Noname manuscript No.
(will be inserted by the editor)

Creation and Evaluation of Large Keyphrase Extraction Collections with Multiple Opinions

Lucas Sterckx · Thomas Demeester · Johannes Deleu · Chris Develder

Received: date / Accepted: date

Abstract While several Automatic Keyphrase Extraction (AKE) techniques have been developed and analyzed, there is little consensus on the definition of the task and a lack of overview of the effectiveness of different techniques. Proper evaluation of keyphrase extraction requires large test collections with multiple opinions, currently not available for research. In this paper, we (i) present a set of test collections derived from various sources with multiple annotations (which we also refer to as *opinions* in the remained of the paper) for each document, (ii) systematically evaluate keyphrase extraction using several supervised and unsupervised AKE techniques, (iii) and experimentally analyze the effects of disagreement on AKE evaluation. Our newly created set of test collections spans different types of topical content from general news and magazines, and is annotated with multiple annotations per article by a large annotator panel. Our annotator study shows that for a given document there seems to be a large disagreement on the preferred keyphrases, suggesting the need for multiple opinions per document. A first systematic evaluation of ranking and classification of keyphrases using both unsupervised and supervised AKE techniques on the test collections shows a superior effectiveness of supervised models, even for a low annotation effort and with basic positional and frequency features, and highlights the importance of a suitable keyphrase candidate generation approach. We also study the influence of multiple opinions, training data and document length on evaluation of keyphrase extraction. Our new test collection for keyphrase extraction is one of the largest of its kind and will be made available to stimulate future work to improve reliable evaluation of new keyphrase extractors.

Keywords Automatic Keyphrase Extraction · Test Collections · Annotator Disagreement

Lucas Sterckx
Ghent University – imec
Technologiepark Zwijnaarde 15, 9052 Ghent, Belgium
Tel.: +32 9 331 49 79
Fax: +32 09 331 48 99
E-mail: lucas.sterckx@ugent.be

Thomas Demeester
E-mail: thomas.demeester@ugent.be ·
Johannes Deleu
E-mail: johannes.deleu@ugent.be ·
Chris Develder
E-mail: chris.develder@ugent.be

1 Introduction

Automatic keyphrase extraction (AKE) is the task of automatically extracting the most important and topical phrases of a document (Turney, 2000). Keyphrases are meant to cover all topics and capture the complete content of a document in but a handful of phrases. Applications of keyphrases are rich and diverse, ranging from document summarization (D'Avanzo et al, 2004) to clustering (Hammouda et al, 2005), contextual advertisement (Yih et al, 2006), or simply to enhance navigation through large corpora. While much research has been done on developing supervised (Hulth, 2003; Lopez and Romary, 2010; Kim and Kan, 2009; Bulgarov and Caragea, 2015) and unsupervised methods (Wartena et al, 2010; Liu et al, 2010; Wan and Xiao, 2008; Hasan and Ng, 2014), scores for recall and precision for this task are well below those of standard NLP tasks such as POS-tagging or Named Entity Recognition. This is due to a variety of difficulties faced when extracting keyphrases, including the inherent ambiguity of the task, flaws in evaluation measures (e.g., semantically identical keyphrases are judged as different), the over-generation of keyphrases, etc. One of the most pressing issues in AKE research is the lack of large test collections with multiple opinions. In this paper we aim to address that gap and thus provide initial answers to the still open questions in solving and evaluating AKE, e.g., *What is the agreement on keyphrases among multiple readers? How well do the keyphrase candidates generated using the standard candidate generation procedures match keyphrases assigned by annotators? How do supervised and unsupervised methods compare?*

Keyphrase extraction has a long history, used in libraries for archiving and cataloging purposes. In such a library setting, keyphrases are assigned by trained experts using detailed manuals and rules, such as the Anglo-American Cataloging Rules (AACR) in Encyclopedia of Library and Information Sciences (Bowman, 2003), or the German libraries' "Regeln für den Schlagwortkatalog (RSWK)".[1] However, the setting we consider in our work concerns AKE for popular media articles, where keyphrases will be used by a typically untrained (layman) audience. Thus, also annotators will be laymen, and the keyphrase setting is therefore less constrained and the phrase importance fairly open to interpretation. The key contribution of this paper is the creation of a new dataset (4 corpora of 1000-2000 documents each) and using it to provide a consistent performance comparison of common AKE strategies.

First, in Section 2, we describe the construction of our new set of large and diverse collections of documents annotated with keyphrases. Next, Section 3 gives an overview of common AKE techniques and presents several strategies to include context-dependent features into supervised models, leading to increased precision. In Section 4, we evaluate the performance of the presented supervised and unsupervised AKE techniques, which apply knowledge extracted from background corpora (e.g., under the form of topic models). We study the influence of the amount of training data on AKE performance and point out the relatively low annotation effort needed to train competitive supervised models. In Section 5 we conclude by providing readers with guidelines to keep in mind when researching AKE and evaluating new techniques.

2 Test Collections

The state-of-the-art in AKE is not only diverse in terms of techniques (see further, Section 3), but also in terms of test collections used for evaluation. Indeed, these vary from formal

[1] `http://www.dnb.de/DE/Erwerbung/Inhaltserschliessung/rswk.html`

scientific articles (Augenstein et al, 2017) to more popular content such as mainstream news, or even blogs and tweets (Zhao et al, 2011). Issues with these evaluations are that (i) most collections are fairly limited in size (typically a few hundred documents) and (ii) annotations substantially vary from one collection to the next, since they are performed by either the various authors of the content or a single reader assigning keyphrases to many different documents, with possibly different annotation guidelines or goals from one collection to the next. As Frank et al (1999) noted, "for scientific articles the authors do not always choose keyphrases that best describe the content of their paper, but they may choose phrases to slant their work a certain way, or to maximize its chance of being noticed by searchers." Due to these limitations, and with the existing collections for AKE, it is hard to study how AKE performance may be impacted by annotators and the type or topic of documents.

We set out to systematically construct a rich and diverse set of annotated test collections to investigate that issue. We particularly focus on rather popular content targeted to a diverse, layman audience (e.g., as opposed specialist scientific literature). Our newly created set of test collections (i) is substantial in size with four different collections of 1200–2000 annotated documents each, (ii) comprises different types of news content (online news, online sports, lifestyle magazines, newspaper articles), and (iii) has each document annotated by multiple annotators (on average 6 per document), where annotator guidelines are identical for all collections (i.e., annotators are only informed with the definition and purpose of the keyphrases, regardless of the collection the documents are drawn from).

These collections are available for research purposes.[2]

2.1 Document Collection

In order to procure the test collections, we started from a large collection of candidate documents provided by three major Belgian media companies—each with their own distinct type of content (all in Dutch). The first media company involved, is the public-service broadcaster VRT, who offered two collections: *Online News* and *Online Sports*. The *Online News* collection is a subset of the texts accompanying the videos on its official news channel website De Redactie.[3] Similarly, the *Online Sports* collection represents their specialized sports section Sporza.[4] The second company, Sanoma, is a publishing group owning a selection of lifestyle, fashion, and health magazines, from which we created the *Lifestyle Magazines* test collection. The third company, Belga[5], offers a digital press database comprising content from all Flemish and Dutch newspapers, represented in our *Printed Press* set.

To verify that these collections indeed contain different topics, we use an external multi-label document classifier[6] trained on documents annotated with IPTC media codes[7] to gain insight into the thematic subjects covered. The average contributions of IPTC codes (at the first of three levels in the IPTC codes) are shown in Table 2, and confirm our intuition as humans being familiar with the various document collections: a large focus on sports texts

[2] For information regarding acquiring the test collections, please contact the paper's first author.

[3] `http://www.deredactie.be`

[4] `http://www.sporza.be`

[5] `http://www.belga.be`

[6] Our multi-label classifier is based on methods from top submissions in the "Greek Media Monitoring Multilabel Classification" (`https://www.kaggle.com/c/wise-2014`) and "Large Scale Hierarchical Text Classification" (`https://www.kaggle.com/c/lshtc`) hosted by Kaggle.

[7] `https://iptc.org/standards/media-topics/`

# Annotators	357
# Documents with $\geq$ 1 annotation	7342
max. Annotators/Document	10
min. Annotators/Document	1
$\oslash$ Annotators/Document	6
$\oslash$ Articles/Annotators	140

Table 1 Descriptive statistics regarding the annotations (# = number of; $\oslash$ = average).

for the *Online Sports* collection, mostly political subjects in *Online News*, lifestyle topics in *Lifestyle Magazines* and general news (dominated by sports) in *Printed Press*.

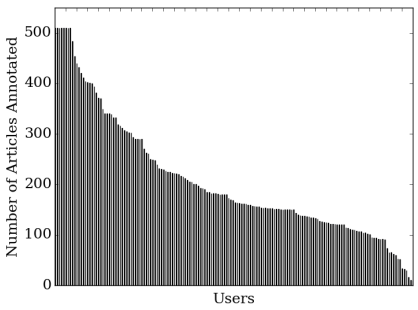
2.2 Collecting Keyphrases

Documents were presented to a panel of 357 annotators of various ages and backgrounds (selected and managed by imec.livinglabs[8]), who were asked to "*select a limited number of short phrases that concisely summarize the document's contents*". Three annotation sessions, of each spanning two weeks, were organized. Annotators were allowed to participate in multiple sessions. For each session an annotator was assigned 140 articles but was not obligated to finish the complete assignment. Compensations were awarded at 60, 100 and 140 articles. The amount of documents annotated by each of the 357 annotators is shown in Figure 1. To ensure overlap, each document was included in ten different annotators' task lists. Depending on the annotators' effort each document received at least one and up to ten different opinions. The final distribution of overlap per document is shown in Figure 2. Overall, 26% of the documents received more than 8 opinions, other descriptive statistics on the annotator panel are shown in Tables 1 and 2. As briefly mentioned in the Introduction, our annotation setup is quite different from traditional keyphrase annotation scenarios, where typically a small number of well-trained assessors provide annotations according to strict rules. In our setting, there was a large number of annotators, sampled from the Flemish media audience, which forms the target group for applications built on the extracted keyphrases. Also, we did not impose strict annotation rules, and instead propose that the results reflect the expectation of what keyphrases should look like for the target audience. The simplicity of our setup could be an important advantage for organizations intending to build a keyphrase extraction system on their own data. On the other hand, lack of strict rules also implies potential issues of disagreement on the chosen keyphrases among different annotators (see Section 2.5).

