Extractive Summarisation Based on Keyword Profile and Language Model

Han Xu **Eric Martin** **Ashesh Mahidadia**

School of Computer Science and Engineering
UNSW, Sydney, NSW, Australia, 2052
hanx,emartin,ashesh@cse.unsw.edu.au

Abstract

We present a statistical framework to extract information-rich citation sentences that summarise the main contributions of a scientific paper. In a first stage, we automatically discover salient keywords from a paper's citation summary, keywords that characterise its main contributions. In a second stage, exploiting the results of the first stage, we identify citation sentences that best capture the paper's main contributions. Experimental results show that our approach using methods rooted in quantitative statistics and information theory outperforms the current state-of-the-art systems in scientific paper summarisation.

1 Introduction and Motivation

Science is not an isolated endeavour, but benefits from and expands on the work of others, with more or less cross fertilisation between disciplines. The interdependent nature of research has naturally resulted in a network of scientific areas with dense interconnections between related fields. Though research is a highly specialised activity, researchers find themselves constantly in need to explore the network further from the core of their research. Tools that can facilitate understanding the key contributions of papers in those parts of the network being explored can only prove highly valuable.

As an example of such tools, we focus on an application that automatically extracts information-rich sentences describing the main contributions of a given paper. From which corpus the extraction could take place? A natural answer is the abstract of the paper. However, the contributions as perceived by the authors can significantly deviate from those judged extrospectively by the community over time (Mei and Zhai, 2008). Instead, we take as corpus the set of citing sentences to the paper (from other papers). Indeed, those sentences can arguably be deemed as a form of crowd-sourced review of the paper's main contributions. The set of citing sentences is referred to as the *citation summary* of the target paper. Elkiss et al. (2008) carried out a large-scale study and confirmed that citation summaries contain extra information that does not appear in paper abstracts. In addition, they found that the "self-cohesion", measured as the average cosine similarity between sentences, is consistently higher in a paper's citation summary than in its abstract: the former is more focused than the latter in describing papers' main contributions. This work presents our efforts in advancing research along this direction.

Section 2 formally defines the problem we aim to solve: summarise scientific papers using the most informative and diversified part of their citation summaries. It surveys several prominent related studies, and introduces the data used in our experiments and evaluations. In Section 3, we present our statistical framework built upon quantitative statistics and information theory. In Section 4, we evaluate and compare the performance of our method with state-of-the-art systems. We conclude and point to future directions in Section 5.

2 Problem Statement

The problem we tackle in this paper is to generate an *extractive summary* (usually, we will simply say *summary*) from its citation summary. More specifically, we opt for a two stage approach. In the first stage, we automatically discover salient keywords from a paper's citation summary, keywords that are essential in characterising the paper's main contributions. The second stage, exploiting the results of the first stage, identifies citation sentences (to the paper) that best capture the paper's main contributions.

A word of caution: by utilising only citation summaries, one should not expect to obtain well formulated, readily consumable summaries of papers. Indeed, a citation sentence may be not all about the cited paper, but also talk about the citing paper and other co-cited papers, which disqualify citation sum-

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 123–132,
Denver, Colorado, May 31 – June 5, 2015.

maries as a premium source of sentences for building highly readable summaries (Siddharthan and Teufel, 2007). Moreover, a summary built from citing sentences that come for a pool of multiple citing papers is bound to lack coherence. Therefore, it is more appropriate to consider that the output of such a system is to extrinsically gauge a system's effectiveness in indexing information-rich citing sentences containing keywords that facilitate rapidly grasping a paper's important contributions, rather than be treated as a polished, readable summary for human consumption (Qazvinian et al., 2013).

2.1 Related Work

Qazvinian and Radev (2008) first experimented with citation summary based paper summarisations. They proposed a graph-based method, C-LexRank, that first generates a citation summary network for a paper by mapping citing sentences to vertices and creating edges from their lexical similarities. Clusters of sentences capturing the same contribution of the paper are then identified through link-based community detection. Finally, the most central sentence of each cluster is found using a weighted random walk and selected to form a paper summary meant to comprehensively cover the paper's main contributions. Mohammad et al. (2009) further adapted the C-LexRank to multi-document summarisation in an attempt to generate surveys for scientific paradigms.

In a later paper, Qazvinian et al. (2010) proposed a more computationally efficient summariser that does not require clustering citing sentences. As a first step, key phrases are automatically identified as significant n-grams with positive point-wise divergence (Tomokiyo, 2003) from a foreground language model estimated using the citation summary of a paper w.r.t. a background language model built from a large set of paper abstracts. A greedy algorithm is subsequently applied to select citing sentences and form a summary that maximises key phrase coverage.

