A Semantic Metadata Generator for Web Pages Based on Keyphrase Extraction

Dario De Nart, Carlo Tasso, Dante Degl'Innocenti

Artificial Intelligence Lab
Department of Mathematics and Computer Science
University of Udine, Italy
{dario.denart,carlo.tasso}@uniud.it, dante.deglinnocenti@spes.uniud.it

Abstract. The annotation of documents and web pages with semantic metatdata is an activity that can greatly increase the accuracy of Information Retrieval and Personalization systems, but the growing amount of text data available is too large for an extensive manual process. On the other hand, automatic keyphrase generation and wikification can significantly support this activity. In this demonstration we present a system that automatically extracts keyphrases, identifies candidate DBpedia entities, and returns as output a set of RDF triples compliant with the Opengraph and the Schema.org vocabularies.

1 Introduction

In the last few years we have witnessed the rapid growth of the Semantic Web and all its related technologies, in particular the ones that allow the embedding of semantic data inside the HTML markup of Web pages, such as RDFa. Recent studies highlight how a significant part of the most visited pages of the Web is annotated with semantic data and this number is expected to grow in the near future. However, up to now, the majority of such metadata is manually authored and maintained by the owners of the pages, especially those associated with textual content (such as articles and blog posts). Keyphrase Extraction (herein KPE) and Wikification can greatly ease this task, by identifying automatically relevant concepts in the text and Wikipedia/DBpedia entities to be linked. In this demonstration we propose a system for semantic metadata generation based on a knowledge-based KPE and Wikification phase and a subsequent rule-based translation of extracted knowledge into RDF [1]. Generated metadata adhere to the Opengraph and the Schema.org vocabularies which currently are, according to a recent study [2], wide-spread on the Web.

2 Related Work

Several authors in the literature have already addressed the problem of extracting keyphrases (herein KPs) from natural language documents and a wide range of

[1] A live demo of the system can be found at http://goo.gl/beKJu5 and can be accessed by logging as user “guest” with password “guest”

approaches have been proposed. The authors of [11] identify four types of KPE strategies:

- *Simple Statistical Approaches*: mostly unsupervised techniques, considering word frequency, TF-IDF or word co-occurency [8].
- *Linguistic Approaches*: techniques relying on linguistic knowledge to identify KPs. Proposed methods include lexical analysis [1], syntactic analysis [4], and discourse analysis [6].
- *Machine Learning Approaches*: techniques based on machine learning algorithms such as Naive Bayes classifiers and SVM. Systems such as KEA [10], LAKE [3], and GenEx [9] belong to this category.
- *Other Approaches*: other strategies exist which do not fit into one of the above categories, mostly hybrid approaches combining two or more of the above techniques. Among others, heuristic approaches based on knowledge-based criteria [7] have been proposed.

Automatic semantic data generation from natural language text has already been investigated as well and several knowledge extraction systems already exist [5], such as OpenCalais [2], AIDA[3], Apache Stanbol[4], and NERD[5].

3 System Overview

The proposed system includes three main modules: a Domain Independent KPE module (herein DIKPE), a KP Inference module (KPIM), and a RDF Triple Builder (RTB). Our KPE technique exploits a knowledge-based strategy. After a candidate KP generation stage, candidate KPs are selected according to various features including statistic (such as word frequency), linguistic (part of speech analysis), meta-knowledge based (life span in the text, first and last occurrence, and presence of specific tags), and external-knowledge based (existence of a match with a DBpedia entity) ones. Such features correspond to different kinds of knowledge that are involved in the process of recognizing relevant entities in a text. Most of such features are language-independent and the modular architecture of DIKPE allows an easy substitution of language-dependent components, making our framework language-independent. Currently English and Italian languages are supported.

