On Handling Textual Errors in Latent Document Modeling

Tao Yang Dongwon Lee

College of IST, The Pennsylvania State University, University Park, PA 16802, U.S.A.

{tyang,dlee}@ist.psu.edu

ABSTRACT

As large-scale text data become available on the Web, textual errors in a corpus are often inevitable (e.g., digitizing historic documents). Due to the calculation of frequencies of words, however, such textual errors can significantly impact the accuracy of statistical models such as the popular Latent Dirichlet Allocation (LDA) model. To address such an issue, in this paper, we propose two novel extensions to LDA (i.e., TE-LDA and TDE-LDA): (1) The TE-LDA model incorporates textual errors into term generation process; and (2) The TDE-LDA model extends TE-LDA further by taking into account topic dependency to leverage on semantic connections among consecutive words even if parts are typos. Using both real and synthetic data sets with varying degrees of "errors", our TDE-LDA model outperforms: (1) the traditional LDA model by 16%-39% (real) and 20%-63% (synthetic); and (2) the state-of-the-art N-Grams model by 11%-27% (real) and 16%-54% (synthetic).

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms, Experimentation

Keywords

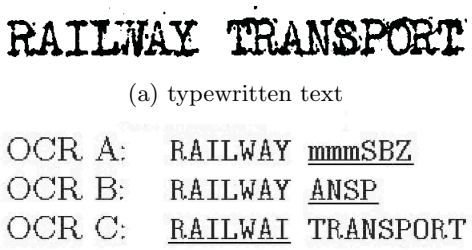
Topic Models, Textual Errors, Topic Dependency

1. INTRODUCTION

Using topic models for representing documents has recently been an area of tremendous interests in data mining and machine learning. Probabilistic topic models are stochastic models for text documents that explicitly model topics in document corpora. Because probabilistic topic models are "generative", they describe a procedure for generating documents using a series of probabilistic steps. One of the popular paradigms of topic models, characterized by the **Latent Dirichlet Allocation (LDA)** model, consists of a series of probabilistic document models and extensions where topics are modeled as hidden random variables. The LDA model is a widely used Bayesian topic model which can model the semantic relations between topics and words for document corpora. The LDA model assumes that text documents are mixtures of hidden topics and applies Dirichlet prior distribution over the latent topic distribution of a document having multiple topics. In addition, it assumes that topics are probability distribution of words and words are sampled independently from a mixture of multinomials. Since the LDA model was introduced in [4], it has quickly become one of the most popular probabilistic document modeling techniques in data mining and also has inspired a series of extensions (e.g., [18, 6, 12, 15, 20, 1, 14]).

Despite tremendous advancement in document modeling, however, we believe that two major limitations still remain during the document modeling process.

First, *the LDA model assumes that the entire document corpus is error-free to ensure accurate calculation of frequencies of words.* However, an increasing number of new large-scale text data are often machine-generated, and thus inevitably erroneous. For instance, speech recognition softwares can turn audio data into textual transcripts with varying error rates. Similarly, Optical Character Recognition (OCR) engines, despite great success in recent attempts such as Google Books or Internet Archive, are not without problems, and often produce error-abundant text data. [13] pointed out that although researchers are having increasing levels of success in digitizing hand-written manuscripts,



(a) typewritten text

(b) OCR output

Figure 1: Three examples of erroneous OCR outputs for a poor quality typewritten text (taken from [21]). Erroneous outputs are underlined.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505555.

Table 1: Top words (selected by LDA) for five topics of a small sample of `Unlv` OCR data set (erroneous words are in italic).

Top words
school, *stu*, district, teacher, angel, lo, board, *educ*
sto, food, *res*, *servic*, low, leonard, *temperatur*, retail
air, *airlin*, *fli*, american, *engin*, subject, *threate*, *pil*
mln, *dlrs*, year, net, quarter, share, *dlr*, *ln*
mcknight, vista, *de*, fleetwood, brown, *davi*, san, *democr*

error rates remain significantly high. Consider our illustrations below.

Example 1. As an illustration, consider Figure 1 that shows three examples of OCR outputs for a poor-quality typewritten text "RAILWAY TRANSPORT." All three popular OCR engines (i.e., ABBYY FineReader, OmniPage Pro, and Google Tesseract) generated outputs with one erroneous word for each. It is known that the accuracy of the LDA model often declines significantly as word error rates increase [21]. Now, consider Table 1 that shows some top words (selected by the LDA model) for five topics of a small sample of `Unlv`[1] OCR data set. From the list, we can see that there exist a lot of erroneous words in the selected top words. In addition, the words are not representative and the differences between the topics are difficult to identify. This example shows that the performance of traditional LDA model greatly suffers when documents contain erroneous words. □

Second, *since the LDA model does not consider the order of the topics and words, during parameter estimation and inference, the topics and the words are assumed to be exchangeable*. The LDA model relies on the bag-of-words document prototype. It assumes each word in a document is generated by a latent topic and explicitly models the word distribution of each topic as well as the prior distribution over topics in the document. However, we argue that the ordering of words and phrases are often critical to capture the meaning of texts in data mining tasks. Successive words in the same document are more likely to belong to the same topic. For example, a phrase "social network" is a term in modern information society under Web 2.0 while "social" is a term from traditional sociology and "network" refers to a particular term in computer science. Often, the ordering of terms carries special meanings in addition to the appearance of individual words. Therefore, incorporating topic dependency is important to learn topics and also to disambiguate words which may belong to different topics. More importantly, considering the ordering of consecutive terms can often help in dealing with errors found in parts. For instance, despite the typo "betwork" in the middle from a phrase "social betwork analysis", surrounding correct words "social" and "analysis" still have common semantic connections that could be exploited.

Motivated by the above two observations, in this paper, we introduce our novel models to tackle the issues of noisy data. In particular, we propose a new LDA model termed as TE-LDA to deal with textual errors in document corpora. We further extend it to a new TDE-LDA model in order to take into account topic dependency in the document generation process. Through a set of comprehensive experiments, the efficacy of our proposed models is validated using both real and synthetic data sets.

In summary, with respect to the document modeling problem with varying degrees of *noisy* corpora and using the perplexity as an evaluation metric, our second proposal with a better result, TDE-LDA, outperforms: (1) the traditional LDA model by 16%-39% using real data and by 20%-63% using synthetic data; and (2) the state-of-the-art N-Grams model [23] by 11%-27% using real data and by 16%-54% using synthetic data.

