Topic Stability over Noisy Sources

Jing Su[1], Derek Greene[2] and Oisín Boydell[1]

[1]Centre for Applied Data Analytics Research, University College Dublin
[2]School of Computer Science & Informatics, University College Dublin
{jing.su, derek.greene, oisin.boydell}@ucd.ie

Abstract

Topic modelling techniques such as LDA have recently been applied to speech transcripts and OCR output. These corpora may contain noisy or erroneous texts which may undermine topic stability. Therefore, it is important to know how well a topic modelling algorithm will perform when applied to noisy data. In this paper we show that different types of textual noise can have diverse effects on the stability of topic models. On the other hand, topic model stability is not consistent with the same type but different levels of noise. We introduce a dictionary filtering approach to address this challenge, with the result that a topic model with the correct number of topics is always identified across different levels of noise.

1 Introduction

Topic modelling techniques are widely applied in text retrieval tasks. Frequently these techniques have been applied to explore high-quality texts such as news articles (Newman et al., 2006) and blog posts (Yokomoto et al., 2012) which have low levels of noise (i.e. few missing, misspelled, or incorrect terms and phrases). However, with the reduction in the cost of automatic speech transcription and optical character recognition (OCR) technologies, the range of sources that topic modelling can now be applied to is growing. One challenge with such textual sources is dealing with their inherent noise. In speech to text transcriptions, humans in general manage a WER of 2% to 4% (Fiscus et al., 2007). When transcribing with a vocabulary size of 200, 5000 and 100000 terms, the reported corresponding word error rates are 3%, 7% and 45% respectively. The best accuracy for broadcast news transcription is 13% (Pallet, 2003), but this drops below 25.7% in conference transcription and gets worse in casual conversation (Fiscus et al., 2007). These results indicate that the difficulty of automatic speech recognition increases with vocabulary size, speaker dependency, and level of crosstalk.

Noise aside, many of these newly available sources contain rich and valuable information that can be analysed through topic modelling. For example, automatic speech transcription applied to call centre audio recordings is able to capture a level of detail that is otherwise unavailable unless the call audio is manually reviewed which is infeasible for large call volumes. In this case topic modelling can be applied to transcribed text to extract the key issues and emerging topics of discussion.

In this study we propose a method for simulating various types of transcription errors, for the purpose of examining the effect of noise on topic modelling algorithms. We then test the robustness of probabilistic topic modelling via Latent Dirichlet Allocation (LDA) using a topic stability measure (Greene et al., 2014) over a variety of corpora. The stability of a clustering model refers to its ability to consistently produce similar solutions on data originating from the same source (Lange et al., 2004) (Ben-David et al., 2007). A high level of agreement between the resulting clusterings indicates high stability, in turn suggesting that the current model fits the data well. Consequently, we measure stability of probabilistic topic models over noise-free dataset and its corresponding noisy dataset. A high agreement score indicates an appropriate selection of topic model which is robust against textual noise.

Proceedings of the 2nd Workshop on Noisy User-generated Text,
pages 85–93, Osaka, Japan, December 11 2016.

2 Topic Modelling and Metrics

Topic models aim to discover the latent themes or topics within a corpus of documents, which can be derived by studying the distribution of co-occurrences of words across the documents. Popular approaches for topic modeling have involved the application of probabilistic algorithms such as Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003). For the evaluation of topic models, we follow the approach by Greene et al. (2014) for measuring topic model agreement. We can denote a topic list as $S = \{R_1, ..., R_k\}$, where R_i is a topic with rank i. An individual topic can be described as $R = \{T_1, ..., T_m\}$, where T_l is a term with rank l belong to the topic. Jaccard index (Jaccard, 1912) compares the number of identical items in two sets, but it neglects ranking order. Average Jaccard (AJ) similarity is a top-weighted version of the Jaccard index used to accommodate ranking information. AJ calculates the average of the Jaccard scores between every pair of subsets in two lists. Based on AJ, we can evaluate the agreement of two sets of ranked lists (topic models). The topic model agreement score between S_1 and S_2 is a mean value of the top similarity scores between each cross pair of R. The agreement score is solved using the Hungarian method (Kuhn, 1955) and is constrained in the range [0,1], where a perfect match between two identical k-way ranked sets results in 1 and a score 0 for non-overlapping sets (Greene et al., 2014).

