Multi-ERSS and ERSS 2004

Sabine Bergler, René Witte, Zhuoyan Li, Michelle Khalifé,
Yunyu Chen, Monia Doandes, and Alina Andreevskaia
The CLaC Laboratory
Department of Computer Science
Concordia University

Abstract

Last year, we presented a system, ERSS, which constructed 10 word summaries in form of a list of noun phrases. It was based on a knowledge-poor extraction of noun phrase coreference chains implemented on a fuzzy set theoretic base. This year we present the performance of an improved version, ERSS 2004 and an extension of the same basic system: Multi-ERSS constructs 100-word extract summaries for clusters of texts. With very few modifications we ran ERSS 2004 on Tasks 1 and 3 and Multi-ERSS on Tasks 2, 4, and 5, scoring generally above average in all but the linguistic quality aspects.

1 Introduction

Fundamentally, summarization requires to rank information in a (collection of) text(s) and to produce a coherent subset of the highest ranking information depending on the expected length of the summary. Extract-based summaries most often use a keyword ranking scheme as developed for information retrieval in order to achieve this ranking, thus they look at all terms and score them according to their frequency and importance (often approximated by rarity over some corpus). Keywords are then located in the sentences and the most salient part of the sentence is included in the output. This gross oversimplification of the actual process (compare for instance [3, 8] for better characterizations) serves here to contrast with a more language inspired approach, where the linguistic structure of the text is assumed to package information in a form that can be exploited by the summarizer. Litkowski attempts full parsing and analysis of discourse markers [11]; other systems embed partial linguistic analysis into more complex systems [7, 2]. Our interest was in exploring how far the most shallow linguistic analysis, done in a knowledge-poor environment, would carry us: ERS, the base system underlying both ERSS 2004 and Multi-ERSS, attempts to extract NP coreference chains based on a few heuristics motivated in [1]. Summarization is then guided by the simple idea that the longest coreference chains represent the most important entities for the task and should therefore appear in the summary.

The success of our systems is due mostly, however, to the way the basic idea is implemented. The coreference chains are formed from NP chunks using fuzzy set theory [13]. And the implementation makes crucial use of (partly improved versions) of the ANNIE components shipped with GATE, for instance the Gazetteer for named entity recognition [4]. The summary conclusion of this year's DUC competition is that knowledge-poor coreference resolution can perform similar to both systems based on information extraction technology and systems based on more elaborate strategies in both the 10-word indexing tasks and the multi-document summarization tasks.

Last year's ERSS produced 10-word summaries of newspaper texts based on a knowledge-poor way of computing coreference chains built using fuzzy set theory. That system ranked slightly below average and was run only on one DUC task. ERSS-2004 has been based on a more rigorous use of the fuzzy set theoretic reasoning component and by more extensive use of ANNIE components available through GATE and participated in Tasks 1 and 3 and placed in the upper third (except for the MT input track of Task 3).

Multi-ERSS is the evolution of ERSS-2004 to produce extract summaries of multiple documents in a single summary of roughly 100 words. The documents were preclustered according to some topic, which was not known beforehand. Multi-ERSS participated in Tasks 2, 4, and 5. It scored average for Task 2 and 5 in coverage, but placed second for the manually computed *responsiveness* score comparing systems against each other, not against a human generated model summary.

We used the same analysis techniques and largely the same summarization strategy for all five tasks. Because our technique is based on a knowledge-poor determination of noun phrase coreference, we felt that our system should be impervious to the fact that the text was machine translated for tasks 3 and 4. This was only partially true, the fact that the translation did not preserve the noun phrases of the original text decreased readability and usefulness. We have not tested to what degree it would be possible to improve the performance and given the superior performance of Lakhas, the only system that worked from the original Arabic texts [6], don't feel it worthwhile to compensate for translation errors.

Task 5 was the most specific task, where the topic of the summary was the description of a person. We chose not to build a special system, but to only adjust the weights of our regular summarization parameters. Task 5 proved an unexpected success in responsiveness, and scored average in coverage.

