Published in Proceedings of DUC Workshop at HLT/NAACL Conferente May 2-7, 2004, Boston, USA.

Keyphrase Extraction for Summarization Purposes: The LAKE System at DUC-2004

Ernesto D'Avanzo
ITC-irst
38040, Povo (TN), Italy
davanzo@itc.it

Bernardo Magnini
ITC-irst
38040, Povo (TN), Italy
magnini@itc.it

Alessandro Vallin
ITC-irst
38040, Povo (TN), Italy
vallin@itc.it

Abstract

We report on ITC-irst participation at Task 1 (very short document summaries) at DUC-2004. We propose to exploit a keyphrase extraction methodology in order to identify relevant terms in the document. The LAKE algorithm first considers a number of linguistic features to extract a list of well motivated candidate keyphrases, then uses a machine learning framework to select significant keyphrases for a document. With respect to other approaches to keyphrase extraction, LAKE makes use of linguistic processors such as multiword and named entities recognition, which are not usually exploited.

1 Introduction

ITC-irst participated in the DUC-2004 evaluation exercise, task 1 (*very short single document summaries*, limited to 75 bytes), with a system, called LAKE, which exploits the role of Keyphrase Extraction (hereafter KE) as a useful approximation to summarization. Our decision to participate was mainly motivated by the fact that some features of Task 1, i.e. the length limit of the output summaries and the fact that summaries could be returned as lists of disjointed items, seemed to fit well in a KE approach. This is the first participation of ITC-irst at DUC.

Keywords, or keyphrases[1], provide semantic metadata that characterize documents, producing an overview of the subject matter and contents of a document. Keyword extraction is a relevant technique for a number of text-mining related tasks, including document retrieval, Web page retrieval, document clustering and summarization, Human and Machine Readable Indexing and Interactive Query Refinement (see (Turney, 2000) and (Gutwin et al., 1998)).

There are two major tasks exploiting keyphrases: keyphrase assignment and keyphrase extraction (see (Turney, 1999)). In a keyphrase assignment task there is a predefined list of keyphrases (i.e, a *controlled vocabulary* or *controlled index terms*). These keyphrases are treated as classes, and techniques from *text categorization* are used to learn models for assigning a class to a given document. A document is converted to a vector of features and machine learning techniques are used to induce a *mapping* from the feature space to the set of keyphrases (i.e. labels). The features are based on the presence or absence of various words or phrases in the input documents. Usually a document may belong to different classes.

In keyphrase extraction (KE), keyphrases are selected from the body of the input document, without a predefined list. When authors assign keyphrases without a controlled vocabulary (*free text keywords* or *free index terms*), typically about 70% to 80% of their keyphrases appear somewhere in the body of their documents (Turney, 1997). This suggests the possibility of using author-assigned free-text keyphrases to train a KE system. In this approach, a document is treated as a set of candidate phrases and the task is to classify each candidate phrases as either a keyphrase or non-keyphrase (Turney, 1997; Frank et al., 1999). A feature vector is calculated for each candidate phrase and machine learning techniques are used to learn a model which classifies each candidate phrase as a keyphrase or non-keyphrase.

The paper is organized as follows. In Section 2 we report on the general architecture of our system, which combines a machine learning approach with a linguistic processing of the document. Section 3 shows the results obtained by the system and discusses the evaluation carried out with ROUGE. We conclude suggesting

[1] Throughout this document we use the latter term to subsume the former.

possible future improvements.

2 The LAKE System

In this section we describe LAKE (*Learning Algorithm for Keyphrase Extraction*), a keyphrase extraction approach developed at ITC-irst based on a supervised learning approach that makes use of a linguistic processing of the documents. The system works in two phases (see Figure 2): first it considers a number of linguistic features to extract a list of well motivated candidate keyphrases from a given document; then it uses a machine learning framework to select significant keyphrases for that document.

More in detail, *candidate phrases* consist of words or sequences of words that match a set of previously manually defined linguistic patterns (see section 2.2). Then (see section 2.3), a supervised learning algorithm is used to score the head of each phrase, according to features (e.g. TF/IDF and the position within a document) that signal the relevance of the head in the whole document collection. The motivation for considering the heads of the candidate phrases instead of the phrase itself is that phrases do not appear frequently enough in the collection. Finally, the score of the head is assigned to the whole candidate phrase, and the best scored phrases that fill the 75 bytes required by the task are given as output.

