# Mixed Approach to Headline Extraction for DUC 2003

Maria Fuentes*, Marc Massot*, Horacio Rodríguez§, Laura Alonso†

*Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona
{maria.fuentes,marc.massot}@udg.es

†TALP Research Center
Software Department
Universitat Politècnica de Catalunya
horacio@lsi.upc.es

‡ Departament de Lingüística General
Universitat de Barcelona
UdG Visiting Researcher
lalonso@{lingua.fil.ub.es,ima.udg.es}

**Abstract**

We present a summarization system that produces extractive headlines from single documents. A machine learning approach is applied for detecting the most relevant sentence in a document, which is then compressed by manual rules to obtain a grammatical headline. A first version of the system participated in DUC 2003. Results are analyzed and some improvements are proposed accordingly.

## 1 Introduction

A headline is a highly concise representation of the most relevant points contained in a document. It can consist of a sentence, either extracted from the document or automatically generated, or, sometimes, of a list of relevant terms. The main characteristic of a headline is its extremely small length (usually between 5 and 10 words). So, in the case of extractive Headline Extraction (HE), an aggressive condensation of the extracted sentence(s) has to be performed.

We present a HE system that combines a Machine Learning approach with manual rules to obtain informative, readable headlines at parametrisable length. The system participated in DUC 2003 contest (DUC, 2003) for 10 word summaries. Results have been analyzed to find the weak points at different stages of the process.

The rest of the paper is structured as follows. The next section outlines previous work in HE. Section 3 describes the basic system, and Section 4 presents the analysis of the results obtained at DUC. After weak points are identified, possible improvements are proposed in Section 5.

## 2 Previous Work in Headline Extraction

A lot of different techniques have been applied to single document summarization for:
i) locating the relevant fragments,
ii) ranking them by relevance and
iii) producing the summary.

Most of these techniques have been applied as well to HE. We can find systems ranking from purely statistical approaches to others including different Natural Language Processing tasks. However, three main issues are present in most HE systems:

1. Identification of relevance signalling words.
2. Combination of the results proposed by simpler individual methods.
3. Some form of compression or simplification of the extracted sentences.

(Schiffman, Mani, and Concepcion, 2001) face the identification of importance-signalling words, considering as key issues the lead property (words occurring more frequently in the lead sentences of documents), the specificity of the verbs and the use of "Concept Sets" derived from WordNet instead of words.

(Kraaij, Spitters, and Hulth, 2002) aim to identify the most informative topical (for a cluster of documents) noun phrase. Sentences are ranked by a Their hybrid model that merges i) a unigram language model, mixing cluster and document models, for scoring sentences and ii) a Bayesian model on content features like cue phrases, position, length, etc. A Trigger Word Pool is built from the ranked sentences and the most salient trigger word is selected. Then, all maximal NPs containing the trigger word are found, and the one in the highest ranked sentence is taken as a headline.

(Daumé III et al., 2002) go beyond words or NP and perform a full parsing of the document for getting the main entities and their relations. From them the headline is built.

(Zajic, Door, and Schwartz, 2002) use a Noisy Channel Model (NCM), implemented by means of a HMM, for selecting a stream of headline words from a stream of story words. The first is modelled by a bigram language model and the latter by unigram one. The headline is generated by a simple Viterbi decoder constrained by the model. A set of penalty constraints (length, position, gap and string) has been added for improving the quality of the headlines.

As regards the way of combination, (Lin, 1999) uses a simple linear combination schemata with weights empirically set while (Aone et al., 1997) use Decision trees (DT). (Nobata et al., 2001) consider different variants of 4 scoring functions (sentence position, sentence length, term frequency and distance to the title or headline). The final score of each sentence is computed by a linear combination of the individual scores.

Finally, as regards the way of compression, (Knight and Marcu, 2000) face the problem of sentence compression as the translation of a sentence from a source language (full text) to a target language (summary). Two different approaches are followed: a NCM and a DT. Alternatively (Lal and Rueger, 2002), employ lexical simplification and addition of background knowledge for modifying the initial summaries.

