MASTER IN COGNITIVE SCIENCE AND LANGUAGE

MASTER THESIS

07 2024

Applying semantic prosody for machine translation improvement on English-Chinese passive sentences

by *Xinyue Ma*

Under the supervision of:

Mireia Farrús Cabeceran

Universitat de Barcelona

Abstract: Passive sentences are widely used in English and have a neutral semantic prosody, while in Chinese they are much less frequent and with negative semantic prosody. Thus, many times English passives should be translated into active Chinese sentences and passive voice should be mainly reserved for unfavorable content. This work uses focused datasets to fine-tune a neural network machine translation model and improves its performance on translating English *BE* passives to Chinese. Evaluation after fine-tuning suggests training on English-Chinese sentence pairs that contain *BE* passives that are translated into actives by human translator, but to *BEI* passives by the model leads to considerable improvement of model performance. Also, data shows that the semantic orientation score of *BE* passives translated into Chinese passives by human translators are significantly different from those translated into active sentences, providing quantifiable way to address this issue.

Keywords: semantic prosody, contrastive linguistics, machine translation, passive sentence, semantic orientation

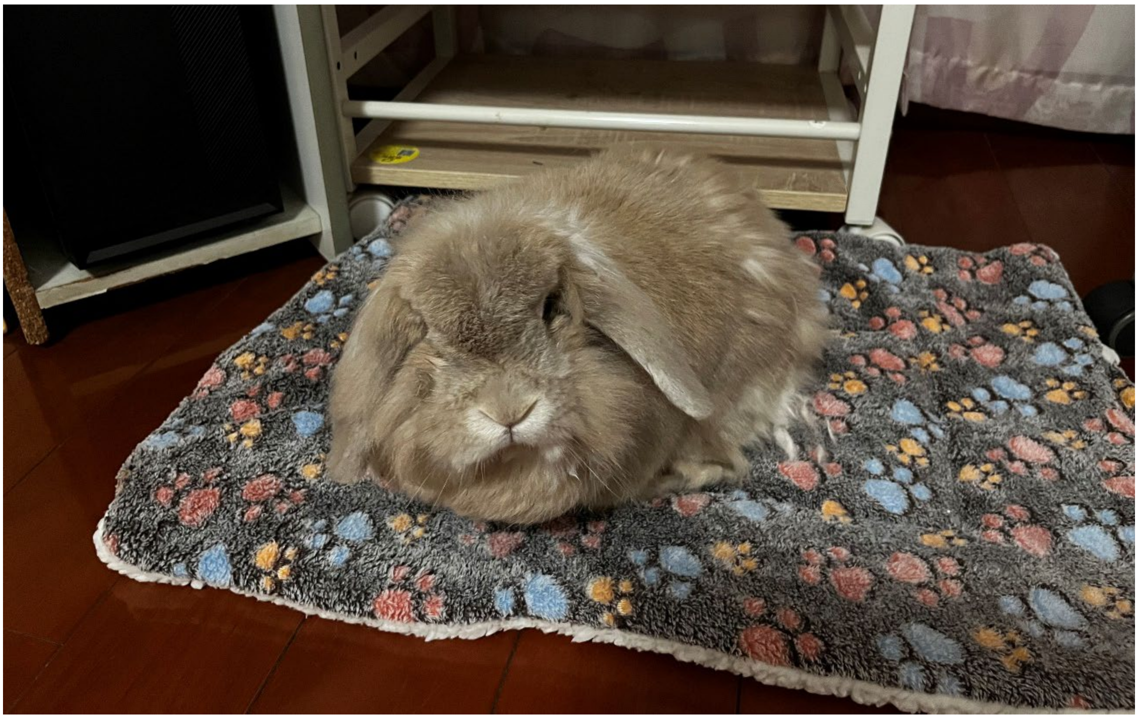
Acknowledgements

I would like to thank my advisors Prof. Mireia Farrús Cabeceran for her guidance and advice for conducting this work. Every discussion with her was inspiring. She pointed me in the right direction when I was hesitant about how to move forward with my thesis, and offered me a lot of help with thesis writing.

I would like to thank Pol Pastells Vilà, Research Support Technician of El Servei de Tecnologia Lingüística (STeL) for his assistance with the fine-tuning code and process of machine translation model. Without his knowledge of programming this work can never be realized.

I would like to thank my parents for supporting me both mentally and financially to study abroad and see more aspects of this world for myself.

Last but not least, many thanks to my rabbit 三七 for keeping me sane this year! The rest of this page belongs to him!



List of abbreviations

BNC: British National Corpus

CCL: Centre for Chinese Linguistics

CECPC-Core: China English-Chinese Parallel Corpus-Core

COCA: Corpus of Contemporary English in the United States

FLOB: Freiburg-LOB corpus

LCMC: The Lancaster Corpus of Mandarin Chinese

NEG: negation

OPUS: The Open Parallel Corpus

SO-CAL: semantic orientation calculator

List of tables

Index

1 Introduction

The concept “semantic prosody” arose in corpus linguistics in 1990s. Semantic prosody is the semantic coloring of a word (the node) that is contingent upon its semantic preferences, which involves semantic transfer. It may reveal speaker’s attitude under certain circumstances (Sinclair, 1996), but is inaccessible to a speaker’s conscious (McEnery, Xiao & Tono, 2006; Stewart, 2010). It is the result of the analyst’s judgement and interpretation of the corpus data and is covert in nature. Currently, one way to categorize it is to divide it into three categories: positive, neutral and negative semantic prosody. As the name suggests, words whose collocations are mostly positive gain a positive semantic prosody; the same is true for negative semantic prosody and neutral prosody respectively. Cooccurrence of a word with a predominantly positive or negative context may foster subtle associative meaning.