2.3 Annotation Tool

A web application was built for the test panel to perform annotations using a web browser from home. The annotation process works as follows. Annotators log in to the application using a personalized password. Each annotator was then directed to a briefing on the meaning and purpose of keyphrases and how to use the application to enrich articles with keyphrases. These guidelines are explained in detail in the following Section. The application chooses an article from the total collection stored in a database and presents it to the annotator. Articles

[8] `https://www.iminds.be/en/succeed-with-digital-research/go-to-market-testing/proeftuinonderzoek`



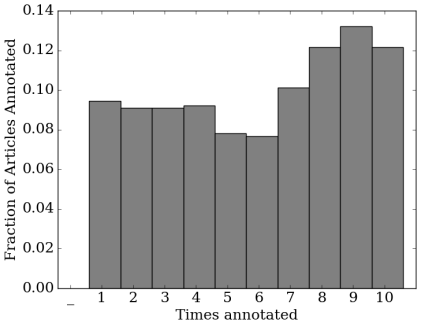


Fig. 1 Amount of annotated documents per annotator. **Fig. 2** Distribution of overlap per document.

are selected to increase overlap of annotators per document as fast as possible. A first document gets ranked first in the task list of 10 different annotators, a second document is then ranked second for these ten annotators. This is repeated until each annotator received a task list of 140 articles to be annotated during the two week session. Keyphrases are selected by sliding the mouse pointer over a selection of words. Unlike many keyphrase extraction tasks where authors assign free-form phrases to a document, this means keyphrases are guaranteed to appear in the text of the document. A theoretical upper bound for an extractor solely from the text thus would be 100%. One of the reasons this is often imposed, is the intended use of the keyphrases to highlight the most important phrases in the articles themselves. While this confines the task in certain sense, as some key concepts do not explicitly appear in the text, the extraction of keyphrases in unrestricted forms is a problem requiring different strategies, we thus should not penalize a keyphrase extractor for not being able to recognize these keyphrases. Documents are tokenized before annotation and annotators' highlighting is confined to token boundaries to facilitate annotating and reduce matching errors afterwards. The annotator is then prompted to add the selection as keyphrase, after which the keyphrase is shown in a list next to the article with other assigned keyphrases.

All keyphrases are highlighted after selection. Figure 3 shows the application as displayed in the web browser. The annotator is also provided with a button to send a form to provide feedback on a specific article, e.g., to indicate confusing cases or articles not suitable for annotation (such as articles featuring two entirely different topics or stories, tables of sports results, cooking recipes, etc.).

2.4 Keyphrase Annotation Guidelines

The full guidelines section of the annotation tool is shown in Figure 4. Next to guidelines, videos and multiple examples of annotated documents were provided to instruct the annotators. We keep to the standard definition of keyphrases (Turney, 2000) and impose no constraints on keyphrase form. Whereas the annotation procedure differs from one collection in literature to the next, all of the currently presented test collections were created the same way. No prior limits or constraints were set on the amount, length, or form of the keyphrases. This allows us to study the interpretation by annotators and their disagreement, as well as investigate the form of typical keyphrases that candidate generation approaches should produce. Figure 5 shows an example of a (translated) Dutch lifestyle article about music artist *Anastacia* and her recovery from breast cancer, where bold text represents keyphrases as

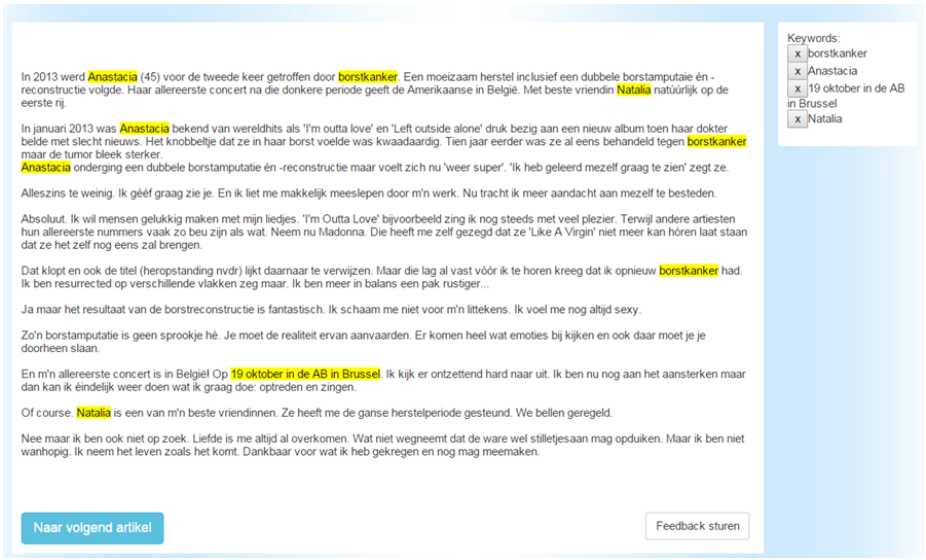


Fig. 3 Web interface for annotation of keyphrases.

annotated by 10 different annotators. A superscript of i indicates the identifier of the annotator who selected this phrase as keyphrase. This example is a first demonstration of the lack of consensus on keyphrases: only few keyphrases are selected by all of the annotators. We expand on this disagreement issue in Section 2.5.

In Table 2 we present descriptive statistics on the length of the documents, the amount of assigned keyphrases, the amount of keyphrase candidates per document (candidate generation is presented in the following section), the entities per document, the distribution over n-grams in keyphrases, predicted topics and POS-tags, for the four different test collections. As Table 2 indicates, the largest difference between the test collections is their thematic content. Articles from the collections are relatively short, with *Printed Press* featuring slightly longer articles than the three other collections. *Online Sports* articles contain more entities, with notably more entities that are seen as keyphrase by the annotators. On average, a single annotator assigns 5 keyphrases to each document. The union of all annotations per document on average contains 15 keyphrases.

2.5 Annotator Disagreement

Multiple annotations for each document show that the notion of "*what is a keyphrase?*" remains subjective. Figure 6 shows the fraction of annotated keyphrases for different ratios of overlap by the complete set of annotators. This shows that the largest fraction of all keyphrases ($\geq 50\%$) are selected by less than 20% of all the annotators that assigned keyphrases to the document. This is due to different interpretations of the article, but also due to keyphrases with equal semantics appearing in different forms. This has important consequences for training models on keyphrases annotated by a single annotator: in such a setting, many alternate candidate phrases that other annotators would pick, would be considered as negative training data. The performance on evaluation sets can thus greatly vary depending on the annotator of the test set. Studying disagreement and the effect of training

Dear Annotator,
Thank you for participating in the Steamer Bootcamp! The next 2, 4 or 6 weeks, you will read a lot of news reports and select the most important keywords or keyphrases in them.

Your aid is of major importance
Depending on your choice of keywords, we will develop a system that automatically recognizes these keywords and adds them to documents. These keywords are not only very useful for many applications, they also make it easy for search engines to improve the automatic recommendation of other relevant articles -for you-.

Keyphrases?
The keywords or 'keyphrases' are defined as "a selection of short, significant expressions consisting of one or more words that can summarize the article very compactly."

Too complicated?
Below you find some videos back with a quick guide.
<Link to instruction video> Reading articles.
<Link to instruction video> Indicating keywords.

Method
There are some tips to get to the best keywords:
Ask yourself, "What words summarize the content of the article?" Or "What words are most representative of the contents of the article?" This can be an event or an object, the crucial entities, or organizations that are mentioned in the article. Try to keep the keyphrase as short as possible. Words that do not contribute may be omitted to the meaning of the keyphrases. The number of keywords per article depends largely on the length of the article and the various topics discussed in it. It is rare to select more than 10 keywords per article.

We demonstrate this with an example:
"Higher education is bracing itself. Once it had ample offer, now it calls for a economization of supply. The Flemish coalition agrees that the universities themselves must make proposals to achieve a constrained and transparent offer. In interviews, the Minister of Education, Hilde Crevits (ISA), indicates that the offer can be safely pruned to one hundred of majors. "
in this example "higher education" , "economization of supply" and " Hilde Crevits" would be appropriate keywords.

Still not clear?
Click <Link to more examples> for more examples.

Select keywords or keyphrases
You select a keyphrase by clicking a phrase in its first word and sliding to the last word in the keyphrase. The tool will then ask if the selected keyword should be added. Afterwards all selected keywords are displayed in the left column. If you've changed your mind, then a chosen keyword can be removed by clicking on the red cross. Think you have selected all the keywords? Save the article and go to the next article.

Ready?
Then you can begin! The more articles you read, the faster you will start to find the keywords. It's a little hard at first, but hang in there, it gets better.
Click "Start" and you can start!

Any questions?
Take a look at our FAQ page or use the feedback button. Good luck!
Greetings,
The Steamer research team

Fig. 4 Instructions shown at the main page of the keyphrase annotation tool.