2 System's Overview

The basic system's architecture underlying both ERSS 2004 and Multi-ERSS has been described in [13]. Here, we outline some changes.

2.1 Chunking

Named Entity (NE) recognition is an important part of our noun phrase recognition phase. We extended our use of components of the ANNIE system that comes as an example application with the GATE framework. These consist of grammars for dates, person names, etc. and an extensive word list, their Gazetteer. We extended and amended these tools slightly.

Our noun phrase extractor (NPE) uses a context-free NP grammar and an Earley-type chart parser to extract minimal noun phrases, i.e., NPs without any attachments. It relies heavily on the various named entities (names, dates, and so on) and only falls back to part-of-speech tags if the input tokens have not been marked by any of the NE transducer grammars. This pre-processing of NPs boosts recall and precision compared to chunking all tokens, mainly by removing ambiguities. When compared to manually annotated NPs, we can retrieve up to 99% of the marked NPs[1] when scored leniently, that is when marked NPs that overlap with retrieved NPs score as a hit.

This year, in an additional step, the minimal NPs are joined into long NPs by attaching certain grammatical features, like conjunctions, prepositions, appositions, or relative clauses. The extended NPs are used in the apposition heuristic of Task 5.

2.2 Fuzzy NP Coreference Resolution

Fuzzy-ERS groups the NPs extracted by NPE into *coreference chains*, ordered sets of NPs that refer to the same entity. Details on our fuzzy coreferencer and its algorithms can be found in [13] and [12]. Here, we only describe the core idea of the fuzzy resolution algorithm and the enhancements we added compared to last year's system.

We use fuzzy set theory to acknowledge the inherent uncertainty in coreference resolution. In this, we do not only avoid setting (or learning) thresholds based on a limited training corpus, but retain (and have used) the possibility to change our thresholds from stricter to more lenient, thus providing "soft scrolling" between recall and precision in different contexts. For the 10-word summaries we did last year, for instance, more lenient values led to better summaries: here recall outweighed precision, false positives usually did not surface, but the length of the chain being the major criterion for inclusion of the chain representative in the summary means chopping a chain into several more precise subchains deteriorates performance substantially.

The output of our coreference algorithm is a set of fuzzy coreference chains, similar to classical resolution systems. Each chain holds all noun phrases that refer to the same conceptual entity. However, unlike for classical, crisp chains, we do not have to reject inconsistent information out of hand, so we can admit a noun phrase as a member of more than one chain, with a varying degree of certainty for each.

2.3 Fuzzy Heuristics and Anti-Heuristics

The fuzzy coreference resolution is based on a number of heuristics for establishing coreference, each focusing on a particular linguistic phenomenon. Examples for fuzzy heuristics are pronominal coreference, synonym/hypernym-coreference, or substring coreference. Unlike crisp heuristics' binary decisions, fuzzy heuristics compute a degree of certainty varying between 0 (impossible) and 1 (certain) for a given noun phrase pair. Formally, a fuzzy heuristic $\mathscr{H}_i$ takes as input a noun phrase pair $(\mathrm{np}_j, \mathrm{np}_k)$ and returns a fuzzy set $\mu^{\mathscr{H}_i}_{(\mathrm{np}_j,\mathrm{np}_k)}$ that indicates the certainty of coreference for the noun phrase arguments.

Similar to the positive heuristics, *anti-heuristics* compute a degree of certainty between two NPs, but here the degree indicates how certain the two NPs do *not* corefer. The concept of anti-heuristics allows us to encapsulate exceptions to the general heuristics described above, without overloading each of them individually. One anti-heuristic already in use in ERSS was to prevent measurements to all corefer (*2 million* and *350 million*) and to discourage all cities to corefer to each other because they share a hypernym. Now we use anti-heuristics more pervasively and for smaller penalties, not only to block certain errors.

2.4 Inter-document Coreference

The changes discussed so far concern both, ERSS 2004 and Multi-ERSS. Multi-ERSS summaries are built from sentence extracts determined by the longest inter-document (that is cluster) coreference chain. Coreference strategies differ within a document and across documents. Thus, while we use the same coreference algorithm we use only a subset of the coreference heuristics for cross-document coreference resolution. For intra-document coreference, only NPs from the same document are compared, for inter-document coreference we only examine NPs from different documents (i.e., never two NPs

[1] Marked NPs are all NPs, not only named entities. Marked NPs are not limited to minimal NPs.

from the same document).