Both the linguistic patterns and the features used in the classification phase have been defined considering the DUC-2003 material.

In the following sections we describe each of these components.

2.1 Linguistic Pre-Processing

The document collection provided by DUC was preprocessed according to the following three steps: (i) part of speech tagging, (ii) multiwords recognition, (iii) named entities recognition.

Part of Speech Tagging. We use the Tree tagger (Schmid, 1994) POS tagger developed at the University of Stuttgart. The tagged text is given to a module that recognizes multiword expressions.

Multiwords Recognition. Sequences of words that are considered as single lexical units are detected in the input document according to their presence in WordNet (Fellbaum, 1998). For instance, the sequence "Christmas trees" is transformed into the single token "christmas_tree" and assigned to the part of speech found in WordNet.

Named Entities Recognition. As for Named Entities recognition we used NERD (Magnini et al., 2002), a multilingual Named Entity Recognizer for Italian and English. The system has been designed for the identification and the categorization of entity names (persons, locations and organizations), temporal expressions (dates and times) and certain types of numerical expressions (measures, monetary values, and percentages) in written texts. NERD relies on the combination of a set of language-dependent rules (approximately 350 for English and 400 for Italian) with a set of language-independent predicates, defined on the WordNet hierarchy, for the identification of both proper nouns and *trigger words*. The system, integrated into the DIOGENE Question Answering architecture, has been successfully used for the ITC-irst participation to the last two editions of the TREC QA main task, and to the first edition of the CLEF multiple language QA track.

Figure 1 reports a sample input document and its corresponding output after it has been processed by NERD.

```
<DOC>
<DOCNO> APW19980314.0392 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 03/14/1998 10:36:00 </DATE_TIME>
<BODY>
<HEADLINE>
Russian show of power in World Cup finale
</HEADLINE>
<TEXT>
OSLO, Norway (AP) _ Russia's Alexey Prokurorov had
no real challengers as he won the 50-kilometer
classical-style cross country World Cup race in
Holmenkollen Saturday, clocking 2 hours,
32 minutes, 25.3 seconds.
</TEXT>
</BODY>
</DOC>
```

Output format:

```
<DOC>
<DOCNO> APW19980314.0392 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 03/14/1998 10:36:00 </DATE_TIME>
<BODY>
<HEADLINE>
Russian show of power in World Cup finale
</HEADLINE>
<TEXT>
<b_enamex type="LOCATION">OSLO<e_enamex>,
<b_enamex type="LOCATION">Norway<e_enamex>
(<b_enamex type="ORGANIZATION">AP<e_enamex>) _
<b_enamex type="LOCATION">Russia<e_enamex>'s
<b_enamex type="PERSON">Alexey Prokurorov<e_enamex>
had no real challengers as he won the
<b_numex type="MEASURE">50-kilometer<e_numex>
classical-style cross country World Cup race in
<b_enamex type="LOCATION">Holmenkollen<e_enamex>
<b_timex type="DATE">Saturday<e_timex>
<b_timex type="DURATION">2 hours, 32 minutes<e_timex>
<b_timex type="DURATION"> 25.3 seconds <e_timex>.
</TEXT>
</BODY>
</DOC>
```

Figure 1: Input and output of the NERD system

Figure 2: Architecture of the LAKE system.

2.2 Candidate Phrase Extraction

The selection of relevant phrases has been strongly task-oriented. Within the framework of the DUC summarization task, the concept of relevance was heuristically defined: we considered significant the syntactic patterns that described either a precise and well defined entity, or concise events/situations. In the former case we focussed on uni-grams and bi-grams (for instance *Named Entity*, *noun*, *adjective+noun*, etc.), while in the latter we considered longer sequences of parts of speech, often containing verbal forms (for instance *noun+verb+adjective+noun*). Despite the important role heuristics play in our approach, the choice of relevant patterns was also linguistically motivated: sequences like *noun+adjective*, that are not allowed in English, were not taken into consideration. We discarded the patterns (uni-grams, bi-grams and tri-grams) that appeared with low frequency (less than 100 times) in the training corpus and we eliminated those patterns containing punctuation.

We manually selected a restricted number of PoS sequences that could have been significant in order to describe the setting, the protagonists and the main events of a newspaper article. To this end, particular emphasis was given to named entities, proper and common names. In order to outline three different settings (i.e. runs), we wrote three versions of our PoS-pattern filter: one containing 121 patterns (without any verbal form), one containing 223 patterns (with a few verbal constructions) and the largest one that consisted of 654 patterns (with verbs, adverbs and prepositions). Once we had extracted all the uni-grams, bi-grams, tri-grams, and four-grams from the PoS-tagged output of NERD, we filtered them with the patterns we previously defined.

