ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

On-the-fly Topic Adaptation for YouTube Video Transcription

Kapil Thadani, Fadi Biadsy,† Daniel M. Bikel †*

*Department of Computer Science, Columbia University, New York, NY
kapil@cs.columbia.edu
†Google Inc., New York, NY
{biadsy,dbikel}google.com

Abstract

Automatic closed-captioning of video is a useful application of speech recognition technology but poses numerous challenges when applied to open-domain user-uploaded videos such as those on YouTube. In this work, we explore a strategy to improve decoding accuracy for video transcription by decoding each video with a language model (LM) adapted specifically to the topics that the video covers. Taxonomic topic classifiers are used to determine the topic content of videos and to build a large set of topic-specific LMs from web documents. We consider strategies for selecting and interpolating LMs in both supervised and unsupervised scenarios in a two-pass lattice rescoring framework. Experiments on a YouTube video corpus show a 3.6 absolute reduction in WER over generic single-pass transcriptions as well as a statistically significant 0.8 absolute improvement over rescoring with a very large non-adapted LM built from all the documents.

1. Introduction

The popularity and ubiquity of online streaming video services in recent years has spurred interest in the task of producing closed captions of videos using automated speech recognition. Generating manual transcriptions is labor-intensive and expensive and, as a result, this is usually not feasible for online video producers who are often amateur videographers. Automated closed-captioning systems therefore fill a need for many real-world applications including the assistance of hearing-impaired viewers and the indexing and browsing of large video collections. For this reason, this task is also the focus of many commercial software systems.

YouTube, in particular, ranks among the largest and most diverse collection of user-generated videos. In order to transcribe every video containing speech, we have to tackle numerous challenges: acoustic models must cope with a variety of recording environments, noise, sound effects and multiple speakers, while language models need to be able to decode conversational (often spontaneous) speech on a wide range of topics. This final problem is the subject of the work presented here.

We hypothesize that decoding the speech of a particular video using an LM adapted to the topics of that video will improve automatic transcription. Our approach is to use a general-purpose topic taxonomy developed at Google, along with text and video classifiers that assign topics to documents and videos according to this taxonomy. First, we build LMs specific to each topic. Next, we perform topic classification for each video. Finally, we dynamically interpolate amongst our topic-specific LMs when decoding the video's speech.

Incorporating topic information into LMs has a fairly long history. Previous efforts [1, 2] used document clustering using an unbounded amount of history in a document to model its topic but are not directly applicable to speech recognition. A major strain of related research focuses on modeling topics as latent variables [3, 4, 5, 6, 7, 8, 9]; however, our setting assumes the availability of a comprehensive topic taxonomy. The video transcription scenario specifically involves the problem of dynamic unsupervised LM adaptation, which has previously been considered by Khudanpur and Wu [10, 11], who use maximum entropy models to incorporate n-gram and topic constraints, as well as more recent approaches based on latent Dirichlet allocation [4, 6, 7, 8, 9]. Broadly speaking, these techniques assign topics to utterances using n-best recognition hypotheses produced under generic models and then obtain a final transcription under a topic-adapted model. We employ a similar hypothesis-guided technique for our unsupervised adaptation strategy.

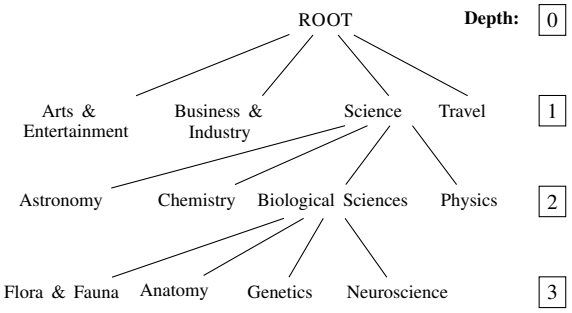

Figure 1: A sample of some topics from the taxonomy.

The main contributions of this paper include on-demand adaptation strategies for video transcription via linear interpolation of topic-specific LMs under both supervised and unsupervised scenarios. Given a corpus of transcribed videos, we discuss a nearest neighbor technique to recover interpolation weights from optimal interpolation weights generated from transcribed utterances. We also construct a simple but effective unsupervised approach in which mixture weights are optimized to accurately reproduce a first-pass transcript decoded with a generic LM. Finally, our experiments consider different choices in selecting models for interpolation: using video metadata, the first-pass transcript and the taxonomy structure. The proposed adaptation techniques generate significant gains against a strong baseline which employs a very large non-adapted LM.

