Agreement in human factoid annotation for summarization evaluation

Simone Teufel* **and Hans van Halteren**†

*Computer Laboratory University of Cambridge, UK
Simone.Teufel@cl.cam.ac.uk

†Department of Language and Speech
University of Nijmegen, The Netherlands
hvh@let.kun.nl

Abstract

Factoid analysis was introduced by (van Halteren and Teufel, 2003) as an objective, yet semantics-oriented way of measuring overlap of information rather than surface strings in summaries. In this paper, we report on annotation experiments with two sets of summaries, and on a factoid-pairing program which finds correlations between factoids semi-automatically.

1. Introduction

Measuring the quality of summaries is extremely hard. In the past years, there has been quite a lot of summarisation work that has effectively aimed at finding viable evaluation strategies (Spärck Jones, 1999; Jing et al., 1998). Large-scale conferences like SUMMAC (Mani et al., 1999) and DUC (2002) have unfortunately shown weak results in that current evaluation measures could not distinguish between automatic summaries – though they are effective enough to distinguish them from human-written summaries. Extrinsic evaluations (measuring performance on the task for which the summary was meant in the first place) are so time-consuming to set up that they cannot be used for the day-to-day evaluation needed during system development – therefore, we concentrate here on *intrinsic* evaluation, ie. the properties of the summary itself are examined, independently of its application.

The problem with intrinsic summarisation evaluation is the absence of a gold standard: in summarisation, there is no single "best" result, but rather various "good" results. We argue that an acceptable gold-standard comparison must therefore include far more than one single human summary.

Another problem occurs with the information unit on which similarity measures are based. Previous approaches have measured similarity either as overlap in sentences (Rath et al., 1961; Jing et al., 1998; Zechner, 1996) or in words (Lin and Hovy, 2002; Saggion et al., 2002). The method of counting how many identical sentences were chosen from a text by two summarisation agents (systems or humans) has the disadvantage that the same information can be expressed in a slightly different sentence elsewhere in the text, or that a summarisation agent could produce a new sentence which does not occur anywhere in the text; even if the meaning of these sentences is the same, the evaluation measures would not be able to detect this. The method of basing the similarity on various measures of word-based similarity between two sentences, while being able to deal with this particular problem, still cannot adequately treat synonymy, polysemy, or any grammatical variations between the two sentences which are compared.

Another approach, introduced by DUC (DUC, 2002), uses information overlap judgements as the main metric, reflecting the intuition that human judgements of shared "meaning" of two texts should in principle be superior to surface-based similarity.

DUC assessors judge the informational overlap between "model units" (elementary discourse units (EDUs), i.e. clause-like units, taken from the gold standard summary) and "peer units" (sentences taken from the participating summaries) on the basis of the question: "How much of the information in a model unit is contained in a peer unit: 100%, 80%, 60%, 40%, 20%, 0%?" Weighted recall measures report how much gold standard information is present in the summaries. The downside of this approach are the fact that humans don't agree much on these non-qualitative overlap judgements (Lin and Hovy, 2002), and that non-qualitative judgements cannot provide feedback to system builders on the exact information their systems fail to include or included superflously.

Our suggestion is to base the gold standard comparison on *factoids*, a new representation of the text, which measures information rather than string similarity. Factoids are defined in data-driven manner, and our hypothesis was that they are thus more objective than wholesale DUC-style information overlap judgement; the high annotation consensus we found supports this claim.

In (van Halteren and Teufel, 2003), we present preliminary results based on this factoid–annotated data: (1) Ranking with regard to a single gold standard summary is insufficient as rankings based on any two randomly chosen summaries are very dissimilar (correlations average $\rho = 0.20$) (2) Stability of a consensus summary requires a larger number of summaries (in the range of 30-40 summaries); and (3) Similarity measurement using unigrams shows a similarly low ranking correlation when compared with factoid-based ranking. These results were based on factoid annotations by two annotators, using 50 summaries of the same newspaper text (about the murder of the Dutch politician Fortuyn), based on written guidelines which prescribe how factoid definition should be performed. We present here additional data from a new set of 20 summaries of a different text (about the invasion by Iraq of Kuwait). Both data sets were annotated by two human judges.

In this paper we address an important question regarding the stability of such resources. In how far do humans agree when defining and annotating the factoids? We also report on an algorithm which helps us in determining whether or not two potentially related factoids identified by two annotators are in fact related or not.