3 Datasets

In this paper, we make use of two previously-studied text datasets which differ in terms of content and documents lengths. The *bbc* corpus includes 2225 general news articles assigned to five ground truth topics. The *wikilow* corpus is a subset of 4986 Wikipedia articles, where the articles are labeled with 10 fine-grained WikiProject sub-categories. In both datasets the topics consist of distinct vocabularies which we would expect LDA to detect. For example, the topics in *bbc* datasets are *business*, *entertainment*, *politics*, *sport*, and *technology*.

Table 1: Corpora used in the experiments, including the total number of documents n, terms m, and the number of labels k in the associated "ground truth".

Corpus	n	m	$\hat{k}$	Description
bbc	2,225	3,121	5	General news articles from the BBC (Greene and Cunningham, 2005).
wikipedia-low	4,986	15,441	10	A subset of Wikipedia. Articles are labeled with fine-grained WikiProject sub-groups (Greene et al., 2014).

3.1 Simulated Corpus with Word-level Noise

We artificially introduce noise into the above datasets to approximate the performance of topic modelling over naturally noisy sources. We measure noise using word error rate (WER), a common metric for measuring speech recognition accuracy. Moreover, WER has been used as a salient metric in speech quality analytics (Saon et al., 2006) and spoken dialogue system (Cavazza, 2001). WER is defined as the fraction between the sum of the number of substitutions S, the number of deletions D, the number of insertions and the number of terms in reference N:

$$WER = \frac{S + D + I}{N} \tag{1}$$

The experiments investigate the robustness of topic models against each type of noise, and at which noise levels the output of a topic model is consistent with that of the original corpus. *Deletion* noise is introduced by randomly removing a portion of text in the corpus. The proportion of deletion ranges from 0% to 50% and the term selection is based on uniform distribution. *Insertion* is introduced by adding a portion (0% to 50%) of frequent terms from a list of frequent English words with 7726 entries[1]. The probability of sampling of a certain term from the list is based on the term frequency.

[1] http://ucrel.lancs.ac.uk/bncfreq/flists.html

3.1.1 Metaphone Replacement

We simulate speech recognition errors using Metaphone, a phonetic algorithm for indexing English words by their pronunciation (Philips, 1990). Here we use the Double Metaphone (Black, 2014) algorithm in replacement and the replacement is on a one-to-one basis. This may not simulate the full range of errors produced by ASR systems, in which the substitution may be a one-to-many or many-to-one mapping, but it was deemed sufficient for the current experiments.

Table 2: Double metaphone dictionary where terms are ranked with descending frequencies.

Metaphone	Matching terms
ANTS	**industry**, units, induced, wound, ...
KRTF	grateful, **creative**, Cardiff, ...

Table 3: An example of 30% metaphone replacement in the *bbc* dataset.

Original text	Replaced text
We **are** hoping to understand the **creative industry** that has a **natural** thirst for **broadband** technology	We **air** hoping to understand the **Cardiff induced** that has a **neutralist** thirst for **portable** technology

In this study we map Metaphone codes to frequent English words (see examples in Table 2). Then in a given text document, we randomly select X percent terms and replace each by a term in the Metaphone map. The candidate terms sharing the same metaphone symbol are selected based on term frequencies. A frequent term has higher probability to be selected over a rare term (see Table 3).

