A New Multi-lingual Knowledge-base Approach to Keyphrase Extraction for the Italian Language

Dante Degl'Innocenti, Dario De Nart and Carlo Tasso
Artificial Intelligence Lab, Department of Mathematics and Computer Science, University of Udine, Udine, Italy

Keywords: Keyphrase Extraction, Information Extraction, Italian Language, Natural Language Processing, Text Analysis, Text Classification, Text Summarization.

Abstract: Associating meaningful keyphrases to text documents and Web pages is an activity that can significantly increase the accuracy of Information Retrieval, Personalization and Recommender systems, but the growing amount of text data available is too large for an extensive manual annotation. On the other hand, automatic keyphrase generation can significantly support this activity. This task is already performed with satisfactory results by several systems proposed in the literature, however, most of them focuses solely on the English language which represents approximately more than 50% of Web contents. Only few other languages have been investigated and Italian, despite being the ninth most used language on the Web, is not among them. In order to overcome this shortage, we propose a novel multi-language, unsupervised, knowledge-based approach towards keyphrase generation. To support our claims, we developed DIKpE-G, a prototype system which integrates several kinds of knowledge for selecting and evaluating meaningful keyphrases, ranging from linguistic to statistical, meta/structural, social, and ontological knowledge. DIKpE-G performs well over English and Italian texts.

1 INTRODUCTION

Due to the growth of the amount of unstructured text data available on the Web and in digital libraries, the demand for automatic summarization and real-time information filtering has rapidly increased. However, such systems need metadata that can precisely and compactly represent the content of a document. Even though a huge number of different metadata formats has been proposed and Semantic Web technologies have grown bigger and bigger over the last few years, the most common way to represent these metadata is still constituted by *KeyPhrases*. A KeyPhrase (herein KP) is a short phrase, typically made of one to four words which identifies an entity, a concept, or a generic topic of interest. Such representation bears several advantages: it is simple to understand, yet expressive and less exposed to polisemy issues than a single-term-keyword representation; moreover it has an high cognitive plausibility, since it is proven (Silverstein et al., 1999) that humans often think in terms of KPs rather than single term keywords or network representations such as concept maps. Associating meaningful KPs to a text is a trivial task for humans, however, even by exploiting social Web collaborative technologies, one cannot expect the whole Web to be manually annotated, therefore automatic KP generation techniques are highly desirable. As shown in section 2, several authors have already addressed the problem of KP generation in English texts, but little work has been done with other languages. Italian, in particular, though being the ninth most used language on the Web (W3Techs, 2014) has never received much attention. In this work, we present DIKpE-G an experimental system specifically built for performing KP Extraction and Inference from Italian and English documents. The proposed system exploits a knowledge-based approach combining various classes of knowledge, in part language-dependent, in part independent and it is designed to emulate some of the cognitive processes that are exploited when a human expert is asked to summarize or classify a text.

The paper is organized as follows: in Section 2 we briefly illustrate some related work; in Section 3 we present our keyphrase generation approach; in Section 4 we give a brief description of the DIKpE-G prototype, in Section 5 we expose some experimental results and, finally, in Section 6 we conclude the paper.

Degl'Innocenti D., De Nart D. and Tasso C..
A New Multi-lingual Knowledge-base Approach to Keyphrase Extraction for the Italian Language.
DOI: 10.5220/0005077100780085
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 78-85
ISBN: 978-989-758-048-2

2 RELATED WORK

Several authors in the literature have already addressed the problem of extracting keyphrases from natural language documents and a wide range of approaches have been proposed. The authors of (Zhang, 2008) identify four types of keyphrase extraction strategies:

- *Simple Statistical Approaches:* these techniques assume that statistical information is enough to identify keywords and KPs, thus they are generally simple and unsupervised; the most widespread statistical approaches consider word frequency, TF-IDF or word co-occurency (Matsuo and Ishizuka, 2004). It is important to note how TF-IDF based methods require a closed document corpora in order to evaluate inverse frequencies, therefore they are not suitable to an open world scenario, where new items can be included in the corpora at any time.
- *Linguistic Approaches:* these techniques rely on linguistic knowledge to identify KPs. Proposed methods include lexical analysis (Barker and Cornacchia, 2000), syntactic analysis (Fagan, 1987), and discourse analysis (Krapivin et al., 2008).
- *Machine Learning Approaches*: since KP extraction can be seen as a classification task, machine learning techniques can be used as well (Frank et al., 1999), (Turney, 2000) and (Hulth, 2003). The usage of Naive Bayes, SVM and other supervised learning strategies has been widely discussed and applied in systems such as KEA (Witten et al., 1999), LAKE (DAvanzo et al., 2004), and GenEx (Turney, 2000).
- *Other Approaches:* other strategies exist which do not fit into one of the above categories and most of the times they are hybrid approaches combining two or more of the above techniques. Among others, heuristic approaches based on knowledge-based criteria (Liu et al., 2009), and meta-knowledge over the domain (Danilevsky et al., 2013) have been proposed.

