Actes du huitième défi fouille de texte, DEFT2012, Grenoble, France, 8 juin 2012. Pages 19-28.
Proceedings of the Eight DEFT Workshop, DEFT2012, Grenoble, France, 8th July 2012. Pages 19-28.

Key-concept extraction from French articles with KX

Sara Tonelli[1] Elena Cabrio[2] Emanuele Pianta[1]
(1) FBK, via Sommarive 18, Povo (Trento), Italy
(2) INRIA, 2004 Route des Lucioles BP93, Sophia Antipolis cedex, France
satonelli@fbk.eu, elena.cabrio@inria.fr, pianta@fbk.eu

Résumé

Nous présentons une adaptation du système KX qui accomplit l'extraction non supervisée et multilingue des mots-clés, pour l'atelier d'évaluation francophone en fouille de textes (DEFT 2012). KX sélectionne une liste de mots-clés (avec leur poids) dans un document, en combinant des annotations linguistiques de base avec des mesures statistiques. Pour l'adapter à la langue française, un analyseur morphologique pour le Français a été ajouté au système pour dériver les patrons lexicaux. De plus, des paramètres comme les seuils de fréquence pour l'extraction de collocations, et les index de relevance des concepts-clés ont été calculés et fixés sur le corpus d'apprentissage. En concernant les pistes de DEFT 2012, KX a obtenu de bons résultats (Piste 1 - avec terminologie : 0.27 F1 ; Piste 2 : 0.19 F1) en demandant un effort réduit pour l'adaptation du domaine et du langage.

Abstract

We present an adaptation for the French text mining challenge (DEFT 2012) of the KX system for multilingual unsupervised key-concept extraction. KX carries out the selection of a list of weighted keywords from a document by combining basic linguistic annotations with simple statistical measures. In order to adapt it to the French language, a French morphological analyzer (PoS-Tagger) has been added into the extraction pipeline, to derive lexical patterns. Moreover, parameters such as frequency thresholds for collocation extraction and indicators for key-concepts relevance have been calculated and set on the training documents. In the DEFT 2012 tasks, KX achieved good results (i.e. 0.27 F1 for Task 1 - with terminological list, and 0.19 F1 for Task 2) with a limited additional effort for domain and language adaptation.

Mots-clés : Extraction de mots-clés, patrons linguistiques, terminologie.

Keywords: Key-concept extraction, linguistic patterns, terminology.

1 Introduction

Key-concepts are simple words or phrases that provide an approximate but useful characterization of the content of a document, and offer a good basis for applying content-based similarity functions. In general, key-concepts can be used in a number of interesting ways both for human and automatic processing. For instance, a quick topic search can be carried out over a number of documents indexed according to their key-concepts, which is more precise and efficient than full-text search. Also, key-concepts can be used to calculate semantic similarity between documents and to cluster the texts according to such similarity (Ricca *et al.*, 2004). Furthermore, key-concepts provide a sort of quick summary of a document, thus they can be used as an intermediate step in *extractive* summarization to identify the text segments reflecting the content of a document. (Jones *et al.*, 2002), for example, exploit key-concepts to rank the sentences in a document by relevance, counting the number of key-concept stems occurring in each sentence. In the light of the increasing importance of key-concepts in several applications, from search engines to digital libraries, a recent task for the evaluation of key-concept extraction was also proposed at SemEval-2010 campaign (Kim *et al.*, 2010)

In this work, we present an adaptation of the KX system for multilingual key-concept extraction (Pianta et Tonelli, 2010) for the French text mining challenge (DEFT 2012) task. A preliminary version of KX for French took part in the DEFT 2011 campaign on "Abstract – article matching" (Tonelli et Pianta, 2011), and achieved good performances in both tracks (0.990 and 0.964 F1 respectively).

Compared to the previous version of KX, we have now integrated into the extraction pipeline a French morphological analyzer (Chrupala *et al.*, 2008). This allows us to exploit morphological information while selecting candidate key-concepts, while in the version used at DEFT 2011 the selection was made using regular expressions and black lists.

The paper is structured as follows : in Section 2 we detail the architecture of KX (i.e. our key-concepts extraction tool), providing an insight into its parameters configuration. In Section 3 we present the setting defined and adopted for the DEFT 2012 task, while in Section 4 we report the system performances on the training and on the test sets. Finally, we draw some conclusions, and discuss future improvements of our approach in Section 5.

