Integrating Methods and LRs for Automatic Keyword Extraction from Open Domain Texts

Panunzi Alessandro, Fabbri Marco, Moneglia Massimo

University of Florence, Italian Department
Piazza Savonarola 1, Florence, Italy
alessandro.panunzi@unifi.it, fabbri@lablita.dit.unifi.it, moneglia@unifi.it

Abstract

The paper presents a tool for keyword extraction from multilingual resources developed within the AXMEDIS project. In this tool lexical collocations (Sinclair, 1991) internal to documents are used to enhance the performance obtained through standard statistical procedure. A first set of mono-term keywords is extracted through the TF.IDF algorithm (Salton, 1989). The internal analysis of the document generates a second set of multi-term keywords based on the first set, rather than on multi-term frequency comparison with a general resource (Witten et al. 1999). Collocations in which a mono-term keyword occurs as the head are considered a multi-term keywords, and are assumed to increase the identification of the content. The evaluation compares the results of the TF.IDF procedure and the ones obtained with the enhanced procedure in terms of 'precision'. Each set of keywords received a value from the point of view of a possible user, regarding: (a) overall efficiency of the whole set of keywords for the identification of the content; (b) adequacy of each extracted keyword. Results show that multi-term keywords increase the content identification with a 100% relative factor and that the adequacy is enhanced in 33% of cases.

1. Introduction

1.1. Brief state of the art

Keyword detection procedures pointed out two main strategies, that have been merged in various ways according to practical needs: (a) the use of statistical-based technologies (exploiting both external-comparative analysis and internal one); (b) the use of linguistic-driven tools and databases.

The statistical-based procedures mostly rely on the comparison with a general corpus (Drouin, 2003). Salton (1989) suggested the TF.IDF algorithm to capture the "weight" of a word in a document comparing the internal frequency of a term with its distribution over a reference corpus. This kind of procedures can be defined as "external" or "comparative analysis". Other well-known proposed methods in statistical analysis of word occurrences in documents rely on the mere "internal analysis" of a single text (see, as example, Matsuo and Ishizuka 2004). In this perspective, various lexical measures have been developed, namely Mutual Information (MI), log-likelihood measure, χ^2 measure.

In general, linguistic-driven techniques start to run after a first statistical analysis. For example, in Van der Plas et al. (2004) there is a first stage of statistical analysis based on RFR algorithm, and a second stage of semantic analysis, based on lexical databases such as WordNet or EDR. Linguistic procedures are exploited also for collocation extraction in which the grammatical information provided by PoS-tagger is exploited (see Fung 1998).

1.2. Aims

The paper presents a tool for open domain keyword extraction, that has been developed within the AXMEDIS project (www.axmedis.org), for automatic indexing of textual information from multilingual multimedia resources (Panunzi, Fabbri, Moneglia 2005). In this approach, basic knowledge regarding the structure of lexical association in natural language performance is used to improve the performance of automatic extraction that has been obtained through standard statistical procedure.

The procedure exploit the mono-term keywords extracted by the standard statistic procedures. As far as those words represent a relevant argument of the content, they may be also the head of recurrent multi-words that specify qualities of this content. Therefore the procedure combines the statistic analysis of the document (for mono-term extraction) and the internal analysis of lexical associations. The procedure selects those keywords that are the head of recurrent multi-words (collocations), according to the following steps:

1) comparative analysis
   - mono-term keywords extraction
   - mono-term keywords semantic disambiguation
   - domain detection

2) internal analysis
   - multi-term keywords detection

The aim of this paper is to evaluate the difference of identification value obtained from a keyword extraction technique in which the lexical associations are taken into account.

Results of evaluation show that the complex keywords are considered more descriptive of the document content and more adequate to be selected as keyword than single words. Both content identification and keyword adequacy highly increase their value, according to the judgment of the evaluators.

2. Comparative analysis: mono-term keyword extraction

The first part of the procedure aims to extract mono-term keywords from the text. This process is performed through an integration of resources and standard algorithms, by means of comparison between the document and the referring universe, represented by a general corpus.

