Constructing of Multimedia Resources for Second Language Teaching Based on Intelligent Information Processing of Movie Resources

http://dx.doi.org/10.3991/ijet.v8i5.3045

Hua Liu
Jinan University, Guangzhou, China

***Abstract*—Based on the topic models database and time density of caption's dialogue, designed a adaptive, iterative incremental learning algorithm, which could meanwhile carry out topic detection, words clustering and calculating words' used degree. Lastly, based on words' used degree, by calculating degree of difficulty of caption's dialogue, graded videos' corresponding to captions, constructed a large multimedia database for oral Chinese teaching.**

***Index Terms*—Videos, Intelligent Information Processing, Chinese Teaching.**

I. INTRODUCTION

"Colloquial, communicative, topicalized and interesting" Chinese teaching based on scene communicative function is the development trend of Chinese teaching. Video resources can meet these four requirements. Video subtitles are continuous oral text of dialogue flow and a collection of topics based on various communicative situations. Meanwhile, videos are multimedia and interesting audio-visual resources.

This article, based on video resources (subtitle texts and videos), employs Topic Detection and Tracking (TDT), conducts topic detection, word clustering and grading with word clustering and application frequency computing method, segments video resources, grades them based on difficulty, and constructs a multimedia teaching resource databases for Chinese oral topics of scene communicative function, including topic database, topic vocabulary database, large-scale subtitle text database graded based on difficulty corresponding to the topics, and video clip database.

II. A REVIEW OF RELEVANT OVERSEAS AND DOMESTIC RESEARCHES

A. Topic Detection and Tracking (TDT)

Most topic detection algorithms are conducted based on clustering algorithm, by describing news and topics with vector space model, calculating the similarity and clustering them with certain strategy. Popular algorithms are hierarchical clustering based on average grouping and single-pass clustering.

Topic tracking methods are roughly divided into two types: those based on text categorization methods (such as KNN and decision tree approach) and those based on information retrieval methods (such as Rocchio). Unsupervised Adaptive topic tracking (ATT) with self-learning ability has gradually become the new research trend in TDT field, mainly including two aspects: methods based on content (such as GE R&D) and methods based on statistics (DRAgon, Umass and LIMSI).

We believe that TDT performance can be improved in the following two aspects:

1. Methods based on linguistic discourse analysis. Linguistic characteristics of processing objects of a certain type, such as video subtitles, are analyzed, in terms of time sequence, reference, named entities, topic switching mechanism and the like. Character extraction and representation are optimized and topic/report description model is further optimized.

2. Adaptive learning strategy of topic model. Topic detection and tracking effect are improved, with known expert knowledge base as heuristic knowledge and iterative and incremental learning strategy employed. Two key factors are how to reduce impact of topic shifting caused by pseudo feedback in the process of learning and how to improve recall rate while resolving topic shifting and raising accuracy rate.

Different from independent event news, video subtitles are continuous dialogue flow of weak event pattern, so topic segmentation and detection are more difficult to do. At present, there is no report on topic detection and tracking for video dialogues in educational circles.

B. Word categorization and clustering

Noted knowledge bases, such as WordNet and HowNet, are mainly manually constructed by experts. Rule-based methods can be used to gain field words from large-scale classified corpus with matching method and the manually constructed template. Statistics-based methods fall into three types. First, distance (or similarity) between different elements is expressed by heuristic measurement. Second, distance measurement (likelihood function, for example) and category total of cluster results are set with the statistical model. Third, distance measurement is set with the statistical model in the same way, but some measurement (such as perplexity and MDL) is added to control increase and decrease of the category number during the clustering process. The first algorithm is not as effective as the second or the third.

This research needs to gain topic word cluster by clustering each topic. Topics, different from each other, are classified into different types. We believe that it is an effective approach of improving word clustering effect to

employ distinctive features of different topic corpuses and distribution differences of words in topics.

At present, there is no research on word cluster and word cluster grading for Chinese teaching construction, other than researches by the writer.

C. *Word application frequency calculation and word grading*

The essence of word grading lies in word application frequency calculation. Relevant solutions are Word Choice Function proposed by Liang Nanyuan and Liu Yuan, application frequency formula proposed by Yin Binyong and Fang Shizeng, circulation degree formula proposed by Zhang Pu, scatter coefficient and application frequency calculation formula proposed by Sun Maosong and so on.

