A LDA-BASED METHOD FOR AUTOMATIC TAGGING OF YOUTUBE VIDEOS

Mohamed Morchid and Georges Linarès

Laboratoire d'Informatique d'Avignon, University of Avignon
Avignon, France
{mohamed.morchid,georges.linares}@univ-avignon.fr

ABSTRACT

This article presents a method for automatic tagging of Youtube videos. The proposed method combines an automatic speech recognition (ASR) system, that extracts the spoken contents, and a keyword extraction component that aims at finding a small set of tags representing a video. In order to improve the robustness of the tagging system to the recognition errors, a video transcription is represented in a topic space obtained by a Latent Dirichlet Allocation (LDA), in which each dimension is automatically characterized by a list of weighted terms. Tags are extracted by combining the weighted word list of the best LDA classes.

We evaluate this method by employing the user-provided tags of Youtube videos as reference and we investigate the impact of the topic model granularity. The obtained results demonstrate the interest of such model to improve the robustness of the tagging system.

Index Terms— audio categorization, structuring multimedia collection, speech recognition, keyword extraction

1. INTRODUCTION

Video sharing platforms have been strongly increased over the last few years. The stored collections are generally difficult to exploit due to the lack of structuring and reliable information related to the video contents. The indexing process employed by the user is relied essentially on keywords and document titles given by the users. These meta-data are often incomplete or wrong. Sometimes, users choose keywords to get a better popularity even if this set of keywords does not match fairly with the videos content.

This article proposes a method that automatically extracts keywords from the spoken content of a video. This method relies on a two-step process that respectively consists of transcribing the spoken contents by using an ASR system and of applying a keyword extraction method to the ASR outputs.

One of the major issue in such a cascading of extraction and analysis process is due to the ASR component, that frequently fails on Web data: speech recognition systems are usually trained on very large databases that are extracted from newspapers and transcriptions of meetings or news. In most of the cases, the topics, speech styles and acoustic conditions of user-generated videos are far from the ASR training conditions and the recognition precision may be very low.

Two ways are typically followed to deal with speech recognition errors. The first one consists of improving the ASR accuracy. Such an approach usually requires task-specific speech materials and costly annotations. The second way consists of improving the robustness of the analysis component to recognition errors. This article presents a robust keyword extraction strategy that remains effective when applied to highly erroneous automatic transcriptions.

Our proposal starts from the idea that lexical level is dramatically sensitive to recognition errors, and that an abstract representation of spoken contents could limits the negative impact of ASR errors on the keyword extraction component. Following this idea, we propose to estimate a topic space, by using a Latent Dirichlet Allocation (LDA), in which each document may be viewed as a mixture of latent topics. The tags are then searched into this topic-level representation of automatically transcribed videos. We expect, from such a passage through a well structured semantic space, an improvement of the system robustness to recognition errors.

The remainder of the paper is organized as follows: the related works and discussion about their relevance for ASR output processing are detailed in Section 2. The proposed approach is described in Section 3. The experimental setup and results are shown in Section 4 and are discussed in Section 5. Finally, conclusions and future work are presented in Section 6.

2. RELATED WORKS

Keyword extraction is a classical issue of natural language processing. This task from spoken documents presents difficulties due to the specificities of spoken language and to the use of ASR systems for the extraction of linguistic contents. Some works proposed high level approaches, based on ontologies and linguistic knowledge. In [1], the authors use WORDNET and EDR electronic dictionaries for proper noun extraction from meeting dialogues. This method relies on a first step of text tagging that follows a concept level scoring.

Other approaches are based on statistical models that demonstrated their efficiency on various speech processing tasks. [2] uses the LSA (Latent Semantic Analysis) technique to extract the most relevant phrases from a spoken document. In [3], the authors apply LSA to an encyclopedic database for keyword extraction.