3.2 Simulated Corpus with Image-level Noise

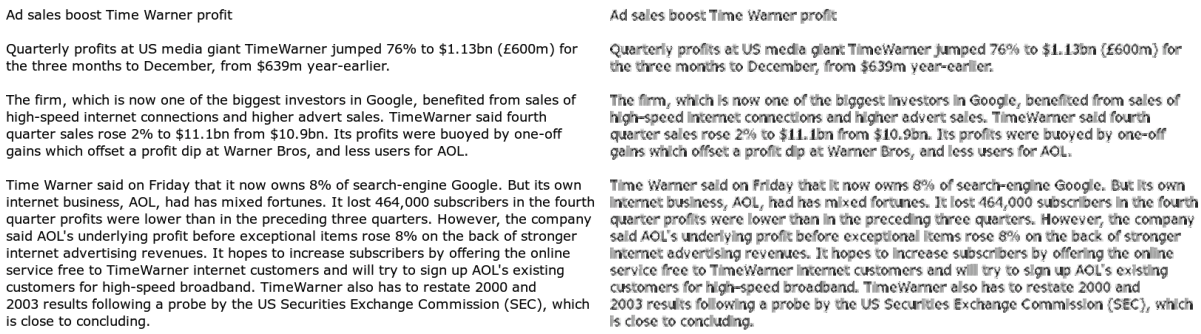
In the last section we simulatde word-level noise with deletion, insertion and metaphone replacement errors. As an alternative, we now introduce image-level random noise and simulate term errors from OCR. *ImageMagick*[2] is a popular software tool which can apply transformations and filters to images. *Tesseract*[3] is a widely-used open source OCR engine. We combine ImageMagick and Tesseract in sequence to generate noisy document images and then re-scan the content from those images. The level of image noise can be specified as a parameter to ImageMagick. The number of error terms from OCR is assumed to be proportional to the level of image noise that has been added. However, we observe that the robustness of OCR is highly influenced by the choice of text font. For instance, Tesseract scans of text in a monospace (fixed-width) font tend to be identical across a range of increasing noise until a certain level is reached, above which almost no characters are retrieved. Therefore, we opt to use variable-width font *Verdana* in this study, as its noise resistance performance is close to linear.

Figure 1 shows two snippets of ImageMagick generated images with noise level 1 and 3. The source texts are from the same document from the *bbc* corpus. The text in the image with level 1 noise remains clearly legibile. However it is rather difficult to read the content from the image with level 3 noise. Figure 2 shows texts scanned by Tesseract from the images of Figure 1. We can observe systematic errors on both images. On Figure 2 (a), `sales` is translated as `saies`, and `Google` is translated as `Googie`. On Figure 2 (b), `which` is retrieved as `wildi`, and `the` is retrieved as `die`. The word `Time` is read as `1Ine`, `Tine` or `The`. These are only part of the errors from OCR output, but they show both one-to-one mapping and one-to-many mapping.

Figure 3 shows the Recall (i.e. how many instances in the scanned texts match the instances of words in original texts) of Tesseract OCR against original texts in *bbc* and *wikilow* corpora. The recall score at each noise level is a mean value of recall from each pair of documents (one original document and its corresponding scanned text). In both corpora the highest recall value is observed from noise level 2.

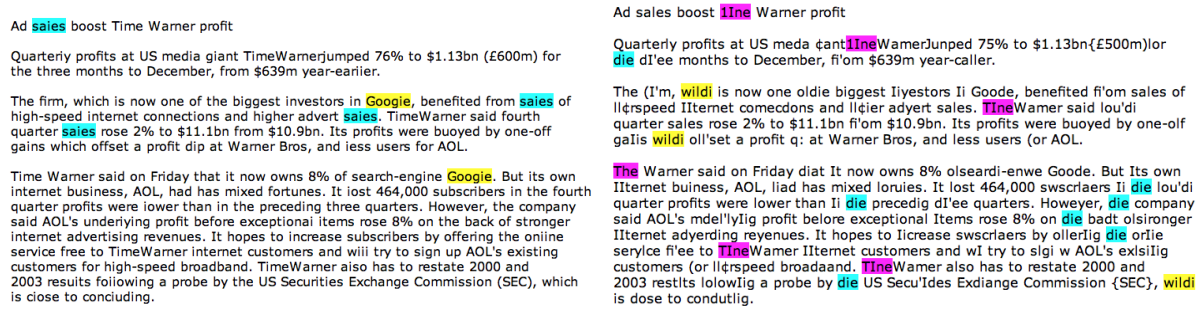
[2] `http://www.imagemagick.org`

[3] `https://github.com/tesseract-ocr`



(a) noise level 1 (b) noise level 3

Figure 1: Snippets of ImageMagick generated images with noise level 1 (a) and level 3 (b).