2 Key-concept extraction with KX

This section describes in details the basic KX architecture for unsupervised key-concept extraction. KX can handle texts in several languages (i.e. English, Italian, French, Finnish and Swedish), and it is distributed with the TextPro NLP Suite[1] (Pianta *et al.*, 2008). KX architecture is the same across all languages, except for the module selecting multiword expressions, that is based on PoS tags (this is the only language-dependent part of the system). In order to perform this selection, a morphological analyzer/PoS tagger has been integrated for each of the five languages, and some selection rules have been manually defined. More details on the French rules are reported in Section 2.2 and in Section 3.

2.1 Pre-processing of the reference corpus

If a domain corpus is available, the extraction of key-concepts from a single document can be preceded by a pre-processing step, during which key-concepts are extracted from the corpus and their inverse document frequency (IDF) at corpus level is computed by applying the standard formula :

[1] `http://textpro.fbk.eu/`

$$IDF_k = log \frac{N}{DF_k}$$

where N is the number of documents in the corpus, and DF_k is the number of documents in the corpus that contain the key-concepts k. The IDF of a rare term tends to be high, while the IDF of a frequent one is likely to be low. Therefore, IDF may be a good indicator for distinguishing between common, generic words and specific ones, which are good candidates for being a key-concept. For DEFT 2012, we have used as a *reference corpus* all the documents contained in the training and in the test sets (468 documents in total).

2.2 Key-concept extraction

Figure 1 shows KX work-flow for the key-concept extraction process : starting from a document, a list of key-concepts ranked by relevance is provided as the output of the system. The same work-flow applies both to *i)* the extraction of key-concepts from a single document, and to *ii)* the extraction of different statistics including IDF from a *reference corpus*, which can be optionally used as additional information when processing a single document. For more information, see above and Section 2.3.

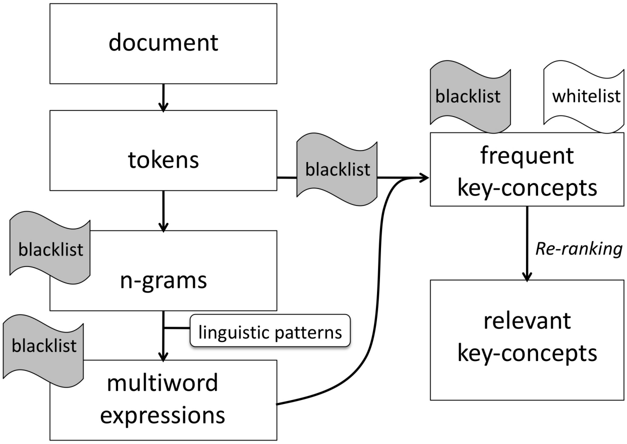

FIG. 1 – Key-concept extraction workflow with KX

As a first step, the system takes a document in input and tokenize the text. Then, all possible n-grams composed by any token sequence are extracted, for instance ‘Éclipse de soleil’, ‘tous les’, ‘ou chacun’. The user can set the max length of the selected n-grams : for DEFT 2012 we set such length to six.

Then, from the n-gram list a sublist of *multiword expressions (MWE)* is derived, i.e. combinations of words expressing a unitary concept, for example ‘procès de travail’ or ‘économie politique’.

In the selection step, the user can choose to rely only on local (document) evidence or to make use also of global (corpus) evidence. As for the first case, a frequency threshold called $MinDoc$ can be set, which corresponds to the minimum number of occurrences of n-grams in the current document. If a reference corpus is also available, another threshold can be added, $MinCorpus$, which corresponds to the minimum number of occurrences of an n-gram in the corpus. KX marks an n-gram in a document as a multiword term if it occurs at least $MinCorpus$ times in the corpus or at least $MinDoc$ times in the document. The two parameters depend on the size of the reference corpus and the document respectively. In our case, the corpus was the set of documents used in the training and in the test set (see Section 2.1).

A similar, frequency-based, strategy is used to solve ambiguities in how sequences of contiguous multiwords should be segmented. For instance, given the sequence 'retour des bonnes manières' we need to decide whether we recognize 'retour des bonnes' or 'bonnes manières'. To this purpose, the strength of each alternative MWE is calculated as follows, and then the stronger one is selected.

$$Strength_{colloc} = docFrequency * corpusFrequency$$

In the next step, the single words and the MWEs are ranked by frequency to obtain a first list of key-concepts. Thus, frequency is the baseline ranking parameter, based on the assumption that important concepts are mentioned more frequently than less important ones. Frequency is normalized by dividing the number of key-concept occurrences by the total number of tokens in the current document.

