SCRIPT-DESCRIPTION PAIR EXTRACTION FROM TEXT DOCUMENTS OF ENGLISH AS SECOND LANGUAGE PODCAST

Hyungjong Noh, Minwoo Jeong, Sungjin Lee, Jonghoon Lee and Gary Geunbae Lee
Pohang University of Science and Technology (POSTECH), Korea

Keywords: Text extraction, ESL.

Abstract: One of the best effective way to learn a language is having a conversation with a native speaker. However it is often very expensive way. A good alternative way is using Dialog-Based Computer Assisted Language Learning (DB-CALL) systems. The feedback quality in DB-CALL systems is very important. Therefore, to provide various expressions as feedback information, we propose a method which extracts script and their description sentence pairs from English as a Second Language (ESL) podcast web site. A linear CRFs classifier is used to find the corresponding description sentences and several features are selected according to the characteristics of the ESL text documents. The experimental results show that the performance of our system is acceptable.

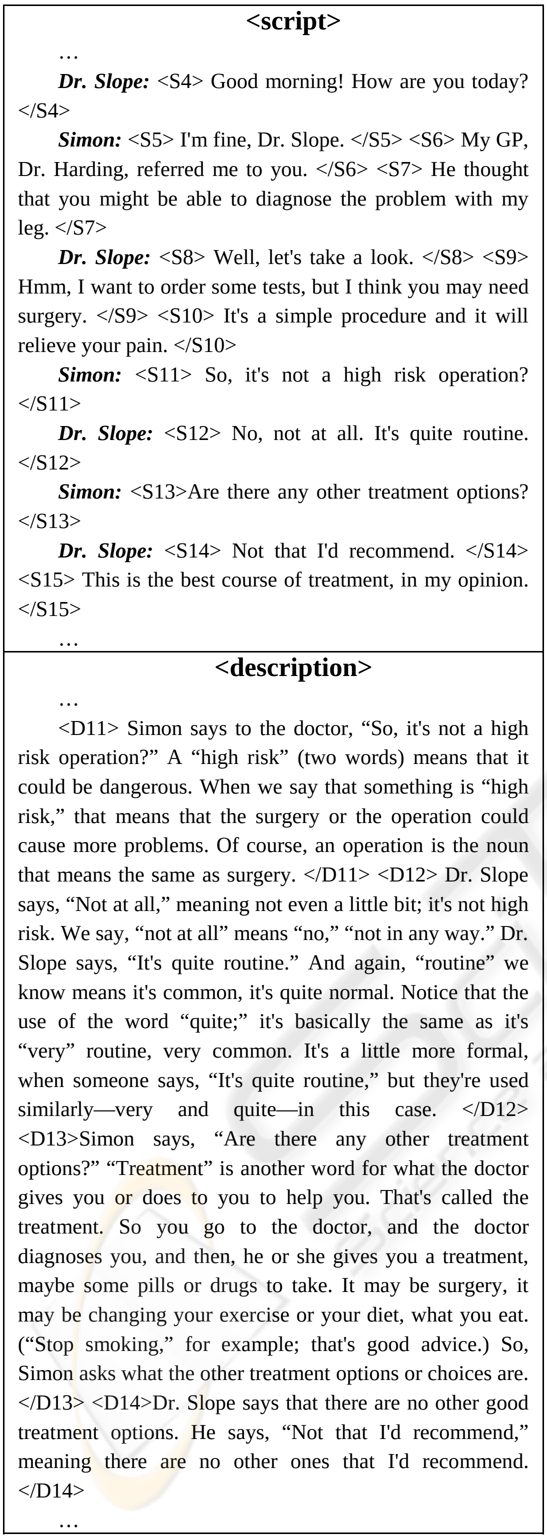
1 INTRODUCTION

The importance of English education increases with the needs of conversation ability in English. Especially, many students learn English not as a foreign language, but as a second language. One of the best way to learn English as a Second Language (ESL) is talking with a native speaker. One can learn abilities to listen and speak practical expressions by having a conversation in English. However, the most serious problem of this approach is that the conversational education is very expensive. Although the conversational education is one of the most effective way to learn languages, the cost can be a obstacle to take this approach. Therefore some alternative ways are needed to take the advantage of conversational education with reducing costs.