Due to language variation, a word in English may have a counterpart in Chinese that have the same semantic meaning, but the two words doesn’t necessarily have the same semantic prosody, which should be taken into account in order to reach semantic/pragmatic equivalence in translation. For example, the semantic meaning of English word *BECAUSE* and Chinese word “由于 *YOUYU*” are equivalent, yet their semantic prosodies are not. *BECAUSE* has a neutral semantic prosody while *YOUYU* has a negative one (Wu & Lan, 2020). Another example would be *INSIST ON,* which is often used to describe annoying stubbornness, its literal translation in Chinese “坚持 *JIANCHI*” has positive semantic prosody (Dong, 2020). This divergence in semantic prosody even caused English learners with Mandarin as L1 to use *INSIST ON* to give encouragement, while this usage never appears in Corpus of Contemporary English in the United States (COCA).

Currently in machine translation, this problem hasn’t been taken under consideration. For example, if we ask a machine to translate “I was praised by my teacher” into Chinese, Google Translation would give “我被老师表扬了”; DeepL and ChatGPT-4o would give “我受到了老师的表扬”. Both translations are using passive voice, which, despite the fact that in English it mainly occurs in neutral contexts (80% of the time for *be passive*) (Xiao & McEnery, 2005), has a negative semantic prosody in Chinese (Wu, 2022; Dong *et.al.*, 2023). As the standard and most common passive structure in Chinese, “被 *BEI* +

verb” structure has obvious negative prosody, which is not the best option to translate a sentence talking about “being praised” or any other favorable situation.

This study includes an attempt to improve the performance of a neural network machine translation model on passive sentences through proving with a focused training dataset. I also try to apply a semantic orientation calculator (later referred to as SO-CAL) (Taboada *et al.*, 2011) in order to see whether the semantic orientation of English passives translated into Chinese passives by human translators are significantly different from those translated into active sentences. If so, semantic orientation calculation score of the source text may be used as an indicator for choosing active or passive voice in a task that translates English passives to Chinese.

In this article, starting in Section 2, I give a literature review on previous studies about semantic prosody and its application in translation. In Section 3, I first explain the characteristics of passive sentences in Chinese. And then I focus on the most canonical passive structure *BEI* + verb, providing information about the semantic prosody of itself and its negation, the diachronic change of its meaning and usage, and a comparison to its English counterpart *BE* passives. Section 4 provides methodology. I use three different focused datasets to fine-tune “opus-mt-en-zh”, a neural network machine translation model (Tiedemann & Thottingal, 2020) and try to make it learn the negative semantic prosody of *BEI* passives. Three datasets include: 1) *BE* passives that are translated into actives by human translator, but to *BEI* passives by the model (negative evidence); 2) *BE* passives that are translated into *BEI* passives by human translator (positive evidence); 3) a combination of the previous two datasets. Then I evaluate the performance of these three models on *BE* passives translation task. Apart from the models, SO-CAL is also used to see if there is a significant difference between the semantic orientation of the first and second training dataset (just for the source text in English). Section 5 provides the results of the evaluation and discussion, and Section 6 conclusions.

2 Literature review

2.1 Semantic prosody

Over the last thirty years, semantic prosody has aroused considerable attention within corpus linguistics, yet its definition is not undisputed. Interest in the subject was initially kindled by Sinclair's observations about the lexico-grammatical environment of the phrasal verb *SET IN* (1991). Sinclair first noticed that the phrasal verb *SET IN* tends to have unpleasant states of affairs as its subject, such as *ROT*, *DECAY* and *DESPAIR*. Sinclair has begun to refer to this phenomenon as "semantic prosody" (in personal communication with Louw in 1988), applying the term "prosody" in an analogical way, which is in the same sense that Firth (in Palmer 1968:40) used the word to refer to phonological coloring which was capable of transcending segmental boundaries. An example would be the "nasal prosody" of vowels in word "Amen": the vowels are imbued with a nasal quality because of their proximity to "m" and "n", which are nasal consonants. (Louw, 1993) The term 'semantic prosody' itself first gained currency in Louw (1993). According to Louw, "the habitual collocates of the form *SET IN* are capable of coloring it, so it can no longer be seen in isolation from its semantic prosody, which is established through the semantic consistency of its subjects". It should be noted that this process of semantic coloring, as described by Louw, is a gradual one which over time would alter the meaning of *SET IN*. Louw defined semantic prosody as a "consistent aura of meaning with which a form is imbued by its collocates". His examples of lexical items with prosodies include *UTTERLY*, *BENT ON* and *SYMPTOMATIC OF*, for all of which he claimed negative prosodies.

In 2005, Whitsitt pointed out that there are three different definitions for semantic prosody and he laid criticism on the first definition offed by Louw (1993), saying that observations made from a synchronically designed corpus cannot account for the diachronic change of a word in meaning. The second definition, Sinclair's definition (1996) of semantic prosody, puts it on the pragmatic side of the semantic/pragmatic interface, and claims that it is attitudinal. This pragmatic feature of being evaluative and expressing speaker attitude caused Stubbs (2001) to prefer the term "discourse prosody". The third definition seems to confuse semantic prosody with connotation. Partington wrote that semantic prosody is "an aspect of expressive connotation" (1998). Although

Louw (2000) later intended to revise the definition of semantic prosody and distinguished it from connotation, which shows the community attitude towards a word/expression and is involved with notions of appropriateness in language use (Allan, 2007), Whitsitt (2005) states that his attempt was not successful, since he paid little attention to semantic transfer, which, according to Whitsitt, should best define what semantic prosody refers to and best distinguish it from a concept like connotation.

