UvT: The UvT Term Extraction System in the Keyphrase Extraction task

Kalliopi Zervanou
ILK / TiCC - Tilburg centre for Cognition and Communication
University of Tilburg, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
K.Zervanou@uvt.nl

Abstract

The UvT system is based on a hybrid, linguistic and statistical approach, originally proposed for the recognition of multiword terminological phrases, the C-value method (Frantzi et al., 2000). In the UvT implementation, we use an extended noun phrase rule set and take into consideration orthographic and morphological variation, term abbreviations and acronyms, and basic document structure information.

1 Introduction

The increasing amount of documents in electronic form makes imperative the need for document content classification and semantic labelling. Keyphrase extraction contributes to this goal by the identification of important and discriminative concepts expressed as keyphrases. Keyphrases as reduced document content representations may find applications in document retrieval, classification and summarisation (D'Avanzo and Magnini, 2005). The literature distinguishes between two principal processes: keyphrase extraction and keyphrase assignment. In the case of keyphrase assignment, suitable keyphrases from an existing knowledge resource, such as a controlled vocabulary, or a thesaurus are assigned to documents based on classification of their content. In keyphrase extraction, the phrases are mined from the document itself. Supervised approaches to the problem of keyphrase extraction include the Naive Bayes-based KEA algorithms (Gordon et al., 1999) (Medelyan and Witten, 2006), decision tree-based and the genetic algorithm-based GenEx (Turney, 1999), and the probabilistic KL divergence-based language model (Tomokiyo and Hurst, 2003). Research in keyphrase extraction proposes the detection of keyphrases based on various statistics-based, or pattern-based features. Statistical measures investigated focus primarily on keyphrase frequency measures, whereas pattern-features include noun phrase pattern filtering, identification of keyphrase head and respective frequencies (Barker and Cornacchia, 2000), document section position of the keyphrase (e.g., (Medelyan and Witten, 2006)) and keyphrase coherence (Turney, 2003). In this paper, we present an unsupervised approach which combines pattern-based morphosyntactic rules with a statistical measure, the C-value measure (Frantzi et al., 2000) which originates from research in the field of automatic term recognition and was initially designed for specialised domain terminology acquisition.

2 System description

The input documents in the Keyphrase Extraction task were scientific articles converted from their originally published form to plain text. Due to this process, some compound hyphenated words are erroneously converted into a single word (e.g., "resourcemanagement" vs. "resource-management"). Moreover, document sections such as tables, figures, footnotes, headers and footers, often intercept sentence and paragraph text. Finally, due to the particularity of the scientific articles domain, input documents often contain irregular text, such as URLs, inline bibliographic references, mathematical formulas and symbols. In our approach, we attempted to address some of these issues by document structuring, treatment of orthographic variation and filtering of irregular text.

The approach adopted first applies part-of-speech tagging and basic document structuring (sec. 2.1 and 2.2). Subsequently, keyphrase candidates conforming to pre-defined morphosyntactic rule patterns are identified (sec. 2.3). In the next stage, orthographic, morphological and abbreviation variation phenomena are addressed

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 194–197,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

(sec. 2.4) and, finally, candidate keyphrases are selected based on C-value statistical measure (sec. 2.5).

2.1 Linguistic pre-processing

For morphosyntactic analysis, we used the Maxent (Ratnaparkhi, 1996) POS tagger implementation of the openNLP toolsuite[1]. In order to improve tagging accuracy, irregular text, such as URLs, inline references, and recurrent patterns indicating footers and mathematical formulas are filtered prior to tagging.

2.2 Basic document structuring

Document structuring is based on identified recurrent patterns, such as common section titles and legend indicators (e.g., "Abstract", "Table..."), section headers numbering and preserved formatting, such as newline characters. Thus, the document sections that the system may recognise are: Title, Abstract, Introduction, Conclusion, Acknowledgements, References, Header (for any other section headers and legends) and Main (for any other document section text).

2.3 Rule pattern filtering

The UvT system considers as candidate keyphrases, those multi-word noun phrases conforming to pre-defined morphosyntactic rule patterns. In particular, the patterns considered are:

$M^+ \, N$

$M \, C \, M \, N$

$M^+ \, N \, C \, N$

$N \, P \, M^* \, N$

$N \, P \, M^* \, N \, C \, N$

$N \, C \, N \, P \, M^* \, N$

$M \, C \, M \, N$

$M^+ \, N \, C \, N$

where M is a modifier, such as an adjective, a noun, a present or past participle, or a proper noun including a possessive ending, N is a noun, P a preposition and C a conjunction. For every sentence input, the matching process is exhaustive: after the longest valid match is identified, the rules are re-applied, so as to identify all possible shorter valid matches for nested noun phrases. At this stage, the rules also allow for inclusion of potential abbreviations and acronyms in the identified noun phrase of the form:

$M^+ \, (A) \, N$

$M^+ \, N \, (A)$

where (A) is a potential acronym appearing as a single token in uppercase, enclosed by parentheses and tagged as a proper noun.