In the current version, the algorithm works only on English texts and the BNC has been used as Reference Corpus. The set of resource exploited for the tool is the following:

(a) language resources, mainly reference corpora and frequency lexica for the treated languages (BNC for English is already implemented, other reference corpora are in development and integration for Italian, French, German, Spanish), stop-word lists, rules for collocation and multiword identification;

(b) a multilingual PoS-tagger; the system integrates the TreeTagger, developed by Helmut Schmid within the TC project at the Institute for Computational Linguistics of the University of Stuttgart, and implemented for German, English, French and Italian;

(c) other existing semantic resources, namely WordNet and WordNetDomains, and frequency lexicons from English, Italian, French, German, Spanish general corpora.

The procedure foresees a statistic keyword extraction followed by a semantic processor that provides disambiguation of the extracted keywords given their assignment to a specific semantic domain.

2.1. Statistic keyword extraction

After tokenization, the input document is PoS tagged and nouns are extracted. All the following steps depend on the assumption that keywords have to be identified within the nominal lexicon.

The term frequency (TF) of all nouns in the document is compared to their distribution in a reference general corpus (inverse documents-frequency, IDF) through the standard TF.IDF algorithm. For a term *i* in a document *j*, the weight of the term in the document is:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

where $tf_{i,j}$ is the number of occurrences of *i* in *j*, df_i is the number of documents of the general corpus containing *i*, and *N* is the total number of documents in the corpus.

As they are considered more specific, words that are more frequent in the input document and less spread over the different documents of the corpus are the best candidates to represent the document itself, and they receive an higher score of "key-ness".

2.2. Semantic component and domain detection

The output of the previous step is a list of all nouns in the text, associated and ordered with respect to their TF.IDF value. Since they are potentially ambiguous with respect with their semantics, a word sense disambiguation (WSD) procedure is run over these words. In the WordNet (WN) database, a word can be associated with one or more synsets, that describe the possible meanings of a single lemma. Semantic Similarity (SS) among synsets related to the keyword candidates is then estimated, through the Lesk's distance-measure (Lesk 1986), on the WN lexical database. The SS is exploited to operate WSD: for each candidate, the synset with the highest SS score is preferred and associated to the keyword, generating a list of unambiguous lexical concepts suitable for translation.

WordNet Domains database is also exploited to determine the "area of discussion" to which each keyword belongs, so providing other keys for content identification.

WSD for translation and domain identification are feature of the object in the in the AXMEDIS multilingual framework. Results on these aspect will be not discussed in this paper.

3. Internal analysis: multi-term keyword detection

The internal analysis of the document relies on the previous results. The extracted keywords are further refined from the point of view of the accuracy, of the content identification, and of the value of the descriptors, referring to language properties of word association: high frequency collocations (Sinclair 1991) within the text are considered more definite and highly representative of its content. The underlying idea is that the more definite the keyword is, the more significant it will be for document identification.

Collocations in which a single selected keyword occurs as the head is considered a multi-term keywords, and therefore they increase the predictability of keywords for the identification of the content. This approach differs from others present in literature (see Witten et al. 1999), in which a statistical comparison between multi-terms in the document and the ones in a reference corpus is performed estimating TF.IDF value of phrases (instead of single terms). We think that this kind of measure is not suitable to identify lexical associations which may represent the document content, while it is useful for extraction of collocations in terms of linguistic analysis of preferred argument selection. Lexical association which constitutes keywords for a text are dependent on the document topic, i.e. on internal properties, and not on general language distribution properties.

The output list must consider both the multi-term and the mono-term keywords, in a unique list in order to produce a coherent list, key-ness scores of mono- and multi-term keywords must be balanced, using a normalizing value.

3.1. Grammatical driven selection of n-gram

In the procedure, the n-grams (bi- tri- and quadri-grams) of the terms in the document are produced, and then the relevant ones are selected through a linguistic filter that identifies only the possible multi-keyword configurations. The linguistic information provided through the PoS-tagging is further exploited to prevent non-grammatical n-grams (see Merkel and Andersson 2000). To be selected as potential multi-keyword, an n-gram must follow three conditions:

(1) the n-gram must contain a noun;

(2) the pattern has to be acceptable as multiword or collocation: a sequence "noun + preposition", for example, is a bi-gram that cannot represent itself a multi-keyword, while the sequences "noun + noun" or "adjective + noun" can;

(3) the n-gram must occur more than once in the document. This constraint is needed to avoid that hapax legomena multi-terms key-ness value obtains an overestimated score (see formula in the next paragraph and conclusions).