Our contributions are as follows:

- To the best of our knowledge, this is the first attempt to solve the noisy data problem in document modeling. We formally incorporate textual errors into the document generation process and show how to apply it to the model formulation.
- We discard the bag-of-words assumption in the LDA model. Instead, we assume that successive words in the document are more likely to have the same topic. We model the topics in a document to form a Markov chain with a transition probability and show how to incorporate dependency of topics into the generative process.
- We apply our proposed models to both real and synthetic data sets and compare the performance against the traditional LDA model and the state-of-the-art N-Grams model, and report promising results of our proposal in terms of perplexity.

The rest of this paper is organized as follows. Section 2 briefly reviews the related research work in document modeling. Section 3 gives a general overview of the traditional LDA model. Section 4 introduces our proposed models and presents the detailed model formulation. Section 5 presents the results of extensive experimental evaluations of applying our document models to both real and synthetic data sets. Finally, section 6 concludes the paper and discusses the directions of future work.

2. RELATED WORK

Probabilistic document modeling has received tremendous attention in the data mining community. A series of probabilistic models have been introduced to simulate the document generation process. These models include the Naive Bayesian model and the Probabilistic Latent Semantic Indexing (PLSI) model [11]. The LDA model has become most popular in the data mining and information retrieval community due to its solid theoretical statistical foundation and promising performance. A wide variety of extensions of LDA model have been proposed for different modeling purposes in different contexts. For example, the author-topic model [18, 20] uses the authorship information with the words to learn topics. The correlated LDA model learns topics simultaneously from images and caption words [6]. The Link-LDA model and Topic-link LDA model [12] represent topics and author communities using both content words and links between documents.

Most topic modeling techniques require the bag-of-words assumption [4]. They treat the generation of all words independently from each other given the parameters. It is true

[1] `http://code.google.com/p/isri-ocr-evaluation-tools/updates/list`

that these models with the bag-of-words assumption simplified the problem domain and enjoyed a big success, hence attracted a lot interests from researchers with different backgrounds. Some researchers tried to drop this assumption to assume that words are generated dependently. For example, [22] developed a bigram topic model on the basis of the hierarchical Dirichlet language model, by incorporating the concept of topic into bigram models. [23] proposed a topical n-grams model to automatically determines whether to form an n-gram based on the surrounding context of words. [1] developed a probabilistic time series model to capture the evolution of topics in large document corpora. [10] proposed a hidden topic Markov model (HTMM) to incorporate a hidden Markov structure into LDA. However, their model is based on the assumption that all words in the same sentence must have the same topic and imposes a sentence boundary for words. [2] proposed a correlated topic model which allows for correlations between topic assignments and draws a topic proportion from a logistic normal instead of a Dirichlet distribution. [9] proposed the HMMLDA model as a generative composite model which considers both short-range syntactic dependencies and long-range semantic dependencies between words. [5] proposed a probabilistic model to match documents at both general topic level and specific word level in information retrieval tasks.

Recently, a number of researchers proposed topic segmentation models which are closely related to topic models. Topic segmentation is to split a text stream into coherent and meaningful segments. For example, the aspect hidden markov (HMM) model proposed in [3] models unstructured data which contains streams of words. In the aspect HMM model, documents are separated into segments and each segment is generated from a unique topic assignment and there is no mixture of topics during the inference. [17] proposed a Bayesian approach to linear topic segmentation which assumes some numbers of hidden topics are shared across multiple documents. [8, 7] further extended this work by marginalizing the language models using the Dirichlet compound multinomial distribution, and applied the model to both linear topic segmentation and hierarchical topic segmentation for the purpose of multi-scale lexical cohesion. [19] proposed a statistical model that combines topic identification and segmentation in text document collections, and the model is able to identify segments of text which are topically coherent and cluster the documents into overlapping clusters as well. Note that the Markov transition is based on segments with each being generated from a linear combination of the distributions associated with each topic.

Most topic modeling techniques require clean document corpora. This is to prevent the models from confusing patterns which emerge in the noisy text data. Recent work in [21] is the first comprehensive study of document clustering and LDA on synthetic and real-word Optical Character Recognition (OCR) data. The character-level textual errors introduced by OCR engines serve as baseline document corpora to understand the accuracy of document modeling in erroneous environment. As pointed out by these researchers, even on clean data, LDA will often do poorly if the very simple feature selection step of removing stop-words is not performed first. The study shows that the performance of topic modeling algorithms degrades significantly as word error rates increase. Our work in this paper is a substantial extension of our preliminary work [25] with a novel model proposed, a comparison to state-of-the-art model, and a much more comprehensive empirical study.

Table 2: Notations

Symbol	Description
D	total number of documents
W	total number of word tokens
T	total number of topics
N_d	total number of words in document d
$w_{d,i}$	ith word in document d
$z_{d,i}$	latent topic at ith word in document d
$\theta_{d,i}$	probability of topic i in document d
$\phi_{t,w}$	probability of word w in topic t

3. THE LDA MODEL

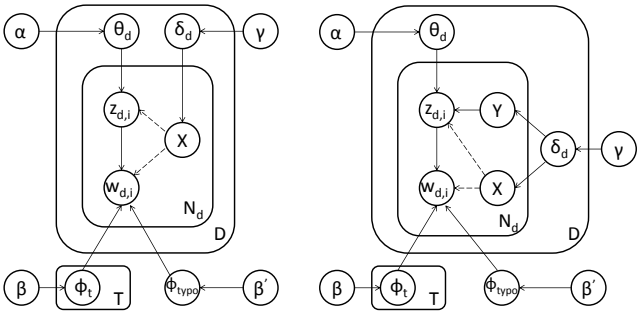
In this section, we give a brief overview of the *Latent Dirichlet Allocation* (LDA) model. [4] introduced the LDA model as a semantically consistent topic model, which attracted considerable interest from both the statistical machine learning and natural language processing communities. LDA models documents by assuming that a document is composed by a mixture of hidden topics and that each topic is characterized by a probability distribution over words.

The model is known as a graphical model for topic discovery. The notations are shown in Table 2. θ_d denotes a T-dimensional probability vector and represents the topic distribution of document d. ϕ_t denotes a W-dimensional probability vector where $\phi_{t,w}$ specifies the probability of generating word w given topic t. $Multi(.)$ denotes multinomial distribution. $Dir(.)$ denotes Dirichlet distribution. α is a T-dimensional parameter vector of the Dirichlet prior distribution over θ_d, and β is a W-dimensional parameter vector of the Dirichlet prior distribution over ϕ_t. The process of generating documents is shown in Algorithm 1.

Algorithm 1: The LDA Model.

