Examining the consensus between human summaries: initial experiments with factoid analysis

Hans van Halteren
Department of Language and Speech
University of Nijmegen, The Netherlands

Simone Teufel
Computer Laboratory
Cambridge University, UK

Abstract

We present a new approach to summary evaluation which combines 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 a gold standard consensus summary, in our case based on 50 individual summaries of one text. Even though future work on more than one source text is imperative, our experiments indicate that (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) a stable consensus summary can only be expected if a larger number of summaries are collected (in the range of at least 30-40 summaries), and (3) similarity measurement using unigrams shows a similarly low ranking correlation when compared with factoid-based ranking.

1 Introduction

It is an understatement to say that measuring the quality of summaries is hard. In fact, there is unanimous consensus in the summarisation community that evaluation of summaries is a monstrously difficult task. 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; Donaway et al., 2000). 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.

In principle, the best way to evaluate a summary is to try to perform the task for which the summary was meant in the first place, and measure the quality of the summary on the basis of degree of success in executing the task. However, such *extrinsic* evaluations are so time-consuming to set up that they cannot be used for the day-to-day evaluation needed during system development. So in practice, a method for *intrinsic* evaluation is needed, where the properties of the summary itself are examined, independent of its application.

We think one of the reasons for the difficulty of an intrinsic evaluation is that summarisation has to call upon at least two hard subtasks: selection of information and production of new text. Both tasks are known from various NLP fields (e.g. information retrieval and information extraction for selection; generation and machine translation (MT) for production) to be not only hard to execute, but also hard to evaluate. This is caused for a large part by the fact that in both cases there is no single "best" result, but rather various "good" results. It is hence no wonder that the evaluation of summarisation, combining these two, is even harder. The general approach for intrinsic evaluations, then (Mani, 2001), is to separate the evaluation of the form of the text (quality) and its information content (informativeness).

In this paper, we will focus on the latter, the intrinsic evaluation of informativeness, and we will address two aspects: the (in)sufficiency of the single human summary to measure against, and the information unit on which similarity measures are based.

1.1 Gold standards

In various NLP fields, such as POS tagging, systems are tested by way of comparison against a "gold standard", a manually produced result which is supposed to be the "correct", "true" or "best" result. This presupposes, however, that there is a single "best" result. In summarisation there appears to be no "one truth", as is evidenced by a low agreement between humans in producing gold standard summaries by sentence selection (Rath et al., 1961; Jing et al., 1998; Zechner, 1996), and low overlap measures between humans when gold standards summaries are created by reformulation in the summarisers' own words (e.g. the average overlap for the 542 single document summary pairs in DUC-02 was only about 47%).

But even though the non-existence of any one gold standard is generally acknowledged in the summarisation community, actual practice nevertheless ignores this. Comparisons against a single gold stan-

dard are widely used, due to the expense of compiling summary gold standards and the lack of composite measures for comparison to more than one gold standard.

In a related field, information retrieval (IR), the problem of subjectivity of relevance judgements is circumvented by extensive sampling: many different queries are collected to level out the difference humans have in suggesting queries and in selecting relevant documents. While relevance judgements between humans remain different, Voorhees (2000) shows that the relative rankings of systems are nevertheless stable across annotators, which means that meaningful IR measures have been found despite the inherent subjectivity of relevance judgements.

Similarly, in MT, the recent Bleu measure also uses the idea that one gold standard is not enough. In an experiment, Papineni et al. (2001) based an evaluation on a collection of four reference translations of 40 general news stories and showed the evaluation to be comparable to human judgement.

Lin and Hovy (2002) examine the use of a multiple gold standard for summarisation evaluation, and conclude "we need more than one model summary although we cannot estimate how many model summaries are required to achieve reliable automated summary evaluation". We explore the differences and similarities between various human summaries in order to create a basis for such an estimate, and as a side-effect, also re-examine the degree of difference between the use of a single summary gold standard and the use of a compound gold standard.

1.2 Similarity measures

The second aspect we examine is the similarity measure to be used for gold standard comparison. In principle, the comparison can be done via co-selection of extracted sentences (Rath et al., 1961; Jing et al., 1998; Zechner, 1996), by string-based surface measures (Lin and Hovy, 2002; Saggion et al., 2002), or by subjective judgements of the amount of information overlap (DUC, 2002). The rationale for using information overlap judgement as the main evaluation metric for DUC is the wish to measure the meaning of sentences rather than use surface-based similarity such as co-selection (which does not even take identical information expressed in different sentences into account) and string-based measures. In the DUC competitions, 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: all of it, most, some, any, or none." This overlap judgement is done for each system-produced summary, and weighted recall measures report how much gold standard information is present in the summaries.