2.5 Summarization

Our summaries consist of a sequence of text extracts. A summarization framework allows the development of different summarization strategies. For each strategy, features are extracted from the document's annotations (for example, the length of a coreference chain), the features are weighted, resulting in a rank for an annotation. Based on this rank, we then extract the selected annotation(s), for example a list of NPs or sentences. A more detailed description of the different summarization strategies for the different tasks follows.

3 ERSS 2004 – Very Short Summaries of Single Documents

Very short summaries (75 characters) of single documents are required in Task 1 (summarization of single English newspaper articles) as well as Task 3 (summarization of manual and automatic translations from Arabic newspaper articles).

We participated in Task 1 for calibration purposes: how did our changes to the fuzzy set reasoner affect the 10-word summary performance and where does our system stand with respect to this year's texts and participants?

As mentioned, we expected the same system to score similarly in Task 3.

3.1 Summarization Strategy

Our very short summaries may better be called indices or keywords, since we provide a list of noun phrases only, not proper headlines. We rank all NPs of a single document by two features: (1) the length of the coreference chain they appear in (NPs appearing in longer chains receive a higher rank) and (2) whether the NP appears within the first two sentences. Both features are equally weighted. For the summary, we then extract the highest-ranking NPs until the length limit of 75 characters has been reached.

Some simple post-processing is performed on the resulting set of NPs to remove determiners and other fillers, and to remove redundant (overlapping) NPs.

Examples of Task 1 summaries for four texts from the same cluster are given in Figure 1.

3.2 Tasks 1 and 3

Tasks 1 and 3 required very short summaries of single documents.

Since the summarization strategy of ERSS 2004 depends only on the text NPs, our initial hypothesis was that the strategy should work just the same (with only a small performance penalty due to translation imprecision)

DOCREF="APW19981018.0423"
Castro; London; dictator Augusto Pinochet; Pinochet's arrest; Ibero-America

DOCREF="APW19981019.0098"
arrest; diplomatic immunity; London; Pinochet's release; government officia

DOCREF="APW19981020.0241"
Margaret Thatcher; husband; home; 82-year-old Pinochet; genocide; minister'

DOCREF="APW19981021.0557"
formal extradition request to British authorities; dictator Augusto Pinoche

Figure 1: Sample Task 1 summaries for four texts from the same cluster

on translated texts. We were partly proven right, as Figure 2 shows for manual translations of the Arabic texts.

The results were, however, quite unsatisfactory for the machine translated versions, as can be seen in Figure 3. Here, the problems of different newspaper style, translation errors, and the high dependency of ERSS 2004 summaries on complete NPs which can form reasonable coreference chains compound to produce incomprehensible gibberish as in the third summary of Figure 3.

DOCREF="AFA19981006.0000.0038"
King Hussein; phase; Chemotherapy; Fayez Al-Tarawneh; node cancer; Amman

DOCREF="AFA19981013.0000.0030"
Jordanian monarch; Marwan Al Muasher; sixth phase; Jordanian Ambassador; Un

DOCREF="AFA19981020.0000.0010"
treatment; King Hussein; Mayo Clinic Hospital spokeswoman; Jordanian monarc

DOCREF="AFA19981029.0000.0003"
Rochester; King Hussein's Treatment Nearly Reaches End; Mayo Clinic Hospita

Figure 2: Sample Task 3 summaries of four manually translated texts from the same cluster

Subsequent experiments showed that using an alternative set of machine translated text would have given slightly better performance in the Rouge score, yet to the human reader the improvement is still marginal. Given the superior performance of Lakhas, the only system that worked from the original Arabic texts [6], we don't feel it worthwhile to compensate for translation errors but would rather spend the effort on developing resources to work on the Arabic texts directly and translate only the results.

