SUTRA TRANSLATION USING RECENT ADVANCES IN ARTIFICIAL INTELLIGENCE

by Khanh T. Tran*

ABSTRACT

Artificial intelligence (AI), especially machine learning, has begun to impact many activities in our daily life. This paper will focus on the application of recent advances in machine translation to Buddhism, namely the translation of Buddhist sutras, especially those in the Mahayana or Northern Tradition Tripitaka, from Chinese into English. The Taisho Tripitaka (Three Baskets) is composed of over 70 million Chinese characters and 2372 texts divided into sutras, vinayas (precepts) and sastras (commentaries).

In recent years, many international organizations in Japan and USA have translated Mahayana texts. Yet there are still too many sutras that have not yet been translated. Since 2006, Tuệ Quang Wisdom Light Foundation has committed to translate the Taisho texts into Vietnamese, English and French. At the present time, we have developed computer software based on the word substitution approach that performs the rough translation of the entire Taisho Tripitaka in less than 60 hours.

Our rough translations based on word substitution are more accurate than those from Google Translate but they are still full of grammar errors and, hence, require intense manual editing. For the last several years, we have sponsored several editors to develop the Vietnamese version of the Taisho Tripitaka which is nearly completed. For the English version, we plan to use the latest advances in artificial intelligence (AI) to enhance the accuracy of the computer translation. This paper will review the recent

* Tuệ Quang Wisdom Light Foundation, Henderson, Nevada, USA

advances in machine learning, especially the neural machine translation (NMT). NMT translates a sequence of words and entire phrases using large artificial neural networks, much like neurons in human brain. In addition to a dictionary, NMT learns from a large database of Chinese-English parallel texts. This critical Buddhist corpus is taken from well-known translated sutras such as the Amitabha Sutra, The Diamond Sutra, the Lotus Sutra and the Flower Ornament Sutra. NMT is expected to be more accurate than the word substitution approach and the expense of much more computer resources. We will develop a NMT app based on the library TensorFlow developed by Google Brain and present some preliminary NMT translations at the conference.

We fully realize that the translation of an English Tripitaka requires a multi-year effort from numerous experts and editors. However, we also believe that all Buddhists, lay or monastic, can participate and contribute to this important Buddha-work. To encourage the wide participation of all Buddhists, we are making the English computer translations available online at The Compassion Network of Rev. Guo Cheen. We hope to expedite the Tripitaka translation by the active participation of other Buddhists through Open Wiki.

INTRODUCTION

Artificial intelligence (AI), especially machine learning, has begun to impact many activities in our daily life. In recent years, major advances have occurred in several areas, from medical diagnostics to autonomous driverless cars. Since 2015, rapid developments have also been realized in the field of machine translation (MT). This paper will focus on the application of recent MT advances to Buddhism, namely the translation of Buddhist sutras, especially those in the Mahāyāna or Northern Tradition Tripiṭaka, from Chinese into English.

THE TAISHŌ TRIPIṬAKA

Mahāyāna texts are organized into Sutras (discourses), Vinayas (precepts) and *Ś*āstras (commentaries). Collectively they are known as Tripiṭaka (Three Baskets). These texts have been primarily translated from Sanskrit into Chinese for over 1200 years, from the Later Han dynasty (2nd century) until the Yuan dynasty (13th

century). The first translated text is the Sutra in Forty Two Sections in 76 BCE. Among numerous translators, the most famous ones include Kumārajīva (344-413) and Xuanzang (599-644).

Among several Tripiṭaka versions, the most widely used today is the Taishō Tripiṭaka. For nine years (1924-1932), this Tripiṭaka was compiled by two scholars at the University of Tokyo, Takakusu Junjirō (1866-1945) and Watanabe Kaikyoku (1872-1932). It was published in 85 volumes under the Taishō reign and, hence, its name Taishō Tripiṭaka. Mahāyāna texts are primarily in volumes 1-55 and 85 (the other volumes belong to Japanese Buddhism). With over 70 million Chinese characters, they are organized into 9035 fascicles and 2372 texts. The length of texts varies greatly, from the short Heart Sutra (the version by Xuanzang has only 260 words) to the voluminous Great Prajñā Sutra (600 fascicles).