However, Lin and Hovy (2002) report low agreement for two tasks: producing the human summaries (around 40%), and assigning information overlap between them. In those cases where annotators had to judge a pair consisting of a gold standard sentence and a system sentence more than once (because different systems returned the same sentence), they agreed with their own prior judgement in only 82% of the cases. This relatively low intra-annotator agreement points to the fact that the overlap judgement remains a subjective task where judges will disagree. Lin and Hovy show the instability of the evaluation, expressed in system rankings.

We propose a gold standard comparison based on factoids, a pseudo-semantic representation of the text, which measures information rather than string similarity, like DUC, but which is more objective than DUC-style information overlap judgement.

2 Data and factoid annotation

Our goal is to compare the information content of different summaries of the same text. In this initial investigation we decided to focus on a single text. The text used for the experiment is a BBC report on the killing of the Dutch politician Pim Fortuyn. It is about 600 words long, and contains a mix of factual information and personal reactions. Our guidelines asked the human subjects to write generic summaries of roughly 100 words. We asked them to formulate the summary in their own words, so that we can also see which different textual forms are produced for the same information.

Knowledge about the variability of expression is important both for evaluation and system building, and particularly so in in multi-document summarisation, where redundant information is likely to occur in different textual forms.

We used two types of human summarisers. The largest group consisted of Dutch students of English and of Business Communications (with English as a chosen second language). Of the 60 summaries we received, we had to remove 20. Summaries were removed if it was obvious from the summary that the student had insufficient skill in English or if the word count was too high (above 130 words). A second group consisted of 10 researchers, who are either native or near-native English speakers. With this group there were no problems with language, format or length, and we could use all 10 summaries. Our total number of summaries was thus 50.

2.1 The factoid as atomic information units

We use atomic semantic units called *factoids* to represent the meaning of a sentence. For instance, we

represent the sentence

The police have arrested a white Dutch man.

by the union of the following factoids:

```
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
```

Note that in this case, factoids correspond to expressions in a FOPL-style semantics, which are compositionally interpreted. However, we define atomicity as a concept which depends on 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. If we had found, e.g., that there is no summary that mentions only one of FP25 and FP26, those factoids would be combined into one new factoid "FP27 The suspect is a Dutch man".

Our definition of atomicity means that the "amount" of information associated with one factoid can vary from a single word to an entire sentence. An example for a large chunk of information that occurred atomically in our texts was the fact that the victim wanted to become PM (FV71), a factoid which covers an entire sentence. On the other hand, a single word may contain several factoids. The word "gunman" leads to two factoids: "FP24 The perpetrator is male" and "FA20 A gun was used in the attack".

The advantage of our functional, summary-set-dependent definition of atomicity is that the definition of what counts as a factoid is more objective than if factoids had to be invented by intuition, which is hard. One possible disadvantage of our definition of atomicity (which is dependent on a given set of summaries) is that the set of factoids used may have to be adjusted if further summaries are added to the collection. In practice, for a fixed set of summaries for experiments, this is less of an issue.

We decompose meanings into separate (compositionally interpreted) factoids, if there are mentions in our texts which imply information overlap. If one summary contains "was murdered" and another "was shot dead", we can identify the factoids

```
FA10 There was an attack
FA40 The victim died
FA20 A gun was used
```

The first summary contains only the first two factoids, whereas the second contains all three. That way, the semantic similarity between related words can be expressed.

2.2 Compositionality, generalisation and factuality

The guidelines for manual annotation of summaries with factoids stated that only factoids which are explicitly expressed in the text should be marked. When we identified factoids in our actual summary collection, most factoids turned out to be independent of each other, i.e. the union of the factoids can be compositionally interpreted. However, there are relations between factoids which are not as straightforward. For instance, in the case of "FA21 Multiple shots were fired" and "FA22 Six shots were fired", FA22 implies FA21; any attempt to express the relationship between the factoids in a compositional way would result in awkward factoids. We accept that there are factoids which are most naturally expressed as generalisations of other factoids, and record for each factoid a list of factoids that are more general than it is, so that we can include these related factoids as well. In one view of our data, if a summary states FA22, FA21 is automatically added.

