# Latent Topic Model based Representations for a Robust Theme Identification of Highly Imperfect Automatic Transcriptions

Mohamed Morchid[1], Richard Dufour[1], Georges Linarès[1], and Youssef Hamadi[2]

[1]LIA - University of Avignon, (France)
{firstname.lastname}@univ-avignon.fr
[2] Microsoft Research, Cambridge (United Kingdom)
youssefh@microsoft.com

**Abstract.** Speech analytics suffer from poor automatic transcription quality. To tackle this difficulty, a solution consists in mapping transcriptions into a space of hidden topics. This abstract representation allows to work around drawbacks of the ASR process. The well-known and commonly used one is the topic-based representation from a Latent Dirichlet Allocation (LDA). During the LDA learning process, distribution of words into each topic is estimated automatically. Nonetheless, in the context of a classification task, LDA model does not take into account the targeted classes. The supervised Latent Dirichlet Allocation (sLDA) model overcomes this weakness by considering the class, as a response, as well as the document content itself. In this paper, we propose to compare these two classical topic-based representations of a dialogue (LDA and sLDA), with a new one based not only on the dialogue content itself (words), but also on the theme related to the dialogue. This original Author-topic Latent Variables (ATLV) representation is based on the Author-topic (AT) model. The effectiveness of the proposed ATLV representation is evaluated on a classification task from automatic dialogue transcriptions of the Paris Transportation customer service call. Experiments confirmed that this ATLV approach outperforms by far the LDA and sLDA approaches, with a substantial gain of respectively 7.3 and 5.8 points in terms of correctly labeled conversations.

## 1 Introduction

Automatic Speech Recognition (ASR) systems frequently fail on noisy conditions and high Word Error Rates (WERs) make difficult the analysis of the automatic transcriptions. Solutions generally consist in improving the ASR robustness or/and the tolerance of speech analytic systems to ASR errors. In the context of telephone conversations, the automatic processing of these human-human interactions encounters many difficulties, especially due to the speech recognition step required to transcribe the speech content. Indeed, the speaker behavior may be unexpected and the train/test mismatch may be very large, while speech signal may be strongly impacted by various sources of variability: environment and channel noises, acquisition devices...

One purpose of the telephone conversation application is to identify the main theme related to the reason why the customers called. In this considered application, 8 classes corresponding to customer requests are considered (*lost and founds, traffic state, timelines...*). Additionally to the classical transcription problems in such adverse conditions, the theme identification system should deal with class (*i.e.* theme) proximity. For example, a *lost & found* request, considered as the main conversation theme, can also be related to itinerary (*where the object was lost?*) or timeline (*when?*). As a result, this particular multi-theme context makes identification of the main theme more difficult, ambiguities being introduced with the secondary themes.

An efficient way to tackle both ASR robustness and class ambiguity is to map dialogues into a topic space abstracting the ASR outputs. Dialogues classification will then be achieved in this topic space. Many unsupervised methods to estimate topic-spaces were proposed in the past. Latent Dirichlet Allocation (LDA) [1] was largely used in speech analytics applications [2]. During the LDA learning process, distribution of words into each topic is estimated automatically. Nonetheless, the class (or theme) associated to the dialogue is not directly taken into account in the topic model. Indeed, the classes are usually only used to train a classifier at the end of the process. As a result, such a system separately considers the document content (*i.e.* words), to learn a topic model, and the labels (*i.e.* classes) to train a classifier. Nonetheless, in the considered theme identification application, a relation between the document content (words) and the document label (class) exists.

This word/theme relation is crucial to efficiently label unseen dialogues. The supervised LDA [3] works around this drawback by considering the class belonging to a document, as a response during the learning process of the topic space. However, this representation could not substantially evaluate the relation between document content (words) and each theme. Indeed, these relations are evaluated through relations between topics and classes of the sLDA model. Moreover, these models (LDA and sLDA) need to infer an unseen document into each topic space to obtain a vectorial representation. The processing time during the inference phase as well as the difficult choice of an efficient number of iterations, do not allow us to evaluate effectively and quickly the best theme related to a given dialogue.