Here I summarize a single definition that I intend to apply for this study: a semantic transfer from habitual collocates to the node that is capable of coloring it gives rise to the semantic prosody of the node. Later in section 3.3 I will further explain the alignment between this definition and the study this paper aims to conduct.

2.2 Semantic prosody in translation

As for studies on semantic prosody in translation, Xiao and McEnery (2006) made a comparison of prosodies of near-synonyms across English and Chinese, and Berber Sardinha (2000) conducted an analysis of English and Portuguese. Both studies conclude that collocational behaviour and semantic prosodies of near-synonyms are unpredictable across the two language pairs, in some cases being quite similar and in others quite different. Partington (1998) claims that perfect equivalents across English and Italian are few and far between because even words and expressions that are 'look-alikes' (e.g., correct vs. the Italian corretto) may have very different lexical environments. Also, there are many case studies discussing the more appropriate translation of certain word or phrase, such as the study of Wang and Ge (2021), claiming that "事已至此 (the matter has come to this)" is a better translation of "It is what it is", compared to "情况就是这样 (this is the situation)".

Currently, applying semantic prosody to translation and translation pedagogy is under investigation, yet no study has tried to incorporate it to machine translation. Considering its growing importance in translation equivalence, it should be a feasible way to improve machine translation performance.

3 Passive voice in Chinese

The use of passive voice is one of the most obvious differences between Chinese and English. It differs in frequency, distribution in different genres and semantic prosody. Dong et al. (2023) showed through data of a self-built corpus of recent material (literature, news and papers published between January 1st and October 20th of 2021) that in English, passive voice is approximately 8 times more frequent than in Chinese, and is used to express neutral content, while in Chinese it is mainly used for negative content. This is also in line with its higher frequency in English news and papers than in English novels, since the former require more objectivity. In Chinese, the frequency of passive voice doesn't vary much according to genre.

In Chinese syntactic passive sentences there are multiple passive markers, such as: "被 *BEI*", "让 RANG" and "给 GEI", among which *BEI* is the only fully grammatical passive marker without semantical meaning, and also the most frequent one. RANG, GEI, "叫 JIAO" and some other verbs are partially grammatical markers with semantical meaning. All these markers have a negative semantic prosody and are mostly used under negative circumstances (but not necessarily). See the following examples:

(1) 我**被**逮捕了。

Wo ***BEI*** dai bu le

I was arrested.

(2) 我**让**人逮住了。

wo **rang** ren dai zhu le

I got caught.

(3) 我**给**人打了。

wo **gei** ren da le

I was beaten up.

Apart from syntactic passive sentences, there is another kind of passive sentences named notional passives, which contain verbs with passive meaning but are not categorized as passive markers, such as "挨 AI" (suffer), "受 SHOU" (undergo) and "遭 ZAO" (suffer). Since searching for passive sentences without markers and with partially grammatical passive markers is relatively difficult and may get a lot of noise in the result, this paper focus only on the use of passive sentences with marker *BEI* in translation.

3.1 Semantic prosody of *BEI* in Chinese fiction

The frequency of *BEI* passive sentence in fiction is higher than that in other genres, appearing 153 times per 100K words in literary texts, while only 94 times for news and even less times for scientific paper and Miscellaneous. Meanwhile, its semantic prosody is also the most negative in literary, with 66% of all cases being negative use, while the percentage for news is 51.5% (Xiao & McEnery, 2005). Of course, there are examples with positive, neutral and negative prosody. See the following examples from *To Live*, a Chinese novel written by Yu Hua:

(4) 那天新娘**被迎**进村里来时，穿着大红的棉袄，哧哧笑个不停。(positive)

The day the bride **was welcomed** into the village, she was wearing a quilted red jacket and couldn’t stop her nervous giggling.

(5) 田里的棉花已**被收起**. (neutral)

The cotton in the fields had already **been harvested**.

(6) 家珍**被拖出去时**，双手紧紧捂着凸起的肚子，那里面有我的儿子呵。(negative)

As Jiazhen **was carried out**, her hands firmly clasped her protruding belly, which held my son.

In *To Live*, most of the words that collocate with *BEI* are negative, among which we find “包围 *BAOWEI*” (surrounded by enemy), “打死 *DASI*” (beaten to death) and “俘虏 *FULU*” (captured) to be the most frequent ones. As for more general data, I collected all the sentences with the structure “*BEI* + verb” (distance between the maker and the verbs should be no more than three words) and found 3262 matches in fiction genre between 2000 and 2020 of CCL (Centre for Chinese Linguistics) Corpus. After analyzing with Wordless (Ye, 2024), data shows that among the 500 most frequent collocations (1392 sentences in total) of *BEI*, 58.84% are negative verbs, 34.05% are neutral and 7.11% are positive. This result is in line with the data of Hu & Zeng (2010), which showed that in all the passive sentences with *BEI* in LCMC (The Lancaster Corpus of Mandarin Chinese), 51.50% are negative, 37.80% are neutral and 10.70% are positive.

Collocation	Frequency
摧毁 (destroy)	40
打 (beat)	27
发现 (find out)	25
安置 (settle)	24
枪毙 (shoot dead)	24
当成 (regard as)	19
遗忘 (forget)	18
安排 (arrange)	18
镇压 (suppress)	17
切断 (cut off)	16

Table 1: Ten most frequent collocation of *BEI*

Using *BEI* passives for positive expressions often occurs in favor-accepting type of passive sentences, which is an isolated type compared with other types of passive sentences. There are three most representative collocations: "被授予 *BEI SHOUYU*" (be awarded), "被评为 *BEI PINGWEI*" (be recognized as) and "被列入 *BEI LIERU*" (be listed/included in). This type of passive sentences is mainly used for presenting achievements, awards, promotions, etc. See the following example:

(7) 2012 年诺贝尔经济学奖被授予美国学者阿尔文•罗思和劳埃德•沙普利。

The Nobel Prize for economics has been awarded to Alvin Roth and Lloyd Shapley.