The extraction of keywords may be viewed as an extreme form of summarization. Our approach is different from a summarization task. The authors, in [4], employs the Clustering By Committee (CBC) Model and LDA to extract a set of words that summarize a set of documents. The obtained results have demonstrated the efficiency of LDA and seems robust to recognition errors. This is a critical point of speech analysis systems, especially in adverse and unexpected conditions as in Youtube videos. Our proposal is to investigate LDA-based methods for robust tagging of video hosted by a video sharing platform, without any assumption about the video sources, acoustic quality of recording or topics.

This work was funded by the ANR project SuMACC (ANR-10-CORD-007) in CONTINT 2010 program, led by the LIA (University of Avignon).

3. PROPOSED METHOD

The global process is shown in figure 2 and consists of mapping the automatic transcription of the video into a topic space estimated by LDA. This mapping allows us to select the most representative LDA classes, considered as topics. Each of these classes is represented by a set of weighted words. The best tags are searched in the intersection of the best-classes word set.

Concretely, the proposed method proceeds with 5 successive steps:

1. off-line estimation of a LDA model on a large corpus D; this step produces the topic space T_{spc}
2. automatic transcription t of each video document v
3. representation of t, with a vocabulary $\mathbf{V}$, as a feature vector W^t
4. projection of W^t into T_{spc} and selection of a subset $S^z \subset T_{spc}$ of the best LDA classes (each of these classes being implicitly associated to a topic)
5. extraction of a subset of the best tags $S^w \subset \mathbf{V}$ from S^z regarding W^t.

The next sections describe in-depth the main parts of this process.

3.1. Estimation of a topic space

We estimate off-line a LDA model on a large corpus D; this step produces the topic space T_{spc}. In the following sections we describe this process.

Latent Dirichlet Allocation (LDA):

LDA is a generative probabilistic model which considers a document as a *bag of words*. Word occurrences are linked by latent variables determining the distribution of topics in a document. This decomposition model of documents offers good generalization abilities compared to other generative models that are commonly used in automatic language processing such as Latent Semantic Indexing (LSI) or Probabilistic Latent Semantic Indexing (PLSI) [5, 6].

All these methods require a set of data to build a global model. Our training corpus D is composed by documents from Wikipedia and the AFP (Agence France Presse) collection of newswire. These corpuses represent respectively 1 and 4.5 GB, corresponding to about 1 billion of words and 3 million of documents. These corpuses are lemmatized using the TreeTagger tool and are filtered by a stop list.

One of the critical points of LDA models lies in the number of classes. This number results from a prior choice which significantly impacts the final model: high number of classes lead to a fine-granularity model, where each class is supposed to represent a specific topic. Moreover, the estimation of LDA models is a quite heavy process.

On the other hand, a configuration of few classes leads to wide-covered classes that may be poorly relevant to precisely identify the main latent topics of a specific video.

We tested various configurations of the topic space T_{spc} by varying the number of classes (from 200 to 15000), that determines the granularity of the resulting topic model.

Topics representation:

After the estimation of the background topic model, we have to project the document in this semantic space and select the nearest topics of the document. This subset of topics is named S^z and is considered as the most representative of the main underlying idea of the document. A topic z, associated with an LDA class, is represented by a vector V^z. The *i*th ($i = 1, 2, \ldots, |\mathbf{V}|$) coefficient of this vector represents the probability of the word w_i knowing the topic z:

$$V_i^z = P(w_i|z)$$

Even if the dimensionality of these vectors is high (equals to the size of the lexicon associated to the LDA training corpus), most of the coefficients are close to zero, corresponding to words that are poorly related to the topic. For simplicity and efficiency, we limit the representation of a class to the twenty words of maximum weight.

3.2. Automatic transcription

Videos document can not be projected in T_{spc}. Thus, an automatic speech recognition (ASR) system (in our case, the LIA system *Speeral* [7]) is used to extract spoken contents. This text document is processed as a bag of words in order to obtain a feature vector W_i^t. The details of *Speeral* are presented in Section 4.2.