Also the problem of defining multi-language approaches has been discussed by several authors. In (Litvak et al., 2010) it is presented a multilingual approach towards sentence extraction for summarization purposes based on a machine learning approach. The authors of (Paukkeri et al., 2008) introduce a multilingual KP extraction system exploiting a statistical approach based on word frequency and a reference corpus in 11 different European languages, including Italian. The performance of such system, however, relies on the quality of the reference corpus since phrases not included in the corpus will never be extracted from the text. Moreover, its accuracy proved to be highly variable over the 11 considered languages and overall poor. The authors of (El-Beltagy and Rafea, 2009) propose a more sophisticated approach based on a set of heuristic rules for identifying a set of potentially good candidate KPs; candidate KPs are then selected according to a TF-IDF based score metric. The system exploits two language dependant resources: a stopwords list and a stemmer. Upon a suitable substitution of such language dependant resources, the system proved to perform well in different languages.

Keyphrase extraction from Italian texts has received little attention. The authors of (Ferragina and Scaiella, 2010) propose TAGME, a system whose purpose is to annotate documents with hyperlinks to Wikipedia pages by identifying *anchors* in the text. The task of identifying text anchors can be seen as a naive KP extraction technique and is capable to identify and propose KPs only if they are also in Wikipedia. The system by (Paukkeri et al., 2008), previously mentioned, is also capable of extracting KPs from Italian text, however it features a very limited accuracy.

3 A KNOWLEDGE-BASED APPROACH TO KEYPHRASE GENERATION

In order to accomplish our goals and to take into consideration our previous work on keyphrase extraction for English texts (Pudota et al., 2010), we propose here a *Knowledge-Based* KP extraction technique based upon (i) exploitation of several kinds of knowledge, (ii) consideration of the specific languages addressed, and (iii) typical/common writing styles. An initial design work of knowledge engineering allowed us to identify four classes of knowledge which can be exploited to recognize meaningful phrases in a text:

1. *Statistical Knowledge:* this knowledge deals exclusively with the quantitative aspects of natural language, such as the frequency of a given word in a text or its inverse document frequency in a corpus; though lacking of a clear semantic meaning, it can be useful to identify terms and phrases that characterize a text.
2. *Linguistic Knowledge*: this knowledge comes from the specific language considered and deals with morphological and grammatical aspects of the text; examples of linguistic knowledge are

Part-Of-Speech (POS) tags, the information on whether a given word is a stopword or not, or whether a given sequence of words is constituted by an acceptable pattern of POS tags for a KP (such as, for instance: "noun-noun" or "adjective-noun").

3. *Meta/Structural Knowledge*: this knowledge consists of heuristics over the general structure of the text and typically deals with the position of a phrase in the considered document; an example of meta-knowledge is knowing that phrases appearing in the abstract of an article may be more representative than the ones included in its body. This knowledge corresponds to various writing styles exploited by the author of the text. Another example of exploitable meta-knowledge is constituted by some specific metadata inserted in a document by the author (such as the "topic" meta-tag in Web pages and the "subject" meta-tag in a *PDF* file).
4. *Semantic/Social Knowledge*: this knowledge comes from sources external to the considered text. Semantic knowledge deals with the meaning of the terms present in the candidate KPs and with the typical conceptual context where they are used. An ideal source of semantic knowledge is constituted by ontologies, which describe concepts, their properties, and their mutual relationships, together with the natural language terminology usually exploited for linguistically referring to them. Other common sources of such kind of knowledge are dictionaries, thesauri, classification schema, etc. This knowledge is useful for recognizing terms belonging to a specific jargon and for resolving polysemic words. Other relevant examples of sources of semantic knowledge, which are becoming more and more popular in the participative Web (Web 2.0), are fast growing collaborative dictionaries, thesauri and knowledge bases, such as DBpedia. They feature a very wide conceptual coverage and they provide a way to socially validate candidate KP: for a candidate KP being an entry of one of these sources, means that other humans have already identified it as a meaningful way to linguistically refer to the underlined concept. This is the reason why we consider appropriate to attach to this kind of knowledge also the term "social".

