EXCEPTIONS IN LANGUAGE AS LEARNED BY THE MULTI-FACTOR SPARSE PLUS LOW-RANK LANGUAGE MODEL

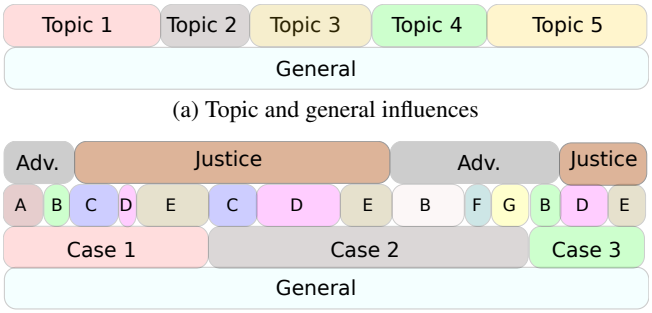
Brian Hutchinson, Mari Ostendorf and Maryam Fazel

Electrical Engineering Department, University of Washington

ABSTRACT

Word usage is influenced by diverse factors, including topic, genre and various speaker/author characteristics. To characterize these aspects of language, we introduce the "Multi-Factor Sparse Plus Low Rank" exponential language model, which allows supervised joint training of arbitrary overlapping factor-specific model components. This flexible architecture has the advantage of being highly interpretable. The elements of sparse parameter matrices can be viewed as factor-dependent corrections (e.g. topic- or speaker-dependent phenomena). In topic modeling experiments on conversational telephone speech, we obtain modest perplexity reductions over an n-gram baseline and demonstrate topic-dependent keyword extraction that leads to a 13% (absolute) improvement in precision over TF-IDF. We also show how keywords can be jointly learned for speakers, roles and topics in a study of Supreme Court oral arguments.

***Index Terms*—** Language modeling, sparse plus low rank decomposition, topic models, keyword extraction



(a) Topic and general influences

(b) Role, speaker, case and general influences

Fig. 1. Two examples of overlapping scopes of influence: topic in conversational telephone speech (a) and several factors in Supreme Court transcripts (b).

1. INTRODUCTION

The probabilities of word sequences in language are influenced by numerous factors, such as topic, genre, formality, as well as the role, intention and idiosyncrasies of the speaker/author. Furthermore, within a corpus, the scopes of these different influences will vary; for example, in a collection of newswire text the discussion of professional sports is likely to be concentrated in a subset of the documents. These scopes of influence can also be arbitrarily overlapping, as would be the case if you have several speakers/authors covering different sets of topics, or formal and informal examples of language in the form of both written and spoken documents. For example, in Fig. 1 (a), two influences are active at a given time: one topic-dependent and one topic-independent (language-wide). In Fig. 1 (b), the model is more complicated, as the speaker, the court case and the role of the speaker all augment language-wide factors in an overlapping fashion.

In most language models (LMs), different sources of variation are not explicitly accounted for. Instead, training data from different sources are combined in a mixture model, e.g. [1, 2], or via count merging, e.g. [3], or domain adaptation techniques are used to leverage a general language model in the context of limited in-domain training data [4, 5]. More recently, the impact of topic has been explored using non-parametric Bayesian models, e.g. [6, 7, 8], which use a Dirichlet (or other) prior in unsupervised learning of latent topic variables. In [9], a similar approach is used with latent variables for both topic and role. While most of this work has focused on unigram language models for computational reasons, n-gram variants of the non-parametric Bayesian topic model are described in [10]. Despite their clear appeal, non-parametric Bayesian approaches have not been widely adopted; they have a relatively high computational cost and their non-parametric nature makes them somewhat more difficult to interpret.

We propose an alternative approach for characterizing different sources of variation in language: a Multi-Factor Sparse + Low Rank (SLR) exponential language model. At the base of the model is a low rank component that, as in [11], induces continuous representations of words and histories to get a smooth model capturing general syntactic-semantic language behavior. Added to that in the parameter space are arbitrarily many factor-dependent sparse components, each specializing in some phenomenon (e.g. capturing the idiosyncrasies of a speaker or topic) which may overlap in different ways with other factors. By regularizing these components to be sparse, we emphasize the most salient differences and discourage overfitting. In this light, each of the factor-dependent components can be seen as an additive correction to a global model. The model provides a flexible framework for adaptation to a new domain: depending on the nature and the extent of the mismatch, some factors can be updated, some kept intact, and others thrown out entirely.