In addition to generality, there are two further complicated phenomena we had to deal with. The first one is real inference, rather than generalisation, as in the following cases:

```
FL52   The scene of the murder had tight security
       checks
FL51   The scene of the murder was difficult to
       get into
FL50   It is unclear how the perpetrator got to
       the victim
```

FL52 implies (in the sense of real inference) FL51, which in turn implies FL50. We again record inference relations and automatically compute the transitive closure of all inferences, but we do not currently formally distinguish them from the simpler generalisation relations.

The second phenomenon is the description of people's opinions. In our source document, quotations of the reactions of several politicians were given. In the summaries, our subjects often generalised these reactions and produced statements such as

Dutch as well as international politicians have expressed their grief and disbelief.

As more than one entity can be reported as saying the same thing, straightforward factoid union is not powerful enough to accurately represent the attribution of opinions, as our notation does not contain variables for discourse referents and quoted statements. We therefore revert to a separate set of factoids, which are multiplied-out factoids that combine the statement (what is being said) together with a description of who said it. Elements of the description can be interpreted in a compositional manner.

For instance, the above sentence is expressed in

our notation as

```
OG10   Grief was expressed
OG60   Dutch persons or organizations expressed
       grief
OG62   International persons or organizations
       expressed grief
OG40   Politicians expressed grief
OS10   Disbelief was expressed
OS60   Dutch persons or organizations expressed
       disbelief
OS62   International persons or organizations
       expressed disbelief
OS40   Politicians expressed disbelief
```

Another problem with attribution of opinions is that there is not always a clear distinction between fact and opinion. For instance, the following sentence is presented as opinion in the original *"Geraldine Coughlan in the Hague says it would have been difficult to gain access to the media park."* Nevertheless, our summarisers often decided to represent such opinions as facts, ie. as *"The media park was difficult to gain entry to."* – in fact, in our data, every summary containing this factoid presents it as fact. For now, we have taken the pragmatic approach that the classification of factoids into factual and opinion factoids is determined by the actual representation of the information in the summaries (cf. FL51 above, where the first letter "F" stands for factual, the first letter "O" for opinion).

The factoid approach can capture much finer shades of meaning differentiations than DUC-style information overlap does – in an example from Lin and Hovy (2002), an assessor judged *some* content overlap between "*Thousands of people are feared dead"* and "*3,000 and perhaps ... 5,000 people have been killed.*" In our factoid representation, a distinction between "killed" and "feared dead" would be made, and different numbers of people mentioned would have been differentiated.

2.3 Factoid annotation

The authors have independently marked the presence of factoids in all summaries in the collection. Factoid annotation of a 100 word summary takes roughly half an hour. Even with only short guidelines, the agreement on which factoids are present in a summary appears to be high. The recall of an individual annotator with regard to the consensus annotation is about 96%, and precision about 97%. This means that we can work with the current factoid presence table with reasonable confidence.

Whereas single summaries contain between 32 and 55 factoids, the collection as a whole contains 256 different factoids. Figure 1 shows the growth of the number of factoids with the size of the collection (1 to 40 summaries). We assume that the curve is Zipfian. This observation implies that larger numbers of summaries are necessary if we are looking for a definitive factoid list of a document.

