ENGtube: an Integrated Subtitle Environment for ESL

Chi-Ho Li[†], **Shujie Liu**[‡], **Chenguang Wang**[⊢] and **Ming Zhou**[†]
[†]Microsoft Research Asia, Beijing, China
{chl, mingzhou}@microsoft.com
[‡] Harbin Institute of Technology, Harbin, China
shujieliu@mtlab.hit.edu.cn
[⊢]Peking University, Beijing, China
wangcg.pku@gmail.com

Abstract

Movies and TV shows are probably the most attractive media of language learning, and the associated subtitle is an important resource in the learning process. Despite its significance, subtitle has never been exploited effectively as it could be. In this paper we present ENGtube, which is a video service for ESL (English as Second Language) learners. The key component of this service is an integrated environment for displaying the video clips, the source subtitle and the translated subtitle with rich information at users' disposal. The rich information of subtitle is produced by various speech and language technologies.

1 Introduction

Language learning has always been area to apply NLP technology. Among language teachers, it is widely accepted that movies and TV shows, in comparison with books, dictionaries, etc. are the most interesting and attractive media in learning language. When watching movies and TV programs in a non-native language, subtitle is an important aid to the learners. In this paper we will use 'video' as a subsuming term to refer to movies, TV programs and other similar content. We will also use the term 'transcript' to refer to the transcribed text of the monologue or dialogue in video content, and 'subtitle' to refer to a portion of the transcript to be shown at a time when the video is displayed.[1]

[1] There is a related term 'closed caption', which is usually defined as a kind of subtitle for people with hearing difficulties. Closed caption is not covered in this paper.

According to Micola *et.al.* (2009), people in (non-English speaking) countries where movies and TV programs are shown with their original English subtitle generally achieve higher English proficiency, and such difference is amount to 4 to 20 years of formal education in school. Despite its significance, video subtitle has never been exploited effectively as it could be; subtitle has been nothing more than a piece of plain text when a video is displayed. Thus it is an important task to unlock the potential of video subtitle for second language learning.

This paper introduces ENGtube, a video service with an integrated subtitle environment, aiming at maximizing the learning experience of ESL learners from English video. We assume that English is the second language to be learned and Chinese is the native language of the ESL learners. ENGtube comprises a front-end user interface and back-end data processing modules, which will be elaborated in detail in the next two sections.

2 Front-end User Interface

Figure 1 illustrates the basic outlook of the front-end user interface of ENGtube, which is currently implemented as a web application. The upper right area of the interface is the video player. Traditionally, subtitle is displayed at the bottom of the screen of the video player. However, in order to provide more flexible manipulation of, and users' interaction with, the subtitle, ENGtube displays subtitle in a separate area (the upper left area of the interface). The subtitle area displays two kinds of subtitle: the source English subtitle transcribed from the video soundtrack, and the subtitle translated into the user's own native language.

As mentioned before, a subtitle of a movie or TV program is a line to be shown on screen at a

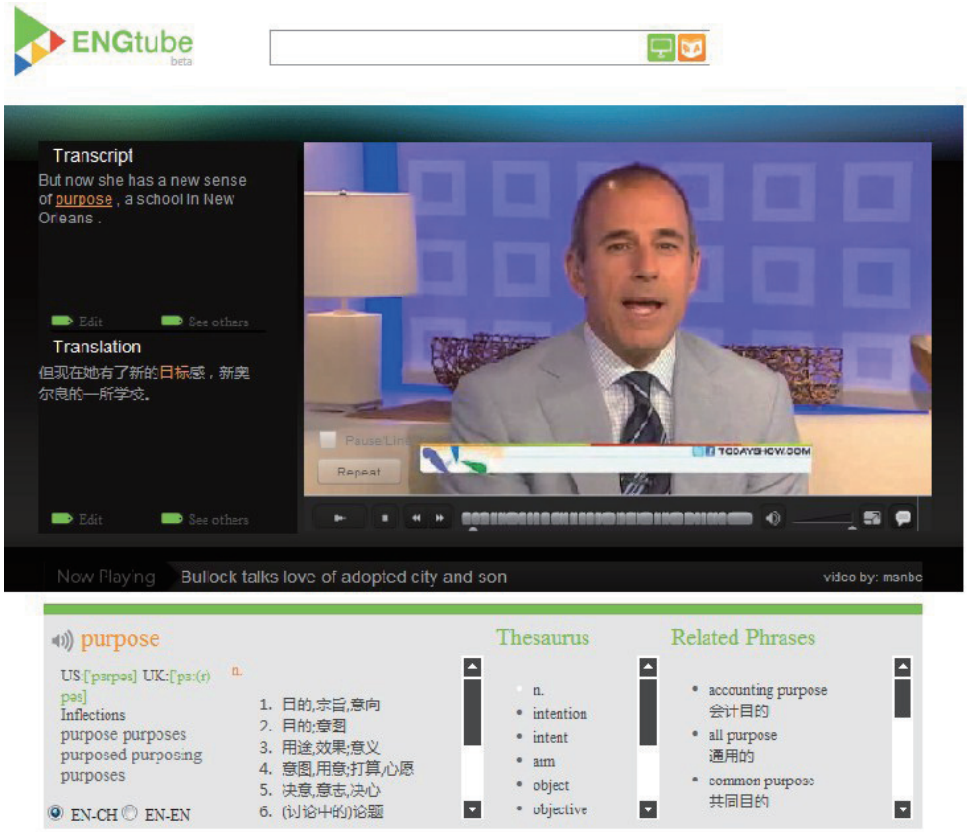

Figure 1. A snapshot of the ENGtube front-end user interface.