1. For each of the T topics t, sample a distribution over words ϕ_t from a Dirichlet distribution with hyperparameter β;
2. For each of the D documents d, sample a vector of topic proportions θ_d from a Dirichlet distribution with hyperparameter α;
3. For each word $w_{d,i}$ in document d, sample a topic $z_{d,i}$ from a multinomial distribution with parameters θ_d;
4. Sample word $w_{d,i}$ from a multinomial distribution with parameters $\phi_{z_{d,i}}$.

Performing exact inferences for the LDA model is intractable due to the choice of distribution and the complexity of the model. The existing approximate algorithms for parameter estimation and inference of the LDA model include variational methods [4], EM algorithm [11] and Markov Chain Monte Carlo (MCMC) [16]. One assumption in the generation process above is that the number of topics is given and fixed. LDA model considers documents as "bags of words", i.e., there is no ordering between words and all words as well as their topic assignments in the same document are assumed to be conditionally independent. Furthermore, finding good estimates for the parameters of LDA model requires accurate counts of the occurrences and co-occurrences of words, which in turn requires a "perfect" corpus with errors as few as possible.



(a) TE-LDA Model (b) TDE-LDA Model

Figure 2: Proposed Models.

Algorithm 2: The TE-LDA Model.

For each of the D documents d , sample $\theta_d \sim$ Dir(α)and $\delta_d \sim$ Beta(γ);
For each of the T topics t, sample $\phi_t \sim$ Dir(β);
Sample $\phi_{typo} \sim$ Dir(β');
foreach N_d *words* $w_{d,i}$ *in document* d **do**
  Sample a flag $X \sim$ Binomial(δ_d);
  if $X = 1$ **then**
    Sample a topic $z_{d,i} \sim$ Multi(θ_d);
    Sample a word $w_{d,i} \sim$ Multi($\phi_{z_{d,i}}$);
  end if
  if $X = 0$ **then**
    Sample a word $w_{d,i} \sim$ Multi(ϕ_{typo});
  end if
end foreach

4. PROPOSED MODELS

To account for textual errors in the traditional LDA topic model, in this section, we propose a new LDA model termed as TE-LDA (LDA with **T**extual **E**rrors) to take into account noisy data in the document generation process. We further extend it to a new TDE-LDA (LDA with **T**opic **D**ependency and textual **E**rrors) model in order to take into account topic dependency in the document generation process. We explain the details of our proposed models in the following.

4.1 TE-LDA

In this model, we distinguish the words in the documents and separate them as tokens and typos. Given a document, each word has a probability to be an error and we want to capture this probability structure in the term generation process. In order to reflect the nature of textual errors in the generative model, we adopt a switch variable to control the influence of errors on the term generation.

The proposed model is illustrated in Figure 2(a). Here we introduce some notations used in the graphical model: D is the number of documents, T is the number of latent topics, N_d is the total number of words in document d (with $N_d = N_{term} + N_{typo}$, the sum of all the true terms and typos). α, β and β' are parameters of Dirichlet priors, θ_d is the topic-document distribution, ϕ_t is the term-topic distribution. ϕ_{typo} is the term distribution specifically for typos. We include an additional binomial distribution δ with a $Beta$ prior of γ which controls the fraction of errors in documents.

For each word w in a document d, a topic z is sampled first and then the word w is drawn conditioned on the topic. The document d is generated by repeating the process N_d times. To decide if each word is an error or not, a switch variable X is introduced. The value of X (which is 0 or 1) is sampled based on a binomial distribution δ with a $Beta$ prior distribution of γ. When the sampled value of X equals 1, the word w is drawn from the topic z_t which is sampled from the topics learned from the words in document d. When the value of X equals 0, the word w is drawn directly from the term distribution for errors. Overall, the generation process for TE-LDA can be described in Algorithm 2.

4.2 Topic Dependency

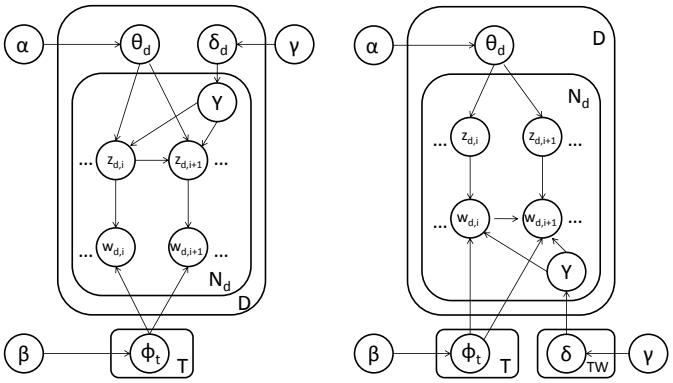
As we mentioned in the introduction section, LDA relies on the bag-of-words assumption. However, in many data mining tasks, words are often connected in nature and successive words in the document are more likely to be assigned the same topic. Therefore, incorporating topic dependency is important to capture the semantic meaning of texts and also to disambiguate words which may belong to different topics. Even in noisy text corpora, consecutive words may be dependent to each other regardless of textual errors. For example, in a phrase "text dat mining" with textual error "dat" as typo of word "data", the correct word "text" and "mining" still have semantic connections and both words belong to the same topic of data mining. Hence, incorporating this correlation gives a more realistic model of the latent topic structure and we expect to obtain better generalization performance quantitatively. To apply topic dependency and drop the bag-of-words assumption, we assume the topics in a document form a Markov chain with a transition probability that depends on a transition variable Y. When Y equals 0, a new topic is drawn from θ_d. When Y equals 1, the current topic of word w_i is equivalent to the previous topic of word w_{i-1}.

[23] proposed a topical n-grams model to automatically determine whether to form an n-gram based on the surrounding context of words. The n-grams model is an extension of the bigram topic model, which makes it possible to decide whether to form a bigram for the same two consecutive words depending on the nearby context. As a result, the n-grams model imposes a Markov relation on the word set. In contrast, topic dependency considers the relation between consecutive *topics* instead of words. That is, the Markov relation is on the topic set instead of the word set. Figure 3(a) shows an alternative graphical model for applying topic dependency to LDA. The n-grams model is illustrated in Figure 3(b). We incorporate topic dependency in our proposed TE-LDA model in the following.

4.3 TDE-LDA

We extend our TE-LDA model and further incorporate topic dependency into one unified model, named as TDE-LDA. The proposed model is illustrated in Figure 2(b).