# 3 System Overview

In our approach, headline extraction is carried out in four steps (see Figure 1):

1. **Enrichment**: the document is segmented in Textual Units (TUs) and enriched with features relevant to the task in three phases:
   (a) **Pre-processing**: general NLP tasks that provide information necessary for further processes, namely: Sentence Segmentation, Tokenization and Morphological Analysis, Named Entity Recognition, POS Tagging and Semantic Tagging (by attaching WordNet synsets, with no attempt to Word Sense Disambiguation). The DUC 2002 segmenter has been used for segmentation, details of the other tasks can be found in (Fuentes and Rodríguez, 2002).

(b) **Lexical Chainer**: computes lexical and NE chains (Section 3.1).

(c) **Feature Extraction**: extracts the features needed for classification of each TU. Currently the system uses the features described in Table 1. Numeric features are discretized, the number and limits of the intervals have been empirically set with the training data.

2. **Classification**: each TU is classified as belonging to the summary or not, according to its features and a set of classification rules induced from a training corpus (see Section 3.2). A confidence score is assigned to each decision, based on the confidence associated to the rule applied, and the set of summary TUs is ranked accordingly.

3. **Summary Content**: from the set of ranked TUs, only one is selected for compression (Section 3.3).

4. **Simplification**: the selected TU is parsed and compression rules are applied to achieve the targeted length and maintain informativity (Section 3.4).

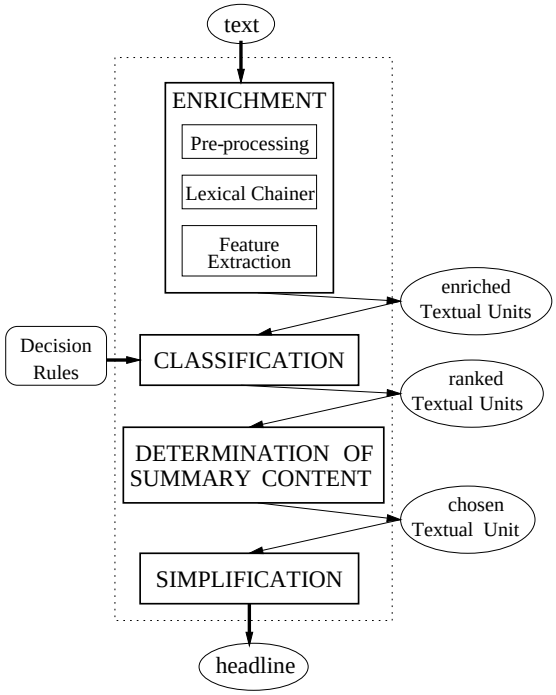


Figure 1: Architecture of the System for Headline Extraction

## 3.1 Obtaining Lexical Chains

To obtain Lexical Chains (LC), the work of (Morris and Hirst, 1991) and (Barzilay, 1997) is followed. We have adapted to English the lexical chainer for Spanish described in (Fuentes and Rodríguez, 2002).

Chain candidates are common nouns and Named Entities. For each chain candidate, three kinds of relations are considered, as defined by (Barzilay, 1997): Extra-strong, Strong and Medium-strong. The strength of chains has been scored using length and kind of relation between their words. Only strong chains have been taken as descriptive features for TUs.

| Feature Types | Feature Names | Value Type |
|---|---|---|
| Length | words, characters, relative_length | one of 5 possible intervals |
| Position | pos_d | one of 6 possible intervals depending on position of TU in document |
| Unigram Overlap | uni_1, uni_2, uni_3, uni_4, uni_5 | number of TUs in document with unigr_overlap with current TU within interval |
| Bigram Overlap | bi_0, bi_1 | number of TUs in document with not null bigram overlap with current TU |
| Simple Cosine | scos_1, scos_2, scos_3, scos_4, scos_5 | number of TUs in document with cosine with current TU within interval |
| Weighted Cosine | cos_1, cos_2, cos_3, cos_4, cos_5 | number of TUs in document with weighted cosine with current TU within interval |
| Lexical Chains | strong_lex_chains | number of strong lexical chains crossing current TU (numeric value) |

Table 1: Features describing TUs for classification as belonging to summary.