3.3 Performance and Evaluation

As Figure 1 illustrates, the summaries extracted by ERSS 2004 are useful to humans who have information on

DOCREF="AFA19981006.0000.0038"
every stage; King Hussein; Jordanian monarch; Prime Minister; medical supervision

DOCREF="AFA19981013.0000.0030"
King Hussein; Ambassador Jordanian in Washington Marwan Al-Muashar; Phase IV

DOCREF="AFA19981020.0000.0010"
treatment make; behalf hospital; Jordanian monarch; King Hussein; doctors

DOCREF="AFA19981029.0000.0003"
King Hussein; hospital; treatment; statement; end; Jordanian monarch; cancer

Figure 3: Sample Task 3 summaries of four machine translated texts from the same cluster

	Baseline	**ERSS-2004**	**Rank**
Task 1	0.22	0.2	6/18
Task 3 (manual)	0.14	0.256	2/10
Task 3 (MT IBM)	0.14	0.184	7/10
Task 3 (MT ISI)	0.14	0.2	-

Figure 4: Rouge-1 scores for 75 byte summaries

the given topic. The most explicit summary is the fourth one, which is by the same token not very revealing about the particular text—how is it distinguished from the other texts in the cluster? Here again, the potential user profile and purpose of the system play a crucial role in evaluating the output.

Unfortunately, manually derived SEE scores for coverage and usefulness are not available for Tasks 1 and 3. To assess the improvement in performance on very short summaries from 2003 to 2004, ERSS 2004 was run on the DUC 2003 data and results compared to the Rouge scores (see below) for ERSS on the same dataset. While the bulk of summaries was only slightly changed (and change went both ways, improving low scoring summaries and degrading high scoring ones), some that had shown no overlap with model summaries now had some overlap. Overall, there was a 10% increase in Rouge score, which corresponds to coverage. In DUC 2003 usefulness was generally higher than coverage for ERSS yet the expectation that usefulness remains unchanged when coverage increases cannot be shown without human intervention.

DUC 2004 conducted for the first time mainly automatically scored evaluations called *Rouge* [9]. Rouge-n is fundamentally an n-gram matching scheme between a peer summary and a model summary. Rouge-1 scores for our very short summaries are summarized in Figure 4. Our Rouge-2 scores were much weaker than the Rouge-1 scores and are not reported here. On Task 1, ERSS-2004 scored just slightly below the baseline (the 75 first bytes of the text)[2], whereas in Task 3 it was above the baseline on all data sets: manual, IBM, and ISI translations. Note that the summaries on the ISI data set were not submitted to competition, but obtained post-DUC. The best relative performance was achieved on the manual translations, the IBM translation is chunk to chunk and produces more incomplete and ambiguous (that is adjacent) NPs, whereas the ISI translation performs a syntax tree transformation that lessens the resulting ambiguity perceived by our chunker (on average the number of parsed units dropped by ca 15% for MT input).

[2]Only one system beat the baseline for Task 1.

4 Multi-ERSS – Short Multiple-document Summaries

This section describes Multi-ERSS, an extension of ERSS-2004 to compute NP coreference chains across documents and to select the most important NPs in the most important chains to extract the sentences for the summary.

Multi-ERSS participated in Task 2 (clusters of English documents) and Task 4 (clusters of translated Arabic documents).

Again, the system was expected to be largely independent of the language of the source documents and degrade only slightly in the presence of a focus question.

4.1 Summarization Strategy

Inter-document coreference chains are at the center of Multi-ERSS. Inter-document coreference differs significantly from intra-document coreference, for instance no pronoun resolution should be allowed across documents and synonym and hypernym relations are insufficient to indicate coreference across documents. The coreference heuristics used across documents are thus only a subset of the heuristics used within documents, relying mostly on named entities.

Multi-ERSS constructs short summaries (665 bytes) by ranking the NPs from all documents based on two simple features: the length of the cross-document coreference chains and within a cross-document coreference chain the length of the NP itself (encoding the assumption that longer NPs provide more information about an important entity). Eliminating sentences with material in double quotes and repetitions, the sentences with the highest-ranking NPs are extracted until the length limit has been reached and then the sentences are sorted first by document, then by their order within the documents.

