

REPLICATION OF THE KEYWORD EXTRACTION PART
OF THE PAPER
“WITHOUT THE CLUTTER OF UNIMPORTANT WORDS”:
DESCRIPTIVE KEYPHRASES FOR TEXT VISUALIZATION

Jason Chuang and Christopher D. Manning and Jeffrey Heer,
Stanford University

Replicated by: Shibamouli Lahiri (University of Michigan; Winter 2015)
lahiri@umich.edu

Abstract

“Keyword Extraction” refers to the task of automatically identifying the most relevant and informative phrases in natural language text. As we are deluged with large amounts of text data in many different forms and content – emails, blogs, tweets, Facebook posts, academic papers, news articles – the task of “making sense” of all this text by somehow summarizing them into a coherent structure assumes paramount importance. Keyword extraction – a well-established problem in Natural Language Processing – can help us here. In this report, we construct and test three different hypotheses (all related to the task of keyword extraction) that take us one step closer to understanding how to meaningfully identify and extract “descriptive” keyphrases. The work reported here was done as part of replicating the study by Chuang et al. [3].

1 Introduction

Keyword extraction has been a long-standing problem in Natural Language Processing and Information Retrieval where the goal is to identify (and rank) a set of phrases given a natural language document that best describe the contents of that document in a coherent manner. Usually, a *budget* is imposed on the number of keywords¹ to be returned, and the length of the phrases in number of words. Since a document usually talks about several different ideas, topics, and thoughts, potentially interweaving them into a single, coherent body of text, it becomes hard to disentangle it and sieve out only the most important phrases, and still respecting the length and number budgets as described.

When faced with such a task, it is instructive to look into how human beings solve this challenging optimization problem. It has been generally observed that human annotators tend to select certain topics first from the set of topics they feel best describe a given document, and then they pick out phrases that they feel best describe a given topic [14]. This two-step procedure yields a set of phrases that are both “locally coherent” (in that they represent the topical structure of a document), and “globally adequate” (in that they are representative of the whole rather than a part of the document).

However, given that human judgment is rather subjective, and the length and number budgets so restrictive, it often becomes the case that given the same piece of text, two annotators come up with two radically different sets of keyphrases – leading to a very low inter-annotator agreement (20-30% F-score) [14]. While this paper suggests “relaxed matching” as a way to counteract and alleviate this problem, the inherent subjectivity of keyword extraction has long been a challenge [9].

This paper addresses the subjectivity and budget issues of keyword extraction by attacking the problem from an information visualization perspective. The authors – coming from two different backgrounds (Natural Language Processing and Human-Computer Interaction) – define “descriptive keyphrases” as phrases that best visualize a given document in terms of

¹In this report, we use the terms “keywords” and “keyphrases” interchangeably.

tag clouds.² If, among two sets of keyphrases, one has better visualization capability than the other, then the former is regarded as a “better set of keyphrases”. Note that “good visualization” is a subjective term itself, and evaluation of visualizations must somehow take into account that subjectivity.

The key contribution of this paper lies in using online surveys to gather data from many human annotators regarding what they think would constitute valid keyphrases for a document,³ then selecting factors that were most prevalent in human selection of keyphrases, constructing statistical models using those factors to classify phrases as “key” vs “non-key”, and finally evaluating tag clouds (generated by the statistical models) by Amazon Mechanical Turk.⁴ This process of systematic data collection, modeling, and evaluation gets rid of the subjectivity problem by collecting data from *a lot* of annotators instead of a few (as is customary in keyword extraction⁵). The authors also counteracted the two budget problems (length and number of keyphrases) by allowing annotators to select as many keyphrases as they wanted (minimum of five), and as long as they wanted. They were further allowed to choose their own vocabulary in building the keyphrases. This is what is known in Natural Language Processing as *abstractive summarization*, where you summarize a document’s content using your own words rather than using words/sentences from the document itself (*extractive summarization*).

The authors analyzed human-generated keyphrases to identify several interesting patterns. They showed that bigrams (two-word phrases) were most commonly chosen by humans as keyphrases, followed by unigrams, followed by n-grams with $n \geq 2$. It was further shown that phrases that were not too rare and not too common, were most frequently chosen as keyphrases, and the majority of keyphrases were technical terms that followed certain part-of-speech patterns. We shall come back to and examine some of these findings in more detail

²See, e.g., <http://www.wordle.net/>.

³The chosen domain was scientific abstracts, but the general methodology would work in any other domain.

⁴<https://www.mturk.com/mturk/welcome>.

⁵For example, the biggest annotation effort in academic keyword extraction – the SemEval 2010 Keyphrase Extraction Task – used 50 annotators [9], whereas the current paper uses 69. The more annotators you have, the more confidence you can expect in the obtained keyphrases.

in Section 4.1.

This report is organized as follows. We describe and explain our research hypothesis in Section 2, along with refinements and the final version. The project proper is detailed in Section 3, where we identify and resolve several challenges, including but not limited to lack of data, annotations, and tools. The results are presented in Section 4, where we also tie them together with our hypothesis, and show that the latter is indeed true under the assumptions we made. We conclude the report in Section 5.

2 Research Hypothesis

Our initial research hypothesis consisted of four parts:

1. *Multi-word phrases are better descriptors of a scientific abstract than single words (unigrams); for multiple abstracts, this observation needs to be traded off against the better aggregation afforded by unigrams. “Better”-ness is measured by analyzing frequency histograms on a 144-abstract corpus.*
2. *Probabilistic measures such as G^2 outperform raw term frequency and tfidf in terms of precision and recall on scientific abstracts, when logistic regression is used as classifier.*
3. *Keyphrase quality improves by including part-of-speech and positional features; parse-tree-based features do not lead to additional improvement. Quality is measured by precision and recall.*
 - **Setting:** *logistic regression classifier on scientific abstracts, using frequency, commonness, probabilistic, part-of-speech-based, parse-tree-based, and positional features.*
4. *Reducing redundancy (by coalescing semantically identical phrases into a single phrase) helps identify more meaningful phrases, especially from a visualization perspective. “Meaningful”-ness is measured by average user satisfaction on Amazon Mechanical Turk.*

These four parts – we could call them “four hypotheses” – are distilled from the authors’ recommendations in Section 8.1 of the paper. We shall first explain the background in more detail, before getting into refinements, and the final version of the hypothesis.

Keyphrase extraction can be tackled as an unsupervised (ranking) problem, or as a supervised (classification) problem. In the unsupervised setting, a set of candidate phrases is first extracted from a document (or a set of documents), and then ranked by a graph-based ranking model (e.g., TextRank [16]) or an informativeness-based model [20]. Since

the ranking model cannot control the number of returned keyphrases (usually a heuristic cutoff is set at about five or ten top-ranked phrases), it becomes critical that the initial keyphrase selection be as good (and picky) as possible. The selection phase is guided by named entity recognition [5, 17] and noun phrase chunking, and often accompanied by part-of-speech patterns [8]. While noun phrases and named entities can by no means define the complete set of keyphrases for a document (see, e.g., Section 4.1), they do help in identifying potentially good candidate phrases and filtering out obviously incorrect choices early on in the process.

In *supervised keyword extraction*, on the other hand, candidate phrases are classified as *keyphrase* or *not*. More accurately, a candidate phrase is classified as a *valid keyphrase* for a document vs. *not valid*. Note that this latter setting allows the use of document-based features (e.g., positional, frequency-based, and grammatical) that are not possible in the former setting. Also to be noted is the fact that although most studies in supervised keyphrase extraction treat the problem as binary classification, a few used *sequence labeling* methods such as Conditional Random Fields (CRF) to achieve good performance [21]. Unsupervised phrase ranking functions (e.g., phrase length, phrase centrality in a network, informativeness) often serve as good features in supervised keyphrase extraction.

Both supervised and unsupervised keyphrase extraction need a final filtering and redundancy reduction step to cope with semantically related phrases that are generated as part of the process. Unsupervised approaches are particularly susceptible to generating keyphrases that are semantically related (often due to poor candidate selection), such as “Obama” and “Barack Obama” [6].

In this paper, keyphrase extraction is tackled as a supervised problem (i.e., binary classification). Six different types of features were used (frequency, commonness, probabilistic, part-of-speech-based, parse-tree-based, and positional; see Section 3.2). The first three types of features need what is known as a *background corpus* in keyphrase extraction literature. The document(s) we are extracting keywords from constitute(s) the *foreground corpus*, whereas the documents being used to extract certain categories of features, are called the

background corpus. This paper used 144 Ph.D. thesis abstracts from Stanford University as its foreground corpus, and 9,068 Ph.D. theses (presumably abstracts only) as its background corpus. A second variant of commonness features was extracted on a separate background corpus – Web1T [2].

The usage of Ph.D. thesis abstracts as foreground corpus characterizes and re-inforces a long line of research in *academic keyphrase extraction*. The reason academic keyphrase extraction has received more attention in the literature than other domains, is mostly because of data availability. There are three separate academic keyword extraction datasets (cf. [11], Section 4), whereas there is only one dataset for keyword extraction from meeting transcripts, and one dataset for keyword extraction from news articles (cf. [7], Section 2). The SemEval 2010 Keyphrase Extraction task was specifically devoted to scientific articles [9].

The authors of this paper created their own dataset from Stanford thesis abstracts. They looked into both single-document as well as multi-document keyphrase extraction, the latter being tried for the first time in keyword extraction literature (as far as we know). Another novel aspect of this paper is the authors’ use of *commonness* as a feature in keyword extraction. Commonness quantifies how “popular” or “rare” a particular phrase is with respect to the background corpus. Usage of Amazon Mechanical Turk in keyphrase validation is a novelty as well, but that was inspired by the fact that human-computer interfaces (tag cloud visualizations in this case) ultimately need to be evaluated by human annotators.