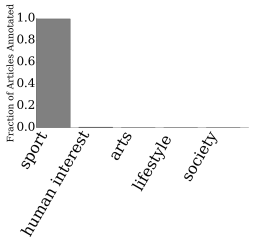
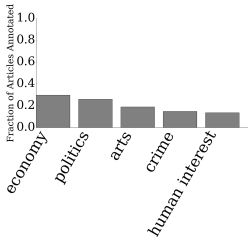
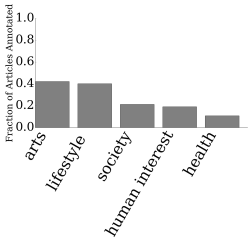
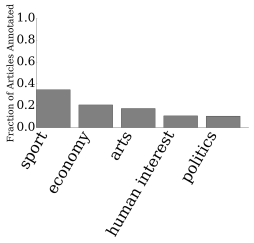
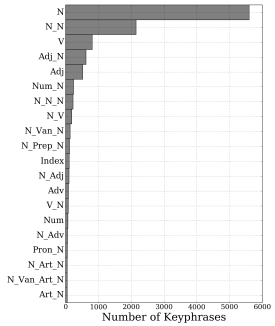
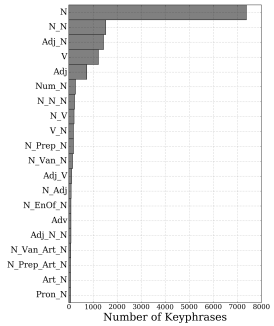
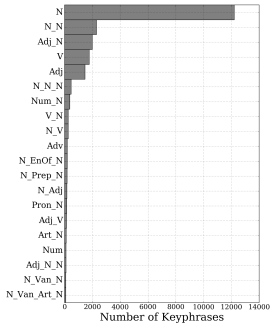
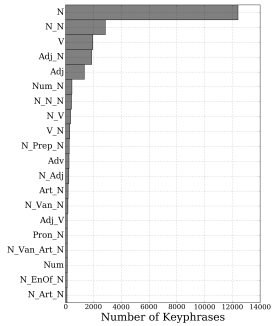
Name	Test Collections: Online Sports	Online News	Lifestyle Magazines	Printed Press
Type (date range generated content)	Sports Articles (2012-2014)	Television News (2012-2014)	Fashion, Lifestyle… (2010-2013)	News Articles (2009-2014)
# Documents	1,252	1,259	2,202	2,196
# Keyphrases	14,544	19,340	29,970	31,461
⊘ Keyphrases/Annotator	4.6	5.7	4.7	4.8
⊘ Keyphrases/Document	11.6	15.4	13.7	14.4
⊘ Tokens/Keyphrase	2.1	2.3	2.0	1.8
⊘ Tokens/Document	288	332	284	399
⊘ Tokens/Candidate phrase	1.8	2.0	1.9	1.5
⊘ Candidate Keyphrases/Doc.	43	52	49	67
⊘ Entities/Document	10.6	5.7	4.1	8.3
⊘ Entities/Keyphrases (%)	30.6	12.7	13.4	18.1
1/2/3/3+ -gram distribution (%)	52 / 35 / 8 / 5	55 / 27 / 9 / 9	58 / 25 / 9 / 8	57 / 28 / 8 / 7
Max. POS-filter Recall (%)	65.9%	65.5%	59.5%	64.2%
IPTC Theme Distribution				
Distribution of Keyphrase POS-Tags				

Table 2 Corpus statistics for the four annotated keyphrase datasets used in this paper. We describe the amount of documents and keyphrases, average length of the documents and the keyphrases, the tokens in the keyphrases, the average amount of keyphrases assigned to documents, the distribution of keyphrases over n-grams, total and average amount of entities present in keyphrases. Plots show the distribution of the topics detected in the collection by a multi-label classifier and the distribution of POS-tag sequences of keyphrases.

In 2013 **Anastacia**[1,2,3,4,5,6,7,8,9,10](45) was struck for the second time with **breast cancer**[1,2,3,4,5,6,7,8,9,10]. A difficult recovery, including a double **mastectomy**[4] and reconstruction followed. She gave her **first concert**[2] in Belgium after this dark period, with her best friend **Natalia**[3,4,8,9,10] in the front row. In January 2013 Anastacia was known for world hits like '**I'm outta love**'[3] and '**Left Outside Alone**'[3]. Busy working on a new album, her doctor called with bad news. The lump she felt in her breast was cancerous. Ten years earlier she had already been treated for breast cancer, but the tumor appeared stronger. Anastacia underwent a **double mastectomy**[10] and reconstruction, but now feels "great again". "I've learned love to see myself," she says. "I am happy, you see. And I let myself be carried away by my work easily. Now I try to pay more attention to myself. Absolutely. I want to make people happy with my songs. 'I'm Outta Love', for example, I still sing that with pleasure. While other artists are often tired of their first songs. Take Madonna. Who said herself that she can't hear 'Like A Virgin', let alone sing it herself. That's right, and even the title ('**Resurrection**[7]', ed) seems to refer to it. But this was already determined before I was told I had cancer again. I am resurrected in different areas, so to speak. I'm more balanced, a lot calmer ... Yes, but the result of the **breast reconstruction**[5] is fantastic. I'm not ashamed of my scars. I still feel sexy. Such a **mastectomy**[8] is not a fairy tale, huh. You must accept the reality. There are a lot of emotions involved, and you have to beat you through. And my **first concert**[3,10] is in **Belgium**[2] On October 19, in the AB in Brussels I can finally do what I love to do: act and sing. Of course, Natalia is one of my best friends. She has supported me the whole recovery period. We call regularly. No, but I'm not searching. Love has always happened to me. Which does not mean that true love may turn up little by little. But I'm not desperate. I take life as it comes and I am grateful for what I've got."

Fig. 5 Example of annotated article, with indication of keyphrase annotations by 10 different annotators using superscripts.

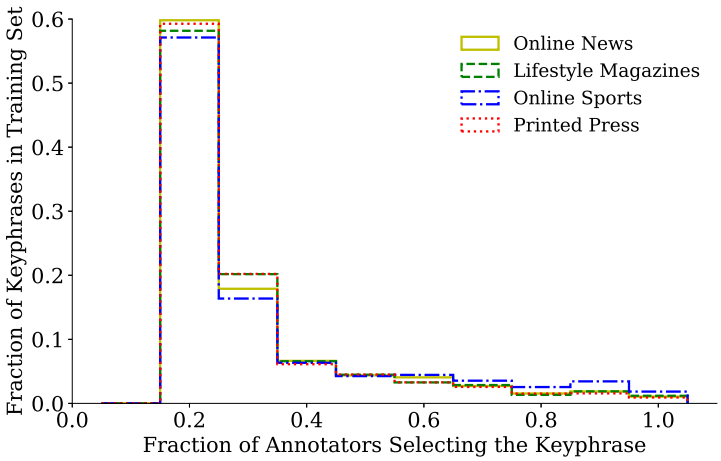


Fig. 6 Illustration of annotator disagreement on keyphrases. The X-axis shows the fraction of annotators that agree on selecting a single keyphrase for a given document. For example, if we were to restrict keyphrases to those selected by 50% of the annotators, this shows that we would retain less than 5% of all keyphrases.

data by different annotators on the evaluation confidence is a valuable direction for future research. In our evaluation of automatic extractors, we use the aggregated set of keyphrases as a reference but also report on the average and standard deviation on scores for different reference keyphrase sets assigned by different annotators.

As metric for inter-annotator disagreement we report the Fleiss' kappa score (Fleiss, 1971). We define a Fleiss' kappa, κ_F, for each document as

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \quad (1)$$

Collection	Online Sports	Online News	Lifestyle Magazines	Printed Press
Avg. κ_F per doc.	0.235	0.189	0.193	0.186

Table 3 Annotator agreement on keyphrases, quantified with Fleiss' kappa κ_F, is quite low.

Here, $1 - \bar{P}_e$ measures the degree of agreement that is attainable above chance, and, $\bar{P} - \bar{P}_e$ measures the degree of agreement achieved above chance. For these formulas, we consider the N generated keyphrase candidates as rated items, that are scored by each of the n annotators with one of $k = 2$ possible scores, to represent the cases where a given phrase is annotated as a keyphrase or a non-keyphrase by a given annotator. To find $\bar{P}$ and $\bar{P}_e$, first p_j, the proportion of all assignments as keyphrase ($j = 1$) or non-keyphrase ($j = 2$), is calculated:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}. \tag{2}$$

Then P_i is calculated, the extent to which annotators aggree on the i-th keyphrase candidate.

$$P_i = \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^{k} n_{ij}^2 \right) - n \right]. \tag{3}$$

$\bar{P}$ is the mean of the P_i's and $\bar{P}_e$ is the mean of the squared p_j's, which are then used to calculate κ_F.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i, \quad \text{and} \quad \bar{P}_e = \sum_{j=1}^{k} p_j^2. \tag{4}$$

Median fleiss kappa across all documents is low at 0.19 with slightly higher values for *Online Sports'* articles. As presented in Table 3, higher agreement for sports articles might be due to entities being central to many of these articles. The annotators were recruited from among the target audience in the Flemish media landscape by experienced Living Lab researchers (Lievens et al, 2014), including both residential consumers and media professionals. For the experiments presented in this work, we pooled all annotators. However, the dataset contains additional information that would allow making a distinction between different types of annotators, in order to study different use cases. For example, we quantified the difference in annotation behavior for the most active half versus the least active half of the annotators. The 50% most active annotators are responsible for 82% of all sets of keyphrases, assigning on average 4.1 keyphrases per document, each on average 2.2 tokens long. The other half assigned on average 5.2 keyphrases per document, but slightly longer ones, with on average 2.9 tokens.

3 Keyphrase Extraction Techniques

This section provides a brief overview of common AKE techniques. After explaining how candidate keyphrases are selected (Section 3.1), and how well these common heuristics cover the annotated keyphrases, the most prominent unsupervised methods are introduced (Section 3.2). Next, supervised keyphrase extractors and feature design for AKE are presented (Section 3.3).