No post-processing is performed on these summaries due to time constraints.

4.2 Tasks 2 and 4

Task 2 is to summarize 10 English documents of a topical cluster in less than 665 bytes. The result should not be repetitive, there should be no dangling references and the summary should be grammatical. The baseline of Task 2

consists of the first 665 bytes of the text of the most recent document in the cluster.

Multi-ERSS produces summaries of the form given in Figure 5.

> President Yoweri Museveni insists they will remain there until Ugandan security is guaranteed, despite Congolese President Laurent Kabila's protests that Uganda is backing Congolese rebels attempting to topple him. After a day of fighting, Congolese rebels said Sunday they had entered Kindu, the strategic town and airbase in eastern Congo used by the government to halt their advances. The rebels accuse Kabila of betraying the eight-month rebellion that brought him to power in May 1997 through mismanagement and creating divisions among Congo's 400 tribes. A day after shooting down a jetliner carrying 40 people, rebels clashed with government troops near a strategic airstrip in eastern Congo on Sunday.

Figure 5: Sample TASK 2 summary for Multi-ERSS (Rouge score: 0.36, average)

Multi-ERSS does well on clusters 31043, 31009, 30036 and 30040, where each single document within the cluster is in the same style and on the same topic. Multi-ERSS performs badly on clusters where the individual articles present different aspects of a common topic. Natural disasters are a case in point (e.g. cluster 30002 on Hurricane Mitch), as are summary topics like the United Stated Midterm Election (cluster 30050), where we find nine articles reporting on different elections in nine different states. These clusters demonstrate topics for which this summarization strategy is not well suited.

Task 4 is the same task performed on translated Arabic texts. Again, there was manually and machine translated data. An example of a summary from machine translated data is given in Figure 6.

> said Jacobs the speaking on behalf hospital May in Rochester in Minnesota that "the treatment make as expected . said the Jordanian monarch in the United States, where receive treatment in a telephone call by with him television official Jordanian yesterday evening Friday " with regard to the chemotherapy ended last stage during the the first 10 days recent and there is no impact of the disease same item . " in his statement to the Palestinian people Jordanian emplacement television official, he said Prince El Hassan brother King Hussein of the smallest " while aid to my words this be Hussein had left the hospital) . . (and may recovery and discovery of disease .

Figure 6: Sample Task 4 summary from machine translated text.

4.3 Performance and Evaluation

We ran identical versions of Multi-ERSS on both, Task 2 and Task 4. The Rouge score in the DUC 2004 competition is reported in Figure 7.

	Baseline	**Multi-ERSS**	**Rank**
Task 2	0.32	0.36	8/16
Task 4 (manual)	0.33	0.39	7/11
Task 4 (autom. IBM)	0.33	0.36	7/11
Task 4 (autom. ISI)	0.33	0.36	-

Figure 7: Rouge-1 scores for 665-byte summaries

Not unexpectedly, performance on the machine translated data was worse, but not unreasonably so, given that the system has not been tuned to this type of data at all. The Rouge scores for Tasks 2 and 4 illustrate nicely that absolute Rouge score is not meaningful. The rank (top two thirds) indicates that all systems suffered from the quality of the input data, but that some compensated for it. Multi-ERSS can be said to be robust under degraded input, but intuitively not very useful.

DUC 2004 provided two evaluation streams for Task 2: the manual SEE evaluation as well as the automatic ROUGE score. The manual SEE evaluation ranks the degree of overlap of peer summary subunits with the model summaries. SEE evaluations compute coverage and linguistic quality scores for the systems (see `http://duc.nist.gov/duc2004/protocol.html` for the protocol description).

Mean coverage for Multi-ERSS on Task 2 was ranked 7/16 at 24%, with the mean for all systems at 20% and the mean for manual summaries at 45%.