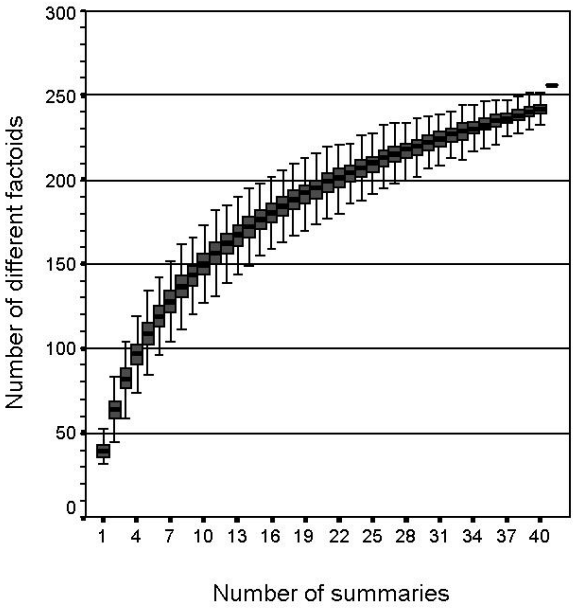

Figure 1: Average number of factoids in collections of size 1–40

The maximum number of possible factoids is not bounded by the number of factoids occurring in the document itself. As we explained above, factoids come into existence because they are observed in the collection of summaries, and summaries sometimes contain factoids which are not actually present in the document. Examples of such factoids are "FP31 The suspect has made no statement", which is true but not stated in the source text, and "FP23 The suspect was arrested on the scene", which is not even true. The reasons for such "creative" factoids vary from the expression of the summarisers' personal knowledge or opinion to misinterpretation of the source text. In total we find 87 such factoids, 51 factual ones and 36 incorrect generalisations of attribution.

Of the remaining 169 "correct" factoids, most (125) are factual. Within these factoids, we find 74 generalisation links. The rest of the factoids concern opinions and their attribution. There are 18 descriptions of opinion, with 11 generalisation links, and 26 descriptions of attribution, with 16 generalisation links. For all types, we see that most facts are being represented at differing levels of generalisation. Some of the generalisation links are part of 3- or 4-link hierarchies, e.g. "FV40 Victim outspoken about/campaigning on immigration issues" (26 mentions) to "FV41 Victim was anti- immigration" (23) to "FV42 Victim wanted to close borders to immigration" (9), or "FV50 Victim outspoken about race/religion issues" (17 mentions) to "FV51 Victim outspoken about Islam/Muslims" (16) to "FV52 Victim made negative remarks about Islam" (14) to "FV53 Victim called Islam a backward religion" (9).

It is not surprising that more specific factoids are less frequent than their generalisations, but we expect interesting correlations between a factoid's importance and the degree and shape of the decline of its generalisation hierarchy, especially where factoids about the attribution of opinion are concerned. This is an issue for further research.

3 Human summaries as benchmark for evaluation

If we plan to use human summaries as a reference point for the evaluation of machine-made summaries, we are assuming that there is some consensus between the human summarisers as to which information is important enough to include in a summary. Whether such consensus actually exists is uncertain. In very broad terms, we can distinguish four possible scenarios:

1. There is a good consensus between all human summarisers. A large percentage of the factoids present in the summaries is in fact present in a large percentage of the summaries. We can determine whether this is so by measuring factoid overlap.
2. There is no such overall consensus between all summarisers, but there are subsets of summarisers between whom consensus exists. Each of these subsets has summarised from a particular point of view, even though a generic summary was requested, and the point of view has led to group consensus. We can determine whether this is so by doing a cluster analysis on the factoid presence vectors. We should find clusters if and only if group consensus exists.
3. There is no such thing as overall consensus, but there is a difference in perceived importance between the various factoids. We can determine whether this is the case by examining how often each factoid is used in the summaries. Factoids that are more important ought to be included more often. In that case, it is still possible to create a consensus-like reference summary for any desired summary size.
4. There is no difference in perceived importance of the various factoids at all. Inclusion of factoids in summaries appears to be random.

3.1 Factoid frequency and consensus

We will start by examining whether an importance hierarchy exists, as this can help us decide between scenario 1, 3 or 4. If still necessary, we can check for group consensus later.

If we count how often each factoid is used, it quickly becomes clear that we do not have to worry about worst-case scenario 4. There are clear differences in the frequency of use of the factoids. On the other hand, scenario 1 does not appear to be very likely either. There is full consensus on the inclusion of only a meager 3 factoids, which can be summarised in 3 words:

Fortuyn was murdered.

If we accept some disagreement, and take the factoids which occur in at least 90% of the summaries, this increases the consensus summary to 5 factoids and 6 words:

Fortuyn, a politician, was shot dead.

Setting our aims ever lower, 75% of the summaries include 6 further factoids and the summary goes up to 20 words:

Pim Fortuyn, a Dutch right-wing politician, was shot dead before the election. A suspect was arrested. Fortuyn had received threats.

A 50% threshold yields 8 more factoids and the 47-word summary:

Pim Fortuyn, a Dutch right-wing politician, was shot dead at a radio station in Hilversum. Fortuyn was campaigning on immigration issues and was expected to do well in the election. He had received threats. There were shocked reactions. Political campaigning was halted. The police arrested a man.

