Context-Sensitive Error Correction: Using Topic Models to Improve OCR

Michael L. Wick Michael G. Ross Erik G. Learned-Miller
University of Massachusetts Amherst
Computer Science and Psychology Departments
Amherst, MA, USA
mwick@cs.umass.edu, mgross@psych.umass.edu, elm@cs.umass.edu

Abstract

Modern optical character recognition software relies on human interaction to correct misrecognized characters. Even though the software often reliably identifies low-confidence output, the simple language and vocabulary models employed are insufficient to automatically correct mistakes. This paper demonstrates that topic models, which automatically detect and represent an article's semantic context, reduces error by 7% over a global word distribution in a simulated OCR correction task. Detecting and leveraging context in this manner is an important step towards improving OCR.

1. Introduction

As researchers and the general public become more reliant on computer-searchable document databases, paper documents that have not been translated into computer strings are in grave danger of being forgotten [1]. Optical character recognition (OCR) software has made great strides over the past few decades, but the translation of documents into searchable strings still requires that humans manually proofread and correct the output. This paper presents a new algorithm for automatically correcting errors in OCR output. By detecting the semantic context of OCRed documents, our algorithm can use topic-specific word frequency information to correct corrupted words.

While there has been much focus on improving the accuracy of OCR by incorporating language models to guide error detection and correction, these models are typically global and treat each document equivalently, even though vocabulary usage varies between documents. For example, the distribution of words in Car & Driver differs from the distribution of words in the New England Journal of Medicine. Because these two publications use jargons derived from separate domains, using word frequencies observed in one periodical to correct OCR results from the other could be disastrous.

Imagine proof-reading the result of a document extracted by OCR and encountering the string "tonque". Given no contextual evidence, it is reasonable to believe that the software might commonly mistake the letter 'q' for 'g', and that the actual word should be "tongue". However, once we learn that the article is about sports cars, we may want to change our beliefs; perhaps the word is more likely to be "torque" than "tongue". A major problem with a global language model is its inability to adapt to the idiosyncrasies of particular domains.

One possible solution is to create many independent topic-specific vocabulary models, but that imposes high training costs and requires end users to semantically classify every article prior to OCR. Additionally, it does not solve the problem of OCRing documents that contain multiple categories. A more promising solution should minimize human involvement by automatically deducing all categories present in each document.

In the fields of social network analysis and document corpus modeling, these questions are often addressed with *topic models*. Topic models can automatically describe a document as a mixture of semantic topics, each with an independent vocabulary distribution. These models can be trained in an unsupervised manner, and can dynamically determine the context of new documents without user input. The utility they bring to language modeling offers the prospect of improved OCR results and reduced reliance on human error correction.

This paper describes the use of a topic model to correct simulated OCR output and demonstrates that it outperforms a global word probability model across a substantial data set. This use of contextual modeling is the first step towards a number of promising new techniques in document processing.

2. Related Work

Topic models [8] come in a number of varieties. This work uses Latent Dirichlet Allocation (LDA) developed by Blei et al. [2]. LDA is a generative model that represents each document as a "bag of words" in which word order is discarded and only word frequencies are modeled. A corpus is represented by a Dirichlet distribution that indicates the probabilities of different topic mixtures. Under such a model, a new document is generated by first selecting a topic mixture — for example, the document might be 80% about music, 10% about computers, and 10% about politics. This defines a document-specific multinomial distribution. To generate individual words, repeatedly draw a topic from this distribution and then sample from the multinomial word probability distribution associated with that topic.

LDA models can be learned in an unsupervised fashion from unlabeled document collections and later exploited to infer the topics present in a novel document. No user input is required, a crucial difference between these techniques and those used by Strohmaier et al. [9] to correct OCR output with topic-specific dictionaries. Furthermore, LDA allows a document to contain any mixture of topics, avoiding the need to artificially divide articles into fixed categories. Finally, these models have been successful in many areas. For example, Wei and Croft [10] have demonstrated that useful LDA models can be built from large corpora.

There have been many previous efforts to use language models to improve OCR results. Zhang and Chang [11] post-processed OCR output with a linear combination of language models to correct errors. Hull [4] used a hidden Markov model to incorporate syntactic information into character recognition.

3. Topic Modeling for Error Correction

3.1 Model construction

The error correction algorithm consists of two models: a topic model that provides information about word probabilities and an OCR model that represents the probability of character errors.

The LDA topic model is trained from a collection of unlabeled documents using Andrew McCallum's MALLET software [7]. We assume that these documents are free of OCR errors, and the output of the training is two sets of probability distributions: the Dirichlet prior over topic mixtures and a set of per-topic multinomial word distributions (as discussed in Section 2). During the error correction process these distributions will be used to detect the topic mixture present in each OCR document, which will consequently enable estimation of the relative probabilities of possible word corrections.