To do this, many researchers develop Dialog-Based Computer Assisted Language Learning (DB-CALL) systems. These systems can talk with a user and the user learns the speaking and listening skills through the conversational practice. The system SPELL (Hazel, 2005) is one of the famous DB-CALL systems. Some technologies such as virtual environment, speech recognition and synthesis are used in this system. Another DB-CALL system, ISLAND (Ian, 2007), is developed based on the spoken dialog system of MIT. This system shows the recognition results including misrecognized utterances because they can be used to improve the user's pronunciation. The system DEAL (Wik, 2007) combines the concept of computer games and dialog systems. The user tries to complete the mission given by the system.

One of the essential issues of DB-CALL systems is giving feedback information to promote learners' comprehension. When the system responses to the user, the user may not understand the exact meaning of the response sentence or some expressions in the utterance. Then some additional information need to be presented to the user to help comprehension. The additional information can include descriptions for the expressions or some example sentences which have the expressions. With these feedback information, the user can acquire better comprehension ability. Therefore, if the system does not provide these operations, the users will not be effectively able to improve their conversational skill.
As a resource of feedback information, we concentrate on ESL podcast web sites. On these sites, many useful scripts and descriptions are provided for ESL education (Figure 1). Scripts can be a simple dialog or a short newspaper article.
A native speaker explains the scripts in following description part.

Noh H., Jeong M., Lee S., Lee J. and Geunbae Lee G. (2010).
SCRIPT-DESCRIPTION PAIR EXTRACTION FROM TEXT DOCUMENTS OF ENGLISH AS SECOND LANGUAGE PODCAST.
In *Proceedings of the 2nd International Conference on Computer Supported Education*, pages 5-10
DOI: 10.5220/0002773000050010



<script>

…

Dr. Slope: <S4> Good morning! How are you today? </S4>

Simon: <S5> I'm fine, Dr. Slope. </S5> <S6> My GP, Dr. Harding, referred me to you. </S6> <S7> He thought that you might be able to diagnose the problem with my leg. </S7>

Dr. Slope: <S8> Well, let's take a look. </S8> <S9> Hmm, I want to order some tests, but I think you may need surgery. </S9> <S10> It's a simple procedure and it will relieve your pain. </S10>

Simon: <S11> So, it's not a high risk operation? </S11>

Dr. Slope: <S12> No, not at all. It's quite routine. </S12>

Simon: <S13>Are there any other treatment options? </S13>

Dr. Slope: <S14> Not that I'd recommend. </S14> <S15> This is the best course of treatment, in my opinion. </S15>

…

<description>

…

<D11> Simon says to the doctor, “So, it's not a high risk operation?” A “high risk” (two words) means that it could be dangerous. When we say that something is “high risk,” that means that the surgery or the operation could cause more problems. Of course, an operation is the noun that means the same as surgery. </D11> <D12> Dr. Slope says, “Not at all,” meaning not even a little bit; it's not high risk. We say, “not at all” means “no,” “not in any way.” Dr. Slope says, “It's quite routine.” And again, “routine” we know means it's common, it's quite normal. Notice that the use of the word “quite;” it's basically the same as it's “very” routine, very common. It's a little more formal, when someone says, “It's quite routine,” but they're used similarly—very and quite—in this case. </D12> <D13>Simon says, “Are there any other treatment options?” “Treatment” is another word for what the doctor gives you or does to you to help you. That's called the treatment. So you go to the doctor, and the doctor diagnoses you, and then, he or she gives you a treatment, maybe some pills or drugs to take. It may be surgery, it may be changing your exercise or your diet, what you eat. (“Stop smoking,” for example; that's good advice.) So, Simon asks what the other treatment options or choices are. </D13> <D14>Dr. Slope says that there are no other good treatment options. He says, “Not that I'd recommend,” meaning there are no other ones that I'd recommend. </D14>

…

Figure 1: An example of an ESL podcast document. The description sentence <Di>_</Di> explains the script sentence <Si>_</Si>.

Basically scripts and descriptions are given as speech audio files, but the transcription text files are also provided. Though these files are good sources for ESL education for their own good, we can extract more valuable information from the files for DB-CALL systems. In ESL podcast files, the speaker explains each sentence used in the script part and many phrases in detail. If these descriptions can be extracted with corresponding script sentences or phrases, the extracted pairs can be used as a database for feedback information. When the user who uses a DB-CALL system wants to know the meaning of the sentence or the phrase generated by the system, the system can present similar expressions and their descriptions that are gathered from ESL podcast files. These descriptions can help the user’s understanding better than simple word dictionary explanations, because the descriptions can give practical usage examples and alternative expressions which are used in real world conversations. For example, a user may not understand the meaning of a sentence: “It was quite great”. If the system detected the word ‘quite’ is the point of understanding, it searches the script and description parts related to ‘quite’, <S12>_</S12> and <D12>_</D12> in Figure 1. With these explanations, the user can learn the detailed meaning of the word ‘quite’.

