Smart Subtitles for Vocabulary Learning

X X X
X X X
X X X
X X X

ABSTRACT

Language learners often use subtitled videos to help them learn the language. However, standard subtitles are suboptimal for vocabulary learning, as translations are nonliteral and made at the phrase level, making it hard to find connections between the subtitle text and the words in the video. This paper presents Smart Subtitles, which are interactive subtitles tailored towards vocabulary learning. Smart Subtitles can be automatically generated from common video sources such as subtitled DVDs. They provide features such as vocabulary definitions on hover, and dialog-based video navigation. Our user study shows that students studying Chinese learn over twice as much vocabulary with Smart Subtitles than with dual Chinese-English subtitles. Learners' self-assesed enjoyment of the viewing experience, as well as their comprehension of the video, both self-assessed and as indicated by independent evaluations of their summaries, remain unchanged.

Author Keywords

subtitles; interactive videos; language learning

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation: Graphical User Interfaces

INTRODUCTION

Students studying foreign languages often wish to enjoy authentic foreign-language video content. For example, many students cite a desire to be able to watch anime in its original form as their motivation for starting to study Japanese [8]. However, the standard presentations of videos are not accommodating towards language learners. For example, if a learner were watching anime, and did not recognize a word in the dialog, the learner would normally have to listen carefully to the word, and look it up in a dictionary. This is a time-consuming process which detracts from the enjoyability of watching the content. Alternatively, the learner could simply watch a version that is dubbed, or a version with subtitles in their native language to enjoy the content. However, they might not learn the foreign language effectively this way.

We aim to build a foreign-language video viewing tool that maximizes vocabulary learning, while ensuring that the learner fully understands the video and enjoys watching it.

BACKGROUND

Videos in foreign languages have been adapted for foreign viewers and languages learners in many ways. These are summarized in Figure 1.

Presenting Videos to Foreign Viewers

One approach used to adapt videos for viewers who do not understand the original language is *dubbing*. Here, the original foreign-language voice track is removed, and is replaced with a voice track in the viewer's native language. Because the foreign language is no longer present in the dubbed version, this medium is ineffective for foreign language learning [11].

Another approach is to provide *subtitles* with the video. Here, the foreign-language audio is retained as-is, and the native-language translation is provided in textual format, generally as a line presented at the bottom of the screen. Thus, the learner will hear the foreign language, but will not see its written form. Therefore, they must pay close attention to the audio to learn the foreign language. Subtitles have had mixed reactions in the context of language learning. Some studies have found them to be beneficial for vocabulary acquisition, compared to watching videos without them [6]. That said, other studies have found them to provide little benefit to language learners in learning vocabulary [5]. Additionally, the presence of subtitles are considered to detract attention from the foreign-language audio and pronunciation [16]. The mixed results that studies have found on the effects of subtitles on language learning suggests that their effectiveness depends on factors such as the experience level of the learners [2].

Presenting Videos to Language Learners

In addition to subtitles, other video comprehension aids have been experimented with in the context of language learning:

With a *transcript*, also known as a *caption*, the video is shown along with the text in the language of the audio, which in this case is the foreign language. Transcripts are generally used to assist hearing-impaired viewers. However, they can also be beneficial to language learners for comprehension, particularly if they have better reading ability than listening comprehension ability [6]. However, transcripts can harm comprehension, unless they are used by advanced learners with good reading ability [2].

Figure 1. Ways that a Chinese video can be presented to English-speaking viewers and language learners. Note that GliFlix does not actually support Chinese; these are mockups. The mockup for Smart Subtitles does not show some features, such as dialog-based navigation.

With *reverse subtitles* [5], the video has an audio track and a single subtitle, just as with regular subtitles. However, in reverse subtitles, the audio is in the native language, and the subtitle shows the foreign language. This takes advantage of the fact that subtitle reading is a semi-automatic behavior [7], meaning that the presence of text on the screen tends to attract people's eyes to it, causing them to read it. Therefore, this should attract attention to the foreign- language text. The presentation of the foreign language in written form may also be helpful with certain learners whose reading comprehension ability is stronger than their listening comprehension. That said, because the foreign language is presented only in written form, the learner may not end up learning the pronunciation, particularly with a language with a non- phonetic writing system, such as Chinese.

With *dual subtitles*, the audio track for the video is kept as the original, foreign language. However, in addition to the subtitle displaying the foreign-language, they also display the viewer's native language as well. In this way, a learner can both read the written representation, as well as hear the spoken representation of the dialog, and will still have the translation available. Thus, of these options, dual subtitles provide the most information to the learner. Indeed, dual subtitles have been found to be at least as effective for vocabulary acquisition as either captions or subtitles alone [17].

