IIRG-UCD at DUC 2006

William Doran, John Dunnion and Joe Carthy.
Intelligent Information Retrieval Group,
School of Computer Science and Informatics,
College of Mathematical and Physical Sciences,
UCD Dublin, Ireland

{William.Doran, John.Dunnion, Joe.Carthy}@ucd.ie,

Abstract

This paper describes a user-centric multi-document summarisation system. This system creates 250 word summaries of a collection of documents that satisfy a set of questions. Solving this involves a linguistic approach to question reformulation and analysis. These questions are passed to a question answering system that uses word occurrence statistics, semantic entity extraction and grammatical similarity to extract answers. We also present the results of the official DUC evaluation of our system.

1 Introduction

The task of this year's Document Understanding Conference (DUC) is to provide a 250-word summary of a collection of related documents that answers a set of supplied questions. This task has evolved over the past number of years so that now questions are used to focus the summary instead of returning a generic summary. It stems from a merging of several research areas such as summarisation, question answering and document comprehension.

The task also addresses changing information needs of the user; where the difference between searching for specific information and browsing large information sources has become less apparent. It no longer suffices for a user to read a document to find the answer to their information need. We now want to be able to find specific details about our desired need taken from a wide range of documents. This year's task is a step in that direction.

Question Answering (QA) researchers and summarisation researchers have struggled for years to find/construct representations of information that express the meaning of a document or the answer to a question. Both are similar and related problems that spawn the basis of this task. The first key element of this task is to find the information that answers the presented question, the second task is to take this information and present it in the correct context in a coherent and meaningful passage.

Question answering traditionally involves the difficult task of returning fact-based answers to single questions. To further compound this, this year's task involves a set of questions that not only look for specific facts, but also facts related to the answers, and even facts that relate to the answers of previous questions. These facts must then be presented in a coherent, structured and concise summary. This adds the problems of anaphora resolution, inference and elaboration of facts, among others.

Our research has focussed on the decomposition of these sets of questions into independent questions that can then be used as input to a simple QA system. We carry out syntactic analysis of the questions to find out what information is required to answer the question. We carry out similar syntactic analysis on the sentences of the documents along with some additional statistical investigation and semantic entity extraction. This forms an intermediate representation of the questions and document sentences. The QA system then uses very simple matching strategies to find the most suitable candidates for answer selection. A ranked list of sentences is returned based on the combination of the scores assigned to the sentence based on different matching functions.

The next section of this paper provides more details on our system implementation, section 3 explains the evaluation of the summaries and presents the official evaluated results of our system. In section 4, we discuss our results and finally we make our conclusions and propose further work to be undertaken.

2 System Overview

In this section we present the approach we used to tackle the user focussed multi-document summarisation task in this years DUC. To fulfil this task one must provide a summary that answers a set of given questions to a required level of granularity. These questions are in a narrative form and are often inter-related and interdependent on each other. To address this problem we chose a modular approach that would break this problem into smaller manageable chunks. We firstly describe the pre-processing that was required, then we will discuss the procedure of breaking the narrative into manageable questions and finally describe the answer selection mechanism.

2.1 Pre-processing:

There are two main types of document involved in this years task; there are the SGML formatted news articles and a question file that includes the narrative-style questions for each topic. Each of these files goes through the pre-processing stage with the question file undergoing a further pre-processing stage laid out in the next section.

Several steps are required to convert the documents from their raw SGML format to the format to be used in our system. The first step is to remove the tags and meta-information and to convert the text into a single-sentence-per-line format. This standard format is necessary for the modular design of the system ensuring that all modules receive the documents in the same manner. We input the documents to a part of speech (POS) tagger which removes the SGML tags and gives part of speech information to the words in the document. Following this we use the POS information to place one sentence per line on the document.

The next step is to resolve any anaphoric references that occur through the document. Although anaphoric reference is in itself a difficult task, we felt that is was a useful addition to the system, given that one of the evaluation criteria is referential clarity. We resolve the references using the GUITAR[9] system, this is a probabilistic based system that reports 67% accuracy, and it performs reasonably well for this task. Using some additional scripting, we replace all referents with their antecedents.