Mei and Zhai (2008) presented a sophisticated generative approach that frames summarisation under an Information Retrieval (IR) context. Specifically, an impact language model for a paper is first built as a mixture of a language model estimated from the paper's own text, and a weighted citation language model based on its collective citation contexts, using a compound coefficient reflecting both a sentence's proximity to the citation label (anchor) in the citing paper and the citing paper's authority calculated from the citation network using PageRank (Brin and Page, 1998). Finally, documents (sentences in the target paper) that are closest to the query (the impact language model of the target paper) are extracted to form a summary using ad-hoc document retrieval. Note that Mei and Zhai (2008) utilised extra information (i.e., paper full texts and citation networks) to produce summaries that consist of sentences from papers' own texts rather than their citation summaries, making their task related to but different to ours.

2.2 Data

The experiments and evaluations presented here have been based on Qazvinian's single paper summarisation corpus[1]. The dataset consists of 25 highly cited papers in the ACL Anthology Network (AAN) (Radev et al., 2009) from 5 different domains: Dependency Parsing (DP), Phrase Based Machine Translation (PBMT), Text Summarisation (SUM), Question Answering (QA) and Textual Entailment (TE). There are two files provided for each paper: a citation summary file containing all citing sentences to it, and a manually constructed key fact file containing its main contributions hand picked by human annotators after reading the citation summary. The manual annotation has been performed independently by annotators, and a phrase needed to be marked by at least 2 annotators to be qualified as capturing a paper's key fact (Qazvinian and Radev, 2008). This corpus represents a gold standard in research paper summarisation and it has been widely used in system evaluations (Qazvinian and Radev, 2008; Qazvinian et al., 2010).

3 Our Approach

In this section, we first introduce our quantitative statistical method to automatically construct a *keyword profile* of a paper and statistically capture a paper's main contributions in terms of words from its citation summary. We then discuss how we construct a *keyword profile language model*. Finally, we elaborate on how we cast the task of sentence selection from the citation summary as language model divergence based IR in a probabilistic framework.

3.1 Paper Keyword Profile

As indicated in Section 1, the citation summary of a paper can be deemed a collective review of its contributions. Therefore, the main contributions of a

[1] http://www-personal.umich.edu/~vahed/data.html

paper are salient keywords, those keywords which are commonly used by its citers to refer to it and are statistically over-represented in the paper's citation summary w.r.t. the overall distribution of such words across other papers' citation summaries. Put another way, the salience of a word in characterising a paper's main contributions is qualified along *over-representedness* and *exclusiveness* dimensions. Clearly, a proper statistical model of words distribution is required in order to measure words' salience in a paper's citation summary. Consider five papers, $D_1, \ldots, D_5$ with citation summaries $CS_1, \ldots, CS_5$. We aim at identifying salient keywords from D_1's citation summary CS_1, that map to D_1's main contributions. To decide whether a word W is a characterising keyword of D_1, we first collect all n citing sentences containing W from $CS_1, \ldots, CS_5$; suppose there are $n = 20$ of them. Then for each citing sentence S amongst those 20, we perform the binary test: success iff S belongs to CS_1. Suppose that there are $k = 18$ successes and 2 failures. This represents a surprising observation: one would expect a word of no characterising power to appear in roughly the same number of sentences in $CS_1, \ldots, CS_5$, assuming all citation summaries have the same number of sentences[2]. So one would heuristically conclude that W is a good candidate keyword for D_1, a keyword that is likely to represent a main contribution.

The previous process can be abstracted as sampling without replacement from a finite set whose elements can be classified into mutually exclusive binary categories, which itself follows a Hypergeometric distribution. Let N be the total number of citing sentences in citation summaries for papers belonging to collection C, K be the number of sentences in paper D's citation summary, n be the total number of citing sentences containing a certain word W, and X be the number of citing sentences containing W in D's citation summary. The probability of observing exactly k citing sentences in D's citation summary containing W is:

$$H(X{=}k|N,K,n){=}\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad (1)$$

We can then calculate a p-value to the observed number of x citing sentences in D's citation summary that contain word W using the Hypergeometric test, which in turn is used to measure word W's salience in characterising D's main contributions:

$$S(W) \stackrel{def}{=} P(X{\geq}x){=}1{-}\textstyle\sum_{i=0}^{x-1} H(X{=}i|N,K,n)) \quad (2)$$

[2]This assumption is only made to simplify the discussion.

The smaller the value of $S(W)$, the more salient W is. Also, words not appearing in D's citation summary have a maximum p-value of 1.0, and common words appearing in many papers' citation summaries are expected to have larger p-values than words that are more exclusively used when citing paper D.