A key feature of our language model is its interpretability: the elements of the sparse factor-dependent components correspond to keywords that represent salient factor-dependent differences. Unlike past work leveraging topic in exponential models [5, 12], identifying topic-related n-gram keywords is a byproduct; no separate pre-processing step is used to find them. Further, topic characteristics can be learned jointly with other factors such as genre, speaker, or speaker role. With multiple factors accounted for, the keywords are more meaningful. We explore the use of the learned keywords for topic summarization, and show that they can be used to identify salient characteristics of speaker roles and idiosyncrasies of individual speakers. In contrast to [13], where a sparse plus low rank decomposition of word-document matrices was shown to be effective

at identifying document keywords, we need no stop word filtering and support arbitrarily many overlapping factors.

2. SPARSE + LOW RANK LM REVIEW

Here we briefly review the SLR language model before discussing extensions in Sec. 3. The SLR-LM extends the standard maximum entropy exponential language model by reparameterizing the model weights and using regularization to produce a novel weight structure. The conditional probability $P(x|h)$ (word given history) is defined to be

$$P(x|h) = \frac{\exp\left(\psi(x)^T(L+S)\phi(h)\right)}{\sum_{x'}\exp\left(\psi(x')^T(L+S)\phi(h)\right)}, \quad (1)$$

where $\psi(x) \in \mathbb{R}^{d_\psi}$ and $\phi(h) \in \mathbb{R}^{d_\phi}$ are feature functions that map words x and histories h, respectively, to some (typically sparse, high-dimensional) feature representation, such as n-gram indicators used here: unigram for x and orders $n-1$ and lower for h. $\phi(h)$ also has an always-on "0-length" feature, which allows the model to include unigram probabilities. Two matrices, L and S, form the weights of the model. In contrast to the standard maximum entropy model, where the probabilities are determined by a log-linear function of a feature vector and weight vector, the SLR-LM uses a log-bilinear function. As noted in [11], if $S = 0$ then the SLR-LM can be viewed as a convex, feature-based version of the model termed simply the "log-bilinear model" that uses an explicitly factored matrix with a particular structure [14].

The first component is a low rank matrix, L, which has the effect of inducing continuous low-dimensional representations of words and histories, and is effective at exploiting the similarities that exist between words and between histories. This interpretation can be seen by noting that if L has rank r, then

$$\begin{aligned}\psi(x)^T L\phi(h) &= \psi(x)^T U\Sigma V^T\phi(h) &(2)\\ &= \left(U^T\psi(x)\right)^T \Sigma\left(V^T\phi(h)\right) &(3)\\ &= \tilde{\psi}(x)^T\Sigma\tilde{\phi}(h) &(4)\end{aligned}$$

Here $U\Sigma V^T$ is the compact singular value decomposition of L, and U^T and V^T have the effect of projecting $\psi(x)$ and $\phi(h)$ down to continuous, r-dimensional representations. As with other continuous language models, this provides a very natural form of smoothing: words that function similarly will be projected near to each other in this low-dimensional representation, naturally pooling information between words to more robustly estimate probabilities.

While the low rank component is effective at identifying and exploiting similarities that exist in the data, not all sequential language fits into regular patterns. Some n-grams are more frequent than the component words suggest (e.g. "taco bell" is more frequent than one's knowledge of tacos and bells would predict). Other n-grams are less frequent than the individual words would suggest (e.g. despite the general similarity between "really" and "very," the n-gram "really much" is not commonly used like "very much"). Rather than burden the low rank component, which excels at finding similarities, with these exceptions, we add a sparse component, S, that provides corrections as needed. In [11], we learned several distinct types of exceptional n-grams in the sparse component, including names, topic n-grams, and common multiword expressions.

