Semantic Title Evaluation and Recommendation Based on Topic Models

Huidong Jin[1,2], Lijiu Zhang[2], and Lan Du[3]

[1] CSIRO Mathematics, Informatics and Statistics, Acton ACT 2601, Australia
[2] Research School of Computer Science, CECS, the Australian National University, Acton ACT 2601, Australia
[3] Department of Computing, Macquarie University, NSW 2109, Australia
Warren.Jin@csiro.au, Lijiu.Zhang@anu.edu.au, Lan.Du@mq.edu.au

Abstract. To digest tremendous documents efficiently, people often resort to their titles, which normally provide a concise and semantic representation of main text. Some titles however are misleading due to lexical ambiguity or eye-catching intention. The requirement of reference summaries hampers using traditional lexical summarisation evaluation techniques for title evaluation. In this paper we develop semantic title evaluation techniques by comparing a title with other sentences in terms of topic-based similarity with regard to the whole document. We further give a statistical hypothesis test to check whether a title is favourable without any reference summary. As a byproduct, the top similar sentence can be recommended as a candidate for title. Experiments on patents, scientific papers and DUC'04 benchmarks show our Semantic Title Evaluation and Recommendation technique based on a recent Segmented Topic Model (STERSTM), performs substantially better than that based on the canonical model Latent Dirichlet Allocation (STERLDA). It can also recommend titles with quality comparable with the winners of DUC'04 in terms of summarising documents into very short summaries.

Keywords: Topic models, semantic, evaluation, hypothesis test.

1 Introduction

Text mining techniques have been sought after in order to make informed decisions efficiently based on tremendous textural information [1]. For a lot of documents, a good title, which gives a concise and semantic representation of contents in main text, often provides a shortcut for readers to digest documents. However, due to various reasons like lexical ambiguity (say, polysemy), eye-catching intention or being prepared by an inexperienced writer, a lot of documents come with titles whose semantics are away from their main texts. For example, "Learning to fly" can be a title of a book for flight training or an autobiography for Victoria Beckham. These motivate us to develop automatic techniques to evaluate to what degree a title captures the main contents of its associated document. As a byproduct, our title evaluation techniques can be used to recommend a title-worthy sentence from which a quality title could be generated.

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 402–413, 2013.

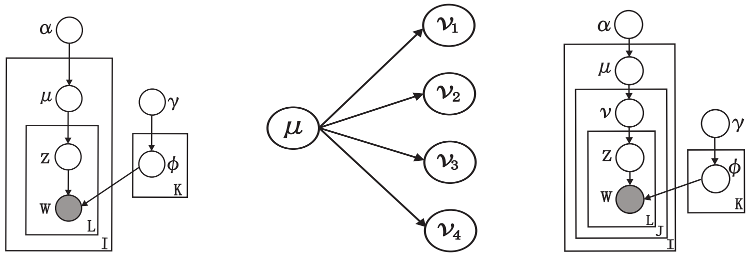
Two issues hamper adjusting traditional document summarisation evaluation techniques including ROUGE for title evaluation. To evaluate a title, these techniques require a, normally human generated, reference summary [12]. In addition, the evaluation is mainly based on whether words in a title appear or not in the reference summary. It is often not enough, especially considering polysemy and synonymy. To overcome these two issues, we will propose to compute semantic similarity in a space spanned by latent topics learnt by topic models, and then to use a statistical hypothesis test to check or classify whether a title is poor.

Our title evaluation and recommendation techniques are relevant to extractive text summarisation. Extractive summarisation only chooses information (words/sentences) from documents to compose concise representation for them [13]. There are mainly two types of approaches [13]. One type of approaches first derive an intermediate representation like topic words, TF*IDF, Latent Semantic Analysis (LSA), and Bayesian topic models, for documents that captures the contents in main text. Sentences are then scored for importance. Because of their modelling generalisability to unseen documents and short textual units [2, 8, 9], we choose topic models to learn a topic representation of a document and its sentences/title. In this way, not only can one handle polysemy and synonymy [5], but also make a title, sentences, a document directly comparable. We will develop and compare two semantic title evaluation techniques based on either a recent Segmented Topic Model (STM) [9] or the standard Latent Dirichlet Allocation (LDA) [2].

