TermITH-Eval: a French Standard-Based Resource for Keyphrase Extraction Evaluation

Adrien Bougouin[1], **Sabine Barreaux**[2], **Laurent Romary**[3], **Florian Boudin**[1], **Béatrice Daille**[1]

[1] Université de Nantes, LINA, 2 rue de la Houssinière, 44322 Nantes, France
[2] INIST-CNRS, 2, allée du parc de Brabois, 54519 Vandoeuvre-lès-Nancy, France
[3] Humboldt-Universität zu Berlin, Dorotheenstraße 24, 10117 Berlin, Germany
E-mail: adrien.bougouin@univ-nantes.fr, sabine.barreaux@inist.fr, laurent.romary@inria.fr,
florian.boudin@univ-nantes.fr, beatrice.daille@univ-nantes.fr

Abstract

Keyphrase extraction is the task of finding phrases that represent the important content of a document. The main aim of keyphrase extraction is to propose textual units that represent the most important topics developed in a document. The output keyphrases of automatic keyphrase extraction methods for test documents are typically evaluated by comparing them to manually assigned reference keyphrases. Each output keyphrase is considered correct if it matches one of the reference keyphrases. However, the choice of the appropriate textual unit (keyphrase) for a topic is sometimes subjective and evaluating by exact matching underestimates the performance. This paper presents a dataset of evaluation scores assigned to automatically extracted keyphrases by human evaluators. Along with the reference keyphrases, the manual evaluations can be used to validate new evaluation measures. Indeed, an evaluation measure that is highly correlated to the manual evaluation is appropriate for the evaluation of automatic keyphrase extraction methods.

Keywords: TermITH-Eval, structured resource, automatic evaluation, keyphrase extraction.

1. Introduction and Motivation

Keyphrases are textual units (words and phrases) that represent the most important topics of a document. Keyphrase extraction is the task of automatically detecting those topics in the content of a document. The common practice to evaluate the performance of keyphrase extraction systems is to compute the number of exact matches between extracted keyphrases and (human assigned) reference keyphrases (Hasan and Ng, 2014). However, this leads to overly pessimistic scores since variations in the extracted keyphrases that might be judged as correct cannot be taken into account (Zesch and Gurevych, 2009).

Producing a more reliable estimate of system performance is not an easy task as assessing whether a textual unit is a keyphrase is highly subjective (Kim et al., 2010). Yet, a handful of attempts have been made in this direction (Zesch and Gurevych, 2009; Kim et al., 2010) but with limited success. The initiating work of Zesch and Gurevych (2009) stated the need for partial matching instead of exact matching but did not show the effectiveness of their measure compared with a human evaluation. Kim et al. (2010) improved the measure of Zesch and Gurevych (2009) and evaluated the correlation of both the original and improved measures with human evaluations. Computing the correlation between an automatic evaluation measure and human evaluators is an effective way of measuring how close the automatic judgment is to human judgment. However, the results shown by Kim et al. (2010) were not significant enough to influence automatic evaluation in recent work. Also, Kim et al. (2010) did not provide the manual evaluation data they used to correlate the evaluation measures with the manual evaluation. Researchers would benefit from such data and the problem would be more effectively addressed.

This paper describes the construction of a corpus for which the outputs of three keyphrase extraction systems were manually evaluated[1]. More specifically, our work has three main contributions. First, we present evaluation guidelines for manual keyphrase evaluation regarding two aspects: appropriateness and silence (Section 2.). Second, we propose a structured format to ease data access and analysis (Section 3.). Finally, we provide an analysis of the manual evaluations and show why it is important to work on new evaluation measures for automatic keyphrase extraction (Section 4.).

2. TermITH-Eval Dataset

We selected 400 French bibliographic records from the FRANCIS and PASCAL databases of the French Institute for Scientific and Technical Information (Inist). The records cover four specific-domains (100 each): Linguistics, Information Science, Archaeology and Chemistry. Every bibliographic record contains a title, an abstract, author keyphrases and reference keyphrases assigned by professional indexers. This work only take into account the titles, abstracts and keyphrases assigned by professional indexers (see Figure 1).

The following subsections present the guidelines given to professional indexers for assigning the reference keyphrases, the three keyphrase extraction methods used to automatically extract keyphrases and the guidelines given to professional indexers for the evaluation of automatically extracted keyphrases.

