Usability Evaluation of Two Chinese Segmentation Methods in Subtitles to Scaffold Chinese Novice

Chih-Kai Chang

Department of Information and Learning Technology
National University of Tainan, Taiwan
chihkai@mail.nutn.edu.tw

Abstract. Recently the number of people who learn Chinese as a Foreign Language (CFL) increased. New comers, international students, and denizened spouses all need to improve their Chinese reading fluency and listening comprehension for daily communication and work requirements. However, not everyone gets opportunity for formal education in a language school. Thus, informal learning is very important for CFL learners in Taiwan. For novice Chinese learners, they should first master a skill to grouping Chinese words into meaningful chunks, i.e. Chinese segmentation. For instance, "老師對教育的貢獻" (teachers' contribution in education). After Chinese word segmentation, the sentence becomes "老師 (teachers)/ 對 (P)/ 教育 (education)/ 的 (DE)/ 貢獻 (contribution)" from "老/師/對/教/育/的/貢/獻". Consequently, this study used two Chinese segmentation methods to highlight meaningful and important word chunks in subtitles of Chinese videos and evaluate its usability for CFL learners. The first method adopted the top 800 and 1600 high-frequency words from an analysis report based on Academia Sinica Balanced Corpus of Modern Chinese to identify proper word segmentation in video subtitles and analyze its performance based on the forward maximum matching method. The statistical results show that most Chinese subtitles still remain unsegmented (62.3%) which means the Chinese subtitles in the videos are not appropriately segmented based on the corpus that contains the top 800 high frequency words. However, with the integration of the top 1600 high frequency words in the corpus, approximately 60% of the subtitles in each video are effectively segmented, and numerous unknown words still remain. Active phrases, idioms, and short phrases in Chinese subtitles may lead to the difficulty in word segmentation; moreover, the usability testing result of using high frequency words to conduct word segmentation is not significant.

The second method used natural language processing technique to split Chinese subtitles into its separate morphemes. The study adopted CKIP Chinese parser, which is a word segmentation tool for Chinese, to split subtitles according their part-of-speech tagging (i.e. grammatical tagging). The statistical results show that 97.26% subtitles are split, but the usability testing shows that subjective satisfaction is not good enough. To further investigation, we asked subjects to identify the "improper" word segmentation. For instance, the subtitle "接受治療很久了" (treated for a long time) will be split into "接受/治療/很/久/了", but most novices think that the proper segmentation should be "接受/治療/很久了". The "improper" rate is about 22.30% on average. In other words, the

A. Marcus (Ed.): DUXU/HCII 2013, Part II, LNCS 8013, pp. 28–37, 2013.

segmentation results from Chinese parser based on natural language processing technique are not best scaffolding for Chinese novice while watching videos with Chinese subtitles. The preliminary results of usability testing show that the second method can provide effective scaffolding for novice, but the granularity of chunked words may be too fine to read fluently sometimes (i.e. less than thirty percentage in results). Consequently, adaptation mechanism is required for learners to achieve the balance point of provided scaffolding between aforementioned two methods. For example, the Chinese function words, such as 很 and 了, serve only grammatical functions (i.e. they have no meaning by themselves). Those function words should not be separated out from subtitles for learning purpose. Further work is necessary to find out the proper granularity for chunking words, design adaptation mechanism of segmentation, and prevent segmentation errors in new or unknown words.

Keywords: Chinese as a foreign language, Chinese segmentation, subtitle manipulation, natural language processing, computer-assisted language learning.