3. THE MULTI-FACTOR SPARSE + LOW RANK LM

In the new Multi-Factor SLR-LM, the monolithic sparse component is replaced with a variable number of factor-dependent sparse components. At any given point in the document, only a subset of all sparse components will be "active." In the example of Fig. 1 (a), each n-gram will be associated with a set of three matrices: the (general) low rank component L, the (general) sparse component S_0, and a topic-dependent sparse component S_t. The low rank L exists to capture topic-independent linguistic regularities as before; the general sparse S_0 captures topic-independent exceptions (e.g. genre artifacts like "yeah yeah" or common place names like "new york"); the topic-dependent sparse matrices S_t capture topical exceptions (e.g. "black lab" and "pure bred" for the topic "Pets").

$Adv.$	1	1							1	1	1	1		
$Justice$			1	1	1	1	1	1					1	1
A	1													
B		1							1			1		
C			1			1								
D				1			1						1	
E					1			1						1
F										1				
G											1			
$Case1$	1	1	1	1	1									
$Case2$						1	1	1	1	1	1			
$Case3$												1	1	1
$General$	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Fig. 2. A binary "scope" matrix K defining which sparse components (rows) are active in which segments of the document (columns). This key corresponds to the example in Fig. 1 (b).

Let $\mathcal{C}_i$ denote the set of components active at word token x_i in the document. We refer to the set of word tokens x_i (and corresponding histories h_i) that have the same set $\mathcal{C}_i$ of active components as a "segment"; for example, in Fig. 1 (a) there are five segments, while in Fig. 1 (b) there are fourteen. Let $\mathcal{C}^{(t)}$ denote the shared set $\mathcal{C}_i$ for all word tokens x_i in segment t; i.e., $\mathcal{C}_i = \mathcal{C}^{(t)}$ for all word tokens i in segment t. The sets $\mathcal{C}^{(t)}$ can be equivalently represented in a binary "scope" matrix, K. The rows of K correspond to the sparse components in the model, while the columns correspond to segments. Fig. 2 shows the scope matrix for the Supreme Court example of Fig. 1 (b), with sparse components for each speaker, for each case, as well as a sparse component for "justices" and one for "advocates." Then the set $\mathcal{C}^{(t)}$ is just the set of rows $\{j : K_{jt} = 1\}$.

The Multi-Factor LM thus consists of a general low rank L, a general sparse S_0, and C additional sparse components (e.g. the other rows in Fig. 2). The average log likelihood $\mathcal{L}$ of the full dataset $\mathcal{X}$ (with N word tokens) is

$$\mathcal{L}(\mathcal{X}; L, S_0, S_1, \ldots, S_C) = \frac{1}{N}\sum_{i=1}^{N}\log P_{\mathcal{C}_i}(x_i|h_i), \quad (5)$$

$$P_{\mathcal{C}_i}(x|h) = \frac{\exp\left(\psi(x)^T(L+\sum_{c\in\mathcal{C}_i}S_c)\phi(h)\right)}{\sum_{x'}\exp\left(\psi(x')^T(L+\sum_{c\in\mathcal{C}_i}S_c)\phi(h)\right)}. \quad (6)$$

Training involves solving a convex optimization problem:

$$\min_{L,S_0,\ldots,S_C} \left(\gamma_0\|L\|_* + \sum_{c=0}^{C}\gamma_{1c}\|S_c\|_1 + \frac{\gamma_2}{2}\|L+\sum_{c=0}^{C}S_c\|_F^2 - \mathcal{L}(\mathcal{X}; L, S_0, \ldots, S_C)\right). \quad (7)$$

The nuclear norm, $\|\cdot\|_*$, is the sum of singular values and is a convex relaxation of rank. We can solve this problem using a modified accelerated proximal gradient descent algorithm, a variant of

the algorithms in [11, 15]; the key difference is that sparsifying line searches are performed in parallel for all sparse components, instead of a single sparse component. Due to the fact that proximal operator for all of the sparse matrices decomposes over individual matrices, the same convergence guarantees apply. We assume that the different sparse components cover different subsets of the data; otherwise, the solution may not be unique.