Generally speaking, passive voice in Chinese is much less frequent and much more negative compared to that in English. As for Chinese translated fiction, it is noted that the frequency of passive voice is lower and the semantic prosody of passive marker *BEI* is more negative in comparison with Chinese original fiction, showing a tendency of conventionalization in translation (Hu & Zeng, 2010).

3.2 Semantic prosody of "NEG-*BEI*" in Chinese fiction

There are mainly two kinds of negation in *BEI* passives: *MEIBEI* (没被, haven't been) and *BUBEI* (不被, not be). In CCL Corpus, sentences contain these two structures in fiction genre between 2000 and 2020 are very limited. In all 26 sentences, 11 of them contain a negative verb (42.3%), 9 sentences contain a neutral verb (34.6%) and the other 6, a positive verb (23.1%). Apart from the positive/negative meaning of collocative verb, it should be noted that the overall meaning of the sentence with a positive verb may be unfavorable, for example: “重视 *ZHONGSHI*” (attach importance to something) is a positive verb, but when it appears with “NEG-*BEI*”, the meaning of the phrase becomes “unappreciated”, which is rather negative. In the 26 sentences mentioned above, 14 of them are sentences giving an unfavorable narration, occupying 53.8% of the total.

(8) 我感觉我们的战士是太伟大了，太可爱了，我不能**不被**他们**感动**得掉下泪来。(positive)

I feel that our soldiers are so great and lovely that I can't **not be moved** to tears by them.

(9) 天津各机关应在思想上、组织上和医疗上均作有效的准备，以保卫首都**不被**鼠疫**侵入**。(neutral)

All departments of Tientsin should make effective ideological, organizational and medical preparations to secure that the capital **is not invaded** by the plague.

(10) 错误思想在开始时总是微小的，但当它**不被**及时**制止和消灭**而得到发展时，则会陷于不可收拾的地步。(negative)

Misthought is always small in the beginning, but when it **is not stopped and eliminated** in time and develops, it slips into an unmanageable situation.

In order to judge the semantic prosody of *NEG-BEI* by a larger amount of data, I collected sentences of all genres in CCL Corpus containing the structure “*BUBEI* + verb” (taking long passive structure as an example, in “passive marker + agent + adverbial + adverbial marker DE + verb”, distance between the marker and the verbs is three words and is taken as the upper limit for data collection.) and found 4136 matches. In the 500 most frequent verbs that collocate with *BUBEI* (3039 sentences in total), 35.83% are negative, 28.76% are neutral and 35.41% are positive. This is very different from the distribution of collocations of *BEI*, with positive verbs and negative verbs appearing at a similar frequency.

But when I turn to the general meaning of the sentence, that is, whether the sentence narrates a favorable or unfavorable event, the data becomes very different from those mentioned above. It should be noted that the combination of *NEG-BEI* and a negative verb does not necessarily yield a positive semantic tendency, but a neutral one in most cases. For example, “保护财产不被夺走” (Protect property from being taken away) is just a neutral statement. Through manually checking the meaning of 2000 sentences containing *NEG-BEI*, only 13.95% are found to be positive. 44.75% of the sentences are neutral, and 41.30% of them are negative. The data show a great contrast in the number of sentences with positive and negative meaning, indicating that *NEG-BEI* has mainly neutral to negative prosody and is rarely used in positive context. Thus, sentences containing *NEG-BEI* are not considered special and will also appear in English-Chinese sentence pairs collected for fine-tuning neural network machine translation model (opus-mt-zh-en).

3.3 A diachronic view of *BEI* passives

In section 2 a few flaws in the definition of semantic prosody were mentioned. In this and next sections, I intend to discuss them one by one to clarify whether these uncertainties in definition and some other factors would affect this study.

For *BEI* passives, there are diachronic records to be consulted in order to deem whether there is any semantic change. Passives first appeared in Old Chinese during the Spring and Autumn period (approximately from 770 to 481 B.C.) and *BEI* passives germinated around the end of the Warring Period (approximately from 476 to 221 B.C.) (Wang, 2013). During the Han Dynasty, the use of *BEI* passives gradually became increasingly common.

Initially, *BEI* as a noun means "quilt, cover", and from this meaning, two different uses as a verb have derived: the first meaning is "(actively) cover, place on the body", and the second is "(passively) be imposed with, suffer from". The auxiliary *BEI* in passive structure comes from the second use. According to Wang (2013), examples dating back to Han Dynasty show that *BEI*, both as a verb (second meaning mentioned above) and as an auxiliary, has been mainly used in unfavorable contexts. Apart from *BEI*, other passive markers like "WEI" also have the same distribution. After The May Fourth Movement, *BEI* started to appear more in neutral and positive context due to the influence of Indo-European languages. But this change is only shown in written Chinese. Oral Chinese seems to be resistant to such influence.

The fact that *BEI* once was a negative verb doesn't necessarily explain why *BEI* in passives has a negative semantic prosody, since there are other auxiliaries with neutral prosody that also derive from a verb with unfavorable meaning. One example would be "哉 *ZAI*" in Classical Chinese, which initially meant "to traumatize; to initiate" as a verb but later turned into an auxiliary that is sometimes used as a spacer in a sentence, but more often used at the end of a sentence to express an exclamation or question (Zhou, 2018).