2. Taxonomic Topic Categorization

Throughout this work, we employ a taxonomy of generic categories to describe video and textual content. This taxonomy is used widely within Google and features 1112 categories in a hierarchical tree with up to seven levels ranging from depth 0 (the root) to depth 6. In this work, we apply this general-purpose taxonomy both to videos on YouTube as well as two types of text: generic web documents and transcribed utterances. Figure 1 shows a sample of topics and paths in the taxonomy, and Figure 2 shows the number of nodes it contains at each level.

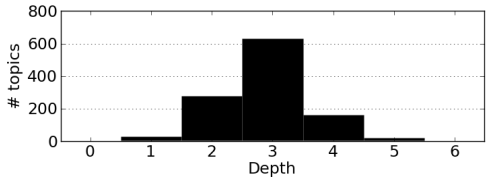

Figure 2: The number of nodes in the topic taxonomy at each depth starting from the root (depth 0).

2.1. Constructing topic-specific LMs

To build LMs associated with the categories in this taxonomy, we make use of a large corpus of crawled webpages along with a multiclass classifier that assigns topics from the above taxonomy to them. The classification system is based on binary linear SVMs with text features and is trained on a human-labeled corpus. Each document is added to the training corpus of its most probable category, as well as the training corpora of all ancestors of that category. As a result, taxonomy nodes at lower depths are strictly assigned more documents than nodes at higher depths. For example, documents classified under the depth-3 topic *Arts & Entertainment* → *Music & Audio* → *Pop Music* would also be included in document sets for the depth-2 topic *Arts & Entertainment* → *Music & Audio*, the depth-1 topic *Arts & Entertainment* as well as the depth-0 pseudo-topic ROOT. All documents in the corpus are necessarily also assigned to ROOT; the root LM therefore functions as a background LM.

2.2. Topic classification for videos

In addition to document classification, we also make use of topic classification of individual videos in order to guide the selection of LMs at decoding time. The multiclass classifier used here is an implementation of prior work in taxonomic video categorization that employs both textual and video content features [12, 13]. Each video is assigned up to three topics that are the most confident labels under the classifier.

3. Adaptation of Topic-specific LMs

Formally, we employ a generic taxonomy $\mathcal{T}$ of topic categories $C_t \in \mathcal{T}, t = 1, \ldots, |\mathcal{T}|$, each associated with a category-specific LM denoted by G_t. Each candidate video v is assigned topics from $\mathcal{T} \backslash \{\text{ROOT}\}$; we denote the set of category assignments for v by $\mathcal{C}(v)$. Our goal is to produce an LM adapted to $\mathcal{C}(v)$ at decoding time by estimating weights for linear interpolation of the LMs associated with the topics in $\mathcal{C}(v)$.

3.1. On-demand LM interpolation

Given a set of $m = |\mathcal{T}|$ backoff LMs $\mathcal{G} \triangleq \{G_1, \ldots, G_m\}$, a vector of mixture weights $\boldsymbol{\lambda} \triangleq (\lambda_1, \ldots, \lambda_m)^T$ and a vocabulary Σ, a linear interpolation of $\mathcal{G}$ by $\boldsymbol{\lambda}$ is defined as the LM $H_{\boldsymbol{\lambda}}$ assigning the following conditional probability for word $w \in \Sigma$ given context $h \in \Sigma^*$.

$$p_{H_{\boldsymbol{\lambda}}}(w|h) = \sum_{t=1}^{m} \lambda_t \, p_{G_t}(w|h) \tag{1}$$

On-demand linear interpolation of $\mathcal{G}$ using (1) directly may be inefficient because the model may need to back off several times in up to m LMs for any given (w, h) pair. To address this, the interpolated LM can be reformulated as a single backoff model [14]:

$$p_{H_{\boldsymbol{\lambda}}}(w|h) = \begin{cases} \sum_{t=1}^{m} \lambda_t \, p_{G_t}(w|h), & \text{if } hw \in S(\mathcal{G}) \\ f(\boldsymbol{\lambda}, \boldsymbol{\alpha}_h) \, p_{H_{\boldsymbol{\lambda}}}(w|h'), & \text{otherwise} \end{cases} \tag{2}$$

where $\boldsymbol{\alpha}_h \triangleq (\alpha_h(G_1), \ldots, \alpha_h(G_m))^T$ is a vector of backoff weights for the context h under every LM G_t, $S(\mathcal{G}) \triangleq \cup_{t=1}^{m} S(G_t)$ where each $S(G_t)$ is the set of all unpruned context-word sequences observed in the construction of G_t, and h' is the longest common suffix of h. Although a closed-form expression for $f(\boldsymbol{\lambda}, \boldsymbol{\alpha})$ exists in order to normalize the model in (2) to be equivalent to (1), simpler approximations such as $f(\boldsymbol{\lambda}, \boldsymbol{\alpha}) = \boldsymbol{\lambda}^T \boldsymbol{\alpha}$ can be used in practice. Additionally, since the set of models $\mathcal{G}$ is known in advance in our setting, $S(\mathcal{G})$ can be precomputed.

