4	0.34
	B-J	0.89	0.9	0.18	0.28	0.31	0.35
base		0.66	0.75	0.44	0.29	0.28	0.68

Table 3: Inter-judge agreement for the four sentence set.

comparisons are used as a baseline.

The inter-judge agreement is most of the time higher than the baseline. Only for the Rushdie sentence set the agreement between Judge_B and Judge_H is lower for V_{beta} and $V_{0.5}$ if the bucket cluster method is used.

As explained in section 6 the two methods for adding sentences that were filtered out by the judges have a notable influence on the values of the evaluation measures. When adding singletons to the set of clusters the inter-judge agreement is considerably higher than with the bucket cluster method. For example the agreement between Judge_A and Judge_B is 0.98 for V_{beta} and $V_{0.5}$ and 0.04 for NVI when singletons are added. Here the judges filter out the same 185 sentences which is equivalent to 74.6% of all sentences in the set. In other words 185 clusters are already considered to be homogen and complete, which gives the comparison a high score. Five of the 15 clusters Judge_A created contain only sentences there were marked as irrelevant by Judge_B. In total 25 sentences are used in clusters by Judge_A which are singletons in Judge_B's set. Judge_B included nine other sentences that are singletons in the set of Judge_A. Four of the clusters are exactly the same in both sets, they contain 16 sentences. To get from Judge_A's set to the set of Judge_B 37 sentences would have to be deleted, added or moved.

With the bucket cluster method Judge_A and Judge_H for the Rushdie sentence set have the best inter-judge agreement. At the same time this combination receives the worst $V_{0.5}$ and NVI values with the singleton method. The two judges agree on 22 irrelevant sentences, which account for 21.35% of all sentences. Here the singletons have far less influence on the evaluation measures then the first example. Judge_A includes 7 sentences that are filtered out by Judge_H who uses another 11 sentences. Only one cluster is exactly the same in both sets. To get from Judge_A's set to Judge_H's cluster 11 sentences have to be deleted, 7 to be added, one cluster has to be split in two and 11 sentences have to be moved from one cluster to another.

Although the two methods of adding irrelevant sentences to the sets of cluster result in different values for the inter-judge agreement, we can conclude that the agreement between the judges is good and (almost) always exceed the baseline. Overall Judge_B seems to have the highest agreement throughout all sentence sets with all other judges.

8 Conclusion and Future Work

In this paper we presented a gold standard for sentence clustering for Multi-Document Summarization. The data set used, the guidelines and procedure given to the judges were discussed. We showed that the agreement between the judges in sentence clustering is good and exceeds the baseline. This gold standard will be used for further experiments on clustering for Multi-Document Summarization. The next step will be to compared the output of a standard clustering algorithm to the gold standard.

References

Ramiz M. Aliguliyev. 2006. A novel partitioning-based clustering method and generic document summarization. In *WI-IATW '06: Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology*, Washington, DC, USA.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence Fusion for Multidocument News Summariation. *Computational Linguistics*, 31(3):297–327.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The Second Release of the RASP System. In *COLING/ACL 2006 Interactive Presentation Sessions*, Sydney, Australien. The Association for Computer Linguistics.

Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. 2001. SIMFINDER: A Flexible Clustering Tool for Summarization. In *NAACL Workshop on Automatic Summarization*, pages 41–49. Association for Computational Linguistics.

Andreas Hess and Nicholas Kushmerick. 2003. Automatically attaching semantic metadata to web services. In *Proceedings of the 2nd International Semantic Web Conference (ISWC 2003)*, Florida, USA.

Christopher D. Manning, Prabhakar Raghavan, and Heinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Daniel Marcu and Laurie Gerber. 2001. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, Pittsburgh, PA.

Marina Meila. 2007. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Martina Naughton. 2007. Exploiting structure for event discovery using the mdi algorithm. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 31–36, Prague, Czech Republic, June. Association for Computational Linguistics.

William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical Recipies in C: The art of Scientific Programming*. Cambridge University Press, Cambridge, England.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *In ANLP/NAACL Workshop on Summarization*, pages 21–29, Morristown, NJ, USA. Association for Computational Linguistics.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *American Statistical Association Journal*, 66(336):846–850.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *Proceedings of the EACL workshop on GEometrical Models of Natural Language Semantics*.

Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 307–314, New York, NY, USA. ACM.

Hongyuan Zha. 2002. Generic Summarization and Keyphrase Extraction using Mutual Reinforcement Principle and Sentence Clustering. In *Proceedings of the 25th Annual ACM SIGIR Conference*, pages 113–120, Tampere, Finland.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota. (Technical Report #01-40).