Keyword extraction systems are often evaluated (and compared) using area under the precision-recall curve, and this paper follows the same approach. A second approach is to use the F-score (harmonic mean of precision and recall) at different cutoff points, and see which system gives better F-scores. The best keyphrase extraction systems in the SemEval 2010 Shared Task achieved F-scores (at cutoff 15) between 20% and 30%, by using several complex features, external knowledge bases such as Wikipedia, and sophisticated classifiers such as multi-layered perceptrons and bagged decision trees [1, 9, 13]. This paper on the other hand uses a much smaller and simpler set of features with logistic regression classifier, and shows that the generated keyphrases can still be very good from a visualization standpoint. We

shall describe the features in Section 3.2.

We refined and qualified our initial hypothesis in light of our revelations from studying the paper in depth, and our experimental findings. We also abandoned the fourth part of the hypothesis because it was more related to visualization and Human-computer Interaction than Natural Language Processing. In the end, our final hypothesis looks as follows:

While doing single-document supervised keyphrase extraction from scientific abstracts:

- 1. Two-word phrases (bigrams) are most frequently chosen by humans, followed by unigrams, trigrams, and higher-order n-grams.*
- 2. Log-odds Ratio outperforms raw term frequency and tfidf in terms of area under the precision-recall curve, when logistic regression is used as classifier.*
- 3. Area under the precision-recall curve improves by including part-of-speech and positional features; parse-tree-based features lead to very little additional improvement when part-of-speech features are included.*
 - **Setting:** frequency, commonness, probabilistic, part-of-speech-based, parse-tree-based, and positional features; logistic regression classifier.*

Note that the initial hypothesis included a multi-document keyphrase extraction component. We abandoned it in the final version, because grouping documents into 48 topically similar clusters (with exactly three abstracts in each) proved to be a very challenging optimization problem.⁶ We also abandoned a part of the paper (not the hypothesis) that related extracted keywords to the annotators' *level of familiarity* with the scientific topic, because it was very difficult to control for annotators' level of familiarity on Amazon Mechanical Turk

⁶Please see the following links for discussion on this issue: <http://stats.stackexchange.com/questions/8744/clustering-procedure-where-each-cluster-has-an-equal-number-of-points>, <http://stackoverflow.com/questions/5452576/k-means-algorithm-variation-with-equal-cluster-size>.

without running into privacy and adversarial issues. Details on our annotation study appear in Section 3.1.

3 Project

In this section we shall describe the steps we followed to replicate the study, and gather evidence in favor of our hypothesis. The steps can be broadly classified into three categories – data collection and annotation, feature extraction, and performance evaluation – as detailed in the three following subsections.

3.1 Data Collection and Annotation

A big part of any empirical project (all Natural Language Processing tasks included) is to collect data, and this paper is no exception. Recall from Section 2 that we needed two different datasets (*corpora*) – a foreground corpus, and a background corpus. The authors used Stanford thesis abstracts for both purposes, and Web1T as a second background corpus [2]. The problem, however, is that the Stanford dataset was not available to us due to privacy reasons, and Web1T – although available in our research group – was too time-consuming to use in an efficient manner (owing to the large number of Web1T queries we needed to run for each experiment).

In the end, we resorted to the SemEval 2010 Keyphrase Extraction Task dataset as our foreground corpus [9]. More specifically, we used the training set of this corpus as the foreground. The training set has 144 academic papers collected from the ACM Digital Library,⁷ from four 1998 ACM classifications: C2.4 (Distributed Systems), H3.3 (Information Search and Retrieval), I2.11 (Distributed Artificial Intelligence – Multiagent Systems), and J4 (Social and Behavioral Sciences – Economics). The training set is balanced among the

⁷<http://dl.acm.org/dl.cfm>

four classes. To make the setup similar to the original paper (which used 144 thesis abstracts as the foreground corpus), we collected the title and abstract of these 144 SemEval papers from ACM Digital Library and copied them into text files.

Our background corpus comes from the **Inspec** database of 2,000 abstracts – first described by Anette Hulth, then at Stockholm University [8]. The abstracts are from 1998 to 2002, from journal papers, and from the disciplines *Computers and Control*, and *Information Technology*. The set of abstracts has associated with them two distinct sets of keywords assigned by a professional indexer – one comprising *controlled terms* (terms that were present in a thesaurus), and another comprising any terms (*uncontrolled terms*) that the indexer deemed appropriate. In our experiments, we only considered the text portion of the abstracts as our *background corpus*, and ignored the keywords because they were not relevant to feature extraction.

The reason we chose Inspec as our background corpus is because it consists of scientific abstracts that are similar in content and length to the SemEval abstracts (the *foreground*). Furthermore, the corpus is large enough to serve as background (i.e., has enough information to gather term statistics and extract features), and yet small enough to be quickly indexed and processed without the overhead of an online query mechanism – as would be the case with Web1T [2]. Furthermore, Inspec has been used as a benchmark in several previous keyword extraction studies [7, 11, 16].

Since we did not have access to the original keyword annotations (again, due to privacy reasons), we resorted to Amazon Mechanical Turk to annotate the foreground corpus (144 abstracts) with keyphrases.⁸ The original study collected 69 human-participant responses, so we went ahead and made 69 assignments on Mechanical Turk for each foreground abstract. In Mechanical Turk, every abstract is associated with a *human intelligence task* (HIT) that needs to be completed 69 times (called *assignments* in Mechanical Turk parlance) by 69 different people. In the end, we received 9,936 assignments in total, submitted by 214

⁸Code and data can be found at http://web.eecs.umich.edu/~lahiri/replication_of_keyword_extraction_part_of_the_paper_by_Chuang_etal_data_and_code.zip.

unique Turkers. Each assignment consisted of five to 16 keyphrases that could be either abstractive or extractive with no length restrictions imposed on them (in keeping with the spirit of the original study). In total, we received 17,640 unique keyphrases (15,187 after lowercasing).⁹

The reason we received so many responses, is because Turkers are creative and they come up with new phrases to describe a document (esp. abstractive phrases). We suspect that Turkers are more creative than the human judges of the original paper – potentially owing to the fact that the original judges were more educated on average, and therefore, more familiar with academic abstracts. Note that we enforced quality control by making sure that only those Turkers who completed 200 or more HITs with 95% or higher approval rate could work on our assignments. But even with quality control, Turkers seemed to select arbitrarily long phrases (even sentences and sentence fragments!) to describe abstracts. This indicates a fundamental lack of understanding of the underlying problem. One reason this was the case, is that Turkers were given minimal instructions (in keeping with the spirit of the original study), which probably confused them into over-/under-generating phrases. Verifying this hypothesis is beyond the scope of the present work, but it will be an interesting direction to investigate in future.

Since it was impossible to manually verify all 17,640 unique keyphrases (that was one reason of doing the Mechanical Turk study in the first place), we implemented a simple filter to weed out obviously incorrect or spurious answers.¹⁰ This filter was constructed manually by looking through responses, and coming up with a list of “spurious phrases” (cf. Table 1). After filtering, we are left with 17,621 unique keyphrases (15,173 if lowercased) – still too large in comparison with the original study. We further restricted our attention to only those phrases that are of length five words or shorter (in keeping with the original), thus ending up with 12,810 keyphrases. Focusing on the phrases that were *extractive* (i.e., the ones that appeared in their corresponding abstracts), we finally obtained 7,816 keyphrases.

⁹Contrast this with 5,611 responses received from 69 human participants in the original study.

¹⁰We also filtered out “keyphrases” that were essentially the whole title of the paper, or the whole abstract – copied and pasted verbatim.

3.2 Feature Extraction

We built logistic regression models on the 7,816 human-annotated keyphrases. These 7,816 serve as *positive examples* to the classifier. We obtained 78,160 negative examples by randomly sampling n-grams ($1 \leq n \leq 5$) from the SemEval corpus without replacement. These negative examples were each given a weight of 0.1 in accordance with the original study.¹² Four types of features – frequency, commonness, grammatical, and positional – were extracted for each example phrase (positive and negative). Note that probabilistic features (G^2 , BM25, weighted log-odds ratio) are subsumed under frequency features, and grammatical features consist of part-of-speech features and parse-tree-based features.

Frequency features have been summarized in Table 3 of the original paper, which we reproduce here for conceptual clarity (Table 2). Note that although frequency features are often used in keyphrase extraction, their original intent was to help assess *document relevance* in information retrieval. Note also that probabilistic features – G^2 , BM25, weighted log-odds ratio – are not used in keyphrase extraction that often. We will later show that one probabilistic feature – weighted log-odds ratio – outperforms other frequency-based features (Section 4.2).

Commonness features (briefly mentioned in Section 2) encode the *relative rarity* of a phrase against a background corpus. *Commonness* of a term (n-gram) is defined as $\frac{\log(tf_{bg})}{\log(tf_{max})}$, where tf_{bg} is the frequency of the term in the background corpus, and tf_{max} is the frequency of the most frequent n-gram (with the same number of words as the term) in the background corpus. The authors used two background corpora to extract commonness information – Stanford thesis abstracts and Web1T, and experimented with up to 20 *bins* of commonness (which we shall replicate in Section 4.2). We used **Inspec** as our sole background corpus, so the commonness information comes from Inspec.

We extracted ten binary grammatical features (yes/no) – six of them based on parse trees,

¹²The rationale behind choosing ten times as many negative examples as there are positives and then under-weighting the former, was never explained in the paper.