To construct these resources as a database for DB-CALL systems, we propose a method which extracts each script sentence and its description from the ESL podcast text files. The method must be semi-automatic to reduce the construction cost. For each sentence in the script part, the corresponding description sentences are classified by a linear Conditional Random Fields (CRFs) (Lafferty, 2001) classifier. Using the classifier, we can reduce human effort. Several features are selected to train the classifier. The experimental results show that the proposed method can extract each pair of a script sentence and corresponding descriptions successfully.

The remainder of this paper is as follows: Section 2 presents related work. Section 3 describes our proposed method and features. Section 4 explains the experimental environments. Section 5 shows the evaluation results of our method. Finally, we conclude this paper.

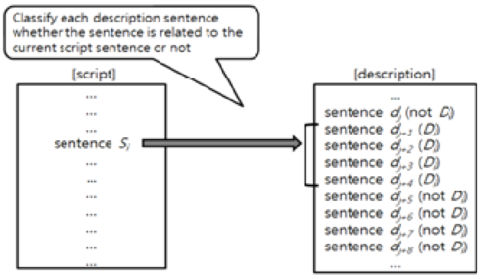

Classify each description sentence whether the sentence is related to the current script sentence or not
[script]
sentence S_i
[description]
sentence d_j (not D_i)
sentence d_{j+1} (D_i)
sentence d_{j+2} (D_i)
sentence d_{j+3} (D_i)
sentence d_{j+4} (D_i)
sentence d_{j+5} (not D_i)
sentence d_{j+6} (not D_i)
sentence d_{j+7} (not D_i)
sentence d_{j+8} (not D_i)

Figure 2: Extracting the script-description pair by binary classification.

2 RELATED WORK

Some researchers studied for extracting information or knowledge from texts, though their researches are not directly related to ESL education. Barbara et al. classified documents into two classes, relevant or not relevant, given a topic of interest. Rege et al. presented an approach that co-clusters document-word. Sindhwani et al. proposed an algorithm about document-word co-regularization for sentiment analysis. Their works are related to our work in terms of pair extraction from documents. However our goal is not exactly the same with these researches because we want to extract descriptions for all script sentences, not a given topic.

We found another research that is similar to our study. Sun et al. proposed a method that conducts topic segmentation and alignment with shared topics on multiple documents.

The most similar work is the one applied on online forums. Cong et al. proposed a method to detect a question in a forum thread, and detect the answer. Ding et al. proposed a framework which is based on various CRFs to detect contexts and the answers of questions. They used various algorithms that are used for natural language processing to extract questions and corresponding answers and contexts from online forum documents.

Even though there are many researches to extract documents with given topics, to our best knowledge, none of the previous works were conducted on ESL education domain. We think that our attempt to extract valuable knowledge from ESL podcast documents would be the first trial to contribute to language learning.

3 EXTRACTING SCRIPT AND DESCRIPTION PAIR

3.1 CRF as a Classifier

To extract information from documents, we need to know the characteristics of the documents. The ESL podcast documents have script part and in subsequent parts the native speaker explains them sequentially according to the order of the script sentences. This means that to which script sentence a description sentence corresponds highly correlates with the correspondences of the adjacent descriptions. Within a description block (Figure 1), the corresponding script sentence of the first description sentence should be the one that the subsequent description sentences correspond to.

Considering these characteristics of the documents, we select CRF as a classifier. CRF has strength in using sequential context information to predict labels of nodes. In our problem, the description sentences can be treated as nodes and the corresponding script sentence can be treated as the label of each description sentence. We use a linear CRF to match the script and description sentences. The following part explains the problem definition in detail with notations.

3.2 Problem Definition

Given the *m* script sentences, each sentence is labeled as $S_1, S_2, \ldots, S_m$, sequentially. The task is to classify the all *n* description sentences, $d_1, d_2, \ldots, d_n$, into the class $D_1, D_2, \ldots, D_m$. Each description sentence has its own label, one of $D_1, D_2, \ldots, D_m$. Description sentences which have the class D_i are considered as the description part of the *i*th script sentence, S_i. This is basically a multi-class sequence labeling problem because the script sentences are treated as labeled classes.