Guiding documents for word grading of Chinese teaching are National Syllabus of Graded Words and Characters for Chinese Proficiency (2001) and Syllable, Chinese Characters and Phrase Grading for International Chinese Education (2010). However, these documents are merely about grading, providing no data basis for word grading, such as word frequency and application frequency.

Application frequency calculation formulae are all strongly correlated to corpus composition and categorical distribution. Different results can be gained for different corpora. We believe that application frequency is closely related to spatial and temporal word distribution. Temporal word distribution is reflected in constant state in development, and spatial word distribution is reflected in distribution uniformity of user types and application scopes. Therefore, it is the key for calculating word application frequency how to effectively simulate special and temporal word distribution uniformity.

III. OVERALL ALGORITHM DESIGN

A. *Movie resource bank construction*

A movie resource bank, including subtitle dialogue text (containing start and stop time information) and video files (with merely samples for experiment are downloaded), is being constructed, based on the principle of contemporariness, lifestyle and multi-theme of films and TV programs. The movie resource bank is expected to contain 1,000 films and TV programs, in which the subtitle corpus contains about 200,000,000 words. By far, a part of the movie subtitle corpus has been completed.

B. *Expert construction of topic resource database and topic model base*

With reference to the Table of Suggested Topics and Content for Chinese Teaching and the Table of Examples of Topics and Content of Chinese Teaching in the International Chinese Teaching Universal Curriculum and based on large-scale oral Chinese, intensive reading textbooks and the topic and function syllabus, Chinese teaching experts have established a topic database and most commonly used seeded words for each topic, set the initial weight, and constructed the initial TDT topic model base.

C. *Overall algorithm of topic detection and word clustering of heuristic, iterative and incremental learning*

1. Goal description

(1) Topic detection and segmentation

Since Chinese teaching topics are known, with the topic model constructed in advance, the nature of topic detection of this subject is actually topic tracking. For each Chinese teaching topic in the International Chinese Teaching Universal Curriculum, topic tracking needs to be done in the subtitle corpus containing 1,000 movies, so as to find the dialogue flow collection corresponding to the topic and segment the dialogue flows.

(2) Word clustering and grading

Basic resources of communicative Chinese teaching are topics and topic word clusters (topic word cluster for Ordering A Meal includes “restaurant, menu, ordering dishes and so on”). Therefore, the goal of word clustering is to collect common topic word clusters for each topic.

In addition, words in a word cluster differ from each other in terms of difficulty, and they need to be graded based on application frequency.

2. Overall algorithm design idea: integral algorithm of heuristic, iterative and incremental learning

We have preliminarily designed a heuristic integral algorithm of topic detection and word clustering based on iterative and incremental learning strategy:

①. Pretreatment. Time interval between two subtitle dialogue sentences exceeding a certain threshold value (6 seconds, for example) can be set as natural segmentation mark between topics, with which movie subtitles can be roughly segmented. Word segmentation and part of speech tagging are automatically done.

②. Topic tracking is conducted (refer to the following content for more details), with adaptive topic tracking method based on dialogue density time column model, and with the initial topic model base as heuristic knowledge. Dialogue flows related to the topic are extracted and the dialogue flow database of the topic is updated.

③. Based on dialogue flow databases, word clustering is done (refer to the following content for more details) with text categorization feature extraction method, topic word cluster (containing weight) of the topic is extracted, the topic word cluster is then added to seeded words, so as to expand seeded words, update the weight and reconstruct the initial topic model base.

④. ② and ③ are iterated, until the extracted topic word cluster changes little or the dialogue flow database increases little.

The entire algorithm process is a learning process of mutual promotion and correction. Dialogue flow databases gained through topic detection are taken as word clustering data, and word clustering results are then taken as knowledge base for topic detection. Meanwhile, new word clustering results function as the basis of weight correction of previous knowledge base, and dialogue flow databases are adjusted and updated accordingly. In addition, this is also a process of reviewing learning and incremental learning. The initial knowledge base is reconstructed and expended, after word clustering for several times. The first dialogue flow base gained through

topic detection is updated and incremented after topic detection for several times.