It is worth pointing out that the above formulation can be equivalently expressed as applying the one-tailed Fisher's exact test to measure strengths of statistical associations between words and paper's citation summaries at the sentence level. Our choice of this statistical procedure has been informed by (Moore, 2004). Prior to this work, Dunning (1993) was pointing out that some commonly used methods such as the Pearson's χ^2 test are inappropriate for measuring textual associations due to the fact that the underlying normality assumption is usually violated in textual data. He was subsequently introducing the log-likelihood ratio test (LLR) and showing that it can yield more reliable results. The LLR was then and has since been widely adopted in statistical NLP as a measure of strength of association (Moore, 2004). For instance, Lin and Hovy (2000) successfully applied LLR in mining "topic signatures" of pre-classified document collections. But to further verify LLR's validity applied to rare events, Moore (2004) performed an empirical study comparing results obtained using LLR and Fisher's exact test on bilingual word association and found that albeit being a good approximation to Fisher's exact test, LLR can still introduce a substantial amount of error and the author went on to advocate the use of Fisher's exact test where computationally feasible. Recall that we measured associational strengths at the sentence level. This resulted in marginal frequencies in the order of only hundreds for Qazvinian's small corpus. We therefore followed this empirical advice and used the one-tailed Fisher's exact test (i.e., Hypergeometric test) as our measure of textual association to perform keyword profiling of a scientific paper.

To obtain a set of keywords likely to map to a paper's main contributions, one can simply sort all words according to their statistical significance and pick the top few (e.g., 10 words with the smallest p-values). A more statistically tenable scheme would be to identify the keywords of a paper as all words appearing in its citation summary with p-values below some significance level. A technicality here is that in the identification of keywords, multiple Hypergeometric tests have been performed. For example, all unique words that appeared in the collection

of citation summaries have been individually tested for their salience in a target paper's citation summary in succession. The significance level used to qualify a word as a keyword thus requires correction for multiple tests to reduce type I errors. However, we shall show that the rigid statistical significance is not crucial in our subsequent building of a keyword language model for a paper, and so we did not perform multiple tests corrections, but simply used the raw p-values in subsequent analysis.

Another technicality is special handling of citation anchors. Cited authors' names, almost systematically appearing in citing sentences, are bound to be identified as salient keywords. We thus substituted all citation anchors appearing in a paper's citation summary with the pseudo token "targetanchor" if they refer to the target paper, and "otheranchor" if they refer to other co-cited papers.

Furthermore, our keyword profiling approach allows for a flexible control of the level of selectiveness in its statistical procedure through the choice of the benchmarking collection C. For example, we can choose to use a heterogeneous collection of papers covering multiple domains. Words that are salient in characterising a domain may then evaluate to a high salience for a paper in C on that domain (e.g., word "parsing" for domain Dependency Parsing (DP)). We can also choose C to be a homogeneous collection of papers from the same domain. Only words that are salient in characterising a single paper will then be evaluated to a high salience for that paper (e.g., if C is on DP, "parsing" will not show up as a salient word for any paper in C). Recall from Section 2.2 that we use as data papers from five domains. We exploited the homogeneity of this data and performed keyword profiling intra-domain. This effectively made the keyword profiling all the more selective that the keywords identified for a paper only characterise its *unique* contributions w.r.t. its domain, using five highly cited papers. We shall show in the next section that it is this high selectiveness in keyword profiling that bestows our approach its high discriminative power.

For paper P05-1013, Table 2 lists the top 10 keywords identified from its citation summary using our method, while Table 1 lists the humanly selected gold standard key facts (Qazvinian and Radev, 2008). It can be seen that our method is highly effective in identifying the paper's main contributions which closely mirror those picked by human experts. We term our word list ranked by p-values the *keyword profile* of the paper; it statistically and objectively captures words' salience (measured along the dimensions of over-representedness and exclusiveness) in characterising the paper's main contributions using the statistical surprise given by Hypergeometric tests. While only unigram keywords were considered here, our method can be easily extended to cope with higher order n-gram "key phrases". This is left for future work.

Fact id	Fact	Occurence	Pyramid tier
1	non-projective pseudo-projective projectivizing projective graphs projectivization transformation	15 6 1 1 1	19
4	czech swedish	6 5	8
2	data-driven training data	4 2	6
5	maltparser	4	4
3	nonterminal categories in constituency	1	1

Table 1: Gold standard key facts of P05-1013 (Qazvinian and Radev, 2008) ordered by importance. The pyramid tier might not be the sum of the occurrences of facts, as multiple facts can appear in the same sentence.

Salience rank	Word	P-value
1	non-projective	1.54e-08
2	pseudo-projective	5.61e-06
3	transformation	4.47e-05
4	transformations	1.26e-04
5	maltparser	3.48e-04
6	swedish	7.53e-04
7	danish	1.56e-03
8	following	2.64e-03
9	arcs	2.64e-03
10	dependencies	4.43e-03

Table 2: Extracted keywords for P05-1013, ranked by decreasing Hypergeometric test significance.

3.2 Keyword Profile Language Model

Each sentence in a paper's citation summary covers keywords (possibly none) that map to the paper's main contributions. Intuitively, a good summarisation should be short, and consist of citing sentences that maximise keywords coverage w.r.t. an arbitrarily imposed summary length limit (Qazvinian and Radev, 2008). A good summariser should thus pick citing sentences that contain as many non-redundant keywords as possible. We have shown in the last sec-

tion that not all keywords are of equal importance, so a good summariser should favour sentences covering the most important ones. Intuitively, the keyword profile of a paper containing valuable information on words' salience in characterising the paper's main contributions should be utilised to drive such a discriminative sentence selection process.