It is important to point out how such classes of knowledge differ from each other in terms of domain and language dependency: as shown in Figure 1 statistical knowledge is both domain and language independent, linguistic knowledge is domain independent, but language dependent, meta/structural knowledge is domain dependent, and, finally semantic/social knowledge may be both domain and language dependent. Domain and language dependency are very different. Domain dependency can be sensibly reduced by considering only general assumptions, such as assuming that most of the interesting concepts of a document will be introduced in its first section. It can also be turned down by taking into account information gathered from dictionaries or ontologies with a very broad scope (such as Wikipedia). Language dependency, on the other hand, cannot be relaxed: language dependent knowledge, indeed, needs dedicated modules and/or knowledge bases.

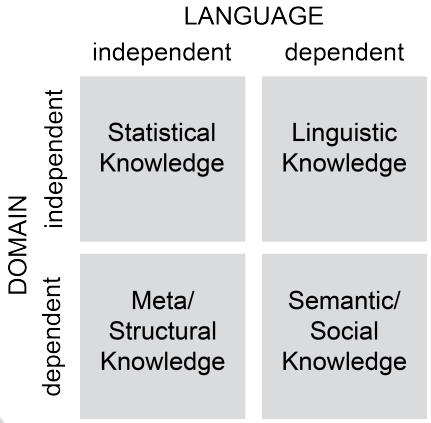

Figure 1: Dependencies of the various kinds of knowledge considered.

When reading a text with the purpose of extracting relevant concepts a human expert typically performs various kinds of evaluations and we believe that, in order to match the performance of a human, an automatic system should try to follow the same process. To this purpose, the overall KP extraction process is organized into three stages: in the first phase, the text is analysed in order to identify all the possible candidate KPs to be possibly extracted from the text. Later, in a second phase, each candidate KP is scored by associating it to a set of features which are the result of applying the various kinds of knowledge described above to the specific candidate KP. More specifically, each class of knowledge is mapped into one or more features and the final selection criterion of candidate KPs takes into account all the features. The chosen features are then combined to produce a final decision associated to the candidate KP: this can be performed, for instance, by means of a unique score or of a multi-dimensional classification technic. This knowledge based approach can be used both in a supervised and an unsupervised scenario. In a supervised scenario the feature combination function could be the result of a training activity of a machine learning algorithm (e.g.: Bayesian classifier, Support Vector Machine, Artificial Neural Network, etc.), while in an unsuper-

vised approach it is explicitly known and may be the result of a knowledge engineering activity. Finally, in the third phase other relevant KPs are generated once the major concepts included in the text have been extracted. In this stage, a domain-dependent inference process takes place, able to identify other (usually more general or related) concepts that are derived starting from the concepts (KPs) extracted in the first two stages and by exploiting external semantic/social knowledge.

4 SYSTEM OVERVIEW

In order to support our claims we have developed *DIKpE-G*, a revised extended version of the system presented in (Pudota et al., 2010) and (De Nart and Tasso, 2014). DIKpE-G stands for *Domain Independent Keyphrase Extractor - Generator*. Figure 2 shows the overall organization of the system.

The data workflow mimics the 3-phase cognitive process described in the previous section. First of all the text is read and the *KP Extraction Module* (*KPEM*) discovers and ranks concepts (KPs) that appear in the text, then the *KP Inference Module* (*KPIM*) augments the set of extracted KPs with new linked, related or implied concepts. Operation of DIKpE-G is also supported by *External Knowledge Sources* (*EKS*): in the current implementation we exploit *Wikipedia*[1] and *Wordnik*[2]. The generated KPs represent tacit and explicit knowledge because part of them is explicitly contained in the text and the rest of them are inferred starting from the ones already present in the text.

