Multilingual Open Domain Key-word Extractor Proto-type

Alessandro Panunzi
Marco Fabbri
Massimo Moneglia
University of Florence

Automatic Keyword extraction is now a mature language technology. It enables the annotation of large amount of documents for content-gathering, indexing, searching and for its identification, in general. The reliability of results when processing documents in a multilingual environment, however, is still a challenge, particularly when documents are not limited to one specific semantic domain. The use of multi-term descriptors seems to be a good mean to identify the content. According to our previous evaluations (Panunzi et al. 2006a, 2006b), the availability of multi-term keywords increases the performance with respect to mono-term keywords of 100% relative factor. The LABLITA tool presented in this demo works now in a multilingual environment, as well. The demo calculates on the fly the number of mono-term and multiword keywords of parallel documents in English, Italian, German, French and Spanish, and will allow the audience to judge: a) the enhancement bared by multiword keywords for the identification of content; and b) the comparability of performance obtained by the tool processing different languages.

1. The demonstration

Automatic Keyword extraction is now a mature language technology. It enables the annotation of large amount of documents for content gathering, indexing, searching and more in general for its description. Since the multi-term identification constitutes a central issue in such a task, keyword extraction can be also used for lexicographical aims. In particular, the automatic extraction of informational-relevant complex lexical-units can be useful in highlighting those multiwords which are the most relevant to understand the content of text documents, i.e. which are relevant in the actual language usage, production and understanding.

However the reliability of results when processing documents in a multilingual environment is still a challenge, in special when documents are not limited to one specific semantic domain. The use of multi-term descriptors seems a good mean to identify the content. According with our previous evaluations (Panunzi et al. 2006a, 2006b) the availability of multi-term keywords increases the performance with respect to mono-term keyword of 100% relative factor. The LABLITA tool presented in demo version is now implemented for working in a multilingual environment. The demo is able to extract on the fly mono-term and multi-term keywords from documents written in English, Italian, German, French and Spanish.

The strategy implemented in this tool is: 1) to identify Keywords within the nominal lexicon; 2) to retain single terms from the statistic comparison against the reference corpus (the relevant key-concepts); 3) to use general statistics derived from corpora that are roughly comparable to the BNC; 3) to estimate lexical associations within the analyzed document to grasp the relevant specifications of key-concepts. This paper will briefly sketch the algorithms implemented in this tool and will focus on the evaluation of the results at two levels: a) the enhancement brought by using multiword keyword for the identification of the content; b) the comparability of performance obtained by the tool processing different languages.

2. Comparative analysis and mono-term keyword extraction

The procedure extracts first a set of mono-term keywords, by means of comparison between the word frequency in the document and the referring universe, represented by a general corpus

(BNC for English and comparable Italian, French, German, Spanish corpora, that have been designed to approximate the BNC standard).

PoS Tagging. Keywords are crucially identified within the nominal lexicon. The identification of nominal lexicon is performed by the PoS tagger (TreeTagger). After tokenization, the input document is PoS tagged and nouns are extracted.

TFIDF. Key-Nouns are extracted using a revised version of the standard TF.IDF algorithm (Salton 1989). The term frequency (TF) of all nouns in the input document is compared by the tool to their distribution in the general corpus (inverse documents-frequency, IDF). For a word *w* in a given document, the weight of the term in the document is represented by the formula in Figure 1:

$$TF.IDF(w) = TF(w) \log\left(\frac{N}{DF(w)}\right) \qquad (1)$$

Figure 1. TF.IDF algorithm

where TF(*w*) is the number of occurrences of *w* in the input document, DF(*w*) is the number of documents of the general corpus containing w, and N is the number of documents in the corpus.

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.

The following table presents the four higher ranked mono term key-words extracted from *quasi*-parallel texts about the Great Depression of 1929 (taken from Wikipedia).

EN	IT	DE	SP	FR
economy	crisi	Krise	depresión	marché
bank	borsa	Weltwirtschaftskrise	economía	crise
government	uomini	Deflationspolitik	guerra	PIB
depression	guerra	Bearbeiten	trabajador	krach

Table 1. Mono term key-words generated from multilingual Wikipedia texts about the Great Depression of 1929

3. Analysis of lexical associations in the input document

Multi-term keywords identification relies on the result of the first procedure. The second procedure exploits lexical collocations of the extracted keywords and then combines the statistic analysis of the document (for mono-term extraction) and the internal analysis of lexical associations (for multi-term extraction).

Strategy. Lexical associations are measured only considering the analyzed document, with no reference to the general resource (Matsuo et al. 2004). This approach differs from others in literature (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).

Indeed, lexical associations which constitute keywords for a text are dependent on internal properties of the document, and not on the distribution of the association itself in the reference corpus. The tool retains single terms from the statistic comparison against the reference corpus (as relevant key-concepts), and then estimates their associations within the analyzed document (as relevant specifications of the concepts).

Procedure. In the procedure, the n-grams (2-3- and 4-grams) of all names in the document are produced, and the relevant ones are selected through a linguistic filter that identifies only the possible multiword configurations. The linguistic information provided through the PoS-tagging is further exploited to prevent non-grammatical n-grams (Merkel et al. 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.