In addition to this we carry out semantic entity extraction on the document sentences using the ANNIE system taken from the GATE framework [13]. This system extracts semantic entities such as Person, Place, Job title, and Organisation. To bolster our anaphoric resolution efforts we carry out partial resolution on the entities discovered by ANNIE so that all entities mentioned in the document are replaced by the longest occurrence of that entity. i.e. *Clinton* is replaced by *President Bill Clinton.*

Further to all this we carry out syntactic analysis of the sentences in the documents by using the RASP system [2]. This system parses the sentence and give us a syntactic representation of the sentence which is useful in the reformulation of questions and in the selection of candidate answers. A full explanation of this process is given below.

2.2 Narrative Decomposition:

The narrative style of the questions in this task make it very difficult to apply normal question answering heuristics without some modification. We propose that breaking the narrative questions into separate questions would simplify the task and make it easier to deal with in a question answering context.

Each narrative contains several questions; some of these questions look for direct answers (e.g. who, where, when etc.) while others require more detail and inference (eg. Identify, explain etc.). There are also grammatical issues that arise with conjunctions, comparisons, clauses and elaborating sentences. Other issues that can arise are that of anaphora, and references in questions to the answers of previous questions. All these issues must be dealt with in order to successfully translate the narrative into a set of useful questions.

A further description of the different types of questions is given below.

- "wh-"-These are the standard QA type questions of who, what, where etc.. E.g. "In what countries are MAGLEV rail systems being proposed?"
- Imperative- A question in the imperative that usually asks for particular details or specific information about something. E.g. "Explain the industrial espionage case involving VW and GM."
- Elaboration – These questions have a lead-in sentence or a subsequent elaborating sentence to give more information about the question. E.g. " Nobel prizes are awarded each year for achievements in the sciences (physics, chemistry, physiology and medicine) and economics. Who are the Nobel prize winners

in the sciences and in economics and what are their prize-winning achievements?"

- Boolean- These questions look for a yes/no answer and can also ask to choose between several cases. E.g. "Are journalists specifically targeted?"
- Conjunctions- These questions contain conjunctions of items that need to be split up into separate questions. E.g. "Have diplomatic, economic, and military relations been restored?"
- Clausal- These questions contain several clauses that need to be refined into separated questions. E.g. "Where have poachers endangered wildlife, what wildlife has been endangered and what steps have been taken to prevent poaching?"
- Other- There are other types of question that can occur for example; including questions with prepositions and conjunction, and comparisons with previous answers. E.g. "What other factors affect the disputes?"," What rules have been imposed regarding food labelling and by whom?"

To convert the narratives into usable independent questions, we firstly describe the algorithm for splitting the sentences. The algorithm recurses several times to ensure that questions are split and reformulated correctly and also to ensure that any newly formed questions are also split and reformulated if required.

We begin with a queue of questions in single-line format with any anaphora resolved. The first step is to look for surface grammatical clues of conjunction of question clauses. These are usually two or more "wh-" questions joined with a conjunction. We split the sentence as long as we have the following regular expression",? and|or|but wh[o|ere|en]|how". The next step is to look for questions that do not contain any of the common "wh" type question words. These types of sentences are generally either an imperative style question or an elaborating sentence. If the sentence begins with an imperative key word (explain, name, describe, identify, define etc.), then it is deemed to be an imperative question. If the sentence contains no such keywords at the start then it is deemed to an elaborating sentence. At present we haven't linked these sentences to their elaborated question, but intend to use our previous work on lexical cohesion analysis[4] and anaphoric resolution to create a useful question from the elaborating sentence and question.

The next type of question to look for is the Boolean question. These questions usually have some form of the verb "to be" at the beginning of the question. E.g. is there, is, are, was etc.. The answer for these types of question is often a yes/no answer but there are also cases where you are asked to decide which case is true, e.g. is A, B or C true. We decided to reformulate these types of questions into statements. We did this by swapping the subject and the auxiliary of the verb. The premise behind this was to try to find grammatically similar sentences in the documents to the reformulated statements.

The next type of question are those that contain conjunctions. There are several different sub-class of these; the first being of the kind "Provide information on A, B and C", this splits into three separate questions as the conjunction is used in an enumerative context. To separate these questions we generally look for a proposition followed by a conjunction of items. We then match the individual conjuncts to stem of the question to form separate questions. This stem usually occurs before the conjunction but it can also happen afterwards and this must be taken into account.