3.1 Candidate Selection

To avoid spurious keyphrase instances, and to limit the number of candidates, extractors choose a subset of phrases which are selected as candidate keyphrases. Especially for long documents, the resulting list of candidates can be long and hard to rank. Current state-of-the-art mainly adopts part-of-speech (POS) filters, typically after stopword removal. Other heuristics for selecting candidates only allow keyphrases from a curated, fixed list (Medelyan and Witten, 2002) or Wikipedia article titles (Grineva et al, 2009), which drastically reduces the amount of possible candidates. Here, we quantitatively address the open question as to what extent such POS-filtered keyphrases correspond to those assigned by human annotators. For that purpose, we calculated the measure "Maximum POS-filter Recall" shown in Table 2 for each collection, defined as the recall attained by the most common POS-filter based on the rules defined in (Kim and Kan, 2009) over the set of human annotator keyphrases. More specifically, the filter is defined by the following regular expression[9]:

$$(<\text{Adj}|\text{Num}>^* <\text{N}>+<\text{IN}|\text{Van }>)?<\text{Adj}|\text{Num}>^*<\text{N}>+ \quad (5)$$

Applying this common filter to our data sets shows that, if we were to use it to select candidate phrases from the text, we would maximally reach a recall of about 66% when considering *all* the annotated keyphrases as gold standard, as listed in Table 2. This relatively low coverage demonstrates the mismatch between POS-filters and the interpretation of the keyphrase concept by the (layman) annotators. POS-filters also extract longest matching sequences of tokens, while keyphrases might be subsequences. Figure 7 shows the complete distribution of POS patterns assigned to the annotated keyphrases versus those of extracted candidates by the POS-tagger. While the majority of keyphrases are Noun Phrases, it is shown that a considerable fraction of keyphrases are not extracted by the standard POS-filter, such as lone verbs and adjectives. This indicates that people also tend to see actions or events, denoted by a verb, central to an article's content. A topic for future research is to maximize the coverage of keyphrases by candidates while limiting the total amount of extracted candidates, i.e., , the trade-off between recall (as the maximum achievable amount increases), and precision (as keyphrase extraction becomes harder with more candidates to rank correctly). What POS-filter is the most effective (for optimal recall versus precision) also depends on the type of document: while news articles describe events, typically requiring entities and verbs as keyphrases, we expect this to be less the case for scientific articles where domain specific technological terms are more common. We advise to adapt the candidate generation procedure appropriately.

3.2 Unsupervised Keyphrase Extraction

A disadvantage of supervised approaches is the requirement of training data and the resulting bias towards the domain of the training data, undermining their ability to generalize well to new, unseen domains. This limitation is bypassed by unsupervised approaches that focus on word-frequency or centrality in graph transformations (Mihalcea and Csomai, 2007; Liu et al, 2010; Sterckx et al, 2015a,b). Note that unsupervised approaches have been reported as state-of-the-art on many test collections (Hasan and Ng, 2014). Because most of these test collections do not supply a training and test data split, comparisons of these models

[9] POS-tag definitions used here: Adj = adjective, N = nouns (including singular and plural), IN, Van = preposition or subordinating conjunction and Num = quantity expressions.

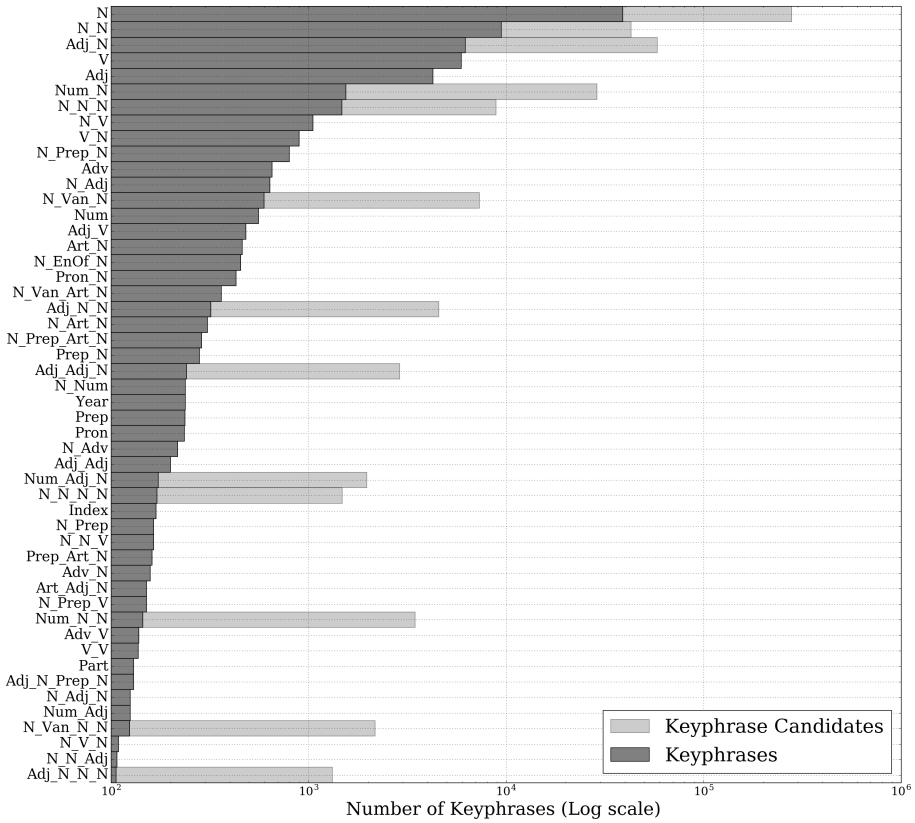


Fig. 7 POS-tags of extracted keyphrase candidates by filters versus complete distribution of all the annotated keyphrases from all collections.

with supervised models is missing. The most important unsupervised AKE approaches are (variations on) the following baseline methods:

- **TF*IDF** (Salton and Buckley, 1988) is the most common strategy for ranking keyphrases for a given document in both unsupervised and supervised models (Grineva et al, 2009; Zhang et al, 2005). The TF*IDF weight consists of two factors: TF is the frequency of the considered keyphrase. The second factor, IDF, is the Inverse Document Frequency, computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific phrase appears. The IDF factor is incorporated to diminish the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.
- **TextRank** is a completely within-document AKE technique (Mihalcea and Tarau, 2004), which represents the document as a graph. Each word corresponds to a node in the graph, edges are created between words co-occurring within a window of pre-defined width. Centrality of the nodes is then calculated using PageRank. Keyphrases are generated using high-scoring nodes and by merging co-occurring terms. In our evaluation we use the **SingleRank** variant (Wan and Xiao, 2008), in which edges are weighted according to the number of times they co-occur within the window. The score for word w_i is

computed iteratively until convergence using the recursive formula:

$$S(w_i) = \lambda \cdot \sum_{j:w_j \to w_i} \left(\frac{e(w_j, w_i)}{O(w_j)} \cdot S_z(w_j) \right) + (1 - \lambda) \tag{6}$$

where $S(w_i)$ is the PageRank score for word w_i, $e(w_j, w_i)$ is the weight of the edge $(w_j \rightarrow w_i)$, the number of outbound edges is $O(w_j) = \sum_{w'} e(w_j, w')$ and λ is a damping factor $\in [0, 1]$ indicating the probability of a random jump to another node in the word graph.

– **Topical PageRank**, as described in Liu et al (2010), calculates a PageRank score separately for each topic in a pre-trained topic model and boosts the words with high relevance to the corresponding topic. That topic-specific PageRank score for word w_i is defined as follows:

$$S_z(w_i) = \lambda \cdot \sum_{j:w_j \to w_i} \left(\frac{e(w_j, w_i)}{O(w_j)} \cdot S_z(w_j) \right) + (1 - \lambda) \cdot P_z(w_i), \tag{7}$$

where $S_z(w_i)$ is the PageRank score for word w_i in topic z. A large $S_z(w_i)$ indicates that w_i is a good candidate keyword in topic z. The topic specific preference value $P_z(w_i)$ for each word w_i is the probability of arriving at this node after a random jump, thus with the constraint $\sum_{w \in \nu} P_z(w) = 1$ given topic z. In TPR, the best performing value for $P_z(w_i)$ is reported as being the probability that word w_i occurs given topic z, denoted as $P(w_i|z)$. This indicates how much that topic z is focused on word w_i. With the probability of topic z for document d $P(z|d)$, the final ranking score of word w_i in document d is computed as the expected PageRank score over that topic distribution, for a topic model with K topics,

$$S(w_i) = \sum_{z=1}^{K} S_z(w_i) \cdot P(z|d). \tag{8}$$

We apply the more efficient, equally effective, single-PageRank variant proposed in (Sterckx et al, 2015a). Other graph based methods using background information are based on relatedness between candidates in the document in thesauri (Gazendam et al, 2010).

3.3 Supervised Keyphrase Extraction

Supervised methods recast the extraction problem as a binary classification task, where a model is trained to decide whether a candidate phrase (generated from the candidate generation procedure discussed in Section 3.1) is a keyphrase or not (Turney, 2000; Frank et al, 1999; Hulth, 2003). Treating automatic keyphrase extraction as a supervised machine learning task means that a classifier is trained using documents with known keyphrases. While the decision is binary, a ranking of phrases can be obtained using classifier confidence estimates, or alternatively, by applying a learning-to-rank approach (Jiang et al, 2009).

3.3.1 Feature Design and Classification

An important aspect of supervised approaches is feature design. In previous work, many features have been designed and reported as being effective on different occasions (Turney, 2000; Hulth, 2003; Park et al, 2002; Kim and Kan, 2009; Lopez and Romary, 2010; Bulgarov and Caragea, 2015). In these studies, several types of features can be distinguished:

- **Statistical Features:** Features such as the term frequency, TF*IDF (discussed in the following section) and keyphraseness (the total amount of times a keyphrases occurs in a training collection).
- **Structural Features:** Features characterizing the position of a term with respect to the document structure (first location, last location, occurrence in title, etc.),
- **Content:** Features characterizing the keyphrase, such as the lexical cohesion of the keyphrase (Dice Coefficient of the tokens in the keyphrase and the complete keyphrase), length, the POS-pattern, capitalization, etc.
- **External Resource Based Features:** information is added using external resources or dictionaries such as terminological resources (Medial Subject Headings (MeSH), the Gene Ontology, etc.), linguistic resources (WordNet), thesauri (Gazendam et al, 2009), Wikipedia, topic models, or tags from a Named Entity Recognizer.

After features are extracted, a learning algorithm is applied on the training collection to distinguish keyphrases from non-keyphrases. Many different statistical classifiers have been applied for this task, including Naive Bayes (Witten et al, 1999), bagging (Hulth, 2003), max-entropy (Yih et al, 2006), multilayered perceptron (Lopez and Romary, 2010),support vector machine (Jiang et al, 2009) and (boosted) decision trees (Lopez and Romary, 2010). A detailed comparison with each of the designed features in existing supervised techniques is not in the scope of this paper. As for many supervised machine learning tasks, a classifier needs to be developed, evaluated and separately optimized for each collection (also known as the *no-free-lunch theorem* (Wolpert and Macready, 1997)). We propose a different approach and develop a baseline supervised extractor, compare with unsupervised techniques, propose several features modeling the context of the document, and study the influence of the background collection on the AKE effectiveness.