The result of this KPE phase is a set of relevant KPs including DBpedia matches, hence providing a partial wikification of the text. Such knowledge is used by the KPIM for a further step of KP generation, in which a new set of potentially relevant KPs not included in the text is inferred exploiting the link structure of DBpedia. Properties such as *type* and *subject* are considered in order to discover concepts possibly related to the text. Finally, the extracted and the inferred KPs are used by the RTB to build a set of Opengraph and Schema.org triples. Due to

[2] http://www.opencalais.com/

[3] www.mpi-inf.mpg.de/yago-naga/aida/

[4] https://stanbol.apache.org/

[5] http://nerd.eurecom.fr/

the simplicity of the adopted vocabularies, this task is performed in a rule-based way. The rdf fragment to be generated, in fact, is considered by the RTB as a template to fill according to the data provided by the DIKPE and the KPIM.

4 Evaluation and Conclusions

In order to support and validate our approach several experiments have been performed. Due to the early stage of development of the system and being the KP generation the critical component of the systems, testing efforts were focused on assessing the quality of generated KPs. The DIKPE module was benchmarked against the KEA algorithm on a set of 215 English documents labelled with keyphrases generated by the authors and by additional experts. For each document, the KP sets returned by the two compared systems were matched against the set of human generated KPs. Each time a machine-generated KP matched a human-generated KP, it was considered a correct KP; the number of correct KPs generated for each document was then averaged over the whole data set. Various machine-generated KP set sizes were tested. As shown in Table 1, the DIKPE system significantly outperformed the KEA baseline. A user evaluation of the perceived quality of generated KPs was also performed: a set of 50 articles was annotated and a pool of experts of various ages and gender was asked to assess the quality of generated metadata. Table 2 shows the results of the user evaluation.

Table 1. Performance of DIKPE compared to KEA.

Extracted	Average number of correct KPs	
Keyphrases	KEA	DIKpE
7	2.05	3.86
15	2.95	5.29
20	3.08	5.92

Table 2. User evaluation of generated keyphrases.

Evaluation	Frequency
Good	56,28%
Too Generic	14,72%
Too Specific	2,27%
Incomplete	9,85%
Not Relevant	9,85%
Meaningless	7,03%

Evaluation is, however, still ongoing: an extensive benchmark with more complex Knowledge Extraction systems is planned, as well as further enhancements such as inclusion of more complex vocabularies and integration with the Apache Stanbol framework.

References

1. Barker, K., Cornacchia, N.: Using noun phrase heads to extract document keyphrases. In: Advances in Artificial Intelligence, pp. 40–52. Springer (2000)
2. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of rdfa, microdata, and microformats on the web–a quantitative analysis. In: The Semantic Web–ISWC 2013, pp. 17–32. Springer (2013)
3. DAvanzo, E., Magnini, B., Vallin, A.: Keyphrase extraction for summarization purposes: The lake system at duc-2004. In: Proceedings of the 2004 document understanding conference (2004)
4. Fagan, J.: Automatic phrase indexing for document retrieval. In: Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 91–101. SIGIR '87, ACM, New York, NY, USA (1987), http://doi.acm.org/10.1145/42005.42016
5. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: The Semantic Web: Semantics and Big Data, pp. 351–366. Springer (2013)
6. Krapivin, M., Marchese, M., Yadrantsau, A., Liang, Y.: Unsupervised key-phrases extraction from scientific papers using domain and linguistic knowledge. In: Digital Information Management, 2008. ICDIM 2008. Third International Conference on. pp. 105–112 (Nov 2008)
7. Liu, Z., Li, P., Zheng, Y., Sun, M.: Clustering to find exemplar terms for keyphrase extraction. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1. pp. 257–266. EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA (2009), http://dl.acm.org/citation.cfm?id=1699510.1699544
8. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13(01), 157–169 (2004)
9. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2(4), 303–336 (2000)
10. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: Practical automatic keyphrase extraction. In: Proceedings of the fourth ACM conference on Digital libraries. pp. 254–255. ACM (1999)
11. Zhang, C.: Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems 4(3), 1169–1180 (2008), http://eprints.rclis.org/handle/10760/12305