We changed the problem to binary class prediction problem (Figure 2.). In other words, we consider only one script sentence at a time. Given a script sentence S_i at each time, we classify all the description sentences into positive (D_i) or negative (not D_i). Positive means that the description sentence is explaining the script sentence S_i, while negative means that it is not. The classification process is repeated *m* times with each script sentence $S_1, S_2, \ldots, S_m$. Because the features for classification, to be explained below, are related to similarity between the description sentences and corresponding script sentence, we must consider only one script sentence at each classification process. We want to build the model that considers relation between a script sentence and description sentences.

3.3 Features Selected

We selected some features to build the model which reflects the essential relation between the script part and description part. Our feature selection can be justified with statistics acquired from the ESL podcast documents.

First feature is the lexical similarity between the script sentence and the description sentence. This selection is reasonable because the description sentence uses the words which occur in the script sentence as a necessity. The duplicated words will increase the lexical similarity. We computed the measure like Term Frequency – Inverse Document Frequency (TF-IDF) value for each word which occurs in all documents and formed TF-IDF vectors of the script sentence and the description sentence. The measure is defined as in the following equation:

$$\frac{\text{(The ratio of word occurence in the sentence)}}{\text{(\# of sentences which have the word occurence in all documents)}} \quad (1)$$

Calculating the cosine similarity between two vectors, we can measure how similar the two sentences are. From training data, we got the mean similarity of 0.13589 between the script sentence (a sentence S_i) and their actual description sentences (sentences which have the label D_i), and 0.00620 between the script sentence (a sentence S_i) and description sentences which explain another script sentence (sentences which have the label D_i, $i \neq j$). This result shows that the similarity feature between the script and description sentences is useful to improve the classification performance.

Second feature is the same similarity value between two adjacent sentences in description part. We assume that the similarity may be relatively low when the label changes and high when it does not. The statistics from the training data show that our assumption is acceptable (Table 1). The similarity is highest (0.06357) when two adjacent description sentences explain the same script sentence. The similarity is relatively low (0.01429, 0.01279) when two sentences explain different sentences.

These similarity differences according to the label change can be clues to the label classification.

We also used the semantic similarity feature. This feature is also measured for two cases; between script sentence and description sentences, and between two adjacent description sentences.

Table 1: The mean similarity values according to the labels of adjacent description sentences.

		Current Sentence (d_p)	
		D_i	not D_i
Previous	D_i	0.06357	0.01429
Sentence (d_{p-1})	not D_i	0.01279	0.05903

To calculate semantic similarities, we expand each word to its three hypernym levels with WordNet (Fellbaum, 1998). For example, the word “pencil” has hypernyms “writing implement”-“implement”-“instrumentality, instrumentation”. All hypernyms are counted for a sentence and we construct a semantic vector that represents which hypernym occurs in the sentence. The cosine similarity of the semantic vectors of two sentences means the semantic similarity. We found that the mean semantic similarity between the script sentence and their actual description sentences is 0.28955, which is higher than the similarity between the script sentence and other description sentences (0.07084).

The last feature we used is that the difference between relative positions of the script sentence and the description sentence within each part. When a script sentence occurs early in script part, its description may occur early too because the structure of the description part is sequential according to the script part in ESL podcast documents (Figure 1). To use this characteristic, we defined the relative position (RP) measure. A script sentence S_i has RP value of i/n, and a description sentence d_j has value of j/m. All RP values are between 0 < RP < 1. We used the difference of two RP values: $|\mathrm{RP}(s_i) - \mathrm{RP}(d_j)|$. The statistics from training data give us the value of 0.13820 for the actual description sentences and the value of 0.35026 for description sentences which explain another script sentence. This means that the RP of the script sentence and its corresponding description sentence is similar.

4 EXPERIMENTAL SETUP

4.1 Evaluation Measure

We used precision, recall, and F1-score as performance measures. F1-score is the harmonic mean of precision and recall. They are widely used in information retrieval researches. Higher value means higher performance. The equations of the measures are as below:

$$\text{precision} = \frac{(\#\text{ of actual positive sentences predicted as positive})}{(\#\text{ of sentences predicted as positive})} \quad (2)$$

$$\text{recal} = \frac{(\#\text{ of actual positive sentences predicted as positive})}{(\#\text{ of actual positive sentences})} \quad (3)$$

$$\text{F1} - \text{score} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}. \quad (4)$$

4.2 Corpus Information

We acquired 100 documents that have their own script and description part from ESL podcast web site. Eight annotators tagged the documents for matching each script sentence to the corresponding sentences. The agreements between annotators for tagging classes is 94.6%. The documents (24267 sentences) are divided into two corpora: training data (80 documents, 18906 sentences) and test data (20 documents, 5361 sentences). A set of linear CRFs classifier is trained using the training data, and test data is used to validate the model.