The second instance of this type of question occurs in the following form, "Are the proposals for short_A or long_B haul_C", where the conjunction joins two modifiers (A and B) of the head noun (C). In this case we must split the modifiers and rejoin them separately with the head noun to form two new questions. This case is greatly simplified when the modifiers are antonyms of each other. The final case we looked at was when we have a question followed by another short question, for example "where was Kennedy killed and by whom?". In this case we need to split the sentence at the conjunction of clauses as before, but we also need to check that the resulting sentences are formed correctly. If the second sentence is too short (one or two words) E.g. "by whom", then we create a new question by swapping the subject of the verb with the subject in the second. E.g. "by whom was Kennedy killed?".

Following are some examples of the reformulation and breaking up of conjunctions of items.

Are the proposals for short or long haul? ?
The proposals are for long haul?
The proposals are for short haul?

We continue to process the questions on the queue until there are no new questions formed. This then leaves us with a list of independent questions that are converted into an intermediate representation of the information need expressed in the question. This conversion is undertaken by means of Grammatical Relation Annotation [1]. This is a syntactic parser that labels grammatical relations between syntactic constituents of a sentence. These relations allow us to extract triples of the main verbs in a sentence along with their

corresponding subjects and objects. We can then use these triples to match against a similar representation of the document sentences. Further details of this process are indicated in the following section.

2.3 Candidate Answer Representation.

The selection of a suitable candidate answer representation is crucial to the success of any question answering system. This difficulty of this stage is further compounded in this task by not only looking for direct answers to questions but also to other pieces of information that are related to the answer. It is for this reason that one strategy will not work reliably for all types of question and thus we require several, often very different, methods for providing the sought answer in correct context. In this section we will describe the strategy we implemented using Grammatical Relation Annotation and statistical analysis.

Above we explained how we spilt the narrative questions and converted them to an intermediate representation of the information need of the question. We also carry out this analysis of the sentences to convert them into subject-verb-object triples. In addition to this, for each sentence we also look at modifiers to the word triples to find additional information to add to the representation. We can also attribute semantic information to the word triples by looking at the modifiers of the triples and the semantic entities extracted by ANNIE [13]. For example, we can give some location context to those triples that are modified by entities tagged as being locations. E.g. "Castro trained forces in Cuba", where Castro is the subject of the verb trained, the forces are the object and it is modified by *in Cuba* which is a location.

The lists of semantic entities also allow us to carry out statistical analysis of the document collection. We observe that entities that occur in a large proportion of the documents are likely to be of importance to the central theme of the documents. Thus sentences that have similar subject-verb-object triples to the questions, and are modified by, or contain common occurring entities, are more likely to be candidate answers. We determine the document frequency for the following entities: people, places, job titles and organisations. We also calculate the document frequency for verbs that occur throughout the documents. We employ a stopword list of verbs to remove verbs that have low information content, e.g. *make, do, find, etc.* Verbs with high information content and high document frequency can prove valuable in uncovering word triples that are important in a document collection. i.e. *arrest, smuggle, murder etc.*

Preliminary examination of the entity lists can allow some level of abstraction between entities. For example, if a story mentions many countries, such as Panama, Mexico and Nicaragua, it is conceivable to replace this enumeration of countries in a summary with their common hypernym, *Central America*.

It is also conceivable to investigate the focus of the stories based on the distribution of the different entity class. If a story cluster is based around events that occur in a particular location then you would expect the location entities to be quite narrowly distributed. On the other hand if the cluster is based around a particular person, you would expect that person entities to be narrowly distributed around that person and the other entities to be widely distributed. It warrants further investigation to see if this is the case. One possible stumbling block to this idea is a cluster made of very loosely related documents, however using these statistics it may be possible to create sub-clusters that are aligned with location information, person information etc.

We use this statistical information along with the word triple representations of the sentences to try to match candidate sentences to the question representations explained in Section 2.2. We perform a pair-wise comparison of all questions and document sentences. We compare them by firstly looking for common verbs, subjects and objects that occur between the question and the candidate sentence. We assign a score for each match that occurs between the sentence and question representations. We then look for matches of supplementary information that we previously added to the sentence triples representation. This supplementary information represents words that modify the main word triples and often contain relationships between events (verbs) and locations and people (subjects/objects). We also award a score to words that have a higher document frequency than a pre-set threshold, allowing us to bias sentences that have entities that occur in a majority of the documents in the collection.

We use an Edmundsonian [14] approach in which use a linear combination of the weights for each of the different matching criteria to arrive at a total score for each sentence. For each question we obtain a ranked list of sentences and return sentences from these lists until the 250-word limit is reached. In the next section we will present the official results of our participation in this years DUC.