There were seven linguistic quality questions (the rank of Multi-ERSS is given in parentheses). Question 1 asks whether the summary builds to a coherent body of information (10/16). Question 2 asks whether there is useless, confusing, or repetitive text (9/16). Question 3 asks whether the same content is repeated (3/16). Question 4 asks whether noun phrase referents are clear (12/16). Question 5 asks whether entities are re-mentioned (9/16). Question 6 asks whether there are ungrammatical sentences(14/16). Question 7 asks whether there are datelines in the text or other formatting or capitalization errors (8/16).

The quality questions show clearly the weaknesses of Multi-ERSS: the references of the NPs are not clear, because no coherence is built up in the summary—sentences are ordered by the ordering of the texts and by their occurrence within the text. This was expected. The high ungrammaticality score compared to other systems on the other hand is unexpected in an extract-based system.

In summary, Multi-ERSS on Task 2 proved to be slightly above average in coverage and slightly below average in linguistic quality. Since Multi-ERSS has no linguistic processing to improve coherence of the summary text this is an unexpectedly good performance.

Task 4, like Task 3, is based on Arabic to English translations. Surprisingly, Multi-ERSS performs at the same level for both, data from the manual and machine translations.

5 "Who is . . . " Summaries

Task 5 asks for a summary from a cluster of texts focused by a query in form of a single named entity. Multi-ERSS produced the following summary to the question "Who is Stephen Hawking?"

> Hawking, 56, is the Lucasian Professor of Mathematics at Cambridge, a post once held by Sir Isaac Newton. Hawking, 56, suffers from Lou Gehrig's Disease, which affects his motor skills, and speaks by touching a computer screen that translates his words through an electronic synthesizers. Stephen Hawking, the Cambridge University physicist, is renowned for his brains. Hawking, a professor of physics an mathematics at Cambridge University in England, has gained immense celebrity, written a best-selling book, fathered three children, and done a huge amount for the public image of disability. Hawking, Mr. Big Bang Theory, has devoted his life to solving the mystery of how the universe started and where it's headed.

Figure 8: Multi-ERSS summary for the question "Who is Stephen Hawking?"

5.1 Summarization Strategy

We use Multi-ERSS almost unchanged by adding the query NP as another single document to the cluster. The summary is produced by extracting only those sentences that contain the most important NPs of the chains that include NPs[3] from the query.

We adopt two strategies to solve Task 5: *simple sentence* selection and fuzzy coreference *clustering*. We also submitted different settings of the IgnoreQuotes[4] parameter for a total of three runs for Task 5:

Run	Strategy	IgnoreQuotes
Priority-1	simple sentence	True
Priority-2	clustering	True
Priority-3	clustering	False

In the Simple Sentence Selection strategy, the chains that include the query NP(s) are ranked by three features: (1) the chain length, (2) the NP's length, and (3) whether the NP appears within an *apposition* construct. *Chain Length* has a factor of 1.0, *Apposition* has a factor of 3.0. Apposition is an important text feature for the characterization of persons, since it typically introduces or elaborates on the named entity and thus provides the most useful information for this kind of focused summary. We select the highest ranking NP from each chain (from each document) and select the sentence it belongs to. The extracted sentences are sorted by their order within a document and the order of the documents within a cluster, but again no post-processing or smoothing of the summary was performed due to time constraints.

The second strategy relies on NP clustering. The clusters are sorted by size and those that do not contain references to the query are removed. Here too, we rank the NPs in each cluster by (1) NP length and (2) apposition, and select the sentence with the highest ranking NP.

5.2 Performance and Evaluation

The Rouge score places Multi-ERSS into rank 7/14, which seems in keeping with its performance on Tasks 2 and 4. Manual SEE evaluation for coverage, similarly, places Multi-ERSS at the systems' mean (incidentially also the baselines' mean). The seven linguistic questions show again clearly strengths and weaknesses of Multi-ERSS, but they do cluster differently from Task 2. It scores below average in Questions 3 (same content repeated, 14/14), Question 5 (entities rementioned, 13/14, Question 6 (ungrammatical sentences, 11/14), Question 2 (useless, confusing, repetitive text, 10/14), and Question 1 (builds coherent body of information, 10/14). For Question 7 (datelines, formatting, capitalization problems) Multi-ERSS scored average (7/14) and on Question 4 (trouble identifying noun phrase referents) 5/14. The difference to Task 2 for Question 4 (12/16) is unclear, the fact that the seed NPs used to select sentences all had to be present in the focus question may have helped to increase cohesion. Note that despite the low rank in Question 1, for instance, the difference between Multi-ERSS' performance and the first ranked performance was not statistically significant.