1 Research Background

The study aims to design and develop a system for people who learn Chinese as a foreign language (CFL learners). In recent years, the rapid economic development in the Chinese region brings the worldwide craze of learning Chinese becoming the second international language after English. In addition to the formal education, learning Chinese is not limited to regular educational settings and textbooks. Many scholars pointed out that the development of technology makes the way of learning a second language or foreign languages become more diverse. The integration of technology in teaching Chinese internationally and the application of multimedia in language learning are booming. Currently, learning from watching videos is a popular trend; subtitles of the videos can effectively benefit second language learners' reading, vocabulary, and listening comprehension [1, 2, 3]. Researches indicated that the integration of subtitles in audiovisual teaching materials has been verified as an effective teaching strategy to promote listening and reading comprehension of a second language [4, 5, 6, 7]. Subtitles can aid learners to visualize messages of what they hear, especially for people whose language ability is unable to comprehend those messages; subtitles can increase learners' language comprehension ability. Moreover, subtitles can assist language learning because audio and visual messages of the brain can be transformed into a message map which is also a process of language learning [8]. Since 1980, subtitles have been considered a tool to enhance concentration, reduce anxiety, increase motivation, and help learners to instantly confirm messages of what they hear [9, 10, 11, 12]. Therefore, many studies verified that whether watching videos with subtitles can benefit learners more than those without subtitles [13, 14, 15, 16]; the results showed that learners who watched videos with subtitles performed better than those who didn't on the comprehension test at the time. With the features of the rapid spread of the network and the convenience of the subtitle software, many universities combine traditional language teaching methods with online resources in the U.S. [17, 18, 19, 20]. Therefore, learners can learn vocabulary and

syntax under a natural and relaxed environment to lower cultural shock, reduce learners' psychological panic and anxiety, and minimize the feelings of rejection by integrating features of multimedia [21]. Chinese language and videos with subtitles respectively have great potential for development; it is anticipated that combining those two elements together will have the synergistic effect of increasing the pleasure, motivation and learning ability of the learners.

With different cultures and language proficiency levels, learners may have difficulty to understand the contents of movies and learn a language. The basic semantic unit of movie subtitles, machine translation, full-text search index, and sentence comprehension is a word which plays an important role in understanding of words and reading comprehension for learners. However, Chinese sentences are sequences of words delimited by white spaces; in Chinese text, sentences are represented as strings of Chinese characters without similar natural delimiters. Learners must possess the word segmentation ability to identify the sequence of words in a sentence and mark boundaries in appropriate places. Word segmentation is to divide a string of written sentences into component words so that readers can accurately understand its meaning. For instance, "老師對教育的貢獻"(teachers' contribution in education). After Chinese word segmentation, the sentence becomes "老師(teachers)/對(P)/教育(education)/的(DE)/貢獻(contribution)" from "老/師/對/教/育/的/貢/獻" (Fang, 2008). "Word ambiguity" and "unknown word" are two major segmentation problems that affect the accuracy of Chinese word segmentation performance. The first problem is associated with typical ambiguity problems that may lead to unexpected segmentation results [22]. For example, the sentence, "下雨天留客天留我不留", has several explanations because of the different ways of word segmentations. Therefore, word segmentation becomes the first task when processing Chinese text. The text of the Chinese corpus has been segmented, and words are separated by white spaces which can enhance and increase learners' ability to identify words and help learners to combine those words after understanding their meanings to solve the inability of recognizing words.

The study uses the maximum matching algorithm and the top 800 and 1600 high-frequency words from the Academia Sinica Balanced Corpus of Modern Chinese to segment Chinese subtitles in videos so that learners can understand the meaning of Chinese subtitles and learn Chinese language [23]. The Maximum Matching (MM) algorithm, a most commonly used dictionary-based approach, is used to initially segment the text by referring to a pre-compiled dictionary. The algorithm starts at the first character in a sentence and attempts to find the longest matching word in the text starting with that character. If a word is found, the maximum-matching algorithm marks a boundary at the end of the longest matching word, then, begins the same longest match search starting at the character following the match. If no match is found in the text, the MM algorithm simply skips that character and begins the search starting at the next character to obtain an initial segmentation. The accuracy of the MM algorithm is expected to be more than 90%. The forward maximum matching method starts with the beginning of the sentence which is from left to right, attempting to find the longest matching word in the given sentence and then repeating the process until it reaches the end of the sentence. The backward maximum matching

approach starts with the end of the sentence, finding the longest matching word in the database and repeating the process until it reaches the beginning of the sentence (See Table 1). Since the MM algorithm segments words based on a pre-compiled dictionary, it is unable to deal with unknown words, not listed in the dictionary. For example, "亞洲巨星五月天" (Asian superstar Mayday) is segmented as "亞洲(Asia)/巨星(superstar)/五(five)/月(month)/天(day)" (See Table 1).