From a diachronic point of view, we do not see much change in the context of *BEI* passives, which is mainly negative. Although there is no evidence for the process of semantic transfer, I believe the negative collocates of *BEI* have played a significant role so that instead of becoming an auxiliary like *ZAI*, *BEI* remains in a mainly negative context.

3.4 Other variables that affect semantic prosody

Semantic prosody of a node can vary according to different basic meanings (Bublitz, 1996). Sinclair (1991) claimed that *HAPPEN* has an unfavorable semantic prosody, but Bublitz pointed out that this does not apply to "by-chance-meaning" of *HAPPEN* (e.g., 'I happen to know his work'). *BEI* passives should be free of this trouble, since as an auxiliary, *BEI* has no other use than passive voice marker. And as a noun, a suffix "子 *ZI*" that clearly denotes word class is frequently added in Standard Mandarin, so that "被子 *BEI ZI*" (quilt) can be easily distinguished from the auxiliary.

"Local prosodies" is another concept that should be taken into consideration. According to Tribble (2000), words in certain genres may establish local semantic prosodies that only occur in these genres, or analogues of these genres. For example, in recipes, *CHOPPED* mainly collocates with *FINELY, FRESH, PARSLEY, ONION, GARLIC, et. al.,* but in other text-types, it has a greater tendency to combine with *OFF*, *UP* and *DOWN* and involve violence to humans (Stubbs, 2001). As a matter of fact, genre has a significant influence on the semantic prosody of *BEI* passives. *BEI* passives in press reviews and adventure fiction have the most negative semantic prosody, with negative *BEI* passives up to 54% and 65% respectively, while the percentage drops to around 20% for reports, official documents and academic prose (Xiao & McEnery, 2005). Since the local prosody of fiction genre fits better the canonical use of *BEI* passives in Chinese, this is where this study will focus on.

3.5 A comparison between Chinese passives and English passives

The structure *BE* + past particle can be considered as the norm for English passives (Xiao & McEnery, 2005) and it is the structure this study will focus on. There are also other copular verbs that can replace BE in this structure, such as *GET*, *BECOME*, *FEEL*, *LOOK*, among others, but they are much less frequent compared to *BE* passives in corpus data. *BE* passives appeared 9908 times in FLOB (Freiburg-LOB corpus, an update of Lancaster-Oslo-Bergen corpus of British English that contain texts published in 1991-1992), while *GET* passives only appeared 59 times.

English *BE* passives and Chinese *BEI* passives show great divergence in semantic orientation. According to Xiao and McEnery (2005), unlike *BEI* passives, 80.3% of *BE* passives in FLOB and BNCdemo (a demographic sampled component of the British National Corpus, the World edition) express neutral content. It is worth noting that *GET* passives are typically used for events with a consequence that is negative or viewed as unfortunate by the speaker, which means they can be translated into *BEI* passives and maintain its negative semantic prosody, and will not be discussed further in this work.

4 Methodology

Chinese passives have a negative semantic prosody while English passives are mainly neutral. Thus, English passives with positive or neutral content should be translated into Chinese active sentences, and the passive voice should be kept mainly for unfavorable events. The primary research method for my study is to fine-tune an English-to-Chinese machine translation model with focused training data to make it consider semantic prosody when translating English passive sentences into Chinese. And a semantic orientation calculator is also used to quantify negativity of sentences.

4.1 Datasets for fine-tunning

In order to teach opus-mt-en-zh model about the negative semantic prosody of *BEI* passives in Chinese, I created a focused training dataset. Training data is collected from the fiction genre of The Babel English-Chinese Parallel Corpus (244,696 words in total) created by Richard Xiao, China English-Chinese Parallel Corpus-Core (CECPC-Core, 5,499,591 words in total) created by Kefei Wang, BFSU and Yiyan English-Chinese

Parallel Corpus (1,169,970 words in total) created by Xiuling Xu & Jiajin Xu (All accessable on CQPweb of Beijing Foreign Studies University at http://114.251.154.212/cqp/), since the negativity of Chinese passives is most obvious in fiction, compared to that in scientific papers and documents (Xiao & McEnery, 2005).

As mentioned before, there are three datasets used. In total, 900 English-Chinese sentences pairs are collected for training and this dataset is split into two subsets. One contains 424 English *BE* passives that are translated to active sentence by human translators, but are translated to *BEI* passives by the model (the corresponding Chinese translation, of course, is also included in the subset). This subset is intended to attenuate the degree of correspondence between the two passives and will be later referred to as "negative evidence". The other subset contains the remaining 476 pairs in which *BE* passives translated to *BEI* passives by human translator and express a negative content. They are selected to reinforce the relation between *BEI* passives and negativity and will be later referred to as "positive evidence".

4.2 Machine translation model

The model used is opus-mt-en-zh, a Transformer model of Language Technology Research Group at the University of Helsinki (https://huggingface.co/Helsinki-NLP/opus-mt-en-zh). This model is based on MarianMT model and was originally trained on data from The Open Parallel Corpus (OPUS, http://opus.nlpl.eu), with a BLEU score of 31.4 and chrF2 score of 26.8.

Currently, the model many times translates a positive/neutral passive English sentence into a passive Chinese sentence, without considering the negative semantic prosody may interfere with comprehension. Some examples of unproper use of passive voice in model translation are as follows:

(11) A chair was offered to him, and he **was invited** to the feast.

a. 立地便有一张椅子给他，**请**他就席。(human translation)

"(Someone) offered him a chair immediately, and **invited** him to the feast"

b. 他得到了一张椅子给他，他**被邀请**参加盛宴。(model translation)

"He got a chair to him and he **was invited** to the feast."