(a) noise level 1 (b) noise level 3

Figure 2: Snippets of Tesseract scanned texts with noise level 1 (a) and level 3 (b).

Therefore we drop the assumption that the error rate of scan is linearly scalable to noise levels. Another observation is that recall scores in both corpora are close to each other on each level of image noise, although the contents as well as scan errors of two corpora are quite different.

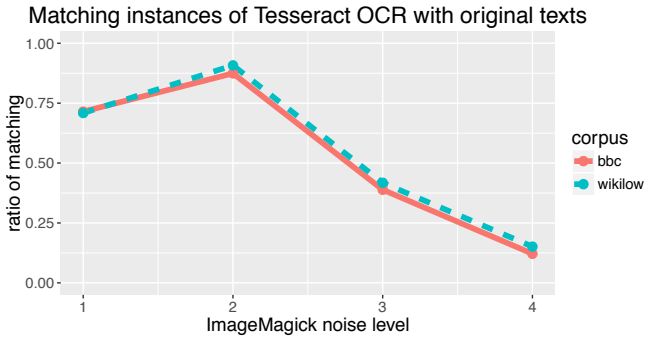
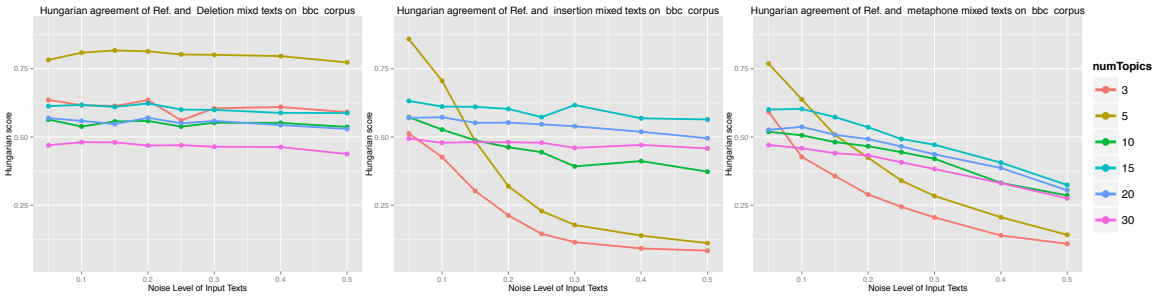


Figure 3: Recall of Tesseract OCR against original texts in *bbc* and *wikilow* corpora

4 Experiments

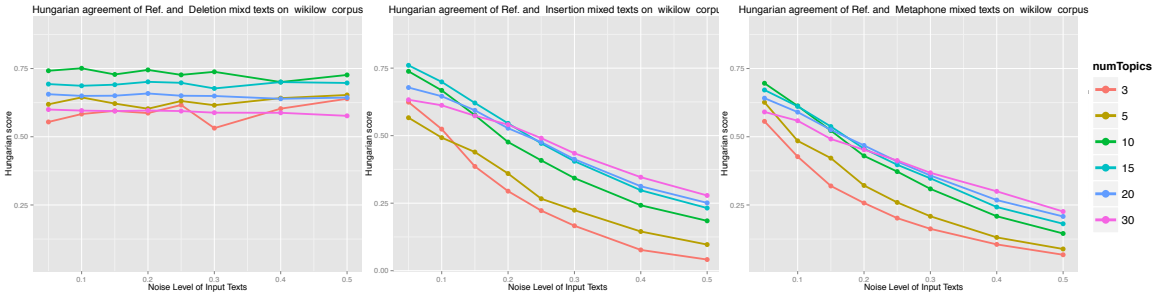
In our experiments with LDA, we aim to test topic stability over different levels of noise and different numbers of topics. In other words, when there is noise added to a text corpus, would the generated topic models be consistent with those from original corpus? If not, how much discrepancy exists between the two? Two sets of experiments are designed to introduce word-level noise and image-level noise to the original data. Our aim is to find guidelines for topic modelling over real noisy text sources in practice.