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 basic key-ness value for a single word, K(w), is defined as below (Figure 2):

$$K(w) = \begin{cases} TF.IDF & \text{if } w \text{ is a noun} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Figure 2. Key-ness value for single words *(w)*

A multi-term keyword is defined as an n-gram containing at least one noun. To estimate the key-ness value of an n-gram, K(*ng*), three parameters are taken into account: 1) the relative frequency of the multi-word; (compared to the frequency of the single words which compose it); 2) the K value, of each noun within the n-gram; 3) a normalizing value represented by the mean of TF.IDF values. These parameters are related together in the following formula (Figure 3), where C(*ng*) is the number of occurrences of the n-gram, $C(w_i)$ 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:

$$K(ng) = \left(\sum_{i=1}^{n} \frac{C(ng)}{C(w_i)} K(w_i) \right) \overline{TF.IDF} \qquad (3)$$

Figure 3. Key-ness value for n-grams *(ng)*

The algorithm has been implemented in two versions:

1) in C++ for processing multimedia contents within the AXMEDIS framework;
2) in Java (freely distributed for research purpose from the LABLITA web site).

4. The evaluation

For the evaluation typical web contents have been used. A set of parallel Wikipedia articles have been downloaded in the five languages. Articles have been selected considering both the ontological category and the semantic domain of the argument and then edited in order to process texts of similar length. These arguments range over different domains and semantic fields and also vary from the more specific to the more general for what regards their ontological level. The following is the list of arguments, followed by their domain and ontological level.

ARGUMENT	*Great Depression of 1929*	*The atomic bombing of Hiroshima and Nagasaki*	*Nouvelle vague*	*Savana*	*Pasta*	*Mammals*
DOMAIN	History	History	Art	Natural Environment	Artefacts	Natural Category
SEMANTIC FIELD	Economy	War	Cinema	Place	Food	Animals
ONTOLOGICAL LEVEL	Specific Fact	Specific Event	Specific Concept	Basic level concept	Basic level concept	Higher level concept

Table 2. Wikipedia Articles: Argument, Domain, Semantic field and Ontological level of the concept

The following table presents the five higher ranked mono and multi- term key-words extracted from the parallel texts regarding the Great Depression of 1929.

EN	IT	DE	SP	FR
money supply	crisi	Krise	gran depresión	marché
bank failure	economia americana	Weltwirtschaftskrise	economía	crise
stock market crash	bene di consumo	Deflationspolitik	año depresión	évolution du pib
gold standard	crollo della borsa	Bearbeiten	problema de solvencia	marché boursier
stock market	causa della recessione	Bank	guerra mundial	taux de chômage

Table 3. Mono term key-words generated from multilingual Wikipedia texts about the Great Depression of 1929

Comparing this set with the list of mono-term keywords it turns out that the predictive value is strongly enhanced and the result appears sufficiently predictive in all implemented languages. According with our previous evaluations (Panunzi et al. 2006a, 2006b) the availability of multi-term keywords increases the performance with respect to mono-term keyword of 100% relative factor. It has to be noticed that a good selection of multi-terms which have a highly descriptive value for many different documents could be a useful basis for lexicographical researches on complex lexical-units.

The evaluation must take into account the peculiar nature of keyword extraction task. This task is not achieved selecting "the set of words" which uniquely define a given document, but rather those words that are the most representative in accordance with the interest and background knowledge of the user. Humans and Machines do not follow the same strategies. While machines work on frequencies of words in a text, humans can work on inferences. 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. Therefore the set of "all keywords" of a text is under-defined, and for this reason Recall cannot be estimated.

The evaluation estimates the extracted keywords from the point of view of a potential user. For each language two groups of six evaluators have been selected (results reported do not consider French). Each group tested the keywords from opposite perspectives: a) their adequacy to represent the argument of a document that is known by the evaluator; b) their efficiency for predicting the topic of an unknown text.

a) *Adequacy*. The first group have been asked to read the Wikipedia articles and to evaluate the keywords in two tasks:
    i. KEYWORDS KEY-NESS. The task is to judge whether each keyword is adequate or not to represent the content. In this respect, a keyword can be judged:
        - Adequate
        - Inadequate
        - Vague
    ii. A-POSTERIORI PREDICTIVE ADEQUACY . The task is to judge whether the keyword set sufficiently identify the content of the document. Four degrees have been considered:

        A = Very good

        B = Sufficient to good

        C = Insufficient

        D = Bad

b) *Predictivity*. The evaluators of the second group do not know the text and do not assess the keywords. After reading a keyword set, they try to figure out the possible argument of the text and specify it through a definition. Such definition has been mapped onto the following values:

- Right
- Too generic (chosen argument represents a hypernym of the right one)
- Too specific (chosen argument represents a hyponym of the right one)
- Wrong

5. Results

The results of the evaluation must consider two factors that affect the performance of the tool in different languages: 1) documents are *quasi* parallel; i.e. although they strictly share the function and the argument, as it is frequently the case on the web, they are not each other's translations; 2) although the reference corpora of each language have a high level of representativeness, they cannot be parallel by definition as they are not derived from parallel resources. Therefore, no strict equivalence among the keyword extracted from quasi-parallel in different languages can be expected. More specifically, BNC is the source of statistics for English. The other reference corpora approximate this standard at different levels. Two 100 millions tokens sampling of comparable web corpora have been used for German and Italian (WaCky Web Corpora), and two smaller 25 millions tokens newswire corpora have been adopted for Spanish and French (respectively Compiled at LLI-UAM and at LABLITA).