For these reasons, this paper is based on the work presented in [4] in which the authors proposed to use the Author-topic (AT) model [5] to represent a document instead of the classical LDA approach. The contribution of the paper is to go beyond this previous work by comparing the proposed AT model, called Author-topic Latent Variables (ATLV) representation, with the supervised LDA (sLDA) representation, which is an interesting alternative for classification tasks. For sake of comparison, results using the classical LDA topic-based representation [4] will also be reported. This robust ATLV representation takes into consideration all information contained into a document: the content itself (*i.e.* words), the label (*i.e.* class), and the relation between the distribution of words into the document and the label, considered as a latent relation. From this model, a

vectorial representation in a continuous space is built for each dialogue. Then, a supervised classification approach, based on SVM [6], is applied. This method is evaluated in the application framework of the RATP call centre (Paris Public Transportation Authority), focusing on the theme identification task [7] and compared to LDA and sLDA approaches.

The rest of this paper is organized as follows. Topic model representations from document content are described in Section 2, by introducing LDA, sLDA and ATLV representations. Section 3 presents the experimental protocol and results obtained while finally, Section 4 concludes the work and gives some perspectives.

## 2 Topic-model for automatic transcriptions

Dialogues, automatically transcribed using an Automatic Speech Recognition (ASR) system, contain many errors due to noisy recording conditions. An elegant way to tackle these errors is to map dialogues in a thematic space in order to abstract the document content. The most known and used one is the Latent Dirichlet Allocation (LDA) [1] model. The LDA approach represents documents as a mixture of latent topics. Nonetheless, this model does not code statistical relations between words contained into the document, and the label (*i.e.* class) that could be associated to it.

Authors in [3] proposed the supervised Latent Dirichlet Allocation (sLDA) model. This model introduces, in the LDA model, a response variable associated with each document contained into the training corpus. This variable is, in our considered context, the theme associated to a dialogue. In the sLDA model, the document and the theme are jointly modeled during the learning process, in order to find latent topics that will best predict the theme for an unlabeled dialogue of the validation data set. Although this model codes relation between the response variable and topics, this relation is not effective for theme identification task. Indeed, sLDA allows to relate a response (or theme) to a topic which is related itself to a set of words, and does not code strongly the relation between the document content (words occurrences) and the themes directly.

To go beyond LDA and, more importantly, sLDA [3] limits, an adapted Author-topic (AT) model is proposed here. The proposed Author-topic Latent Variables (ATLV) representation links both authors (here, the label) and documents content (words). The next sections describe LDA, sLDA, and ATLV based representations.

### 2.1 Latent Dirichlet Allocation (LDA)

In topic-based approches, such as Latent Dirichlet Allocation (LDA), documents are considered as a *bag-of-words* [8] where the word order is not taken into account. These methods demonstrated their performance on various tasks, such as sentence [9] or keyword [10] extraction. In opposition to a multinomial mixture model, LDA considers that a theme is associated to each occurrence of a

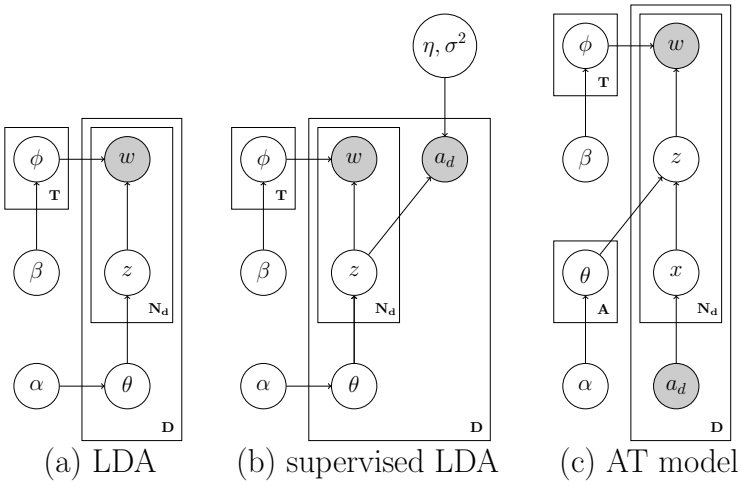


**Fig. 1.** Generative models for documents in plate notation for Latent Dirichlet Allocation (LDA) (a), supervised LDA (b) and Author-Topic (AT) (c) models.

word composing the document. Thereby, a document can change of topics from a word to another. However, the word occurrences are connected by a latent variable which controls the global respect of the distribution of the topics in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them. LDA models have been shown to generally outperform other topic-based approches on information retrieval tasks [11].