3.2. Supervised adaptation via nearest neighbors

Let $\mathcal{U}$ be a reasonably large training corpus consisting of videos whose utterances have been manually transcribed. We now aim to generate interpolation weights $\boldsymbol{\lambda}$ that can effectively leverage the information in $\mathcal{U}$. Recall that interpolation weights are not independent and so generating mixture weights $\boldsymbol{\lambda}(v)$ for an unseen test video v cannot be reduced to estimating each component $\lambda_t(v)$ separately. Our approach relies on the assumption that videos with similar topic classifications will yield similar mixture weights. We therefore estimate weights for an unseen video from its most similar transcribed videos, as determined by a distance metric over the topic-based characterization.

The training corpus is first preprocessed to determine the optimal mixture weights $\boldsymbol{\lambda}^*(u)$ for each training video $u \in \mathcal{U}$. Optimization is performed with a standard iterative EM-based approach over the true transcript of u in which each $\lambda_t(u)$ component is set proportional to the fraction of the probability that G_t contributes to the overall mixture:

$$\lambda_t^{(i+1)} = \frac{1}{n} \sum_{j=1}^{n} \frac{\lambda_t^{(i)} \, p_{G_t}(w_j|h_j)}{\sum_{r=1}^{m} \lambda_t^{(i)} \, p_{G_r}(w_j|h_j)} \tag{3}$$

To avoid slow convergence when the number of LMs m is large, we consider only a subset of the components of each $\boldsymbol{\lambda}(u)$ for optimization. Specifically, we restrict $\lambda_t(u)$ to be non-zero only when $C_t \in \mathcal{C}(u) \cup \{\text{ROOT}\}$, i.e., at most four $\lambda_t(u)$ are non-zero. The inclusion of the ROOT category background LM avoids limiting the decoding to potentially small LMs; instead, the optimization procedure must capture *how* topic-specific a given video is relative to the background model.

We indicate a video's topic content by a vector of per-category confidence scores $\mathbf{s}_u \in \mathbb{R}^{|\mathcal{T}|}$ in which the t'th component contains the confidence of the taxonomic classifier that $C_t \in \mathcal{C}(u)$. The estimation of $\boldsymbol{\lambda}(v) \triangleq (\lambda_1(v), \ldots, \lambda_{|\mathcal{T}|}(v))^T$ for an unseen video v then proceeds in two steps:

1. Find the k training videos $u_1, \ldots, u_k \in \mathcal{U}$ with smallest $d(\mathbf{s}_v, \mathbf{s}_{u_i})$
2. Set $\lambda_t(v) = \sum_{i=1}^{k} \beta_i \, \lambda_t^*(u_i) \;\; \forall t$

where $d(\mathbf{s}_v, \mathbf{s}_{u_i})$ is a distance function between two topic characterizations[1] and $\boldsymbol{\beta} \triangleq (\beta_1, \ldots, \beta_k)^T$ is a vector of linear coefficients which can be set to preferentially weight the influence of similar videos,[2] i.e., setting $\beta_i \propto 1/d(\mathbf{s}_v, \mathbf{s}_{u_i})$. The first step can be performed through either explicit search (for small $\mathcal{U}$), approximate search [15, 16] or locality-sensitive hashing [17].

3.3. Unsupervised adaptation to first-pass transcripts

Obtaining manual transcriptions for a sufficiently large training corpus of online videos can be both expensive and time-consuming, prompting the question of whether interpolation

[1] $\mathbf{s}_v \in \mathbb{R}^{|\mathcal{T}|}$ in our implementation so we use Euclidean distance. Potential alternatives include divergence measures for distributions and learned metrics.

[2] We report results for a uniform $\boldsymbol{\beta} = \mathbf{1}^k$ in our experiments.