2.1. Indexing Guidelines

Indexing is the process of describing and identifying a document in terms of its subject content, in order to facilitate the retrieval of information from a collection of documents. Professional indexers at the Inist work in their own specialized fields and follow five principles to ensure quality

[1] https://github.com/termith-anr/TermITH-Eval

La cause linguistique *Linguistics*

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Reference keyphrases: Français; interprétation sémantique; conjonction; expression linguistique; concept linguistique; relation syntaxique; cause.

Figure 1: Example of the content of a bibliographic record

indexing: conformity, exhaustivity, consistency, specificity and impartiality.

Conformity relies on a domain terminology (indexing language). Bibliographic records from a given research area are mainly indexed in accordance with the same indexing language and its usage rules.

Exhaustivity completes keyphrases obtained when focusing on conformity. Professional indexers must identify every keyphrase, for a document, that has potential value for information retrieval. Professional indexers also need to include implicit keyphrases if they are useful for the contextualisation of a given keyphrase.

Consistency increases the quality of document indexing and retrieval. If the same concept is important in two bibliographic records of the same domain, then the concept must be represented by the same keyphrase (preferably from the indexing language).

Specificity relies on the term hierarchy in the domain. As a rule, keyphrases must be as specific as possible and more general ones can be added to point their place within the domain (e.g. *Français* – French – in Figure 1).

Impartiality is a required quality for professional indexers to posses. Keyphrases associated to documents must not convey the personal opinion of the indexer regarding the bibliographic record.

To cope with the increasing amount of documents to be referenced, Inist indexers are helped by a pre-indexing system which proposes keyphrases to be validated and enriched. The pre-indexing system relies on pattern matching between text and predefined expressions related to potential keyphrases. The predefined expressions requires constant updating in order to generate appropriate keyphrases.

2.2. Automatic Keyphrase Extraction

We selected three keyphrase extraction methods to extract 30 keyphrases (10 each) per bibliographic record. The methods cover the main techniques used for automatic keyphrase extraction: the statistical method TF-IDF (Salton et al., 1975), the classification method KEA (Witten et al., 1999) and the graph-based method TopicRank (Bougouin et al., 2013).

TF-IDF is a simple and common keyphrase extraction method that ranks the textual units of a document according to their TF-IDF score, frequently used in Information Retrieval. The idea is to give a high importance score to textual units which are both frequent in the document and specific to it. The specificity of a textual unit regarding a document is obtained using a collection of documents. The lower the number of documents in which a textual unit occurs, the more specific this textual unit is.

KEA also relies on simple statistics. According to KEA, a keyphrase can be recognized by its importance (TF-IDF) and the position of its first occurrence within the document. Indeed, Witten et al. (1999) observed that keyphrases tend to appear earlier than later in a document. The two properties (TF-IDF and first position) are used as features of a Naive Bayes classifier that labels either the class of "*keyphrase*" or "*non keyphrase*" to every textual unit of the document.

TopicRank is a graph-based method that ranks topics by importance and extracts one representative keyphrase for each important topic. Topics are clusters of textual units which "contain" the same concept and the representative keyphrase for each topic is its textual unit that appears first within the document.

For comparison purposes, we implemented each method and integrated them on top of the same preprocessing tools. Every document is first segmented into sentences, sentences are tokenized into words and words are labelled according their morphological class (Part-of-Speech tagging — POS tagging). We performed sentence segmentation with the PunktSentenceTokenizer provided by the Python Natural Language ToolKit (NLTK)(Bird et al., 2009), word tokenization using the French tokenizer Bonsai included with the French POS tagger MElt (Denis and Sagot, 2009), which we use for POS tagging.

2.3. Manual Evaluation Guidelines

Four evaluators took part in the manual evaluation. Being chosen for their indexing experience and their expertise in the selected scientific disciplines, evaluators have been asked to follow the guidelines described below.

After reading the title and the abstract of a bibliographic record, evaluators needed to assess if the automatically extracted keyphrases were relevant to the bibliographic record. This assessment is made regarding two aspects: appropriateness and silence.