Frequency Feature	Definition
$\log(\text{tf})$	$\log(t_{Doc})$
tf.idf	$\frac{t_{Doc}}{t_{Ref}} \log(\frac{N}{D})$
G^2	$2(t_{Doc} \log(\frac{t_{Doc} T_{Ref}}{T_{Doc} T_{Doc}}) + t_{\overline{Doc}} \log(\frac{t_{\overline{Doc}} T_{Ref}}{T_{\overline{Doc}} T_{Doc}}))$
BM25	$\frac{3t_{Doc}}{t_{Doc} + 2(0.25 + 0.75 \frac{T_{Doc}}{r})} \log(\frac{N}{D})$
WordScore	$\frac{t_{Doc} - t_{Ref}}{T_{Doc} - T_{Ref}}$
Weighted log-odds ratio	$\frac{\log(\frac{t'_{Doc}}{t_{\overline{Doc}}} - \log(\frac{T'_{Doc}}{T_{\overline{Doc}}})}{\sqrt{t'^{-1}_{Doc} + t'^{-1}_{\overline{Doc}}}}$

Table 2: Frequency features. Note that we dropped WordScore because the definition of T_{Ref} was not clear from the paper. The other definitions are explained as follows. Given a document from a reference corpus with N documents (in our case, Inspec with 2,000 documents), the score for a term is given by these formulas. t_{Doc} and t_{Ref} denote term frequency in the document and reference corpus; T_{Doc} and T_{Ref} are the number of *words* in the document and reference corpus; D is the number of documents in which the term appears; r is the *average word count* per document; t' and T' indicate measures for which we increment term frequencies in each document by 0.01; terms present in the corpus but not in the document are defined as $t_{\overline{Doc}} = t_{Ref} - t_{Doc}$ and $T_{\overline{Doc}} = T_{Ref} - T_{Doc}$. Among the family of tf.idf measures, a *reference-relative form* was chosen as shown. For BM25, the parametrization of $k_1 = 2$ and $b = 0.75$ has been suggested in previous literature. A *term* is any analyzed phrase (n-gram). When frequency statistics are applied to n-grams with $n = 1$, the terms are all the individual words in the corpus. When $n = 2$, scoring is applied to all unigrams and bigrams in the corpus, and so on.

and rest based on parts-of-speech.¹³ The Stanford Parser was used to extract grammatical features [10]. We implemented several Java functions within Stanford Parser code to make it work according to our specifications. The grammatical features are as follows (first six are parse-tree-based, and the rest are part-of-speech-based):

- **Is full noun phrase?** If a phrase matches a full noun phrase, we mark this feature as a one; otherwise, it is zero.
- **Is full verb phrase?** If a phrase matches a full verb phrase, we mark this feature as a one; otherwise, it is zero.
- **Is partial noun phrase?** If a phrase is part of a full noun phrase, we mark this feature as a one; otherwise, it is zero.
- **Is partial verb phrase?** If a phrase is part of a full verb phrase, we mark this feature as a one; otherwise, it is zero.
- **Is optional leading word?** If the first word of a phrase is an *optional leading word* (leading word in a noun phrase whose part-of-speech is a cardinal number (CD), a determiner (DT), or a pre-determiner (PDT)), we mark this feature as a one; otherwise, it is zero.
- **Is head noun?** If the last word of a phrase is a *head noun* (last word of a noun phrase), we mark this feature as a one; otherwise, it is zero.
- **Is technical term?** If a phrase matches one of the part-of-speech patterns for a technical term, we mark this feature as a one; otherwise, it is zero.
- **Is compound technical term?** If a phrase matches one of the part-of-speech patterns for a compound technical term, we mark this feature as a one; otherwise, it is zero.
- **Is partial technical term?** If a phrase is part of a technical term, we mark this feature as a one; otherwise, it is zero.

¹³Note that it was not clear from the paper whether these features were binary or categorical, but based on further pondering, it seemed to us that binary will be the best choice.

- **Is partial compound technical term?** If a phrase is part of a compound technical term, we mark this feature as a one; otherwise, it is zero.

Technical terms and compound technical terms are identified by the following regular expression productions on parts-of-speech:

$$T = (A|N)^+ (N|C) | N \tag{1}$$

$$X = (A|N)^* N \textit{ of} (T|C) | T \tag{2}$$

where T is a technical term, X is a compound technical term, A is an adjective (corresponding to part-of-speech tags JJ, JJR, JJS), N is a noun (corresponding to part-of-speech tags NN, NNS, NNP, NNPS), and C is a cardinal number (corresponding to part-of-speech tag CD).¹⁴ All part-of-speech tags are in the standard Penn Treebank format (cf. Table 2 of [15]).

Finally, we extracted three positional features – *absolute first occurrence* (real number between 0 and 1), *relative first occurrence* (real number between 0 and 1), and *presence in first sentence* (binary yes/no) – defined as follows:

- **Absolute first occurrence:** Earliest position (i.e., word index) of the document where a phrase first appeared, normalized by the total number of words in the document.
- **Relative first occurrence:** $(1 - a)^k$, where a is the absolute first occurrence of a phrase, and k is the number of times the phrase appeared in the document. Relative first occurrence penalizes frequently occurring phrases’ first occurrences.
- **Presence in first sentence:** If a phrase is present in the first sentence, we mark this feature as a one; otherwise, it is zero. We used NLTK for sentence segmentation [12].

¹⁴Note that all technical terms are compound technical terms, but not the other way around. Note also that *of* is a terminal in production (2).

3.3 Performance Evaluation

Recall from Section 2 that the authors performed *supervised keyphrase extraction* by feeding positive and negative example phrases (along with their features) to a classifier. Upon closer inspection, it seemed to us that what was really done, was *supervised ranking* (i.e., learning to rank) using a regression model rather than supervised classification. This observation could not be confirmed (or otherwise) with the first author, because most details of the original study were lost, leaving us to work with best guesses.

It was clear, however, that logistic regression was used as the learning model, and it also became clear, upon further inspection, that R was used as the library of choice for working with logistic regression models. R has some nice parametrizations to weight each sample differently, and that facility was used to under-weight negative examples (cf. *glm()* in [19]). Furthermore, precision-recall curves seemed also to have been generated using R. In particular, we used the package ROCR to generate precision-recall curves from raw predictions [18].

The choice of using precision-recall curves (i.e., area under the curves) to measure and compare performance of different models may seem arbitrary at first blush, but do note that precision-recall curves show a clear trade-off between retrieving good phrases at top ranks vs. retrieving as many good phrases as possible, no matter what the rank is. This trade-off is not so obvious with other measures of performance such as F-score at rank k , or *mean average precision*. Also to be noted is the fact that precision-recall curves have a firm grounding in Information Retrieval community in general, and keyword extraction literature in particular [7].

We experimented with five different logistic regression models – the first four corresponding (almost) exactly to the four sub-figures of Figure 4 in the original paper, and the last one to assess the impact of commonness bins. The models are constructed on the following features:

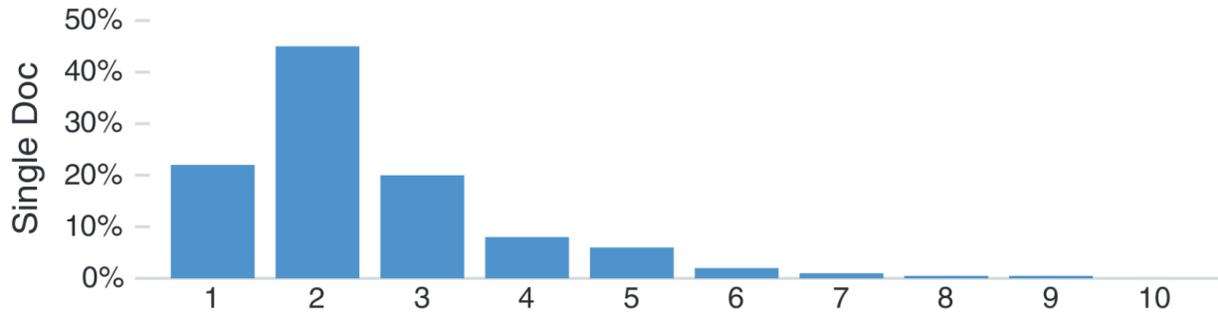
- Frequency and probabilistic features.
- Adding commonness to raw term frequency, and comparing the joint model with raw term frequency and G^2 .
- Adding grammatical features to frequency and commonness features.
- Adding positional features (to the rest of the features) to create the final models.