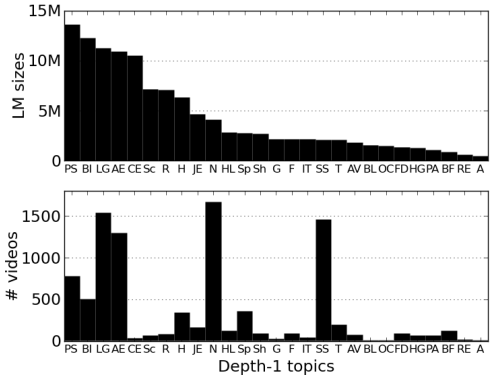

Figure 3: Proportion of topics in (a) the set of LMs $\mathcal{G}$, as seen by the aggregated size of LMs in terms of number of n-grams, and (b) the total number of corresponding topic assignments in the YouTube video corpus. Depth-1 categories are used to represent the assignment of all their taxonomic descendents in both plots.

weights $\boldsymbol{\lambda}$ for topic-specific LMs $\mathcal{G}$ can be estimated without the use of human-generated transcriptions. Our unsupervised adaptation strategy employs a two-pass approach to decoding. The first pass is performed with a generic LM F which is built from a large multi-domain document corpus. Interpolation weights $\boldsymbol{\lambda}$ for each video are then optimized to maximize the likelihood of the transcript x_F generated in the first pass. A second decoding pass is then performed with the interpolated model $H_{\boldsymbol{\lambda}}$ to yield the final transcription.

Optimizing towards a noisy first-pass transcript is analogous to the maximum likelihood linear regression (MLLR) approach for speaker adaptation in acoustic modeling [18]. We assume that F correctly transcribes enough of a video's utterances on average to provide a reasonable signal of its topic content. $H_{\boldsymbol{\lambda}}$ is then adapted to better model this topic distribution.

Since optimizing a mixture of all $|\mathcal{T}|$ LMs is problematic due to slow convergence, we only target mixture weights for a subset of topics derived from $\mathcal{C}(v)$ and its taxonomic ancestors. We can also obtain taxonomic categories $\mathcal{C}(x_F)$ from the first-pass transcript x_F using the same text-based topic classifier from §2.1; these topics provide a more direct characterization of the language in the target transcription and can therefore augment or replace those in $\mathcal{C}(v)$, which was produced by the video classifier. This offers a robustness to inconsistent topic semantics between the text classifier and the video classifier.

4. Experiments

The speech recognition system used was a speaker-adaptive system using maximum mutual information (MMI) training, along with MLLR and constrained MLLR (CMLLR) adaptation. For interpolation experiments, we make use of a two-pass rescoring framework in which word lattices obtained by decoding with the generic LM F are rescored using on-demand interpolation of the topic-specific LMs $\mathcal{G}$ with mixture weights $\boldsymbol{\lambda}$.

4.1. Corpora

The vast majority of documents for building our first-pass LM F were web pages obtained from a web crawl, but this model also contained the manually generated transcripts of training videos. We treated the first-pass LM as a black box as it had been trained and used in a speech recognizer prior to this work.

The corpus on which we performed all our speech recognition experiments is a set of YouTube videos containing news broadcast–style material downloaded in 2008. Colleagues identified a fairly broad array of YouTube news "channels" from which videos would be drawn. To ensure that both short and long videos were well represented, videos from each channel were sampled according to duration: the ith video was drawn from a duration-sorted list of n videos according to the normal distribution $\mathcal{N}(\frac{n}{2}, s\sqrt{\frac{n}{2}})$, where s was a user-controlled scaling factor. The corpus of sampled videos consisted of 3643 training videos and 77 test videos, constituting roughly 200 hours of audio. Videos were automatically segmented into short utterances (158942 in total) based on pauses between speech. The audio was transcribed at high quality by humans trained in the task. Topic assignments for videos were obtained by taxonomic classification as described in §2.2; each video was automatically assigned up to three categories based on classifier confidence scores. We chose a corpus of news-like videos in order to diminish the factors of acoustic environment, sound effects and music on our evaluation; however, the distribution of topic categories obtained on our video corpus is quite diverse (see figure 3).

Interpolation strategy	**WER**	**Relative reduction from**	
		F	G_{ROOT}
First-pass decoding with F	34.64	0.00%	-
Rescoring with G_{ROOT}	31.84	34.83%	0.00%
Supervised NN (k=10)	31.19	42.91%	12.40%
Optimization over first-pass transcript	31.07	44.40%	14.69%

Table 1: WER for proposed adaptation techniques. Relative WER reduction from baselines is computed under the *oracle* condition (see §4.2). Utterance-level and video-level gains from adaptation are statistically significant compared to baselines at $p \leq 0.05$ under a paired t-test and Wilcoxon's signed rank test.