As shown in Figure 1, the first key-concepts list is obtained by applying *black and white lists* almost at every step of the process. A black list is applied to discard n-grams containing one of the language-specific stopwords defined by the user, for example 'avons', 'peut', 'puis', 'parce'. Also single words corresponding to stopwords are discarded when the most frequent tokens are included into the first key-concept list. For example, in French we may want to exclude all key-concepts containing the words 'toi', 'très', 'finalement', etc.

When deriving multiword expressions (MWEs) from the n-gram list, KX applies another selection strategy. This selection is crucial because only MWEs are selected as candidate key-concepts, since they correspond to combinations of words expressing a unitary concept, for example 'régime de despotisme familial' or 'reproduction matérielle'. The n-grams are analyzed with the Morfette morphological analyzer (Chrupala *et al.*, 2008) in order to select as multiword expressions only the n-grams that match certain lexical patterns (i.e. part-of-speech). This is the so-called linguistic filter. For example, one of the patterns admitted for 3-grams is the following :

$$[SP] - [O] - [SP]$$

This means that a 3-gram is a candidate multiword term if it is composed by a single or plural noun (S and P respectively), followed by a preposition (defined as O), followed by another noun. This is matched for example by the 3-gram 'procès [S] de [O] travail [S]'.

Finally, black and white lists can be manually compiled also for key-concepts, to define expressions that should never be selected as relevant key-concepts, as well as terms that should always be included in the key-concept rank. For example, the preposition 'de' is very frequent in documents, so it can happen that it is selected as single-word key-concept. In order to avoid this, 'de' can be included in the key-concept black list.

2.3 First key-concept ranking

Different techniques are used to re-rank the frequency-based list of key-concepts obtained in the previous step according to their relevance. If a reference corpus is available, as in our case, additional information can be used to understand which key-concepts are more specific to a document, and therefore are more likely to be relevant for such document.

In order to find the best ranking mechanism, and to tailor it to the type of key-concepts we want to extract, the following parameters can be set :

Key-concept IDF : This parameter takes into account the fact that, given a data collection, a concept that is mentioned in many documents is less relevant to our task than a concept occurring in few documents. To activate it, a reference corpus must undergo a pre-processing step in which the key-concepts are extracted from each document in the corpus, and the corresponding inverse document frequency (IDF) is computed, as described in Section 2.1.

When this parameter is activated, for each key-concept found in the current document, its IDF computed over the reference corpus is retrieved and multiplied by the key-concept frequency at document level.

Key-concept length : Number of tokens in a key-concept. Concepts expressed by longer phrases are expected to be more specific, and thus more informative. When this parameter is activated, the frequency is multiplied by the key-concept length. For example, if ‘expression verbale’ has frequency 6 and ‘expression verbale des èmotions’ has frequency 5, the activation of the key-concept length parameter gives ‘expression verbale’ = 6 * 2 = 12 and ‘expression verbale des émotions’ = 5 * 4 = 20. In this way, the 4-gram is assigned a higher ranking than the 2-gram.

Position of first occurrence : Important concepts are expected to be mentioned before less relevant ones. If the parameter is activated, the frequency score will be multiplied by the $PosFact$ factor computed as :

$$PosFact = \left(\frac{DistFromEnd}{MaxIndex}\right)^2$$

where $MaxIndex$ is the length of the current document, and $DistFromEnd$ is $MaxIndex$ minus the position of the first key-concept occurrence in the text.

A configuration file allows the user to independently activate such parameters. The key-concept relevance is then calculated by multiplying the normalized frequency of a key-concept by the score obtained by each active parameter. We eventually obtain a ranking of key-concepts ordered by relevance. The user can also set the number of top ranked key-concepts to consider as best candidates.