As an example, we can consider the document APW19981023.1166, that reports on the possible extra-diction of Pinochet from London (where he is recovering from an operation) to Spain. Table 1 shows some of the candidate phrases that our largest filter accepted as candidates from this document.

2.3 Scoring Candidate Phrases

In this phase we assign a score the each candidate phrases in order to rank them and to make it easier their selection for the final 75 bytes output. The basic idea was to implement a binary classifier that, given a candidate phrase, is able to classify it either as a relevant keyphrase for a given document or as not relevant for that document. The classifier was trained on two features: $TF \times IDF$ (i.e. the product between the frequency of a candidate phrase in a certain document and the inverse frequency of the phrase in all the documents belonging to the same cluster) and *First Occurrence*, i.e. the distance of the candidate phrase from the beginning of the document in which it appears. Two reasons led us to choose these features: they are very easy to calculate and many systems performing at the state-of-the-art make use of them.

However, since the frequency of a candidate phrase in the whole collection was not significant, we decided to estimate the values of the two features using the head of the candidate phrase, instead of the phrase itself. According to the *principle of headedness* (Arampatzis et al., 2000), any phrase has a single word as head. The head is the main verb in the case of verb phrases, and it is a noun (the last noun before any post-modifiers) in noun phrases.

As for the learning algorithm we used the Naive Bayes Classifier provided by the WEKA package (Witten et al., 1999) and publicly available at the University of Waikato Web site [4]. The classifier was trained on the DUC-2003 material in this way. From the document collection we extracted all the nouns and all the verbs. Each of them was marked as a positive example of relevant keyphrase for a certain document if it was present in the assessor's judgment of that document, otherwise it was marked as a negative example. Then the two features (i.e. $TF \times IDF$ and first occurrence) were calculated for each word and the learner was run.

As a result, the classifier prefers the candidate phrases whose head both maximizes its $TF \times IDF$ and tends to occur at the beginning of a document. For example, the head *dictator* receives an high $TF \times IDF$ relatively to the document APW19981023.1166, a fact that suggests that a candidate phrase such as *Chilean dictator*, selected by a linguistically motivated pattern, is likely to be relevant for that document.

3 Results and Discussion

Participating groups at DUC-2004 were allowed to submit up to three runs. Submissions were ranked according to their priority, i.e. according to the confidence in the results. As mentioned above (section 2.2), we designed three different settings, corresponding to the three versions of linguistic patterns we used. We assigned the highest priority to the run that exploited all the 654 patterns we manually extracted, because it is able to mark as candidate phrases also complex sequences including verbs. The shortest filter, containing no verbs, was used in the run with priority 2, while the intermediate one was chosen as priority 3.

Submissions were automatically evaluated using the ROUGE program (Lin and Hovy, 2003). Table 2 shows the results obtained on the three runs, each with a different ROUGE score for uni-gram (Rouge-1), bi-gram

[4] http://www.cs.waikato.ac.nz/ml/weka/

Table 1: A sample of candidate phrases extraction. We use the following abbreviations (borrowed from Penn Treebank Project[3] (see (Marcus et al., 1994))): NE stands for Named Entity, JJ for Adjective, NN for a singular noun, CC for a conjunction, VDB for a verb at the past tense, MD for a modal auxiliary, VBN for a verb at the past participle, IN for a preposition, NNS for a plural noun, TO for the preposition TO, VB for a verb in its base form, DT for a determiner.

	Pattern	**Example**
Uni-grams	NE	*London*
	NE	*1973*
Bi-grams	JJ+NN	*Chilean dictator*
	JJ+NN	*Spanish magistrate*
	JJ+NN	*urinary infection*
Tri-grams	NN+CC+NN	*genocide and terrorism*
	NN+VBD+NE	*newspaper reported Friday*
	NN+VBD+NN	*room lacked television*
Four-grams	NE+MD+VB+VBN	*Augusto Pinochet would be extradicted*
	VBN+IN+JJ+NNS	*detained by British police*
	NN+TO+VB+NN	*extradition to stand trial*
	NN+VBD+JJ+NN	*dictatorship caused great suffering*
	VBN+IN+DT+NE	*detained in the London*
	NN+VBD+VBN+NN	*family had asked police*

(Rouge-2), tri-gram (Rouge-3), four-gram (Rouge-4), and the longest common substring (Rouge-L). The best result has been obtained by LAKE on the second run, using a set of 223 linguistic patterns.