3 DUC evaluation

The evaluation procedure for this year's task covered the intrinsic or internal qualities required by a summary. The evaluation focussed on the linguistic quality and information content of the summaries. Several tools, both automatic and semi-automatic were used to evaluate the information content and linguistic quality of the submitted summaries. The linguistic quality is determined by marks given out of five for various linguistic properties; 1) grammaticality, 2) non-redundancy, 3) referential clarity, 4) focus, 5) structure and coherence. The marks are awarded by human assessors in response to the five linguistic quality questions. The average scores for the linguistic quality questions are presented in table 1.

Peer	1	2	3	4	5
25	3.2	4.1	2.8	3.4	2.2

Quality Score (1 = very poor...5 = very good)

Table 1: Average Linguistic Quality Marks.

The information content is measured using several different methods. The first being responsiveness; this is a human-assigned ranking that reflects the information content of a summary with the respect to the information need expressed in the topic. There is also a score for the overall content and readability of the summaries produced by a system. Again these scores are marked from 1 to 5 as above.

Peer	Content	Overall
25	2.34	2.06

Table 2: Average Responsiveness Scores.

The second information content measuring metric is the fully automatic ROUGE evaluation system[5]. ROUGE scores systems based on the number of N-gram matches between the summary and several model summaries. The official ROUGE scores for this year are the ROUGE-2 metric (OR2), which is a measure of bi-gram overlap, and ROUGE-SU4 (ORSU4) which allows for matches that bridge a gap of four or less words. The official scores have been averaged to offset the bias given to topics that are evaluated with a greater number of model summaries. These scores are presented in table 3. We have also tabulated the raw averaged f-scores for the remaining rouge n-gram metrics, Rouge1 (rR1), Rouge2(rR2), RougeLCS, (rRLCS), RougeW1.2 (rRW) and finally RougeSU4 (rRSU4). These ROUGE results are laid out below in table 3.

OR2	ORSU4	rR1	rR2	rRLCS	rRW	rRSU4
0.07	0.125	0.37	0.07	0.34	0.12	0.12

Table 3: Raw Macro-averaged ROUGE

The next evaluation metric uses the Basic Elements (BE) framework [12]. The BE framework is divided into three distinct parts: the splitting function, the matching function and the scoring function. The splitting function breaks up model summaries and peers into usable units for comparison. The matching function determines the amount of overlap between the model units and the peer units and the scoring function evaluates a score for a particular peer based on the amount of matched overlap. In the BE package different functions can be used for all three parts of the framework. In this years evaluation, the summaries are split using the MINPAR parser, are matched using the Head-Modifier criteria and scored using Rouge. In fact both ROUGE and the final evaluation model, the pyramid model, are both instances of the BE framework. For more details of this refer to the paper [12].The BE score for our system 25, was 0.034.

The final information content metric is the Pyramid Model Evaluation method[7]. This is the second time this method has been used on such a large scale. The Pyramid Model relies on two things; the first being the identification of Summarisation Content Units (SCUs), and secondly the organisation of these SCUs into a weighted Pyramid structure. This Pyramid structure places higher importance on SCUs that occur in many model summaries than on SCUs that occur less frequently. Using this hierarchical technique we can compare a candidate summary to the Pyramid and highlight which SCUs occur in summary and the Pyramid. The score for the summary then depends on how many of its SCUs occur in the Pyramid, what level they occur at, the amount of repetition that occurs and the number of non-pyramid SCU's that occur in the summary. The identification of SCU's and building of Pyramids is done by hand. The annotation of summaries is also done by hand and the scoring is done automatically based on the annotations. The time constraints involved in the pyramid evaluation meant that only a subset of the topics were used in the evaluation. The official Pyramid scores for this are tabulated in table 4.

Peer 25	Pyramid Score	No. SCU/Summary
Average	15.8	5.5

Table 4: Official Pyramid Evaluation scores.

4 Discussion

Above we have presented the official results from our participation in this year's DUC. Overall some results are positive and some are disappointing. We performed well on the grammaticality and non-redundancy fields, this was slightly surprising as we had not implemented any of our previous multi-doc redundancy techniques. The grammaticality was no real surprise as we used an extractive sentence selection method. We felt the grammaticality may have suffered due to anaphora resolution but this doesn't seem to be the case. In light of this the referential clarity scores are good reflecting that anaphora resolution step is a worthwhile task.