GliFlix [18] is a variant on the conventional, native-language subtitle, which adds translations to the foreign language for the most common words that appear in the dialog. For example, for a French dialog, instead of "This is a line of dialog", GliFlix would show "This is (*est*) a line of dialog", showing that *is* in French is *est*. In user studies with learners beginning to study French, they attain larger rates of vocabulary acquisition compared to regular subtitles, though not dual subtitles. Compared to dual subtitles, GliFlix has the disadvantage that because it shows only the most common vocabulary words in a dialog, then learners may not learn all the vocabulary in the video. Additionally, because GliFlix presents the foreign vocabulary in the order of the viewer's native language, this approach is likely less beneficial than dual subtitles for other language-learning tasks such as learning pronunciation and grammar.

SMART SUBTITLES INTERFACE

We developed a video viewing tool, Smart Subtitles, which displays subtitles to language learners to enhance the learning experience. It does so by providing support for dialog-level navigation operations, as well as vocabulary-learning features, which are shown in Figure 2. Smart Subtitles can be automatically generated for any video, provided that a caption is available.

Navigation Features

We developed the navigation features of our interface based on foreign language learners' video viewing patterns. In interviews conducted with language learners who enjoyed watching subtitled foreign-language videos, they reported that they often reverse-seeked to the beginning of the current line of dialog to review the portion that had just been said. Therefore, we aimed to make this process as seamless as possible. In our interface, clicking on a section of the dialog will seek the video to the start of that dialog.

Another activity that language learners reported doing was attempting to locate the line of dialog where a particular word or phrase had been said. Therefore, we enable easy seeking through the video based on dialog. The transcript is prominently shown, and can be navigated by pressing the up/down keys, or scrolling. It is also possible to search the video for occurrences of particular words.

Vocabulary Learning Features

Figure 2. Screenshot of the Smart Subtitles system, with callouts pointing out features that help users learn vocabulary and navigate the video.

The vocabulary learning features of our interface are aimed towards the use case where the viewer encounters an unknown word, and would like to look up its definition. The interface allows a user to hover over any word, and it will show the definition.

In addition, for languages such as Japanese and Chinese, which have non-phonetic writing systems, the interface also shows the phonetic representations for learners. For Chinese, it shows *pinyin*, the standard romanization system for Chinese. For Japanese, it shows *hiragana*, the Japanese phonetic writing system.

Sometimes, word-level translations are not enough for the learner to comprehend the current line of dialog. To address these cases, we include an button that shows learners a translation for the currently displayed line of dialog when pressed.

IMPLEMENTATION

Smart Subtitles are automatically generated from captions with the assistance of dictionaries and machine translation. The Smart Subtitles system is implemented as 2 main parts: a system to extract subtitles and captions from videos, as well as an interactive interface to display them to learners.

Extracting Subtitles from Videos

Our system takes digital text captions in either the SubRip (SRT) [25] or Web Video Text Tracks (WebVTT) formats [22] as input. These are plain-text formats that specify a time range for each line of dialog, and the text that should be displayed. We can download these from various online services, such as Universal Subtitles. However, many possible sources of subtitles either do not come with timing information, or are in non-textual formats, so we have developed a subtitle extraction system so that Smart Subtitles can be generated from a broader range of videos. An overview of the subtitle extraction process is shown in Figure 3.

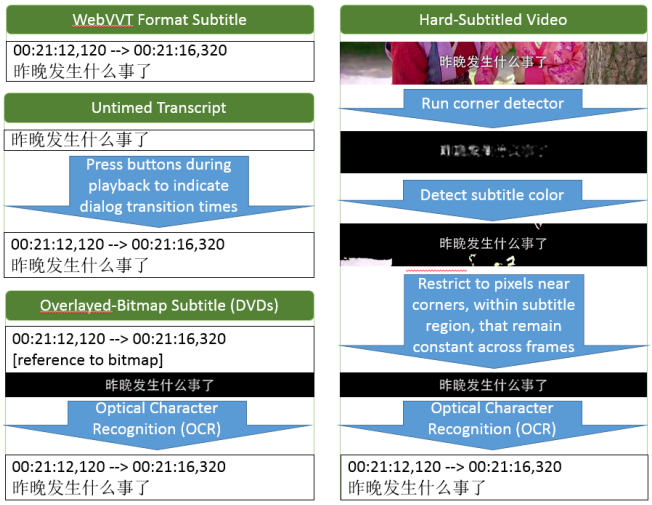

Figure 3. Smart Subtitles uses subtitles in the WebVTT format by default, but it can extract subtitles from various other sources.