3.3.2 Supervised Model

As baseline supervised keyphrase extractor, we extract a number features based on prior work presented in the previous section. Features effective during development and used in the baseline model are: (i) keyphrase frequency, (ii) number of tokens in the keyphrase, (iii) length of the longest term in the keyphrase, (iv) a binary feature which indicates whether the keyphrase contains a named entity, (v) relative position of the keyphrase's first occurrence in the full article, (vi) relative position of last occurrence and (vii) span (relative last occurrence minus relative first occurrence).

Apart from features extracted from the document the keyphrase appears in, we calculate two features based on background corpora: (viii) TF*IDF ($f_{\text{TF*IDF}}$) and (ix) Topical Word Importance (f_{LDA}), which is based on context. TF*IDF consists of the multiplication two factors: TF is the frequency of the considered keyphrase relative to the document length. The second factor IDF is the Inverse Document Frequency, computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific phrase appears. Topical word importance was introduced in Sterckx et al (2015a) and is based on topic modeling, for which we use standard Latent Dirichlet Allocation (Blei et al, 2003). Topical word importance is the similarity between the word-topic probability and the topic-probability from a topic model trained on the background corpora. This similarity, as a feature for a word-document pair (w, d), is determined as the cosine distance between the vector of topical word probabilities $\mathbf{P}(w|Z) = [P(w|z_1), \ldots, P(w|z_K)]$ and the document topic probabilities, $\mathbf{P}(Z|d) = [P(z_1|d), \ldots, P(z_k|d)]$:

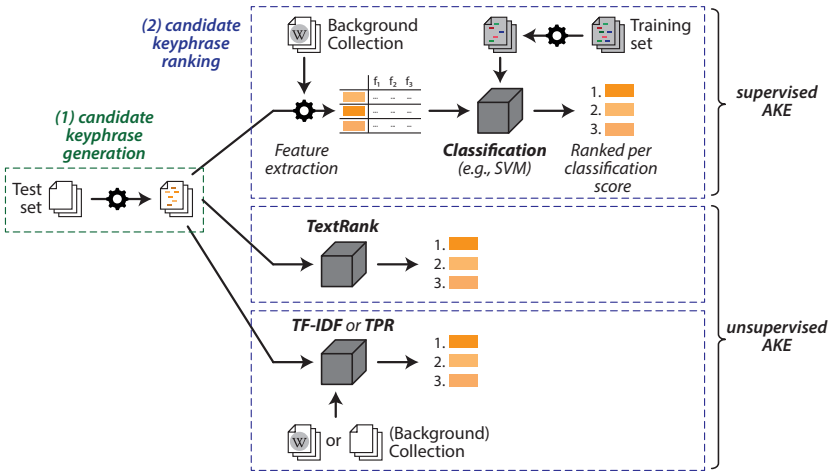


Fig. 8 Schematic representation of the experimental set-up.

$$f_{\text{LDA}}(w,d) = \frac{\mathbf{P}(w|Z)\cdot\mathbf{P}(Z|d)}{\|\mathbf{P}(w|Z)\|\cdot\|\mathbf{P}(Z|d)\|}. \tag{9}$$

This score is usually included as a weight in a biased graph-ranking algorithm, here we also include it as contextual feature. Both context dependent features stem from unsupervised techniques. Features established in literature, but which were found to be ineffective include *keyphraseness* and *keyphrase cohesion* (Lopez and Romary, 2010). In all experiments, we use a support vector machine (SVM) classifier with a linear kernel from the *libsvm* library (Chang and Lin, 2011) and gradient boosted decision trees implemented in the *XGBoost* package (Chen and Guestrin, 2016). For token-based features we use the sum and average of the TF*IDF and Topical Word Importance values of the tokens constituting the keyphrase. IDF values and topic models are trained on large background collections stemming from the same source and on a more *general* Wikipedia corpus.

4 Systematic Evaluation

In this section, we describe our experimental set-up, followed by an evaluation of different unsupervised and supervised models, with a careful study of the effect of multiple opinions during evaluation and training.

4.1 Experimental set-up

Creating training and test set

In Figure 2, we showed the amounts of annotations per document. Before separating documents into train and test collections, we remove documents annotated by less than five annotators. From documents with more than five opinions we randomly select five opinions

	Collections			
Name	Online Sports	Online News	Lifestyle Magazines	Printed Press
# Background Documents	325,438	325,437	976,318	976,316
# Training Documents	312	275	981	957
# Test Documents	500	500	500	500

Table 4 Number of documents in training and test collections after filtering documents having fewer than five opinions.

as reference annotations. This is to avoid bias towards more frequently annotated documents, while keeping a reasonable amount of documents to produce meaningful results. From these filtered collections of documents, we sampled 500 documents as test collection for each of the sub-collections, the remaining ones were used for training and development. The amounts of training and test documents are shown in Table 4. The amount of training documents ranges from around three hundred to close to a thousand for the different sub-collections, but as will be shown in Section 4.5, the effectiveness of the studied supervised approaches saturates beyond a few hundreds of training documents. All annotated documents for the different sub-collections, with listings of document IDs in the training and test collections, as well as extracted candidate keyphrases are made available upon request for research purposes.[10]

System architecture

Figure 8 outlines the experimental set-up. After filtering the annotated data and separating train and test set, all documents were further pre-processed by extracting POS-tags using the rule-based Fast Brill Tagger (Brill, 1992), implemented in the NLTK (Bird, 2006) package trained on the CONLL-2002 training set (accuracy of about 95% on the CONLL test sets). Keyphrase candidates were extracted using the POS-filter presented in Section 3.1.

For the supervised models, features for each of the candidate phrases and keyphrases (as detailed in Section 3) were calculated, and the models were trained on the training subsets.

Contextual information (i.e., IDFs, 1,000-topic LDA-models) was derived for each of the collections individually from a non-annotated background corpus provided by the corresponding media company (ranging from 325,000 documents (*Online Sports, Online News*) to 976,000 (*Printed Press, Lifestyle Magazines*)), as well as from a more universal background corpus, i.e., a 2014 Dutch Wikipedia dump. The Dutch Wikipedia corpus contains 1,691,421 articles, amounting to a total of 226,080,236 tokens.

When processing the documents in the test collection, each AKE approach provides a confidence score to the phrases generated during the candidate generation procedure. For unsupervised models this is the TF*IDF score or PageRank score, whereas for supervised models this is the predicted score, i.e., probability of the candidate being a keyphrase. The candidate keyphrases of the test documents are ranked according to these scores (shown in orange in Figure 8). For converting predicted scores to binary decisions on whether or not to retain the keyphrases, the cut-off scores with highest F_1 score on the training set are used (also for the unsupervised approaches).

Hyperparameter tuning was kept to a minimum, using standard values for the unsupervised graph algorithms. For wordgraph algorithms, Textrank and Topical PageRank, length of the sliding window was set to 10 tokens and a damping factor of 0.85 for TextRank and 0.7

[10] Due to copyright issues, the data cannot be published publicly: researchers only can obtain the data (including annotations and candidate keyphrases) after contacting the authors and signing a non-disclosure agreement.

for Topical PageRank was chosen. Development of classifiers was done by crossfolding the trainingdata four times. Hyperparameters for boosted trees are the number of trees and their depth, for the SVM classifier we tune the regularization. We optimized for micro-averaged F_1 scores on the held-out folds.

Note that some of the keyphrases assigned by the annotators are not extracted by the candidate generator. These are filtered out from the train set (as indicated by the processing step in the top right of Figure 8), to prevent classifiers from overfitting on forms of keyphrases that are not generated by the candidate generator. Such keyphrases that do not match the candidate generation pattern are however included in the ground truth set of keywords for the test collection, so for most documents a recall of 100% is not attainable.

4.2 Evaluation Setup using Multiple Opinions

Several options arise when evaluating keyphrases for documents with multiple opinions and depending on the goal or application of the keyphrases, different requirements are to be met or preferred by the keyphrase extractor's output. We propose two quite different evaluation scenarios, as well as a short motivation for both of them, followed by experimental results. As we will demonstrate, deciding on one of these scenarios strongly influences the scores and may lead to differently ranked keyphrase extractors.

A straightforward way is to create a reference set of keyphrases by pooling the annotated keyphrases by the different judges. The main advantage of this pooling approach is the increased robustness when measuring precision. Different annotators can select different representatives for identical concepts in an article. This way, the precision of the keyphrase extractor remains when making a specific choice of keyphrases, provided it covers the central concepts present in the reference set. However, this scenario suffers from two drawbacks. First, it does not penalize a possible lack of diversity among the predicted keyphrases, and second, it is hard to interpret the resulting metrics based on aggregated keyphrases from the point of view of a single annotator. This scenario is preferable when keyphrases are applied for visual purposes, e.g., to provide an overview of content by highlighting all keyphrases, or to get an estimate of the overall precision of the extractor regardless of redundancy in the output of the extractors. The second scenario, discussed next, avoids these drawbacks.

In the second scenario, each set of keyphrases from a specific annotator is treated as an independent target set. In this case, averaging over the obtained evaluation metrics corresponds to measuring the expected performance for a random annotator, if we can assume that the annotator population is represented by the set of annotators. This evaluation scenario comes closer to the purpose of summarization by keyphrases, which is a more common goal of keyphrases and and in line with the instructions given to the annotators. In this setting, extractors are rewarded for the extraction of small but diverse sets of keyphrases as annotated by different annotators. We perform this annotator based evaluation by averaging the scores over different annotators per document and measuring the standard deviation on this average.