3.3. Video representation

Let C be a corpus of n_d documents d and $|\mathbf{V}|$ ($|\mathbf{V}|$ is about $2,8$ million of unique words in our experiments) be the vocabulary size. The corpus can be represented by a matrix of size $n_d \times |\mathbf{V}|$. This representation permits to evaluate the Inverse Document Frequency (IDF) for vocabulary words. A Youtube video d is automatically transcribe to a document t. Each transcription t can be represented as a point of $\mathbb{R}^{|\mathbf{V}|}$ by a vector W_i^t of size $|\mathbf{V}|$ where the *i*th feature ($i = 1, 2, \ldots, |\mathbf{V}|$), combines: the Term Frequency (TF), the Inverse Document Frequency (IDF) and the Relative Position (RP) [8] of a word w_i of t:

$$W_i^t = tf_i \times idf_i \times rp_i$$

where

$$tf_i = \frac{|\{w_i : w_i \in t\}|}{|t|},\ idf_i = log\frac{|C|}{|\{d : w_i \in d\}|}, rp_i = \frac{|t|}{fp(w_i)}$$

Here, $|\cdot|$ is the number of elements in the corresponding set and fp_i is the position of the first occurrence of w_i.

3.4. Projection of W_i^t into T_{spc}

Each video v has a feature vector W_i^t. Thus, the subset S^z of the nearest topics of v can be determined by using a similarity measure between its representation W_i^t and each topic $z \in T_{spc}$. Then, the topics with the best similarity regarding the document transcription t are kept in S^z. The similarity between a video transcription d and the semantic space T_{spc} is evaluated for each topic z of T_{spc} by using a simple cosine metric:

$$cos(t, z) = \frac{\sum\limits_{w_i \in t} V_i^z . W_i^t}{\sqrt{\sum\limits_{w_i \in z} {V_i^z}^2 . \sum\limits_{w_i \in t} {W_i^t}^2}}. \quad (1)$$

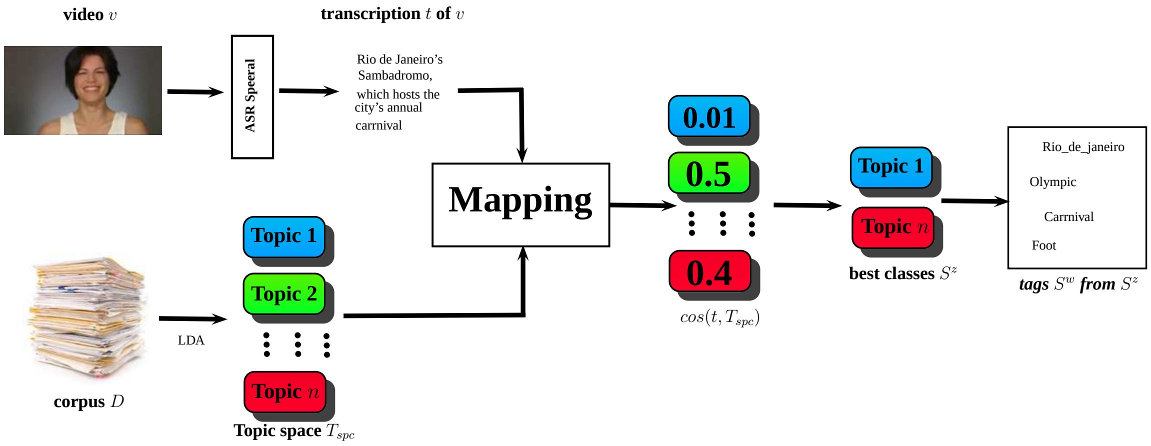

Fig. 1. Overall process of the automatic tagging method by LDA-based representation of speech contents

3.5. Extraction of the best tags S^w from S^z

Each video v is associated with a subset of topics S^z. The next step is to extract a subset S^w of the most representative words from S^z. In our experiments, we compare our method with the following subsets of words: TF-IDF-RP and S^w.