In the second type of summarisation approaches, indicator representation approaches, the text is represented by a diverse set of possible importance indicators that do not aim at discovering topicality [13]. These indicators are combined, using graph-based ranking methods, say PageRank [10] or machine learning techniques, say classification [6], to score the importance of each sentence. Different from Bayesian topic models, these approaches normally require extra information [10, 13, 14, 15], such as costly training data [6, 14], WordNet [6], Wikipedia [14, 15], or search query logs [14]. Our experiments show our title recommendation techniques can give very short summaries with quality comparable with the winners of DUC'04, including [6, 10].

The basic procedure of our Semantic Title Evaluation and Recommendation (STER) techniques is as follows. (1) We use topic models to generate latent topics, each of which is a probability distribution over words, from documents as well as sentences/titles in them. Based on two topic models STM and LDA, we have STERSTM and STERLDA techniques respectively. (2) Each document/sentence/title is represented as a mixture of latent topics. (3) Semantic similarity values are calculated between documents and its sentences/title based on their topic distributions. (4) Being compared with other sentences in a document, a title with a statistically significantly low similarity is regarded as unfavourable. The top similar sentence is recommended as a title worthy sentence.

In Section 2, we first brief Bayesian topic modelling techniques. Section 3 presents STERSTM and STERLDA. Experimental results of STERSTM, and



(a) LDA (b) Topic hierarchy (c) STM

Fig. 1. LDA [2], hierarchical structure within a document, and STM [9]

comparison with STERLDA and the methods participated in DUC'04 are reported in Section 4, followed by concluding comments in Section 5.

2 Background of Bayesian Topic Modelling Techniques

In order to estimate semantic coverage of a title, we need to compute semantic similarity between a title and its whole document in the same latent topic space. Because of better generalisation capability to unseen documents and short textual units than LSA and its variants [9], we learn this topic space using Bayesian topic models that specify a probabilistic process by which text documents can be generated.

The canonical topic model, LDA [2], is a latent variable model of documents, where a document is regarded as a mixture of K latent topics, each of which is a probability distribution over words. Following [9], documents are indexed by i $(i = 1, \cdots, I)$, and words $\boldsymbol{w}$ are observed data, each is indexed by l $((l = 1, \cdots, L))$. The latent variables are $\boldsymbol{\mu}_i$ (*the topic distribution or topic proportion* for a document) and $\boldsymbol{z}$ (the *topic assignments* for observed words), and the model parameter of $\boldsymbol{\phi}_k$'s (*per-topic word distributions*). This generative model, as illustrated in Fig. 1(a), is as follows:

$$\begin{array}{llll} \boldsymbol{\phi}_k \sim \text{Dirichlet}_W(\boldsymbol{\gamma}) & \forall\, k; & \boldsymbol{\mu}_i \sim \text{Dirichlet}_K(\boldsymbol{\alpha}) & \forall\, i; \\ z_{i,l} \sim \text{Multinomial}_K(\boldsymbol{\mu}_i) & \forall\, i,l; & w_{i,l} \sim \text{Multinomial}_W\left(\boldsymbol{\phi}_{z_{i,l}}\right) & \forall\, i,l. \end{array}$$

$\text{Dirichlet}_K(\cdot)$ is a K-dimensional Dirichlet distribution, and W is the number of different words. The hyper-parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are Dirichlet priors for word and topic distributions respectively.

Since LDA was introduced, topic models have been widely extended in the text mining community (see [5, 8] and references therein). Topic models have been successfully used in document summarisation [13], opinion mining [16], sequential topic evolution [8, 7], etc. Via leveraging hierarchical structure within a document, such as a document consisting of sentences (Fig. 1(b)), STM can generate much more accurate topics than LDA and its variants [9]. In addition, it models a document and its sentences in the same topic space, which is required by our semantic title evaluation. In fact, in STM, topic proportions of sentences distribute around the topic proportion of the whole document, as it is described by a Poisson-Dirichlet Process (PDP). Conditioned on the model parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Phi}$ and PDP parameters a, b (called *discount* and *strength* respectively, $0 \leq a < 1$, $b > -a$), STM that we used in this paper assumes the following generative process (graphical view see Fig. 1(c)):