To match predicted keyphrases with the reference keyphrases, we follow the traditional evaluation scheme for keyphrase extraction (Hasan and Ng, 2014) by exactly matching keyphrases from the golden answer set with those provided by the automatic extractors without stemming, and apply a standard rank- or set-based metric. We measure the micro-averaged precision for the 5 (precision@5) top ranked or most confident keyphrases. Note that 5 is approximately the average amount of keyphrases assigned by a single annotator to a document and is the number of keyphrases the content providers agreed to assign to

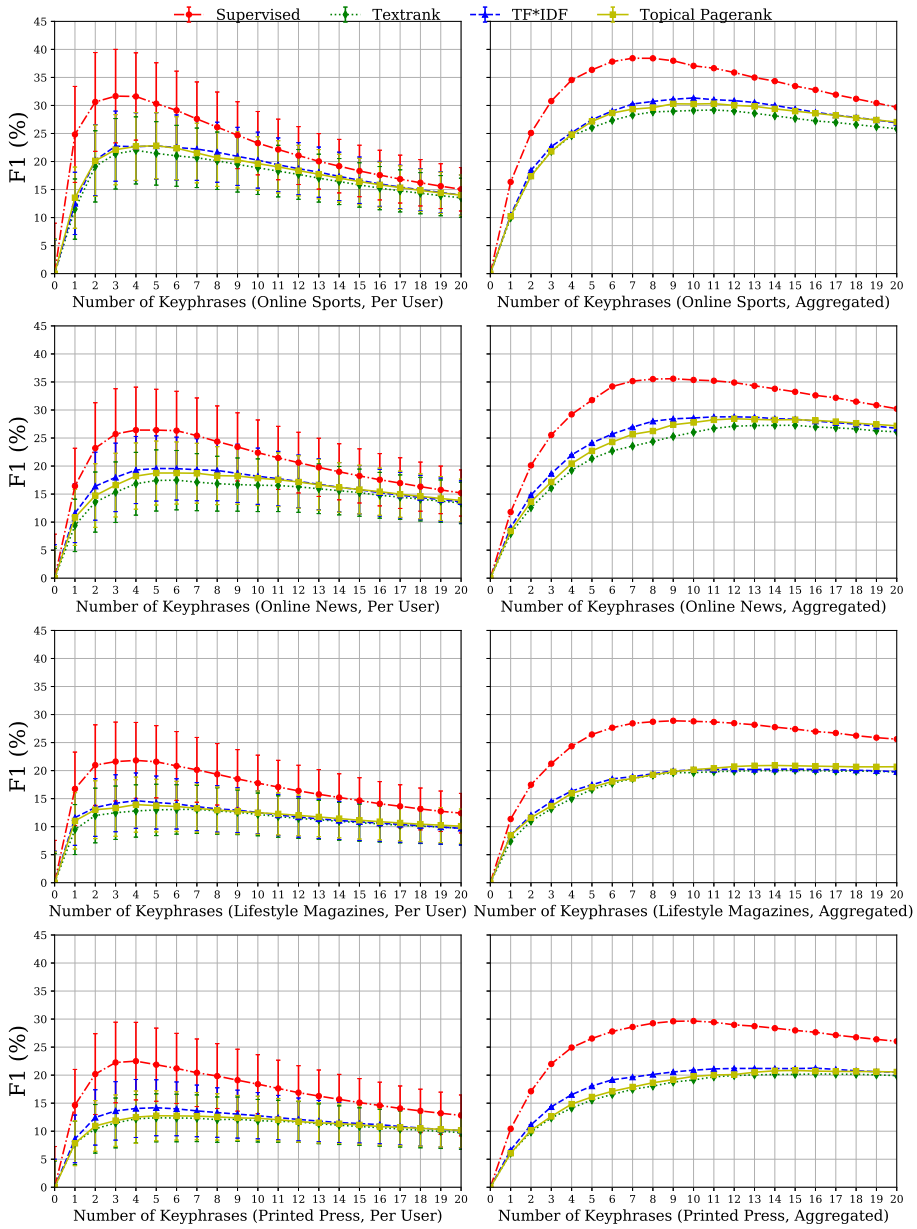


Fig. 9 Plots on the left show micro-averaged F_1 scores (with error bars showing standard deviation) for different fixed amounts of assigned keyphrases for different annotators. The right column shows the same models evaluated on aggregated collections of keyphrases.

each document. To evaluate from a set-based perspective, we measure the micro-averaged F_1 from precision and recall per document after setting a threshold (the same for all documents) on the confidence values predicted by the extractors that optimizes F_1 scores on the development set .

In Tables 5 and 6, we show results for these two different approaches to evaluation, i.e., the first scenario with aggregated target collections (*Aggr.*), and the second scenario with scores averaged per annotator (*Av.*$\pm$*Stdv.*) for the precision at 5 extracted keyphrases per document (precision@5) in Table 5 and F_1 scores at a tuned threshold in Table 6. We define precision, recall and F_1 as follows:

$$\text{precision} = \frac{|\text{annotated keyphrases} \cap \text{extracted keyphrases}|}{|\text{extracted keyphrases}|} \tag{10}$$

$$\text{recall} = \frac{|\text{annotated keyphrases} \cap \text{extracted keyphrases}|}{|\text{annotated keyphrases}|} \tag{11}$$

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{12}$$

In Figure 9, we plot F_1 as a function of a fixed amount of selected keyphrases per document to further illustrate the difference between these two ways of evaluating. Graphs in the left column show results for averaging over sets of keyphrases by the different annotators with standard deviations, graphs on the right show evaluations on the aggregated set of keyphrases.

4.3 Comparison of different techniques

A first observation is that, in terms of performance of different AKE approaches, *supervised* models using baseline features (see Section 3.3.2) outperform each of the standard unsupervised techniques on every test collection by a margin for the different metrics. These results highlight the (perhaps unsurprising) need for supervision and feature design. Models show improvement using contextual information ($f_{\text{TF*IDF}}$, f_{LDA}). Also for unsupervised models, techniques like TF*IDF and Topical PageRank which include background information generally perform better than those that do not, like TextRank. For statistical classifiers, the gradient boosted decision tree outperforms the linear classifier on each occasion.

For the TF*IDF and Topical PageRank models, using IDFs or topic models outperform those inferred from the more general Wikipedia background collection. This is less the case when they are used as features in the supervised models.

4.4 Comparison of different test collections

Between test collections there is a clear distinction in performance by keyphrase extractors for the different types of content. Precision@5 and F_1 scores for keyphrases predicted on *Online Sports* content can be up to 10% higher than those for *Lifestyle* and *Printed Press*. An explanation for this may be the focus on entities in these Sports documents, and higher annotator agreement for these documents. These types of keyphrases are covered well by candidate generators and are modeled well by the features. Scores for *Online News* are overall better than *Lifestyle* and *Printed Press*.

Test Collection→	**Online Sports**		**Online News**	
Model ↓	Aggr.	Av.±Stdev.	Aggr.	Av.±Stdev.
TextRank	39.6	18.8 ±5.1	37.5	16.3 ±5.2
TF*IDF	41.9	20.0 ±5.4	42.8	18.4 ±5.7
Topical PageRank	41.4	20.0 ±5.3	40.0	17.6 ±5.5
XGBoost	(57.9) 55.2	(27.5) 26.7 ±6.9	(56.6) 56.0	(24.3) 24.6 ±7.2
XGBoost + $f_{\text{TF*IDF}}$	(59.2) 56.2	(28.0) **27.0** ±**6.9**	(57.4) 56.0	(24.6) 24.9 ±7.2
XGBoost+ $f_{\text{TF*IDF}}$, f_{LDA}	(60.5) 55.8	(28.2) 26.9 ±6.9	(59.3) **56.4**	(25.3) 24.9 ±7.3
SVM + $f_{\text{TF*IDF}}$, f_{LDA}	(53.6) 51.8	(25.6) 24.7 ±6.5	(52.2) 52.4	(22.9) 23.1 ±6.8
Wikipedia Background:				
TF*IDF	40.3	19.6 ±5.4	40.1	17.2 ±5.6
Topical PageRank	40.8	19.6 ±5.4	38.1	16.7 ±5.3
XGBoost + $f_{\text{TF*IDF}}$, f_{LDA}	(60.0) **56.2**	(28.3) 26.9 ±6.9	(59.8) **56.4**	(25.4) **25.0** ±**7.2**
Test Collection→	**Lifestyle Magazines**		**Printed Press**	
Model ↓	Aggr.	Av.±Stdev.	Aggr.	Av.±Stdev.
TextRank	28.8	11.8 ±4.2	28.0	11.5 ±4.1
TF*IDF	30.3	13.1 ±4.4	32.7	13.3 ±4.8
Topical PageRank	29.2	12.5 ±4.3	28.6	11.7 ±4.2
XGBoost	(47.1) 45.6	(20.6) 19.4 ±6.2	(48.3) 45.8	(20.8) 19.4 ±6.3
XGBoost + $f_{\text{TF*IDF}}$	(47.2) 45.1	(20.8) 19.3 ±6.0	(48.2) 46.8	(20.5) 19.8 ±6.3
XGBoost+ $f_{\text{TF*IDF}}$, f_{LDA}	(48.1) 46.0	(20.9) 19.7 ±6.1	(49.8) **47.5**	(21.1) **20.2** ±**6.3**
SVM + $f_{\text{TF*IDF}}$, f_{LDA}	(43.3) 42.0	(19.3) 18.0 ±5.7	(42.6) 41.0	(18.0) 17.2 ±5.7
Wikipedia Background:				
TF*IDF	26.9	11.2 ±4.0	29.6	12.1 ±4.4
Topical PageRank	27.5	11.6 ±4.1	28.4	11.4 ±4.2
XGBoost + $f_{\text{TF*IDF}}$, f_{LDA}	(48.8) **46.4**	(21.3) **20.0** ±**6.3**	(49.4) 46.8	(21.2) 19.8 ±6.4

Table 5 Experimental results over different unsupervised and supervised models. The **precision at 5** selected keyphrases is evaluated on an aggregated set of keyphrases from different annotator (Aggr.) and for scores averaged over different annotators with standard deviation (Av.±Stdv.). Development scores for supervised classifiers are included between brackets.

The optimal number of keyphrases for aggregated target sets is eight, and for per-annotator sets is four. This difference seems reasonable, given that a single annotator on average assigns about 5 keyphrases, whereas the per-annotator sets are aggregates of 5 annotations.