4 Results

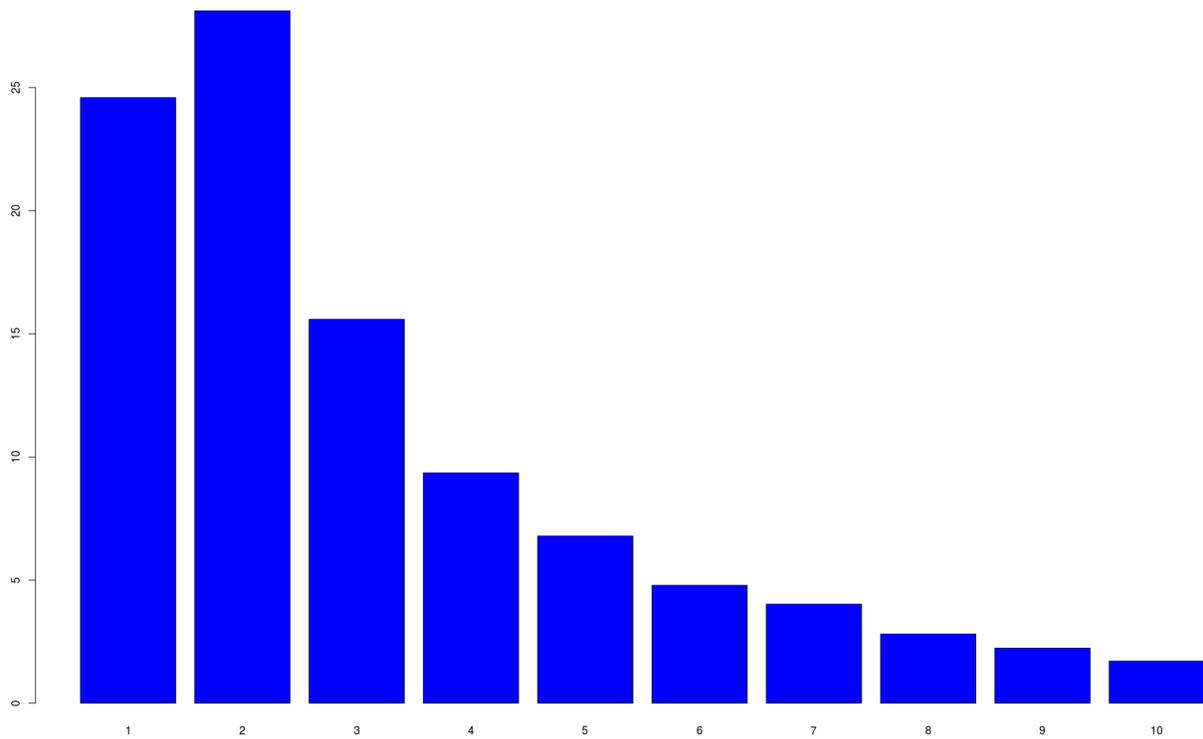
We describe the results in two subsections – the first one is an exploratory analysis on Amazon Mechanical Turk responses, and the second one illustrates the performance of logistic regression models for supervised keyphrase extraction under different feature combinations, using precision-recall curves. We shall also show evidence in these two subsections that goes in favor of our research hypothesis (cf. the final hypothesis in Section 2), and show that the three parts of our hypothesis are indeed true, on our dataset, annotations, features, and models.

4.1 Exploratory Analysis

The paper spends considerable amount of space to detail the results of their exploratory analysis, for good reasons. *A priori*, it is hard to define or validate keyphrases on scientific abstracts (or on any other domain), let alone “descriptive” keyphrases. The authors took the exploratory analysis as an opportunity to look into people’s minds to see what they usually think are the most important components of a document. Recall from Section 3.1 that the annotation instructions were minimal (“summarize the content using five or more keyphrases, using any vocabulary”). This was done to ensure that people could freely choose what they thought would best describe a given document, without being biased by experimenters’ (often arbitrary) decisions.



(a) Original



(b) After replication

Figure 1: Distribution of phrase length. X-axis is the phrase length (in number of words), and Y-axis is the percentage of (human-generated) phrases with that length.

While the original intent of minimal instructions was clear from the paper, note that we used Amazon Mechanical Turk, which somehow calls for more control on our part. However, we shall see in the remainder of this report that even without sufficiently detailed instructions, we obtained results that matched the hypothesis of the original paper.

As a starter, we looked into the frequency distribution of n-grams in human-generated keyphrases. The paper says: “bigrams are the most common response, accounting for 43% of all free-form keyphrase responses, followed by unigrams (25%) and trigrams (19%).” After replication, the numbers we obtained are: 24.5% bigrams, 21.4% unigrams, and 13.6% trigrams. **This validates the first part of our final hypothesis (Section 2).** The full distributions are shown in Figure 1. There are several points to note here. Figure 1a shows only *one* histogram from the three reported in Figure 2 of the original study, because we did not experiment with *multiple-document keyphrase extraction* (reasons explained in Section 2). Also to be noted is the fact that we have a much heavier and longer tail in the histogram. For illustration purposes, we capped the tail at 10 words per phrase. In fact, while the original study reports <5% of longer-than-five-words responses, we have 26.45% of responses that are longer than five words. The reason, as explained in Section 3.1, is that Turkers selected very long phrases (even sentences and sentence fragments) because of our minimal instructions. Another point is that we have a substantially lower peak at bigrams, and a somewhat lower peak at trigrams, the mass being shifted to longer keyphrases (there are many keyphrases longer than 10 words, but we capped them due to visual clarity).

The authors found that 65% of keyphrase responses were *extractive*, i.e., they were present in the original document. For us, the number was slightly higher: 66.55%. They also found that “22% of keyphrases occur in the first sentence, even though first sentences contain only 9% of all terms.” We found that 18.6% of keyphrases occurred in the first sentence, whereas first sentences contained 25.33% of all unique terms. To be noted is the fact that our foreground corpus is different than the authors’ (SemEval vs. Stanford thesis abstracts), and the former potentially contains smaller abstracts (conference paper abstracts)

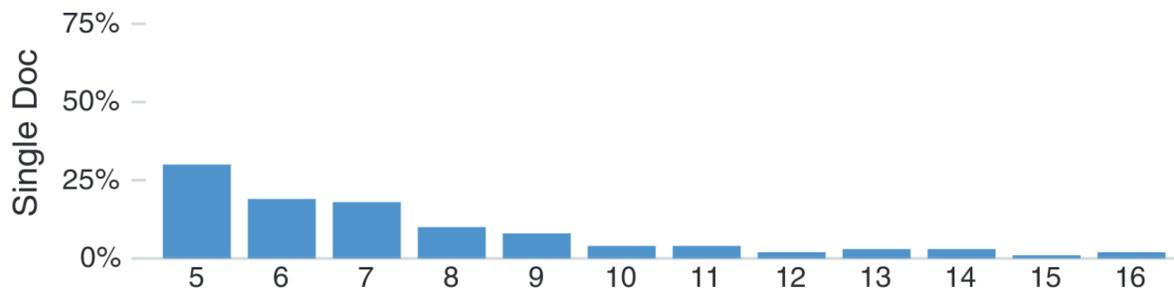
than theses, thus concentrating many important terms in first sentences.¹⁵ Further, the original study reported that nearly “two-thirds of keyphrases found in the document” (i.e., *extractive* keyphrases) “are part of a noun phrase”, “7% are part of a verb phrase”, and “over 80% . . . are part of a technical term”.¹⁶ Note that these numbers add up to more than 100%, which is fine because many technical terms are actually noun phrases, so they are double-counted. After replication, we found that 38.73% unique keyphrases are part of a noun phrase, 43.18% are part of a verb phrase, and 25.62% are part of a technical term. That we have many more verb phrases, and many fewer noun phrases and technical terms, can be tentatively explained as follows:

- We provided minimal instructions, thereby confusing Turkers.
- Turkers on average are less educated than the annotators of the original study, and potentially much less familiar with academic jargon.
- Turkers over-emphasized the *actions* (verb phrases) rather than *content* (noun phrases and technical terms), perhaps owing to their lack of familiarity with the domain.
- Our abstracts were shorter than thesis abstracts, and less laden with technical terms.

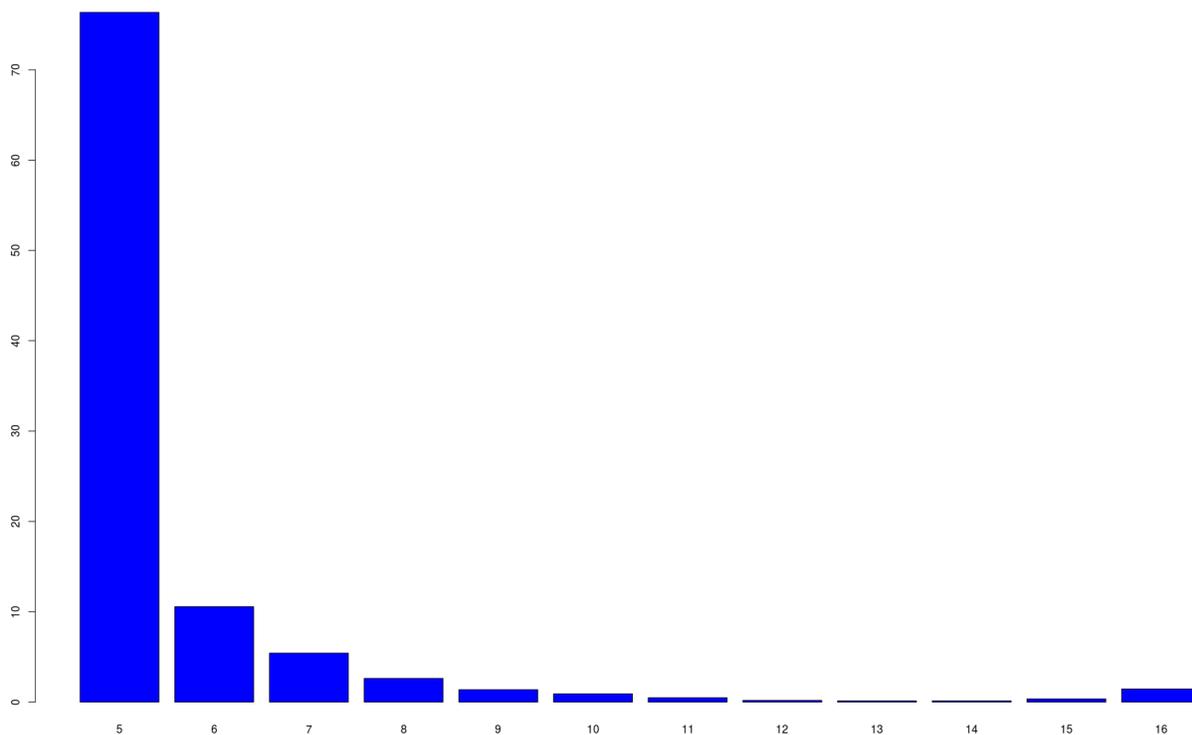
An interesting aspect of the original study was to see how many keyphrases people use to describe a given scientific abstract. It was observed that for the *single-document case*, most people used five phrases, and the number gradually decreased thenceforth (cf. Figure 1 in the paper), tapering off at 16 phrases. This is why we chose 5-16 as our number of keyphrases in the Mechanical Turk study. In our case, the percentages look as in Figure 2b. Figure 2a is a duplicate of the first part of Figure 1 of the original. Note that the five-phrase peak is much higher in our case than in the original, and other peaks are much lower. This points to the fact that in most cases, Turkers preferred not to spend any time at all on providing extra information (i.e., optional information). The general shape of the histogram also closely follows the original, thereby validating the authors’ comments: “the peak at five

¹⁵Our SemEval corpus contains 6.76 sentences on average (median = 6, standard deviation = 2.42).