Based on the previous considerations, we use a paper's keyword profile to build a discriminative unigram language model that directly encodes words' salience as pseudo generative probabilities to facilitate the seamless incorporation of such information into a generic probabilistic framework. More specifically, we directly translate words' salience (in the form of p-values) into a discriminative unigram language model of a paper that assigns high probabilities to its characterising keywords. The pseudo generative probability of word W according to a paper D's keyword profile language model M_{kp} is:

$$P(W|M_{kp}) = -\frac{1}{Z}\log(S(W)) \quad (3)$$

where $S(W)$ denotes the salience of word W in characterising paper D calculated using (2), and Z is a normalisation factor. An intuitive interpretation of (3) is to deem $-\log(S(W))$ a pseudo word count of W, where more salient words have higher pseudo counts; this makes Z the total length of the pseudo document generated from the paper's keyword profile. We disregard actual word counts to make the keyword profile language model directly encode words' salience. Also, in the previous step, keyword profiling had already implicitly taken such information into account, providing another justification for this design decision. Table 3 shows a miniature example to illustrate how a keyword profile language model is built. In this example, W_5 is automatically eliminated from the resulting language model because it has lowest salience in characterising the imaginary document. Any word S with salience value $S(W)$ close to but strictly less than 1.0 would still have a tiny pseudo probability in the resulting keyword profile language model (e.g., W_4). Words with low salience are not necessarily stop words (*e.g.*, W_4 and W_5), and neither is the reverse true: a content word can possibly be used across the document collection and thus evaluate to a very low salience (and so have a nul or low pseudo generative probability in the resulting keyword profile language model) for the document under consideration. For example, "parsing" would have a low salience for any paper in a collection on Dependency Parsing. It can be seen that our method amounts to a highly adaptive data driven term weighting framework. For brevity, from now on, we use KPLM to refer to keyword profile language model.

Word	**Salience** $S(W)$	**Pseudo count** $-\log(S(W))$	$P(W\|M_{kp})$
W_1	0.01	4.61	0.605
W_2	0.10	2.30	0.303
W_3	0.50	0.69	0.091
W_4	0.99	0.01	0.001
W_5	1.00	0.00	0.000

Table 3: Keyword profile language model built for an imaginary document consists of only 5 distinct words.

Although implicitly conveyed in the formulation of KPLM above, it should be made clear that the KPLM is a pseudo language model that encodes words' salience in the form of pseudo generative probabilities, which functions as a language model, yet should not be interpreted as a true language model under the traditional definition. A traditional unigram language model is constructed using the actual term frequencies in the document, the resulting model capturing generative probabilities. In contrast, the KPLM of a document is built using pseudo term frequencies that directly encode words' salience in characterising a document's contents, measured using a sophisticated quantitative statistical procedure. It can thus be interpreted as a probabilistic description of the document's keywords with significantly boosted discriminative power. Having clarified the nature of KPLM, we treat it as a language model in the rest of the paper.

3.3 KPLM Based Summarisation

3.3.1 Sentence Selection

The KPLM of a paper is a discriminative generative model that incorporates words' salience in characterising a paper's main contributions. It thus represents an effective language model from which a model citing sentence covering the paper's main contributions could be sampled from[3]. So by measuring the statistical surprise between the realistic language model estimated from each citing sentence with the KPLM of a paper, we can select the set of citing sentences that conform best to the optimal model given by the the KPLM and build a summary that well captures keywords. More specifically, we adopt the negative cross entropy retrieval model (Zhai, 2008), use the KPLM of a paper as the

[3] A pseudo citing sentence sampled from KPLM in this manner would simply be a bag of words, not a grammatical sentence. So here "model" has the favour of keywords coverage.

sole document model, and measure the cross entropy of multiple query models from it (one for each citing sentence in that paper's citation summary). Citing sentences whose Maximum Likelihood Estimation (MLE) language models are closest to the paper's KPLM are taken as building blocks of the summary.

Formally, let S be a citing sentence and let $c(W, S)$ denote the number of occurrences of word W in S. The MLE language model M_{mle} of S is the relative frequency of word W in S:

$$P(W|M_{mle}) = \frac{c(W,S)}{|S|} \quad (4)$$

Subsequently, the score for a citing sentence S is given by its negative cross entropy with the M_{kp}:

$$\begin{aligned} Score(S) &= -H(M_{mle}||M_{kp}) \\ &= \sum_{W \in V} P(W|M_{mle}) \log(P(W|M_{kp})) \end{aligned} \quad (5)$$

The larger a citing sentence's score, the closer it is to the cited paper's KPLM, thus the higher the citing sentence would be ranked. To summarise a paper, one can just pick the top k ranked citing sentences where k is the imposed summary length limit.