1. For each document documents D_i ($i \in \{1, \cdots, I\}$), draw a document topic proportion or distribution $\boldsymbol{\mu}_i \sim \text{Dirichlet}_K(\boldsymbol{\alpha})$
2. For each sentence $S_{i,j}$ ($j \in \{1, \cdots, J_i\}$)
   (a) Draw sentence topic proportion $\boldsymbol{\nu}_{i,j}$ around $\boldsymbol{\mu}_i$, i.e., $\boldsymbol{\nu}_{i,j} \sim \text{PDP}(a, b, \boldsymbol{\mu}_i)$
   (b) For each word $w_{i,j,l}$, where $l \in \{1, \ldots, L_{i,j}\}$
      i. Select a topic $z_{i,j,l} \sim \text{Multinomial}_K(\boldsymbol{\nu}_{i,j})$
      ii. Generate a word $w_{i,j,l} \sim \text{Multinomial}_W(\boldsymbol{\phi}_{z_{i,j,l}})$

3 Semantic Title Evaluation and Recommendation

The procedure of our semantic title evaluation methods is given as follows. It first represents a document, its sentences and title using the same set of latent topics learned by a topic model. The semantic similarity between a title/sentence and the document is computed based on their topic proportion (i.e., distributions) vectors. Via comparing the title's similarity value with those of sentences in main text, we use a hypothesis test to compute p-Value to check how semantically good a title is. As a byproduct, p-Values for those sentences can also be used to recommend a top one for a title candidate, from which a title can be generated quickly.

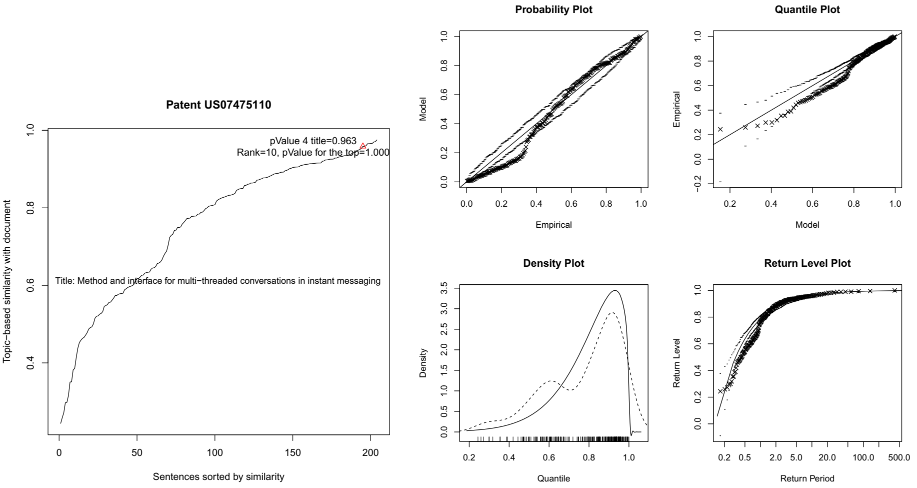
Algorithm 1 outlines our Semantic Title Evaluation and Recommendation method based on STM (STERSTM). In the preprocessing step (Step 1), a document is first split into its constituent sentences by a Perl programme (Lingua:en:sentence package) [3] based on a regular expression and a list of abbreviations. Hereinafter, a title is treated as a separate sentence for the sake of

Algorithm 1 STERSTM

Input: One corpus $\mathcal{D}$ with one or multiple documents, and the number of topics K.

```
1. Document preprocessing: Split documents $D_i$ ($\in \mathcal{D}$) into sentences $S_{i,j}$, and
   then split sentences $S_{i,j}$ into words $w_{i,j,l}$; remove most and least frequent words
2. Build a STM, and estimate its parameters using the collapsed Gibbs sampler in [9]
3. Infer topic proportions $\boldsymbol{\mu}_i$ for documents and $\boldsymbol{\nu}_{i,j}$ for sentences based on STM
4. FOR each document $D_i$ in $\mathcal{D}$ DO
5.     Compute similarity $s_{i,j}$ between $D_i$ and sentence $S_{i,j}$ using $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_{i,j}$
6.     Fit a GEV distribution $G(s; \boldsymbol{\theta}_i)$ over $s_{i,j}$ via maximising likelihood
7.     Compute p-Value $G(s_{i,j}; \boldsymbol{\theta}_i)$ for each sentence $S_{i,j}$        /*Hypothesis test*/
8.     Categorise and rank sentences based on their p-Values
```

Output: p-Value for titles, and the sentence with largest p-Value for each document



(a) Similarity, rank and p-Value

(b) Diagnosis plots for a GEV distribution

Fig. 2. Semantic title evaluation result of STERSTM for Patent US07475110

simplicity. Sentences are then split into words. After that, all stop-words, extremely frequent (*e.g.*, top 30 in our experiments) words, and least frequent (*e.g.*, less than 5 times) words are removed. We do not stem words in order to keep post-processed sentences with an acceptable length.