¹⁶Please see Section 3.2 for the definition of technical terms and compound technical terms.



(a) Original



(b) After replication

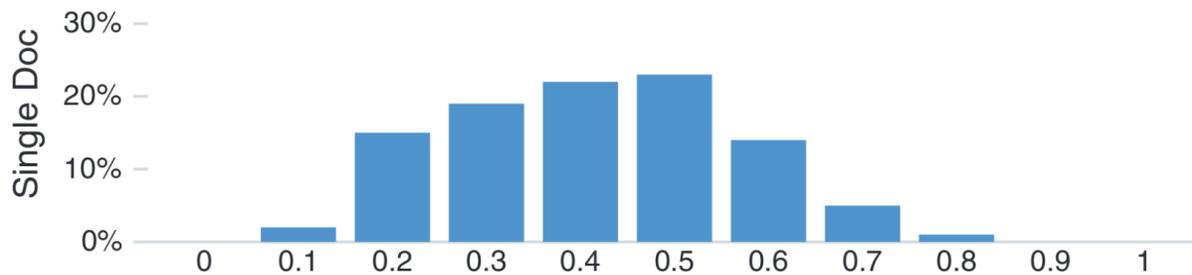
Figure 2: Distribution of the number of human-assigned keyphrases. X-axis is the number of keyphrases, and Y-axis is the percentage of documents/assignments with that number. See text for details.

...suggests that subjects might respond with fewer without this requirement. However, it is unclear whether this reflects a lack of appropriate choices or a desire to minimize effort.” In our case, the much sharper peak at five phrases, and much lower peaks at others suggest that the latter might be the real reason.

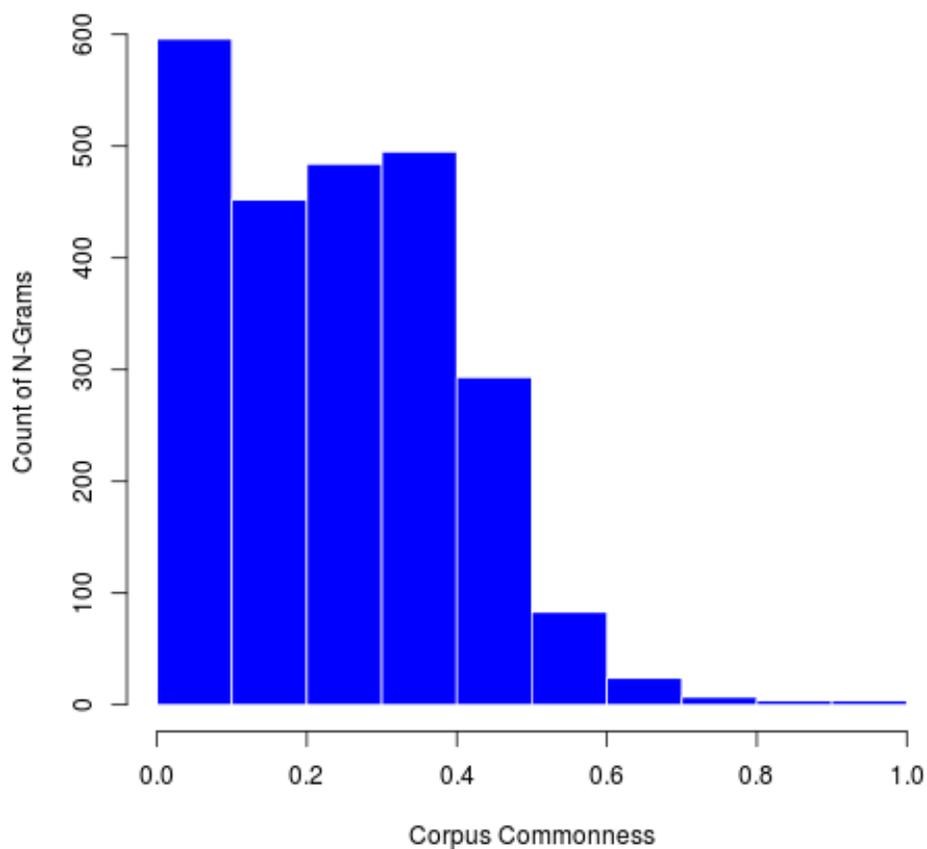
Another difference with the original is that since we have a varying number of keyphrases from each assignment (recall from Section 3.1 that an *assignment* is a Mechanical Turk term for a single piece of human intelligence task (HIT) to be completed by a single person in a single session), we plotted the histogram in Figure 2b *on the assignments* rather than on the abstracts (which presumably is the case with Figure 2a), thus potentially inflating the number of five-phrase responses, and deflating others.

Exploration of *commonness* was an important aspect of the paper.¹⁷ The authors found that annotators tend to prefer phrases that are not too rare, and not too common. They called those phrases *mid-commonness*. Validation of this argument came from Figure 3 of the paper, the first part of which is repeated in Figure 3a. Our replicated version appears in Figure 3b. There are several things to note here. First, we computed commonness on our background corpus (*Inspec*) rather than Web1T as was done in the paper (cf. Section 3.1). So, our commonness peaks are much higher than Web1T peaks (because on the Web, which is a huge corpus, all terms are sufficiently rare; whereas on *Inspec*, which is way smaller, most terms are not very rare). Second, we got a peak – the largest peak in fact – at the *lowest* (leftmost) commonness bin. The reason for this is that Web1T excludes extremely rare terms – terms that appeared less than 40 times [4] – which is why there is no peak at low commonness values. While this does make sense for an enormous background corpus like Web1T, for our much smaller *Inspec* corpus, deleting such terms does not seem to be worthwhile. Now, given that, it seems to be the case that the authors did miss a point: annotators do not prefer the *mid-commonness terms* as they suggested; they instead prefer the *rarest terms*. Third, we were able to re-construct the pattern of the original histogram (Figure 3a) by deleting the rarest terms from *Inspec* corpus (cf. Figure 4). This shows

¹⁷Please see Section 3.2 for the definition of commonness.

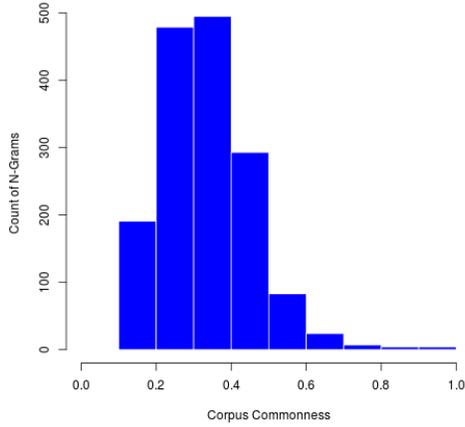


(a) Original (on Web1T)

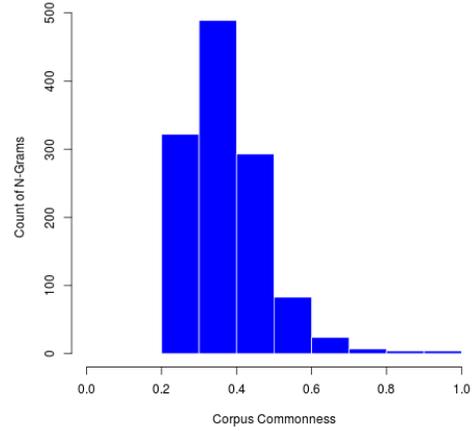


(b) After replication (on Inspec)

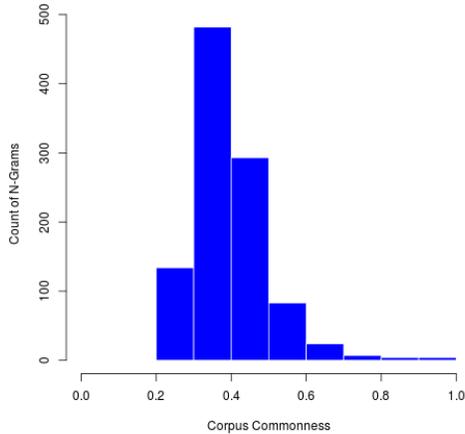
Figure 3: Distribution of commonness. X-axis is commonness, and Y-axis is the percentage/number of unique (human-selected) n-grams with that commonness.