Extracting Subtitles from Untimed Transcripts

For many videos, a transcript is available, but the timing information stating when each line of dialog was said is unavailable. Examples include transcripts of lectures on sites such as OpenCourseWare, as well as lyrics for music videos.

It is possible to add timing information to videos automatically based on speech recognition techniques, which is called *forced alignment* [10]. However, we found that existing software for doing forced alignment yields poor results on certain videos, particularly those with background noise and in non-English languages.

Thus, to generate timing information, we wrote an interface where the user views the video, and presses the down button whenever a new line of dialog starts. We gather this data for several users to guard against user errors, and use it to compute the timing information for the transcript.

Extracting Subtitles from Overlayed-Bitmap Formats

Overlayed-bitmap subtitles are pre-rendered versions of the text which are overlayed onto the video when playing. They consist of an index mapping time-ranges to the bitmap image which should be overlayed on top of the video at that time. This is the standard subtitle format used in DVDs, where is it called VobSub.

Because we cannot read text directly from the overlayed-bitmap images in DVDs, Smart Subtitles uses Optical Character Recognition (OCR) to extract the text out of each image. Then, it merges this with information about time ranges to convert them to the WebVTT subtitle format. Our implementation can use either the Microsoft OneNote [15] OCR engine, or the free Tesseract [19] OCR engine.

Extracting Subtitles from Hard-Subtitled Videos

Many videos come with hard subtitles, which include the subtitle as part of the video stream. Hard subtitles have the advantage that they can be displayed on any video player. However, hard subtitles have the disadvantage that they are non-removable. Additionally, hard subtitles are difficult to extract machine-readable text from, because the subtitle must first be isolated from the background video, before we can apply OCR to obtain the text. Existing tools that perform this task, such as SubRip, are time-consuming, as they require the user to specify the color and location of each subtitle line in the video [25].

That said, hard-subtitled videos are ubiquitous, particularly online. Chinese-language dramas on popular video-streaming sites such as Youku are frequently hard-subtitled in Chinese. Thus, to allow Smart Subtitles to be used with hard-subtitled videos, we devised an algorithm which can identify Chinese subtitles in hard-subtitled videos and extract them out.

Our hard-subtitle extraction algorithm takes advantage of the properties of subtitles which we have found to hold true in the Chinese-language hard-subbed material we have observed:

1. Subtitles in the same video are of the same color, with some variance due to compression artifacts.
2. Subtitles in the same video appear in the same vertical region.
3. Subtitles remain static on-screen, so they do not move around and are not animated.
4. Characters in the subtitle have many corners. This is a Chinese-specific assumption, owing to the graphical complexity of Chinese characters.

QUESTION do we actually want to describe the algorithm in this much detail? It's pretty complex and is probably of little interest to CHI audiences...

Our hard-subtitle extraction algorithm first attempts to determine the color of the subtitle. To do so, it first runs the Harris corner detector [9] on each frame of the video. Then, it computes a histogram of color values of pixels near corners, buckets similar color values, and considers the most frequent color to be the subtitle color. This approach works because Chinese characters contain many corners, so corners will be detected near the subtitle, as illustrated in Figure 3.

Next, the algorithm determines which region the subtitle is displayed on the screen. Possible vertical regions are given scores according to how many of the pixels within them match the subtitle color and are near corners, across all video frames. A penalty is given to larger vertical areas, to ensure that it does not grow beyond the subtitle area. We consider the vertical region that scores the highest under this metric to be the subtitle area.

Next, the algorithm determines where each line of dialog in the subtitle starts and ends. For each frame, it considers the set of pixels within the subtitle area, which match the subtitle color, and are near the corners detected by the Harris corner detector. We will refer to such pixels as *hot pixels*. If the number of hot pixels in the frame is less than an eighth of the average number of hot pixels across all frames, then we consider there to not be any subtitle displayed in that frame. If the majority of hot pixels match those from the previous frame, then we consider the current frame to be a continuation of the line of dialog from the previous frame. Otherwise, the current frame is the start of a new line of dialog.

Next, we come up with a reference image for each line of dialog, by taking hot pixels which occur in the majority of frames in that line of dialog. This eliminates any moving pixels from the background, using our assumption that the subtitle text remain static on screen.

Next, we extract the text from the reference images generated for each line of dialog, via OCR. We merge adjacent lines of dialog for which the OCR engine detected the same text. We eliminate lines of dialog for which the OCR engine failed to detect text. Finally, we output the subtitle in WebVTT format.

The accuracy of our hard-subtitle extraction algorithm depends on the resolution of the video and the font of the subtitle. It generally works best on videos with 1280x720 or better resolution, and with subtitles that have distinct, thick outlines. The choice of OCR engine is also crucial - using Tesseract instead of OneNote more than tripled the character error rate, as Tesseract is much less resilient to extraneous pixels in the input.