Table 1. Forward and backward maximum matching examples

Example	Forward MM	Backward MM
才能夠完成	才能/夠/完成	才/能夠/完成
家庭和諧	家庭/和諧	家庭/和諧
亞洲巨星五月天	亞洲/巨星/五月/天	亞洲/巨星/五/月/天

The completeness of the dictionary to a large extent determines the degree of success for segmenting words using this approach. Therefore, learners can easily understand the contents of the videos when the Chinese subtitles have appropriate word segmentations so that their learning effectiveness of Chinese language can be increased.

2 Research Method

2.1 The System Design

The study uses a programming language, Python, to develop a word segmentation system to segment Chinese subtitles according to the corpus that contains the top 800 and 1600 high frequency words so that learners can understand the contents of Chinese videos and learn Chinese language. Figure 1 displays the framework of the system. The system functions are to select subtitles, segment words, select high frequency words, display segmented subtitles, collect and analyze data. The users (CFL learners) learn Chinese as a foreign language. They can select Chinese subtitles from the given videos for word segmentation, and the system displays the segmented subtitles so that they can easily comprehend the contents of the videos and increase their motivation to learn Chinese.

The study randomly selected two videos with Chinese subtitles that introduce Taiwan as samples from a university's library. The CFL learners can select one of the two provided videos with the SRT subtitle format. The study adopts the top 800 and 1600 high frequency words from the "Word List with Accumulated Word Frequency in Sinica Corpus" from Institute of Linguistics of Academia Sinica to develop a corpus for Chinese word segmentation [24, 25]. The SRT subtitle format mainly consists of three parts, serial numbers, timelines, and texts of the subtitles. The function of "select subtitles" can extract serial numbers and timelines and then process the

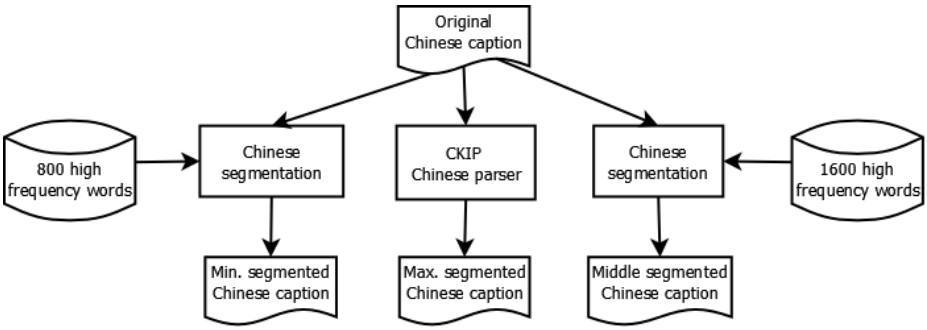

Fig. 1. The flowchart of the process to generate adaptive Chinese captions

remaining strings of subtitles. Subtitles within each timeline may have multiple lines of texts required to separate them into multiple strings of subtitles for further word segmentation in each subtitle file. The study adopts the forward maximum matching method to carry out word segmentation of subtitles in Chinese videos. The algorithm starts from the a given point in the sentence, finding the longest string of words that matches a word entry in a corpus and then repeating the process until it reaches the end of the subtitles. For example, if a string of words is "相當重要的生態資源" (a very important ecological resource), the system matches it with the longest word entry in the corpus. If the system cannot find a match, the system repeats the process until it finds the string of words, "相當", in the corpus. Then, the system removes it, "相當", from the original sentence and starts to match the remaining string of words, "重要的生態資源", with the word entries in the corpus until the end. The anticipated result is "相當/重要/的/生態/資源". The corpus for Chinese word segmentation is developed based on the "Word List with Accumulated Word Frequency in Sinica Corpus" from Institute of Linguistics of Academia Sinica. The word lists of the corpus allows users to search for the word frequency ranking, the frequency by words, the words by frequency, and the cumulative frequency; the searching results display information with regard to the high frequency words, their rankings, frequencies, percentages, and cumulative percentage (See Table 2).