The training algorithm requires computing the gradient of the smooth part of the objective (the last two terms) with respect to each problem variable. These gradients can be computed efficiently in one pass over the data. Let $\nabla_{A_t}\mathcal{L}$ denote the gradient of average log-likelihood with respect to the sum $A_t = L + \sum_{c\in\mathcal{C}^{(t)}} S_c$, then

$$\nabla_{A_t}\mathcal{L} = E_{\hat{P}(x,h)}[\psi(x)\phi(h)^T] - E_{P_{\mathcal{C}_t}}[\psi(x)\phi(h)^T] \quad (8)$$

where $\hat{P}$ is the unnormalized empirical joint distribution of words and histories occurring in segment t. ($\hat{P}$ sums to the number of words in segment t over the number of words in the corpus.) $P_{\mathcal{C}_t}$ is similarly unnormalized. Then, the gradients of the smooth part of the objective with respect to the sparse components, S_c, are simply

$$\nabla_{S_c} f_{\text{smooth}} = \sum_{\{t:K_{ct}=1\}} \nabla_{A_t}\mathcal{L} + \gamma_2(L + \sum_{j=0}^{C} S_j). \quad (9)$$

That is, one can do a single pass from $t = 1, \ldots, T$ and accumulate each of the $\nabla_{S_c} f_{\text{smooth}}$ along the way.

4. EXPERIMENTS AND DISCUSSION

We conducted several experiments to measure the Multi-Factor LM, including its performance in terms of perplexity with joint training and adaption scenarios and in terms of the quality of keywords learned in the sparse components. The first set of experiments uses the Fisher corpus of conversational telephone speech; the second set uses a corpus of Supreme Court transcripts.

4.1. Conversational Speech

The Fisher corpus consists of telephone conversations between strangers on 40 pre-assigned topics. We split (by conversation) each topic into training, development and test sets, yielding 5.5M word tokens of training data, 1.9M word tokens of development data, and 2.0M word tokens of test data. Our language model vocabulary consists of the most frequent 9.7K word tokens appearing in the training set (all out-of-vocabulary words are mapped to a dedicated OOV token). Due to our motivation to analyze the exceptions learned, we restrict ourselves to bigram language models in all experiments, which are sufficient for most topical keywords.

4.1.1. Topic-Dependent Language Model

We first consider the joint training case, where our training data consists of the first 20 Fisher topics, split by topic, and we evaluate test set perplexities on each of the same 20 topics; specifically, we report the average test set perplexity over all 20 topics. Using a Multi-Factor LM with sparse component topology analogous to that in Fig. 1 (a) (but with 20 topics) we trained a joint model on the training set. Parameters γ_0, γ_{1c} and γ_2 were tuned using coarse grid search on the development set (for simplicity we set γ_{1c} equal for all c). The model we use to compute perplexity on test set topic t is the matched topic-dependent model with parameters $L + S_0 + S_t$.

	mKN	Multi-Factor LM	$L + S_0$
Joint Training	81.5	**79.7**	93.1
Adaptation	84.6	**83.3**	98.9

Table 1. Jointly trained and adapted test set perplexities, averaged over topics 1-20 (joint) or topics 21-40 (adaptation).

We also consider another common scenario: the training data and evaluation data have some type of mismatch; specifically, we consider topic mismatch. Our training data consists of the same first 20 topics of the Fisher data used before, while we treat each of the next 20 topics (21-40) as new domains. We adapt our Multi-Factor LM to new test topic t' as follows: 1) from the model trained in Sec. 4.1.1, we keep general L and S_0, but discard all training topic-specific models S_t, 2) we parameterize the adapted model with weights $L + S_0 + S_{t'}$, and 3) we estimate the new $S_{t'}$ by solving the following convex adaptation optimization problem

$$\min_{S_{t'}} \quad \gamma_1\|S_{t'}\|_1 + \frac{\gamma_2}{2}\|S_{t'}\|_F^2 - \mathcal{L}(\mathcal{X}_{t'}; L, S_0, S_{t'}) \quad (10)$$

This is solved by a straightforward variant of the proximal gradient algorithm employed used for training the Multi-Factor LM. There are a few points to note about adaptation. First, the low rank L and general sparse S_0 components are preserved, which assumes that they are capturing topic-independent information; this is a reasonable assumption in our case because topic-dependent n-grams ended up in the various S_t, by design. Second, the adapted models are learned independently from other new topics, rather than jointly. Finally, the adaptation problem is significantly faster than the original model training, because no low rank component is being learned.