In order to identify the KPs, the KPEM relies on a series of *Language Specific Resources* (*LSR*). They consist of a *POS-Tagger* module, a *Stemmer* module and two repositories: one for stopwords and one for POS-Patterns that typically characterize KPs. Decoupling the language dependent part from the rest of the architecture allows us to easily port the system to other languages. All the necessary language dependent modules are in fact widely available for all major languages: for example, the *Snowball stemmer* library[3] provides functionality for over twenty languages and the *TreeTagger*[4] provides POS tagging for over fifteen languages.

The extraction task is organized in two steps: the candidate KPs selection and the ranking phase. In the first step all possible sequences of one, two, three, and four words are considered, but only the ones matching a valid POS pattern are chosen as candidate KPs. Identification of valid POS patterns is a knowledge engineering task and can be carried out by considering widely used patterns (indicated as "valid") in a large enough set of human generated KPs (human generated such as the author KPs included in scientific papers). The number of POS patterns depends on the considered tag set. Currently we have a dozen POS patterns for the Italian language and about 40 for the English language. The difference is due to the different granularity of the employed TAG set.

In the following second step, each candidate KP is assessed by means of a set of features, which are computed by exploiting the various classes of knowledge previously described in Section 3. In the current implementation of DIKpE-G, we are experimenting the set of features introduced in (De Nart and Tasso, 2014). More specifically, in Figure 3, we show, for the various steps of the extraction, the different classes of knowledge taken into account, the relative features considered and, for each of them, their purposes and value range.

As it can be noticed in Figure 3, each feature has a value varying in various ranges. Once for each KP a specific set of values have been computed for its features, a final ranking step is performed, which is aimed at producing a final global rank for each KP. The result is a ranked list of KPs: the highest ranked are proposed as relevant keyphrases for the input text. In our vision, the ranking step can be performed in various ways, ranging from (i) a strictly numerical approach to (ii) a more sophisticated and general knowledge-based assessment based on both qualitative and quantitative reasoning. The highly modular architecture of DIKpE-G, allows a seamless substitution of the modules and submodules devoted to ranking, permitting in such a way the experimentation of alternative approaches. The current DIKpE-G prototype follows the approach proposed in (Pudota et al., 2010), which adheres to a numerical approach: each feature is given a numerical value and all the features are then combined in order to compute a unique index called *keyphraseness*, which represents how much a candidate KP is considered suitable and significant for representing the content of the input text. The keyphraseness index is computed in the current DIKpE-G prototype as a weighted linear combination of the features values. The features weights are currently experimentally obtained. However we are exploring new approaches, namely (i) rule based reasoning for mapping the various features in an n-dimensional space, where different regions of space

[1] www.wikipedia.org
[2] www.wordnik.com
[3] snowball.tartarus.org
[4] www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger

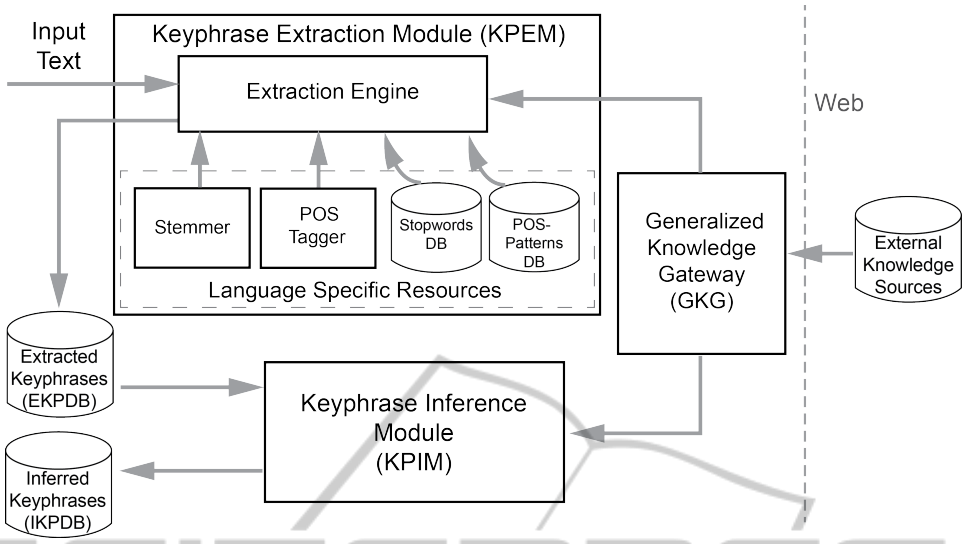

Figure 2: Architecture of the DIKpE-G System.