If we want to arrive at a 100-word summary (actually 104), we need to include 26 more factoids, and we need to allow all factoids which occur in at least 30% of the summaries:

Pim Fortuyn was shot six times and died shortly afterwards. He was attacked when leaving a radio station in the (well-secured) media park in Hilversum. The Dutch far-right politician was campaigning on an anti- immigration ticket and was outspoken about Islam. He was expected to do well in the upcoming election, getting at least 15% of the votes. Fortuyn had received threats. He expected an attack and used bodyguards. Dutch and international politicians were shocked and condemned the attack. The Dutch government called a halt to political campaigning. The gunman was chased. The police later arrested a white Dutch man. The motive is unknown.

We conclude that the extreme scenarios, full consensus and full absence of consensus, can be rejected for this text. This leaves the question whether the partial consensus takes the form of clusters of consenting summarisers.

3.2 Summariser clusters

In order to determine whether the summarisers can be assigned to groups within which a large amount of consensus can be found, we turn to statistical techniques. We first form 256-dimensional binary vectors recording the presence of each of the factoids in each

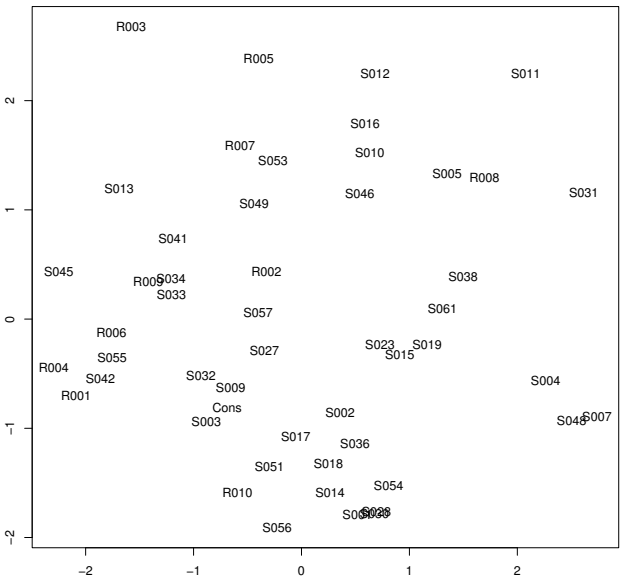

Figure 2: Classical multi-dimensional scaling of distances between factoid vectors into two dimensions

summariser's summary. We also added a vector for the 104-word consensus summary above ("Cons").

We then calculate the distances between the various vectors and use these as input for classical multi-dimensional scaling. The result of scaling into two dimensions is shown in Figure 2.

Only a few small clusters appear to emerge. Although we certainly cannot conclude that there are no clusters, we would have expected more clearly delimited groups of summarisers, i.e. different points of view, if scenario 2 described the actual situation. For now we will assume that, for this document, scenario 3 is the most likely.

3.3 The consensus summary as an evaluation tool

Two of the main demands on a gold standard generic summary for evaluation are: a) that it contains the information deemed most important in the document and b) that two gold standard summaries constructed along the same lines lead to the same, or at least very similar, ranking of a set of summaries which are evaluated.

If we decide to use a single human summary as a gold standard, we in fact assume that this human's choice of important material is acceptable for all other summary users, which it the wrong assumption, as the lack of consensus between the various human summaries shows. We propose that the use of a reference summary which is based on the factoid importance hierarchy described above, as it uses a less subjective indication of the relative importance of the information units in the text across a population of summary writers. The reference summary would then take the form of a consensus summary, in our case the 100-word compound summary on the basis of factoids over the 30% threshold.

The construction of the consensus summary would indicate that demand a) will be catered for, but we still have to check demand b). We can do this by computing rankings based on the F-measure for included factoids, and measuring the correlation coefficient ρ between them.

As we do not have a large number of automatic summaries of our text available, we use our 50 human summaries as data, pretending that they are summaries we wish to rank (evaluate).

If we compare the rankings on the basis of single human summaries as gold standard, it turns out that the ranking correlation ρ between two "gold" standards is indeed very low at an average of 0.20 (variation between -0.51 and 0.85). For the consensus summary, we can compare rankings for various numbers of base summaries. After all, the consensus summary should improve with the number of contributing base summaries and ought to approach an ideal consensus summary, which would be demonstrated by a stabilizing derived ranking.