Previous ATT learning mainly adopted sequential one-way learning with follow-up report. Our algorithm employs iterative reviewing learning, which requires cautious and accurate learning. Less knowledge is learned each time, but the goal can be finally achieved through iteration and reviewing. Therefore, high filtration indexes are adopted in step ② and step ③, to reduce impact of topic shifting caused by pseudo feedback (added subsequent dialogue flow and word cluster, not necessarily correct) in the process of learning and improve accuracy rate. Meanwhile, recall rate is continuously increased through iterative reviewing learning.

3. ATT based on dialogue density time column model

Topics in movie subtitle dialogues tend to have the following characteristics:

Scene change is usually needed at the beginning of a dialogue, with a longer time interval from the previous topic. A dialogue usually starts from a short sentence (greeting or opening remarks). With the dialogue going deeper, sentence length increasing and time interval shortening, dialogue density gradually increases, so that many short dialogues occur within a short period of time or dialogues even overlap in the same period of time. When communicative tasks are completed, sentences gradually get short again (saying goodbye, closing), with a longer time interval to the next topic. The dialogue density characteristics can be simulated through the following time column model:

A unit (20 seconds, for example) is set and the total of lengths of all the sentences in the unit form a column, indicating dialogue density. The higher the dialogue density, the higher the column is. There might be no column, or there might only be an extremely short column between two dialogues. Within a dialogue, it may be "short column—long column—short column".

Topic tracking can be completed with integral algorithm of heuristic, iterative and incremental learning introduced in the above content, based on topic segmentation function of the time column model and the initial topic model base.

4. Word clustering based on text categorization feature extraction and application frequency calculation

Word clustering can be done with feature extraction method in text categorization, based on categorical attributes of the dialogue flow databases classified on the basis of topics.

Word application frequency refers to the frequency of a word appearing in each topic, and it reflects whether a word is frequently used and provides the basis for word learning order. Word application frequency can be simulated and calculated with the used of distribution uniformity.

Please refer to Section 3.1 for specific algorithm of word clustering and application frequency calculation.

D. Topic word cluster grading based on vocabulary level syllabus and application frequency

After topic word cluster is gained through topic clustering, words in the word cluster are divided into four levels, based on the National Syllabus of Graded Words and Characters for Chinese Proficiency and Syllable, Chinese Characters and Phrase Grading for International Chinese Education and the words on each level are then arranged in descending order based on their application frequency. If a certain word does not appear in either vocabulary, it is arranged at the end, as beyond the syllabus, based on its application frequency.

E. Construction and difficulty grading of video clips corresponding to topics

After topic clustering, each topic corresponds to multiple dialogue flows, and each dialogue flow has its corresponding start and stop time information. Therefore, corresponding video clips can be conveniently segmented based on this start and stop time information.

Foreign students of different Chinese levels need video resources of different difficulty levels. Therefore, video clips need to be graded based on difficulty. Video clips correspond to sentence groups in dialogue flows. As a result, the problem is now sentence grading based on difficulty. The mean value of sentence difficulty in a sentence group of a certain dialogue flow is calculated, so that the video clips can be graded based on difficulty.

Please refer to Section 3.1 for specific algorithm of sentence difficulty grading.

IV. EXPERIMENT AND ANALYSIS

By far, 100 video subtitle corpuses and 10 video topic and topic model bases have been constructed, and relevant rough segmentation, word clustering and application frequency calculation algorithm and procedures have been completed.

A. Word clustering and realization of application frequency calculation algorithm

1. Realization of word clustering algorithm

TFIDF formula itself reflects words' ability to distinguish document types and topics. If feature vectors are arranged in reverse order after calculation of TFIDF value of a certain topic, words with high topic-distinguishing ability will be at the front. Therefore, word clustering can be conducted with feature extraction method in text categorization, based on categorical attributes of dialogue flow databases classified on the basis of topics.