All LMs used for the second pass rescoring stage were constructed using a text corpus obtained from a very large random web crawl performed in 2010. Each web document was cleaned of extraneous material (such as HTML tags) and then passed through an automatic filter to determine if the document was primarily in English. The full corpus contains approximately 59 billion words. We follow the hierarchical strategy described in §2.1 for building topic-specific LMs from this corpus. The root topic LM G_{ROOT}, which is built from all the documents in this collection, contains over 19 billion unique n-grams while the smallest topic-specific LM contains in excess of 400K n-grams. Every LM built was a 4-gram model using Kneser-Ney smoothing, compactly encoded as an FST.

4.2. Evaluation

Table 1 lists word error rates (WER) for decoding experiments[3] on the YouTube corpus described above. The baselines featured are single-pass decoding with F and the two-pass rescoring approach with the very large aggregate LM G_{ROOT}. The latter significantly outperforms naïve topic-based techniques such as rescoring with the single LM corresponding to the highest-confidence assigned topic (WER 32.65) or with a confidence-proportional interpolation of LMs for all topics assigned (WER 32.02). This LM therefore represents a strong baseline for adaptation strategies. Furthermore, the best-case (i.e., *oracle*) WER from first-pass lattices is just 26.6; this high oracle error rate illustrates the challenges posed by the YouTube video transcription task. We account for errors introduced specifically by our rescoring strategies (as opposed to oracle errors) when reporting relative WER reduction from the baselines in Table 1.

Both adaptation strategies exhibit statistically significant

[3]For the supervised approach, interpolation weights for each training video were estimated using LMs associated with $\mathcal{C}(u) \cup \{\text{ROOT}\}$. First-pass optimization for each video v used all LMs associated with a combination of the $\mathcal{C}(v)$ topics from video-based classification, $\mathcal{C}(x_F)$ topics from text-based classification of the first-pass transcript, and all their ancestors in the taxonomy.

Topic selection strategy	First-pass transcript	True transcript
$\mathcal{C}(v) \cup \{\text{ROOT}\}$		31.27
Ancestors of $\mathcal{C}(v)$		31.28
$\mathcal{C}(v) \cup \mathcal{C}(x_F) \cup \{\text{ROOT}\}$	31.13	30.99
Ancestors of $\mathcal{C}(v) \cup \mathcal{C}(x_F)$	31.07	30.98
Supervised NN (k=10)	31.24	31.13

Table 2: WER for unsupervised optimization using different topic selection strategies and optimization targets (first pass vs. true transcript).

G_{ROOT}	by replanting forests that were cut down **from my game**
NN	by replanting forests that were cut down for **binding**
Opt	by replanting forests that were cut down for mining
G_{ROOT}	**the master springs** was **on the eastside** you know the ignition
NN	**the master springs** was only the start you know the ignition
Opt	damascus **springs** was only the start you know the ignition

Table 3: Selected fragments of system-generated transcriptions from the test corpus. Boldfaced words indicate errors.

absolute WER improvements (approx. 3.5 over F, 1.0 over G_{ROOT}) at $p \leq 0.05$ when averaging at the utterance level and the video level. 65% of videos show a WER reduction of 1.3 on average over the stronger G_{ROOT} baseline using the supervised approach, while 56% of videos show an improvement of 1.7 on average with unsupervised optimization. 97% of videos improve over the first-pass F baseline using either adaptation approach. Table 3 shows some examples of the type of errors addressed by the adaptation techniques.

Table 2 compares WER results for optimizing towards the first-pass transcripts and human-generated gold transcripts on the test corpus. Comparing just approaches for optimization towards true transcripts, we observe that the inclusion of topics from $\mathcal{C}(x_F)$ significantly improves WER. We speculate that this is due to the topic selection better fitting the actual target domain and bridging the gap between the semantics of categories for videos and categories for text. This may also account for the relatively weaker performance of a hybrid method which uses the supervised NN technique for selecting LMs: it doesn't make use of $\mathcal{C}(x_F)$ and only indirectly considers $\mathcal{C}(v)$.

Finally, the relatively small gains seen when optimizing to the true transcript instead of the first pass (although significant under Wilcoxon's signed rank test at $p \leq 0.05$) appear to support the hypothesis that noise in first-pass transcripts does not preclude a reasonable topic characterization. This is also seen in Figure 4 which shows only a slight average increase in WER for utterances whose first-pass WER was near perfect (0–10), and a decrease in all other cases.