For each word w in a document d, a topic z is sampled first and then the word w is drawn conditioned on the topic. The document d is generated by repeating the process N_d times. To decide if each word is an error or not, a switch variable X is introduced. The value of X (which is 0 or 1) is sampled based on a binomial distribution δ with a $Beta$ prior distribution of γ. When the sampled value of X equals 1, the word w is drawn from the topic z_t which is sampled from the topics learned from the words in document d. To decide if the current topic is dependent to the previous topic



(a) Topic dependency (b) N-Grams

Figure 3: Comparison of topic dependency and term dependency.

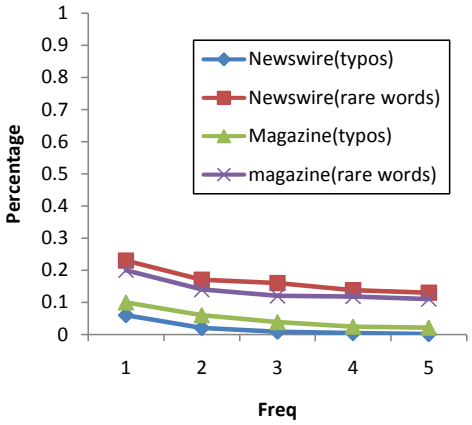


Figure 4: Comparison of percentages of typos and rare words.

or not, a switch variable Y is introduced. The value of Y (which is 0 or 1) is sampled based on a binomial distribution δ with a *Beta* prior distribution of γ. When the sampled value of Y equals 1, the topic z_i is assigned to be identical to the previous one z_{i-1} to reflect the dependency between them. When the value of Y equals 0, the topic z_i is sampled from the topics learned from the words in document d. And the word w is drawn from the topic z_t. When the value of X equals 0, the word w is drawn directly from the term distribution for errors. The generation process for TDE-LDA can be described in Algorithm 3.

Algorithm 3: The TDE-LDA Model.

For each of the D documents d, sample $\theta_d \sim \text{Dir}(\alpha)$and $\delta_d \sim \text{Beta}(\gamma)$;
For each of the T topics t, sample $\phi_t \sim \text{Dir}(\beta)$;
Sample $\phi_{typo} \sim \text{Dir}(\beta')$;
foreach N_d *words* $w_{d,i}$ *in document* d **do**
  Sample a flag $X \sim \text{Binomial}(\delta_d)$;
  if $X = 1$ **then**
    Sample a flag $Y \sim \text{Binomial}(\delta_d)$;
    if $Y = 1$ **then**
      Assign a topic $z_{d,i} = z_{d,i-1}$;
    end if
    if $Y = 0$ **then**
      Sample a topic $z_{d,i} \sim \text{Multi}(\theta_d)$;
    end if
    Sample a word $w_{d,i} \sim \text{Multi}(\phi_{z_{d,i}})$;
  end if
  if $X = 0$ **then**
    Sample a word $w_{d,i} \sim \text{Multi}(\phi_{typo})$;
  end if
end foreach

4.4 Discussion

In this section, we discuss two important issues on our proposed models.

Rare Words vs. Textual Errors
In terms of frequency of words, note that it is difficult to differentiate between rare-but-correct-English words and typos because both appear rather seldom in the corpus. Without prior knowledge of grammar and syntax of human language or helps of dictionary, that is, machines cannot solely rely on the word frequency to tell the difference between a textual error and a rare word. To illustrate this point, we selected the `Reuters` newswire data set (to be explained in Section 5.1) and combined two OCR `Magazine` data sets. We calculated the percentages of words that appear from once to five times in the corpus. In Figure 4, the percentage curves of both typos and rare words exhibit very similar patterns in both corpora, making a computation-based differentiation hard. Therefore, our models adopt a supervised approach to distinguish rare words and textual errors in the document modeling process. One may use linguistic characteristics to differentiate typos in an unsupervised fashion. However, since the immediate goal of this paper is first to evaluate the validity of incorporating textual errors into document modeling process, we rather leave the development of more sophisticated modeling methods for future work.

Topic vs. Term Dependency
The bigram topic model and n-grams model we mentioned in section 2 determine whether to form a bigram or an n-gram based on the surrounding words in the document. Although these models show better generalization performance over LDA, we argue that incorporating term dependency is not suitable in noisy text data for two reasons. First, in noisy document corpora, simply forming bigram or n-gram between consecutive words will increase the overall error rate. This is because an erroneous word will impact both the previous word and the succeeding word in terms of term combination. But it only impacts itself under the bag-of-words assumption for documents. Secondly, even though our TE-LDA model has a mechanism to distinguish between textual errors and correct words, by skipping typos the document model may generate incorrect bigram or n-gram word combinations which, in turn, decreases the accuracy of generalization performance. Therefore, we only consider topic dependency in order to capture the semantic relation of words. As a result, in this paper, we select the traditional LDA model and the n-grams model without error modeling as baselines.

5. EXPERIMENTAL VALIDATION

In order to validate our proposed models, we applied it to the *document modeling* problem. We trained our new models as well as the traditional LDA model on both synthetic

Table 3: Summary of data sets.

Name	Error	Domain	# of documents	# of unique terms	AVG document length
OCR Business	real	business documents	220	4,556	252
OCR Magazine	real	magazine articles	320	9,842	462
OCR Legal	real	legal documents	300	4,608	339
OCR Newspaper	real	newspaper articles	240	5,948	346
OCR Magazine2	real	magazine documents	200	10,485	872
OCR BYU	real	communique documents	600	33,749	529
TREC AP	synthetic	newswire articles	16,333	23,075	458
NIPS	synthetic	proceedings	1,740	13,649	2,843
Reuters	synthetic	newswire articles	12,902	12,112	223

and real text corpora to compare the generalization performance of these models. The documents in the document corpora are treated as unlabeled and the goal is to achieve high likelihood on a held-out test data [4]. In our experiments, each model was trained on 90% of the documents in each data set with fixed parameters α=0.5, β=0.01, β'=0.01 and γ=0.1 for simplicity and performance. The trained model was used to calculate the estimate of the marginal log-likelihood of the remaining 10% of the documents.

5.1 Data Sets

Table 3 shows the summary of both real and synthetic data sets that we used in our experiments.

First, we prepared real data sets that contain varying degrees of errors in texts. From the PDF images in the data set, Unlv, using one of the most popular OCR engines (Google Tesseract), we converted PDF images to a textual document corpus. Since Unlv has the full texts as the ground truth, by comparing the transcript generated from OCR, we can exactly pinpoint which words are errors. In the end, we prepared five subsets: Business, Magazine, Legal, Newspaper, Magazine2. Similarly, we prepared another real corpus called BYU[2] which consists of 600 of the Eisenhower World War II communiques. This data set contains the daily progress of the Allied campaign until the German surrender. Example documents from Newspaper data set and BYU data set are shown in Figure 5. The quality of these originals is quite poor, hence the error rate is pretty high for the outputs of OCR engine. Note that in these real data sets, we cannot control the degrees of errors, and the error rates are determined by the OCR engine.