Weight of a feature word in a certain topic is mainly influenced by two factors: occurrence frequency of the word in the current topic (application frequency) and difference between occurrence frequencies of the word in different topics (distribution uniformity). The calculation formula is provided as follows:

$$w(w_i,c_j)=\sqrt{\sqrt{\sum_j\left(p_{ij}-\overline{p_i}\right)^2\Big/\sum_j pi_j}\times\left(\log(N(w_i)/N)\right)^2}\times\sqrt[n]{p_{i_j}}$$

In which, $p_{i_j}=T_{ij}/L_j$, L_j is the total of occurrence numbers of all the words contained in type c_j, T_{ij} is occurrence number of the word i in type c_j. $\overline{p_i}=\sum_j p_{ij}/m$, in which m is the number of types, $N(w_i)$ refers to occurrence number of word w_i in the

training corpus and N is the total of occurrence numbers of all the words in the training corpus; n>=1。

The word clustering system is highly accurate, with accuracy rate of the first 2,000 words as high as 95.8%

2. Realization of word application frequency algorithm

Word application frequency is closely related to word distribution, mainly reflected in time axis and space axis. Temporal word distribution is reflected in constant state in development, and spatial word distribution is reflected in distribution uniformity (of user types and application scopes). Frequently used words are words frequently used in a certain period of time by broad population in broad field.

Word application frequency can be simulated, with distribution uniformity (IWF*IWF and product of variance and generative capacity of a word itself:

$$U_{w_i} = \sqrt{\sqrt{\sum_j \left(p_{ij} - \overline{p_i}\right)^2} \Big/ \sum_j pi_j} \times \left(\left|\log(\overline{p_i})\right|\right) \times \left(\log(N(w_i)/N)\right)^2 \times \left(\left|\log(p_{di})\right|\right)$$

In which, p_{di} is the frequency of left or right collocation of a word, to simulate generative capacity of a word.

Word application frequency has been calculated in the constructed super-sized balanced corpus (sampling from 1919 to 2010, containing 15 categories, 2,300,000,000 words). Each word is endowed with a weighted value of application frequency.

3. Realization of sentence application frequency algorithm

The following factors influence sentence difficulty: sentence length, mean value of application frequencies of all the words in a sentence and application frequency of the most difficult word in a sentence (learners may find it difficult to understand the sentence meaning due to an infrequently used word). Sentence difficulty can be simulated through these three factors.

Sentence difficulty is simulated by "mean value of application frequency of all the words in the sentence + application frequency of the most frequently used word in the sentence + sentence length (number of words in the sentence) × 3".

The computational formula is provided as follows:

$$L_s = \sum_{i=1}^{j} U_{w_i} \Big/ j + U_{w_i \max} + j \times 3$$

In which, L_s represents difficulty of sentence S, $\sum_{i=1}^{j} U_{w_i} / j$ means the mean value of application frequency of all the words in the sentence, $U_{w_i \max}$ refers to application frequency of the word in the sentence of the highest application frequency, j is the sentence length, namely, the number of words in the sentence, and indicates application frequency of the word Wi.

B. Topics and topic model base construction

Topic database and topic vocabulary database for teaching of Chinese as a foreign language have been constructed, based on large-scale corpus of textbooks and syllabuses for teaching Chinese as a foreign language. 5 types of topics are finally determined: daily life topics, business topics, education topics, culture topics and special topics. Each type contains 40 level-one topics and 50 level-two topics, as are provided in the following table (with examples):

TABLE I.
EXAMPLES OF TOPIC DATABASE (SOME TOPICS OF THE DAILY LIFE CATEGORY)

Level-one topics	Level-two topics
Personal information	Basic personal information, hobbies and special skills, character
Housing	Buying a house, paying house loan, renting a house and accommodation
Transportation	A、 (Traveling) by train, by air, by taxi, by ship, by metro, by bus B、Asking for direction， showing the way
Shopping	Buying fruits, buying clothes, buying books
Seeking medical advice	Getting sick, seeing a doctor, buying medicine
Dining	Food and beverage, cooking methods, dining out
Marriage and family	Love, marriage and family
Entertainment	Exercising, surfing the Internet, going to movies, watching TV, traveling, visiting a park

Altogether 90 seeded topic words and 2,401 seeded entries are collected. Examples of words in the topic vocabulary are listed as follows (arranged based on application frequency, for "daily life—dining—dining out"):

Eating, hygienic, inexpensive, fresh, hotel, taste, restaurant, mild, home delivery, authentic, economic, waiter, nutritious, salty, check, fast food restaurant, menu, bill, wrapping leftovers, delicious, light, tavern, paying the bill, tasty, ordering dishes, spicy, tips, unsavory, coupons, napkin tissue, serving dishes, snack bar, napkin, greasy, meal delivery service, ordering food, dishes, sweet and delicious, avoiding certain food, go Dutch, table number, paying by card, specials, special snacks, take-out, paying in cash, signature dishes.