Overall, on a set of 4 high-resolution Chinese hard-subtitled 5-minute video clips, the algorithm recognized roughly 80% of the dialog lines completely correctly. Overall, roughly 95% of all characters were correctly recognized. 2% of the errors at the dialog line level were due to the algorithm missing the presence of a line of dialog, as the OCR engine often failed to recognize text on lines consisting of only a single character or two. The remaining dialog-level errors were due to characters that were misrecognized by OCR.

Listing Vocabulary Words in a Line of Dialog

The subtitles generated by our subtitle extractor provide us with the text of each line of dialog. For many languages, going from each line of dialog to the list of words it includes is fairly simple, since words are delimited by spaces and punctuation. For European languages supported by Smart Subtitles (English, French, Spanish, and German), we list vocabulary words in each line of dialog using the tokenizer included in the Natural Language Toolkit (NLTK) [3].

A particular issue which occurs with Chinese and Japanese is that the boundaries between words are not indicated in writing. To determine what words are present in each line of dialog in these languages, we instead use statistical word segmenters. We use the Stanford Word Segmenter [21] for Chinese, and JUMAN [13] for Japanese.

Listing Word Definitions and Romanizations

Now that we have determined what the words in each line of dialog are, we need to obtain word definitions and romanizations. These will be displayed when the user hovers over words in the dialog.

For languages such as Chinese that lack conjugation, the process of obtaining definitions and romanizations for words is simple: we look them up in a bilingual dictionary. The dictionary we use for Chinese is CC-CEDICT [14]. This dictionary provides both a list of definitions, as well as the pinyin for each word.

Obtaining definitions for a word is more difficult for languages that have extensive conjugation, such as Japanese. In particular, bilingual dictionaries, such as WWWJDIC [4], the dictionary we use for Japanese, will only include information about the infinitive, unconjugated forms of verbs and adjectives. However, the words which result from segmentation will be fully conjugated, as opposed to being in the infinitive form. For example, the Japanese word meaning "ate" is 食べた [tabeta], though this word does not appear in the dictionary. Only the infinitive form "eat" 食べる [taberu] is present. In order to provide a definition, we need to perform *stemming*, which is the process of deriving the infinitive form from a conjugated word. Rather than implementing our own stemming algorithm for Japanese, we adapted the one that is implemented in the Rikaikun Chrome extension [20].

For the other supported languages, instead of implementing additional stemming algorithms for each language, we instead observed that Wiktionary for these languages tends to already list the conjugated forms of words with a reference back to the original [24]. Therefore, we generated dictionaries and stemming tables by scraping this information from Wiktionary.

For a given foreign-language word, there can be many possible translations depending on the context the word is used in. Hence, we wish to determine the most likely translation for each word, based the contents of the line of dialog it appears in. This problem is referred to as *translation-sense disambiguation* [1]. Smart Subtitles can optionally use translation-sense disambiguation to rank the word definitions displayed to users, putting more likely definitions of a word higher on the definition list. However, because the translation-sense disambiguation feature was not yet implemented at the time of our user study, users were instead shown word definitions ranked according to their overall frequency of usage, as stated by the dictionary.

Getting Translations for Full Lines of Dialog

Translations for full lines of dialog are obtained from a subtitle track in the viewer's native language, if it was provided to the program. For example, if we gave Smart Subtitles a Chinese-language DVD that contained both English and Chinese subtitles, then it would extract translations for each line of dialog from the English subtitles. Alternatively, if we only have a transcript available, and not a subtitle in the viewer's native language, we rely on a machine translation service to obtain a translation. Either Microsoft's or Google's translation service can be used.

USER STUDY

Our user evaluations for Smart Subtitles was a within-subjects user study that compared vocabulary learning with this system, to the amount of vocabulary learning when using parallel English-Chinese subtitles. Specifically, we wished to compare the effectiveness of our system in teaching vocabulary to learners, compared to dual subtitles, which are believed to be among the best ways to learn vocabulary while viewing videos [17].

Materials

The video we showed was the first 5 minutes, and the next 5 minutes, in the first episode of the drama 我是老師 (I am a Teacher). This particular video was chosen because the vocabulary usage, grammar, and pronunciations were standard, modern spoken Chinese, as opposed to historical videos, which are filled with archaic vocabulary and expressions from literary Chinese. Additionally, the content of these video clips, consisting of conversations in classroom and household settings, was everyday, ordinary settings, so while there was still much unfamiliar vocabulary in both clips, cultural unfamiliarity with the video content would not be a barrier to comprehension. The Chinese and English subtitles were extracted from a DVD and OCR-ed to produce WebVTT-format subtitles.