2.4 Final key-concept ranking

Section 2.3 described the first set of ranking strategies, that can be optionally followed by another set of operations to adjust the preliminary ranking. Again, such operations can be independently activated through a separate configuration file. The parameters have been introduced to deal with the so-called *nested* key-concepts (Frantzi *et al.*, 2000), i.e. those that appear within other longer candidate key-concepts. After the first ranking, which is still influenced by the key-concept frequency, *nested* (shorter) key-concepts tend to have a higher ranking than the containing (longer) ones, because the former are usually more frequent than the latter. However, in some settings, for example in scientific articles, longer key-concepts are generally preferred over shorter ones because they are more informative and specific. In such cases, the user may want to adjust the ranking in order to give preference to longer key-concepts and to reduce or set to zero the score of nested key-concepts. These operations are allowed by activating the following parameters :

Shorter concept subsumption : It happens that two concepts can occur in the key-concept list, such that one is a specification of the other. Concept *subsumption* and *boosting* (see below) are used to merge or rerank such couples of concepts. If a key-concept is (stringwise) included in a longer key-concept with a higher frequency-based score, the score of the shorter key-concept is transferred to the count of the longer one. For example, if ‘expression verbale’ has frequency 4 and ‘expression verbale des èmotions’ has frequency 6, by activating this parameter the relevance of ‘expression verbale des èmotions’ is 6 + 4 = 10, while the relevance of ‘expression verbale’ is set to zero. The idea behind this strategy is that nested key-concepts can be deleted from the final key-concept list without losing relevant information, since their meaning is nevertheless contained in the longer key-concepts.

Longer concept boosting : This parameter applies in case a key-concept is (stringwise) included in a longer key-concept with a lower relevance. Its activation should better balance the ranking in order to take into account that longer n-grams are generally less frequent, but not less relevant, than shorter ones. The parameter is available in two different versions, having different criteria for computing such boosting. With the *first option*, the average score between the two key-concepts relevance is computed. Such score is assigned to the less frequent key-concepts and subtracted from the frequency score of the higher ranked one. With the *second option*, the longer key-concepts is assigned the frequency of the shorter one. In none of the two variants key-concepts are deleted from the relevance list, as it happens by activating the *Shorter concept subsumption* parameter.

For example, if ‘expression verbale’ has score 6 and ‘expression verbale des émotions’ has score 4, by activating the first option of this parameter the relevance of ‘expression verbale’ becomes 6 - ((6 + 4) / 2) = 1, while the relevance of ‘expression verbale des émotions’ is set to 5, i.e. (6 + 4) / 2 .

With the second option, both the relevance of ‘expression verbale des émotions’ and of ‘expression verbale’ is set to 6.

The examples above show that these parameters set by the user can change the output of the ranking by deleting some entries and boosting some others. After applying one cycle of subsumption/boosting, the order of the concepts can dramatically change, producing the conditions for further subsumption/boosting of concepts. The user can set the number of iterations for the application of this re-ranking mechanism, and each cycle increases the impact of the re-ranking on the key-concept list. The parameters can be activated together and in different combinations. If all parameters are set, the short concept subsumption procedure is applied first, then the longer concept boosting is run on the output of the first re-ranking, so that the initial relevance-based list goes through two reordering steps.

3 KX configuration for the DEFT 2012 task

As introduced before (Section 2.2), to port KX to the French language and, in particular, to adapt it to the DEFT 2012 task, the Morfette morphological analyzer (Chrupala *et al.*, 2008) has been integrated into the system, to select as multiword expressions only the n-grams matching certain lexical patterns (i.e. part-of-speech). Such lexical patterns are learned on the gold standard, and manually formalized and added into the system as a linguistic filter. In order to speed up this process, we took advantage of the set of lexical patterns defined for Italian, and we checked if they could be applied also for French. Moreover, new patterns were added in compliance with DEFT training data requirements. For example, the following n-grams have been added as allowed patterns (i.e. candidate multiword terms) :

- 6-grams : [SP]-[O]-[SP]-[O]-[S]-[JK], where S and P correspond to singular or plural nouns, O to the prepositions (also in combination with the article), and J and K to singular or plural adjectives (e.g. 'soulèvement [S] des [O] Métis [S] dans l' [O] Ouest [S] canadien [J]') ;
- 5-grams : [SP]-[O]-[SP]-[O]-[P], (e.g. 'gestion [S] des [O] troupeaux [P] de [O] rennes [P]') ;
- 4-grams : [SP]-[JK]-[0]-[S], (e.g. 'histoire [S] canonique [J] de la [0] traduction [S]') ;
- 3-grams : [S]-[SP]-[JK], (e.g. 'français [S] langue [S] première [J]').

