

Automatic Extraction of Knowledge from Greek Web Documents

Fotis Lazarinis
Technological Educational Institute of Mesolonghi
30200 Mesolonghi, Greece
0030-26310-58148
lazarinf@teimes.gr

ABSTRACT

Extracting textual data from Greek corpuses poses additional difficulties than in English texts as inclinations and intonation differentiate terms of equal information weight. Pre-processing and normalization of text is an important step before the extraction procedure as it leads to fewer rules and lexicon entries, thus to less execution time and greater success of the mining process. This paper presents a system accessible via the Web which automatically extracts data from Greek texts. The domain of conference announcements is utilized for experimentation purposes. The success of the extraction procedure is discussed on the basis of an evaluative study. The conclusions and the techniques discussed are applicable to other domains as well.

Categories and Subject Descriptors

H.2.8 [Database Application]: Data mining

H.3 [Information Systems]: Information Search and Retrieval

Keywords

Web mining, information extraction, XML storage, multilingual retrieval

1. INTRODUCTION

Some recent studies showed that common search engines supporting Greek do not actually understand specific characteristics of the language [7, 8] so utilizing a general purpose search engine to discover specific information such as dates, keywords or even general purpose terms demand more effort by the user resulting also to lower success. This is mainly due to differences in Greek terms caused by inclinations, intonation and lower and upper case forms.

In this paper we present a tool for extracting the title, keywords, event date, submission deadline and location of conference announcements. This tool is based on the identification of patterns and on knowledge lexicons (dictionaries) for extracting the previously mentioned data. Pre-processing and normalization of text is an important step before the extraction procedure as it leads to fewer rules and lexicon entries and to greater success of the mining process. Our main aim is not simply to build a system with extraction capabilities but to explore additional inconveniences and present solutions applicable in mining data from Greek corpuses which show considerable grammatical diversity although they carry the same information weight. The conclusions of this work could be applied to other spoken languages with similar characteristics to the Greek language.

2. EXTRACTING TEXTUAL DATA

Information extraction systems analyze unrestricted text in order to extract specific kind of information. They process documents trying to identify pre-defined entities and the relationships between them, filling a structured template with the mined information. Such systems have been implemented to extract data such as names and scientific terms from chemistry papers [2, 12]. Gaizauskas and Robertson [4] used the output of a search engine as input to a text extraction system. Their domain was management succession events and their scenario was designed to track changes in company management.

More contemporary work uses co-occurrence measurement in order to identify relationships and to extract specific data from Web pages [9]. Han et al [5] extract personal information from affiliation, such as emails and addresses, based on document structure. Efforts on Greek information extraction are recorded as well. In [11] a rule based approach to classify words from Greek texts was adapted. Rydberg-Cox [14] describes a prototype multilingual keyword extraction and information browsing system for texts written in Classical Greek. This system automatically extracts keywords from Greek texts using term frequency.

Our approach differs from the ones described in the previous paragraphs in that it tries to identify specific information based on rules and on vocabularies of rule activation terms. Also a technique for recognizing term relationships is explored. Additionally classic IR techniques such as suffix and stopword removal [1] are utilized and evaluated in Greek texts.

Proceedings of the sixth Dutch-Belgian Information Retrieval workshop (DIR 2006)

©: the author(s)

3. SYSTEM OVERVIEW

The relevant work done so far, focus mainly on English text neglecting other languages, which are more demanding and challenging in terms of recognition of patterns. In languages like Greek the same information may appear in many different forms, e.g. 11 Μαΐου 2005 or 11 ΜΑΙΟΥ 2005 or Μάιο 11 or 11 Μάη 2005 (11 May 2005), and still convey exactly the same meaning.