		Knowledge Class	Feature	Purpose	Value Range
KP EXTRACTION	Candidate KP Identification	Linguistic Knowledge	POS Tag patterns	Excluding certain patterns	
			Stop-word list	Excluding certain words	
			Stemming	Working on common stems	
	Candidate KP Scoring	Linguistic Knowledge	POS Tag patterns	Preferring typical patterns	0-1
		Statistical Knowledge	Frequency	Preferring most frequent terms	0-1
			Co-Occurrence	Preferring common co-occurrent patterns	0-1
		Meta/Structural Knowledge	Phrase depth	Preferring concepts appearing at the beginning of the text	0-1
			Phrase last Occurrance	Preferring concepts mentioned till the end of the text	0-1
			Life Span	Preferring concepts appearing in a large part of the text	0-1
		Semantic/Social Knowledge	Flag of presence in EKS	Preferring KPs appearing in ontologies, dictionaries, thesauri, ...	boolean
			Flag of presence in Web 2.0 EKS	Preferring concepts recognized by other human actors	boolean
KP INFERENCE		Semantic/Social Knowledge	Navigation paths in EKS	Inferring new KPs related to many extracted KPs	
			Navigation paths in EKS	Disambiguating polysemic inferred KPs	

Figure 3: Usage of the various classes of knowledge proposed in DIKpE-G.

are associated to different levels of the keyphraseness index and (ii) machine learning techniques for associating (by means of training based on ad-hoc annotated data sets) the set of the features' values of the single

KPs to the corresponding level of keyphraseness.

The final phase is devoted to inferring new KPs (i.e. KPs which are not already present in the input text) starting from the topmost ranked extracted KPs. The KPIM considers each extracted KP in order to match it against the entries of the available EKSs: if a match is found (i.e. the considered KP is also an entry of a specific EKS), all the concepts (terms) present in the EKS and linked to the matching entry are considered as candidate *inferred* KPs. All the candidate inferred KPs collected from all the extracted KPs are then ranked according to the sum of the keyphraseness values of the extracted KPs from which they have been derived. Note that inferred KPs can be obtained both from hi-ranked or low-ranked extracted KPs. For instance the system can infer a KP that is linked to a large number of low-ranked KPs rather than a KP that is linked to a little number of hi-ranked ones. The top-n inferred KPs are finally returned as output together with the extracted KPs identified by the KPEM.

5 EVALUATION

In order to support and validate our approach several experiments have been performed. To evaluate the performance when considering English texts, the original version (Pudota et al., 2010) was benchmarked against the KEA algorithm on a set of 215 English documents labelled with keyphrases generated by the authors and by additional experts. The comparison was performed only on the KP extraction capabilities and not on the inference ones. 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 and the improvement increases as the KPs set size increases.

Table 1: Performance of DIKpE compared to KEA.

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

When the DIKpE prototype has been extended into the current DIKpE-G prototype, we have added knowledge bases in order to cover also the Italian language. The initial experimental evaluation activity has concerned the Italian language and it has shown very encouraging results. Due to the lack of extensive labelled corpora and available baseline systems, the evaluation of DIKpE-G on the Italian language has followed so far a qualitative approach. A set of 50 papers was gathered, and 11 to 16 KPs were automatically extracted from each paper. A dozen of human experts of various ages and gender were then asked to read all the texts and to assess the quality of extracted KPs. The main goal of the experiment was to identify common pitfalls of the KP extraction process and to classify unsatisfactory KPs extracted. Table 2 shows the seven classes identified and their relative frequency. A significant number of KPs were

Table 2: Results of user evaluation.

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

perceived as “too generic” by our experts; in particular these KPs are generally made of a single word with a very generic meaning such as “catene” (chains) or “funzione” (function) and often were included in other KPs made of multiple words (such as “catene montuose”, that means “mountain ranges”). Another frequent flaw in the extracted KPs by DIKpE-G was the presence of incomplete phrases such as “spaziale Orion”. However also these KPs were often part of a longer phrase that was returned as well (“navicella spaziale Orion”). These observations led us to introduce a simple heuristic consisting in not returning short phrases which are included in longer ones already in the extracted set. This simple mechanism allowed us to significantly increase to 75% the fraction of good KPs as they were presented again to the expert pool.