The OCR model represents the probability of different character corruptions in the documents. It is clear that some corruptions are much more likely than others — for example, OCR software is more likely to mistake 'i' for 'j' than to confuse 'x' and 'l'. Therefore the OCR model is non-uniform. We expect OCR to produce the correct result on most character instances, so the probability of a correct recognition is relatively high. The notation $P(l^f|l^s)$ designates the probability that the OCR software generates letter l^f given that the truth is letter l^s. This model is used both to generate simulated OCR output for testing purposes and as part of the correction process. Statistics from actual character recognition output could be used to construct an OCR model that would enable our method to be used as a post-processor for real-world OCR software.

3.2 Error-correction algorithm

The algorithm takes an OCR document and a list of its incorrect words. Currently, the incorrect word list is provided by an oracle, but many OCR packages are capable of indicating low certainty words to their users.

For each incorrect word w_i in the document, we generate a list of all strings that differ from w_i by zero, one, or two characters. Due to the combinatorial explosion of this method, we do not consider words that are three or more characters apart from the original string. For each word w_c in this candidate list, we assign a score based on the particular model that is used and the letters that are flipped. For the simple global frequency approach, this combines the OCR model and the probability of the candidate word into

$$\text{Score}(w_c) = P(w_c) \prod_{j}^{N} P(l_j^f|l_j^s)$$

where $P(w_c)$ is the probability of the word, N is the number of letters in the word, and $P(l_j^f|l_j^s)$ is the probability that letter l_j^s was mistaken for l_j^f. For a topic model, the probability of a word is

$$P(w) = \sum_{k}^{M} P(w|t_k)P(t_k)$$

where w is a word, M is the number of topics in the model, and t_k is a topic. $P(t_k)$ is computed by applying the trained topic model to the correctly recognized words in the document.

After the scores of all candidates are computed, the word is corrected by substituting the highest-scoring candidate. Ties are broken randomly and corrections only occur if the selected string scores strictly higher than the original.

Newsgroups				
Models	**2**	**4**	**6**	**8**
Global	67.2	63.9	65.2	64.2
30 Topics	69.6	65.8	67.6	65.4

Table 1. Error correction accuracy for global and topic models on multi-domain newsgroup data

Most common words in top topics				
10	22	8	11	2
car	science	writes	post	posting
cars	writes	people	judas	nntp
engine	article	article	death	host
drive	objective	mark	center	message
oil	values	read	policy	idea

Figure 1. These are the five most common words in the five most probable topics for the example rec.autos document. Note that the words in most of the topics are related — topic 10 is clearly the "car" topic for example.

4. Experiments

4.1 Data

For our experiments we use the publicly accessible 20 Newsgroups data corpus available at http://people.csail.mit.edu/jrennie/20Newsgroups/. This data set is well suited for our experiments as it contains documents from various domains. For the experiments, we used documents from the alt.atheism (480 documents), comp.graphics (588 documents), sci.space (594 documents), talks.politics.guns (549 documents), talks.politics.mideast (569 documents), talks.politics.misc (467 documents), rec.autos (595 documents), and religion.misc (377 documents) newsgroups.

We tested our system on corpora containing two (comp.graphics and talk.politics.mideast), four (adding sci.space and talk.politics.guns), six (adding alt.atheism and talk.politics.misc) and eight (adding talk.religion.misc and rec.autos) newsgroups. In each case, the documents were randomly divided, setting aside 100 testing documents and using the remainder for training. The testing documents were corrupted by the OCR error model described previously and lists of the corrupted words in each document were provided to the correction algorithm.

The same model parameters were used throughout the experiments to demonstrate that no extensive parameter tuning is necessary for this method. The number of topics was fixed to 30 — even though we never test on 30 newsgroups, each newsgroup might cover several distinct, although related, topics.

Using the algorithm described previously, we evaluated two word models: a global word frequency model and an LDA topic model. The only difference between the models is in the calculation of $P(w_c)$. The global model used the same multinomial distribution for every correction of every document, while the topic model used the correctly recognized words to determine the topic probabilities and adapt $P(w_c)$ to the local context.

4.2 Results

Table 1 displays the error correction results for both global and topic-based language models while varying the number of newsgroups the documents are drawn from. The topic model outperforms the global model for every tested combination of newsgroups, reducing error by an average of 7%.

An example from the rec.autos newsgroup demonstrates how the topic model enables this improvement in error correction. It is possible to qualitatively understand the topics in the model by looking at the most probable words under each one's distribution. In Figure 1, we see several of the most probable topics given the correct words in a particular rec.autos document. Clearly, topic 10 contains words related to cars, while the other topics seem to relate to other subjects such as science or religion.

Figure 2 shows the probabilities of each of the Figure 1 topics given the rec.autos document. Topic 10, the "cars" topic, clearly dominates this distribution. In Figure 3, we see that the topic model was able to correct several corrupt car-related strings while the global model made incorrect substitutions indicating that this success was the result of the document-specific contextual information provided by the topic model.