## 3.2 Learning the classifier

The classification task is carried out by means of a set of decision rules automatically extracted from a DT. The DT has been learned using the Sipina shell (SIPINA, ). The training corpus was a set of 147 documents with human built extracts, obtained from the DUC 2001 data (Conroy et al., 2001). In total 6933 sentences have been used as examples for training.

Each sentence in this training corpus was described by the features in Table 1, plus the additional feature of belonging or not to the extract (the classification task is binary). Several learning schemata (ID3, C4.5, ...) and parameter sets, available in Sipina, have been experimented with no significative differences between them. The final choice, performing slightly better, has been C4.5. The resulting tree has been transformed into decision rules represented as Prolog clauses. From the training set a total of 84 rules have been extracted with a global accuracy over the training corpus of 94%. Each rule has been scored taking into account its individual accuracy and coverage.

## 3.3 Determining the summary content

The best scored TU among those classified as belonging to the extract and coming through a set of filters (e.g. a filter discards TUs shorter than 12 words, because they tend to convey information about the document and not about the event) is returned by this module. However, as TUs are classified in an independent way, it could be the case that, for a given document, no TU is classified as belonging to the summary. In this case, if there are not classified TUs, the one with lowest position is selected. If all the TUs are classified as not belonging to the summary, then the one with less score is selected.

## 3.4 Simplification

The TU considered as more likely to be included in a summary of the document was parsed by MINIPAR (MINIPAR, ) and compression rules were applied on the obtained parse. Compression rules proceed as follows:

1. find the main verb(s)
2. take syntactically required arguments of main verb(s): subject and objects, but not lexically required ones, like collocative or semantic arguments

3. take complements of main verb(s) that were necessary from the point of view of truth value, for example negative particles
4. take complements of verbal arguments that may specify their truth value, like lexical modifiers
5. take discursively salient sentence constituents, namely, adjuncts marked by a discursive particle signalling relevance
6. fulfill well-formedness requirements

# 4 Analysis of the results

This system participated in DUC 2003 Task 1, which consisted in providing 10 word single document summaries for pieces of news. Figure 2 displays results for coverage and length-adjusted coverage of automatic systems, our system is number 24.

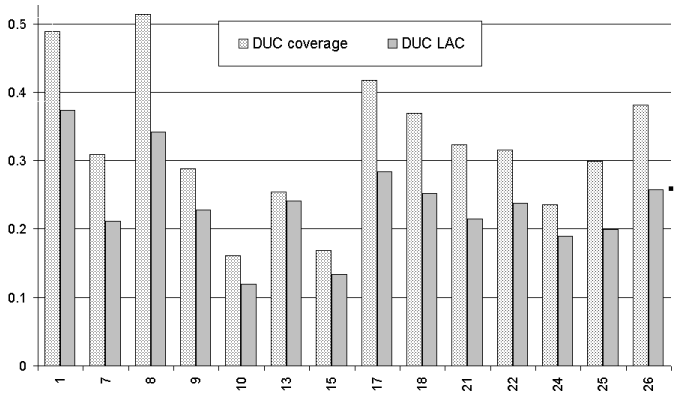


Figure 2: Results for coverage and length adjusted coverage of the systems submitted to DUC 2003 [7 - 26] and a Baseline [1].