time. As ENGtube is aimed at language learning, it ensures that each subtitle corresponds to a complete sentence or clause. A subtitle in ENGtube should be long enough to convey complete meaning yet also short enough to suit non-native speakers' English proficiency.[2] Thus at each moment the subtitle area displays only one sentence or clause for both the source and translated subtitle. Besides, the video player inherently divides the video clip into segments corresponding to these sentences/clauses, and it provides several controls to allow the user switch to a particular sentence/clause.

The source and translated subtitles are not only sentence/clause-aligned but also word-aligned. If the user moves the mouse pointer over a particular English word, then both the English word and the corresponding Chinese word will be highlighted. Such word alignment among both versions of subtitle is implemented to improve the user's understanding of each individual word.

In addition to the alignment between the two versions of subtitle, ENGtube also implements the alignment between subtitle and soundtrack. That is, a source subtitle dynamically changes the colors of words in response to the progress of the soundtrack. For each line, those words which have been spoken out in the video will be highlighted in a different color. That is, the display of the source subtitle is in the same style of Karaoke.

One more aid to the user is dictionary information on user's request. When the user clicks on a particular English word in the source subtitle, then the corresponding dictionary information will be displayed in an area under the video player and subtitle area. The dictionary information includes pronunciation, brief definitions/explanations, synonyms, and related phrases.

In sum, the ENGtube front-end arranges the video and its transcript into many sentences/clauses (shortest possible meaningful units), and provides multi-level alignment information, alongside dictionary information and exercises. While ENGtube

[2] For example, if the sentence at hand is as short as "Mary is going to cinema", then the entire sentence will be shown at once; if the sentence at hand is as long as "Mary is going to cinema near the park after she has finished her work and dinner", then the clause "Mary is going to cinema near the park" will be shown at the first moment, and the clause "after she has finished her work and dinner" at the second.

is currently build for Chinese users to learn English, the technology itself is language independent, and can be extended to support other language pairs in the future.

3 Back-end Data Processing

3.1 Source Data

Currently we have licensed a number of news video clips from MSNBC, and there are two reasons behind this decision. First of all, we need to negotiate on licensing terms with one data provider only. Moreover, an Internet news provider like MSNBC hosts a huge number of news content, including politics, finance, sports, and celebrities. Such a wide variety of topics is hopefully attractive enough to most ESL learners.

3.2 System Architecture

Given a video clip, there are five phases of the back-end data processing.

The first phase is the ***transcription*** of the monologue/dialogue in the video clip. Such transcript is the basis of all other modules. Currently, as the video clips in ENGtube are provided by MSNBC, which also provides the source English transcripts, we are free from errors by automatic transcription. One problem of these video clips is that not all the video clips are suitable for language learning, such as those video clips without much linguistic content, or those where the speakers speak too quickly for an ESL learner. In the first phase, we introduce video filters to discard all these videos.

The second phase is ***sentence segmentation***; that is, to segment the transcript in phase 1 into sentences or clauses, such that at each moment ENGtube displays a subtitle which is long enough to be meaningful but also short enough to be understood by a non-native speaker. (c.f. Section 2.) The details of the sentence segmentation module will be elaborated in Section 3.4

The third phase is ***translation***; i.e. to translate the source English transcript into the user's own native language. There are two possible ways, viz. manual translation and machine translation (MT). Currently the source video provider (MSNBC) does not provide translation of transcript. Yet fortunately it is found that our state-of-the-art MT system produces satisfactory translation output for helping ESL learners to grasp the essential meaning of the English transcript.

The fourth phase is ***word alignment***. Word alignment is about the mapping between the English words and the translated Chinese words. If the translation is produced by some MT system, then the required word alignment information can also be produced along with translation output. (c.f. Section 3.5 for explanation.) Otherwise a word aligner is needed. (c.f. Section 3.6)

The fifth phase is ***audio-text alignment***. Audio-text alignment is the alignment between the audio signals in video soundtrack and the transcribed text. Forced alignment is used. (c.f. Section 3.7)

3.3 Video Filters

To filter the unsuitable video clips for language learning, we first calculate the average duration (which is 0.0057 second per word) using all video clips. Any video clip of which the average duration is outside the range 0.0057+-0.001 will be filtered. We also use the per-word cross entropy using *a* 5-gram language model to judge the quality of the transcription. The language model is trained on Gigaword corpus (248M words), which is of the same domain as our MSNBC news video clips, and a corpus of movie subtitle (102M words), which is of the similar colloquial style as the broadcast conversation in the news video. Finally, we discard clips which have a long pause (5 seconds) between two continues words, using the audio-text alignment result. This filter has the desirable side effect of removing manual transcription errors, which are not rare in news video transcript. Our forced alignment system would fail to align audio signals with mis-transcribed words, thus creating long pauses. The video file with incorrect manual transcription will then be discarded by the last filter.