Keywords extraction with TF-IDF-RP:

This method allows a simple extraction of the n most representative words in W_i^t of a transcription t. In our experiments, the IDF is equally estimated by $100,000$ news from Wikipedia and AFP. The system extracts 10 words that have obtained the highest TF-IDF-RP score [9].

Keywords extraction by combination of latent topics (S^w):

At this point, our goal is to extract the best keywords from the projection of ASR outputs into a topic space. Our strategy consists in selecting the top topics of a document, $|S^z|$ being empirically fixed to 100. Considering a topic as a small set of weighed words, tags are searched in the intersection $S^w = \{s(w_1), s(w_2), \ldots, s(w_{|S^w|})\}$ of the main word set topics S^z. Word ranking is obtained, in the intersection, by combining the topic relevance score (cosine Eq.1) of the topic and the weight of the word in the topic to obtain a score s for all words w of $z \in S^z$:

$$s(w) = \sum_{z \in S^z} cos(t, z) \times P(w|z)$$

where $P(w|z)$ represents the probability of w knowing the topic z and $cos(t, z)$ the similarity between z and d the document.

4. EXPERIMENTS

4.1. Evaluation corpus

The test corpus is composed of 138 French videos from the Youtube platform. The average number of tags per video is about 14. A corpus of recent French Wikipedia articles is composed of about 3.7 million of articles. All notes and bibliographical references were removed from this corpus. Finally, this corpus contains around 26 million sentences for a total of about 333 million word occurrences. The Wikipedia vocabulary contains 2.8 million of unique words (at least one occurrence in the corpus). This Wikipedia corpus was used to estimate a set of spaces from 400 to 15,000 topics.

The keyword extraction is considered as a word detecting task, where the user tag set is used as unique reference. Results are estimated classically in terms of *precision at n* tags (recall being inadequate, considering that we produce as many tags as the reference includes). These videos are first processed by the LIA ASR system *Speeral* [7].

4.2. Speech recognition system

This system uses classical acoustic and language models. Acoustic models are context-dependent HMMs with decision free state-tying. These genre-dependent models are estimated on about 250 hours of French broadcast news. A 4-gram language models is estimated from various text sources, mainly from the French newspaper *Le Monde* (about 200 million words) and the *GigaWord* corpus (about 1 billion words) and the manual transcription of acoustic corpus (about 2 million words). The search engine processes two passes, the second one uses speaker-adapted acoustic and a 4-gram language model.

In order to estimate the performance of *Speeral* on such Web data, we transcribed 10 of the 100 test videos (randomly chosen), corresponding to about 35 minutes of speech. On this relatively small sample set, the system obtains a 63.7% WER. As expected, the WER is very high: user provided videos are highly variable and poorly controlled in terms of topics, speech styles, acoustic conditions and acquisition materials, etc. Here, we focus on the robustness of the keyword extraction system to the recognition errors.

5. RESULTS AND DISCUSSION

Figure 1 shows that the proposed method is significantly better than the classical TF-IDF-RP based approach. Indeed, we can see that the precision is twice higher with the LDA approach, this result validating the initial motivation: use of an abstract representation level to limit the negative effects of ASR errors (see table 1).

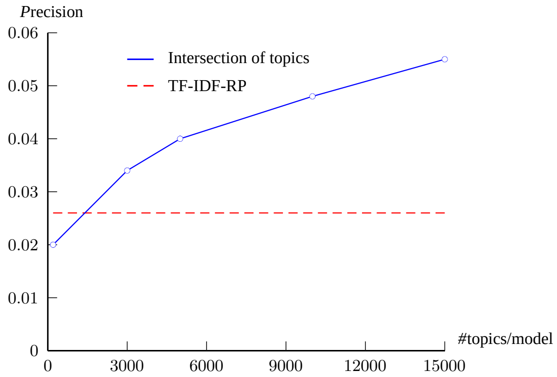

Fig. 2. Precision at n tags of the videos tagging according to the dimensionality of the LDA model.