Participants

Our study participants were 8 students who were enrolled in a third-semester Chinese class. Our study was conducted at the end of the semester, so participants had approximately 1.5 years of Chinese learning experience. 4 of our participants were male, and 4 were female. Participants were paid $20.

Research Questions

The questions our study seeked to answer were:

1. Will users learn more vocabulary using Smart Subtitles than with dual subtitles?
2. Will viewing times differ between the tools?
3. Will viewers' self-assessed enjoyability differ between the tools?
4. Will viewers' self-assessed comprehension differ between the tools?
5. Will summaries viewers write about the clips after viewing differ in quality between the tools?
6. Which of the features of Smart Subtitles will users find helpful and actually end up using?

Procedure

Viewing Conditions

Half of the participants saw the first clip with dual subtitles and the second with Smart Subtitles, while the other half saw the first clip with Smart Subtitles and the second with dual subtitles. For the dual subtitles condition we used the KM-Player video player, showing English subtitles on top and Chinese on the bottom. For the Smart Subtitles condition we used our software.

Before participants started watching each clip, we informed them that they would be given a vocabulary quiz afterwards, and that they should attempt to learn vocabulary in the clip while watching the video. We also showed them how to use the video viewing tool during a minute-long familiarization session on a seperate clip before the session. Participants

were told they could watch the clip for as long as they needed, pausing and rewinding as they desired.

Vocabulary Quiz

After a participant finished watching a clip, we evaluated vocabulary learning via an 18-question free-response vocabulary quiz, with two types of questions. One type of question, shown in Figure 4, provided a word that had appeared in the video clip, and asked participants to provide the definition for it. The other type of question, shown in Figure 5, provided a word that had appeared in the video clip, as well as the context in which it had appeared in, and asked participants to provide the definition for it.

12)
What does the word 数学 mean?
Meaning: _____
Did you already know the meaning of this word before watching this video?

Figure 4. Vocabulary quiz question asking for the definition of a word from the video, without providing the context it had appeared in.

2)
In the following sentence, what does the word 资格 mean?
我没有**资格**当老师
Meaning: _____
Did you already know the meaning of this word before watching this video?

Figure 5. Vocabulary quiz question asking for the definition of a word from the video, providing the context it had appeared in.

For both types of questions, we additionally asked the participant to self-report whether they had known the meaning of the word before watching the video, so that we could determine whether it was a newly learned word, or if they had previously learned it from some external source. This self-reporting mechanism is commonly used in vocabulary-learning evaluations for foreign-language learning [23].

Questionnaire

After completing the vocabulary quiz, we asked them to write a summary of the clip they had just seen, describing as many details as they could recall. Then, they completed a questionnaire where they rated on a 7-point likehardt scale, the following questions:

- How easy did you find it to learn new words while watching this video?
- How well did you understand this video?
- How enjoyable did you find the experience of watching this video with this tool?

Finally, we asked for free-form feedback about the user's impressions of the tool, and whether they would use the tool themselves.

RESULTS

From our study, we found that:

1. Users learned over twice as much vocabulary using Smart Subtitles than with dual subtitles.
2. Viewing times did not differ significantly between the tools.
3. Viewers' self-assessed enjoyability did not differ significantly between the tools.
4. Viewers' self-assessed comprehension did not differ significantly between the tools.
5. Quality ratings of summaries viewers wrote did not differ significantly between the tools.
6. Users made extensive use of both the word-level translations and the dialog-navigation features of Smart Subtitles, and described these as helpful.

Vocabulary Learning

Since the vocabulary quiz answers were done in free-response format, a third-party native Chinese speaker was asked to mark the learners' quiz answers as being either correct or incorrect. The grader was blind as to which condition or which learner the answer was coming from.

As shown in Figure 6, both the average number of questions which were correctly answered, as well as the number of new words learned, was greater with Smart Subtitles than with dual subtitles. We measured the number of new words learned as the number of correctly answered questions, excluding those for which they marked that they had previously known the word. There was no significant difference in the number of words known beforehand in each condition. A t-test shows that there were significantly more questions correctly answered ($t=3.49$, $df=7$, $p < 0.05$) and new words learned ($t=5$, $df=7$, $p < 0.005$) when using Smart Subtitles.

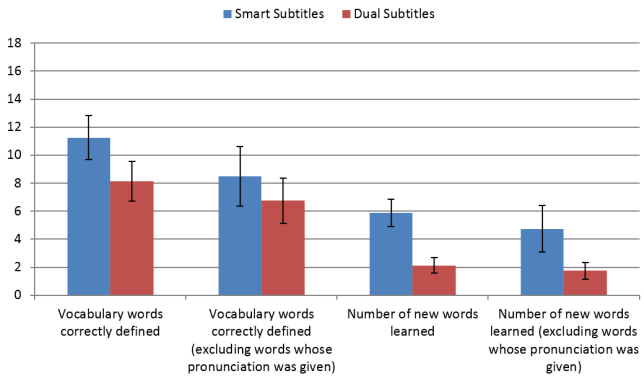

Figure 6. Vocabulary quiz results, with standard error bars.