The generative process corresponds to the hierarchical Bayesian model shown, using plate notation, in Figure 1 (a). Several techniques, such as Variational Methods [1], Expectation-propagation [12] or Gibbs Sampling [13], have been proposed to estimate the parameters describing a LDA hidden space. The Gibbs Sampling reported in [13], and detailed in [14], is used to estimate LDA parameters and to represent a new dialogue $d$ with the $r^{th}$ topic space of size $T$. This model extracts a feature vector $V_d^{z^r}$ from the topic representation of $d$. The $j^{th}$ feature is:

$$V_d^{z_j^r} = \theta_{j,d}^r \ , \tag{1}$$

where $\theta_{j,d}^r = P(z_j^r|d)$ is the probability of topic $z_j^r$ $(1 \leq j \leq T\ )$ generated by the unseen dialogue $d$ in the $r^{th}$ topic space of size $T$.

### 2.2 Supervised LDA (sLDA)

Figure 1 (b) presents the sLDA model into its plate notation. Let $a \in R$ be the response (or theme in our theme identification context), and let fix the model parameters: $T$ topics $\beta 1 : T$ (each $\beta_k$ is a vector of term probabilities), the Dirichlet parameter $\alpha$, and the response (theme) parameters $\eta$ and $\sigma^2$. With the

sLDA model, each document and response arises from the following generative process:

1. Draw topic distribution $\theta|\alpha \sim \text{Dir}(\alpha)$.
2. For each word
   (a) Draw topic assignment $z_n|\theta \sim \text{Mult}(\theta)$.
   (b) Draw word $w_n|z_n \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable (or theme) $a|z_{1:N_d}, \eta, \sigma^2 \sim \mathcal{N}\left(\eta^\top \overline{z}, \sigma^2\right)$.

with $\overline{z} = \frac{1}{N_d} \sum_{n=1}^{N_d} z_n$. The hyper-parameters of the sLDA model are estimated by performing a variational expectation-maximization (EM) procedure, also used in unsupervised LDA [1]. One can find out more about the parameters estimation or, more generally, about the sLDA itself, in [3].

The sLDA approach allows to directly estimate the probability for a theme $a$ (or response) to be generated by a dialogue $d$. Then, the theme $a$ which maximizes the prior $P(a|z_n, \eta, \sigma^2)$ is assigned to the dialogue $d$ with:

$$C_{a,d} = \arg\max_{a \in A} \left\{ P(a|d, z, \eta, \sigma^2) \right\} \tag{2}$$

Thus, each dialogue from the test or development set is labeled with the most likely theme given a sLDA model. This one does not require a classification method, which is not the case for LDA and ATLV representations.

### 2.3 Author-topic Latent Variables (ATLV)

The Author-topic (AT) model, represented into its plate notation in Figure 1 (c), uses latent variables to model both the document content (words distribution) and the authors (authors distribution). For each word $w$ contained into a document $d$, an author $a$ is uniformly chosen at random. Then, a topic $z$ is chosen from a distribution over topics specific to that author, and the word is generated from the chosen topic.

In our considered application, a document $d$ is a conversation between an agent and a customer. The agent has to label this dialogue with one of the 8 defined themes, a theme being considered as an author. Thus, each dialogue $d$ is composed with a set of words $w$ and a theme $a$. In this model, $x$ indicates the author (or theme) responsible for a given word, chosen from $a_d$. Each author is associated with a distribution over topics ($\theta$), chosen from a symmetric Dirichlet prior ($\overrightarrow{\alpha}$) and a weighted mixture to select a topic $z$. A word is then generated according to the distribution $\phi$ corresponding to the topic $z$. This distribution $\phi$ is drawn from a Dirichlet ($\overrightarrow{\beta}$).