3.4 Sentence Segmentation

Sentence/clause segmentation in ENGtube is done by a CRF-based classifier following Liu (2005). At the end of each word of a given video transcript, we ask whether this word should be the last word of a subtitle line. This is a problem of binary classification. The feature templates used by the CRF classifier includes the lexical form and the part-of-speech of the current, preceding, and following words. The training dataset is based on a held-out set of transcription of 100 video clips, with manual

annotation of sentence/clause boundaries. The precision of the classifier is as high as 87%. Speech features will be included in the future to enhance the performance.

3.5 Machine Translation

The machine translation system used in ENGtube is our re-implementation of the hierarchical phrase-based statistical machine translation framework (Chiang, 2007). This system achieves state-of-the-art level performance as shown by its Bleu score of 47.18 on the English-to-Chinese track of the NIST2008 evaluation exercise.

The MT engine does not only produce translation output but also by-product information about the word alignment between the source input and the translation output. Conventionally, after word-aligning the bilingual sentences in MT training data, phrase pairs and rules are collected from the produced alignment matrices. The key here is to keep track of the word alignment links *within* the phrase pairs and rules. During MT decoding, then, all the alignment links for those phrase pairs and rules used by the Viterbi translation candidate are exactly the required alignment information between the source input and translation output. There may be different alignment matrixes for a give phrase pair or rule, and only the most frequent one is maintained.

Note that generative word aligner, such as the widely used tool GIZA++, often produces inconsistent alignment links. While these noisy links may not harm the ultimate translation performance too much, they are disastrous to ESL learners. Therefore ENGtube uses a much more precise word alignment routine as described in the next section.

3.6 Word Alignment

There are two word alignment systems used in ENGtube. The first is our re-implementation of the generative, HMM-based alignment system. This is used both to train a probabilistic bilingual lexicon and to produce preliminary alignment result. The second is a discriminative word aligner based on Moore (2006), but with the training algorithm of Minimum Error Rate Training (Fraser, 2006). The discriminative aligner starts the search space with the HMM-based alignment result, and it also uses the word pair probabilities learned by the HMM-based aligner as features. The ultimate word alignment output of this two-stage word alignment procedure achieves the F-score 92% on our own evaluation dataset, while the GIZA++ baseline achieves 83% only.

3.7 Audio-text Alignment

Speech technology used by ENGtube is forced alignment, which is used for aligning video soundtrack against transcribed text. Forced alignment system is based on the standard HMM-based algorithm that is widely used in the training of ASR.

The accuracy of forced alignment is very promising. The syllable error rate is as low as 7%. Most forced alignment systems are based on the assumption that only one person is speaking in the video clip. However, our forced alignment system can achieve a high performance even on broadcast conversations, where more than one person is speaking.

4 Comparison

The website EnglishCentral[3] was launched when ENGtube was still being developed. This website shares very similar goals and front-end user interface with ENGtube. It should be pointed out that ENGtube provides much more information to the user, such as the multi-level alignment information. Moreover, the back-end data modules of ENGtube allows both manual and automatic processing for many different tasks, while there is only manual processing in EnglishCentral.

5 Conclusion and Future Work

In this paper we have elaborated the motivation, scenarios, architecture, and implementation of the ENGtube system, for helping ESL learners maximize their learning experience from English-speaking video clips. The useful functions in the front-end user interface are supported by the back-end data processing modules, which are in turn backed up by a number of speech and language technologies. While the current performance of these modules is acceptable, there is still much room of improvement for each module. Particularly we have to, in near future, collect more data of the news video genre and topics to re-train each module.

[3] http://www.englishcentral.com

References

David Chiang. 2007. *Hierarchical Phrase-based Translation. Computational Linguistics*, 33(2).

Alexander Fraser and Daniel Marcu. 2006. *Semi-Supervised Training for Statistical Word Alignment.* In Proceedings *of* ACL, Pages:769-776.

Augusto Rupérez Micola, Arturo Bris and Albert Banal-Estanol. 2009. TV or not TV? Subtitling and English skills. Universitat Pompeu Fabra Economic working papers series 1156.

Robert Moore, Wen-tau Yih, and Andreas Bode. 2006. *Improved Discriminative Bilingual Word Alignment.* In Proceedings *of* ACL, Pages: 513-520.