C. Rough segmentation experiment of video subtitles

Subtitle text of the movie A Story of Lala's Promotion is chosen for the experiment, which contains 1,427 dialogues. We expect to roughly segment scenes with rough temporal segmentation method. Time interval between two subtitle dialogue sentences exceeding a certain threshold value (6 seconds are found to be the most appropriate threshold value, after repeated experiments and comparisons) can be set as natural segmentation mark between topics. Therefore, the subtitle text of the movie A Story of Lala's Promotion is roughly segmented into 115 scenes, which are manually verified one by one. The results show that segmentation should have been done at 13 points (which have no impact on subsequent processing and are not considered mistakes) and segmentation shouldn't have been done at 9 points. In general, the accuracy rate is 92.17%.

D. An example of constructing movie topic resources for scene communicative function

What movie resources of oral Chinese topics for scene communicative function can be found, in the movie A Story of Lala's Promotion, for example?

There are altogether 1,472 dialogues in the movie, which are re-segmented into 91 scenes (excluding some clips with correct rough segmentation but too few dialogues). These scenes are manually corresponded to the first category of the above-mentioned constructed topic database. For example, the above-mentioned roughly segmented clips may fall into "daily-life topics-personal information" or "culture topics-social etiquettes".

36 of the 91 video scenes can be finally classified into certain types (one scene may fall into several topic types):

TABLE II.
VIDEO CLIPS AND CORRESPONDING TOPICS (THE FIGURES REPRESENT NUMBERS OF CORRESPONDING CLIPS)

Daily life topics				**Business topics**		**Culture topics**	
Personal information	5	Dining	4	Recruitment and job application	2	Social etiquettes	9
Housing	3	Marriage and family	9	Business negotiation	8	Emotional expression	5
Transportation	3	Entertainment	5	Investment	2	Attitude expression	3
Shopping	2	Contact	2	Marketing	2	Holiday culture	1
Seeking medical advice	2	Time	1				

V. CONCLUSION

The method of constructing multimedia topic resources making use of topic analysis and word clustering methods, to assist oral Chinese topics teaching based on scene communicative function, also applies to teaching of English and other languages. The proposed new methods of topic segmentation, topic tracking and word clustering and grading are of great reference significance for subject analysis and word clustering in natural language processing.

In the future, subtitle text corpus and topic and topic vocabulary databases will be used for automatic topic detection and segmentation and word clustering of large-scale video resources and algorithm research and system implementation of automatic video clip grading.

REFERENCES

[1] Walls, F.; Jin, H.; Sista, S.. "Probabilistic models for topic detection and tracking", *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol1, pp. 521-524, 1999.

[2] He, Qi; Chang, Kuiyu; Lim, Ee-Peng; Banerjee, Arindam. "Keep it simple with time: A reexamination of probabilistic topic detection models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol32, pp.1795-1808, 2010. http://dx.doi.org/10.1109/TPAMI.2009.203

[3] Liu Hua. "Words Clustering Based on Keywords Indexing from Large-scale Categorization Corpora," *5th International Conference on Information Assurance and Security, IAS*, vol1, pp. 407-410, 2009.

[4] Chen Keli. "Balanced Corpus Analysis and Text Categorization Method Based on Large-Scale Real Text", *Advances in Computation of Oriental Languages*. Tsinghua University Press, 2003.

[5] J Allan, J Carbonell, G Doddington, J Yamron and Y Yang. "Topic detection and tracking pilot study: Final report", *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Virginia: Lansdowne, vol5, pp. 194-218, 1998.

[6] Gieve, S. & R. Clark. "The Chinese approach to learning: Cultural trait or situated response? The case of a self-directed learning programme", System, vol33, pp. 217-230, 2005. http://dx.doi.org/10.1016/j.system.2004.09.015

[7] Liu Hua, *Word Computation and Application*, Ji'nan University Press, 2010.

AUTHORS

Hua Liu is with the College of Chinese Language and Culture, Jinan University, Guangzhou, China (e-mail: liuhua0461@sina.com).

Manuscript received 15 July 2013. Published as re-submitted by the author 13 October 2013. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 12JNKY008.