The parameters $\phi$ and $\theta$ are estimated from a straightforward algorithm based on Gibbs Sampling such as LDA hyper-parameters estimation method (see Section 2.1). One can find more about Gibbs Sampling and AT model in [5].

Each dialogue $d$ is composed with a set of words $w$ and a label (or theme) $a$ considered as the author in the AT model. Thus, this model allows one to code

statistical dependencies between dialogue content (words $w$) and label (theme $a$) through the distribution of the latent topics into the dialogue.

Gibbs Sampling allows us to estimate the AT model parameters, in order to represent an unseen dialogue $d$ with the $r^{th}$ author topic space of size $T$. This method extracts a feature vector $V_d^{a_k^r} = P(a_k|d)$ from an unseen dialogue $d$ with the $r^{th}$ author topic space $\Delta_r^n$ of size $T$. The $k^{th}$ $(1 \leq k \leq A)$ feature is:

$$V_d^{a_k^r} = \sum_{i=1}^{N_d} \sum_{j=1}^{T} \theta_{j,a_k}^r \phi_{j,i}^r \tag{3}$$

where $A$ is the number of authors (or themes); $\theta_{j,a_k}^r = P(a_k|z_j^r)$ is the probability of author $a_k$ to be generated by the topic $z_j^r$ $(1 \leq j \leq T)$ in $\Delta_r^n$. $\phi_{j,i}^r = P(w_i|z_j^r)$ is the probability of the word $w_i$ ($N_d$ is the vocabulary of $d$) to be generated by the topic $z_j^r$.

This representation, based on the AT latent variables, is called Author-topic Latent Variables (ATLV) representation in this work.

## 3 Experiments and Results

We propose to evaluate the effectiveness of the proposed approach in the application framework of the DECODA corpus [7, 15, 2].

### 3.1 Experimental protocol

The DECODA project [7] corpus, used to perform experiments on theme identification, is composed of human-human telephone conversations in the customer care service (CCS) of the RATP Paris transportation system. It is composed of 1,242 telephone conversations, corresponding to about 74 hours of signal, split into a train, development and test set, with respectively 740, 175 and 327 dialogues.

To extract textual content of dialogues, an Automatic Speech Recognition (ASR) system is needed. The LIA-Speeral ASR system [16] is used for the experiments. Acoustic model parameters were estimated from 150 hours of speech in telephone conditions. The vocabulary contains 5,782 words. A 3-gram language model (LM) was obtained by adapting a basic LM with the training set transcriptions. A "stop list" of 126 words[1] was used to remove unnecessary words (mainly function words) which results in a Word Error Rate (WER) of 33.8% on the training, 45.2% on the development, and 49.5% on the test. These high WER are mainly due to speech disfluencies and to adverse acoustic environments (for example, calls from noisy streets with mobile phones)

38 topic spaces are elaborated by varying the number of topics from 10 to 200 (step of 5 topics). The topic spaces are learned with a homemade implementation of LDA and AT models, while the version implemented by [17] is used for sLDA.