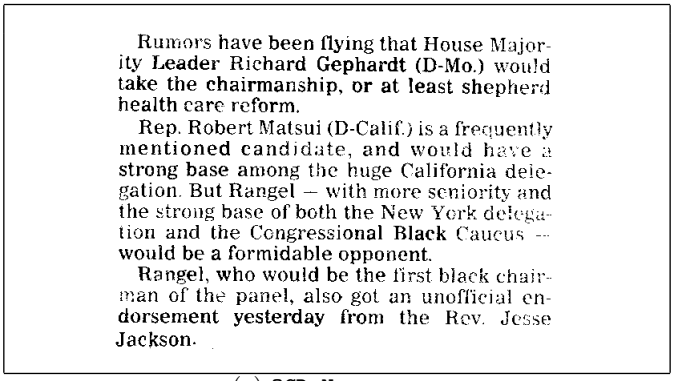
Second, to conduct more controlled experiments, we also prepared synthetic data sets. In particular, we used three well-known benchmark data sets in the document modeling literature: TREC AP, NIPS, and Reuters-21578. The TREC Associate Press (AP) data set[3] contains 16,333 newswire articles with 23,075 unique terms. The NIPS data set[4] consists of the full text of the 13 years of proceedings from 1988 to 2000 Neural Information Processing Systems (NIPS) Conferences. The data set contains 1,740 research papers with 13,649 unique terms. The Reuters-21578 data set[5] consists of newswire articles classified by topics and ordered by their date of issue. The data set contains 12,902 documents and 12,112 unique terms.

[2] http://www.lib.byu.edu/dlib/spc/eisenhower

[3] http://www.daviddlewis.com/resources/testcollections/trecap/

[4] http://www.cs.nyu.edu/~roweis/data.html

[5] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

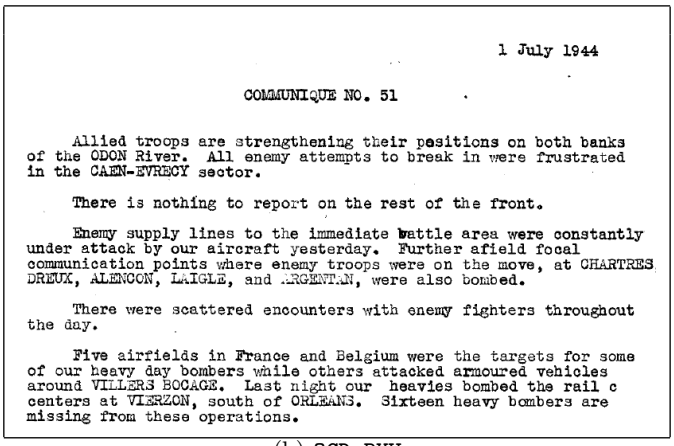


Rumors have been flying that House Majority Leader Richard Gephardt (D-Mo.) would take the chairmanship, or at least shepherd health care reform.

Rep. Robert Matsui (D-Calif.) is a frequently mentioned candidate, and would have a strong base among the huge California delegation. But Rangel – with more seniority and the strong base of both the New York delegation and the Congressional Black Caucus – would be a formidable opponent.

Rangel, who would be the first black chairman of the panel, also got an unofficial endorsement yesterday from the Rev. Jesse Jackson.

(a) OCR Newspaper



1 July 1944

COMMUNIQUE NO. 51

Allied troops are strengthening their positions on both banks of the ODON River. All enemy attempts to break in were frustrated in the CAEN-EVRECY sector.

There is nothing to report on the rest of the front.

Enemy supply lines to the immediate battle area were constantly under attack by our aircraft yesterday. Further afield focal communication points where enemy troops were on the move, at CHARTRES DREUX, ALENCON, LAIGLE, and ARGENTAN, were also bombed.

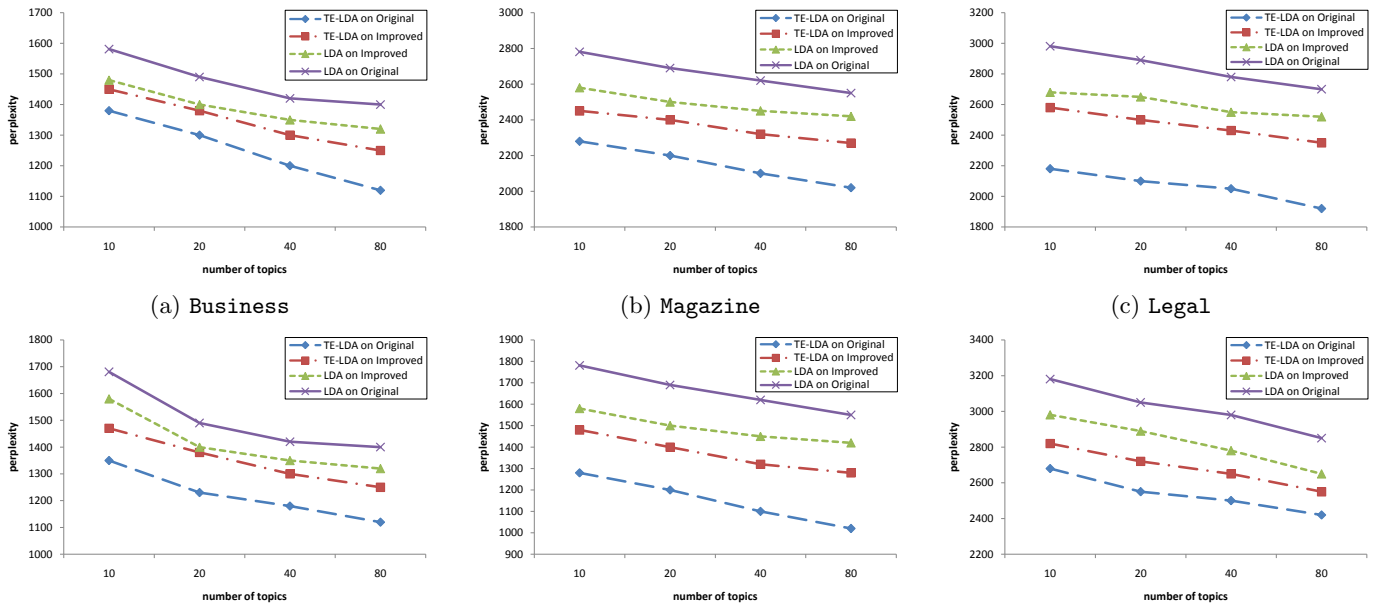
There were scattered encounters with enemy fighters throughout the day.