We are not the first to cast the task of summarisation as document retrieval. Mei and Zhai (2008) pioneered in utilising language models and divergence based IR to select sentences to build summaries. While similar in the fundamental methodology, our approach should be distinguished from this work. First, Mei and Zhai cast the task as ad-hoc retrieval, using the "impact language model" of a paper as sole query, while the paper's sentences are treated as documents whose Kullback-Leibler divergence (Kullback and Leibler, 1951) with the query model is measured in turn. Estimating reliable language models for short documents is challenging due to data sparseness and thus requires prudent smoothing. We purposefully reversed the roles of sentence model and document model, using the shorter sentences as queries and measuring their cross entropy with a sole document model (the KPLM)[4]. This represents a more natural formulation resulting in simpler language models that require fewer parameter estimations. Second, while the impact language model in (Mei and Zhai, 2008) is partially weighted using citing paper authority and sentence proximity to the citation anchor in the citing paper, it is still largely based on actual word occurrences. In contrast, KPLM directly models words' salience in characterising a paper's main contributions using its keyword profile, with expectedly more discriminative power. Last, Mei and Zhai's estimation of an impact language model for a paper assumes the reliable estimation of its citing papers' authority, which cannot always be guaranteed, for example when a paper receives citations from new papers that themselves have not been cited enough. Furthermore, while a citation network can be unavailable, the estimation of KPLM requires only the citation summaries of papers, which is arguably more robust.

[4]Kullback-Leibler divergence, used in (Mei and Zhai, 2008), is unsuitable to our task, as it is not formalised as ad-hoc retrieval (i.e., single query, multiple documents). Instead we compare multiple query models (MLE's of citing sentences) to a single document model (KPLM of the cited paper), making KL-divergence scores not comparable due to query specific entropy terms. See (Zhai, 2008) for a detailed analysis.

3.3.2 Top Sentence Re-ranking

As discussed in Section 3.2, a good summary should capture the most salient keywords of a paper, but also cover as many non-redundant keywords as possible. A summary built using our method is likely to contain citing sentences that concentrate on and repetitively cover salient keywords of the target paper, which may fall short in keywords diversity. Indeed, we can see in the top part of Table 4 that the summary of paper P05-1012 repetitively covers a single keyword, "Minimum Spanning Tree", while it fails to capture other key concepts.

To leverage the diversity in keywords captured in a summary, a simple heuristic is to select the next sentence from a pool of top ranked sentences least similar to the existing summary. From an information theoretic point of view, this amounts to choosing the next sentence that carries the most *extra* information (i.e., statistical surprise), w.r.t. the current contents of the summary. This formulation intuitively suggests that cross entropy, as a natural measure of statistical surprise, could again be employed.

We first need to abstract a citing sentence and the citation summary into probabilistic distributions before their cross entropy can be measured. Again we use unigram language modelling. Since both texts are small in size, data sparseness becomes a major issue, as nul dimensions in the MLE language models would make cross entropy not measurable. Smoothing as a way to alleviate data sparseness is thus required. Another issue that also arises from the texts' small size is the non-negligible amount of cross entropy contributed from non-content words in both texts (English stop words plus the two pseudo tokens: "targetanchor" and "otheranchor"). We therefore remove those non-content words prior to

language model construction to eliminate their noise in the cross entropy calculation. Experiments did support this design decision, and better results have been achieved with non-content words removed.

We perform Dirichlet Prior Smoothing (Zhai and Lafferty, 2001) to both the citing sentence MLE and the summary MLE using the KPLM of the paper as a background model using a Dirichlet Prior (DP) of 20. The choice of 20 has been based on the observation that citing sentences are short (32 words on average) and a large DP is prone to generate overly smoothed language models that are dominated by the KPLM, thus lack discriminative power. Here we choose to use this empirically selected DP parameter without attempting to fine-tune it for best results.

In summary, we implement a top sentence re-ranking heuristic that iteratively selects the next sentence to be appended to the existing summary whose smoothed language model is with the largest cross entropy (so it contains most extra information) with a smoothed language mode built for the summary at its current stage. We shall demonstrate how our top sentence re-ranking method introduces a major performance boost in the next section. For a quick inspection of the effectiveness of this method, compare the summaries constructed for paper P05-1012 with and without sentence re-reranking in Table 4. It shows that the summary constructed with sentence re-ranking covers key facts more comprehensively. The pseudo code for our re-ranking strategy is shown in Algorithm 1. It adopts a straightforward re-ranking approach that simply uses the top k+5 retrieved citing sentences in the previous step as the candidate pool; at each iteration, it selects the best sentence based on its cross entropy with the summary at the current stage. A more sophisticated re-ranking method is to combine the two cross entropy scores in some way (e.g., Maximal Marginal Relevance (Carbonell and Goldstein, 1998)) so that the final score for a citing sentence reflects its value in capturing salient keywords that have not yet been included in the summary. We leave the study of a more sophisticated re-ranking scheme for future work.