Although we did not evaluate pronunciation directly, Smart Subtitles' display of pinyin appeared to bring additional attention towards the vocabulary pronunciations. In our vocabulary quizzes, we gave the participants a synthesized pronunciation of the word, in the event that they did not recognize the Chinese characters. We opted to provide a synthesized pronunciation, as opposed to the pinyin directly, as they would not have been exposed to pinyin in the Dual Subtitles condition. This, predictably, allowed participants to correctly define a few additional words in both conditions. That said, there was a slightly increased level of gain in the Smart Subtitles condition when pronunciation was provided, with an additional 1.1 words correctly answered on average, than in the Dual Subtitles condition, with an additional .3 words correctly answered on average.

We attribute this to certain participants focusing more attention on the pronunciation, and less on the Chinese characters,

in the Smart Subtitles condition. Indeed, one participant remarked during the vocab quiz for Dual Subtitles that she recognized some of the new words only visually and did not recall their pronunciations. We unfortunately did not ask participants to provide pronunciations for words, only definitions, so we cannot establish whether this held across participants.

Viewing Times

As shown in Figure 7, viewing times did not differ significantly between either of the two 5-minute clips, or between the tools. Viewing times were between 10-12 minutes for each clip, in either condition. Interestingly, the average viewing times with Smart Subtitles was actually slightly less than with dual subtitles, which is likely due to the dialog-based navigation features. Indeed, during the user study, we observed that users of Smart Subtitles would often review the vocabulary in the preceding few lines of the video clip by utilizing the interactive transcript, whereas users of Dual Subtitles would often over-seek backwards when reviewing, and would lose some time as they waited for the subtitle to appear.

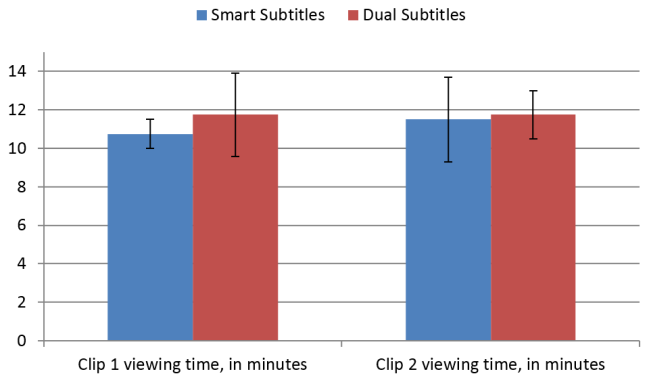

Figure 7. Viewing times, with standard error bars.

Self-Assessment Results

As shown in Figure 8, responses indicated that learners considered it easier to learn new words with Smart Subtitles, ($t=3.76$, $df=7$, $p < 0.005$), and rated their understanding of the videos as similar in both cases. The viewing experience with Smart Subtitles was rated to be slightly more enjoyable on average ($t=1.90$, $df=7$, $p=0.08$). Free-form feedback from participants indicates that an increased perceived ability to follow the original Chinese dialog contributed to the enjoyability result.

Summary Quality Ratings

After watching each video, partcipants wrote a summary describing the clip they had seen. An example is:

It was about a failed teacher whose students don't take him seriously (they leave his class, want to beat him up), and then a president who is angry about his daughter doing poorly at math after hiring an expensive tutor.

To evaluate the quality of the summaries written by our participants, we hired 5 Chinese-English bilingual raters to rate the summaries. The raters were hired from the oDesk contracting site, and were paid $15 apiece. Raters were first asked

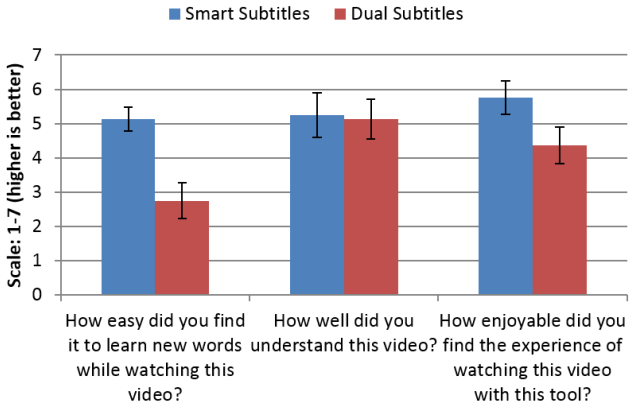

Figure 8. Self-assessment results, with standard error bars.

to view the clips, and write a summary in English to show that they had viewed and understood the clips. Then, we presented them the summaries written by students in random order. For each summary, we indicated which clip was being summarized, but the raters were blind as to which condition the student had viewed the clip under. Raters were asked to rate, on a scale of 1 (worst) to 7 (best):

- From reading the summary, how much does the student seem to understand this clip overall?
- How many of the major points of this clip does this summary cover?
- How correct are the details in this summary of this clip?
- How good a summary of this clip do you consider this to be overall?