After having the word list $w_{i,j,l}$ for each sentence $S_{i,j}$ in document D_i, we run the efficient collapsed Gibbs sampling algorithm [9] to estimate parameters in STM (Step 2). In Step 3, with a sufficient number of samples being drawn from the converged Markov chain for STM, topic distributions of documents and sentences can be estimated by a fixed point estimation with inverting the generative process in Section 2.

Step 5 calculates the semantic similarity between a document and its sentences using their topic proportion vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_{i,j}$. The widely used cosine similarity measures similarity between two vectors by calculating the cosine of the angle between them:

$$s_{i,j} = cosine\_similarity\,(\boldsymbol{\mu}_i, \boldsymbol{\nu}_{i,j}) = \frac{\sum_{k=1}^{K}(\mu_{i,k} \times \nu_{i,j,k})}{\sqrt{\sum_{k=1}^{K} \mu_{i,k}^2} \times \sqrt{\sum_{k=1}^{K} \nu_{i,j,k}^2}} \qquad (1)$$

Because a topic proportion vector also indicates a multinomial distribution, we can also use the Hellinger distance or Kullback-Leibler divergence, which quantify the similarity between two probability distributions [13, 7]. As our preliminary title evaluation experimental results show there is little difference among these similarity metrics, we will only present results for the cosine similarity. Examples of cosine similarities of sentences within two patents and one conference paper could be found in Figs. 2(a) and 3.

Before introducing Steps 6-8, we show that it is not easy to specify a constant threshold for semantic similarities for determining a favourable title through examples in Fig. 3. Similarities for different documents have different value ranges. Comparing with other sentences in a document, 0.95 is reasonably good for the patent's title while just average for the paper's title in Fig. 3(a). Similarly, because the numbers of sentences in a document can range from a few dozens to several thousands, it is difficult to specify a threshold for rank based on similarity or relative rank (e.g., rank of title divided by the total number of sentences within a document). Rank 34th is possibly favourable for a title within a very long document, but doubtable for a short one in Fig. 3(a). A lot of sentences arguably have high semantic similarity with a document. A small change on the title's similarity value may lead to a big change on its rank as well as its relative rank.

We give a statistical mechanism to specify document-specific 'thresholds', as Generalised Extreme Value (GEV) distribution is able to fit well these similarity values in Figs. 2(a) and 3. In the extreme value theorem, the GEV distribution is a limited distribution of properly normalized minima of a sequence of independent and identically distributed random variables [4]. It is a family of continuous probability distributions, and it is a general distribution family, including Weibull and Gumbel distribution families. The GEV distribution we used has a cumulative distribution function:

$$G(x;\boldsymbol{\theta}) = exp\left\{-\left[1-\theta_3\left(\frac{x+\theta_1}{\theta_2}\right)\right]^{-1/\theta_3}\right\} \tag{2}$$

for $1-\theta_3(x+\theta_1)/\theta_2 > 0$, where $\theta_1 \in R$ is the location parameter, $\theta_2 > 0$ the scale parameter and $\theta_3 \in R$ the shape parameter.

In Step 6, parameter $\boldsymbol{\theta}$ of the GEV distribution are estimated via maximising likelihood of all the similarity values within the same document. The parameter estimation can be visually validated via such as probability plot, quantile (Q-Q) plot, density plot or return level plot [4]. Diagnosis plots for a fitted GEV distribution for similarity values in Fig. 2(a) are exemplified in Fig. 2(b).