As a result from the high levels of disagreement between annotators, standard deviations on averaged scores are equally high, even more so for supervised models than unsupervised models. Optimal cut-off confidence thresholds for optimal F_1 are highly dependent on the evaluation setting (aggregated versus annotators based) as is apparent from Figure 9. While scores calculated for aggregated sets of keyphrases (*Aggr.*) are not be compared with those averaged over the different annotators (*Av.*), the difference between them is most notable for the precision@5 metric. Precision@5 scores are generally much higher for aggregated sets as these contain much more keyphrases as they include the same semantic concepts in different forms. On average precision@5 for an automated extractor is around 50%, whereas this value drops to around 25% when evaluated for annotators separately. For the common purpose of summarization by keyphrases, we advocate the use of multiple opinions per document for evaluation. As the task is inherently objective, obtaining scores with low deviations is a desirable aspect of a keyphrase extractor, as this means the keyphrase sets satisfy different opinions better. When keyphrases are used for visual purposes, a better objective is to optimize the score for aggregated sets of keyphrases.

Test Collection→	**Online Sports**		**Online News**	
Model ↓	Aggr.	Av.±Stdev.	Aggr.	Av.±Stdev.
TextRank	26.0	22.1 ±5.9	21.3	17.2 ±5.1
TF*IDF	27.5	22.6 ±5.9	24.1	19.3 ±5.9
Topical PageRank	27.1	22.7 ±5.8	22.8	18.6 ±5.5
XGBoost	(35.9) 36.0	(31.5) 32.4 ±8.3	(29.2) 31.6	(26.0) 26.7 ±7.4
XGBoost + $f_{\text{TF*IDF}}$	(36.7) 36.6	(32.1) 32.1 ±8.1	(29.7) 31.7	(26.3) 26.8 ±7.5
XGBoost+ $f_{\text{TF*IDF}}$, f_{LDA}	(37.5) 36.4	(32.3) 32.6 ±8.2	(30.8) **31.8**	(27.1) **27.1** ±7.7
SVM + $f_{\text{TF*IDF}}$, f_{LDA}	(33.2) 33.9	(29.4) 27.3 ±7.4	(27.1) 29.5	(24.5) 23.8 ±7.2
Wikipedia Background:				
TF*IDF	26.4	22.7 ±6.1	22.6	18.1 ±5.7
Topical PageRank	26.9	22.6 ±6.0	21.7	17.8 ±5.1
XGBoost + $f_{\text{TF*IDF}}$, f_{LDA}	(37.2) **36.7**	(32.4) **33.1** ±**8.6**	(30.9) **31.8**	(27.2) **27.1** ±**7.6**
Test Collection→	**Lifestyle Magazines**		**Printed Press**	
Model ↓	Aggr.	Av.±Stdev.	Aggr.	Av.±Stdev.
TextRank	16.6	13.6 ±5.0	15.6	12.8 ±4.3
TF*IDF	17.5	15.4 ±5.1	18.1	14.2 ±4.8
Topical PageRank	17.0	14.8 ±5.1	16.1	12.9 ±4.3
XGBoost	(26.2) 26.5	(23.0) 23.0 ±7.4	(25.8) 25.6	(22.8) 21.3 ±7.0
XGBoost + $f_{\text{TF*IDF}}$	(26.2) 26.0	(23.2) 23.1 ±7.4	(25.8) 26.1	(22.5) 22.0 ±7.0
XGBoost+ $f_{\text{TF*IDF}}$, f_{LDA}	(26.7) 26.4	(23.3) 23.1 ±7.2	(26.6) **26.5**	(23.2) **22.7** ±**7.1**
SVM + $f_{\text{TF*IDF}}$, f_{LDA}	(24.1) 24.2	(21.6) 19.5 ±6.7	(22.7) 22.4	(19.7) 17.9 ±6.1
Wikipedia Background:				
TF*IDF	15.5	13.0 ±4.7	16.4	13.2 ±4.7
Topical PageRank	16.2	13.8 ±4.7	15.8	12.6 ±4.3
XGBoost + $f_{\text{TF*IDF}}$, f_{LDA}	(27.1) **26.8**	(23.7) **23.3** ±**7.2**	(26.4) 26.1	(23.2) 22.3 ±7.0

Table 6 Experimental results for different unsupervised and supervised models. The macro-averaged $\mathbf{F_1}$ **measure** selected keyphrases is evaluated on an aggregated set of keyphrases from different annotators (Aggr.) and for scores averaged for different annotators with standard deviation (Av.±Stdv.). Development scores for supervised classifiers are included between brackets.

4.5 Training set size

In Section 4.2 the need for supervision in automatic keyphrase extraction was highlighted. In this section we study the annotation effort versus performance. Figure 10 plots the performance of supervised models with contextual features (XGBoost +$f_{\text{TF*IDF}}$ +f_{LDA}) for different amounts of annotated documents in the training data. We use limited sets of training data (10, 25, 50, 100 and 300 documents) and measure the annotator-averaged F_1 score using the models. This demonstrates the rapid increase in supervised performance over the unsupervised models. From a minimum of 25 annotated documents, the F_1 measure exceeds each of the unsupervised models. Another observation is that the optimal performance is reached quite rapidly: a maximum value is obtained for as few as a hundred annotated documents. On the one hand, this shows that the annotation cost for a supervised system that significantly outperforms the best unsupervised systems is quite low. On the other hand, it highlights the need for more descriptive features to further improve supervised keyphrase extraction.

4.6 Training data from multiple opinions

Previous sections focused on the effect of multiple opinions on the *evaluation* of keyphrase extraction. Figure 10 shows the resulting F_1 score for the different subcollections, comparing a supervised method (XGBoost + $f_{\text{TF*IDF}}$+ f_{LDA}) with an unsupervised method (Topical

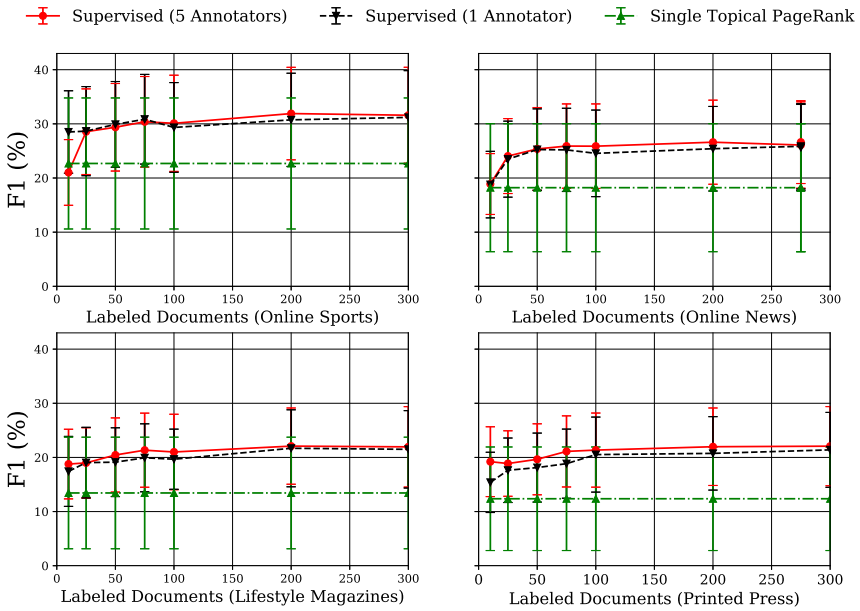


Fig. 10 Supervised model (XGBoost + $f_{\text{TF*IDF}}$+ f_{LDA}) versus various unsupervised model (Topical PageRank) for different amounts of training data.

PageRank), as a function of the number of training items (ranging from 5 to 300 documents). For the supervised method, we also show the influence of multiple opinions during *training* ('5 Annotators' vs. '1 Annotator) The '5 Annotators' case makes use of the aggregated set of annotated keyphrases from all 5 annotators for each document. We observe a limited increase in F_1 performance by aggregating keyphrase sets. This is partly due to the increasing amount of positive training data. An additional explanation is the following: for training collections generated by single readers, or document authors, it is likely that many candidates not tagged as keyphrase, might be seen as keyphrases by others. Yet during training they are implicitly considered as negative cases. When multiple opinions are not available, an elegant solution is to recast the problem into a positive versus unlabeled learning setting (Elkan and Noto, 2008; Sterckx et al, 2016).

4.7 Effect of Document Length

Finally we study the influence of document length on extraction performance. Figure 11 shows F_1 averaged per annotator for the supervised models as a function of document length. The overall trend is that scores get lower for longer documents, most notably for *Lifestyle* content. Intuitively, this is not surprising, since a longer document will produce more candidate phrases and it thus becomes more difficult to pick the (about 5) correct ones.

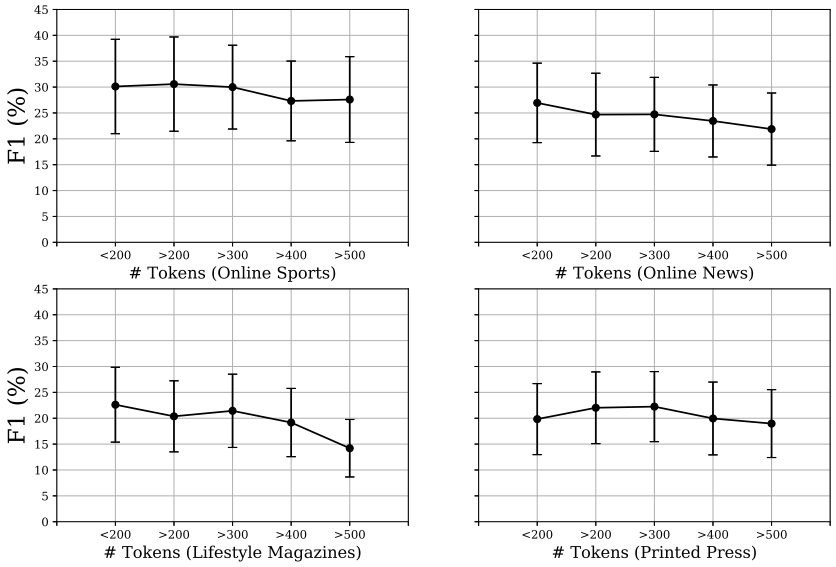


Fig. 11 Annotator-averaged F_1 for the supervised model (XGBoost + $f_{\text{TF*IDF}}$+ f_{LDA}) versus document length.