In order to guide future improvements on the system, the obtained results and the quality judgements provided by NIST were analyzed. To identify weak points clearly, a failure analysis of the two main components of the system was carried out separatedly. First, the performance of the extraction process was analyzed with respect to coverage. To do that, the coverage of the sentences chosen to be compressed was evaluated, as described in Subsection 4.1. Then, the effects on coverage of the compression process were studied. Additionally, grammaticality was also evaluated, although this aspect was not reflected in DUC results, which provided no quality questions for 10 word summaries.

With respect to length, our mean is quite in target: 8.33 words. However, while few of our headlines are longer than the mean, a considerably bigger number are much shorter than 10. Since grammaticality was priorised, these very short summaries contained many more grammatical words than content words, and were therefore not informative. This problem would not be so severe for a list-of-words approach, where all words in a summary are supposed to be content words. Nevertheless, we believe that grammatical summaries may convey more information than lists of content words because they account for semantic relations between these words.

## 4.1 Coverage analysis of the sentence extraction process

To analyze how the sentence extraction process affects coverage of the resulting summary, we evaluated the coverage of the sentences extracted for compression. To begin with, it must be said that a summary was provided for only 91% of the 624 documents, due to restrictive heuristics for

determining the summary content and lack of robustness of compression rules. However, this was not accounted for as a lack of coverage by DUC evaluation.

To assess the coverage of each of the selected sentences before they were compressed, we took advantage of the quality judgements assigned to other summaries of the same document from systems participating in the competition. We decided to calculate **approximate coverage** scores by weighting of the scores assigned to similar submitted summaries, applying $score = \frac{\sum_{i=1}^{3} v_i s_i^2}{\sum_{i=1}^{3} s_i^2}$, where $s$ is the similarity between a summary and the summary to be scored, by unigram overlap[1], and $v$ is the score for coverage assigned to that summary.

Applying this procedure, sentences extracted for compression were assigned an approximate score for coverage. From the 564 summaries we presented, 211 were assigned 0 for coverage at DUC. From these, 210 were assigned less than 0.1 approximate coverage. However, only 97 of the original sentences from which coverage 0 summaries had been obtained were assigned less than 0.1 approximate coverage, which means that in more than 50% of the cases, the compression process caused a total loss of coverage of the original sentence. Possible causes for this decrease in coverage are analyzed within the compression module, in the next Subsection.

## 4.2 Effects of Sentence Compression

The sentence compression module performed a significant reduction in coverage with respect to the whole sentence to be compressed. To identify the causes of this misperformance, a set of 84 sentences was studied, 58 of which were considered insatisfactory, either from the point of view of coverage or grammaticality. In more than half of the cases (34 sentences), the loss of coverage in reducing the original sentence was due to an inadequate treatment of highly informative elements, like Named Entities or words which are part of a Lexical Chain. This is a shortcoming of compression rules, which were based exclusively in structural information, and did not take into account the lexical status of the elements in sentence constituents.

Additionally, compression rules caused grammaticality errors in 26 sentences, and parsing errors affected 15 sentences, as can be seen in Table 2.

| | coverage | compression | parsing |
|---|---|---|---|
| coverage | 21 | | |
| compression | 10 | 12 | |
| parsing | 3 | 3 | 9 |

Table 2: Typology of errors due to the sentence compression process in 58 summaries.

An example of various kinds of misperformance of the compression process can be seen in Example 4.2.

In this example it can be seen that the resulting compression does not convey much of the relevant information in the original sentence. The main errors in the compression process are:

- bad account of verb argumental structure: since there is no information on the arguments required by the verb *feel*, the adjectival phrase *betrayed* is considered optional by compression rules.
- parsing errors: the attachment ambiguity of *surrounding*, which depends from *scandal*, is not well resolved, and it is considered as directly depending from the verb and it is assigned the status of an optional verbal adjunct by compression rules.
- insensitiveness to the lexical status of words: the fact that Named Entities, like *Quebec City*, *Winter Olympics* or *Salt Lake City*, bear much of the content of the sentence is not captured, because they are not in syntactically salient positions (they are not phrasal heads).