Algorithm 1 Top Sentence Re-ranking

```
1: function TopSentenceReranker
2:     k ← summary length limit
3:     top_sent ← top_k_plus_5_sents[0]
4:     es ← top_sent
5:     cp ← top_k_plus_5_sents - top_sent
6:     for s in cp do
7:         cp_lms[s] ← DPSmoothed(s)
8:     for i = 2 to k do
9:         es_lm ← DPSmoothed(es)
10:        s ← argmax_{s∈cp}(CE(cp_lms||es_lm))
11:        es ← es + s
12:        cp ← cp − s
13:        cp_lms ← cp_lms − cp_lms[s]
       return es
```

4 Experimental Setup

4.1 Evaluation Method

Following Qazvinian et al. (2008; 2010), we use the pyramid method (Nenkova and Passonneau, 2004) at sentence level to evaluate our system's performance. The pyramid score is a fact-based evaluation method that has been especially popular in evaluating extractive summarisation systems. It has been widely adopted because it incorporates both fact coverage and fact importance into the scoring process, which resonates well with the goals of summarisation (Qazvinian et al., 2010). More specifically, the pyramid method scores a summary using the ratio between the total facts weights of the facts it covers and that of an optimal summary. First a fact weights pyramid is built using some facts weighting method and each fact is subsequently put into its perspective pyramid tier. Qazvinian et al. (2008; 2010) built a weights pyramid for each paper and assigned each humanly discovered fact into a tier according to the number of citing sentences the fact occurs in that paper's citation summary. For example, fact f_i appearing in $|f_i|$ citing sentences in the citation summary of paper D is assigned to the tier $T_{|f_i|}$ in D's fact weights pyramid P_D. Let F_i denotes the number of facts in the summary ES in tier T_i of P_D. The total facts weights ES covers is calculated as:

$$W(ES)=\sum_{i=1}^{n} i \cdot F_i \quad (6)$$

where n is the highest tier of P_D. Let $ES_{optimal}$ be the optimal summary for D w.r.t. the summary length limit ($ES_{optimal}$ can be found using heuristic-driven exhaustive search). The pyramid score for ES is finally calculated as:

$$Score(ES)=W(ES)/W(ES_{optimal}) \quad (7)$$

Note again that we used exactly the same corpus and evaluation method as in (Qazvinian and Radev, 2008; Qazvinian et al., 2010), which makes our results directly comparable to those described in those papers. Furthermore, both papers report on performance of various baseline methods which are also directly comparable to ours (see next section). We compare our results with the current state-of-the-art; readers are encouraged to refer to (Qazvinian

Rank	Summary
KPLM without sentence re-ranking (Pyramid score: 0.23)	
1	3.1 decoding mcdonald et al (2005b) use the chu-liuedmonds (cle) algorithm to solve the maximum **spanning tree** problem.
2	thus far, the formulation follows mcdonald et al (2005b) and corresponds to the maximum **spanning tree** (mst) problem.
3	while we have presented signi cant improvements using additional constraints, one may won5even when caching feature extraction during training mcdonald et al (2005a) still takes approximately 10 minutes to train.
4	we have successfully replicated the state-of-the-art results for dependency parsing (mcdonald et al, 2005a) for both czech and english, using bayes point machines.
5	the search for the best parse can then be formalized as the search for the maximum **spanning tree** (mst) (mcdonald et al, 2005b).
KPLM with sentence re-ranking (Pyramid score: 0.73)	
1	3.1 decoding mcdonald et al (2005b) use the chu-liuedmonds (cle) algorithm to solve the maximum **spanning tree** problem.
2	to learn these structures we used online **large-margin** learning (mcdonald et al, 2005) that empirically provides state-of-the-art performance for czech.
3	while we have presented signi cant improvements using additional constraints, one may won5even when caching feature extraction during training mcdonald et al (2005a) still takes approximately 10 minutes to train.
4	mcdonald et al (2005a) introduce a dependency parsing framework which treats the task as searching for the **projective tree** that **maximises the sum of local dependency scores**.
5	we take as our starting point a re-implementation of mcdonald's state-of-the-art dependency parser (mcdonald et al, 2005a).

Table 4: Summaries of paper P05-1012 produced using KPLM. Key facts in citing sentences are highlighted and OCR and sentence segmentation errors have been retained as they originally appeared in the corpus.

and Radev, 2008; Qazvinian et al., 2010) for cross-referencing results from a broader set of systems.