3.2. Estimating key-ness value for multi-term keyword

The estimation of the key-ness value of a multi-keyword relies both on TF.IDF score of the noun(s) contained in the multi-word and on the n-gram frequency parameters. The following figures show the formulas for the estimation of the Key-ness value (*K*) of a multi-term keyword. The basic key-ness value for a single word, *K(w)*, is defined as:

$$K(w) = \begin{cases} TF.IDF(w) & \text{if } w \text{ is a noun} \\ 0 & \text{otherwise} \end{cases}$$

A multi-term keyword is defined as an n-gram containing at least one noun; formally: $ng = [w_1 \ldots w_n]$, for $1<n<5$.

To estimate the key-ness value of an n-gram, *K(ng)*, three parameters are taken into account:

- the relative frequency of the multi-word (compared to the frequency of the single words which compose it);
- the *K* value, of each word within the n-gram;
- a normalizing value represented by the mean of TF.IDF values.

These parameters are related together following the formula:

$$K(ng) = \left(\sum_{i=1}^{n} \frac{C(ng)}{C(w_i)} K(w_i) \right) \overline{TF.IDF}$$

where *C(ng)* is the number of occurrences of the n-gram, *C(w)* is the number of occurrences of the noun(s) within the n-gram, and the index *i* varies on the words $[w_1 \ldots w_n]$ which compose the multi-word.

In the following paragraph will be discussed an example of keyword extraction

3.3. Example

Let's consider one example of keyword extracted from a news about a verdict over a trial against tobacco industry (1188 words, source: http://www.timesonline.co-uk). The results of TF.IDF, limited to the first 5 keywords, are shown in the following table:

	Mono-KW
racketeering	18,8546
tobacco	10,6816
industry	10,2369
forfeiture	8,5631
verdict	7,9298

Table 1. TF.IDF values for single keywords.

Although the news content is quite well identified by the set of keyword, someone can notice that the word "industry" is somehow vague, too much general. Moreover, it can be associated to other selected keyword (as in the metaphoric expression "racketeering industry").

In the list obtained by the multi-keyword extractor, shown in table 2, lexical associations reduce the vagueness and potential ambiguity of the extracted keywords:

	Multi-KW
forfeiture__of__profit	25,0803
appeal__court	24,3185
tobacco__industry	20,3713
racketeering	18,8546
cigarette__maker	15,1025

Table 2. Key-ness scores with multi-term keywords.

Once the collocations are considered, on one side new multiword keyword come around with an high score ("appeal court"; "cigarette makers") and on the other keywords become selective through lexical association ("tobacco industry").

4. Evaluation and conclusions

The tool generates keywords from whatever document in plain text. To the end of this paper an evaluation have been performed on a test corpus of English texts, which are representative of the open-domain environment of news.

4.1. Evaluation strategy

The evaluation compares the results of the TF.IDF standard procedure and the ones of the enhanced procedure in terms of (a) content identification value of the whole set of keywords, (b) adequacy of each extracted keyword. These measure are both related to the 'precision' of the extracted keywords.

Recall is not estimated, since the keyword identification is not a "strict" retrieval task, for two independent reasons.

1) The set of "all the keywords" of a text is undefined, and maybe cannot be uniquely defined. While performing a keyword extraction on a text, the task is not to identify "all the words" which are needed to define a document, but to identify a set of words that are as most representative as possible.

2) Humans and machines do not follow similar "strategies" for keyword identification of a document. Automatic keyword extraction on a text tries to identify the most relevant words which occur in the document, while humans are not dependent on the text in identifying the keywords of a document. For example, on a text regarding the life of zebras, elephants and lions, a human can identify "savannah animals" as the main keyword, while this particular word pattern could never occur in the text. While machines work on frequencies within a text (or compared to a general resource), humans work on inferences.

The evaluation tries to estimate the meaning of the two set of keyword from the point of view of a potential user.

Two mother tongue external evaluator with a high level of culture have been asked to read each news in the test corpus and to judge the adequacy of the two keyword sets from two different perspectives.

