Building an English-Chinese Parallel Corpus Annotated with Sub-sentential Translation Techniques

Yuming Zhai[1], Lufei Liu[2], Xinyi Zhong[3], Gabriel Illouz[1], Anne Vilnat[1]
[1] Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France
[2]Université Paris Diderot, 75013, Paris, France
[3]Université Sorbonne Nouvelle - Paris 3, 75005, Paris, France
yuming.zhai@limsi.fr, lufei.liu@etu.univ-paris-diderot.fr, xinyi.zhong@sorbonne-nouvelle.fr
gabriel.illouz@u-psud.fr, anne.vilnat@limsi.fr

Abstract

Human translators often resort to different non-literal translation techniques besides the literal translation, such as idiom equivalence, generalization, particularization, semantic modulation, etc., especially when the source and target languages have different and distant origins. Translation techniques constitute an important subject in translation studies, which help researchers to understand and analyse translated texts. However, they receive less attention in developing Natural Language Processing (NLP) applications. To fill this gap, one of our long term objectives is to have a better semantic control of extracting paraphrases from bilingual parallel corpora. Based on this goal, we suggest this hypothesis: it is possible to automatically recognize different sub-sentential translation techniques. For this original task, since there is no dedicated data set for English-Chinese, we manually annotated a parallel corpus of eleven genres. Fifty sentence pairs for each genre have been annotated in order to consolidate our annotation guidelines. Based on this data set, we conducted an experiment to classify between literal and non-literal translations. The preliminary results confirm our hypothesis. The corpus and code are available. We hope that this annotated corpus will be useful for linguistic contrastive studies and for fine-grained evaluation of NLP tasks, such as automatic word alignment and machine translation.

Keywords: corpus annotation, translation technique, automatic classification

1. Introduction

Translation theorists and linguists have conducted studies on translation techniques for a few decades (Vinay and Darbelnet, 1958; Newmark, 1981; Larson, 1984; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002; Fadaee, 2011; Đorđević, 2017). Translation techniques refer to the specific steps for the sake of accomplishing an acceptable and appropriate translation, which can be divided coarsely into literal translation and non-literal translation at sub-sentential level.

Consider two human non-literal translation examples in table 1: the translation of the first sentence conveys the meaning in a more direct way to help readers' understanding; the translation of the second sentence divides one sentence into two clauses to paraphrase the expression « *unfold out of* », thus the translation is more natural and compact.

EN: *Don't make me* ***go through all of this and not make it***. ZH: 别让我的辛苦白费了。 ("Don't **let my hard work be wasted**.")
EN: *In the east the dawn* ***was unfolding out of the darkness***. ZH: 东方晨曦初现，黑暗渐去。 ("In the east the dawn **was beginning to appear, and the darkness was fading**.")

Table 1: English-Chinese non-literal translations

Non-literal translations between different languages can cause difficulties for automatic word alignment (Dorr et al., 2002; Deng and Xue, 2017), or cause meaning changes in certain cases. However, non-literal translation techniques receive less attention in developing NLP applications. Take the task of paraphrase extraction from bilingual parallel corpora as an example. The assumption is that two monolingual segments are potential paraphrases if they share common translations in another language, and the extraction relies on Machine Translation (MT) related techniques (Bannard and Callison-Burch, 2005; Mallinson et al., 2017). Currently, the largest paraphrase resource, PPDB (ParaPhrase DataBase), has been built based on this method (Ganitkevitch et al., 2013). Nonetheless, Pavlick et al. (2015) revealed that there exist relations other than strict equivalence in PPDB (*i.e. Entailment (in two directions), Exclusion, Other related and Independent*)[1]. Non-literal pivot translations inside the parallel corpora could break the strict equivalence between the candidate paraphrases extracted, whereas they have not received enough attention during this corpus exploration.

In this working context, one of our long term objectives is to have a better semantic control of extracting sub-sentential paraphrases from bilingual parallel corpora. Based on this goal, our hypothesis is that it is possible to automatically recognize different sub-sentential translation techniques (e.g. literal versus non-literal). For this original task, since there is no dedicated data set for English-Chinese, we manually annotated a parallel corpus with translation techniques. To reflect the diversity of textual styles, we constructed a corpus of eleven genres based on existing work. Fifty sentence pairs for each genre have been annotated in order to consolidate our annotation guidelines. Based on this data set, we conducted an experiment to classify between literal and non-literal translations. The preliminary results confirm our hypothesis that we can automatically recognize sub-sentential translation techniques.

[1] Exclusion: X is the contrary of Y; X is mutually exclusive with Y. Other related: X is related in some other way to Y. (*e.g. country / patriotic*). Independent: X is not related to Y.