Step 7 in Algorithm 1 computes p-Values of all the sentences within a document. The p-Values for a sentence/title here can be used for fulfilling a statistical hypothesis test. The p-Value is the probability of the similarity observation under the null hypothesis (H0) which hypothesises that its similarity value based on topics is not extreme in comparison with counterpart sentences. We can reject the null hypothesis if and only if the p-Value is less than the significance level threshold. We will use a conservative threshold, say, 10% in this work. Therefore, if the p-Value for a title is less than the threshold, we reject the null hypothesis and draw a statistically sound conclusion that the title is not semantically good enough (in comparison with other sentences in the associated document). In other words, we can categorise such a title as 'Unfavourable'. Our experiment results, some presented in Section 4.2, show that the sentences with large semantic similarity values can summarise the whole document excellently. As a matter of factor, STERSTM is very close to the runner-up of Task 1 (summarising an English document into a very short summary) in the Document Understanding

Conference (DUC) in 2004 [6]. Thus, we may categorise a title as 'Excellent' if its p-Value is larger than 90%. Other titles, with moderate p-Value ranging from 0.10 to 0.90, will be categorised as 'Average.' Step 8 conducts this categorisation. It also sorts sentences of a document based on their p-Values (equivalently, their semantic similarities). Finally, the p-Values of titles generated by STERSTM can evaluate titles in a statistically sound way without a reference summary. The top sentences with highest p-Value from STERSTM can be recommended as the title-worthy sentence or a title candidate for the document.

Steps of Algorithm 1 could be independently replaced with other techniques to develop new methods. For example, Step 5 can be replaced with other sentence scoring techniques [10, 13]. Steps 2 and 3 can be replaced with a modelling technique as soon as it can represent documents and sentences in the same semantic space. When steps 2 and 3 are replaced with LDA, we call the new method STERLDA. LDA does not consider document structure as STM does. In order to derive topic distributions for both documents and their sentences, we need to run LDA twice, one on the document level, another on the sentence level. However, these two LDAs will come up with two different sets of latent topics due to unsupervised learning. To tackle this problem, the topics generated on the document level are used and fixed in training LDA on the sentence level.

4 Experimental Setting and Results

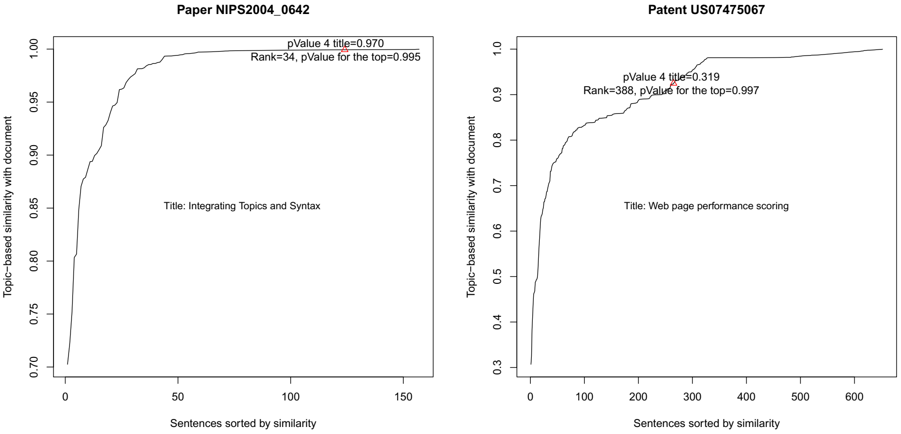
STERSTM can run on a single document, while STERLDA cannot. To facilitate a fair comparison, we ran both of them on a set of documents. We set the number of topics $K = 50$, and priors $\alpha = 0.05$ and $\gamma = 0.01$ for both STM and LDA, and $a = 0.02$ and $b = 10$ for STM in our experiments in this paper.

4.1 Semantic Title Evaluation Experiments

We used two sets of documents for title evaluation experiments. One is Patents-99, where 99 U.S. patents were randomly selected from 5000 U.S. patents [1] granted between Jan. and Mar. 2009 under the class "computing; calculating; counting" with international patent classification (IPC) code G06. After preprocessing, the numbers of post-processed sentences in these patents range from 60 to 2163. The second data set is NIPS-100, in which 100 papers were randomly selected from NIPS conference papers in 2004. These papers contain a lot of equations, which make the preprocessing step harder. The numbers of sentences range from 68 to 207.