2. Factoid definition

Factoids correspond to atomic semantic expressions. Atomicity of a factoid is defined in a data-driven way, depending on the informational distinctions made in the set of summaries we work with. If a certain set of potential factoids always occurs together, this set of factoids is treated as one factoid, because differentiation of this set would not help us in distinguishing the summaries.

For instance, we represent the sentence *"The police have arrested a white Dutch man."* by the union of the following factoids only if for each factoid, there is at least one summary that includes this factoid and one that does not.

```
FP20 A suspect was arrested
FP21 The police did the arresting
FP24 The suspect is white
FP25 The suspect is Dutch
FP26 The suspect is male
```

The question is: how objective is such an instruction? The example above is fairly straighforward, but in the actual summaries we used (collected from students and researchers), we found various difficult cases, e.g. ambiguous expressions, slight differences in numbers and meaning, and inference. There are also issues with statements that are more general than other statements, and with attribution of statements. In order to measure whether there was shared intuition with respect to how to define and annotate factoids, we performed an annotation experiment.

3. Agreement

In our previous work, a "definitive" list of factoids was given (created by one author), and we were interested in whether agreement could be reached on the basis of this list. In the new annotation cycle reported on here, we study the process of factoid list creation, which is more time-consuming. We will discuss agreement in factoid annotation first, as it is more straightforward, even though procedurally, factoids are first *defined* (cf. section 3.2.) and then *annotated* (cf. section 3.1.).

3.1. Agreement of factoid annotation

Assuming that we have the right list of factoids already available, factoid annotation of a 100 word summary takes roughly half an hour, and measuring agreement on the decision of assigning factoids to sentences is relatively straightforward to measure. We calculate agreement in terms of Kappa, where the set of items to be classified are all factoid–summary combinations (e.g. in Figure 1 for Kuwait, N=154 factoids X 20 sentences = 2940), and where there are two categories, either 'factoid is present in summary (1)' or 'factoid is not present in summary (0)'. P(E), probability of error, is calculated on the basis of the distribution of the categories, whereas P(A), probability of agreement, is calculated as the average of observed to possible pairwise agreements per item. Kappa is defined as $k = \frac{P(A)-P(E)}{1-P(E)}$; results are given in Figure 1 for our two texts.

We measure agreement at two stages in the process: entirely independent annotation (Phase 1), and corrected annotation (Phase 2). In Phase 2, annotators see an automatically generated list of discrepancies with the other annotator, so that slips of attention can be corrected. Crucially, Phase 2 was conducted without any discussion. After Phase 2 measurement, discussion on the open points took place and a consensus was reached on which factoids should be associated with which summaries.

Figure 1 includes results for both texts as we have factoid–summary annotations by both annotators for Fortuyn (from the previous annotation round) as well as our new text. The Kappa figures indicate high agreement, even in Phase 1 (K=.87 and K=.86); in Phase 2, Kappas are as high as .89 and .95. Note that there is a difference between the annotation of the Fortuyn and the Kuwait text: in the Fortuyn case, there was no discussion or disclosure of any kind in Phase 1; one author created the factoids, and both used this list to annotate. The agreement of K=.86 was thus measured on entirely independent annotations, with no prior communication whatsoever. In the case of the Kuwait text, the prior step of finding a consensus factoid list had already taken place (as described in section 3.2.), including some discussion.

	Fortuyn text					
	K	N	k	n	P(A)	P(E)
Phase 1	.86	14178	2	2	.970	.787
Phase 2	.95	14178	2	2	.989	.779
	Kuwait text					
	K	N	k	n	P(A)	P(E)
Phase 1	.87	3060	2	2	.956	.670
Phase 2	.89	2940	2	2	.962	.663

Figure 1: Agreement of factoid annotation

3.2. Agreement of factoid definition

We realized during our previous work, where only one author created the factoids, that the task of defining factoids is a complicated process and that we should measure agreement on this task too. Therefore, we only have data on this process for the Kuwait text, and not for the Fortuyn text.