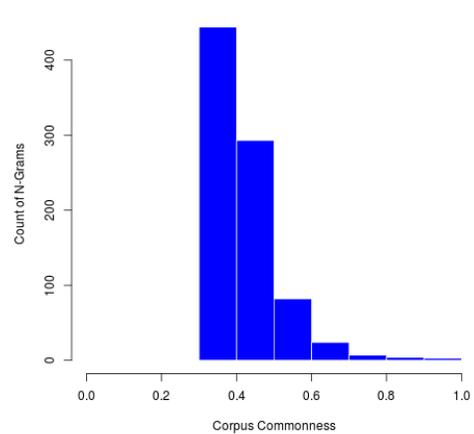
(a) All terms with frequency < 5 are removed.



(b) All terms with frequency < 10 are removed.

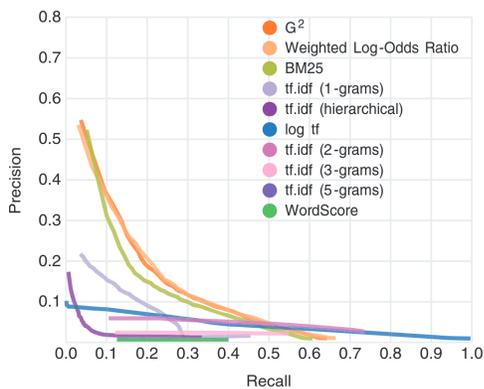


(c) All terms with frequency < 15 are removed.



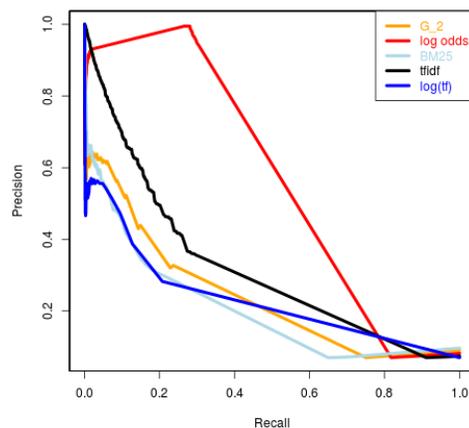
(d) All terms with frequency < 20 are removed.

Figure 4: Distribution of commonness on *Inspec* after the rarest terms are removed. X-axis is commonness, and Y-axis is the number of unique (human-selected) n-grams with that commonness. Note that as we remove more and more rare terms, the figures look more and more similar in shape to Figure 3a – but up to a point. Figure 4d, for example, has removed so many of the rare terms that the general *mid-commonness* shape disappears again.



(a) Frequency statistics.

(a) Original



(b) After replication

Figure 5: Performance of frequency and probabilistic features (best viewed in color).

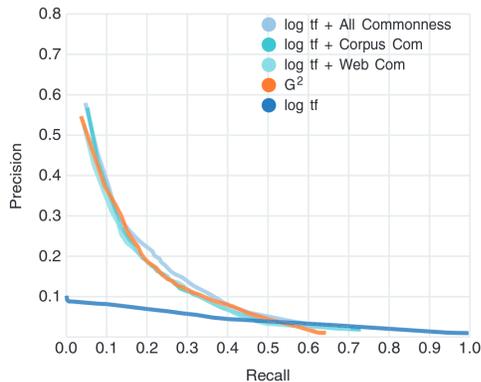
that at least for *corpus commonness*, rarest terms are preferred to mid-frequency terms. Of course, if we remove the rarest terms, then the next highest preferred set of terms comes from mid-commonness range.

4.2 Model Performance

Recall from Section 3.3 that we constructed logistic regression models on the following features:

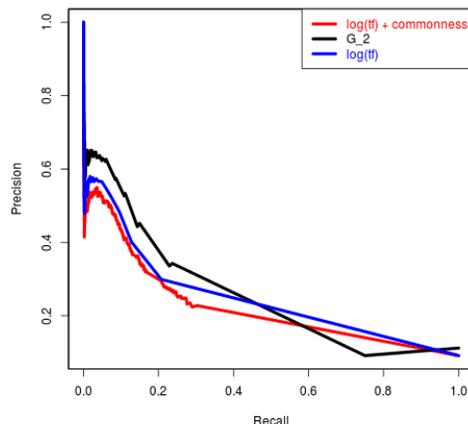
- Frequency and probabilistic features.
- Adding commonness to raw term frequency, and comparing the joint model with raw term frequency and G^2 .
- Adding grammatical features to frequency and commonness features.
- Adding positional features (to the rest of the features) to create the final models.

Performance was evaluated using precision-recall curves (i.e., area under the curves). In the next set of figures, we shall illustrate the results one by one, along with the original. In



(b) Adding term commonness.

(a) Original



(b) After replication

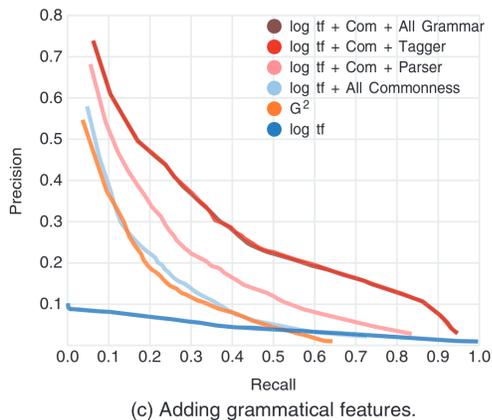
Figure 6: Adding term commonness to term frequency (best viewed in color).

the process, we shall also show evidence in favor of the second and third part of our final hypothesis (cf. Section 2).

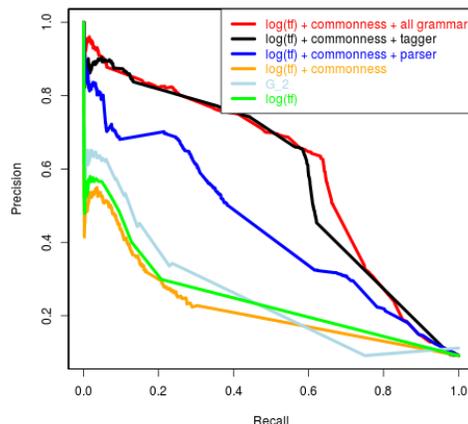
Performance of frequency and probabilistic features is given in Figure 5. Note that we dropped WordScore (cf. caption of Table 2), and did not experiment with the *hierarchical* version of *tfidf*, so we have only five plots in Figure 5b instead of ten in Figure 5a. In Figure 5b, log-odds ratio outperforms $\log(\text{tf})$ and *tfidf* in terms of area under the curve. **This validates the second part of our final hypothesis (Section 2).** The difference with Figure 5a is that in Figure 5a, probabilistic features (G^2 , BM25, log-odds ratio) outperformed all other frequency features. This did not happen in our case. We observed that while log-odds ratio gives a very good performance (see the red curve in Figure 5b), other probabilistic features such as G^2 and BM25 are much worse – sometimes even worse than $\log(\text{tf})$.

Adding term commonness (*corpus commonness* based on the Inspec background corpus) did not help much in our case (as shown in Figure 6b), unlike the original version (Figure 6a). Since the original paper experimented with two forms of commonness (web and corpus), they have five plots in total, whereas we have only three.

Adding grammatical features on top of commonness features, however, gives a big boost



(a) Original



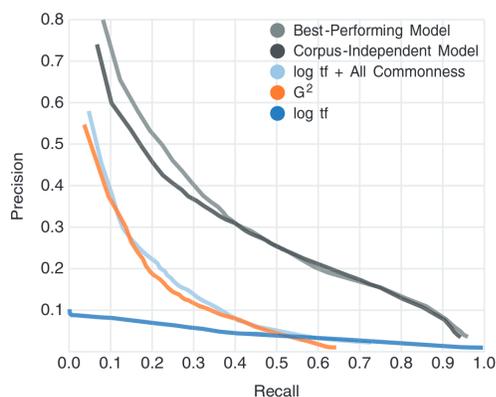
(b) After replication

Figure 7: Adding grammatical features on top of term commonness (best viewed in color).

in performance. Note from Figure 7b that adding grammatical features on top of $\log(\text{tf})$ and commonness outperforms $\log(\text{tf}) + \text{commonness}$ combination, $\log(\text{tf})$ itself, as well as the probabilistic feature G^2 . Similar patterns are observed in Figure 7a from the original. Furthermore, the $\log(\text{tf}) + \text{commonness} + \text{all grammar}$ combination (red curve in Figure 7b) – which includes parse-tree-based features – gives very little additional improvement over the black curve ($\log(\text{tf}) + \text{commonness} + \text{part-of-speech features}$). This is again similar to the original. **This validates the third and final part of our final hypothesis (Section 2).**

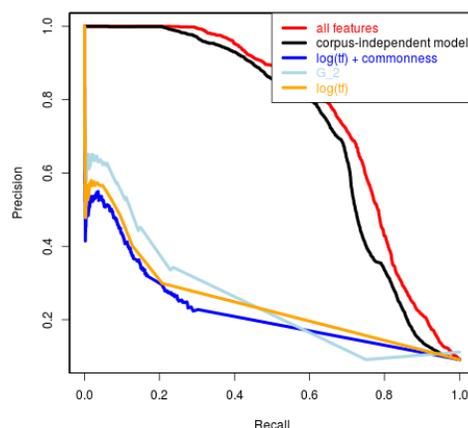
Lastly, we added positional features on top of all other features (the best-performing model) to give a final boost in performance. Results look similar to the original. Compare Figure 8b with Figure 8a. *All features outperform corpus-independent model* (by a thin margin), which in turn outperforms G^2 , $\log(\text{tf})$, and $\log(\text{tf}) + \text{commonness}$ combination. Note that the corpus-independent model only includes $\log(\text{tf})$, positional, and part-of-speech features, and is therefore independent of the *background corpus*.