(12) Finally I won, and **was permitted** (by my stepfather) to go to the school in the day for a few months.

a. 末了，我竟胜利了，后父终于**许**我进日校读这么几个月。(human translation)

“At last, I won, and my stepfather finally **permitted** me to go to the school in the day for a few months.”

b. 最后我赢了, 并**被**我的继父**允许**每天上学几个月。(model translation)

“Finally I won, and **was permitted** by my stepfather to go to school every day for a few months.”

For the fine-tuning, the model is trained on all three datasets mentioned in the previous section for 3 epochs. 80 percent of the data is for training and the rest 20 percent for testing. After training on each dataset, the fine-tuned model is evaluated manually by checking its performance on testing dataset, and later the results of three fine-tuning processes are compared together.

4.3 Semantic orientation calculator (SO-CAL)

Apart from the machine translation model, a semantic orientation calculator (SO-CAL) (Taboada et al., 2011) that has a consistent performance across domains is employed to show the semantic orientation score of all English sentences in the dataset, in order to see if the score of those translated into actives is significantly different from those translated into passives.

To get a semantic orientation score, texts first go through Part-Of Speech tagging using Stanford CoreNLP (Manning et al., 2014). Then SO-CAL can collect sentiment-bearing words (including adjectives, verbs, nouns, and adverbs), and use them to calculate semantic orientation, with special attention to valence shifters (intensifiers, downtoners, negation, and irrealis markers) (Taboada et al., 2011).

5 Results and Discussion

In the following sections I will evaluate the fine-tuned models both by metrics and manually (checking their performance on translating *BE* passives). The results of SO-CAL and analysis will also be provided.

5.1 Results and evaluation for fine-tuned model

After fine-tuning with each dataset, BLEU and chrF2 metrics are used to evaluate the general model performance, which actually has worsened a little bit after fine-tuning. Scores calculated using sacreBLEU (Post, 2018) on a testset that contain 10 000 sentence pairs offered by opus-mt-en-zh repository (https://object.pouta.csc.fi/Tatoeba-MT-models/eng-zho/opus-2020-07-17.test.txt) are shown in table 2:

Model	BLEU	chrF2
opus-mt-en-zh	31.4	26.8
fine-tuned with both evidence	24.7	21.6
fine-tuned with positive evidence	25.3	22.3
fine-tuned with negative evidence	23.8	20.9

Table 2: Evaluation of fine-tuned models

A possible reason for getting lower BLEU and chrF2 score after fine-tuning is that the model is overfitting to the focused training dataset, which only contain *BE* passives in source language, causing a lower score when it is evaluated with general texts.

When it comes to translating *BE* passives into Chinese, all three training datasets have improved model performance on the usage of *BEI* passives. The model fine-tuned with the subset that contains negative evidence only shows the greatest improvement and yields the best result, correctly translating 74% of *BE* passives in negative evidence testset into Chinese active sentences, while the original model only got 4% correct.

Model performance on 180 test sentence pairs is shown in table 3. The translation being a *BEI* passive or not is shown through +/- and follow the order of "human translation/original model translation/fine-tuned model translation". For example, "+++" means that for these *BE* passives, human translator, original model and fine-tuneded model all translate it into *BEI* passives. Accuracy shows in how many cases model translation is in line with human translation in voice, that is, using BEI passive or not.

Case Training data		+++	++-	+-+	+--	+-- *	---	--+ #	-+-	-++	-++ #	Total	Original accuracy	Accuracy after training
Both evidence		77	/	3	14	1	3	2	17	8	55	180	45.55%	55.55%
Positive evidence	Positive testset	79	/	9	7	/	/	/	/	/	/	95	46.11%	50.55%
	Negative testset	/	/	/	/	/	1	3	2	4	75	85		
Negative evidence	Positive testset	55	24	/	13	3	/	/	/	/	/	95	46.11%	65.56%
	Negative testset	/	/	/	/	/	4	/	59	7	15	85		

Table 3: Performance of fine-tuned models on test datasets

Cases marked with "*" are bad translation caused by using active voice in translation without reverse the subject and object, causing the agent to change; those marked with "#" are considered not good because of unproper use of BEI passive, yet still are correct in meaning.

To further explain some special cases, "++-" and "+--" are not marked and considered acceptable, because the overall percentage of passives in Chinese is low and active sentence can be used to describe all kinds of events, regardless of its negativity. As for cases in "-++", the model translation is considered acceptable because the sentences describe unfavorable events.

To clarify the criteria for judging a translation to be "acceptable", in this work, active translation is always accepted unless the model fails to reverse the subject and object, causing a change in meaning. If human translator decided not to use passives, passive translation is only accepted when the content is negative. The criteria are set to better show if the model have learned that *BEI* passives have negative semantic prosody.

After training on both positive and negative evidence for the usage of *BEI* passives, among those cases that human and model translation diverge, the fine-tuned model got 20 cases right (20.41%). These 20 cases are mainly correct translation into active sentence after training.