Since average Jaccard (AJ) score measures the similarity between two ranked lists (topics), we use AJ as an element in evaluating the similarity of two topic models S_x and S_y. If S_x and S_y each contains



(a) Deletion errors (b) Insertion errors (c) Metaphone errors

Figure 4: LDA Hungarian scores against noise levels for the *bbc* corpus (5 reference topics).



(a) Deletion errors (b) Insertion errors (c) Metaphone errors

Figure 5: LDA Hungarian scores against noise levels for the *wikilow* corpus (10 reference topics).

k ranked lists (topics), they are denoted as $S_x = \{R_{x1}, ..., R_{xk}\}$ and $S_y = \{R_{y1}, ..., R_{yk}\}$. We build a $k \times k$ matrix $\mathbf{M}$ to register each similarity score. In this matrix entry M_{ij} indicates the AJ score between topic R_{xi} and topic R_{yj}. The best matching between the rows and columns of $\mathbf{M}$ can be found using the Hungarian method (Kuhn, 1955). The agreement score produced from $\mathbf{M}$ is named as Hungarian agreement score, and is used as a measure of topic model stability in the following experiments.

4.1 LDA with Word-level Noise

In order to produce consistent and repeatable results where each noise generation method relies on a degree of randomness with word deletion, insertion or substitution we generate multiple copies of each modified corpus using 5 random seeds. Similarly we perform 5 iterations of each Mallet LDA (McCallum, 2002) topic model as the algorithm initial state is determined by a random seed. LDA hyperparameters are defined with default values, and each topic is represented by the top 25 terms. The final stability score on each level is a mean value of a number of iterations with fixed seeds.

Figure 4 and Figure 5 show the topic stability achieved by LDA on the bbc and $wikilow$ corpora, which have 5 and 10 reference topics respectively. For each level of topic model complexity, a downward slope indicates decreasing stability of topic models against increasing noise.

For the *bbc* corpus, the stability measure shows clear differences with different types of noise. The model is especially robust against *Deletion* errors. When noise increases from 5% to 50%, the Hungarian agreement score of output topics only drops about 1% (for the fitted model with $K = 5$ in Figure 4(a)). By inspecting each model in Figure 4(a), we can say that the topic models are robust against random *Deletion* noise in the case of the *bbc* corpus.

In Figure 4(b), the model with 5 topics achieves the highest Hungarian agreement score at noise level 5% and 10%, but it drops significantly afterwards. The best and most stable topic model with noise higher than 15% of *Insertion* errors is the model with 15 topics, which is three times of the reference. Similar trend is observed with *Metaphone* replacement errors in Figure 4(c). The topic model with reference number of topics achieves the highest stability when noise level is low. However, there are differences between Insertion and Metaphone errors in *bbc* corpus tests. With 50% of *Insertion* errors, the model

with 15 topics achieves 56.4% agreement with original model, but the agreement is only 32.4% with *Metaphone* errors. In *bbc* corpus *Metaphone* errors are the most challenging case.

For the *wikilow* corpus, we observe similar trends in Figure 4 and Figure 5 on specific types of noise. With *Deletion* errors, the topic model with the reference number of topics is most stable across noise levels. The difference in topic agreement scores is below 2% across the noise levels. With *Insertion* and *Metaphone* errors, the topic model with reference number of topics is almost the best when noise is low, but drops below others when noise is higher than 15%.

Although there are many similarities between Figure 4 and 5, we observe two major differences across corpora. In Figure 5(b) and 5(c), Hungarian scores of different topic models (number of topics) share similar gradient of descending slope. However, a few models for the *bbc* corpus ($K = \{15, 20, 30\}$) achieve quite stable Hungarian scores in Figure *bbc*(b). Another difference is that the most stable topic models against noise levels higher than 20% in Figure 4(b) and 4(c) both have 15 topics, whereas the most stable models in *wikilow* have 30 topics in the same settings. However, if we compare them with corresponding reference topic numbers K, the most stable topic models with high systematic errors all have $K * 3$ topics. Models with topic number higher than $K * 3$ are not optimal in Figure 4(b) and 4(c).