The first evaluator judged to which extent each of the two sets (as a whole) identifies the content. In other words, this evaluation tests whether the keyword set predict the nature of the content. Four degrees have been considered:

A = very good
B = sufficient to good
C = insufficient
D = bad

The second evaluator judged whether each keyword in the two set is adequate or not to represent the content; i.e. if the keyword is a possible expression to be used for searching the content. With respect to the degree of representativeness of the document content, a keyword can be judged as:

- adequate (score 1)
- inadequate (score-1)
- vague (score 0)

Results are shown in the following paragraph.

4.2. Results

The test corpus for the evaluation is constituted by 24 texts of news (from the online version of The Times and New York Times), collected to be representative of the open-domain environment of newspapers (world affairs, business, technology, science, health, education). The judgment given by the evaluators on the different aspects (level of content identification provided by the keyword set and adequacy of the single keywords) are considered separately.

The results show that the overall prediction of the content by the multi-term keyword set is highly increased. In the following table we merged the good results (A-B) and the bad ones (C-D). The percentage of good results using the multi-term keyword set is increased by a relative factor of 100% with respect to the mono-term keyword one, which is very unsatisfactory (from 37,5% to 75%):

judgment on keyword-set	mono-term keywords	multi-term keywords
A-B	37,5%	75,0%
C-D	62,5%	25,0%

Table 3. Evaluation of keyword set identification degree.

Although the value of mono-term keywords have been considered less unsatisfactory by the evaluator, the increase in performance of multi-term keyword extractor reaches a relative factor of 33% (in term of percentage of full adequate keywords)

judgment on single keywords	mono-term keywords	multi-term keywords	relative increase
adequate	48,8%	65,0%	+33,3%
vague	32,5%	27,5%	-15,4%
non adequate	18,8%	7,5%	-60,0%

Table 4. Evaluation of keyword adequacy.

4.3. Conclusion and further steps

Results show that the approach is capable of retrieving multi-terms keywords which have high descriptiveness of the document content. Mixing the classical TF.IDF approach with internal analysis leads to an significant improvement of the adequacy of extracted keywords. These performances depend a lot on linguistic tools that are exploited, in particular on the PoS-tagger precision and on the reference value of the general corpora involved.

To improve the current performances it is planned to:

(1) modify the TF.IDF algorithm taking into account the dispersion of words in each document in the reference corpus;

(2) use log-likelihood measure in order to deal with the hapax legomena multi-terms keyword problem;

(3) exploit the semantic component to increase the key-ness value of words related to the extracted topic domain (this step heavily depends on the goodness of semantic database).

5. References

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1), Benjamins, pp. 99-115.

Fung, P. (1998). Extracting key terms from Chinese and Japanese texts. *The International Journal on Computer Processing of Oriental Language (IJCPOL).* 12(1), WSPC, pp. 99-121.

Lesk, M. (1986). Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream. In *Proceedings of the SIGDOC Conference*. New York, NY: ACM, pp. 24-26.

Matsuo, Y., Ishizuka, M. (2004) Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, *International Journal on Artificial Intelligence Tools (IJAIT)*, 13(1), WSPC, pp. 157-169.

Merkel, M., Andersson, M. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of RIAO 2000 Conference User-Oriented Content-Based Text and Image Handling*, Paris, France: CID-CASIS, pp. 737-746.

Panunzi, A., Fabbri, M., Moneglia, M. (2005). Keyword Extraction in Open-Domain Multilingual Textual Resources. In *Proceedings of 1st International Conference on Automated Production of Cross media Content for Multi-channel Distribution*. Los Alamitos, CA: IEEE Press, pp. 253-256.

Salton, G. (1989). *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Reading, MA: Addison Wesley.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Van der Plas, L. Pallotta, V., Rajman, M., Ghorbel, H. (2004) Automatic Keyword Extraction from Spoken Text. In *Proceeding of LREC 2004*. Paris, France: ELRA, pp. 2205-2208.

Witten, I., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C. (1999) KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*. Berkeley, CA pp. 254-255.

BNC http://www.natcorp.ox.ac.uk/

TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

WordNet http://wordnet.princeton.edu/w3wn.html

WordNetDomains http://wndomains.itc.it/wordnetdomains.html