In our system, information extracting relies on rule formalisms for each identified entity. Each extraction sub-procedure ends up with one of four alternative results:

- (i) identified (IDN)
- (ii) possibly identified (PDN)
- (iii) not identified (NDN)
- (iv) not applicable (NA)

Strong rule paths produce IDN results while weak rule paths end up in PDN. Strong rules are those which definitely identify the information that accurately falls into one of the known and well defined patterns. Weak rules are those who rely on probability and heuristic methods to infer the data.

Failing to identify some entity may be due to one of two reasons:
i. A rule activates but it fails to complete, so the data is not identified because of our system's inability. These cases, denoted as NDN, could be used for retraining the system and eventually improve mining of data.

ii. The detection of an entity is not possible because it does not exist in the announcement. For example in preliminary announcements the exact conference's date is not yet decided. So NA, adopted by Morrissey's work [10], denotes nonappearance of the hunted piece of information. NDN and NA are preferred over null as they provide the system with different semantics which could be utilized for improving the system's functionality and the searching capabilities.

The extracted data form an XML file based on a short DTD. That way data can be presented in many different forms and utilized by other applications. In order to construct rules that will enable the successful extraction of the desired facts, we examined 25 text files, a small part of our collection consisting of 145 meeting announcements. This analysis allowed us to realize the different patterns the desired data follow and construct the rules. The remaining 120 call for papers were used in the evaluation.

3.1 Text Normalization

From the analysis of the textual data it was considered necessary to normalize the data first. Words are capitalized and accents or other marks are removed. In addition, simple suffix removal techniques were applied. The primitive Greek stemmer, which is analytically described in [8] removes final Greek sigma and transforms some endings such as "ει" and "ηκε" to "ω" among other mild transformations. It has been proved that the factors described in the previous paragraph influence searching of the Greek Web space as well [6, 7].

Abbreviations were automatically replaced by their full form. For example, month names appear abbreviated quite often, e.g Jun (Ιουν) stands for June (Ιούνιος). As a final normalization point, multiple spaces, html tags and other elements, which are not useful at this first version of the system, are removed. We should

indicate though that html tags could prove significant especially in correctly identifying the title and the thematic area, as they provide structure to the information.

The normalization procedure leads to fewer rules and vocabulary entries, thus to less execution time and greater success in the mining process. In English text normalization procedure is simpler as there are no differences between upper and lower case forms, there are no inclinations of verbs and nouns (apart from minor differences between singular and plural forms) and accent marks are absent unlike in Greek.

3.2 Title extraction

Extraction of the title of a conference is based on heuristic rules. The basic idea is that titles appear on the top part of an announcement and they follow a "title" format, i.e. words are in capital letters or start with a capital letter, etc. Obviously normalization should be done after the identification of title as the form of words plays an important role here. Another rule employed is based on the surrounding text and in keywords, like conference, symposium, congress and meeting. As we will see in the evaluation section title identification is quite successful, though some extracted titles are truncated.

3.3 Keyword extraction

Correct identification of the title is also important for classifying the meeting. Classification means the detection of some keywords which describe the meeting. At the moment we base the classification on two techniques. We try to identify sort list of terms by discovering terms such as "conference topics".

Furthermore we explored a technique for constructing pairs of terms describing the conference. This technique is based on co-occurring terms [9]. We define co-occurrence of two terms as terms appearing in the same Web page. If two terms co-occur in many pages, we can say that those two have a strong relation and the one term is relevant to the other. Using words from the top part of an announcement we construct a list of pairs of neighboring terms. Then we try to measure the co-occurrence of these pairs. This co-occurrence information is acquired by the number of retrieved results of a search engine using the coefficient measure $r(a, b) = |a \cup b| / (|a| + |b| - |a \cap b|)$. With $|a|$ we symbolize the number of documents retrieved when we search using term a . Similarly $|b|$ is the number of documents relevant to term b and $|a \cap b|$ is the number of pages containing both terms. The co-occurrence is measured for every pair of terms and the top results are kept, based on a fixed cut off value. So if a conference is about New Technologies in Adult Education "in" is removed and the pairs "New Technologies", "Technologies Adult", "Adult Education" are formed. Then these pairs along with the terms "New", "Technologies", "Adult", "Education" are searched in the Web and the coefficient measure of the term pairs is decided.