4.1.1 Discussion

In Section 4.1 we considered topic model stability on two corpora with three types of noise. Here we can define a single measurement of topic stability across different settings. If a level of agreement is set as 70%, LDA is robust against Deletion noise up to 50% in both *bbc* and *wikilow* corpora. However, LDA model reaches this agreement level only on 10% Insertion noise and on 5% Metaphone replacement noise. We see that Metaphone replacement and insertion are more severe challenges to topic models vs. deletion.

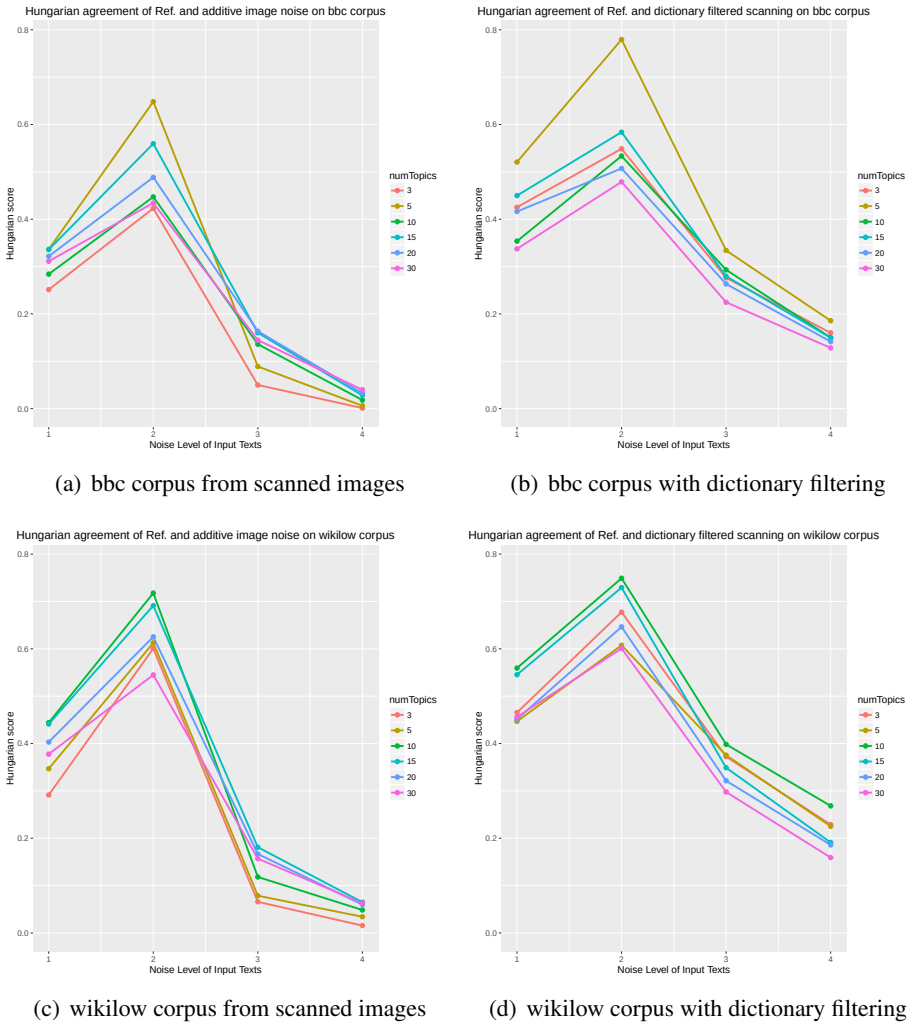
Regarding deletion errors, we observe that the robustness of a topic model is mostly determined by the number of topics. When this matches the number of reference topics, the topic model is the most stable. However, this does not emerge with insertion and metaphone errors. Reference topic models with 5 (bbc) and 10 (wikilow) topics achieve high stability only when noise is $\leq 10\%$. With higher levels of noise, a more complex topic model exhibits higher stability.