Advances in computer technology in recent years allow the development of an electronic Tripiṭaka. Since 1998, the Chinese Buddhist Text Association (CBETA) has published a CD/DVD containing the Mahāyāna texts of the Taishō Tripiṭaka. The most recent version can be downloaded from the CBETA website (http://www.cbeta.org). The availability of digital texts such as CBETA greatly facilitates their translation, especially computer-based, from Chinese into English.

MANUAL ENGLISH TRANSLATION

As described above, the Mahāyāna Tripiṭaka is composed of numerous texts. Many individual scholars and organizations in USA, Japan and elsewhere have been involved in translating these texts. Among these are the Buddhist Text Translation Society (http://www.cttbusa.org/cttb/btts.asp) and BDK America of the City of Ten Thousand Buddhas (http://www.bdkamerica.org). These efforts are labor-intensive, time-consuming and costly. They also require several experts since they are done manually. They have translated several popular texts such as the Amitābha Sutra, the Diamond Sutra, the Lotus Sutra and the Flower Ornament Sutra. A list of translated sutras is available at http://vnbaolut.com/sutras/ and http://fodian.net/world/ . Of 2372 texts in the Taishō Mahāyāna Tripiṭaka, less than 10% of them have been

translated. Thus, there are still too many sutras that have not yet been translated.

COMPUTER-BASED ENGLISH TRANSLATION

Buddhism has been a major religion in Vietnam for over 2,000 years. Yet we do not have a complete Tripiṭaka in Vietnamese. Until the beginning of the 20th century, traditional Chinese was the official writing language. Today, most Vietnamese do not know how to read Chinese. Further, many translations of Buddhist texts still are heavy with Chinese terms that are difficult to understand. Since 2006, Tuệ Quang Wisdom Light Foundation has committed to translate the Taishō texts into Vietnamese, English and French (Tran and Tran 2006). At the present time, we have developed computer software based on the word substitution approach and a comprehensive dictionary of Buddhist terms. In our translation work we have compiled a multi-lingual dictionary of over 40000 Buddhist terms in Chinese, Sanskrit, Vietnamese and English. The translation tool performs the rough translation of the entire Taishō Tripiṭaka in less than 60 hours.

Our preliminary translations based on word substitution are more accurate than those from Google Translate without a specialized Buddhist corpus (see Appendix B). However, they are still full of grammar errors and, hence, require extensive efforts of manual editing. For the last several years, we have sponsored several editors to develop the Vietnamese version of the Taishō Tripiṭaka which is nearly completed (http://vnbaolut.com/daitangvietnam/).

For the English Tripiṭaka, we are using the latest advances in artificial intelligence (AI) to enhance the accuracy of the computer translation. In recent years advances in machine learning and machine translation in particular are focused on the neural machine translation (NMT). NMT is a relatively new approach which translates a sequence of words and entire phrases using large artificial neural networks, much like neurons in human brain. A NMT model often consists of an encoder and a decoder. The encoder extracts a fixed-length representation from a variable-length input sentence, and the decoder generates a correct translation from this

representation. In addition to a dictionary, NMT learns from a large database of Chinese-English parallel texts. We have assembled a Buddhist corpus of over 100000 entries that have been taken from well-known translated sutras such as the Amitābha Sutra, The Diamond Sutra, the Lotus Sutra and the Flower Ornament Sutra. Our translation app is based on OpenNMT from Harvard University (2019) and other Deep Learning algorithms in the library TensorFlow that was developed by Google (2019).

Appendix A shows a sample translation of the Diamond Sutra. As shown, each Chinese line is followed by three translated lines: Han-Viet, Vietnamese and English. From similar texts, the Chinese-English bitext can be extracted for corpus usage. Separate files in Chinese or English can also be easily obtained as UTF-8 text files. NMT is expected to be more accurate than the word substitution approach at the expense of much more computer resources. Since 2016, NMT has been used by online translators from Google and Microsoft. Both of these translation services does a good job in translating newspaper articles and business letters. However, as shown in Appendix B, Google Translate does a poor job in translating the first few sentences of the Diamond Sutra. This poor performance is due to its unfamiliarity with Buddhist terminology and lack of a specialized Buddhist parallel corpus.