Five airfields in France and Belgium were the targets for some of our heavy day bombers while others attacked armoured vehicles around VILLERS BOCAGE. Last night our heavies bombed the rail c centers at VIERZON, south of ORLEANS. Sixteen heavy bombers are missing from these operations.

(b) OCR BYU

Figure 5: Example documents from UNLV and BYU data sets.

For all the above synthetic data sets, we generated erroneous corpora to simulate different levels of *Word Error Rates* (WER) – i.e., the ratio of word insertion, substitution and deletion errors in a transcript to the total number of words. Then, we closely studied the impact of textual errors in document modeling. In our experiments, we used three types of edit operations (i.e., insertion, deletion and substitution) in all the documents as follows: (1) insertion: a number of terms are randomly selected in a uniform fashion to insert a single character into the terms; (2) deletion: a number of terms are randomly selected in a uniform fashion to delete a single character from the terms; (3) substitution: a number of terms are randomly selected in a uniform fashion to change a single character of the terms. Note that multiple edit operations are not allowed for a single word.



(a) Business (b) Magazine (c) Legal

(d) Newspaper (e) Magazine2 (f) BYU

Figure 6: Perplexity of different models in original and improved Unlv and BYU data sets. From (a) to (f), the data sets are Business, Magazine, Legal, Newspaper, Magazine2 from Unlv and BYU. The WER of original data sets are 0.2856, 0.3158, 0.3148, 0.372, 0.3739 and 0.4856, respectively. The WER of improved data sets (using the technique from [24]) are 0.2653, 0.2893, 0.2985, 0.3468, 0.3518 and 0.4438, respectively.

Let S, D and I denote the number of substitution, deletion and insertion operations, and let N denote the total number of words. Then, WER is calculated as follows. The procedure repeats until the desirable WER is achieved.

$$WER = \frac{S + D + I}{N}$$

5.2 Evaluation Metrics

The purpose of document modeling is to estimate the density distribution of the underlying structure of data. The common approach to achieve this goal is by evaluating the document model's generalization performance on new unseen documents. In our experiments, we calculated the *perplexity* of a held-out test set to evaluate the models. In language modeling, the perplexity quantifies the goodness of measuring the likelihood of a held-out test data to be generated from the learned distribution of the trained model. In particular, it is monotonically decreasing in the likelihood of the test data, which means a lower perplexity score corresponds to better generalization performance of the document model. Formally, for a test data of D_{test} documents the perplexity score is calculated as follows [4, 16]:

$$perplexity(D_{test}) = \exp\{\frac{-\sum_{d=1}^{D_{test}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D_{test}} N_d}\}$$

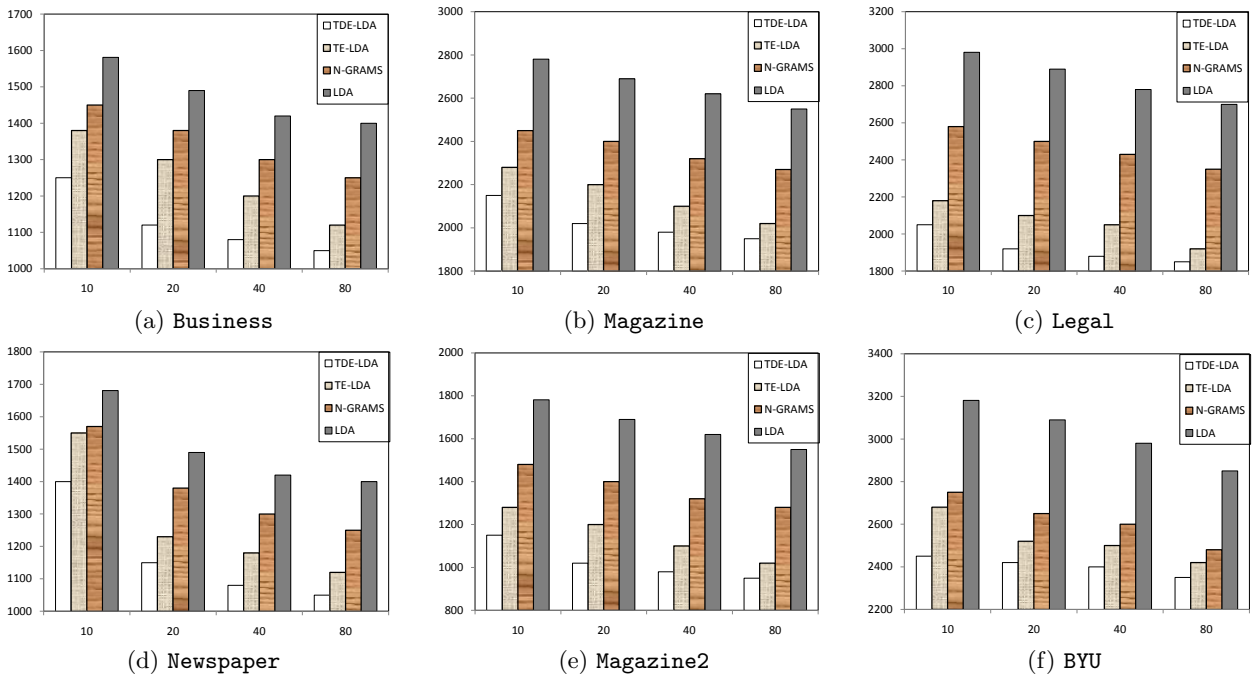
$$p(\mathbf{w}_d) = \sum_{k=1}^{K} p(w_d|z_k) p_{test}(z_k|d)$$

In the above equations, the probability $p(w_d|z_k)$ is learned from the training process and $p_{test}(z_k|d)$ is estimated through an additional Gibbs sampling process on the test data based on the parameters ϕ and δ learned from training data.

5.3 Comparison between TE-LDA and Baseline LDA with Error Correction

We first examine the performance of our TE-LDA model on real OCR data sets. Note that our immediate objective is to evaluate the validity of incorporating textual errors into document modeling process. This is based on the fact that most large-scale text data are machine-generated and thus inevitably contain many types of noise. As a novel solution, our TE-LDA model is developed from the traditional LDA model by adding a switch variable into the term generation process in order to tackle the issue of noisy text data. Hence, in this experiment, we compare the generalization of our TE-LDA model with the traditional LDA on various erroneous OCR text data. For example, each subset of real OCR data Unlv has a fixed WER, determined by the OCR engine. Due to the poor quality of PDF images and imperfect OCR process, WERs range from 0.2856 to 0.3739. That is, about 28–37% of words in the corpus could be erroneous words. Similarly, the WER of real OCR data BYU is as high as 48%.