The authors mentioned *binning* commonness values to see if it has an effect on performance. They did not report precision-recall curves on this, but we went ahead and carried out the experiment nonetheless. Note that our background corpus (Inspec) is way smaller



(d) Adding positional features.

(a) Original



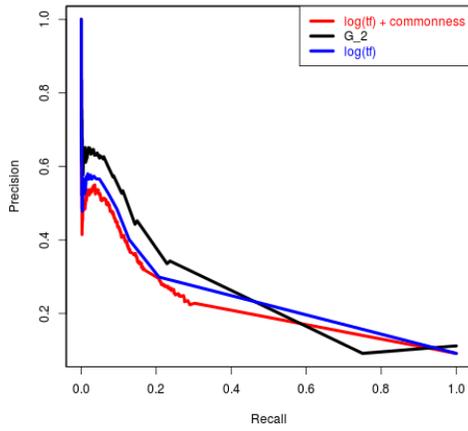
(b) After replication

Figure 8: Adding positional features on top of all other features (best viewed in color).

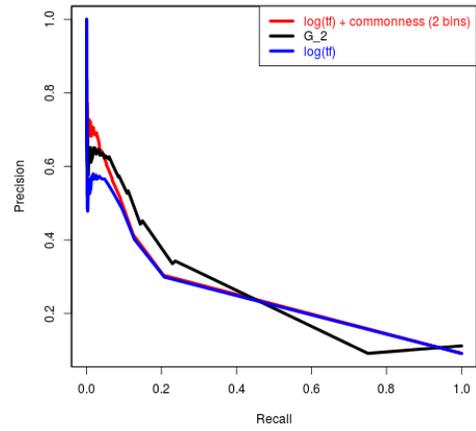
than Web1T, so we are unlikely to see a lot of performance improvement from commonness binning, simply because we do not have a large range of commonness values (like Web1T would have, for example) to bin over. Indeed, we did not see any performance improvement from commonness binning. In fact, performance degraded when the number of bins increased beyond a point. The relevant figures are shown in Figure 9 through Figure 12. We start with the “adding term commonness” model of Figure 6b (no binning; repeated in Figure 9a), then gradually increase the number of bins to 2, 3, 4, \dots , 20. As we can see, performance increases a little when the number of bins is 2, then starts to degrade, and finally reaches a stable (but degraded) configuration when the number of bins is around 6-7. In short, binning did not help much in our case.

5 Conclusion

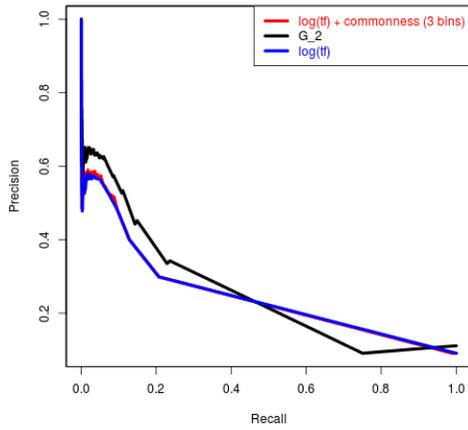
Keyword extraction is a very broad and burgeoning field in Natural Language Processing, and while we do have several references in our bibliography that point to the relevant literature in question, we could by no means justify the exposition of such a vast literature within the scope of this report. There does seem to be very little work done in the specific area of *descriptive*



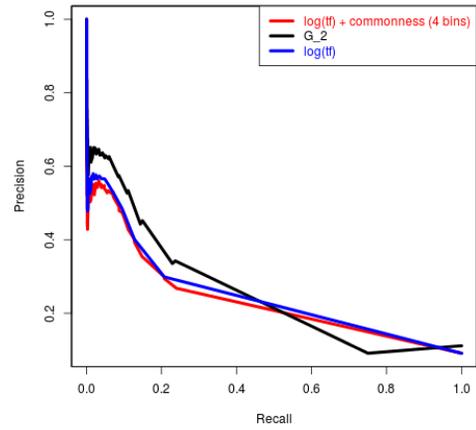
(a) No binning



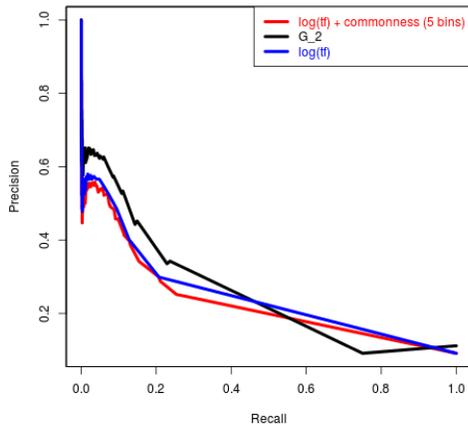
(b) 2 bins



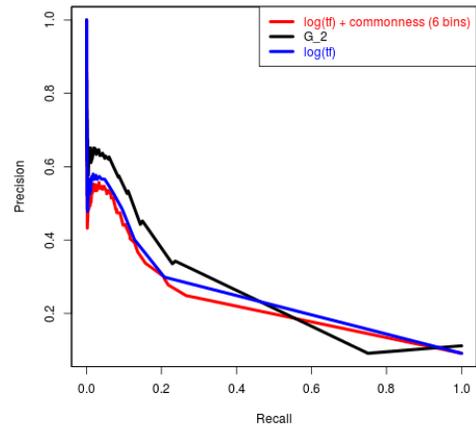
(c) 3 bins



(d) 4 bins

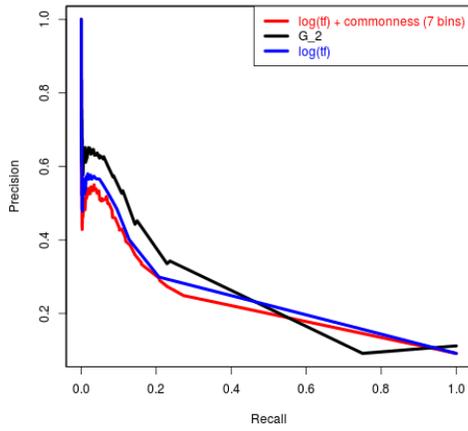


(e) 5 bins

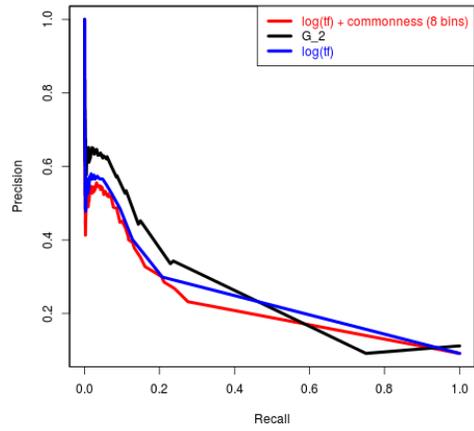


(f) 6 bins

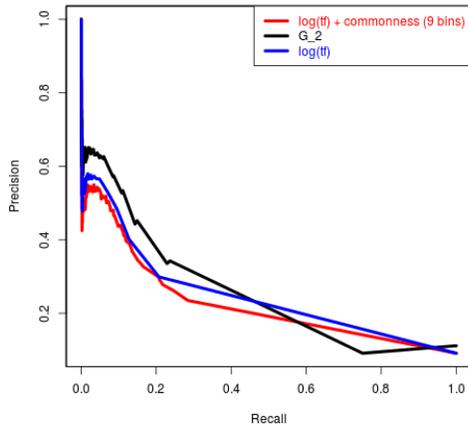
Figure 9: Binning commonness (best viewed in color).



