MediaEval 2012 Tagging Task: Prediction based on One-Best List and Confusion Networks

Yangyang Shi ‡,Martha A. Larson †, Pascal Wiggers ‡, Catholijn M. Jonker‡
‡Interactive System Department
†Multimedia Information Retrieval Lab
Delft University of Technology
yangyangshi@ieee.org

ABSTRACT

In this paper, we describe our participation in the MediaEval 2012 Tagging task, which requires us to predict the genre labels for videos. We use three different types of models: a conventional support vector machine (SVM), probabilistic generative dynamic Bayesian networks (DBN) and probabilistic discriminative conditional random fields (CRF) to classify the videos based on speech transcripts. As the baseline, SVM uses unigram, bigram and trigram features in a bag-of-words strategy. We also apply the DBN and CRF to take advantage of sequence relationship information in the one-best hypothesis. For confusion networks, the possible words at each time slice are applied as the observation attributes for CRF.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]

General Terms

One-best list, Algorithm, Model, Features

Keywords

Support vector machine, Conditional random field, Dynamic Bayesian networks, Text classification

1. INTRODUCTION

Due to the huge amount of multimedia content avaliable on the web, automatic classification systems are needed to support users to more easy discover information. In the MediaEval 2012 Tagging Task, internet video episodes have to be tagged according to their genres [3].

We start with the assumption that information contained in the spoken channel of the video is indicative of genre. Different lexica and syntactic structures occur in different genres. For this reason, we work from the assumption that spoken content is indicative of genre and can be used for genre classification. In order to fully exploit automatic speech recognition transcripts, we evaluated the performance of support vector machines (SVM), Dynamic Bayesian networks (DBN) and conditional random field (CRF) in genre classification. An SVM maps the transcripts to high dimension vectors. The presentation used for input to SVM considers all features to be independent of each other. It assumes that the sequencial relationship from the words doesn't influence the genre. In DBN and CRF, each transcript is treated as a sequence of dependent variables. These two models have an advantage over SVM in that they also take the relation among words into consideration for genre classification.

MediaEval 2012 Workshop, October 4-5, 2011, Pisa, Italy

2. MODELS

Our participation focused on the one-best hypotheses and confusion networks which are provided by LIUM [2]. We applied the non-probabilistic SVM model, the probabilistic generative DBN model and the probabilistic discriminative CRF models in MediaEval 2012 genre tagging task.

2.1 Multi-class support vector machine (1best-SVM)

In representation of the one-best hypotheses for SVM, we extracted unigram, bigram and trigram features from the one-best development data set. The features space dimension is determined by the number of unigrams, bigrams and trigrams occuring more than three times in the training data. In calculating feature values, rather than using the direct frequency counts, we applied the modified version of the term frequency inverse document frequency (tf-idf) metric.

$$\text{weight}(i,j) = \begin{cases} (1+\log(\text{tf}_{i,j}))\text{idf}_i & \text{tf}_{i,j} > 0, \\ 0 & \text{tf}_{i,j} = 0, \end{cases} \quad (1)$$

The term frequency $\text{tf}_{i,j}$ is the number of times term i appears in one-best transcript j. The document frequency df_i is the number of one-best transcripts that contain term i. Inverse document frequency $\text{idf}(i)$ can be calculated by:

$$\text{idf}_i = \log(\frac{N}{\text{df}_i}),$$

where N is the total number of documents. The tf-idf weight is the combination of $\text{tf}_{i,j}$ and idf_i.

The classification into multiple tag classes is achieved by a multi-class SVM [1] using linear kernel with cost parameter $C = 0.5$, which is obtained by grid search over a randomly selected small sample from the development data.

2.2 Dynamic Bayesian networks (1best-DBN)

In representation of one-best hypotheses for DBN, each transcript was treated as a sequence of words. At each position of the sequence, current word information, previous two words information and structure information (sentence, position, n-gram position) were extracted.

Dynamic Bayesian network can model probability distributions of semi-infinite sequences of variables that evolve over time. A dynamic Bayesian network can be represented by a prior model $P(X_1)$ and the following two slice temporal Bayesian network:

$$P(X_t|X_{t-1}) = \prod_{i=1}^{N} P(X_t^i|Pa(X_t^i)) \quad (2)$$

where $\mathbf{X}_t$ is the set of random variables at time t and X_t^i is the ith variable in time step t. $Pa(X_t^i)$ are the parents of X_t^i. In graphical model, parents are the sources of directed edge connecting X_t^i.