NIST also computed responsiveness scores for Task 5, Here, all summaries are compared to each other, not a model to simulate extrinsic evaluation of both form and content. Multi-ERSS ranked 2/13 for responsiveness, meaning that despite the questionable linguistic quality and the average coverage against the model summaries, Multi-ERSS presents an intelligible and useful automatic summary to "Who is . . . " questions.

We are particularly pleased with the results for Task 5, even though on one side they are lower than for other tasks. The competition has shown that focus questions can put special expectations on a summary. This has been acknowledged in the IR community by developing different templates and in the Q/A community by developing different strategies for different types of questions. Several enhancements to the general summarization strategy are possible for Multi-ERSS with the expectation to im-

[3] Note that a named entity recognizer might erroneously split a named entity in two.

[4] Ignore material from sentences that contain double quotes.

prove the score on the linguistic questions. But the high responsiveness ranking for our very general and very simple system is a strong endorsement of the validity of our basic summarization strategy.

6 Lessons Learned

Rouge [9] has emerged as an important tool to compare a system's development (see our comparison of ERSS and ERSS 2004) and to rank the field of systems. While the community still has to develop better intuitions, it seems clear that Rouge is a good indicator of coverage for extract based systems. Systems that achieve higher compression through reformulation would of course be penalized, as are systems that extract information that was not in the model summary but is relevant nonetheless.

Rouge can, however, mask consistent flaws: Multi-ERSS, for instance, sequences the summary sentences in the order of the source document in the cluster and the order of the sentences within a source document. Temporal coherence in particular is lost. Rouge does not penalize for this shortcoming, thus we must ensure human evaluation even during systems development.

A problem that emerged during DUC 2004 is that the systems scored so close to each other as to make statements about their ranking and respective performance almost meaningless. [10] predicted this as a result of the very large space of possible extract based summaries of this length. The manual responsiveness score, where systems were scored against each other rather than against a model summary, provides an interesting alternative that should be carried on to other DUC competitions.

A correlated outcome is that for several systems, including Multi-ERSS, multiple runs did not in fact allow to rate the importance of various systems features, because just as the systems ranked so close together, different versions of the same system, likewise, were not always even distinguishable. Figure 9 shows some post-competition experiments we did with different features.

Strat.	**Fuz. Deg.**	**Ign. Quot.**	**Rouge-1**	**2**	**4**	**L**
Simple	0.6	T	0.36	0.07	0.008	0.37
Simple	0.6	F	0.35	0.06	0.008	0.36
Cluster	0.6	T	0.35	0.07	0.007	0.36
Cluster	0.6	F	0.34	0.07	0.011	0.35
Simple	0.8	T	0.36	0.08	0.010	0.35
Simple	0.8	F	0.34	0.07	0.010	0.35
Cluster	0.8	T	0.34	0.07	0.010	0.35
Cluster	0.8	F	0.34	0.07	0.001	0.34

Figure 9: Correlation of different summarization strategies, different parameter settings, and Rouge algorithms for Task 2. Rouge-3 is omitted, since results in each row were the same (0.2)

Figure 9 is, however, inconclusive. From other experiments we feel that a merge degree of 0.8 is better and that at the moment the simple sentence summarization strategy is more mature. But any experimental variation in Rouge scores is hard to interpret. Although Rouge-L seems to give us best results in this table, we evaluate using only Rouge-1.

One way to make advances might be to provide training corpora with very particular characteristics: multiple texts that report on exactly the same event, for instance, will generate a smaller space of reasonable summaries and allow to hone recall of important (multiply mentioned) events. Sets of texts that elucidate different angles of a common topic, on the other hand, would allow the system to showcase coherence constructing algorithms and connect it to general world knowledge.