Table 2. Example of modern Chinese word frequency ranking from 792 to 800

Rank	Word	Frequency	Percent	Cumulation
792	教師(Na)	748	0.015	56.482
792	要(VE)	748	0.015	56.498
792	否則(Cbb)	748	0.015	56.513
795	重視(VJ)	747	0.015	56.528
795	工具(Na)	747	0.015	56.543
797	實施(VC)	746	0.015	56.559
798	臉(Na)	745	0.015	56.574
798	節目(Na)	745	0.015	56.589
798	法(Na)	745	0.015	56.605

The study adopts the Chinese Natural Language Statistical Toolkit (CNLSTK), a natural language processing search tool for character-based texts, to calculate word frequency. The CNLSTK tags texts and develops a corpus for supporting researches on Chinese language. The corpus is a research resource used for processing natural Chinese language and providing tools for retrieving, mining, and analyzing. In comparison with the Natural Language Toolkit (NLTK), the CNLSTK mainly focuses on statistical functions and places less emphasis on semantics, pronunciation, and part of speech correction. The CNLSTK currently supports UTF-16 encoding; other encodings should be converted to UTF-16. Before manipulating the system, users need to create a folder for saving one or multiple files or folders for saving multiple files. Then, users create indexes (建立索引標籤) for retrieving information. The current search functions of the toolkit are to search for string frequency, to list all files' names, to get the frequency distribution of the given string in files, to retrieve the full text by file names, to retrieve the backward sequence of full text by file names, to get a concordance of strings, to get a backward concordance of strings, to narrow concordance down of the given files, to narrow down backward concordance results, to search for previous or next characters, to get statistical information of previous or next characters, and to have the distribution of two strings in all files. The study uses the CNLSTK to process the corpus that contains the top 800 and 1600 high frequency words to match words of the subtitles in the videos. If the words of the corpus don't match the words of the subtitles at all, those words are removed from the corpus. Therefore, the words of the subtitles match those in the corpus at least once or more times. The remaining words in the corpus are less than the original quantity of the words that may increase the speed of segmenting the subtitles. In addition, the toolkit provides information of the frequency, concordance, location, and distribution of the words showed in the subtitles that help CFL learners to understand semantics of Chinese language and apply those words in making sentences.

3 Results

3.1 Results of Chinese Word Segmentation

The average length of the sample videos is 90 minutes, and the standard deviation is 0. The number of the average segmented sentences in each video is 255.5, and the standard deviation is 145.66. The total number of the average words in the subtitles of the each video is 5786, and the standard deviation is 760.85. The randomly selected two videos that introduce Taiwan are intellectual videos which contain professional terminology used in subtitles that result in unknown words after Chinese word segmentation. The subtitles of the two videos are segmented according to the corpus that contains the top 800 high frequency words. The statistical results show that 37.7% of the subtitles in each video are segmented, and the standard deviation is 0.04; the average percentage of the unchanged sentences in each video is 62.3%, and the standard deviation is 0.04. The results indicate that using the top 800 high frequency words to conduct the word segmentation is not effective and significant.