2.3.1. Appropriateness

Appropriateness is a property of an extracted keyphrase. Appropriate keyphrases suitably represents the subjects and questions discussed in the document described by the bibliographic record. The evaluation of appropriateness is formalized by assigning a score from 2 down to 0, for each extracted keyphrase:

2. The extracted keyphrase is correct, appropriate.

1. The extracted keyphrase represents a subject or question discussed in the document but its textual form is

not the most appropriate. The extracted keyphrase is a synonym, a spelling variant, a morphosyntactic variant, an acronym, an abbreviation or a phrase with the wrong boundaries. In all these cases, the extracted keyphrase is considered as a variant of a preferred form that is present in the text. This preferred form can be proposed as a keyphrase, with a score of 2, and must be linked as the preferred form of the extracted keyphrase with score 1.

0. The extracted keyphrase is inappropriate.

2.3.2. Silence

Silence is the property attached to reference keyphrases. A silence means that the information held by a given reference keyphrase is not represented by one or more extracted keyphrases. In order to evaluate the silence of the keyphrase extraction method, the evaluators need to check every reference keyphrase and determine whether it complements the assessed method or not. The evaluation of silence is formalized by assigning a score from 2 down to 0, for each reference keyphrase:

2. The reference keyphrase is highly complementary to the keyphrase extraction method. The reference keyphrase contains a very important information missing from the extracted keyphrases.

1. The reference keyphrase is moderately complementary to the keyphrase extraction method. The reference keyphrase contains a secondary or implicit information missing (or partially missing) from the extracted keyphrases.

0. The reference keyphrase is not complementary to the keyphrase extraction method. The reference keyphrase has been extracted by the method or cannot be extracted because the notion is absent from the text.

3. TermITH-Eval Format

In the purpose of the TermITH-Eval dataset, we had to tackle the challenge that complex annotations combining automatic extractions, manual annotations, as well as scoring information, would occur within our document. Our choice for dealing with such a complex document structure was to use the TEI guidelines, which particularly offer customization facilities for the identification of an optimal trade-off between full compliance to the TEI architecture and integration of project specific constraints. More precisely we integrated two extensions to the TEI standard representation:

- We used the work done in (Romary, 2014) to complement the TEI guidelines with terminological entries compliant to ISO standard 30042 (TBX, TermBase eXchange). This in turn has now become a proposal to the TEI consortium;

- We heavily experimented with the new proposal (`https://github.com/TEIC/TEI/issues/374`) for an in-document stand-off annotation element, which would allow to class together groups of annotations (e.g. from the same term extraction process). We also added TBX entries as possible body objects (in the sense of the Open Annotation framework) to the stand-off proposal.

All in all, this work of compiling the best of existing but also on-going standardisation efforts, has proved to be highly effective for our project, especially when keyphrase extraction outputs had to be sent for evaluation, and we see this as a possible reference framework for similar projects.

4. TermITH-Eval Analysis

Here, we present and analyse the evaluation scores given by human evaluators regarding the three automatic keyphrase extraction methods applied to each specific domain of our dataset. To allow comparison with automatic keyphrases, Table 1 shows the f1-scores obtained by each method using the standard automatic evaluation approach.

Method	Linguistics	Information Science	Archaeology	Chemistry
TF-IDF	14.0	**13.2**	22.1	12.6
KEA	**14.7**	12.5	**23.9**	**12.8**
TopicRank	11.9	12.1	21.8	11.8

Table 1: Results of the automatic evaluation of TF-IDF, KEA and TopicRank in term of f1-score on each specific domain

Table 2 shows the ratios of appropriateness scores per each method per specific domain of our dataset. To judge if one method outperforms others, we looked for a highest ratio of keyphrases with score 2, a highest ratio of non redundant keyphrases with score 1, a lowest ratio of redundant keyphrases with score 1 and a lowest ratio of keyphrases with score 0. Non redundant and redundant keyphrases with score 1 are distinguished by the `PreferedForm` given by the evaluator. If the extracted keyphrase with a score of 1 has a specified `PreferedForm`, then it is considered redundant because it is similar to another keyphrase that has also been extracted. First, we observe that our guidelines enable a deeper analysis of the methods. Indeed, looking at the results of TopicRank proves that it is less redundant than other methods. The latter observation is one of the main objectives of the author (Bougouin et al., 2013) of TopicRank. However, their evaluation using the standard approach did not show that TopicRank extracts less redundant keyphrases than other methods. Secondly, the ordering of the methods from the best performing to the worst performing changed according to whether evaluation was automatic or manual. With the automatic evaluation, TopicRank is the method that performs the worst yet it performs better than TF-IDF in every case and better than KEA in half of the cases when analysed with manual evaluation. This is due to the fact that automatic evaluation is much more pessimistic than manual evaluation, which deals with subjectivity. As a few researchers have stated (Zesch and Gurevych, 2009; Kim et al., 2010), the automatic evaluation of keyphrase extraction methods must change to enable it to take subjectivity into account, e.g. by accepting variant forms of reference keyphrases.