As a baseline we compare against an n-gram model with modified-Kneser-Ney (mKN) smoothing; to evaluate topic t we linearly interpolate a general model (trained on the first 20 topics) with a topic-dependent model trained only on topic t's data using the SRILM toolkit [16]. (We found linear interpolation to perform better than count merging for this task.) In the "Joint case," the topic training data is accounted for in the general model, and in the "Adaptation case" it is not. The results are presented in Table 1. In perplexity, the Multi-Factor SLR-LM performs similarly to the baseline, slightly edging the modified-Kneser-Ney interpolated models by 2%. In the last column we see that the perplexities using models parameterized by $L + S_0$ only (i.e. omitting the topic-dependent factors) are much worse, suggesting that the topic-dependent factors play a prominent role in capturing the language behavior.

4.1.2. Keyword Extraction

Apart from its role as a language model, the Multi-Factor SLR-LM is of interest for its ability to identifying keywords associated with the factors. Specifically, the sparse entries of the S_t components contain the corrections to the general model for the factor-specific case; that is, they distill out the key differences between general and factor-dependent language. Here we evaluate the quality of this method of keyword extraction; to measure quality, we collect the highest rated 30 entries in each of the 20 sparse components learning and have them annotated as topically-relevant or not. Recall that our "keywords" can be any order of n-gram; because our model in Sec. 4.1.1 is a bigram, the keywords learned here are unigrams and bigrams.

We compare against two other keyword extraction methods, which each make use of a special word-document matrix (technically n-gram-topic matrix - the rows are all bigrams and unigrams

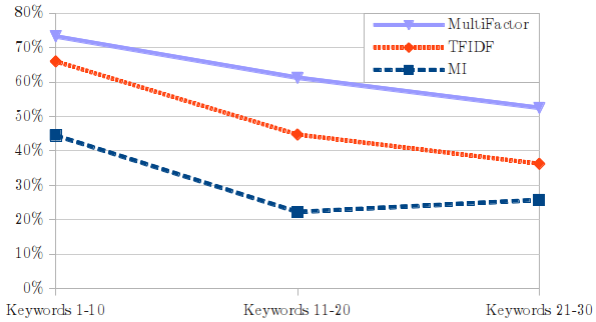


Fig. 3. Percentage of keywords labeled as relevant, for the Multi-Factor, TF-IDF and Mutual Information methods. Results averaged over 20 topics and two annotators.

	"Life Partners"	"Minimum Wage"
MF LM	soul mate, problem solving, physical attraction	food stamps, minimum, flipping burgers
TF-IDF	life partner, partner, life partners	five fifteen, wage, minimum wage
MI	married, life, important	wage, minimum, minimum wage

Table 2. Top three topic keywords learned from the data.

observed in the data and the columns are the 20 topics). The first baseline reweights the matrix using the standard term-frequency inverse-document-frequency (TF-IDF) scheme; after reweighting, the largest 30 entries in each column are used as the keywords. In our second baseline, inspired by feature selection, we use mutual information (MI) between the features (n-grams) and the topics (binary one-vs-rest) to rank the features per topic; the top 30 largest n-grams per topic after stop word filtering are selected as keywords. (In contrast, the Multi-Factor and TF-IDF methods did not require any stop word filtering.)

The keywords from all three methods were combined, with order randomized, and labeled as clearly-topically-relevant or not by two annotators unaffiliated with this research. Fig. 3 plots the percentage of keywords that were rated as clearly-relevant (average over 20 topics and two annotators), in three bins: the top 10 rated keywords per topic, keywords 11-20, and keywords 21-30. While the Multi-Factor model has the highest percentage of relevant keywords at each level, the biggest gains are due to the quality of keywords decaying more slowly in Multi-Factor model than the other baselines. Over all bins, the 62% of the keywords learned by the Multi-Factor model are good, compared to 49% for the TF-IDF method and 31% for the mutual information approach. Some examples of the top keywords by method are listed in Table 2.