However, while the second of these factors is merely due to the lack of available balanced corpora for the above mentioned languages, the first one represent the actual status of the web environment, and allows us to properly judge how the tool runs on real data.

Single keywords have been considered adequate for what regard their key-ness when accepted by 50% of the evaluators or more. Table 5 shows that, from the point of view of the overall comparability among languages, the number of adequate keyword turns out similar. In all languages, good results (four on five) have been achieved in special with basic level concepts, while very specific topics (nouvelle vague) get few adequate keywords and at least two keywords derived from each text have been considered adequate for all languages. The performance of the tool are not affected neither by the domain nor by the semantic field of the topic, in accordance with the open-domain requirement. Therefore the efficiency of the extracted keywords for gathering is promising.

	EN	IT	DE	ES
great_depression	3	3	3	3
a-bombing	4	2	3	2
nouvelle_vague	2	2	2	2
savannah	2	4	4	2
pasta	4	4	4	3
mammals	2	2	3	2

Table 5. Number of adequate key-word
(key-ness)

	EN	IT	DE	ES
great_depression	67%	100%	100%	100%
a-bombing	100%	83%	67%	50%
nouvelle_vague	33%	50%	17%	17%
savannah	100%	100%	100%	33%
pasta	100%	83%	100%	67%
mammals	100%	100%	100%	83%

Table 6. Percentage of A-B scores
(a posteriory predictive adequacy)

	EN	IT	DE	ES
pasta	33%	50%	83%	50%
savannah	17%	50%	67%	0%
mammals	67%	67%	50%	67%
great_depression	33%	83%	83%	50%
a-bombing	33%	33%	50%	33%
nouvelle_vague	0%	17%	0%	0%

Table 7. Percentage of "right"
(predictivity value)

	EN	IT	DE	ES
pasta	83%	100%	100%	83%
savannah	50%	100%	100%	33%
mammals	100%	67%	100%	100%
great_depression	66%	100%	100%	100%
a-bombing	100%	83%	83%	66%
nouvel_vague	50%	67%	33%	50%

Table 8. Percentage of "right" + "hypernym/hyponym"
(predictivity value)

For what regard the predictive value of the keyword sets, data present a better picture in all languages. Table 6 presents the percentage of the evaluations in which they have been considered adequate. Although the results for Spanish are slightly less satisfactory, all topics turn out well identified by the set. Only the very specific argument "Nouvelle vague" is unidentified (cross-linguistically). However, this result is strongly dependent on the evaluator's awareness of the text. When the opposite perspective is taken (Table 7), results are less satisfactory and the positive match of the keyword set seems strongly dependent on the attitude of the evaluators (a quite larger number of evaluators seems needed for leveling this aspect).

Italians and Germans have been more successful in targeting the keyword set. However a more detailed analysis shows that the keyword set is, in most cases, at least sufficient to approximate the topic. When hyperonym and hyponym are considered (Table 8), results mirror those in Table 6. Therefore, the overall efficiency of the automatic keyword-extraction for metadata assignment turns out effective.

References

AXMEDIS [online]. *http://www.axmedis.org* [Access date: 20 Feb. 2008].

BNC [online]. *http://www.natcorp.ox.ac.uk* [Access date: 20 Feb. 2008].

LABLITA Keyword extractor (online). *http://lablita.dit.unifi.it/projects/ke/ca/en/document_view* [Access date: 20 Feb. 2008].

Matsuo, Y.; Ishizuka, M. (2004). "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information". *International Journal on Artificial Intelligence Tools* 13 (1). 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: CID-CASIS. 737-746.

Panunzi, A.; Fabbri, M.; Moneglia, M. (2006a). "Integrating methods and LRs for automatic keywords extraction from open-domain texts". In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Paris: ELRA. CD-rom.

Panunzi, A.; Fabbri, M.; Moneglia, M. (2006b). "Multi-Term Keywords for Indexing Multilingual Textual Repositories: Developing Language Resources and Algorithms". *Proceedings of AXMEDIS 2006 Conference*. Los Alamitos: IEEE Computer Society Press. 173-180.

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

TreeTagger [online]. *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger* [Access date: 20 Feb. 2008].

WaCky Project [online]. *http://wacky.sslmit.unibo.it/doku.php* [Access date: 20 Feb. 2008].

Wikipedia [online]. *http://wikipedia.org/* [Access date: 20 Feb. 2008].

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. 254-255.