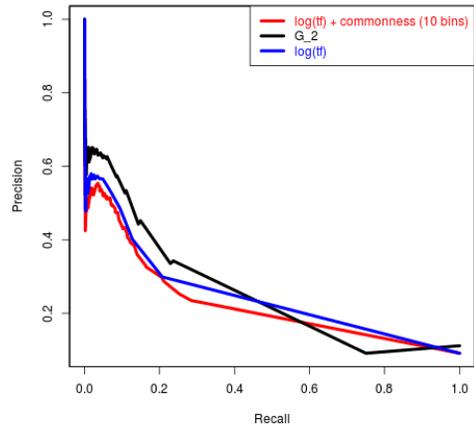
(a) 7 bins



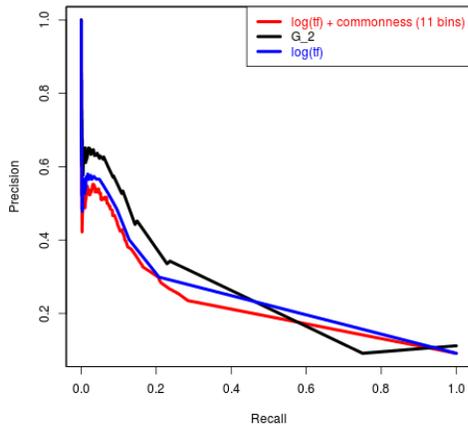
(b) 8 bins



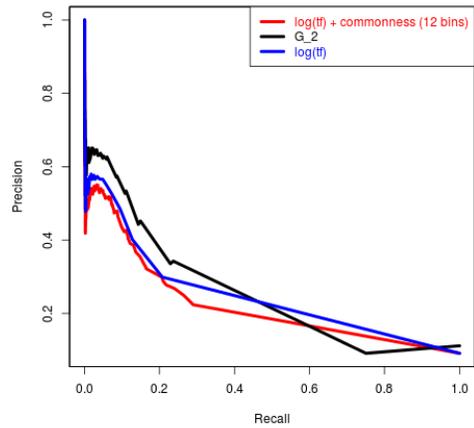
(c) 9 bins



(d) 10 bins

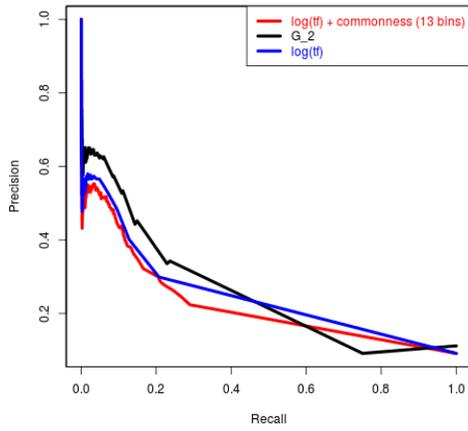


(e) 11 bins

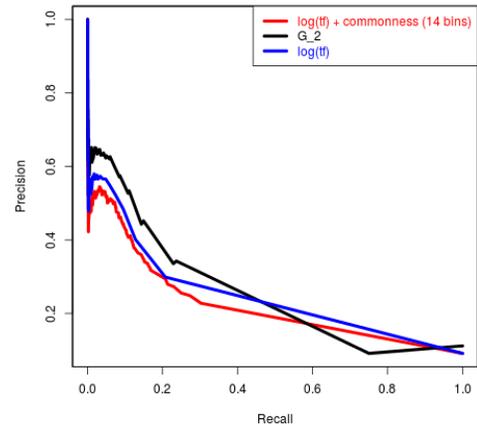


(f) 12 bins

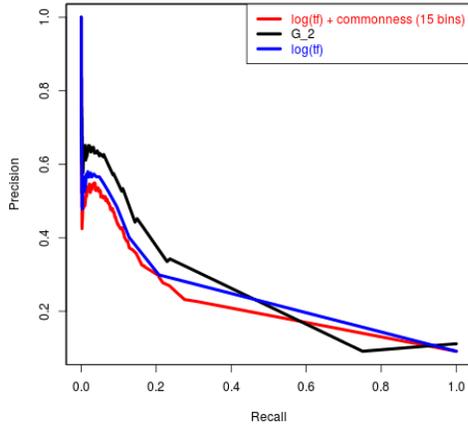
Figure 10: Binning commonness (best viewed in color).



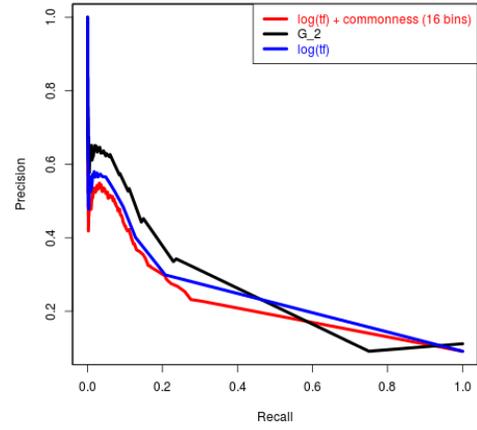
(a) 13 bins



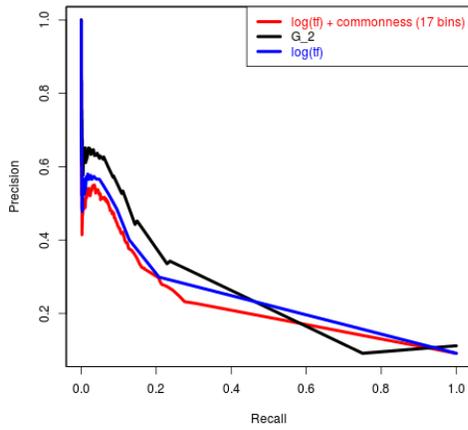
(b) 14 bins



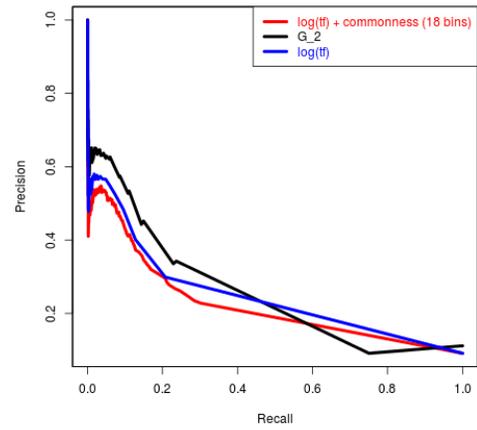
(c) 15 bins



(d) 16 bins



(e) 17 bins



(f) 18 bins

Figure 11: Binning commonness (best viewed in color).

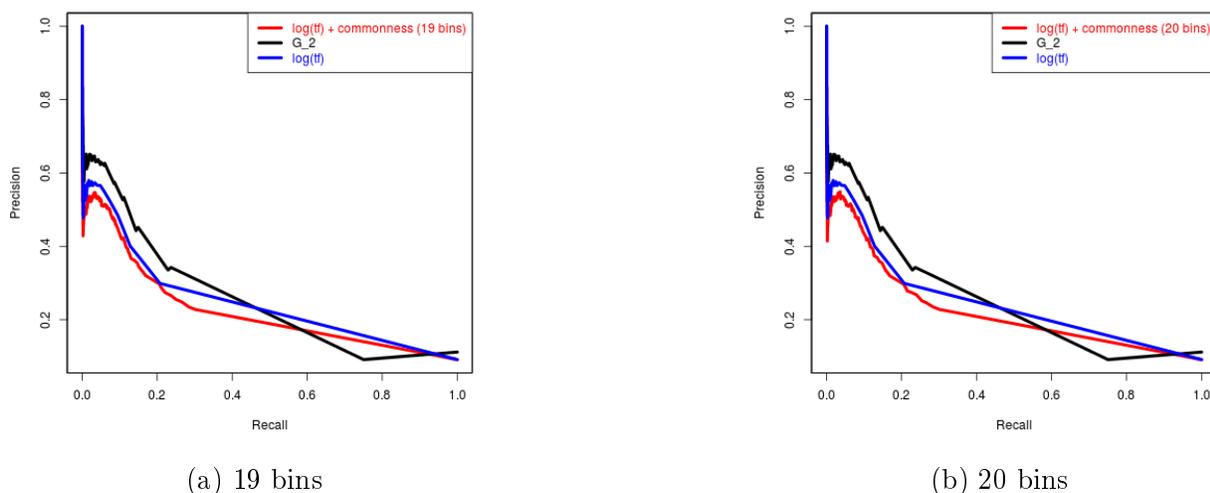


Figure 12: Binning commonness (best viewed in color).

keyphrase extraction, which is the problem tackled by the authors of this paper. There have been some work in information visualization that cited this paper, but none in particular looked at *descriptive keywords*. Our contributions include a re-evaluation and re-assessment of three key hypotheses presented in the paper – one related to the exploratory analysis of human-generated keyphrases, and two related to the types of features that are good at identifying descriptive keyphrases. Along the way, we also pointed out several differences between our replication study and the original, uncovered some potential limitations of the latter (esp. with respect to *web commonness* histograms), and extended the study by investigating the impact of binning *corpus commonness*. We believe that our efforts will result in the elicitation of interest in this nascent body of work that attempts to integrate two apparently disparate lines of research – Natural Language Processing and Human-Computer Interaction – into a unified whole.

References

- [1] Gábor Berend and Richárd Farkas. SZTERGAK : Feature Engineering for Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, July 2010.

- [2] Thorsten Brants and Alex Franz. Web 1T 5-gram corpus version 1. Technical report, Google Research, 2006.
- [3] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. “Without the Clutter of Unimportant Words”: Descriptive Keyphrases for Text Visualization. *ACM Trans. Comput.-Hum. Interact.*, 19(3):19:1–19:29, October 2012.
- [4] Stefan Evert. Google Web 1T 5-grams Made Easy (but Not for the Computer). In *Proceedings of the NAACL HLT 2010 Sixth Web As Corpus Workshop, WAC-6 '10*, pages 32–40, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [6] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. Extracting Key Terms From Noisy and Multi-theme Documents. In *18th International World Wide Web Conference*, pages 661–670, April 2009.
- [7] Kazi Saidul Hasan and Vincent Ng. Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 365–373. Association for Computational Linguistics, 2010.
- [8] Anette Hulth. 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, 2003. Association for Computational Linguistics.
- [9] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 21–26, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

- [10] Dan Klein and Christopher D. Manning. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics.
- [11] Shibamouli Lahiri, Sagnik Ray Choudhury, and Cornelia Caragea. Keyword and Keyphrase Extraction Using Centrality Measures on Collocation Networks. *CoRR*, abs/1401.6571, 2014.
- [12] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [13] Patrice Lopez and Laurent Romary. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10*, pages 248–251, 2010.
- [14] Vanessa Loza, Shibamouli Lahiri, Rada Mihalcea, and Po-Hsiang Lai. Building a Dataset for Summarization and Keyword Extraction from Emails. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).
- [15] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June 1993.
- [16] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

- [17] Jason D. M. Rennie and Tommi Jaakkola. Using Term Informativeness for Named Entity Detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–360. ACM, 2005.
- [18] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):7881, 2005.
- [19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- [20] Takashi Tomokiyo and Matthew Hurst. A Language Model Approach to Keyphrase Extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 33–40. Association for Computational Linguistics, 2003.
- [21] Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, and Bo Wang. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information Systems*, 2008.