[1] http://code.google.com/p/stop-words/

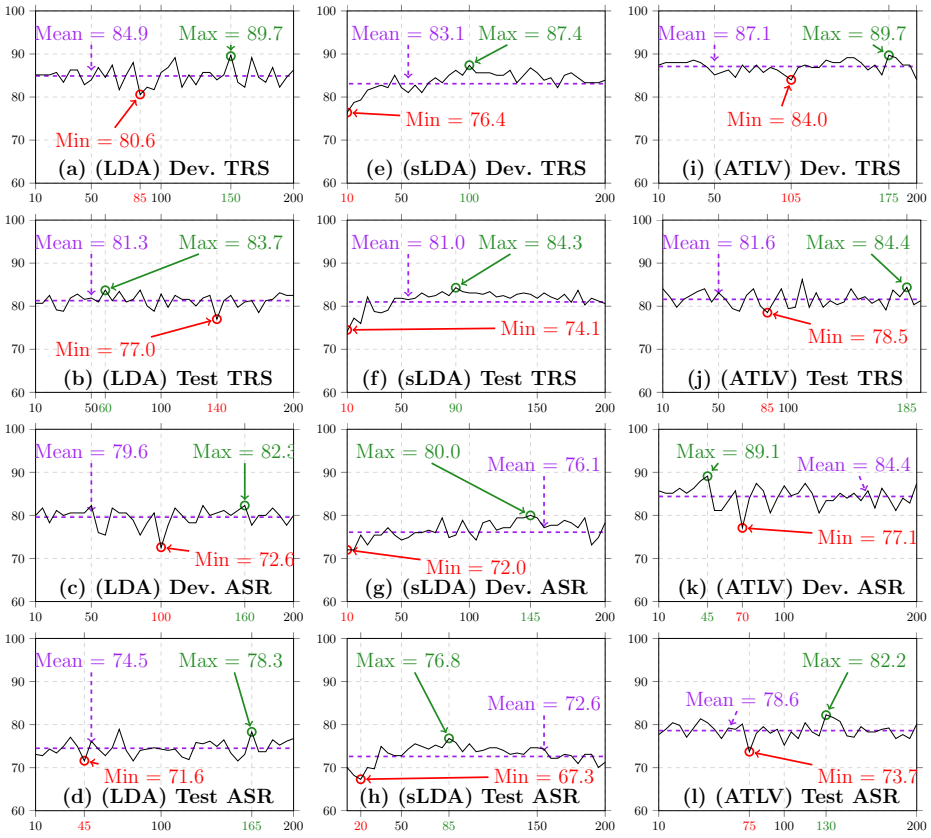


**Fig. 2.** Theme classification accuracies (%) using various LDA topic-based representations on the development and test sets with different experimental configurations. X-axis represents the number $n$ of classes contained into the topic space ($10 \leq n \leq 200$).

A classification approach based on Support Vector Machines (SVM) is performed using the *one-against-one* method with a linear kernel, to find out the main theme of a given dialogue. This method gives a better accuracy than the *one-against-rest* [18]. SVM input is a vector representation of an unseen dialogue $d$ [2].

For sake of comparison, experiments are conducted using the manual transcriptions only (TRS) and the automatic transcriptions only (ASR). The conditions indicated by the abbreviations between parentheses are considered for the development (Dev) and the test (Test) sets. Only homogenous conditions (TRS or ASR for both training and validations sets) are considered in this study. Authors in [2] notice that results collapse dramatically when heterogenous con-

[2] $V_d^{z_j^r}$ for a LDA representation and $V_d^{a_k^r}$ for an ATLV representation.

ditions are employed (TRS or TRS+ASR for training set and ASR for validation set).

**Table 1.** Theme classification accuracies (%) for LDA, sLDA, and ATLV representations. **Best** corresponds to the best operating point obtained on the test data, while **Real** corresponds to the one estimated on the development set and applied to the test set.

| **Topic** | **DATASET** | | **Dev** | | **Test** | |
|---|---|---|---|---|---|---|
| **Model** | **Train** | **Test** | **#topics** | **Best** | **Best** | **Real** |
| LDA | TRS | TRS | 150 | 89.7 | 83.7 | 82.5 |
| LDA | ASR | ASR | 160 | 82.3 | 78.3 | 73.1 |
| sLDA | TRS | TRS | 100 | 87.4 | 84.3 | 83.1 |
| sLDA | ASR | ASR | 145 | 80.0 | 76.8 | 74.6 |
| ATLV | TRS | TRS | 175 | 89.7 | 84.4 | **83.7** |
| ATLV | ASR | ASR | 45 | 89.1 | 82.2 | **80.4** |

### 3.2 Results

The results obtained using manual (TRS) and automatic (ASR) transcriptions with respectively a topic-based representation from LDA, sLDA and ATLV, are presented in Figure 2. One can firstly point out that, for all dialogue representations (LDA, sLDA or ATLV), best results are obtained with manual transcriptions (TRS). Moreover, one can notice that the ATLV representation outperforms LDA, no matter the corpus (development or test) or conditions (TRS/ASR) studied.

In order to better compare performance obtained by all approaches (LDA / sLDA / ATLV), best results are reported in Table 1. Note that these results are given in **Best** and **Real** application condition, *i.e.* the **Real** configuration (number of topics contained into the topic space) being chosen with the **Best** operating point of the development set. As a result, a better operating point could exist in the test set, which could explain the performance difference between results reported in Table 1 and Figure 2.