5. Discussion and Conclusion

Our evaluation primarily considered basic scenarios but parametric adjustments will likely result in further WER gains. For instance, the interpolated rescoring procedure ignores first-pass lattice weights assigned by F in order to make rescoring dominate the evaluation, but we have observed that permitting a tradeoff between F and H_λ generally outperforms either standalone model. Similarly, tuning k and β would almost certainly be beneficial. Analysis of per-utterance results, as in Figure 4, shows that optimization over first-pass transcripts can perhaps be further improved by leveraging the confidence scores associated with each hypothesis; we intend to explore this further.

In summary, we present two strategies for on-demand LM adaptation for video transcription via lattice rescoring. These approaches employ (1) Google's general-purpose taxonomy of topics, (2) text classifiers to build topic-specific LMs from web data, and (3) video classifiers to determine the topic content of videos. Our first approach utilizes training examples and a nearest-neighbor algorithm to predict LM interpolation weights for a new video. The second approach optimizes the interpolation weights directly using the transcripts from first-pass decoding. In addition to being simple to implement, the unsupervised technique achieves the best performance in our experiments: we obtain a WER of 31.1, an absolute improvement of 3.6 over a single-pass baseline and a significant absolute improvement of 0.8 over rescoring with a single huge LM trained on all our data. Relative to the best (oracle) WER obtainable from rescoring (26.6), the latter improvement amounts to a 14.7% reduction in WER. Further analysis suggests that these adaptation techniques are fairly robust and well-suited to the YouTube corpus.

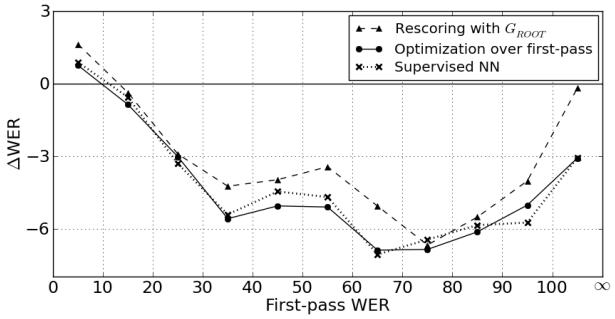

Figure 4: Average absolute change in WER for test utterances after rescoring, binned by WER after first-pass decoding.

6. References

[1] R. Florian and D. Yarowsky, "Dynamic nonlocal language modeling via hierarchical topic-based adaptation," in *Proc. of ACL*, 1999, pp. 167–174.

[2] R.M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 30–39, 1999.

[3] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *European Conference on Speech Communication and Technology*, 1999.

[4] Y.C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. of Interspeech*, 2005.

[5] B.J.P. Hsu and J. Glass, "Style & topic language model adaptation using HMM-LDA," in *Proc. of EMNLP*, 2006, pp. 373–381.

[6] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for Mandarin broadcast conversation transcription," in *Proc. of Interspeech*. 2006, ISCA.

[7] A. Heidel, H. Chang, and L. Lee, "Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm," in *Proc. of Interspeech*, 2007, pp. 2361–2364.

[8] Y. Liu and F. Liu, "Unsupervised language model adaptation via topic modeling based on named entity hypotheses," in *Proc. of ICASSP*, April 2008, pp. 4921 –4924.

[9] S. Watanabe, T. Iwata, T. Hori, A. Sako, and Y. Ariki, "Topic tracking language model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 440–461, Apr. 2011.

[10] S. Khudanpur and J. Wu, "A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition," in *Proc. of ICASSP*. IEEE, 1999, vol. 1, pp. 553–556.

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

[12] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li, "YouTubeCat: Learning to categorize wild web videos," in *Proc. of CVPR*, June 2010, pp. 879–886.

[13] Y. Song, M. Zhao, J. Yagnik, and X. Wu, "Taxonomic classification for web-based videos," in *Proc. of CVPR*, June 2010, pp. 871–878.

[14] B. Ballinger, C. Allauzen, A. Gruenstein, and J. Schalkwyk, "On-demand language model interpolation for mobile speech input," in *Proc. of Interspeech*, 2010, pp. 1812–1815.

[15] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, pp. 891–923, Nov 1998.

[16] C. Silpa-Anan and R. Hartley, "Optimised kd-trees for fast image descriptor matching," in *Proc. of CVPR*, June 2008, pp. 1 –8.

[17] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, pp. 117–122, Jan. 2008.

[18] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.