PROPOSED TRANSLATION PROGRAM

Based on our experience with the development of a Vietnamese Tripiṭaka, we propose the following 5-stage program:

1. Refine our translation software by implementing the latest advances in Chinese-English NMT including context, word order and lexical analysis. A critical and time-consuming input is to increase the size of the Buddhist parallel corpus -- the bigger the better. The refined software will be tested by applying to popular sutras such as the Amitābha Sutra, the Diamond Sutra, the Medicine Buddha Sutra, the Lotus Sutra and the Sixth Patriarch's Platform Sutra;

2. Apply the software to translate the Chinese Tripiṭaka. With the improved software in Step 1, the accuracy of the translated texts will be increased;

3. Edit the translated texts. Due to the large number of texts (9035), it will be necessary to enlist several editors from the United States and elsewhere. They can be drawn from various Buddhist institutes, universities as well as volunteers from Buddhist temples and Dharma practicing groups;

4. Review and approve by the Masters, and

5. Publish the final texts online, by electronic means (CD/DVD) for free distribution.

We fully realize that the translation of an English Tripiṭaka requires a multi-year effort from numerous experts and editors. However, we also believe that all Buddhists, lay or monastic, can participate and contribute to this important Buddha-work. To encourage the wide participation of all Buddhists, we are making the English computer translations available online at The Compassion Network of Rev. Guo Cheen (http://thecompassionnetwork.org/). We hope to expedite the Tripiṭaka translation by the active participation of other Buddhists through Open Wiki. Any Buddhist who abides by the Five Precepts is welcome to register as an editor and help to: 1. Translate from Chinese to English, 2. Review the English against the Chinese, and 3. Edit and proofread the English. Let's pray to the Buddhas for a complete English Tripiṭaka soon!

References

Chinese Buddhist Electronic Text Association (CBETA).The 2018 Taisho DVD. Available from website http://www.cbeta.org

Google Brain (2019). TensorFlow. Available from website https://tensorflow.org

Harvard University (2019). OpenNMT. Available from website http://opennmt.net

Tran, Khanh T. and Tran, Huyen T. (2006). Computer Translation of the Chinese Taishō Tripiṭaka. Available at http://vnbaolut.com/sutras/ComputerTranslationoftheChineseTripiṭaka.pdf

Tue Quang Wisdom Light Foundation, More information are available from Websites

http://vnbaolut.com/sutras/ (English)

http://vnbaolut.com/daitangvietnam/ (Vietnamese)

Appendix A

Sample Translation of the Diamond Sutra

Note: A complete translation of the Diamond Sutra is available at Tue Quang Wisdom Light Foundation website http://vnbaolut.com/sutras/TQtranslate_DiamondSutra.pdf

Taishō Tripiṭaka Vol. 8, No. 235 金剛般若波羅蜜經

#CBETA Chinese Electronic Tripiṭaka V1.13 (UTF-8) Normalized Version

金剛般若波羅蜜經

Kim Cương Bát Nhã Ba La Mật Kinh

Kinh Kim Cương Bát Nhã Ba La Mật

Diamond Prajna Paramita (Perfect Wisdom) Sutra

姚秦天竺三藏鳩摩羅什譯

Diêu Tần Thiên Trúc Tam Tạng Cưu Ma La Thập dịch

Diêu Tần Thiên Trúc Tam Tạng Cưu Ma La Thập dịch

Translated by Indian Tripiṭaka Master Kumarajiva in the Dao Qin Dynasty

如是我聞。一時佛在舍衛國祇樹給孤獨園。

Như thị ngã văn. Nhất thời Phật tại Xá vệ quốc Kì-Thọ Cấp-Cô-Độc viên.

Tôi nghe như vầy. Một thuở nọ, Đức Phật ở nước Xá vệ, trong vườn Kì-Thọ của Ông Cấp-Cô-Độc.