An explanation of the high stability of topic models against *Deletion* error concerns the LDA model. LDA takes term frequency into account. The probability of a word belonging to a topic is high if it appears frequently in one topic and seldom in other topics. Such a word is very likely to be an entry in a topic model. If we randomly delete corpus terms, the scale of frequent terms is influenced trivially and these frequent terms still have a high probability of selection. All rare terms may be removed by deletion, but they have a low chance of appearing in the original topic model anyway. Therefore LDA model has high stability over various levels of deletion errors. *Insertion* and *Metaphone* replacement introduces systematic noise, which changes the distribution of original texts with respect to frequency, thus having more impact on the LDA model. A high portion of general frequent terms may dilute the frequency of characteristic terms and add noisy terms to a topic model. However, a topic model with many more topics than the reference can deal with the effect of systematic errors.

4.2 LDA with Image-level Noise

In Section 3.2 we introduced the concept of image-level random noise. By using this approach, the Tesseract scanned texts contain character-level noise and many of them are systematic errors. For example, many instances of character `l` are retrieved as `i` in Figure 2. In this section we evaluate the stability of LDA topic models over corpora with substantial systematic errors. Mallet LDA topic model is evaluated over the scanned corpora of noise levels 1 to 4. On each noise level, topic model stability is evaluated with a selection of model complexity (the number of topics ranges from 3 to 30). Since Dirichlet prior of a LDA model is set with a random seed, we test one LDA model three times with 3 fixed seeds and get model stability score from the mean value of three iterations. The number of terms in each topic is 25.

Figure 6(a) and 6(c) show Hungarian agreement scores for LDA topic models, as achieved on the original corpus and a corresponding polluted corpus with image-level noises. Figure 6(a) is from the *bbc*



(a) bbc corpus from scanned images

(b) bbc corpus with dictionary filtering

(c) wikilow corpus from scanned images

(d) wikilow corpus with dictionary filtering

Figure 6: LDA Hungarian scores over scanned corpora (a) and (c), and dictionary filtered corpora (b) and (d). In reference dataset, *bbc* corpus has 5 topics and *wikilow* has 10 topics.

corpus and Figure 6(c) is from the *wikilow* corpus. In these two figures Hungarian agreement score drops with increasing image-level noises (noise 2 to 4), but rises from noise 1 to 2. As a comparison, curves in Figure 3 follow similar trends. The stability of LDA model is influenced by character-level noises in the scanned corpus.

For the *bbc* corpus (Figure 6(a)) a model with 5 topics achieves (almost) the highest Hungarian agreement score at noise level 1 and 2, but it drops considerably at noise level 3 and 4. Models with 20 and 30 topics reach the highest Hungarian agreement score at noise level 3 and 4 respectively. For the *wikilow* corpus (Figure 6(c)), the model with 10 topics achieves (almost) the highest scores across noise level 1 to 2, but it drops near to the bottom at noise level 3 and 4. Choosing 15 topics yields the highest Hungarian agreement score at noise level 3 and 4.

4.2.1 Filtering Noisy Corpus with English Dictionary

OCR generates character-level outputs and therefore the scanned corpora contain erroneous terms which are out of English vocabulary (e.g., `saies`). We propose a filtering step over the OCR scanned corpus before the topic modelling step. A general English dictionary with 110k entries is used as a reference set to filter off any undefined terms.

LDA topic models are tested on the dictionary filtered corpora following the same steps used in Section 4.2, and the Hungarian agreement scores are plotted in Figure 6(b) and 6(d). At the first look, we observe similar curves as experiments with unfiltered corpus (Figure 6(a) and 6(c)). However, there are important

differences. In Figure 6(b) topic number 5 reaches the highest agreement scores across noise level 1 to 4. In Figure 6(d) topic number 10 reaches the highest agreement scores across noise level 1 to 4. In both corpora, the topic models with the reference number of topics are most stable with the dictionary filtered texts. On each noise level, if we compare the highest agreement scores across two figures side by side, we can see that dictionary filtering brings an increase of model stability over 10% in most cases and it can be as high as 20% on high level noises.