We compiled black lists both for common French stopwords (containing e.g. articles, prepositions, a few numbers, and functional verbs) and stopphrases (prepositional structures such as 'au sujet de', 'en dehors de', 'en face de'), since we do not want them to be selected as key-concepts.

As for the IDF value mentioned in Section 2.1, it has been computed for 86,419 key-concepts extracted from DEFT 2012 training and test set. Among the key-concepts with the highest IDF (i.e. best candidates for final selection), we find 'inversions culturelles', 'hypertextualité', 'aménagement terminologique'. These are key-concepts that occur only in one document of the reference corpus. Among the key-concepts with a low IDF, instead, we find very common terms and expressions such as 'rapport', 'partie' and 'exemple', which are likely to be discarded as key-concepts.

The standard KX architecture has also been adapted to one of the two tracks of DEFT 2012, namely the one in which a terminological list was provided. For that track, the set of documents to be processed was accompanied by a list of domain terminology. By comparing the gold key-concepts in the training set with this list, we observed that all terms in the terminology were also gold key-concepts. Therefore, we modified KX so that, in the final re-ranking, the candidate key-concepts being present in the terminology list were significantly boosted. This adaptation lead to an improvement of almost 0.8 P/R/F1 on the test set (see Section 4).

4 Evaluation

Since KX does not require supervision, we used the training set to identify the best parameter setting, which was then applied in the test phase. The results obtained on the training and on the test set are discussed in the following subsections.

4.1 System evaluation on the training set

We report in Table 1 the best parameter setting on the training documents. Note that the reported evaluation measures have been computed using our own scorer, which counts as correct each key-concept exactly matching with the gold standard (case-insensitive). The results reported for the test set, instead, have been computed by the task organizers with another scorer, which may apply a slightly different strategy.

We extracted for each document the top k key-concepts, with k being the number of key-concepts assigned to each document in the training set (this number may vary from document to document). For this reason, Precision and Recall are the same.

	Task 1 : with terminology	Task 2 : w/o terminology
KX Parameters		
1. $MinCorpus$	8	8
2. $MinDoc$	3	3
3. Use $corpusIdf$	Yes	Yes
4. Multiply relevance by key-concept length	Yes	Yes
5. Consider position of first occurrence	No	No
6. Shorter concept subsumption	No	No
7. Longer concept boosting	No	No
8. Boost key-concepts in terminology list	Yes	No
P/R/F1 on training set	**F1 0.18**	**F1 0.15**

TAB. 1 – Best parameter combination for training set

The results obtained on the training set suggest that the key-concepts required in this task should not be too specific, since the parameters aimed at preferring specific (i.e. longer) key-concepts are not activated in the best performing setting (we refer to parameters n. 6 and 7 in the above Table). Also the position of the first key-concept occurrence is not relevant, since the parameter n. 5 is not part of the best setting. This is in contrast with KX setting used for Semeval 2010 (Pianta et Tonelli, 2010). In that case, boosting the relevance of specific key-concepts, and of those occurring in the article abstract had a positive effect on the final performance. Note also that the performance measured on French documents in DEFT is around 0.10 points lower than that achieved at Semeval on English scientific articles. We believe that this is not due to a different system performance on the two languages, but rather on the evaluation strategy, because Semeval scorer required the key-concepts to be stemmed and took into account some syntactic variations of the same key-concept (Kim *et al.*, 2010).

4.2 System evaluation on the test set

For each task, we submitted three system runs, testing different parameter combinations. Specifically, for *Task 1* (with terminological list), the three runs had the following configurations :

1. Parameter setting reported in Section 4.1 (with boosting of key-concepts in terminology

list) ;

2. Parameter setting as in Section 4.1 but *Consider position of first occurrence* activated (with boosting of key-concepts in terminology list) ;
3. Parameter setting as in Section 4.1 but terminology list is not taken into account.

As for *Task 2* (without terminological list), the three runs had the following configurations :

1. Parameter setting reported in Section 4.1 ;
2. Parameter setting as in Section 4.1 but *Consider position of first occurrence* activated ;
3. Parameter setting reported in Section 4.1 but system run only on article abstracts.