As we discussed in Section 3, p-Value from a GEV distribution can give us more informative evaluation than ranks etc. When the similarity value of a title has a high rank, it often has a low p-Value. Though the rank and the p-Value are negatively correlated, p-Value takes into account of similarity values of other sentences within the same document, and becomes more informative. For example,

[1] All patents are from Cambia, `http://www.cambia.org/daisy/cambia/home.html`



(a) For Paper 642 in NIPS'04

(b) For Patent US07475067

Fig. 3. Semantic similarity, rank, and p-Value got by STERSTM for two documents

for NIPS paper 579, its title "*Validity estimates for loopy Belief Propagation on binary real-world networks*" has the semantic similarity of 0.921, and is ranked only 116th in comparison with the 131 sentences from the paper, and looks really unfavourable. However, p-Value of 0.417 does not provide evidence statistically significantly to claim this title is unfavourable. Another similar example could be found in Fig. 3(b).

Fig. 3 illustrates semantic similarities between sentences/title and a whole document based on topics learned by STM. For Paper 642 from NIPS'04, the title "*Integrating Topics and Syntax*" has the similarity value of 0.9993, and it is ranked 34th in comparison with 155 sentences from the paper. Its p-Value from the GEV distribution is 0.970. That means this is an excellent title. From Fig. 3(b), we can see the title "*Web page performance scoring*" has the semantic similarity of 0.9247. It is ranked as 388th in comparison with other 650 sentences from the patent. Its p-Value is 0.319, which says the title is not excellent from the viewpoint of covering the whole patent semantically. From its abstract[2], we can see it could be improved if some word related with '*tool*' or '*browser-based tool*' is appended to the title. As another evidence, the top semantically similar sentence chosen by STERSTM is "*More particularly, the invention relates to*

[2] ***Abstract*** *A browser-based tool is provided that loads a Webpage, accesses the document object model (DOM) of the page, collects information about the page structure and parses the page, determines through the use of heuristics such factors as how much text is found on the page and the like, produces statistical breakdown of the page, and calculates a score based on performance of the page. Key to the operation of the invention is the ability to observe operation of the Webpage as it actually loads in real time, scoring the page for several of various performance factors, and producing a combined score for the various factors.*

Table 1. Categorisation of titles for two sets of documents based on p-Values

Title Categorisation		Unfavourable	Average	Excellent
p-Value range		[0,0.1]	(0.1,0.9]	(0.9,1.0]
Patents-99	STERSTM	0	49	50
	STERLDA	6	57	36
NIPS-100	STERSTM	0	55	45
	STERLDA	11	66	23

a tool which analyses the content and structure of Web pages in real time and produces statistics and a performance score."

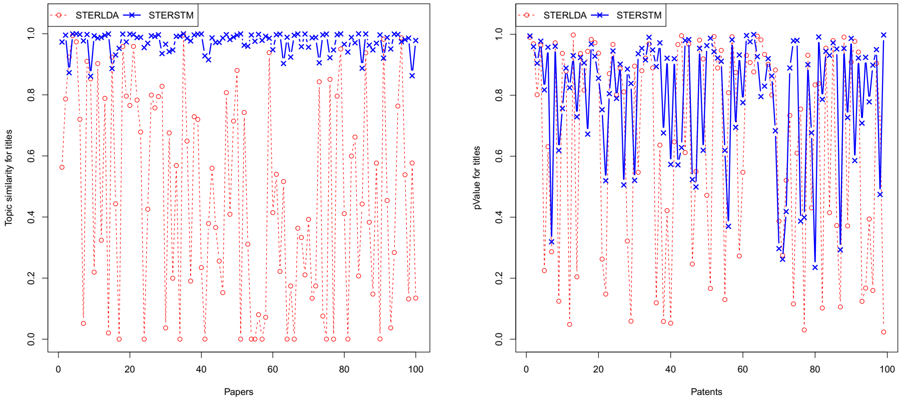
Fig. 4(a) illustrates the semantic similarity values of titles from NIPS-100. The 100 similarity values of these titles generated by STERSTM range from 0.86 to almost 1. They are normally quite high. The similarities by STERLDA range from almost 0 to 0.996 and have a broader value range. It seems that STERLDA generates less reliable evaluation than STERSTM in terms of similarity values. For this document set, according to STERSTM, 45 out of 100 papers have excellent titles, including the one in Fig. 3(a). STERSTM doesn't find any unfavourable title, which is not surprised as all the papers were prepared by experienced researchers. STERLDA surprisingly finds 11 unfavourable titles, and only 23 excellent ones as summarised in Table 1. For example, the title "*Methods Towards Invasive Human Brain Computer Interfaces*" of Paper 443 in NIPS'04 has the p-Value of 0.058 and is inappropriately regarded as unfavourable by STERLDA.