[1]Unigram overlap is normalized by the size of the strings to be compared.

**chosen sentence**

```
TORONTO (AP) Members of the delegation for Quebec City's 2002 Winter
Olympics bid feel betrayed in light of the scandal surrounding the
successful bid by Salt Lake City.
```

**compressed sentence**

```
Members of delegation bid feel.
```

Figure 3: An example of misperformance of the sentence compression process.

- inadequate treatment of multi-word expressions: the lack of relevance of the consituent introduced by *in light of* cannot be found because this expression is not considered as a single discourse marker and cannot be treated consequently.

# 5 Future Improvements on the System

After the analysis of the results, we believe that some simple modifications would improve the performance of the system significantly:

- refining the heuristics for choosing the TU to be compressed will enhance the coberture of the system.
- assigning a coverage status to each lexical item in the chosen TU will prevent coverage loss in the sentence compression process. If words belonging to a strong lexical chain are assigned a high coverage status, decisions as to the inclusion of sentence constituents in the summary can be taken considering not only structural, but also coverage information.
- processing multi-word expressions within MINIPAR, like Named Entities, collocations and discourse markers will allow the inclusion of knowledge associated to these complex linguistic entities.
- inclusion of complex lexical and semantic information, like verb argumental information, will contribute to guarantee grammaticality and improve coverage.
- providing a confidence score to each resulting summary, will allow discarding dubious headlines and switch to simpler methods, like a list of content words, which at least guarantee that some of the content in the original document is covered.

# 6 Conclusions

We have presented a system that automatically produces headlines from single documents. This system uses DT for scoring TUs according to their likelihood to generate the headline. A compression procedure, linguistically guided, is then performed. This double process leads to a good balance between informativeness, compression and readibility of the produced headline.

A first version of the system was submitted and evaluated within DUC-2003. Results leave ample room for improvement. Therefore, a careful analysis of the results has been performed, to guide short-term and mid-term developments of the system.

## References


Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1997. A scalable summarization system using robust NLP. In *Proceeding of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 66–73.

Barzilay, Regina. 1997. *Lexical Chains for Summarization.* Ph.D. thesis, Ben-Gurion University of the Negev.

Conroy, John M., Judith D. Schlesinger, Dianne P. O'Leary, and Mary Ellen Okurowski. 2001. Using HMM and Logistic Regression to generate extract summaries for DUC. In *Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana.

Daumé III, H., A. Echihabi, D. Marcu, D.S. Munteanu, and R. Soricut. 2002. GLEANS: A generator of logical extracts and abstracts for nice summaries. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

DUC. – document understanding conference. http://duc.nist.gov/.

Fuentes, Maria and Horacio Rodríguez. 2002. Using cohesive properties of text for automatic summarization. In *JOTRI'02.*

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, Austin, Texas.

Kraaij, W., M. Spitters, and A. Hulth. 2002. Headline extraction based on a combination of uni- and multidocument summarization techniques. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Lal, P. and S. Rueger. 2002. Extract-based summarization with simplification. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.

Lin, Chin-Yew. 1999. Training a selection function for extraction. In *ACM-CIKM*, pages 55–62.

MINIPAR. www.cs.ualberta.ca/l̃indek/minipar.htm.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational linguistics*, 17(1):21–48.

Nobata, Chikashi, Satoshi Sekine, Masaki Murata, Kiyot aka Uchimoto, Masao Utiyama, and Hitoshi Isahara. 2001. Sentence extraction system assembling multiple evidence. In *Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of C hinese and Japanese Text Retrieval and Text Summarization*, Tokyo.

Schiffman, Barry, Inderjeet Mani, and Kristian J. Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *EACL'01.*

SIPINA. http://eric.univ-lyon2.fr/(ricco/sipina.html).

Zajic, D., B. Door, and R. Schwartz. 2002. Automatic headline generation for newspaper stories. In *Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July, 11-12.