4.2 Results and Discussion

Table 5 shows the pyramid score evaluation results for the 25 papers. To facilitate comparison and cross-referencing, the table has been formatted as close as possible to Table 7 in (Qazvinian and Radev, 2008) with figures in the Gold and C-LexRank columns directly copied over. Note that a Gold pyramid score less than 1 suggests that there are more facts than can be covered using k sentences for that paper's citation summary. It can be seen that KPLM based summarisation achieves quite comparable results (especially in terms of the median score) with C-LexRank, even without top sentence re-ranking. When the re-ranking is introduced, our system outperforms the current state-of-the-art C-LexRank by a measurable margin. Albeit the perceived differences in the results, a one-tailed Wilcoxon signed-rank test indicated that our results are not statistically superior at significance level 0.05 (Z=-1.22, P=0.11). A power analysis reveals that in order to achieve a statistically significant result on this small sample of 25 papers, a system would need to score a medium to large effect size (Cohen's d > 0.53), which is a challenging task considering C-LexRank's strong baseline performance. We hope this analysis can inform future studies using Qazvinian's 25 papers corpus. Nevertheless, it should be pointed out that our approach is not only substantially simpler than C-LexRank, it also yields more interpretable results.

We know of a more recent set of results reported in (Qazvinian et al., 2013), which again confirmed C-LexRank's state-of-the-art status with a mean pyramid score of 0.799 (cf. Table 6 in (Qazvinian et al., 2013)). However those results are not comparable with ours for the following reasons. First, Qazvinian et al. (2013) used a slightly different corpus with 30 papers (5 extra papers from the Conditional Random Field domain). Second, results were based on a summary length limit of 200 words, so roughly equivalent to 6.3 sentences per paper, giving evaluations an extra edge. Both changes boosted system performance in those evaluations, as evidenced by comparing Table 7 in (Qazvinian and Radev, 2008) and Table 6 in (Qazvinian et al., 2013).

Qazvinian et al. (2010) used the same corpus and evaluation method as our work; however the results have been presented as box plots (cf. Figure 1 in (Qazvinian et al., 2010)) from which only the five-number summary (i.e., minimum, lower quartile, median, upper quartile and maximum) of the pyramid scores can be reconstructed and consequently no significance test can be performed. Compared with the best performing variants of the system devised in (Qazvinian et al., 2010) based on unigrams, bigrams and trigrams, our system (KPLM+TSR) achieves a higher median score (0.86 vs. 0.80), as well as a lower score variation across the 25 papers.

An arbitrarily imposed constraint in the evaluations is the summary length limit, which may be changed to suit a specific application context. The summarisation task becomes increasingly more challenging when summary length limit is further tightened as this would require a summariser to pinpoint the best sentences from a potentially large cita-

Domain	Paper	Gold	C-LexRank	KPLM	KPLM+TSR
DP	C96-1058	1.00	0.73	0.33	0.56
	P97-1003	1.00	0.40	0.79	0.79
	P99-1065	0.94	0.67	0.62	0.76
	P05-1013	1.00	0.67	0.66	0.66
	P05-1012	0.95	0.62	0.23	0.73
PBMT	N03-1017	0.96	0.64	0.60	0.60
	W03-0301	1.00	1.00	0.80	0.80
	J04-4002	1.00	0.48	0.86	0.89
	N04-1033	1.00	0.85	0.57	0.86
	P05-1033	1.00	0.85	0.97	0.97
SUMM	A00-1043	1.00	0.95	0.50	0.50
	A00-2024	1.00	0.60	0.60	0.60
	C00-1072	1.00	0.93	0.87	0.93
	W00-0403	1.00	0.70	0.81	0.54
	W03-0510	1.00	0.83	1.00	1.00
QA	A00-1023	1.00	0.86	0.88	1.00
	W00-0603	1.00	0.60	0.44	0.94
	P02-1006	1.00	0.87	0.93	0.93
	D03-1017	1.00	0.85	0.70	0.90
	P03-1001	1.00	0.59	0.94	0.44
TE	D04-9907	1.00	0.94	0.77	0.91
	H05-1047	1.00	1.00	0.83	0.83
	H05-1079	1.00	0.56	0.78	0.89
	W05-1203	1.00	0.71	1.00	1.00
	P05-1014	1.00	0.78	0.89	1.00
	Mean	**0.99**	**0.75**	**0.73**	**0.80**
	Median	**1.00**	**0.73**	**0.79**	**0.86**

Table 5: Summary pyramid score evaluation results with summary length limit $k = 5$.

tion summary. A desirable property of a good summariser is thus the ability in maintaining its performance while the task becomes increasingly demanding. To further evaluate KPLM's performance under increasingly more stringent summary length limits, we gathered the pyramid scores with summary length limit k decreasing from 5 to 1 and visualised the results in Figure 1. We can see that KPLM's performance decays quite gracefully as more stringent limits are imposed. Even under the harshest constraint with the summary length limit sets to 1, our system still managed a mean pyramid score of close to 0.6 across the 25 papers. Indeed, it can be seen that the variance in pyramid scores gradually spreads wider (the dark band in the figure marks out 95% confidence interval of the mean scores), but this phenomenon is expected as the error margin also shrinks along with the summary length limit.