In genre classification, a DBN model [4] from a predefined genre label set T makes a classification decision by seeking a genre label t which maximizes the posterior probability $P(t|w_1, w_2, ..., w_n)$ of the label given the related sequence of words:

$$t^* = \arg\max_{t^i \in \mathbf{T}} \{P(T = t^i|w_1, w_2, ..., w_n)\}, \quad (3)$$

$$= \arg\max_{t^i \in \mathbf{T}} \{\frac{P(T = t^i) \times P(w_1, w_2, ..., w_n|T = t^i)}{P(w_1, w_2, ..., w_n)}\}. \quad (4)$$

In MediaEval 2012, we applied the interpolated trigram dynamic Bayesian networks to do tagging task. The probability of current word given previous history is an interpolation of the probability of current word given previous two words and the probability of current word given the topic. Ninty percent of development data was selected for training the probabilities in DBN. The rest was used to train the trigram interpolation weights.

2.3 Conditional random field

CRF was originally used for sequence data labelling. In the Tagging task, all the words in one transcript were tagged by one genre label. Each position of the sequence is characterized by an attribute vector, which is constituted by the current word, previous two words and next two words. So the conditional probability of the label sequence $(t_1, t_2, ...t_n)$ given the word sequence $W = (w_1, w_2, ..., w_n)$ is:

$$P \propto \exp(\sum_{j}^{J}(\lambda_j f_j(t_{i-1}, t_i, W, i)) + \sum_{k}^{K}(\mu_k s_k(t_i, W, i))), \quad (5)$$

where $f_j(t_{i-1}, t_i, W, i)$ is a transition feature function of the word sequence and the label at position i and $i-1$ in the label sequence; $s_k(t_i, w, i)$ is a state feature function of position i label and word sequence; J and K represents the features size; λ_j and μ_k are the weight for each function.

3. EXPERIMENT AND RESULTS

The rank of each genre label videos in SVM was determined by output probability by the linear logistic regression. In DBN and CRF, the rank is directly determined by the probabilities of the genre label given the transcripts.

As the official results shown in Table 3, the baseline SVM using 1-best list got MAP 0.23. The DBN using 1-best list got the highest MAP 0.25. However, the CRF using 1-best list only obtained a MAP of 0.10, and using confusion networks only achieved a MAP of 0.09 . Even though CRF has been shown in the literature to achieve good performance for sequencial part-of-speech tagging, it didn't yield good results in our Tagging task. The probable reason is that in part-of-speech tagging, each item is a word, in our case each item is a document.

Table 1: Tagging task results

Models	MAP
run2-one-best-SVM	0.23
run2-one-best-DBN	0.25
run2-one-best-CRF	0.10
run2-cf-CRF	0.09

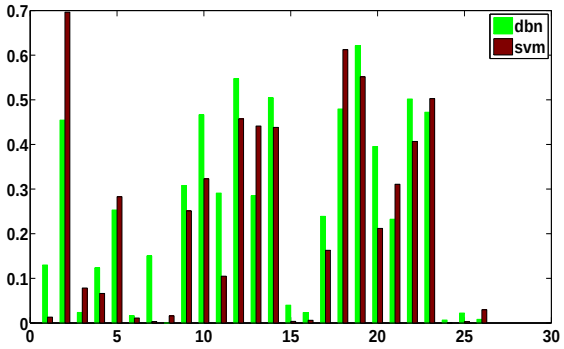

Figure 1: Average precision comparision between DBN and SVM. The horizontal axis represents the label, the vertical axis represents the average precision.

The figure 3 shows the comparisons of DBN and SVM in each category. In most of categories, DBN outperforms SVM. Even in some categories, SVM has a high average precision, in fact it had lower precision than DBN. In "autos and vehicles" the second category, red bar got average precision almost 0.7 . In fact, this category only had 11 samples. The svm predicted 154 results in which 9 were relevant. But the DBN predicted 5 results which were all relevant.

4. CONCLUSIONS AND FUTURE WORK

As demonstrated by our official results on this task taking the word sequence order helped improving the classification. In the future, we can investigate the other advanced language models performance in the Tagging task.

5. REFERENCES

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[2] A. Rousseau, F. Bougares, P. Deleglise, H. Schwenk, and Y. Esteve. Lium systems for the iwslt 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, San Francisco (USA), 8-9 Sept 2011.

[3] S. Schmiedeke, C. Kofler, and I. Ferrane. Overview of MediaEval 2012 Genre Tagging Task. In *MediaEval 2012 Workshop*, Pisa, Italy, October 3-4 2012.

[4] Y. Shi, P. Wiggers, and C. M. Jonker. Dynamic bayesian socio-situational setting classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.