Another important question is the correlation between Rouge-N scores and usefulness. We manually assigned usefulness scores to the seven clusters that were ranked best and to the eight clusters that were ranked worst by Rouge. Usefulness was assessed as a number between 0 and 1, where 0 meant completely useless and 1 perfect. A usefulness of 0.5 meant an average summary. In general, low Rouge score coincided with low usefulness, but there were important outliers, illustrating that Rouge cannot predict usefulness in general. In fact, we have pairs of barely distinguishable 75-byte summaries with drastically different Rouge scores (factor of two) in Figure 10.

Document APW19981016.0240 in d30001t:
Summary with classification header:
People & Politics: country's next president; only other army commander; Syr
ROUGE score: 0.28

Summary without header:
country's next president; only other army commander; Syria; Lebanon; politi
ROUGE score: 0.49

Figure 10: Effect of extra-textual material on Rouge-1 scores.

7 Conclusion

ERSS 2004 and Multi-ERSS are both simple, robust systems that build on a linguistic notion with knowledge-poor, approximative, heuristic methods. Incorporating freely available components from the Web has resulted in a pair of closely related systems that participated successfully in all five tasks in DUC 2004. Their performance is above average overall, demonstrating that heuristics-based methods are competitive. The lack of pre- and post-processing of the texts results in sometimes embarrassingly avoidable glitches, but shows clearly that above average performance can be achieved purely based on con-

tent.

8 Acknowledgments

Our thanks to Jennifer Scott, whose annotations and evaluations are invaluable.

This work was funded in part by an NSERC discovery grant and by a NATEQ nouveau chercheur, volet équipe grant.

References

[1] Sabine Bergler. Towards reliable partial anaphora resolution. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain, July 1997.

[2] J. M. Conroy, J. D. Schlesinger, J. Goldstein, and D. P. O'Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of the HLT-NAACL Workshop on Automatic Summarization, DUC-2004*, Boston, MA, May 6-7 2004.

[3] T. Copeck, N. Japkowicz, and S. Szpakowicz. Text summarization as controlled search. In *Proceedings of the ACM SIGIR Workshop on Text Summarization DUC 2001*, New Orleans, Louisiana, September 13-14 2001. Document Understanding Conference. `http://duc.nist.gov/pubs.html#2001`.

[4] H. Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36:223–254, 2002. `http://gate.ac.uk`.

[5] Document Understanding Conference. *DUC 2002*, Philadelphia, Pennsylvania, USA, July 11-12 2002. `http://duc.nist.gov/pubs.html#2002`.

[6] F.S. Douzidia and G. Lapalme. Lakhas, an Arabic summarization system. In *Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2004*, Boston, MA, May 6-7 2004. Document Understanding Conference. `http://duc.nist.gov/pubs.html#2004`.

[7] S.M. Harabagiu and F. Lacatusu. Generating single and multi-document summaries with gistexter. In *Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2002* [5]. `http://duc.nist.gov/pubs.html#2002`.

[8] W. Kraaij, M. Spitters, and A. Hulth. Headline extraction based on a combination of uni- and multi-document summarization techniques. In *Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2002* [5]. `http://duc.nist.gov/pubs.html#2002`.

[9] Chin-Yew Lin and E.H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference HLT-NAACL 2003*, Edmonton, Canada, May 27 - June 1 2003.

[10] Chin-Yew Lin and E.H. Hovy. The potential and limitations of sentence extraction for summarization. In *Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2003*, Edmonton, Canada, May 31- June 1 2003. Document Understanding Conference. `http://duc.nist.gov/pubs.html#2003`.

[11] K.C. Litkowski. Summarization experiments in duc 2004. In *Proceedings of the HLT-NAACL Workshop on Automatic Summarization, DUC-2004*, Boston, MA, May 6-7 2004.

[12] René Witte. *Architektur von Fuzzy-Informationssystemen*. BoD, 2002. ISBN 3-8311-4149-5.

[13] René Witte and Sabine Bergler. Fuzzy Coreference Resolution for Summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, pages 43–50, Venice, Italy, June 23–24 2003. Università Ca' Foscari. `http://rene-witte.net`.