4.2.2 Discussion

We evaluate the stability of topic models over scanned corpus of noisy images, and we find common observations as the experiments in Section 4.1. In general, the stability of LDA models drops with increasing simulated *insertion* and *Metaphone* replacement errors because these errors are not random terms. More consistent systematic errors (e.g., Figure 2) give higher challenge to topic models. Under these challenges, a topic model with reference number of topics performs best at relatively low noise levels. When noise is as large as 4, LDA models with 30 topics (*bbc*) or 15 topics (*wikilow*) are more stable in both corpora.

When we apply dictionary filtering to scanned texts, LDA stability improves considerably. Moreover, the stability of topic models for the reference number of topics is still highest, despite significant levels of noise. There is no more need to increase topic model complexity for better performance. Although there may be a risk that we may remove terms which correspond to names or specialised terminology, this issue could be overcome by augmenting the dictionary with lists of people, organisations, and domain-specific lexicons.

5 Conclusions

In this paper we investigated how transcription errors affect the quality and robustness of topic models produced over a range of corpora, using a topic stability measure introduced *a priori*. We simulated word-level transcription errors from the perspective of word error rate and generated noisy corpora with *deletion*, *insertion* and *Metaphone* replacement. Topic models produced by LDA showed high tolerance to deletion noise up to a reasonably high level, but low tolerance to insertion and metaphone replacement errors. We also simulated image-level OCR transcription errors which are close to real OCR applications.

In general, we found that the robustness of topic models is mainly determined by the extent of systematic errors which modifies the distribution of words in the original texts. With relatively low levels of systematic errors, topic models with reference number of topics were most stable. However, in the case of high level systematic errors, topic models with many more redundant errors were stable.

In order to increase topic model stability when analysing a noisy corpus, we introduced English dictionary-based filtering. Dictionary filtering was shown to be effective in correcting systematic errors across different noise levels, with the result that topics generated from collections of noisy documents were still informative. In future experiments, we aim to explore more approaches in improving topic model stability over noisy sources.

References

Shai Ben-David, Dávid Pál, and Hans Ulrich Simon, 2007. *Stability of k-Means Clustering*, pages 20–34. Springer Berlin Heidelberg, Berlin, Heidelberg.

Paul E. Black. 2014. "double metaphone", in dictionary of algorithms and data structures [online]. `https://xlinux.nist.gov/dads/HTML/doubleMetaphone.html`, May 2014.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Marc Cavazza. 2001. An empirical study of speech recognition errors in a task-oriented dialogue system. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan G. Fiscus, Jerome Ajot, and John S. Garofolo. 2007. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, pages 373–389.

Derek Greene and Pádraig Cunningham. 2005. Producing accurate interpretable clusters from high-dimensional data. In A. Jorge, L. Torgo, P. Brazdi, R. Camacho, and J. Gama, editors, *Proc. 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, pages 486–494. Springer, October.

D. Greene, D. O'Callaghan, and P. Cunningham. 2014. How Many Topics? Stability Analysis for Topic Models. In *Proc. European Conference on Machine Learning (ECML'14)*.

Paul Jaccard. 1912. The distribution of flora in the alpine zone. *New Phytologist*, 11(2):37–50.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. 2004. Stability-based validation of clustering solutions. *Neural Comput.*, 16(6):1299–1323, June.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Analyzing entities and topics in news articles using statistical topic models. In *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*, ISI'06, pages 93–104, Berlin, Heidelberg. Springer-Verlag.

David S. Pallet. 2003. A look at NIST's benchmark asr tests: Past, present, and future. Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language*, 7(12 (December)):39.

G. Saon, B. Ramabhadran, and G. Zweig. 2006. On the effect of word error rate on automated quality monitoring. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 106–109, Dec.

Daisuke Yokomoto, Kensaku Makita, Hiroko Suzuki, Daichi Koike, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara. 2012. Lda-based topic modeling in labeling blog posts with wikipedia entries. In *Proceedings of the 14th International Conference on Web Technologies and Applications*, APWeb'12, pages 114–124, Berlin, Heidelberg. Springer-Verlag.