5 Conclusion and Future Work

We designed a statistical framework to summarise scientific papers, using methods rooted in quantitative statistics and information theory. We first built a keyword profile for a paper using a quantitative statistical method that captures its characterising keywords that are both overly represented and relatively exclusively used in the paper's citation summary. We then used the keyword profile of a paper to build a discriminative pseudo unigram language model that directly incorporates words' salience in characterising a paper's main contributions into pseudo generative probabilities. Based on the fact that a paper's KPLM represents an effective language model from which pseudo citing sentences with good coverage of important keywords could be sampled, we cast the task of summarisation as language model divergence based IR. Finally, we implemented an information-driven sentence reranking algorithm that can effectively leverage diversity in keyword coverage in summaries produced. Experimental results show that our approach outperforms the current state-of-the-art systems in scientific paper summarisation, which is also with good resilience to more stringent summary length limits.

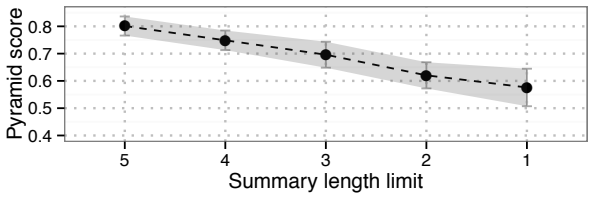

Figure 1: Pyramid scores of KPLM+TSR under different summary length limits.

In the future, we plan to extend our approach to higher order n-grams and see whether larger information units (phrases) would help boost summarisation performance. We also plan to apply our method to the problem of multi-document summarisation. In particular, we are very interested to test our system's performance on automatically generating a technical survey of a scientific paradigm, which thanks to the authors of (Mohammad et al., 2009; Qazvinian et al., 2013), has been established as a well-defined task with high-quality open data. Finally, while we have shown that our approach is effective in summarising a scientific paper's major contributions using its citation summary text, further experiments are required to test our method's effectiveness on more generic summarisation tasks and texts genres.

Acknowledgement

We thank Vahed Qazvinian for making the 25 paper summarisation corpus publicly available; without it, our formal evaluations would have been impossible. We also thank the anonymous reviewers for their highly constructive comments.

References

Brin, Sergey and Page, Larry. 1998 The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web*, pp. 107–117.

Carbonell, Jaime and Goldstein, Jade. 1998 The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the SIGIR*, pp. 335–336.

Elkiss, Aaron and Shen, Siwei and Fader, Anthony and Erkan, Güneş and States, David and Radev, Dragomir. 2008 Blind men and elephants: What do citation summaries tell us about a research article?. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, pp. 51–62.

Kullback, S. and Leibler, R. A. 1951 On information and sufficiency. *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86.

Mei, Qiaozhu and Zhai, Chengxiang. 2008 Generating impact-based summaries for scientific literature. In *Proceedings of the ACL*, pp. 816–824.

Mohammad, Saif and Dorr, Bonnie and Egan, Melissa and Hassan, Ahmed and Muthukrishan, Pradeep and Qazvinian, Vahed and Radev, Dragomir and Zajic, David. 2009 Using citations to generate surveys of scientific paradigms. In *Proceedings of the HLT-NAACL*, pp. 584–592.

Nenkova, Ani and Passonneau, Rebecca. 2004 Evaluating content selection in summarization: The pyramid method. In *Proceedings of the HLT-NAACL*, pp. 145–152.

Qazvinian, Vahed and Radev, Dragomir. 2008 Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 689–696.

Qazvinian, Vahed and Radev, Dragomir and Özgür, Arzucan. 2010 Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 895–903.

Dunning Ted. 1993 Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, vol. 19, no. 1, pp. 61–74.

Lin, Chin-Yew and Hovy, Eduard. 2000 The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational Linguistics*, pp. 495–501.

Moore. Robert C. 2004 On log-likelihood-ratios and the significance of rare events. In *Proceedings of EMNLP*, pp. 333–340.

Qazvinian, Vahed and Radev, Dragomir and Mohammad, Saif and Dorr, Bonnie and Zajic, David and Whidby, Michael and Moon, Taesun. 2013 Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp.165–201.

Radev, Dragomir and Pradeep, Muthukrishnan and Qazvinian, Vahed. 2009 The ACL anthology network corpus. In *Proceedings of the ACL workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, pp. 54–61.

Siddharthan, Advaith. and Teufel, Simone. 2007 Whose idea was this, and why does it matter? In *Proceedings of the NAACL/HLT*, pp. 316–323.

Tomokiyo, Takashi and Hurst, Matthew. 2003 A language model approach to keyphrase extraction. In *Proceedings of the ACL'03 workshop on Multiword expressions*, pp. 33–40.

Zhai, Chengxiang. 2008 Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval*, vol. 2, no. 3, pp. 137–213.

Zhai, Chengxiang and Lafferty, John. 2001 A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334–342.