The performances after training on subsets of only positive or negative evidence are very different. The model trained on positive evidence only performed very well on test dataset of the same kind, yielding all correct or acceptable translation. But on the test dataset of negative evidence, it showed no improvement at all. Examples of improvement on correctly using *BEI* passives in translation are as follow:

(13) Given that the passing countryside **is enveloped** in humidity, darkness and low-hanging cloud, there is only one thing for it and that is to start exploring the train.

a. 列车经过的乡村**被**湿热、黑暗和低垂的乌云所**笼罩** (be enveloped)，我们能做的只是开始熟悉一下自己所乘坐的车厢。(human translation)

b. 路过的乡村**充满**潮湿、黑暗和低悬浮云云，只有一件事，那就是开始探索火车。(original model translation)

"The passing countryside **is full of** damp, darkness and low-hanging clouds. There's only one thing to do and that is to start exploring the train."

c. 鉴于过往的乡村**被**潮湿、黑暗和低悬浮云所**包围**，只有一件事是开始探索火车。(fine-tuneded model translation)

"Given the passing countryside **is surrounded by** damp, darkness and low-hanging clouds, there is only one thing to do and that is to start exploring the train"

Example (13) describes a relatively unfavorable event, namely having nothing to do but to explore the train due to bad weather. Yet still, the translation of the original model is acceptable.

(14) Of course, George had a reason for wanting the bank to **be robbed**.

a.当然，乔治想要这家银行**被抢**是有个理由的。(human translation)

b. 乔治 当然 有 理由 想 **抢** 银行。(original model translation)

George surely has reason want rob bank

"George surely has reason to want to **rob** a bank."

c. 乔治 当然 有 理由 要 银行 **被 抢**。(fine-tuneded model translation)

George surely has reason want bank *BEI* rob

"George certainly has reason to want the bank to **be robbed**."

For example (14), negativity is not the main reason why passive translation is the better one. The original model translates it into an active sentence, failing to show that the subject of "rob" is not mentioned and is not George. Here using active voice would require adding a subject such as "someone" before the verb.

The model trained on negative evidence gives the best performance among all three. The model learned to correctly translate *BE* passives to active sentences in 59 cases out of 85. As for test dataset of positive evidence, it largely maintained the passive translation. In 24 cases it turned from a correct passive translation to an active but still acceptable one.

Combining the performance on both types of test datasets, the third model performed the best and most balanced. Examples of translation maintained correct or improved are shown as follow:

(15) Yet our revolutionary comrades, all warriors against Japan, **have been killed**.

革命的同志，抗日的战士，却**被杀死**了。(human translation)

然而，我们的革命同志们，所有反对日本的战士，都**被杀害**了。(original model translation)

然而，我们的革命同志们，所有反对日本的勇士们，都**被杀害**了。(fine-tuneded model translation)

For example (15), both original and fine-tuned model give good translation. "Comrades and worriers have been killed" is an obviously unfavorable event, which makes passive translation an applaudable choice.

(16) you see, I am travelling on foot -- on this occasion. My trunk is **being sent** after me.

a. 你看我这回是走路来的，我的箱子跟着就**寄来**。(human translation)

"You see, I'm coming on foot this time, and my suitcase **comes in by mail** after me."

b. 你看，我这次是徒步旅行，我的后备箱正**被派来**追我。(original model translation)

"You see, I'm travelling on foot this time, and my boot is **being sent** to chase me"

c. 你看，我这次是徒步旅行，我的行李箱是随我**送来**的。(fine-tuneded model translation)

"You see, I'm travelling on foot this time. My suitcase **comes** with me."

In example (16), "My trunk is being sent after me" is a neutral statement, which is usually in active voice. Also, the translation of the original model is not correct. "Trunk" here means suitcase, not the boot of a car, and it cannot chase (追 *ZHUI*) anyone.

(17) A chair was offered to him, and he **was invited** to the feast.

a. 立地便有一张椅子给他，**请**他就席。(human translation)

"(Someone) offered him a chair immediately, and **invited** him to the feast."

b. 他得到了一张椅子给他，他**被邀请**参加盛宴。(original model translation)

"He got a chair to him and he **was invited** to the feast."

c. 有人向他提供了一把椅子，**请**他来参加盛宴。(fine-tuneded model translation)

“Someone offered him a chair and **invited** him to the feast.”

Example (17) describes a favorable event, namely “a man is treated with respect and invited to feast”, and thus passive sentence is not the best option for translation. Also, in (15b), the cooccurrence of “get” (“得到”) and “to him” (“给他”) gives rise to an error.

5.2 SO-CAL score of datasets

Semantic orientation calculator is used to give a score for all 900 English sentences in the training dataset. Sentences score over 0 are marked “positive”, “negative” if below 0 and “neutral” if exactly 0. The absolute value of the score reflects the degree of positivity or negativity of the sentence. Since I intentionally only collected examples that have negative content for positive evidence subset in order to maximize training efficiency, apart from the training data, another unbiased dataset that collects exhaustively 213 examples from the fiction genre of Yiyan English-Chinese Parallel Corpus was created and also undergoes the calculation. It is obvious that the distribution in two positive evidence subsets and negative evidence subset are significantly different (Mann-Whitney U test was used because the data does not follow normal distribution, $p < 0.005$. Test realized through R (R Core Team, 2023)). The results are shown in table 4.

	Negative	**Positive**	**Neutral**	**Total**	**Average score**
Positive evidence	290	83	103	476	-1.09
Yiyan positive evidence	100	52	61	213	-0.73
Negative evidence	153	145	126	424	-0.21

Table 4: Semantic orientation of different datasets

In Yiyan dataset, negative examples occupy 46.95% of all examples, while the other two kinds only occupy around one-quarter. This is in line with fact that most *BEI* passives have unfavorable content. As for negative evidence dataset, three kinds of examples are of similar amount. Also, positive evidence datasets have lower average score compared to negative evidence dataset. In positive evidence subset, the average score of 290 sentences marked negative is -2.20. This may be used as a minimal standard for allowing *BEI* passives in translation in order to make the translation acceptable for most cases.