Results gathered so far are promising, however development is still in progress and further more systematic evaluation activities are planned: we want to evaluate the KP inference capabilities for both the English and the Italian language.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we present a novel knowledge based multilingual approach for KP generation that can be

easily extended to any given Western language due to the actually large availability of resources such as POS taggers and stemming algorithms. Preliminary evaluation results suggest us that once a satisfactory set of language-specific resources is available, the overall quality of the generated KPs is not affected by the language switch. The four different classes of knowledge considered provide a conceptual framework with a higher level of abstraction than other state-of-the-art systems, featuring a clear separation between language dependent and independent KP selection criteria. Such framework allows us to overcome several shortcomings of the current systems which often consider only one or two classes of knowledge. Moreover, the unsupervised nature of our approach allows our system to accomplish its task with no need of training data, which is a major advantage for non-English languages because of the tremendous lack of annotated data corpora that we are experiencing nowadays.

Results gathered so far show a promising outlook and the system can be effectively employed in several application domains, such as digital libraries and recommender systems.

Our future work will therefore address all the major issues highlighted by the expert evaluation, such as a still high number of KPs perceived as too generic. We also aim at improving the overall underlined conceptual model of human KP generation, by further analysing the four knowledge classes identified and by refining the reasoning process exploited in the system. We plan to observe how experts identify KPs, for instance, by thinking-aloud interviews. The user interaction should be improved as well, since the system actually acts as a black box giving little or no hints to the final user of the process that selected a particular KP, and this encourages distrust in the system. In order to address this issue, the development of an interactive explanation and result tracking interface is ongoing. Finally, specific attention will be devoted to the evaluation issues, both (i) for improving and completing the evaluation of our approach and (ii) for contributing to the development of a methodological standard for evaluating KP extraction and KP inference capabilities systems.

REFERENCES

Barker, K. and Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In *Advances in Artificial Intelligence*, pages 40–52. Springer.

Danilevsky, M., Wang, C., Desai, N., Guo, J., and Han, J. (2013). Kert: Automatic extraction and ranking of topical keyphrases from content-representative document titles. *arXiv preprint arXiv:1306.0271*.

DAvanzo, E., Magnini, B., and Vallin, A. (2004). Keyphrase extraction for summarization purposes: The lake system at duc-2004. In *Proceedings of the 2004 document understanding conference*.

De Nart, D. and Tasso, C. (2014). A domain independent double layered approach to keyphrase generation. In *WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies*, pages 305–312. SCITEPRESS Science and Technology Publications.

El-Beltagy, S. R. and Rafea, A. (2009). Kp-miner: A keyphrase extraction system for english and arabic documents. *Information Systems*, 34(1):132–144.

Fagan, J. (1987). Automatic phrase indexing for document retrieval. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '87, pages 91–101, New York, NY, USA. ACM.

Ferragina, P. and Scaiella, U. (2010). Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., and et al. (1999). Domain-specific keyphrase extraction. In *Proc. Sixteenth International Joint Conference on Artificial Intelligence*, pages 668–673. Morgan Kaufmann Publishers.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krapivin, M., Marchese, M., Yadrantsau, A., and Liang, Y. (2008). Unsupervised key-phrases extraction from scientific papers using domain and linguistic knowledge. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 105–112.

Litvak, M., Last, M., and Friedman, M. (2010). A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 927–936. Association for Computational Linguistics.

Liu, Z., Li, P., Zheng, Y., and Sun, M. (2009). 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*, EMNLP '09, pages 257–266, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matsuo, Y. and Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169.

Paukkeri, M.-S., Nieminen, I. T., Pöllä, M., and Honkela, T. (2008). A language-independent approach to

keyphrase extraction and evaluation. In *COLING (Posters)*, pages 83–86.

Pudota, N., Dattolo, A., Baruzzo, A., and Tasso, C. (2010). A new domain independent keyphrase extraction system. In Agosti, M., Esposito, F., and Thanos, C., editors, *Digital Libraries*, volume 91 of *Communications in Computer and Information Science*, pages 67–78. Springer Berlin Heidelberg.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. In *ACm SIGIR Forum*, volume 33, pages 6–12. ACM.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4):303–336.

W3Techs (2014). Usage of content languages for websites. Available online at: http://w3techs.com/technologies.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., and Nevill-Manning, C. G. (1999). Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255. ACM.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3):1169–1180.