3.1 Discussion on the Linguistic Patterns

We encountered two major difficulties in definining the linguistic patterns:

1. since the relevance of a keyphrase cannot be defined exclusively from a syntactic point of view, the number of relevant patterns was potentially too large to be described. Important information that a human would extract from a document could be contained in every pattern (independently from its part of speech) or in several, non consecutive patterns;

2. it was particularly difficult to determine a priori the syntactic role of the sequences containing verbal forms: we aimed at extracting mainly noun phrases, but we often got truncated and irrelevant sentences.

To sum up, PoS-tagging information proved to be far from exhaustive, introducing a lot of noise. Some candidate phrases turned out to be useless pieces of longer sentences (*newspaper reported Friday*), or irrelevant, though syntactically complete, noun phrases (*urinary infection*, *family had asked police*). The shorter filter, containing no verbs, proved to be the most reliable one, as the automatic evaluation using the ROUGE software showed.

Table 2: Results of the LAKE system for three runs.

	Run 87	Run 88	Run 89
Rouge-1	0.18448	0.18817	0.18362
Rouge-2	0.03638	0.04001	0.03667
Rouge-3	0.00786	0.00933	0.00892
Rouge-4	0.00171	0.00207	0.00189
Rouge-L	0.15412	0.15656	0.15294

3.2 Discussion on the Features

The two features we used to train a classifier for candidate phrase scoring revealed as effective as we thought. However, it might improve the performance of the classifier to consider features able to capture some semantic properties of keyphrases. In particular, the fact that relevant keyphrases in a document tend to be related to the main topic of the document could be captured computing the lexical chains (Barzilay and Elhadad, 1997) of the document, and then considering the membership to such chains as a feature of the candidate phrase.

4 Conclusion and Future Work

In this paper we reported on ITC-irst participation at Task 1 (very short document summaries) at DUC-2004. We have described the LAKE system, which exploits keyphrases extraction for summarization. The system couples a rather sophisticated linguistic analysis of the documents, used for candidate phrases extraction, with

a simple binary classifier, used for assigning a score to candidate phrases. The linguistic processing includes both multiwords and named entities recognition, while the classifier uses as feature both TF/IDF and the position in the document.

Since this was our first participation at DUC, most of our effort was devoted to setting up the basic architecture of the system. For the future we intend to investigate the role of semantic-oriented features and to experiment different machine learning approaches.

5 Acknowledgments

We would like to thank people at TCC division at ITC-irst. In particular we are indepted to Matteo Negri for the NERD system, Claudio Giuliano and Alfio Gliozzo, for the suggestions on the machine learning apparatus, Bonaventura Coppola for his decisive help in the last days before the deadline, and Alberto Lavelli for discussing the overall approach.

References

A. Arampatzis, T. van der Weide, C. Koster, and P. van Bommel. 2000. An evaluation of linguistically-motivated indexing schemes. In *Proceedings of the BCSIRSG '2000*.

R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. In *Intelligent Scalable Text Summarization Workshop, ACL*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Domain-specific keyphrase extraction. In *IJCAI*, pages 668–673.

C. Gutwin, G. Paynter, I. Witten, C. NevillManning, and E. Frank. 1998. Improving browsing in digital libraries with keyphrase indexes. Technical report, Department of Computer Science, University of Saskatchewan, Canada.

Chin-Yew Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of 2003 Language Technology Conference*.

B. Magnini, M. Negri, H. Tanev, and R. Prevete. 2002. A WordNet-Based Approach to Named Entities Recognition. In *Proceedings of the SemaNet'02 workshop on Building and Using Semantic Networks*, Taipei, Taiwan.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

P.D. Turney. 1997. Extraction of keyphrases from text: Evaluation of four algorithms. Technical Report ERB-1051. (NRC #41550), National Research Council, Institute for Information Technology.

P.D. Turney. 1999. Learning to extract keyphrases from text. Technical Report ERB-1057. (NRC #41622), National Research Council, Institute for Information Technology.

P.D. Turney. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval*, 2 (4):303–336.

Ian H. Witten, Eibe Frank Eibe Frank Ian H.Witten Eibe Frank, Eibe Frank Ian H. Witten, and EEibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.