Fig. 4(b) gives the p-Values of these titles within the document set Patents-99. The p-Values of the 99 titles based on STERSTM range from 0.22 to very close to 1. STERSTM finds 45 excellent titles, and it does not find any unfavourable patent titles, as we would expect. In comparison, p-Values from STERLDA range from 0 to close to 1. It finds only 23 excellent titles and 11 unfavourable titles. Thus, STERSTM can evaluate titles more reliably than STERLDA based on the two document sets.

4.2 Semantic Title Recommendation Experimental Results

In this section, we empirically check whether our proposed techniques can recommend a title worthy sentence from the viewpoint of capturing the main idea of a document [11]. Due to limitation of space, we only report results on one set of documents, DUC-2004. DUC-2004 is the benchmarks used for Task 1 (generating a very short summary from a document) in NIST's DUC'04[3]. The corpus consists of 50 sets of documents each contains 10 same topic documents on average. The documents came from the AP newspapers and New York Times newspapers. The short summary generated is peer summary and it is automatically evacuated by one of widely used document summarisation metrics, Recall-Oriented

[3] `http://www-nlpir.nist.gov/projects/duc/guidelines/2004.html`



(a) Similarity values for NIPS-100

(b) p-Values for Patents-99

Fig. 4. Topic similarities values or p-Values from STERSTM and STERLDA

Understudy for Gisting Evaluation (ROUGE) [12]. ROUGE essentially calculates n-gram overlaps between given summaries and previously-written human summaries. A high level of overlap should indicate a high level of shared concepts between the two summaries. There are four reference summary (or model summary) per document in DUC-2004. ROUGE can evaluate a short given summary by comparing it with up to four reference summaries.

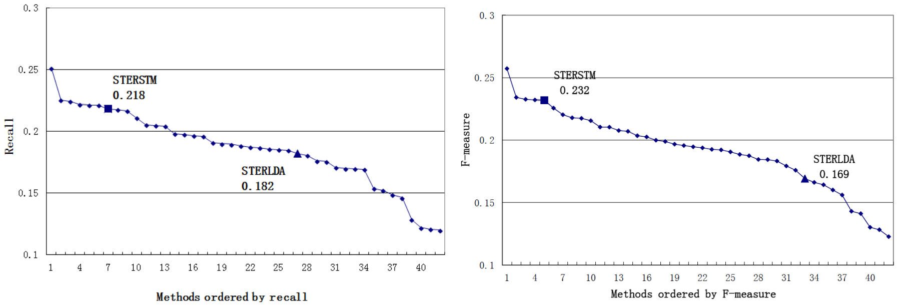
We report evaluation results based on ROUGE-1, i.e., checking unigram overlap between a given summary and a reference summary, partially because both STM and LDA are trained with unigrams. In particular, we use F-measure, which is a weighted harmonic mean of recall and precision.

$$\text{F-measure} = \frac{2 \cdot precision \cdot recall}{precision + recall}, \tag{3}$$

where the recall is the proportion of words in the reference summary appearing in the given sentence, and precision is the proportion of words in the given sentence appearing in the reference summary. Both precision and recall are based on an understanding and measure of relevance. An F-measure score reaches its best value at 1 and worst score at 0.

As we mentioned in Section 3, to facilitate fair comparison, a sentence was trimmed (removing duplicate words, frequent words, and semantically less important words which are not in top 100 word lists of in topic-word distributions) as ROUGE truncates summaries longer than the target length of 75 bytes (alphanumerics, whitespace, and punctuation included) before evaluation for DUC-2004.