We investigate if this assumption is correct by creating pairs of samples of N=5 to 200 base summaries, drawn (in a way similar to bootstrapping) from our original sample of 50. For each pair of samples, we automatically create a pair of consensus summaries and then determine how well these two agree in their ranking. Figure 3 shows how ρ increases with N (based on 1000 trials per N). At N=5 and 10, ρ has a still clearly unacceptable average 0.40 or 0.53. The average reaches 0.80 at 45, 0.90 at 95 and 0.95 at a staggering 180 base summaries.

We must note, however, that we have to be careful with these measurements, since 40 of our 50 starting summaries were made by less experienced non-natives. In fact, if we bootstrap pairs of N=10 base summary samples (100 trials) on just the 10 higher-quality summaries (created by natives and near-natives), we get an average ρ of 0.74. The same experiment on 10 different summaries from the other 40 (100 trials for choosing the 10, and for each 100 trials to estimate average ρ) yields average ρ's ranging from 0.55 to 0.63. So clearly the difference in experience has its effect. Even so, even the 'better' summaries lead to a ranking correlation of ρ=0.74 at N=10, which still is much lower than we would like to see. We estimate that with this type of summaries an acceptably stable ranking (ρ around 0.90) would be reached somewhere between 30 and 40 summaries.

3.4 Using unigrams instead of factoids

Apart from the need for human summaries, the factoid-based comparisons have another problem,

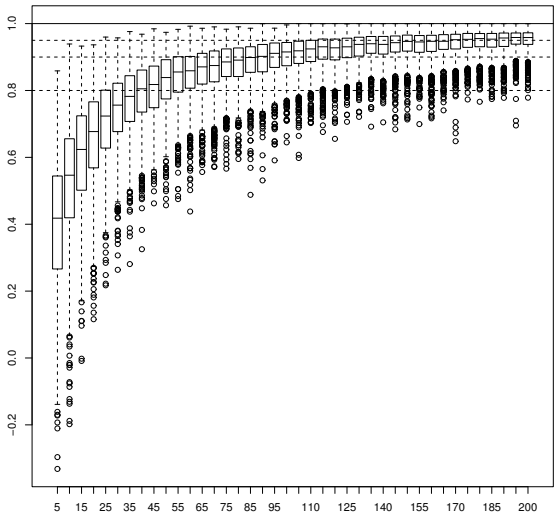

Figure 3: Correlation coefficient ρ between rankings for 50 summaries on the basis of two consensus summaries, each based on a size N base summary collection, for N between 5 and 200

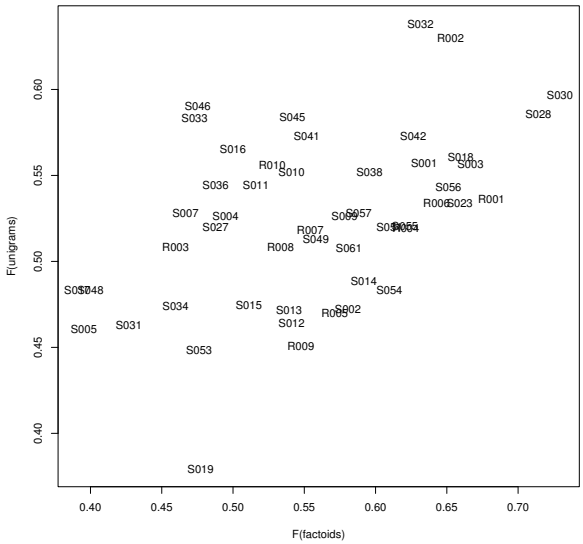

Figure 4: F-measures of summarisers with regard to consensus data: factoid-based versus unigram-based

viz. the need for human interpretation when mapping summaries to factoid lists. The question is whether simpler measures might not be equally informative. We investigate this using unigram overlap, following Papineni et al. (2001) in their suggestion that unigrams best represent contents, while longer n-grams best represent fluency.

Again, we reuse our 50 summaries as summaries to be evaluated. For each of these summaries, we calculate the F-measure for the included factoids with regard to the consensus summary shown above. In a similar fashion, we build a consensus unigram list, containing the 103 unigrams that occur in at least 11 summaries, and calculate the F-measure for unigrams. The two measures are plotted against each other in Figure 4.