Further, the subtitles of the two videos are segmented based on the corpus that contains the top 1600 high frequency words. The statistical results show that 61.55% of the subtitles are segmented in each video, and the standard deviation is 0.09;

the percentage of the unchanged sentences is 38.45%, and the standard deviation is 0.09; the average percentage of the segmented sentences is 61.55% which indicates that 60% of the subtitles can be effectively segmented, and one of the video has 67.58% of the segmented subtitles. In comparison with the top 800 high frequency words, the accuracy of the segmented subtitles of the two videos increase after using the top 1600 high frequency words to conduct word segmentation. The intellectual videos contain professional terminology that results in numerous unknown words. Moreover, active phrases, idioms, and short phrases in Chinese subtitles may lead to the difficulty in word segmentation and the segmentation errors. Therefore, the study should expand the developed corpus to effectively conduct word segmentation.

3.2 Word Segmentation Results of the Subtitles in the Videos

The study found that the length of the segmented subtitles in the two videos displayed on the screen becomes longer than the original ones based on the corpus that contains the top 800 and 1600 high frequency words because of adding segmentation boundaries (See Figure 2 and 3). The statistical results show that the number of the unknown words is more than the number of the properly segmented words in the subtitles which cannot enhance and increase learners' ability to recognize and identify the words because intellectual videos contains numerous professional terminology used in subtitles.

Fig. 2. The original subtitle (see left) and the segmented subtitle based on the corpus that contains the 800 high frequency words (see right)

Fig. 3. The subtitle (see right) based on the corpus that contains the 1600 high frequency words and the subtitle (see right) segmented by CKIP

3.3 Ambiguity Analysis

Mostly word segmentation is referred to the corpus which may not contain all the words so that the segmentation result may not be correct and cause ambiguity problems that usually happen to an unsegmented string of words which can be segmented into different ways based on the semantic structure of an article. For example, the string of words, "我們可以感受到樹幹裏流動的樹液", are segmented as "我們/可以/感/受到/樹幹裏流動的樹液" referred to the corpus. However, the composition ambiguity leads to the misinterpretation of the text; the correct word segmentation is "我們/可以/感受/到/樹幹裏流動的樹液".

4 Conclusion and Implication

The study adopts the forward maximum matching approach to conduct Chinese word segmentation. The system cannot accurately conduct Chinese subtitle segmentation and is unable to precisely identify and recognize Chinese subtitles. The statistical results show that most Chinese subtitles still remain unsegmented (62.3%) which means the Chinese subtitles in the videos are not appropriately segmented based on the corpus that contains the top 800 high frequency words. However, with the integration of the top 1600 high frequency words in the corpus, approximately 60% of the subtitles in each video are effectively segmented, and numerous unknown words still remain. Active phrases, idioms, and short phrases in Chinese subtitles may lead to the difficulty in word segmentation; moreover, the result of using high frequency words to conduct word segmentation is not significant.

It is anticipated to integrate other word segmentation models or algorithms to increase the accuracy rate of the word segmentation; it is worthy to expand the corpus to improve the word segmentation rate. Moreover, combining both forward and backward maximum matching methods is an alternative to discover the segmentation ambiguity and increase learning effectiveness.

The future study will integrate the top 3500 high frequency words in the corpus to increase the accuracy rate of the word segmentation and improve the word segmentation rate. Moreover, the corpus can be expanded by adding specialized terms to reduce the unknown word recognition, increase the rate of indentifying terminology, and enhance the proportion of word segmentation so that CFL learners can easily understand the contents of the videos and learn Chinese. The word segmentation of the current system may yield ambiguity errors that result in misplacement of segmentation boundaries in the subtitles. Therefore, it is necessary to integrate other word segmentation systems in the study to eliminate segmentation ambiguities.

Acknowledgement. The research reported in this paper has been supported by the National Science Council in Taiwan under the research project number NSC 100-2631-S-001-001, NSC 100-2628-S-024-001-MY3, and NSC 101-2511-S-024-007-MY2.