2. Related Work

The first annotation guidelines for manually annotating parallel corpora were established for the project Blinker (Melamed, 1998a; Melamed, 1998b), in order to annotate translational equivalence in English-French Bible verses. More recently, Monti et al. (2015) annotated multiword expressions in an English-Italian parallel corpus of TED Talks[2]. Annotators also indicated whether the generated machine translation is correct, and supplied a correct translation if needed. Ahrens et al. (2018) built an online large database containing English and Chinese political speeches. This corpus is particularly useful for researchers focusing on political speeches and conceptual metaphor analyses.

Concerning non-literal translation techniques, several works have proposed different typologies to categorize them (Vinay and Darbelnet, 1958; Newmark, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002). Our corpus annotation is based on these translation theories. Deng and Xue (2017) built a hierarchically aligned parallel corpora and semi-automatically detected Chinese-English translation divergences, which are caused by non-literal translations and cross-lingual differences. Chen et al. (2018) used attention mechanism scores in an innovative way to detect free translation in English-Chinese parallel corpora. Xu and Yvon (2016) proposed new methodologies for collecting human judgements on bilingual alignment links, which were used to annotate four new data sets. Their observation confirms that a finer categorization than *Sure* and *Possible* word alignment is useful. In our work, we conduct word and segment level alignment, and specify the fine-grained translation technique.

Ahrenberg (2017) compared machine and human translations of an English article translated into Swedish, by using MT metrics and translation techniques. The author pointed out that automatically classifying translation techniques should be a topic for future research. Recently, we have worked on automatically classifying translation techniques for the language pair English-French (Zhai et al., 2019). This present work extends these studies by working on a more distant language pair: English-Chinese.

3. Corpus Presentation

We extend our previous work which focused on annotating an English-French parallel corpus of TED Talks with translation techniques (Zhai et al., 2018). English and French languages are very similar in vocabulary and grammar, while the English-Chinese pair shares far fewer cultural and linguistic similarities. A corpus of eleven genres is constructed based on existing work: *art, literature, law, material for education, microblog, news, official document, spoken, subtitles, science and scientific article*.[3] For our first study of this language pair, we didn't limit ourselves to only one corpus genre, even though the corpus of different genres don't have the same quality. Below we present the origin of each corpus. The translation direction is from English to Chinese, except for the genre of *scientific article*.

UM-corpus (Tian et al., 2014): this corpus has been constructed by the University of Macau, for training machine translation systems. The corpus released contains 2.2M parallel sentences, and is divided into eight genres with a nearly balanced distribution (*law, material for education, microblog, news, science, spoken, subtitles, thesis*). The sentence-level alignments have been manually corrected. However, errors still exist, for example, there are cases where a long segment is not translated in a sentence. We annotated this corpus while filtering out the incomplete or incorrect pairs. The segmentation of Chinese words and the bilingual word alignment are not provided. The corpus is freely available and released with the license *Creative Commons Non-Commercial 4.0.*

UT-corpus (Liu and Sun, 2015): this data set has been constructed by the University of Tsinghua for evaluating the automatic word alignment tool of the authors. It contains 40k sentence pairs. The sentence-level alignments are clean and the word segmentation is provided. The word-level alignments are manually conducted. However, according to the author, the translation direction is sometimes from Chinese to English. The proportion of each genre is unknown (news, subtitles, etc.), but it is sure that the genre *News* occupies a major part. Their corpus is freely available and we can redistribute the annotated corpus.

UB-corpus (Chang and Bai, 2003): this corpus has been constructed by the University of Beijing, mainly for training machine translation systems. The sentence-level alignments have been verified before releasing and the corpus contains a large variety of genres. After signing an agreement, we obtained a corpus of 102k pairs of parallel sentences of genres *Literature, Art* and *Science*, which has been freely provided for research purpose. However, we do not have the right to redistribute this part of the annotated corpus.

UnitedNations-corpus (Ziemski et al., 2016)[4]: this freely available corpus contains official reports and parliamentary documents of the United Nations. Our sub-corpus of genre *Official document* is a sample from this large corpus containing 15M sentence pairs.

For the genre of *scientific article*, after our examination, the quality of the part contained in the UM-corpus is non-satisfactory for annotation. Therefore, we constructed our own corpus by collecting bilingual abstracts from these online journals: *Chinese Linguistics*[5], *Chinese Journal of Software*[6] and *Chinese Journal of Computers*[7]. The translation direction is from Chinese to English. Only those bilingual abstracts offering the same level of content have been retained.

The platform *Linguistic Data Consortium* (LDC)[8] only proposes corpora whose translation direction is from Chinese

[2] https://www.ted.com

[3] These genres are the most used ones in different previous work.

[4] https://cms.unov.org/UNCorpus/en/DownloadOverview

[5] http://www.cuhk.edu.hk/journal/jcl/

[6] http://www.jos.org.cn/josen/ch/index.aspx

[7] http://english.ict.cas.cn/

[8] https://www.ldc.upenn.edu/

to English. Several corpora are aligned at word level, but access is not free. The platform CLARIN[9] provides several English-Chinese corpora whose genres are already covered by the corpus that we mentioned.