For this corpus, the average recall, precision and F-measure of STERSTM are 0.218, 0.250, and 0.232, respectively. For STERLDA, they are 0.182, 0.160, and 0.169, respectively. STERSTM obviously outperforms STERLDA. In comparison with 40 participation methods in the DUC'04 conference, STERSTM did quite well in terms of all the three measures. It is ranked as 7th, 9th and



(a) Average recall

(b) Average F-measure

Fig. 5. Title recommendation results of 42 methods for DUC-2004

5th in terms of average recall (Fig. 5(a)), precision, and F-measure (Fig. 5(b)). One DUC'04 participation method [6] that requires training data and WordNet, has F-measure of 0.234, which is the runner-up in Fig. 5(b). Its average recall is 0.217, quite close to that of STERSTM. The another graph-based document summarisation technique, the winner of several tasks in DUC'04, LexRank [10] has F-measure of 0.208 for this task and is 13th in Fig. 5(b). Thus, in terms of quality of very short summaries generated for DUC-2004, STERSTM is comparable with the top methods participated in the DUC'04 conference.

5 Conclusion and Discussion

Based on a recent topic modelling technique, Segmented Topic Model (STM), this work has presented one Semantic Title Evaluation and Recommendation (STER) technique, STERSTM. Through comparing title/sentences with the whole document in the topic space created by STM, STERSTM computes the semantic similarity of title/sentences, which can estimate the semantic coverage of a title/sentence. Via fitting a Generalised Extreme Value (GEV) distribution over the similarity values of sentences and a title within a document and calculating p-Value under the distribution, STERSTM is able to identify excellent and unfavourable titles without extra information like a human generated reference summary. The sentence with top p-Value is recommended as a title candidate. Experimental results on several different document sets have shown STERSTM can pick up some improvable titles, statistically significantly outperform STERLDA, a counterpart based on the canonical topic model LDA, and generate very short summaries with quality comparable with various document summarisation techniques.

There are several possible extensions of this work. Better trimming techniques to shorten a sentence to a concise and readable title could improve title recommendation [11]. It is appealing to explore more reliable statistical distributions for semantic similarity values, especially for those for small documents. We are also going to extend the proposed techniques for multiple relevant documents, embedding key words or other meta data.

Acknowledgements. Huidong Jin was financially supported by the Environmental and Agricultural Informatics program, CSIRO Mathematics, Informatics and Statistics. Lan Du was supported under the Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593). The authors are grateful to the anonymous reviewers and Dr Peter Caley for their constructive comments.

References

[1] Aggarwal, C., Zhai, C.: Mining Text Data. Springer-Verlag New York Inc. (2012)

[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)

[3] Clough, P.: A perl program for sentence splitting using rules. University of Sheffield (2001)

[4] Coles, S.: An introduction to statistical modeling of extreme values. Springer (2001)

[5] Crain, S., Zhou, K., Yang, S., Zha, H.: Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond. In: [1], ch. 5, pp. 129–161 (2012)

[6] Doran, W., Stokes, N., Newman, E., Dunnion, J., Carthy, J., Toolan, F.: News story gisting at university college dublin. In: The Proceedings of the Document Understanding Conference, DUC (2004)

[7] Du, L., Buntine, W., Jin, H.: Modelling sequential text with an adaptive topic model. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 535–545. Association for Computational Linguistics (2012)

[8] Du, L., Buntine, W., Jin, H., Chen, C.: Sequential latent Dirichlet allocation. Knowledge and Information Systems 31(3), 475–503 (2012)

[9] Du, L., Buntine, W., Jin, H.: A segmented topic model based on the two-parameter Poisson-Dirichlet process. Machine Learning 81, 5–19 (2010)

[10] Erkan, G., Radev, D.: LexRank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. (JAIR) 22, 457–479 (2004)

[11] Jin, R., Hauptmann, A.G.: A new probabilistic model for title generation. In: COLING 2002, pp. 1–7 (2002)

[12] Lin, C., Och, F.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: ACL 2004, p. 605. Association for Computational Linguistics (2004)

[13] Nenkova, A., McKeown, K.: A Survey of Text Summarization Techniques. In: [1], ch. 3, pp. 43–76 (2012)

[14] Svore, K., Vanderwende, L., Burges, C.: Enhancing single-document summarization by combining RankNet and third-party sources. In: EMNLP-CoNLL 2007, pp. 448–457 (2007)

[15] Xu, S., Yang, S., Lau, F.: Keyword extraction and headline generation using novel word features. In: AAAI 2010, pp. 1461–1466 (2010)

[16] Zhai, Z., Liu, B., Xu, H., Jia, P.: Constrained LDA for grouping product features in opinion mining. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS, vol. 6634, pp. 448–459. Springer, Heidelberg (2011)