The ROUGE results are a little disappointing. We finished just outside the top two thirds of the group in both official scores: OR2 23rd and ORSU4 23rd (out of 34). It is a little difficult to decipher these scores as they allow for a gap in overlap to occur. We have vastly improved our raw information overlap (Rouge 1) so we need to investigate how we present and organise this information to improve the official metrics. In the BE category we finish in 22nd place which again is disappointing as we would have hoped that our system for picking out triples would have been similar to the BE framework. For the pyramid scores we finished in 18th out of 21. This is the most dis-heartening of the scores as we hoped that by taking sentences that were biased by document frequencies we would capture elements that occurred higher in the pyramid hierarchy and boost our score. We do take some solace from the fact that all our scores are greatly improved from last year. This provides us with motivation to carry on making improvements and hopefully learning from others.

5 Conclusion and Future work

Overall this was a very challenging task this year and called for a diverse and multi-faceted approach. The questions themselves were very complex and thus needed a complex system to perform well. We feel that this task will continue to challenge for many years to come.

The evaluation using the pyramid model was a very useful exercise and hopefully it will become easier to use and more beneficial once the community converges on a concrete set of regulations for its use. However, a variation of the system may be required to bias SCUs that occur as answers to questions. This would be more beneficial to a Question Answering task instead the generic multi-document task to which it is currently tailored.

Our results were disappointing but we feel we are on right track in trying to break the complex task into smaller manageable chunks. This is a very worthwhile task for us as it incorporates a lot of other work our group carries out. We intend to carry on with the question reformulation and will try to improve and develop upon what we have already achieved. We hope to incorporate simple discourse relations and try to relate the questions to each other and the documents in a more structured manner. We also plan to continue working on methods to find answers to the questions including textual entailment. As well as this we want to implement a plan-b option that will rely on previous work we have done in generic multi-doc summarisation. This will be used when the questions are very general and we hope it will boost our performance in subsequent years.

References:

[1] Ted Briscoe and John Carroll. 2000. Grammatical relation annotation. On-line document. http://www.cogs.susx.ac.uk/lab/nlp/carroll/grdescription/index.html.

[2] E. Briscoe and J. Carroll Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Gran Canaria. 1499-1504. 2002

[3] Document Understanding Conference, http://www-nlpir.nist.gov/projects/duc/index.html

[4] William P. Doran, Nicola Stokes, John Dunnion, Joe Carthy. *Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization.* In the Proceedings of the 5th International conference on Intelligent Text Processing and Computational Linguistics CICLing-2004, 2004.

[5] Lin C-Y, Hovy E.. *Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics.* In Proceedings of HLT-NAACL-2003. 2003

[6] D. Mollá and M. Gardiner. AnswerFinder at TREC 2004 (2005). *The Thirteenth Text REtrieval Conference (TREC 2004).2004*

[7] Ani Nenkova and Rebecca Passonneau, *Evaluating Content Selection in Summarization: the Pyramid Method.* In Procedings of Human Language Technology conference / North American chapter of the Association for Computational Linguistics. (NAACL-HLT),Boston, USA. 2004.

[8] Eamonn Newman, Nicola Stokes, Joe Carthy, John Dunnion. *UCD IIRG Approach to the Textual Entailment Challenge*. In the Proceedings of the PASCAL Recognising Textual Entailment Challenge, April 2005.

[9] Poesio, Massimo and Mijail A. Kabadjov. A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation. In *Proc. of the 4th International Conference on Language Resources and Evaluation* (LREC), Lisbon, Portugal.2004

[10] Porter, M.F., "*An algorithm for suffix stripping*", Program; automated library and information systems, 14(3), 130-137, 1980.

[11] Salton, G., Singhal, A., Mitra, M., and Buckley, C. , Automatic text structuring and summarisation. *Information Processing and Management* 33(2):193–208. 1997

[12] Hovy E., Lin C-Y, *"Zhou L, Evaluating DUC 2005 Using Basic Elements"*, In Proceeding of the Document Understanding Workshop (DUC 2005) held in conjunction with HLT/EMNLP, Vancouver, Canada. October 2005

[13] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. ***GATE**: A **framework** and graphical development environment for robust NLP tools and applications*. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002

[14] H. P. **Edmundson**, "*New methods in automatic extracting*," Journal of the ACM, vol. 16, no. 2, pp. 264--285, 1969.