6 Conclusion

In translation, a word or structure may have a semantically equivalent counterpart in another language, but their semantic prosody may differ according to their respective collocates. Since *BEI* passives in Chinese have a negative semantic prosody and are less frequently used, it is not always the appropriate translation of *BE* passives in English. This study shows that for MarianMT model, which is a neural network machine translation model (NMT model) trained on OPUS data, using sentence pairs containing *BE* passives that it fails to correctly translate into active Chinese sentence as training data can significantly improve its performance in translating such sentences, lower the frequency of using *BEI* passives for translating *BE* passive and reserve *BEI* passives mainly for negative events.

Meanwhile, SO-CAL scores of two positive evidence datasets and negative evidence dataset show significantly different distribution. It can reliably show the semantic orientation of the whole sentence and thus be used to set a threshold value for considering BEI passives applaudable in translation.

This work shows that using focused training data is a feasible way to help NMT model consider semantic prosody when translating *BE* passives. However, the number of “false friends” that have similar or same semantic meaning but different semantic prosody is considerable across all languages. Although fine-tuning with focused training data is effective for improving the usage of *BEI* passives, it is quite time-consuming to collect data, and it may not be possible for other units such as phrases and sayings of low frequency in corpus.

In order to reach the equivalence of semantic prosody between source language and target language, calculating semantic orientation score for both source and target language and try to maintain the score inside a 5-word range on both sides of the node or for the whole sentence during translation might be a more practical and efficient way. However, it should also be kept in mind that although quantificational methods can promote efficiency, they are not as accurate as human native speaker’s judgement and many times cannot pay sufficient attention to context, background information and inference.

Future work may try to find methods for teaching models about the semantic prosody of low-frequency words, idioms and sayings, and create focused datasets that can help a model learn the semantic prosody of multiple units at the same time, which may better avoid overfitting to a specific kind of data. As for promoting efficiency, attention may be paid to creating multilingual dictionaries for "false friends" of inequivalent semantic prosody, offering to machines translation models a quick access and combining statistical method with neural network to achieve better performance.

7 References

Allan, K. (2007). The pragmatics of connotation. Journal of pragmatics, 39(6), 1047-1057.

Dong. D. (2020). A Comparative Research on the INSIST Semantic Prosodic Phrases and Pragmatic Attributes Based on Corpus. Innovation and Practice of Teaching Methods, 3(11), 68-72.

Dong, P., Jiang, C., & Xu, P. (2023). A corpus-based comparative study of English-Chinese passive voice. Overseas English, 16, 53-55+59.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. Text and technology: In honour of John Sinclair, 157, 176.

Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffer & H. Saunston (Eds.), Words in Context. Discourse Analysis Monograph 18 [CD Rom]. Birmingham: University of Birmingham.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

McEnery, T., & Xiao, R. (2005). Passive constructions in English and Chinese: A corpus-based contrastive study. In Corpus Linguistics 2005.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

Palmer, F. R., & Firth, J. R. (1968). Selected Papers of J. R. Firth, 1952-59. Bloomington: Indiana University Press

Partington, A.S. (1998). Patterns and Meanings: Using corpora for English language research and teaching.

Post, M. (2018, October). A Call for Clarity in Reporting BLEU Scores. Proceedings of the Third Conference on Machine Translation: Research Papers, 186–191.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sardinha, T. B. (2000). Semantic prosodies in English and Portuguese: A contrastive study. Cuadernos de filologia Inglesa, 9(1).

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, J. (1996). The Search for Units of Meaning. Textus, IX, 75–106.

Stewart, D. (2010). Semantic prosody: A critical evaluation. Routledge.

Stubbs, M. (2001). Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT). Lisbon, Portugal.

Tribble, C. (2000). 'Genres, keywords, teaching: towards a pedagogic account of the language of project proposals', in L. Burnard and A. McEnery (eds), Rethinking Language Pedagogy from a Corpus Perspective. Papers from the third international conference on Teaching and Language Corpora. Frankfurt: Peter Lang, pp.75–90.

Wu, Z. (2022, July). A quantitative study on the stylistic differences of “Bei” passives in contemporary Chinese. In Computational Social Science: Proceedings of the 2nd International Conference on New Computational Social Science (ICNCSS 2021), October 15-17, 2021, Suzhou, Jiangsu, China (p. 50). Taylor & Francis.

Xiao, R., & McEnery, A. (2006). Near synonymy, collocation and semantic prosody: a cross-linguistic perspective. Applied Linguistics, 27(1), 103-129.

Hu, X., & Zeng, J. (2010). The Frequency, Structure and Semantic Prosody of "Bei" Passives in Chinese Translated Fiction. Journal of Foreign Languages (03),73-79.

Wang, L (2013). History of the Chinese Language. Beijing: Zhonghua Book Company

Wang, T., & Ge, S. (2021). Corpus-based semantic prosody study of English-Chinese translation: taking trump’s popular saying “it is what it is” as an example. In Learning Technologies and Systems: 19th International Conference on Web-Based Learning, ICWL 2020, and 5th International Symposium on Emerging Technologies for Education, SETE 2020, Ningbo, China, October 22–24, 2020, Proceedings 5 (pp. 420-429). Springer International Publishing.

Wu, Z., & Lan, X. J. (2020). The semantic prosody of “Youyu”: evidence from corpora. In Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20 (pp. 654-660). Springer International Publishing.

Ye, L. (2024). Wordless (Version 3.5.0) [Computer software]. Github. https://github.com/BLKSerene/Wordless

Zhou, S (2018). History of word-classes in Chinese. Beijing: China Renmin University Press Ltd