But how should the measurement of agreement on factoid creation proceed? It is difficult to find a fair measure of agreement over set operations like factoid splitting, particularly as the sets can contain a different set of summaries marked for each factoid. For instance, consider the following two sentences: (1) *M01-004 Saddam Hussein said ... that they will leave the country when the situation stabilizes.* and (2) *M06-004 Iraq claims it ... would withdraw soon.*

One annotator (A1) created a factoid "(P30) Saddam H/Iraq will leave the country soon/when situation stabilises" whereas the other annotator (A2) split this into two

	A1	A2
P30{F9.21 – a	1	1
P30{F9.21 – b	0	0
P30{F9.21 – c	1	0
P30{F9.21 – d	0	0
P30{F9.21 – e	1	0
P30{F9.22 – a	1	0
P30{F9.22 – b	0	0
P30{F9.22 – c	1	1
P30{F9.22 – d	0	0
P30{F9.22 – e	1	1

Figure 2: Items for kappa calculation for example

factoids (F9.21 and F9.22). Note that the annotators use their own, independently chosen factoid names.

Our procedure for annotation measurement is as follows. We create a list of identity and subsumption relations between factoids by the two annotators. In the example above, P30 would be listed as subsuming F9.21 and F9.22 (P30{F9.21 and P30{F9.21). It is time-consuming to create such a list; but it is necessary, as we want to measure agreement only amongst those factoids which are semantically related. We use a program which maximises shared factoids between two summary sentences to suggest such identities and subsumption relations, described in section 4.

We then calculate Kappa at Phases 1 and 2. The items are defined as follows: For each equivalence between factoids A and C, create items "A = C" X S (where S is the set of all summaries). For each factoid A subsumed by a factoid B, create items A{B X S.

In our example, given 5 summaries a, b, c, d, e, let's assume that Annotator A1 assigns P30 to summaries **a, c** and **e**, and that Annotator A2 (who has split P30 into F9.21 and F9.22), assigns **a** to F9.21 and **c** and **e** to F9.22. This annotation agrees on the 'P30{F9.21' subfactoid in cases **a** (positive), and **b** and **d** (negative), but disagrees on whether or not **e** and **c** should be assigned to the subfactoid 'P30{F9.21'. Similarly, for the 'P30{F9.22' subfactoid, there is agreement with respect to e and c (both annotators are positive) and b, d (both annotators are negative), and disagreement with respect to a. This creates the items for Kappa calculation given in Figure 2.

Results for our data set are given in Figure 3. For Phase 1 of factoid definition, a rather hard task, a Kappa of .7 indicates still relatively good agreement (but lower than for the task of factoid annotation). Many of the disagreements can be reduced to slips of attention, as the increased Kappa of .81 for Phase 2 shows.

Overall, we observe that this high agreement for both tasks points to the fact that factoid definition/annotation can be robustly performed in naturally occurring text. From our observations, it seems that factoid *annotation* is easier than factoid *definition*.

Disagreements at this stage were mostly due to different interpretation of attribution, disjunctive or collective reading of coordinated subjects, and disagreements about how much variation is allowed to still count as the 'same information'.

	Kuwait text					
	K	N	k	n	P(A)	P(E)
Phase 1	.70	3560	2	2	.91	.69
Phase 2	.81	3240	2	2	.94	.67

Figure 3: Agreement of factoid definition

4. Factoid pairing algorithm

The creation of a factoid–mapping algorithm is needed to measure agreement between two independent factoid annotations of the summaries (as described in the previous session), in order to eventually facilitate the creation of a consensus factoid list. We designed an algorithm to build an initial mapping automatically, which can then be post-processed manually.

The algorithm does not access the description of each factoid, but rather examines the lists of sentences in which the annotators state they have observed the factoids. For each pair of factoids (with each factoid coming from one of the annotators), we calculate a sentence list overlap measure. For now, we use $F_{\beta=1}$, i.e. the harmonic mean of precision and recall. After all overlaps are measured, we go through the pairs in order of decreasing overlap. If the two factoids in the current pair are as yet unlinked, and the overlap exceeds a chosen threshold, we place a strong link (⇔) between them. This creates mutually linked factoids which are likely to represent the same information:

```
PAIRED [P26] [F9.9] | 1.000000
  < P26 there is a claim that elections
    will be held
  > F9.9 Free elections to be organised
```

Sentences:

- = M01-010 Iraq dissolved the parlament and will hold free elections in the future.
- = M05-005 Iraq claimed it invaded at the request of revolutionaries who staged a coup and established "the provisional government of free Kuwait", which was dissolving Parliament and would hold future free elections.
- = M07-006 The "provisional government" announced that it was dissolving Kuwait's parliament and would soon hold "free and honest elections."
- = M14-007 Baghdad announced elections would be held.
- = M18-008 Provisional government announced that would hold free and honest elections at a future date.