Thus have I heard. Once Buddha resided in the country of Śrāvastī, at the Jeta (Victory) Grove of Anathapindika (Provider to the Orphans and the Solitaires).

與大比丘眾千二百五十人俱

Dữ đại bỉ khâu chúng thiên nhị bách ngũ thập nhân câu

Với đại chúng gồm một ngàn hai trăm năm mươi vị Tỳ kheo

With a grand assembly of one thousand two hundred fifty Bhiksus (monks)

爾時世尊食時著衣持鉢入舍衛大城乞食。

Nhĩ thời Thế tôn thực thời trước y trì bát nhập Xá-Vệ đại thành khất thực

Lúc bấy giờ, gần đến giờ ăn, Đức Thế Tôn đắp y cầm bát, vào thành lớn Xá-Vệ khất thực

At that time, near meal time, World-Honored One put on a robe, held his alm bowl and entered the great city of Shravasti to beg for alms

於其城中次第乞已。還至本處飯食訖。

Ư kỳ thành trung thứ đệ khất dĩ. hoàn chí bổn xứ phạn thực cật

Trong thành đó, sau khi khất thực tuần tự từng nhà, Đức Phật trở về tịnh xá. Dùng cơm xong,

In that city, after begging successively from door to door, he returned to his retreat. When he finished eating,

收衣鉢洗足已敷座而坐。時長老須菩提在大眾中。

thu y bát tẩy túc dĩ phu tọa nhi tọa . Thời Trưởng Lão Tu Bồ Đề tại Đại chúng trung.

cất y bát, rửa chân, trải tọa cụ và ngồi xuống. Bấy giờ, Trưởng Lão Tu Bồ Đề (Thiện Hiện), ở trong Đại chúng,

he put away his robe and his alm bowl, washed his feet, spread a seating mat and sat down. At that time, Venerable Subhūti (Good Apparition), in the assembly,

即從座起偏袒右肩右膝著地。

tức tùng toạ khởi thiên đản hữu kiên hữu tất trước địa.

từ chổ ngồi đứng dậy, trịch áo vai phải, quỳ gối phải sát đất,

rose from his seat, uncovered his right shoulder, knelt on the right knee to the ground,

合掌恭敬而白佛言。希有世尊。如來善護念諸菩薩。

hợp chưởng cung kính nhi bạch Phật ngôn. Hi hữu Thế tôn. Như-Lai thiện hộ niệm chư Bồ Tát

cung kính chấp tay và bạch cùng Đức Phật rằng: Hi hữu thay Đức Thế Tôn, Đức Như-Lai hay khéo nâng đở các Bồ Tát,

and, with his palms joined together, respectfully said to Buddha: It's extraordinary, World-honored One, the Thus-Come-One (Tathagata) is well supportive of all Bodhisattvas,

善付囑諸菩薩。世尊。善男子善女人。

thiện phó chúc chư Bồ Tát. Thế tôn. Thiện nam tử Thiện nữ nhân

hay khéo giao phó cho các Bồ Tát. Bạch Thế Tôn, khi Thiện nam Thiện nữ

and entrusting so well all Bodhisattvas. World-honored One, if good men and good women

發阿耨多羅三藐三菩提心。

phát a nậu đa la tam miệu tam Bồ Đề tâm

phát tâm Vô Thượng Chánh Đẳng Chánh Giác

engender the mind of supreme and perfect enlightenment

應云何住云何降伏其心。佛言。善哉善哉。須菩提。如汝所說。

ưng vân hà trụ vân hà hàng phục kỳ tâm. Phật ngôn. Thiện tai Thiện tai. Tu Bồ Đề. như nhữ sở thuyết

thì phải trú ở tâm ấy như thế nào, và phải sửa tâm mình như thế nào? Đúc Phật dạy: Lành thay! Lành thay! Này Tu Bồ Đề , như Ông nói ,

how should they abide there and how should they subdue their mind ? Buddha said: Excellent! Excellent ! Subhūti, just as you say,

如來善護念諸菩薩。善付囑諸菩薩。

Như-Lai thiện hộ niệm chư Bồ Tát. thiện phó chúc chư Bồ Tát

Như-Lai hay khéo bảo hộ và nhớ nghĩ các Bồ Tát, hay khéo giao phó các Bồ Tát

The Thus-Come-One (Tathagata) always protects and is well mindful of all Bodhisattvas and is well enstrusting all Bodhisattvas

汝今諦聽。當為汝說。善男子善女人。

nhữ kim đế thính. đương vi nhữ thuyết. Thiện nam tử Thiện nữ nhân.