Although our first heuristic approach performed well the second technique produced several "bad" instances among some useful two-term keywords. For example in a conference about "Educational Software" the keywords "Educational Games" were produced, which is acceptable and was not stated explicitly in the announcement, but the bizarre keyword "Adult Software" was also produced. Clearly this technique, although promising, needs certain refinements so as to be useful.

3.4 Extraction of dates

3.4.1 Conference's date

The first step in the identification of dates is the construction of a suitable vocabulary containing the normalized month terms that will activate the rules for the extraction of the conference's date. The identification of the date is based on a simple observation. The latest dates, appearing in a call for papers, are most probably the event's start and end dates. Our purpose is to recognize both start and end dates. For example from a date 11-13 June 2005 we extract 11 June 2005 as the start date and 13 June 2005 as the end date.

The date detection procedure initiates when a month or a full date (e.g. 12/05/2006) is found in the text. In that case we first check the succeeding words until the end of the sentence and then the preceding words until the beginning of the sentence. This search aims at identifying the day and the year of the conference and keywords which verify that it is actually the meeting's date. Thus the system needs to be able to keep information preceding and succeeding the rule activation keyword. If more than one date or date range is discovered then the system searches for appropriate keywords.

Rules are a set of *If then else* and *sub ifs*. Document is processed line by line and term by term. At the end of the rule formalism the result is stored in the XML repository. A simplified part of the date extraction procedure in pseudo code is shown below.

```
While not eof and date not identified do
  Separate current line to terms
  While not eof term set do
    Look up Vocabulary
    If month name is found then
      Scan Previous Terms
      Scan Next Terms
    If ... then
      ...
    Else if ... then
      ...
  End
End
End
End
Update conference XML Repository accordingly
```

3.4.2 Submission date

Submission date is trickier than the event's date as is absent in many cases, especially in short announcements. This procedure is complimentary to the previous one as dates which are denoted as meeting's start and end dates should not be checked again. After the extraction of a proper date the surrounding text is scanned for words like deadline (υποβολή), or other synonyms. Clearly these rules are domain dependant and have a high error probability. This procedure ends up mostly with one of the codes PDN, NDN, NA.

3.5 Location extraction

For extracting the location we constructed and utilized an ontology with the major Greek cities and the prefecture in which they belong. This listing also models bordering city and county relations. A city's name will trigger off the rules for the

identification of the desired information. It was proved that normalization of locations names is absolutely essential as they appear in many different forms, e.g. Αθήνα, Αθηνών, Αθήνας (Athens). One problem in the identification of the location arises when a conference is co-organized by more than one institutions. In this case many locations co-exist. Mining is then based on the surrounding context or on the location's tf (term frequency) measured in the whole announcement. If a strong decision is made then the procedure ends up, whereas when a weak decision is made the procedure initiates again when new activation terms appear up.

4. SYSTEM ARCHITECTURE

The system is implemented in Java using JSP and Servlets. For processing the textual information a version of the jflex utility (<http://jflex.de>) is used. A flowchart of the system is shown in figure 1. The conference announcement is submitted either as a url pointing to an html file or it pasted in a text box on the system's web page.

The extracted information is stored in an XML file which is then accessible by the retrieval component of the system. This component, which is currently under development, dynamically forms an index of the processed conferences based on the information found in the XML repository. When projected to the client's browser conferences are classified as open or past and they are categorized based on their date. This tool will also allow multirriteria retrieval of conferences, such as "show me conferences in Athens or near Athens which are about Web mining and will take place this summer". Supporting these queries will be based on the location knowledge base and on the month dictionary.