5. Conclusion and Future Work

We developed an algorithm for applying topic modeling to OCR error correction. This model outperformed a global word distribution on the error correction task on simulated data due to its ability to determine the context of each document and provide a tailored word probability model. Additionally, our method is automatic and does not require additional involvement from the OCR's operator.

The initial success of using topic models to correct simulated OCR output points to a number of exciting avenues

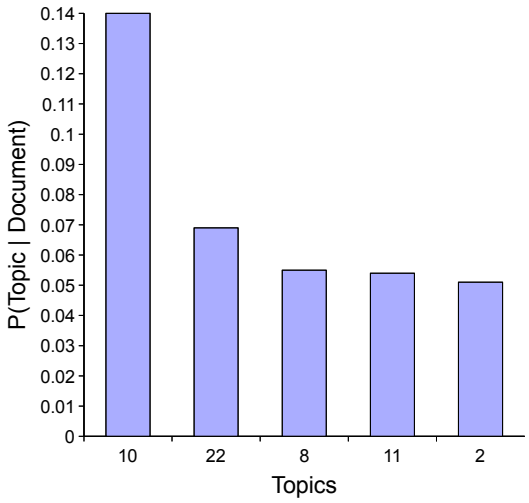

Figure 2. The probability distribution of topics conditioned on the correctly recognized words from the rec.autos example document. Notice that topic 10, the "cars" topic (see Figure 1) is much more probable than any other.

for future work. Applying it as a post-processor to real OCR output will allow us to further validate the approach, as will the collection of larger data sets. We expect that the model's advantages over a global word frequency model will increase with the diversity of the test and training corpora.

Additionally this problem provides an excellent framework for testing advances in topic modeling. Often researchers provide lists of topic words to demonstrate their success, but tasks such as OCR correction could be an objective metric of success.

The topic model approach to OCR correction relies on the first OCR pass identifying some words with high confidence, which enables the model to infer an appropriate topic distribution and, in turn, correct poorly recognized words. It is clear that this process can be easily iterated — the highest confidence corrections can be appended to the recognized word list in each document and the topics can be re-estimated. A better topic distribution should allow additional words to be corrected with high confidence and similarly used in the next round. Also, instead of being used as post-processing step, the topic model probabilities could be integrated with the image processing information and font models already used by OCR software for maximum effectiveness. Additionally, more sophisticated topic modeling schemes such as heirarchical LDA [3] or Pachinko allocation machines (PAM) [6, 5] that nonparametrically adapt to an arbitrary number of topics and relax independence assumptions between them, could potentially contribute to further improvements on OCR correction.

Example corrections

Corrupted word	Global	Topic-model
notor	color	motor
snaw	shaw	snow
deater	center	dealer

Figure 3. These example corrections from the rec.autos sample document show that the topic model provides contextual information that enables it to outperform the global word model.

Topic modeling can also be made practical without an error-free training set of digital documents. Many archival OCR projects involve converting back issues of academic journals so they can be useful for future researchers. Some of these journals are in old fonts or printed on decaying paper stock, so OCR software would only recognize a few words with high confidence. Due to evolutions in vocabulary, there might be very few or no equivalent digital documents for use in topic model training.

However, with a large enough collection of related documents, an initial topic model could be formed from the relatively few words that are confidently recognized. This initial model might allow for high confidence in more words on a second pass, which would in turn lead to a more detailed topic model. Thus a topic model could be bootstrapped from a weak OCR algorithm and result in a strong OCR algorithm for difficult documents.

This iterative style is part of the general *iterative contextual modeling* (ICM) approach to OCR. We believe that ICM can provide a framework for leveraging not only language but also appearance context to advance to new levels of performance on challenging documents.

6 Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0427594, and in part by U.S. Government contract #NBCH040171 through a subcontract with BBNT Solutions LLC. E. Learned-Miller was supported under NSF CAREER award #IIS-0546666. We would also like to thank Andrew McCallum for useful discussion and the use of his MALLET toolkit. Any opinions, findings and conclusions

or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] H. Baird. Digital libraries and document image analysis. In *International Conference on Document Analysis and Recognition*, 2003.

[2] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[3] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems*, 2004.

[4] J. Hull. Incorporating language syntax in visual text recognition with a statistical model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12), 1996.

[5] W. Li, D. M. Blei, and A. McCallum. Nonparametric Bayes Pachinko allocation. In *UAI*, 2007.

[6] W. Li and A. McCallum. Pachinko allocation: A directed acyclic graph for topic correlations. In *NIPS Workshop on Nonparametric Bayesian Methods*, 2005.

[7] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[8] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2006. In press.

[9] C. Strohmaier, C. Ringlstetter, K. Schulz, and S. Mihov. Lexical postcorrection of OCR-results: The web as a dynamic secondary dictionary? In *International Conference on Document Analysis and Recognition*, 2003.

[10] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of SIGIR06*, 2006.

[11] D. Zhang and S. Chang. A Bayesian framework for fusing multiple word knowledge models in videotext recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.