4.2. Supreme Court Transcripts

We also explore the use of a Multi-Factor model with overlapping factors configuration as shown in Fig. 1 (b). For these experiments we use a subset of the Supreme Court corpus[1] consisting of 20 court cases, with 207k words, 58 speakers and two "roles" (justice and advocate). The vocabulary size is 7.3k. Examples of important keywords identified for specific cases include:

[1] www.oyez.org

1. **Rush Prudential HMO, Inc. v. Moran**. An HMO denying a request to cover a surgery: "savings clause," "medical necessity," "h m," "m o," "pilot life."
2. **TRW v. Andrews**. Allegations of violating the Fair Credit Reporting Act: "equitable estoppel," "reporting agency," "misrepresentation exception," "liability arises."
3. **Harris v. United States**. Regarding the sale of illegal narcotics while carrying an unconcealed firearm: "mandatory minimum," "reasonable doubt," "seven years."
4. **Toyota Motor Mfg v. Williams**. A claim of assembly line work leading to carpal tunnel syndrome: "worker's compensation," "assembly line," "life activity."

The model also learned characteristic language associated with the roles of justice, including question words ("why," "how,") and confirmations structured as statements ("you're saying," "your view," "I thought"), and advocate, including deferential language ("your honor," "chief justice," "that's correct") and hedging ("I think"). The per-speaker factors are most reliable for justices, for whom the data covers several cases. In this, we captured speaker idiosyncrasies (e.g. Breyer's habit of starting sentences with "all right" and Scalia's disfluencies) and Rehnquist's expressions that are characteristic of the role of Chief Justice (e.g. "we'll hear," "minutes remaining," "is submitted").

5. DISCUSSION

In summary, we introduced a multi-factor exponential language model that allows supervised learning of overlapping factors that influence sequential language behavior. As a language model, the model provides only small gains in perplexity in a topic adaptation scenario compared to a baseline modified-Kneser-Ney that interpolates general and topic specific models. It is perhaps of greater interest as a mechanism to identify factor-dependent characteristics. In particular, the n-gram elements encoded in sparse parameter matrices give an intuitive way to identify factor-dependent keyword phrases. On a conversational speech task, we demonstrate that human raters prefer topic keywords learned by the multi-factor model over TF-IDF and mutual information baselines. With Supreme Court transcripts, we show qualitatively the ability to learn factor-dependent keywords for different court cases, roles (justice vs advocate), and speakers. In addition to summarization, identification of keyword phrases is of interest for feature selection, learning lexical items, and detecting new or anomalous events. Encoding words with other features (e.g. morphological structure, syntactic dependents) would also make it possible to identify other types of idiosyncratic phenomena.

6. ACKNOWLEDGEMENTS

Supported in part by NSF CAREER grant ECCS-0847077, by DARPA grant # FA8750-09-C-0179, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. REFERENCES

[1] S. Schwarm and M. Ostendorf, "Text normalization with varied data sources for conversational speech language modeling," in *Proc. ICASSP*, 2002, pp. 789–792.

[2] R. Kneser and V. Steinbiss, "On the dynamic adaptation of stochastic language models," in *Proc. ICASSP*, 1993, vol. 2, pp. 586–589.

[3] G. Adda, M. Jardino, and J. Gauvain, "Language modeling for broadcast news transcription," in *Proc. Eurospeech*, 1999, pp. 1759–1762.

[4] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech Communication*, vol. 42, pp. 93–108, 2004.

[5] S. F. Chen, K. Seymore, and R. Rosenfeld, "Topic adaptation for language modeling using unnormalized exponential models," in *Proc. ICASSP*, 1998.

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

[7] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[8] Y.-C. Tam and T. Schultz, "Correlated latent semantic model for unsupervised lm adaptation," in *Proc. ICASSP*, 2007.

[9] S. Huang and S. Renals, "Unsupervised language model adaptation based on topic and role information in multiparty meetings," in *Proc. Interspeech*, 2008, pp. 833–836.

[10] F. Wood and Y. Teh, "A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation," in *Proc. AISTATS*, 2009.

[11] B. Hutchinson, M. Ostendorf, and M. Fazel, "A sparse plus low rank language model," in *Proc. Interspeech*, 2012.

[12] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling," *Computer Speech and Language*, vol. 14, no. 4, pp. 355–372, 2000.

[13] Kerui Min, Zhengdong Zhang, John Wright, and Yi Ma, "Decomposing background topics from keywords by principal component pursuit," in *ACM CIKM*, October 2010.

[14] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *Proc. ICML*, 2007, pp. 641–648.

[15] K. C. Toh and S. W. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," *Pacific Journal of Optimization*, November 2009.

[16] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. ICSLP*, 2002, pp. 901–904.