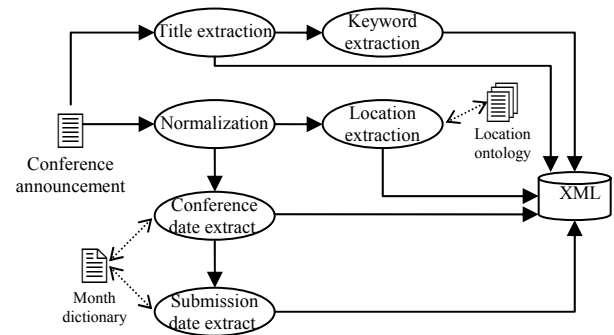


Figure 1. Flowchart of the extraction procedure.

5. EVALUATION

The performance of an information extraction system can be measured using Precision (P) and Recall (R) [13], as in Information Retrieval systems. Precision measures the ratio of the correctly extracted information against all the extracted information. Recall measures the ratio of the correct information extracted from the texts against all the available information. Despite the diversity of the collection the system works adequately well and the employed rules achieve high rates of precision and recall, especially in the attributes where a dictionary is used.

Table 1. Precision and Recall of the extraction procedure

	Title	Keyword	Conf date	Subm date	Location
Correct	77	39	107	89	110
Wrong	29	65	8	19	7
Not extra	14	16	5	12	3
Precision	72,64%	37,50%	93,04%	82,41%	94,02%
Recall	64,17%	32,50%	89,17%	74,17%	91,67%

The results of the evaluation are summarized in table 1. As expected, title and keywords show a higher error percentage. Clearly more sophisticated rules are needed. A possible solution would be the exploitation of tagging information and the usage of lexicons which model domain relationships as well. It should be noted that partially extracted titles, even those with only one not identified word, were accounted as erroneously extracted. So with slight improvements we can achieve higher precision and recall. Date and location rules achieve high precision and recall scores. Their extraction is relying on specific word lists and they follow better structured patterns.

In order to realize the effects of normalization and to get an indication of the additional difficulties posed in Greek we evaluated the system's performance, on date, submission date and location extraction, without extensive normalization. That is words were only capitalized and short forms replaced by their full forms. The evaluation showed that system's precision reduced by more than 30%. It could be argued that in this case more rules should be employed in order to achieve higher precision. While this could be partially true, we need to take into account that more rules means increased execution time as more searches are needed and a higher error probability as more heuristics and weak rules will be employed.

A final evaluation task was performed utilizing Google. A set of five queries concerning specific locations and a second set concerning dates consisting of months and years were run in our collection using Google. Then we evaluated the precision of each query (tables 2 and 3). Clearly Google retrieves many irrelevant files which diminish precision and recall. This is because every file containing the query terms or one of them is retrieved. Furthermore, announcements where terms appear in different forms than the requested ones are not retrieved. In our tool vocabularies act as thesauri as well allowing retrieval of meetings where locations or month names appear in another form or inclination. Of course tables 2 and 3 show an initial estimation. A more thoroughly designed evaluation is needed with more queries to safely reach useful conclusions.

Table 2. Precision and Recall of location queries in Google

Location	Precision	Recall
Query 1	57,50%	76,00%
Query 2	42,86%	83,33%
Query 3	77,78%	83,33%
Query 4	55,88%	64,29%
Query 5	50,00%	65,71%

Table 3. Precision and Recall of date queries in Google

Date	Precision	Recall
Query 1	42,31%	60,00%
Query 2	32,14%	52,38%
Query 3	43,75%	75,00%
Query 4	40,63%	50,00%
Query 5	37,50%	54,29%

6. SYNOPSIS AND FUTURE WORK

This paper presents an under development system which automatically extracts data from Greek conference announcements. Five categories of data are mined utilizing various techniques and approaches. For the first two categories rules are based on text's position, on context surrounding the information and on a coefficient measure. The last three types of data are mined with the utilization of lexicons which contain rule initiation terms. Then the surrounding text is again exploited. It was shown that simple removal of endings and accents and other adjustments, specific to Greek language, improve the extraction procedure and lead to increased Precision and Recall and to less elaborate rules. Vocabularies act as thesauri permitting retrieval of text where terms appear in different forms than the requested ones.