Recently, [24] proposed an algorithm for applying topic modeling to OCR error correction. The algorithm builds two models on an OCR document. One is a topic model which provides information about word probability and the other is an OCR model which provides the probability of character errors. The algorithm can reduce OCR errors by around 7%. We use the same error detecting technique to further correct our six real OCR data sets and then compare the performance of our TE-LDA model with the traditional LDA again. By doing so, we aim at finding out how the behavior of both topic models changes as the error rate changes on real OCR data. Figure 6 shows the perplexity of TE-LDA and LDA as a function of the number of hidden topics



(a) Business (b) Magazine (c) Legal

(d) Newspaper (e) Magazine2 (f) BYU

Figure 7: Perplexity of different models as a function of the number of topics (X-axis) in Unlv and BYU data sets. From (a) to (f), the data sets are Business, Magazine, Legal, Newspaper, Magazine2 from Unlv and BYU. The fixed word error rates (WER) of these data sets are 0.2856, 0.3158, 0.3148, 0.372, 0.3739 and 0.4856, respectively. Note the relatively high WERs due to the poor quality of PDF images in Unlv and BYU data sets.

(e.g., 10, 20, 40, and 80). As we can see, our proposed TE-LDA model consistently outperforms the traditional LDA model on both original and improved Unlv as well as BYU data sets. An interesting finding is that LDA performs better on improved corpora while TE-LDA performs better on original corpora. This is reasonable because our model is specifically designed to deal with textual errors in modeling noisy text documents and can achieve better generalization performance as the word error rates increase.

5.4 Comparison among Different Models

In this section, we systematically evaluate the performance of different models using various real and synthetic data sets. Since our purpose is to understand the performance of document modeling in erroneous environment, we compare the performance of our proposed models and the baseline models without removal of typos in text corpora.

Results using Real Data Sets

We first compare the performance of our proposed models with the traditional LDA model and Wang's n-grams model on the real OCR data sets. Figure 7 shows the perplexity of TE-LDA and TDE-LDA models as a function of the number of hidden topics (e.g., 10, 20, 40, and 80) on the five subsets of Unlv corpus and the BYU corpus. As we can see, despite high WERs and different document themes among these data sets, our proposed TE-LDA and TDE-LDA models consistently outperform the traditional LDA model and the n-grams model. Note also that TDE-LDA is the best among the proposed models and the baseline models, which demonstrates that considering topic dependency improves the generalization performance of topic models in the context of noisy data.

Table 4: Comparison of the selected top words using LDA vs. N-grams vs. our proposed models on a small sample of Unlv OCR data set. OCR-introduced erroneous words are in *italic*.

Model	Top words
LDA	air, *airlin*, *fli*, american, *engin*, subject, *threate*
N-GRAMS	american *airlin*, air flight, *threate* flight, *boe* plane
TE-LDA	air, american, plane, flight, bomb, pilot, airport
TDE-LDA	air, plane, pilot, american, passenger, aboard, bomb

Table 4 shows examples of top words selected by LDA and the n-grams model as well as our models on the topic 3 of Table 1. From the table, note that LDA suffers from choosing many OCR-introduced erroneous words as top words. Furthermore, the n-grams model tends to select several erroneous n-gram words as well. On the contrary, both TE-LDA and TDE-LDA models selected no erroneous top words, highlighting the superiority of our models in dealing with noisy text data. Overall, compared to others, our TDE-LDA model can select meaningful and generic top words or highly related words and make the topic more understandable.

Results using Synthetic Data Sets

We then systematically compare the performance of our proposed models with the traditional LDA model as well as Wang's n-grams model on the synthetically generated erroneous corpora. In this comparison, we simulate different levels of WER (e.g., 0.01, 0.05, 0.1). Figures 8(a)-(c) show the perplexity of TE-LDA and TDE-LDA models as a function of the number of hidden topics in the TREC AP corpus. As we can see from Figures 8(a)-(c), at different levels of WER, our TE-LDA and TDE-LDA models consistently

outperform the traditional LDA model. Furthermore, as WER increases, the margin of improvement increases. This is due to the incorporation of textual errors into the generation of terms in the document modeling process. We can also see that the models with consideration of topic or term dependency outperform the ones without that, regardless of whether we take into account textual errors during term generation. However, TDE-LDA is the best among the models and show better generalization of incorporating topic dependency in noisy text data. This demonstrates the improved performance of topic models with the removal of bag-of-words assumption.

In Figures 8(d)-(f), we fix the number of topics K and demonstrate how the different models perform as the WER increases in the `TREC AP` corpus. An interesting finding here is that the perplexity of both LDA and n-grams models increases as the word error rates increase. This is because these two models do not consider the errors in the term generation where the accuracy of calculation of word frequencies is affected. In contrast, our TE-LDA and TDE-LDA models outperform the other two and the margin of improvement increases as the word error rates increase. The experimental results in the `NIPS` (Figures 8(g)-(l)) and `Reuters` (Figures 8(m)-(r)) corpora show similar perplexity patterns.

6. CONCLUSION AND FUTURE WORK

In this paper, we have proposed two extensions to the traditional LDA model to account for textual errors in latent document modeling. Our work is motivated by the facts that textual errors in document corpora are often abundant and separating words cannot completely capture the meaning of texts in data mining tasks. To overcome these constraints, we proposed our TE-LDA and TDE-LDA models to incorporate textual errors into the term generation process. Both TE-LDA and TDE-LDA adopt a switching mechanism to explicitly determine whether the current term is generated from the topic-document distribution through the general topic generation route or from a special word distribution through the typo processing route. However, TDE-LDA models the transition of topics between consecutive words as a first-order Markov process. Through extensive experiments, we have shown that our proposed models are able to model the document corpus in a more meaningful and realistic way, and achieve better generalization performance than the traditional LDA model and the n-grams model.

Many directions are ahead. First, we plan to infer more complex topic structures and conduct tests of statistically significant differences across all the models. Second, we plan to apply our proposed models to handling textual errors in user-generated contents on social media.

7. ACKNOWLEDGMENTS

This research was in part supported by NSF awards of DUE-0817376, DUE-0937891, and SBIR-1214331.