2.4 Text normalisation

In this processing stage, the objective is the recognition and reduction of variation phenomena which, if left untreated, will affect the C-value statistical measures at the keyphrase selection stage. Variation is a pervasive phenomenon in terminology and is generally defined as the alteration of the surface form of a terminological concept (Jacquemin, 2001). In our approach, we attempt to address morphological variation, i.e., variation due to morphological affixes and orthographic variation, such as hyphenated vs. non-hyphenated compound phrases and abbreviated phrase forms vs. full noun phrase forms.

In order to reduce morphological variation, UvT system uses the J.Renie interface[2] to WordNet lexicon[3] to acquire lemmas for the respective candidate phrases. Orthographic variation phenomena are treated by rule matching techniques. In this process, for every candidate keyphrase matching a rule, the respective string alternations are generated and added as variant phrases. For example, for patterns including acronyms and the respective full form, alternative variant phrases generated may contain either the full form only, or the acronym replacing its respective full form. Similarly, for hyphenated words, non-hyphenated forms are generated.

2.5 C-value measure

The statistical measure used for keyphrase ranking and selection is the C-value measure (Frantzi et al., 2000). C-value was originally proposed for defining potential terminological phrases and is based on normalising frequency of occurrence measures

[1] http://opennlp.sourceforge.net/

[2] http://www.ai.mit.edu/ jrennie/WordNet/

[3] http://wordnet.princeton.edu/

Performance over Reader-Assigned Keywords									
System	top 5 candidates			top 10 candidates			top 15 candidates		
	P	R	F	P	R	F	P	R	F
TF·IDF	17.80%	7.39%	10.44%	13.90%	11.54%	12.61%	11.60%	14.45%	12.87%
NB & ME	16.80%	6.98%	9.86%	13.30%	11.05%	12.07%	11.40%	14.20%	12.65%
UvT	**20.40%**	**8.47%**	**11.97%**	**15.60%**	**12.96%**	**14.16%**	**11.93%**	**14.87%**	**13.24%**
UvT - A	23.60%	9.80%	13.85%	16.10%	13.37%	14.61%	12.00%	14.95%	13.31%
UvT - I	21.20%	8.80%	12.44%	14.50%	12.04%	13.16%	12.00%	14.95%	13.31%
UvT - M	20.40%	8.47%	11.97%	15.10%	12.54%	13.70%	11.40%	14.20%	12.65%
UvT - IC	23.20%	9.63%	13.61%	16.00%	13.29%	14.52%	13.07%	16.28%	14.50%

Performance over Combined Keywords									
System	top 5 candidates			top 10 candidates			top 15 candidates		
	P	R	F	P	R	F	P	R	F
TF·IDF	22.00%	7.50%	11.19%	17.70%	12.07%	14.35%	14.93%	15.28%	15.10%
NB & ME	21.40%	7.30%	10.89%	17.30%	11.80%	14.03%	14.53%	14.87%	14.70%
UvT	**24.80%**	**8.46%**	**12.62%**	**18.60%**	**12.69%**	**15.09%**	**14.60%**	**14.94%**	**14.77%**
UvT - A	28.80%	9.82%	14.65%	19.60%	13.37%	15.90%	14.67%	15.01%	14.84%
UvT - I	26.40%	9.00%	13.42%	17.80%	12.14%	14.44%	14.73%	15.08%	14.90%
UvT - M	24.80%	8.46%	12.62%	17.90%	12.21%	14.52%	14.07%	14.39%	14.23%
UvT - IC	28.60%	9.75%	14.54%	19.70%	13.44%	15.98%	16.13%	16.51%	16.32%

Table 1: UvT, UvT variants and baseline systems performance on the Keyphrase Extraction Task

by taking into consideration the candidate multi-word phrase constituent length and terms appearing as nested within longer terms. In particular, depending on whether a candidate multi-word phrase is nested or not, C-value is defined as:

$$\text{C-value} = \begin{cases} \log_2 |a| f(a) \\ \log_2 |a| (f(a) - \frac{1}{P(T_a)} \sum\limits_{b \in T_a} f(b)) \end{cases}$$

In the above, the first C-value measurement is for non-nested terms and the second for nested terms, where a denotes the word sequence that is proposed as a term, $|a|$ is the length of this term in words, $f(a)$ is the frequency of occurrence of this term in the corpus, both as an independent term and as a nested term within larger terms, and $P(T_a)$ denotes the probability of a term string occurring as nested term.