The reported results show the impact of topic model granularity: the thinner the model is, the higher the precision is. This result suggests that the number of classes should be increased over 15,000 (our largest topic number configuration), but this progression meets a complexity problem: LDA is a costly procedure and the targeted configuration must be tractable, both in terms of CPU time and of amount of training data required for a robust estimate of topic distributions. Nevertheless, increasing the text database is probably one of the promising way to improve the precision of our tagging system.

#Topics	Tag set coverage(%)	Precision(%)
TF-IDF-RP	27.6	2.6
200	27	2
3,000	64	3.4
5,000	71	4
10,000	73	4.8
15,000	75	**5.5**

Table 1. Precision of the LDA-based tagging method according to the topic-model granularity.

Table 1 presents the results (in terms of precision) obtained with the TF-IDF-RP baseline system and with the proposed LDA approach by varying the number of classes (from 200 to 15,000). Another interesting point related to the impact of the granularity is the tag-set coverage of the LDA model, that is reported in the second column of the table 1. It indicates the number of user-tags that are, at least, in one of the 20-best words of the LDA classes. This index clearly shows a limitation of the proposed method that is due to the difficulty of the prior modeling of topics.

Table 1 shows that the use of a semantic space permits to find some tags that do not appear in the transcription d. If a word w belongs to the nearest topic but does not belong to the video, it could be selected with the intersection method and not with the TF-IDF-RP which can only find the words contained in the transcription of the video. The TF-IDF-RP method obtains the same results than the intersection method when the same number of tags can be found ($\approx 27\%$).

In spite of this gain provided by the LDA-based approach, the absolute results remains low. The best configuration reaches 5.5% of precision, while the conventional approach is about 2.6%. Nevertheless, as mentioned in the introduction, the user tags are not the perfect references (if it is possible) and they probably have something unpredictable, depending from the specific up-loader point of view, culture, intents, etc. Table 2 presents an example that shows how much subjective may be the user tagging.

Method	Tags
User's tags	iranian atomic netanyahou livni intel arab
TF-IDF-RP	true ehoud **iranian** jerusalem security iran
LDA-classes	**iranian** foreigner true iran **atomic** jerusalem

Table 2. Example of tags extracted from the topic space, compared to the one obtained by TF-IDF-RP.

6. CONCLUSION AND FUTURE WORK

We proposed a video tagging method that represents a video transcription as a mixture of topics by using the LDA technique. Keywords are extracted from this decomposition into a topic space.

Our experiments demonstrated that such a mapping of a noisy document into a well-structured semantic space improves the robustness of the tagging system to recognition errors. Even if the proposed method significantly outperforms a conventional TF-IDF-RP with a *relative position* based approach, the absolute performance in predicting user provided tags remains low. This is due to subjectivity and imprecision in the human tagging and the high WER of the video transcriptions. These first experiments demonstrated the interest of the passage through an intermediate representation. This way seems interesting for the discovery and the characterization of new or emerging concepts in continuous streams of information. We are currently developing the proposed method in this direction.

7. REFERENCES

[1] L. van der Plas, V. Pallotta, M. Rajman, and H. Ghorbel, "Automatic keyword extraction from spoken text. a comparison of two lexical resources: the edr and wordnet," *Arxiv preprint cs/0410062*, 2004.

[2] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1279–1296, 2000.

[3] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi, "Keyword extraction using term-domain interdependence for dictation of radio news," in *COLING'98*. Association for Computational Linguistics, 1998, pp. 1272–1276.

[4] A. Celikyilmaz and D. Hakkani-Tur, "Concept-based classification for multi-document summarization," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5540–5543.

[5] R. Kubota Ando and L. Lee, "Iterative residual rescaling: An analysis and generalization of lsi," 2001.

[6] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR conference*. ACM, 1999, pp. 50–57.

[7] G. Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *TSD'07*, 2007, pp. 302–308.

[8] G. Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.

[9] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf*idf, lsi and multi-words for text classification," *Expert Systems With Applications*, 2010.