With this real condition, we can note that the ATLV representation allows us to outperform both the LDA and sLDA approaches, with a respective gain of 1.2 and 0.6 points using the manual transcriptions (TRS), and of 7.3 and 5.8 points using the automatic transcriptions (ASR). This confirms the initial intuition that ATLV representation allows to better handle ASR errors than other topic-based ones. Improvements are mostly when ASR documents are used, and outcomes obtained for LDA and sLDA are quite close for both TRS and ASR configurations. One can point out that the results obtained for all topic-based representations in the TRS configuration are similar.

Another interesting point is the stability and robustness of the ATLV curve of the development set in TRS condition, comparatively to the LDA or sLDA

representations. Indeed, the results are mainly close to the mean value (87.1%). The maximum achieved by both representations in TRS condition are the same. Thus, since dialogues are labeled (annotated) by an agent and a dialogue may contain more than one theme, this maximum represents the limit of a topic-based representation in a multi-theme context. Nonetheless, this remark is not applicable to the ASR condition.

## 4 Conclusion

Performance of ASR systems depends strongly to the recording environment. In this paper, an efficient way to deal with ASR errors by mapping a dialogue into a Author-topic Latent Variables (ATLV) representation space is presented. This high-level representation allows us to significantly improve the performance of the theme identification task. Experiments conducted on the DECODA corpus showed the effectiveness of the proposed ATLV representation in comparison to the use of the classic LDA representation or the more elaborated and adapted sLDA, with gains of at least 0.6 and 5.8 points (with the closest representation based on sLDA) respectively using manual and automatic transcriptions.

This representation suffers from the fact that theme distribution could not directly be estimated for an unseen document. Indeed, the proposed approach has to evaluate the probability $P(a_k|d)$ through the document content (words distribution) and the themes distribution. Thus, an interesting perspective is to add a new latent variable into the proposed model, to allow this model to infer effectively an unseen dialogue among all the authors.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research **3** (2003) 993–1022
2. Morchid, M., Dufour, R., Bousquet, P.M., Bouallegue, M., Linarès, G., De Mori, R.: Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In: ICASSP. (2014)
3. Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in neural information processing systems. (2008) 121–128
4. Morchid, M., Dufour, R., Bouallegue, M., Linarès, G.: Author-topic based representation of call-center conversations. In: International Spoken Language Technology Workshop (SLT) 2014, IEEE, to appear (2014)
5. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press (2004) 487–494
6. Vapnik, V.: Pattern recognition using generalized portrait method. Automation and Remote Control **24** (1963) 774–780
7. Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Beze, M., De Mori, R., Arbillot, E.: Decoda: a call-centre human-human spoken conversation corpus, LREC'12 (2012)
8. Salton, G.: Automatic text processing: the transformation. Analysis and Retrieval of Information by Computer (1989)

9. Bellegarda, J.: Exploiting latent semantic information in statistical language modeling. Proceedings of the IEEE **88** (2000) 1279–1296
10. Suzuki, Y., Fukumoto, F., Sekiguchi, Y.: Keyword extraction using term-domain interdependence for dictation of radio news. In: 17th international conference on Computational linguistics. Volume 2., ACL (1998) 1272–1276
11. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **42** (2001) 177–196
12. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (2002) 352–359
13. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences of the United States of America **101** (2004) 5228–5235
14. Heinrich, G.: Parameter estimation for text analysis. Web: http://www. arbylon. net/publications/text-est. pdf (2005)
15. Morchid, M., Linarès, G., El-Beze, M., De Mori, R.: Theme identification in telephone service conversations using quaternions of speech features. In: INTERSPEECH. (2013)
16. Linarès, G., Nocéra, P., Massonie, D., Matrouf, D.: The lia speech recognition system: from 10xrt to 1xrt. In: Text, Speech and Dialogue, Springer (2007) 302–308
17. Wang, C., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1903–1910
18. Yuan, G.X., Ho, C.H., Lin, C.J.: Recent advances of large-scale linear classification. **100** (2012) 2584–2603