5 Conclusion and Guidelines for Automatic Keyphrase Extraction

In this paper, we presented a number of large, new collections for evaluation of Automatic Keyphrase Extraction, with multiple opinions per document. A panel of more than 350 annotators provided sets of keyphrases for different types of content from *online news*, *online sports*, *lifestyle magazines* and *printed press*. We were able to quantify the subjectivity of the keyphrase extraction task in terms of annotator disagreement, thanks to availability of multiple opinions per annotated document. As shown, this has important consequences on evaluation and training of keyphrase extractors. We studied different ways of assessing keyphrase extractor output for a number of existing supervised and unsupervised techniques. Our evaluation experiments demonstrated the importance of a suitable candidate generation strategy, and the superior effectiveness of supervised models over unsupervised models, even for low annotation efforts. When training on documents with multiple opinions, a small increase in performance is found using aggregated sets of keyphrases for training.

Many challenges and opportunities for effective keyphrase extraction remain. To conclude, we present several guidelines one should take into account when automatically extracting keyphrases and future research opportunities.

– **Candidate generation:** Proper candidate generation cannot be underestimated. Depending on the type of the documents, candidate phrases need to cover the keyphrases assigned by annotators while keeping the ratio of candidates versus keyphrases as low as possible. Figure 7 indicates that many valid keyphrases can be lost in this stage by over-filtering or missing crucial keyphrase forms while generating too many candidates which seldom appear as a keyphrase.

- **Feature Design:** A crucial aspect of keyphrase extraction remains supervision and feature design. As Figure 10 shows, performance of supervised classifiers tends to stagnate for a relatively low amount of training data, which indicates limited expressiveness by the features. A possibility for future research is the use of neural network classifiers for more sophisticated representations of keyphrases.
- **Evaluation:** Our evaluation show the frailty of current keyphrase evaluation and potentially large fluctuations in scores depending on what sets of annotations are used for evaluation. As Turney (1999) already noted, a more suitable evaluation for keyphrase extractors would be to let annotators compare sets of keyphrases output by different models. A downside of this setting is that different models need to be evaluated separately, and fine-tuning towards individual opinions is impractical.
- **Task subjectivity:** Low agreement between annotators on keyphrases as demonstrated in Table **??** shows that keyphrase extraction remains a highly subjective natural language processing task with large consequences for evaluation and training.
- **Reranking:** A topic that received less attention in this first evaluation is topic coverage in keyphrase sets. Some keyphrase extraction systems have been proposed with ways to optimize the coverage of topics in sets of keyphrases (Bougouin and Boudin, 2014).

The aim of this work was to underline the importance of these issues, and therefore we make our new test collections available for academic use, to encourage research on better evaluation and extraction techniques of keyphrases, addressing a number of open issues in this area.

Acknowledgements The research presented in this article relates to STEAMER (`http://www.iminds.be/en/projects/2014/07/12/steamer`), a MiX-ICON project facilitated by iMinds Media and funded by IWT (now known as Flanders Innovation & Entrepreneurship) and Innoviris.

References

Augenstein I, Das M, Riedel S, Vikraman L, McCallum A (2017) SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. ArXiv e-prints `1704.02853`

Bird S (2006) NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, pp 69–72

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. JMLR 3(4-5):993–1022, DOI 10.1162/jmlr.2003.3.4-5.993, URL `http://www.crossref.org/jmlr\_DOI.html`

Bougouin A, Boudin F (2014) Topicrank : ordonnancement de sujets pour lextraction automatique de termes-cls. TAL 55(1):45–69, URL `http://www.atala.org/IMG/pdf/2._Bougoin-TAL55-1.pdf`

Bowman J (2003) Essential Cataloguing. Facet Pub., URL `https://books.google.be/books?id=C-7gAAAAMAAJ`

Brill E (1992) A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language, Association for Computational Linguistics, pp 112–116

Bulgarov FA, Caragea C (2015) A comparison of supervised keyphrase extraction models. In: Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume, pp 13–14, DOI 10.1145/2740908.2742776, URL `http://doi.acm.org/10.1145/2740908.2742776`

Chang CC, Lin CJ (2011) Libsvm: A library for support vector machines. ACM Trans Intell Syst Technol 2(3):27:1–27:27

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. CoRR abs/1603.02754, URL http://arxiv.org/abs/1603.02754

D'Avanzo E, Magnini B, Vallin A (2004) Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004. In: Proceedings of the 2004 DUC

Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, pp 213–220, DOI 10.1145/1401890.1401920, URL http://doi.acm.org/10.1145/1401890.1401920

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378

Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-manning CG (1999) Domain specific keyphrase extraction. In: Proceedings of the 16th International Joint Conference on AI, pp 668–673

Gazendam L, Wartena C, Malais V, Schreiber G, de Jong A, Brugman H (2009) Automatic annotation suggestions for audiovisual archives: Evaluation aspects. Interdisciplinary Science Reviews 34(2-3):172–188, DOI 10.1179/174327909X441090, URL http://dx.doi.org/10.1179/174327909X441090, http://dx.doi.org/10.1179/174327909X441090

Gazendam L, Wartena C, Brussee R (2010) Thesaurus based term ranking for keyword extraction. In: Database and Expert Systems Applications, DEXA, International Workshops, Bilbao, Spain, August 30 - September 3, 2010, pp 49–53, DOI 10.1109/DEXA.2010.31, URL http://dx.doi.org/10.1109/DEXA.2010.31

Grineva M, Grinev M, Lizorkin D (2009) Extracting key terms from noisy and multitheme documents. WWW 2009 MADRID! Track: Semantic/Data Web / Session: Mining for Semantics pp 661–670, URL http://dl.acm.org/citation.cfm?id=1526798

Hammouda KM, Matute DN, Kamel MS (2005) Corephrase: Keyphrase extraction for document clustering. In: Machine Learning and Data Mining in Pattern Recognition, Springer, pp 265–274

Hasan KS, Ng V (2014) Automatic keyphrase extraction: A survey of the state of the art. Proceedings of the Association for Computational Linguistics (ACL), Baltimore, Maryland: Association for Computational Linguistics

Hulth A (2003) Improved automatic keyword extraction given more linguistic knowledge. Proceedings of the 2003 conference on Empirical Natural language Processing (2000), URL http://dl.acm.org/citation.cfm?id=1119383

Jiang X, Hu Y, Li H (2009) A ranking approach to keyphrase extraction. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 756–757

Kim SN, Kan MY (2009) Re-examining automatic keyphrase extraction approaches in scientific articles. In: Proceedings of the workshop on multiword expressions: Identification, interpretation, disambiguation and applications, Association for Computational Linguistics, pp 9–16

Lievens B, Baccarne B, Veeckman C, Logghe S, Schuurman D (2014) Drivers For End-users' Collaboration In Participatory Innovation Development And Living Lab Processes. In: 17th ACM Conference on Computer Supported Cooperative Work and Social Computing

Liu Z, Huang W, Zheng Y, Sun M (2010) Automatic keyphrase extraction via topic decomposition. In: Proceedings of the 2010 Conference on EMNLP, pp 366–376

Lopez P, Romary L (2010) Humb: Automatic key term extraction from scientific articles in grobid. In: Proceedings of the 5th international workshop on semantic evaluation, Association for Computational Linguistics, pp 248–251

Medelyan O, Witten I (2002) Thesaurus based automatic keyphrase indexing. In: Proceedings of the 6th ACM/IEED-CS joint conference on Digital libraries, pp 296–297

Mihalcea R, Csomai A (2007) Wikify!: linking documents to encyclopedic knowledge. CIKM07, November 68, 2007, Lisboa, Portugal (July), URL http://dl.acm.org/citation.cfm?id=1321475

Mihalcea R, Tarau P (2004) TextRank: Bringing Order into Texts. In: Proceedings of the 2004 conference on EMNLP, URL http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf

Park Y, Byrd RJ, Boguraev B (2002) Automatic glossary extraction: Beyond terminology identification. In: 19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002, URL http://aclweb.org/anthology/C02-1142

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information processing & management 24(5):513–523

Sterckx L, Demeester T, Deleu J, Develder C (2015a) Topical word importance for fast keyphrase extraction. In: Proceedings of the 24th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp 121–122

Sterckx L, Demeester T, Deleu J, Develder C (2015b) When topic models disagree: Keyphrase extraction with multiple topic models. In: Proceedings of the 24th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp 123–124

Sterckx L, Caragea C, Demeester T, Develder C (2016) Supervised keyphrase extraction as positive unlabeled learning. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, November 2-4, 2016, Austin, Texas

Turney P (1999) Learning to extract keyphrases from text URL http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc\&an=8913245

Turney P (2000) Learning algorithms for keyphrase extraction. Information Retrieval URL http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc\&an=8913713

Wan X, Xiao J (2008) Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2, AAAI 2008, pp 855–860, URL http://dl.acm.org/citation.cfm?id=1620163.1620205

Wartena C, Brussee R, Slakhorst W (2010) Keyword Extraction Using Word Co-occurrence. 2010 Workshops on Database and Expert Systems Applications pp 54–58, DOI 10.1109/DEXA.2010.32, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5592000

Witten I, Paynter G, Frank E (1999) KEA: Practical automatic keyphrase extraction. Proceedings of the fourth ACM conference on Digital libraries URL http://dl.acm.org/citation.cfm?id=313437

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE transactions on evolutionary computation 1(1):67–82

Yih Wt, Goodman J, Carvalho VR (2006) Finding advertising keywords on web pages. WWW 2006, pp 213–222

Zhang Y, Zincir-Heywood N, Milios E (2005) Narrative text classification for automatic key phrase extraction in web document corpora. In: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management, ACM, New York, NY, USA, WIDM '05, pp 51–58, DOI 10.1145/1097047.1097059, URL http://doi.acm.org/10.1145/1097047.1097059

Zhao WX, Jiang J, He J, Song Y, Achananuparp P, Lim EP, Li X (2011) Topical keyphrase extraction from twitter. In: Proceedings of the 49th Annual Meeting of the ACL: HLT-Volume 1, Stroudsburg, PA, USA, HLT '11, pp 379–388, URL http://dl.acm.org/citation.cfm?id=2002472.2002521