8. REFERENCES

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[2] D. M. Blei and J. D. Lafferty. A correlated topic model of science. In *Annals of Applied Statistics*, 2007.

[3] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden markov model. In *SIGIR*, 2001.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, 2003.

[5] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, 2006.

[6] X. Chen, C. Lu, Y. An, and P. Achananuparp. Probabilistic models for topic learning from images and captions in online biomedical literatures. In *CIKM*, 2009.

[7] J. Eisenstein. Hierarchical text segmentation from multi-scale lexical cohesion. In *HLT-NAACL*, 2009.

[8] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In *EMNLP*, 2008.

[9] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*, 2005.

[10] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden topic markov models. In *AISTATS*, 2007.

[11] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.

[12] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: Joint models of topic and author community. In *ICML*, 2009.

[13] W. B. Lund and E. K. Ringger. Improving optical character recognition through efficient multiple system alignment. In *JCDL*, 2009.

[14] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *SIGKDD*, 2008.

[15] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *SIGKDD*, 2006.

[16] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *SIGKDD*, 2008.

[17] M. Purver, T. L. Griffiths, K. P. Kording, and J. B. Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *ACL*, 2006.

[18] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.

[19] M. M. Shafiei and E. E. Milios. A statistical model for topic segmentation and clustering. In *AI*, 2008.

[20] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *SIGKDD*, 2004.

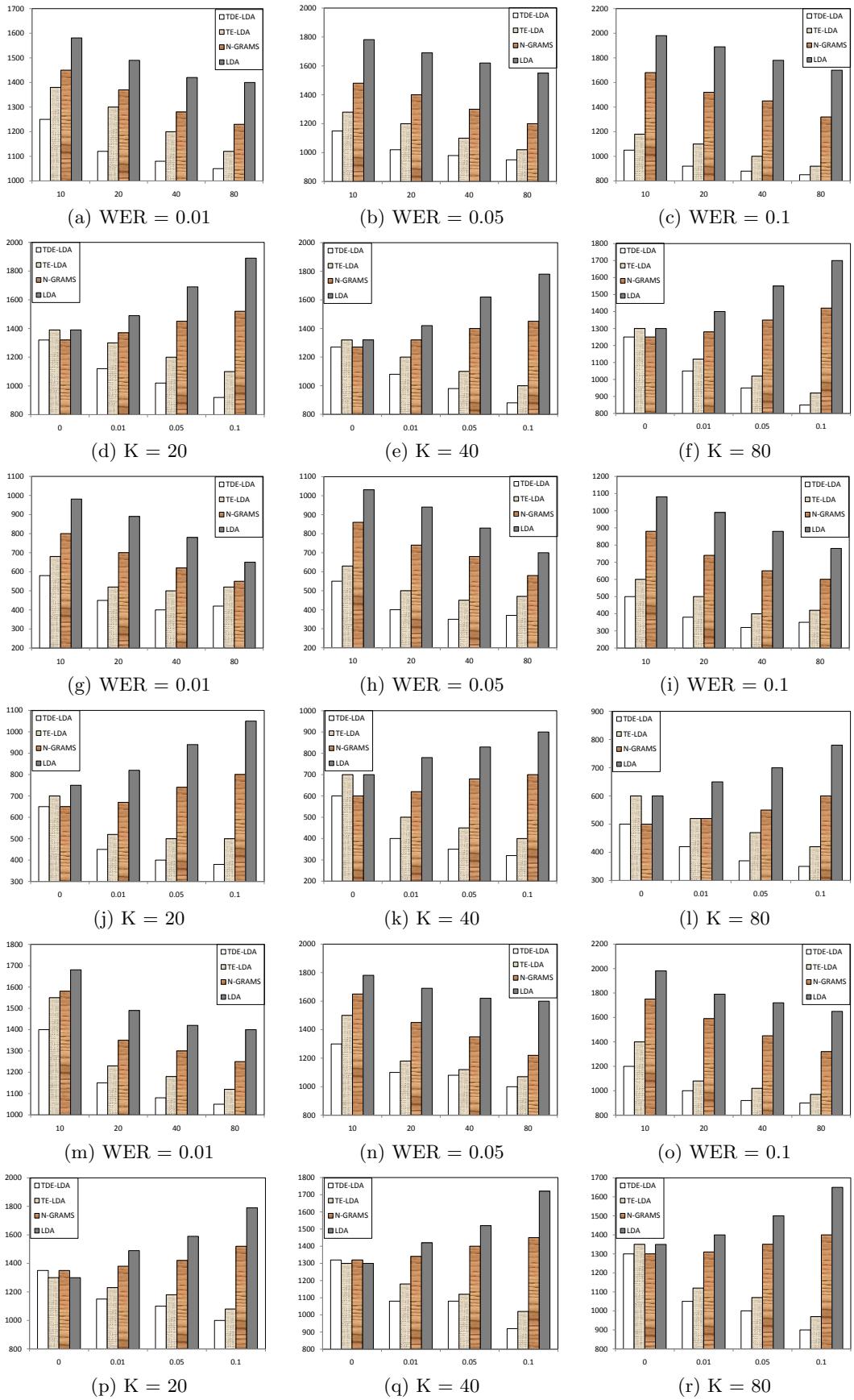
[21] D. D. Walker, W. B. Lund, and E. K. Ringger. Evaluating models of latent document semantics in the presence of ocr errors. In *EMNLP*, 2010.

[22] H. Wallach. Topic modeling: Beyond bag-of-words. In *ICML*, 2006.

[23] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, 2007.

[24] M. Wick, M. Ross, and E. Miller. Context-sensitive error correction: Using topic models to improve ocr. In *ICDAR*, 2007.

[25] T. Yang and D. Lee. Towards noise-resilient document modeling. In *CIKM*, 2011.



(a) WER = 0.01 (b) WER = 0.05 (c) WER = 0.1

(d) K = 20 (e) K = 40 (f) K = 80

(g) WER = 0.01 (h) WER = 0.05 (i) WER = 0.1

(j) K = 20 (k) K = 40 (l) K = 80

(m) WER = 0.01 (n) WER = 0.05 (o) WER = 0.1

(p) K = 20 (q) K = 40 (r) K = 80

Figure 8: Performance summary using three synthetic data sets. Perplexity of different models as a function of the number of topics (X-axis) in (a)-(c) `TREC AP`, (g)-(i) `NIPS` and (m)-(o) `Reuters` data sets respectively. Perplexity of different models as a function of WER (X-axis) in (d)-(f) `TREC AP`, (j)-(l) `NIPS` and (p)-(r) `Reuters` data sets respectively.