In this example, it turns out that P9.9 subsumes P26.

If one factoid of the current pair is already linked, we place a weak link (→) from the unlinked factoid to the linked one. This creates pointers from unlinked factoids to corresponding factoids in the other set, which are likely to be subsuming/subsumed factoids or at least part of a related conglomerate of factoids, e.g. the counterpart of the example above, F9.9 also subsumes P27:[1]

[1] Both annotators made a mistake: one failed to split the factoids even though M14-007 does not specify "free", and the other failed to include M07-006 in P27.

```
UNPAIRED < [P27] BEST > [F9.9] | 0.750000
   < P27 these elections will be free
   > F9.9 Free elections to be organised
```

Sentences:

- = M01-010 Iraq dissolved the parlament and will hold free elections in the future.
- = M05-005 Iraq claimed it invaded at the request of revolutionaries who staged a coup and established ”the provisional government of free Kuwait”, which was dissolving Parliament and would hold future free elections.
- = M18-008 Provisional government announced that would hold free and honest elections at a future date.
- > M07-006 The ”provisional government” announced that it was dissolving Kuwait’s parliament and would soon hold ”free and honest elections.”
- > M14-007 Baghdad announced elections would be held.

We examined the output of the program after it processed the Phase 1 annotations of the Kuwait text, and compared it to the final human-determined relations (see Figure 4). Recall is good, with most of the equal factoid pairs are identified with a strong link (82%) and only 4% are unidentified. Subsumptions are slightly more difficult, with 70% identified by a strong or weak link. Precision is good as well, with only 3% of the strong links and 15% of the weak links being completely useless. The links do not always point to the exact corresponding factoid (although most often), but they do help identify conglomerates of related factoids.

	A ⇔ B	A ← B	A → B	A, B unlinked
A equal to B (A=B)	56	7	2	3
A subsumes B (A{B)	14	16	-	13
A subsumed by B (A}B)	7	1	6	6
A, B refer to different but related information	8	27	15	
A, B refer to different and unrelated information	3	10	3	

Figure 4: Pairing algorithm output compared to factoid relations

5. Conclusions

To summarize, our approach to summary evaluation has two novel aspects, namely (a) content comparison between gold standard summary and system summary via *factoids*, a pseudo-semantic representation based on atomic information units which can be robustly marked in text, and (b) use of larger numbers of model summaries, in our data based on 50 and 20 individual summaries of one text. In this paper, we have presented the agreement study of factoid determination on two datasets, with results in the range of K=.70–.81 for factoid definition, and in the range of K.86–.95 for factoid annotation.

The pairing algorithm as it stands performs adequately, but improvements are still possible. An obvious one is to include some form of description overlap measure.

Also, the reported results correspond to the inituitively set threshold of $F_{\beta=1} \geq 0.8$. A larger set of annotated texts will allow us to determine better settings empirically.

6. References

DUC, 2002. *Document Understanding Conference (DUC)*. Electronic proceedings, `http://www-nlpir.nist.gov/projects/duc/pubs.html`.

Jing, Hongyan, Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad, 1998. Summarization evaluation methods: Experiments and analysis. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*.

Lin, Chin-Yew and Eduard Hovy, 2002. Manual and automatic evaluation of summaries. In (DUC, 2002).

Mani, Inderjeet, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman, 1999. The TIPSTER Summac Text Summarization Evaluation. In *Proceedings of EACL-99*.

Rath, G.J, A. Resnick, and T. R. Savage, 1961. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143.

Saggion, Horacio, Dragomir Radev, Simone Teufel, Wai Lam, and Stephanie M. Strassel, 2002. Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment. In *Proceedings of LREC 2002*.

Spärck Jones, Karen, 1999. Automatic summarising: Factors and directions. In Inderjeet Mani and Mark T. Maybury (eds.), *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, pages 1–12.

van Halteren, Hans and Simone Teufel, 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. In *Proceedings of the HLT workshop on Automatic Summarization*.

Zechner, Klaus, 1996. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of COLING-96*.