Table 3 shows the ratios of silence scores per each method per specific domain of our dataset. To judge if a method

Score	Linguistics			Information Science			Archaeology			Chemistry		
	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank
2	35.3	**37.2**	37.1	34.7	34.2	**36.3**	46.0	49.9	**51.6**	50.9	**54.0**	53.7
1 – non redundant	4.2	**9.8**	5.7	15.3	18.3	**18.5**	14.1	**16.3**	15.4	25.9	24.1	**29.1**
1 – redundant	6.8	8.9	**0.9**	8.1	7.6	**2.8**	4.0	5.7	**0.8**	4.6	5.7	**1.2**
0	53.8	**44.0**	56.3	41.9	**39.9**	42.4	35.9	**28.1**	32.2	18.7	16.3	**16.0**

Table 2: Appropriateness ratios of TF-IDF, KEA and TopicRank on each specific domain

Score	Linguistics			Information Science			Archaeology			Chemistry		
	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank	TF-IDF	KEA	TopicRank
2	20.1	**16.2**	16.8	25.5	22.0	**21.6**	38.2	33.0	**32.8**	22.0	**17.1**	19.2
1	48.5	**45.3**	48.3	25.8	25.8	**25.3**	23.3	23.2	**22.7**	**32.0**	32.2	32.2
0	31.4	**38.5**	35.0	48.7	52.2	**53.1**	38.5	43.9	**44.5**	46.0	**50.7**	48.6

Table 3: Silence ratios of TF-IDF, KEA and TopicRank on each specific domain

outperforms others, we look for a lowest ratio of reference keyphrases with score of 2, a lowest ratio of reference keyphrases with score 1 and a higher ratio of keyphrases with score 0. This new aspect for the evaluation of keyphrases is interesting because it compares the methods regarding the importance of information held by the extracted keyphrases. Once again, the finding vary according to whether the evaluation is automatic or manual. In the future, it would be interesting to see new automatic evaluation measurement techniques that could assess whether a keyphrase extraction method outputs the most important keyphrases first when given ordered reference keyphrase to analyse.

5. Conclusion and Perspectives

This paper presented a new dataset for keyphrase extraction. Unlike any other dataset in the automatic keyphrase extraction community, our dataset includes extracted keyphrases with manual evaluations regarding two aspects: appropriateness and silence. Given that the current automatic keyphrase evaluation measurements is too pessimistic because they cannot manage subjectivity, our dataset or similar can be used to propose and validate new keyphrase evaluation measures.

6. Acknowledgments

This work was partially supported by the French National Research Agency (TermITH project — ANR–12–CORD–0029) and by the French National Center for Scientific Research (GOLEM project — grant of CNRS PEPS FaSciDo 2015, `http://boudinfl.github.io/GOLEM/`).

7. References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 543–551, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Pascal Denis and Benoît Sagot. 2009. Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 110–119, Hong Kong, December. City University of Hong Kong.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland, June. Association for Computational Linguistics.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2010. Evaluating N-Gram Based Evaluation Metrics for Automatic Keyphrase Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 572–580, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laurent Romary. 2014. TBX goes TEI – Implementing a TBX basic extension for the Text Encoding Initiative guidelines. In *Terminology and Knowledge Engineering 2014*, Terminology and Knowledge Engineering, TKE 2014, Berlin, Germany, June.

Gerard Salton, Andrew Wong, and Chungshu Yang. 1975. A Vector Space Model for Automatic Indexing. *Communication ACM*, 18(11):613–620, November.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill Manning. 1999. KEA: Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, New York, NY, USA. ACM.

Torsten Zesch and Iryna Gurevych. 2009. Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of the International Conference RANLP-2009*, pages 484–489, Borovets, Bulgaria, September. Association for Computational Linguistics.