In this processing stage of keyphrase selection, we start by measuring frequency of occurrence for all our candidate phrases, taking into consideration phrase variants, as identified in the *Text normalisation* stage. Then, we proceed by calculating nested phrases frequences and, finally, we estimate C-value.

The result of this process is a list of proposed keyphrases, ranked by decreasing C-value measure, wherefrom the top 15 were selected for the evaluation of the system results.

3 Results

The overall official results of the UvT system are shown in Table 1, where P, R and F correspond to micro-averaged precision, recall and F-score for the respective sets of candidate keyphrases, based on reader-assigned and combined author- and reader-assigned gold standards. Table 1 also illustrates the reported performance of the task baseline systems (i.e., TF·IDF, Naive Bayes (NB) and maximum entropy (ME)[4]) and the UvT system performance variance based on document section candidates (-A: Abstract, -I: Introduction, -M: Main, -IC: Introduction and Conclusion combination). In these system variants, rather than selecting the top 15 C-value candidates from the system output, we also apply restrictions based on the candidate keyphrase document section information, thus skipping candidates which do not appear in the respective document section.

Overall, the UvT system performance is close to the baseline systems results. We observe that the system exhibits higher performance for its top

[4] The reported performance of both NB and ME for the respective gold-standard sets in the Keyphrase Extraction Task is identical.

5 candidate set and this performance drops rapidly as we include more terms in the answer set. One possible reason for its average performance could be attributed to increased "noise" in the results set. In particular, our text filtering method failed to accurately remove a large amount of irregular text in form of mathematical formulas and symbols which were erroneously tagged as proper nouns. As indicated in Table 1, the improved results of system variants based on document sections, such as Abstract, Introduction and Conclusion, where these symbols and formulas are rather uncommon, could be partly attributed to "noise" reduction.

Interestingly, the best system performance in these document section results is demonstrated by the Introduction-Conclusion combination (UvT-IC). Other tested combinations (not illustrated in Table 1), such as abstract-intro, abstract-intro-conclusions, abstract-intro-conclusions-references, display similar results on the reader-assigned set and a performance ranging between 15,6-16% for the 15 candidates on the combined set, while the inclusion of the Main section candidates reduces the performance to the overall system output (i.e., UvT results). Further experiments are required for refining the criteria for document section information, when the text filtering process for "noise" is improved.

Finally, another reason that contributes to the system's average performance lies in its inherent limitation for the detection of multi-word phrases, rather than both single and multi-word. In particular, single word keyphrases account for approx. 20% of the correct keyphrases in the gold standard sets.

4 Conclusion

We have presented an approach to keyphrase extraction mainly based on adaptation and implementation of the C-value method. This method was originally proposed for the detection of terminological phrases and although domain terms may express the principal informational content of a scientific article document, a method designed for their exhaustive identification (including both nested and longer multi-word terms) has not been proven more effective than baseline methods in the keyphrase detection task. Potential improvements in performance could be investigated by (1) improving document structure detection, so as to reduce irregular text, (2) refinement of document section information in keyphrase selection, (3) adaptation of the C-value measure, so as to possibly combine keyphrase frequency with a discriminative measure, such as idf.

References

Ken Barker and Nadia Cornacchia. 2000. Using noun phrase heads to extract document keyphrases. In *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*, pages 40–52, Montreal, Canada, May.

Ernesto D'Avanzo and Bernado Magnini. 2005. A keyphrase-based approach to summarization: the LAKE system. In *Proceedings of Document Understanding Conferences*, pages 6–8, Vancouver, BC, Canada, October 9-10.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The C-Value/NC-value Method. *Intern. Journal of Digital Libraries*, 3(2):117–132.

Ian Witten Gordon, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-manning. 1999. Kea: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM conference on Digital Libraries*, pages 254–256, Berkeley, CA, USA, August 11-14. ACM Press.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Olena Medelyan and Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–297, New York, NY, USA. ACM.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Empirical Methods in Natural Language Processing*, pages 133–142.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

Peter Turney. 1999. Learning to extract keyphrases from text. Technical Report ERB-1057, National Research Council, Institute for Information Technology, February 17.

Peter Turney. 2003. Coherent keyphrase extraction via web mining. In *IJCAI'03: Proceedings of the 18th international joint conference on Artificial intelligence*, pages 434–439, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.