Some correlation is present (r = 0.48 and Spearman's ranking correlation $\rho = 0.45$), but there are clearly profound differences. If we look at the rankings produced from these two F-measures, S054, on position 16 on the basis of factoids, drops to position 37 on the basis of unigrams. S046, on the other hand, climbs from 42nd to 4th place when considered by unigrams instead of factoids. Apart from these extreme cases, these are also clear differences in the top-5 for the two measurements: S030, S028, R001, S003 and S023 are the top-5 when measuring with factoids, whereas S032, R002, S030, S046 and S028 are the top-5 when measuring with unigrams. It would seem that unigrams, though they are much cheaper, are not a viable substitute for factoids.

4 Discussion and future work

From our experiences so far, it seems that both our innovations, viz. using multiple summaries and measuring with factoids, appear to be worth pursuing further. We summarise the results for our test text in the following:

- We observe a very wide selection of factoids in the summaries, only few of which are included by all summarisers.
- The number of factoids found if new summaries are considered does not tail off.
- There is a clear importance hierarchy of factoids which allows us to compile a consensus summary.
- If single summaries are used as gold standard, the correlation between rankings based on two such gold standard summaries is low.
- We could not find any large clusters of highly correlated summarisers in our data.
- Stability with respect to the consensus summary can only be expected if a larger number of summaries are collected (in the range of at least 30-40 summaries).
- A unigram-based measurement shows only low correlation with the factoid-based measurement.

The information that is gained through multiple summaries with factoid-similarity is insufficiently approximated with the currently used substitutes, as the observations above show. However, what we

have described here must clearly be seen as an initial experiment, and there is yet much to be done.

First of all, the notation of the factoid (currently flat atoms) needs to be made more expressive, e.g. by the addition of variables for discourse referents and events, which will make factoids more similar to FOPL expressions, and/or by the use of a typing mechanism to indicate the various forms of inference/implication.

We also need to identify a good weighting scheme to be used in measuring similarity of factoid vectors. The weighting should correct for the variation between factoids in information content, for their different position along an inference chain, and possibly for their position in the summary. It should also be able to express some notion of importance of the factoids, e.g. as measured by the number of summaries containing the factoid.

Something else to investigate is the presence and distribution of factoids, types of factoids and relations between factoids in summaries and summary collections. We have the strong feeling that some of our observations were tightly linked to the type of text we used. We would like to build a balanced corpus of texts, of various subject areas and lengths, and their summaries, at several different lengths and possibly even multi-document, so that we can study this factor. An open question is how many summaries we should try to get for each of the texts in the corpus. It is unlikely we will be able to collect 50 summaries for each new text. Furthermore, the texts of the corpus should also be summarised by as many machine summarisers as possible, so that we can test ranking these on the basis of factoids, in a realistic framework.

A final line of investigation is searching for ways to reduce the cost of factoid analysis. The first reason why this analysis is currently expensive is the need for large summary bases for consensus summaries. There is yet hope that this can be circumvented by using larger numbers of texts, as is the case in IR and in MT, where discrepancies prove to average out when large enough datasets are used. Papineni et al., e.g., were able to show that the ranking with their Bleu measure of the five evaluated translators (two human and three machine) remained stable if only a single reference translation was used, suggesting that "we may use a big corpus with a single reference translation, provided that the translations are not all from the same translator". Possibly a similar averaging effect will occur in the evaluation of summarisation so that smaller summary bases can be used. The second reason is the need for human annotation of factoids. Although simple unigram-based methods prove insufficient, we will hopefully be able to come a long way in automating factoid identification on the basis of existing NLP techniques, combined with information gained about factoids in research as described in the previous paragraph. All in all, the use of consensus summaries and factoid analysis, even though expensive to set up for the moment, provides a promising alternative which could well bring us closer to a solution to several problems in summarisation evaluation.

References

Donaway, Robert L., Kevin W. Drummey, and Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the ANLP/NAACL 2000 Workshop on Automatic Summarization*.

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 Dragomir R. Radev and Eduard H. Hovy, eds., *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, 60–68.

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

Mani, Inderjeet. 2001. *Automatic Summarization*. John Benjamins.

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*, 77–85.

Papineni, K, S. Roukos, T Ward, and W-J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL-02*, 311–318.

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*, 747–754.

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

Voorhees, Ellen. 2000. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management* 36: 697–716.

Zechner, Klaus. 1996. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences. In *Proceedings of COLING-96*, 986–989.