To ensure that the rater was actually reading the summaries and was being consistent in their ratings, we included one of the summaries twice in the list of summaries the raters were asked to rate. Two of our raters did not notice that these summaries were identical and rated them differently, so we eliminated them for inconsistency. Our conclusion about the summary quality not being significantly different between conditions would still have remained the same if we had included the ratings from these two raters. Average rating results from the remaining three raters are shown in Figure 9.

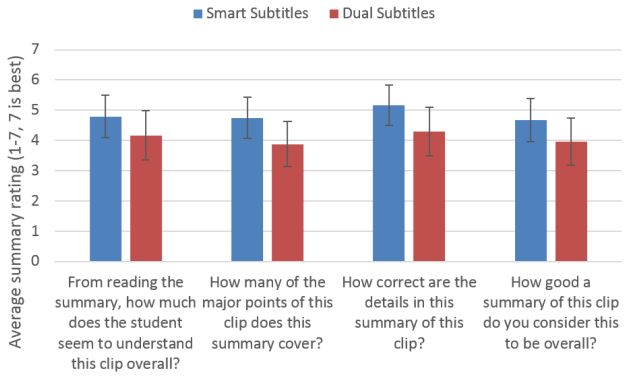

Figure 9. Average ratings given by bilinguals on the quality of the summaries written by learners in each viewing condition.

There was no significant difference in the quality of summaries written by the learners between the Smart Subtitles and Dual Subtitles conditions, according to any of the 4 quality metrics. The Krippendorff's alpha, which measures inter-rater agreement [12], across the raters was 0.7.

Free-form Feedback

We asked users to provide feedback about the watching experience in general, and whether they were interested in using the tool again. Of our 8 users, all expressed interest in using Smart Subtitles again. The written feedback that participants wrote indicated that they found most of the interface features to be helpful. Here is, for example, an anecdote describing the navigation features:

Yes! This was much better than the other tool. It was very useful being able to skip to specific words and sentences, preview the sentences coming up, look up definitions of specific words (with ranked meanings – one meaning often isn't enough), have pinyin, etc. I also really liked how the English translation isn't automatically there – I liked trying to guess the meaning based on what I know and looking up some vocab, and then checking it against the actual English translation. With normal subtitling, it's hard to avoid just looking at the English subtitles, even if I can understand it in the foreign language. This also helped when the summation of specific words did not necessarily add up to the actual meaning

The tone coloring feature in particular was not received as well. The only comment on tone coloring was by one participant who described it as distracting. This would suggest that we may wish to simply remove this feature, or make the tones more salient using another means, such tone numbers, which are more visually apparent than tone marks:

The tone coloring was interesting, but I actually found it a bit distracting. It seemed like I had a lot of colors going on when I didn't really need the tones color-coordinated. However, I think it's useful to have the tones communicated somehow.

Feature Usage during User Studies

During our user studies, we instrumented the interface so that it would record actions such as dialog navigation, mousing over to reveal vocabulary definitions, and clicking to reveal translations for the current line of dialog.

Viewing strategies with Smart Subtitles varied across participants, though all made at least some use of both the word-level and dialog-line translation functionality. Word-level translations were heavily used. On average, users hovered over words in 3/4 of the lines of dialog (standard deviation 0.22). The words hovered over the longest tended to be less common words, indicating that participants were using the feature for defining unfamiliar words, as intended. Participants tended to use dialog-line translations sparingly. On average they clicked on the translate button on only 1/3 of the lines of dialog (standard deviation 0.15). Combined with our observation that there was no decline in comprehension levels with Smart Subtitles, this suggests that word-level translations are often sufficient for learners to understand dialogs.

CONCLUSION AND FUTURE WORK

We have presented Smart Subtitles, an interactive transcript with features to help learners, such as vocabulary definitions on hover and dialog-based video navigation. They can be automatically generated from common sources of videos and subtitles, such as DVDs.