References

1. Chun, D.M., Plass, J.L.: Research on text comprehension in multimedia environments. Language Learning & Technology 1(1), 1–35 (1997)
2. Plass, J.L., Chun, D.M., Mayer, R.E., Leutner, D.: Supporting visual and verbal learning preferences in a second language multimedia learning environment. Journal of Educational Psychology 90(1), 25–36 (1998)
3. Danan, M.: Reversed subtitling and dual coding theory: New directions for foreign language instruction. Language Learning 42(4), 497–527 (1992)
4. Borras, I., Lafayette, R.: Effects of multimedia courseware subtitling on the speaking performance of college students of French. The Modern Language Journal 78(1), 61–75 (1994)
5. Danan, M.: Captioning and subtitling: Undervalued language learning strategies. Meta 49(1), 67–77 (2004)
6. Garza, T.J.: Evaluating the use of captioned video materials in advanced foreign language learning. Foreign Language Annals 24(3), 239–258 (1991)
7. Markham, P.L., Peter, L.: The influence of English language and Spanish language captions on foreign language listening/reading comprehension. Journal of Educational Technology Systems 31(3), 331–341 (2003)
8. Doughty, C.J.: Effect of instruction on learning a second language: A critique of instructed SLA research. In: VanPatten, B., Williams, J., Rott, S. (eds.) Form-Meaning Connections in Second Language Acquisition, pp. 181–202. Lawrence Erlbaum Associates, Mahwah (2004)
9. Burger, G.: Are TV programs with video subtitles suitable for teaching listening comprehension? Zielsprache Deutsch 20(4), 10–13 (1989)
10. Froehlich, J.: German videos with German subtitles: A new approach to listening comprehension development. Die Unterrichtspraxis/Teaching German 21(2), 199–203 (1988)
11. Grimmer, C.: Supertext English language subtitles: A boon for English language learners. EA Journal 10(1), 66–75 (1992)
12. Vanderplank, R.: The value of teletext sub-titles in language learning. English Language Teaching Journal 42(4), 272–281 (1988)
13. Baltova, I.: Multisensory language teaching in a multidimensional curriculum: The use of authentic bimodal video in core French. The Canadian Modern Language Review 56(1), 32–48 (1999)
14. Markham, P.L.: Captioned television videotapes: Effects of visual support on second language comprehension. Journal of Educational Technology Systems 21(3), 183–191 (1993)
15. Markham, P.L.: Captioned videotapes and second-language listening word recognition. Foreign Language Annals 32(3), 321–328 (1999)
16. Neuman, S.B., Koskinen, P.: Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. Reading Research Quarterly 27, 94–106 (1992)
17. Chenoweth, N.A., Murday, K.: Measuring student learning in an online French course. CALICO Journal 20(2), 285–314 (2003)
18. Chenoweth, N.A., Ushida, E., Murday, K.: Student learning in hybrid French and Spanish courses: An overview of Language Online. CALICO Journal 24(1), 285–314 (2006)
19. Sanders, R.F.: Redesigning introductory Spanish: Increased enrollment, online management, cost reduction, and effects on student learning. Foreign Language Annals 38(4), 523–532 (2005)

20. Scida, E.E., Saury, R.E.: Hybrid courses and their impact on student and classroom performance: A case study at the University of Virginia. CALICO Journal 23(3), 517–531 (2006)
21. Chen, L.F.: From file appreciation to the curriculum design and experiment of teaching Chinese. National Taiwan Normal Unversity Mandarin Training Center, Taipei (2007)
22. Fang, S.L.: Segmentation and pronunciation annotation in Mandarin Chinese. Master thesis. National Tsing Huan University, Taiwan (2008)
23. Tang, J.H.: A Chinese speech synthesis system improved by a word segmentation method. Master thesis. National Tsing Huan University, Taiwan (2010)
24. Cheng, C.C.: Word-focused extensive reading with guidance. In: Thirteenth International Symposium on English Teaching, pp. 24–32. Crane Publishing Co., Taipei (2004)
25. Cheng, C.C.: From Digital Archives to Digital Learning: Determining Sentence Readability. In: Bi-Jiaoda Conference on Corpus Linguistics and English Testing, Shanghai, June 13 (2005)