5 EXPERIMENTAL RESULTS

We conducted the experiment using a set of linear CRFs with changing features (Table 2). We classified features as three categories: lexical similarity (F1), semantic similarity (F2), and difference of relative position (F3). In experiments conducted with each category individually, F1, F2 and F3 showed the F-1 score as 0.629, 0.467, and less than 0.05, respectively. This result shows the influence of each feature briefly. The lexical similarity feature between the script sentence and the description sentences is the most effective feature. Table 2 shows the incremental results with the additional features according to the influence order. As expected, the performance which uses full features shows the highest performance (0.698). The interesting thing is the effect of F3. F3 shows very low performance with itself, but it helps improving the performance with combining other two feature categories. In general, precisions are higher than recalls. It means that once the system found some sentences, then they are actually positive with high accuracy. However some positive sentences were not found by the system. We think that it occurs because the system predicts only the sentences which satisfy all three features as positive, but many positive sentences are satisfying only one or two features in the real corpus. Considering another similar experimental results (Ding et al.), we think that our performance (F1-score: 0.698) is acceptable.

Table 2: The performance result with changing features.

Features	Precision	Recall	F1-score
F1	0.685	0.582	0.629
F1 + F2	0.731	0.608	0.663
F1 + F2 + F3	0.750	0.652	0.698

F1: The similarity features;
F2: The semantic features;
F3: The difference of relative positions of the script sentence and the description sentences

6 CONCLUSIONS

We proposed a method which extracts the script sentence and its corresponding description sentences. We think that the performance is good in considering that this study is our first attempt on ESL domain. The pair extraction problem was converted to binary classification and the linear CRF model was a proper choice to conduct the classification process. However there is some room for improvement. Any linguistic information including POS(part-of-speech)-tag was not used to represent the sentences. Our classifier can be also modified to improve the performance. Another limitation was that we extract information from only text documents. There are huge audio resources for ESL education. We need to access these audio files so that more various expressions can be gathered easily. Once we transcribe audio files to text files, then we can apply our method directly to that files. Therefore our future work includes expanding features, improving classification models, and developing a method to extract information from various sources including audio files.

ACKNOWLEDGEMENTS

This research (paper) was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Knowledge Economy (MKE).

REFERENCES

Hazel, M., Mervyn, J., 2005. Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning.*

Ian, M., Stephanie, S., 2007. Immersive Second Language Acquisition in Narrow Domains: A Prototype ISLAND Dialogue System. In *SLaTE-2007.*

Wik, P., Hjalmarson, A., Brusk, J., 2007. DEAL a serious game for CALL practicing conversational skills in the trade domain. In *SLaTE-2007.*

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML.*

Barbara, D., Domeniconi, C., Kang, N., 2003. Mining Relevant Text from Unlabelled Documents. *Third IEEE International Conference on Data Mining.*

Rege, M., Dong, M., Fotouhi, F., 2006. Co-clustering documents and words using Bipartite Isoperimetric Graph Partitioning. *Sixth IEEE International Conference on Data Mining.*

Sindhwani, V., Melville, P., 2008. Document-Word Co-Regularization for Semi-supervised Sentiment Analysis. *Proceedings of IEEE International Conference on Data Mining.*

Cong, G., Wang, L., Lin, C., Song, Y., Sun, Y., 2008. Finding Question-Answer Pairs from Online Forums. *Proceedings of the 31st annual international ACM SIGIR conference.*

Ding, S., Cong, G., Lin, C., Zhu, X., 2008. Using Conditional Random Fields to Extract Contexts and Answers of Questions from Online Forums. *Proceedings of ACL-08: HLT.*

Sun, B., Mitra, P., Zha, H., Giles, C., Yen, J., 2007. Topic Segmentation with Shared Topic Detection and Alignment of Multiple Documents. *Proceedings of the 30th annual international ACM SIGIR conference.*

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

English as a Second Language Podcast. Available: http://www.eslpod.com/