Our user study found that participants learned more vocabulary with Smart Subtitles than dual Chinese- English subtitles, and rated their comprehension and enjoyment of the video as similarly high. Independent ratings of summaries written by participants further confirm that comprehension levels when using Smart Subtitles match those when using dual subtitles. Given that users only viewed dialog-line translations for a third of the dialog lines when using Smart Subtitles, yet matched their comprehension levels with dual Chinese-English subtitles, this suggests that word-level translations are often sufficient for comprehension.

Unlike traditional subtitles, Smart Subtitles currently expect users to actively interact with them. However, we could potentially allow more passive usage by using statistical modelling to predict which words the viewer won't know and automatically showing their definitions.

Much work can still be done in the area of incorporating multimedia into learning. Our current Smart Subtitles system focuses on written vocabulary learning while watching of dramas and movies. However, we believe that augmenting video can also benefit other aspects of language learning. For example, we could incorporates visualizations for helping learn grammar and sentence patterns, and speech synthesis for helping learn pronunciation. We could also pursue further gains in vocabulary learning and comprehension, by dynamically altering the video playback rate, or by adding quizzes into the video to ensure that the user is continuing to pay attention.

Other multimedia forms can likewise benefit from interfaces geared towards language learning, though each form comes with its own unique challenges. For example, the current Smart Subtitles system can easily be used with existing music videos and song lyrics. However, the system would be even more practical for music if we could remove the need for an interactive display, and simply allowed the user to learn while listening to the music. Multimedia that is naturally interactive, such as Karaoke, likewise presents interesting opportunities for making boring tasks, such as practicing prononciation, more interesting to learners.

We hope our work leads to a future where people can learn foreign languages more enjoyably by being immersed and enjoying the culture of foreign countries, in the form of their multimedia, without requiring dedicated effort towards making the material education-friendly or even fully translating it.

REFERENCES

1. Bansal, M., DeNero, J., and Lin, D. Unsupervised translation sense clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Association for Computational Linguistics (2012), 773–782.

2. Bianchi, F., and Ciabattoni, T. Captions and Subtitles in EFL Learning: an investigative study in a comprehensive computer environment.
3. Bird, S., Klein, E., and Loper, E. *Natural language processing with Python*. O'Reilly, 2009.
4. Breen, J. WWWJDIC-a feature-rich WWW-based Japanese Dictionary. *eLEX2009* (2009), 31.
5. Danan, M. Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning 42*, 4 (1992), 497–527.
6. Danan, M. Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal des traducteursMeta:/Translators' Journal 49*, 1 (2004), 67–77.
7. d'Ydewalle, G. Foreign-language acquisition by watching subtitled television programs. *Journal of Foreign Language Education and Research 12* (2002), 59–77.
8. Fukunaga, N. "those anime students": Foreign language literacy development through japanese popular culture. *Journal of Adolescent & Adult Literacy 50*, 3 (2006), 206–222.
9. Harris, C., and Stephens, M. A combined corner and edge detector. In *Alvey vision conference*, vol. 15, Manchester, UK (1988), 50.
10. Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., and Narayanan, S. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research* (2011).
11. Koolstra, C. M., Peeters, A. L., and Spinhof, H. The pros and cons of dubbing and subtitling. *European Journal of Communication 17*, 3 (2002), 325–354.
12. Krippendorff, K. Computing krippendorff's alpha reliability. *Departmental Papers (ASC)* (2007), 43.
13. Kurohashi, S., Nakamura, T., Matsumoto, Y., and Nagao, M. Improvements of japanese morphological analyzer juman. In *Proceedings of The International Workshop on Sharable Natural Language* (1994), 22–28.
14. MDBG. CC-CEDICT Chinese-English dictionary. MDBG (2013).
15. Microsoft. Microsoft Office OneNote 2010. Microsoft (2010).
16. Mitterer, H., and McQueen, J. M. Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one 4*, 11 (2009), e7785.
17. Raine, P. Incidental Learning of Vocabulary through Authentic Subtitled Videos. JALT - The Japan Association for Language Teaching (2012).
18. Sakunkoo, N., and Sakunkoo, P. GliFlix: Using Movie Subtitles for Language Learning. In *UIST 2013 Adjunct*, ACM (2009).
19. Smith, R. An overview of the Tesseract OCR engine. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, IEEE (2007), 629–633.
20. Speed, E. Rikaikun. Google Chrome Web Store (2013).
21. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., and Manning, C. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, vol. 171, Jeju Island, Korea (2005).
22. W3C. WebVTT: The Web Video Text Tracks Format. W3C (2013).
23. Wesche, M., and Paribakht, T. S. Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Language Review 53*, 1 (1996), 13–40.
24. Zesch, T., Müller, C., and Gurevych, I. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *LREC*, vol. 8 (2008), 1646–1652.
25. Zuggy, B. SubRip (2011).