	Task 1 : with terminology	Task 2 : w/o terminology
KX Run 1	0.2682	0.1880
KX Run 2	**0.2737**	**0.1901**
KX Run 3	0.1976	0.1149

TAB. 2 – KX performance on test set

We decided to activate the parameter *Consider position of first occurrence*, even if it was not part of the best performing setting in the training phase, because it achieved good results in the Semeval 2010 challenge on English. The results confirm that, in both tasks, this yielded a (limited) improvement.

In both tasks, the third run was used to exploit configurations that were not tested in the training phase. In Task 1, the third run was obtained without taking into account the terminology list. The difference in performance between Run 1 and Run 3 confirms that this information is indeed very relevant. In Task 2, the third run concerned the extraction of key-concepts only from the abstracts, and not from the whole articles. Also in this case, the initial hypothesis that the abstract may contain all relevant key-concepts proved to be wrong.

At DEFT 2012, 10 teams submitted at least one run in Task 1, and 9 teams in Task 2. The best performing run of KX was ranked $6th$ out of 10 in Task 1 and $5th$ out of 9 in Task 2. In Task 1 the mean F1 for the best submission of each team was 0.3575, the median was 0.3321 and the standard deviation 0.2985, with system performances ranging from 0.0428 (lowest performance) to 0.9488 (best run). In Task 2 the mean F1 for the best submission of each team was 0.2045, the median was 0.1901 and the standard deviation 0.1522, with system performances ranging from 0.0785 (lowest performance) to 0.5874 (best run).

These results show that the use of terminology significantly improves the overall system performance, as confirmed in Table 2. However, KX seems to be more competitive in the second task compared to other systems. This confirms that KX strength lies in its domain-independence and in the fact that is does not require any additional information to achieve a good performance. Furthermore, we believe that the second task is more realistic than the first one : in a real application scenario, it is unlikely that a terminological list, containing only the key-concepts to be identified, is actually available.

5 Conclusions

In this paper, we presented the French version of the KX system, and we described the experiments we carried out for our participation at DEFT 2012. KX achieved good results with few adjustments of the parameter setting and a limited additional effort for domain and language adaptation. Our system requires no supervision and its English and Italian versions are distributed as a standalone key-concept extractor. Its extension, which takes into account a reference terminological list, proved to be effective and achieved a moderate improvement in the first task of the evaluation challenge.

A limitation of our system is that it is not able to identify key-concepts that are not present in the document. This kind of concepts amounted to around 20% of the gold key-concepts in the training set, and this feature strongly affected the outcome of our evaluation. A strategy to exploit external knowledge sources to extract common subsumers of the given key-concepts may be investigated in the future.

Acknowledgements

The development of KX has been partially funded by the European Commission under the contract number FP7-248594, PESCaDO project.

Références

CHRUPALA, G., DINU, G. et van GENABITH, J. (2008). Learning Morphology with Morfette. *In Proceedings of the 6th International Conference on Languages Resources and Evaluations (LREC 2008)*, Marrakech, Morocco.

FRANTZI, K., ANANIADOU, S. et MIMA, H. (2000). Automatic recognition of multi-word terms : the C-value/NC-value. *Journal of Digital Libraries*, 3(2):115–130.

JONES, S., LUNDY, S. et PAYNTER, G. (2002). Interactive Document Summarisation Using Automatically Extracted Keyphrases. *In Proceedings of the 35th Hawaii International Conference on System Sciences*, Hawaii.

KIM, S. N., MEDELYAN, O., KAN, M.-Y. et BALDWIN, T. (2010). SemEval-2010 Task 5 : Automatic keyphrase extraction from scientific articles. *In Proceedings of SemEval 2010, Task 5 : Keyword extraction from Scientific Articles*, Uppsala, Sweden.

PIANTA, E., GIRARDI, C. et ZANOLI, R. (2008). The TextPro tool suite. *In Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

PIANTA, E. et TONELLI, S. (2010). KX : A flexible system for Keyphrase eXtraction. *In Proceedings of SemEval 2010, Task 5 : Keyword extraction from Scientific Articles*, Uppsala, Sweden.

RICCA, F., TONELLA, P., GIRARDI, C. et PIANTA, E. (2004). An empirical study on keyword-based web site clustering. *In Proceedings of the 12th IWPC*, Bari, Italy.

TONELLI, S. et PIANTA, E. (2011). Matching documents and summaries using key-concepts. *In Proceedings of DEFT 2011*, Montpellier, France.