Hãy nghe kỹ ! Ta sẽ vì Ông mà dạy cho hàng Thiện nam Thiện nữ,

Listen carefully! Because of you, I will instruct how good men and good women,

發阿耨多羅三藐三菩提心。

phát a nậu đa la tam miệu tam Bồ Đề tâm

phát tâm Vô Thượng Chánh Đẳng Chánh Giác

when they engender the mind of supreme and perfect enlightenment,

應如是住如是降伏其心。唯然世尊。願樂欲聞。

ưng như thị trụ như thị hàng phục kỳ tâm. Duy nhiên Thế tôn. nguyện lạc dục văn.

được ở tâm ấy và sửa chửa tâm mình. Dạ phải, Đức Thế Tôn, con vui mừng xin muốn nghe.

will be able to abide there and subdue their mind. Yes, World-honored One, I would joyfully want to listen.

佛告須菩提。

Phật cáo Tu Bồ Đề

Đức Phật bảo Ngài Tu Bồ Đề :

Buddha said to Subhūti:

諸菩薩摩訶薩應如是降伏其心。所有一切眾生之類。

Chư Bồ Tát Ma-Ha tát ưng như thị hàng phục kỳ tâm. sở hữu nhất thiết chúng sanh chi loại.

Các Đại Bồ Tát phải sửa chữa tâm mình như thế này. Tất cả chúng sinh.

All Great Bodhisattvas should subdue their mind as follows .All sentient beings

若卵生若胎生若濕生若化生。若有色若無色。若有想若無想。

nhược noãn sanh nhược thai sanh nhược thấp sanh nhược hóa sanh. nhược hữu sắc nhược vô sắc. nhược hữu tưởng nhược vô tưởng.

dù sanh từ trứng, từ bào thai, từ ẩm thấp , từ biến hóa , có hình sắc hay không hình sắc, có tư tưởng hay không tư tưởng,

whether egg-born, womb-born, wetness- born, or born of transformation; whether with form or no form; whether with thought or no thought.

若非有想非無想。

nhược phi hữu tưởng phi vô tưởng

hoặc chẳng có tư tưởng chẳng không có tư tưởng ,

or whether neither with thought nor without thought,

我皆令入無餘涅槃而滅度之。如是滅度無量無數無邊眾生。

ngã giai lệnh nhập Vô-Dư Niết-Bàn nhi diệt độ chi. như thị diệt độ vô lượng vô số vô biên chúng sanh

Ta đều khiến tất cả được nhập Niết-Bàn hoàn toàn mà được diệt độ. Dù diệt độ vô lượng vô số vô biên chúng sinh,

I will lead all to enter the No-Residual (complete) Nirvana to be liberated. Though I have liberated an infinite, countless and boundless number of sentient beings,

實無眾生得滅度者。何以故。須菩提。

thật vô chúng sanh đắc diệt độ giả. hà dĩ cố. Tu Bồ Đề.

mà thật ra không có chúng sinh nào được diệt độ cả . Vì sao? Này Tu Bồ Đề!

in reality not one sentient is getting liberated. Why? Subhūti !

若菩薩有我相人相眾生相壽者相。即非菩薩。

nhược Bồ Tát hữu ngã tướng nhân tướng chúng sanh tướng thọ giả tướng. tức phi Bồ Tát.

Nếu Bồ Tát nào vẫn còn có tướng ngã, nhân, chúng sinh, thọ giả, thì chẳng phải là Bồ Tát.

If a Bodhisattva still has the images of a self, the images of a person, the images of sentient beings or the images of a life span, then he is not a Bodhisattva.

Appendix B

Sample Translation from Google Translate (Jan. 25, 2019)