However more work needs to be done in order to achieve high rates of precision. Tagging and formatting information should be utilized in the identification of complex textual information. Metadata and link tracking, in the case of html or xml files, could be utilized. Links usually point to more detailed announcements in which all the data are applicable. Domain vocabularies are necessary in order to identify classification terms. Also, when fully developed, the system should be evaluated against the existing manual or semi automatic conference engines so as to realize all the advantages of our automated system.

Ultimately we aim at building a more complicate system which continually scans the Web to find future conferences, symposiums and congresses. From this combined system XML descriptions of the events could be produced which in turn could be utilized in automatically constructing conference announcement indices. These Web pages will be thematically sorted and automatically and regularly updated, with advanced searching capabilities thus enabling users to find everything in one place. Many issues related to information retrieval are open in the intended system, from categorization of events to summarization and to multicriteria and multilingual retrieval.

7. REFERENCES

- [1] Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, ACM Press, New York, 1999.
- [2] Chowdhury, G. G., Lynch, M. F. Automatic interpretation of the texts of chemical patent abstracts, part 1: lexical analysis and categorisation. *Journal of Chemical Information and Computer Science*, 32, (1992), 463-467.
- [3] Cowie, J, Lehnert, W. Information extraction. *Communications of the ACM*, 39, (1996), 80-91.

- [4] Gaizauskas, R., Robertson, A. Coupling information retrieval and information extraction: a new text technology for gathering information from the web. In *Proceedings of the RIAO'97 Conference*, (Canada), 1997, 356-370.
- [5] Han, H., Giles, L. C., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.. Automatic document metadata extraction using support vector machines. In *Proceedings of the ACM IEEE Joint Conference on Digital Libraries*, 2003, 37-48.
- [6] Lazarinis, F. Do search engines understand Greek or user requests “sound Greek” to them? In *Open Source Web Information Retrieval Workshop* (in conjunction with IEEE/WIC/ACM International Conference on Web Intelligence & Intelligent Agent Technology, France), 2005, 43-46.
- [7] Lazarinis, F. Evaluating user effort in Greek web searching. In *Proceedings of the 10th PanHellenic Conference in Informatics* (University of Thessaly, Greece), 2005, 99-109.
- [8] Lazarinis, F. Old information retrieval techniques meet modern Greek Web searching. In *Data Mining and Information Engineering Proceedings, 2006* (accepted)
- [9] Mori, J., Matsuo, Y., Ishizuka, M., Faltings, B. Keyword extraction from the web for foaf metadata. In *1st Workshop on Friend of a Friend, Social Networking and the Semantic Web* (1-2 September 2004, Galway, Ireland), 2004.
- [10] Morrissey, M. J. *A treatment of imprecise data and uncertainty in information systems*. PhD Thesis, Department of Computer Science, University College, Dublin, Ireland, 1987.
- [11] Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C. Resolving part-of-speech ambiguity in the Greek language using learning techniques, In *Proceedings of the ECCAI Advanced Course on Artificial Intelligence* (ACAI, Chania, Greece), 1999.
- [12] Postma, G. J., Van der Linden, J. R., Smits, J. R. M., Kateman, G. TICA: a system for the extraction of analytical chemical information from texts. In Karjalainen E J (ed) *Scientific Computing and Automation*. Elsevier, Amsterdam, 1990, 176-181.
- [13] Robertson, S. E. The parameter description of retrieval systems: overall measures. *Journal of Documentation*, 25, 1969, 93-107.
- [14] Rydberg-Cox, A. J. A prototype multilingual document browser for ancient Greek texts. *The New Review of Hypermedia and Multimedia*, 7(1), 2002, 103-113.