We recapitulate in the table 2 different providers of the original corpus.[10] For the first annotation phase, we aim to annotate a sample corpus of 2 200 sentence pairs, which contains 200 pairs for each genre. In this work, we have annotated 50 pairs for each genre to consolidate our annotation guidelines.

Corpus genre	Corpus origin
news	UT-corpus
literature	UB-corpus
art	UB-corpus
scientific article	our own construction
official document	UnitedNations-corpus
law, material for education, microblog, spoken, subtitles, science	UM-corpus

Table 2: Corpus origin for each corpus genre. We take a random sample corpus of each genre for the annotation

4. Typology of Translation Techniques

Based on several previous works which proposed different typologies of translation techniques (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002), we propose a typology of sub-sentential translation techniques for the language pair English-Chinese, established during the manual annotation and the analysis of the phenomena encountered in the corpus.

Figure 1 presents our typology, where the colored blocks represent the categories used for the annotation, and the other blocks serve to establish the hierarchy (*i.e. Non-Literal, Unaligned, No Type*). The annotation of our English-French corpus of TED Talks employs the same typology, which reflects its universality for these two language pairs.

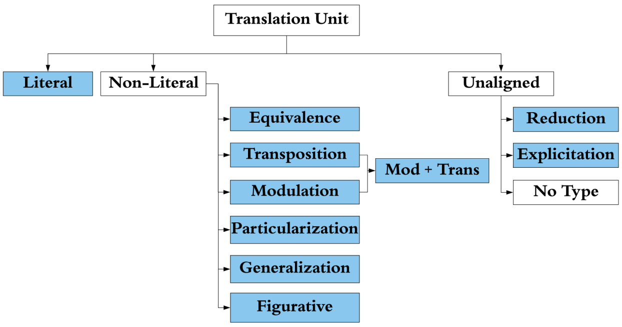


Figure 1: Typology of English-Chinese translation techniques

Compared to proposed typologies in several previous works, our typology presents the following differences:

- The feasibility of the annotation task being taken into consideration, our typology contains less fine-grained categories.
- Certain previous typologies contain the techniques which describe the transformations in two translation directions. In our corpus, the translation direction is from English to Chinese (except for the genre *Scientific article*). Therefore, each category describes the transformation that the Chinese translation has received.
- The translation techniques *calque* and *borrowing* (Vinay and Darbelnet, 1958) are annotated by the category *Literal*.
- The technique *cultural adaptation* is annotated by the category *Equivalence*.
- The category *Transposition* groups together finer categories proposed by Vinay and Darbelnet (1958), for example the *amplification*.
- We have added the combined category *Mod+Trans* and the category *Figurative translation*.
- Since we annotate all words in the corpus, there exist three cases for the unaligned segments: *Reduction*, *Explicitation* and *No Type* (for all the remaining words).

For each category, we present their definition and typical examples in Table 3 and Table 4. In the given examples, the bold part illustrates the translation technique used. For aligned segments, except the eight categories in table 3, we also included three other categories which proved useful during the annotation, but not related to translation techniques : 1) Lexical shift (change of verbal tense, verbal modality or of determiner, differences between plural and singular form, and other minor changes alike); 2) Obvious translation errors; 3) Uncertain cases. The definitions in these two tables are generic, we have completed them with specific rules in our annotation guidelines.

5. Manual Annotation

We have used *Stanford Tokenizer*[11] to tokenize the English corpus, and *THULAC* (Li and Sun, 2009) is used for the Chinese word segmentation. The automatic bilingual word alignment is conducted with *TsinghuaAligner* (Liu and Sun, 2015). These alignments are imported to initialize the annotation, in order to reduce the manual word alignment effort on easy literal word translations. Annotators should verify these automatic word alignments and correct them if needed.

The automatically segmented Chinese corpus contains some errors that could mislead the manual word alignments and the attribution of translation technique categories. Therefore, certain Chinese words need a manual re-segmentation before the annotation, in order to better correspond to English segments. For example, *only is* → 仅仅是 has been corrected to *only is* → 仅仅 *(only)* 是*(is)*. The annotators are told to note down these cases of necessary re-segmentation and the misspellings, which are later corrected in the corpus.

We use the web application *Yawat* (*Yet Another Word Alignment Tool*) (Germann, 2008) for the manual annotation.

[9] `https://www.clarin.eu/resource-families/parallel-corpora`

[10] The redistributing right of our annotated corpus is under negotiation with the corresponding corpus providers.

[11] `http://nlp.stanford.edu/software/tokenizer.shtml`

Translation technique	Definition and important rules
Aligned segments	
Literal	Word-for-word translation: *a* ***bronze*** *ring* → 一 个 **青铜** 戒指 Borrowing words using transliteration: *a cup of* ***coffee*** → 一 杯 **咖啡** Possible literal translation of idioms: ***ivory tower*** → **象牙 塔** Corresponding expression when absolute literal translation does not make sense: ***I give you my word***. → **我 向 你 保证**。 ("I promise you.")
Equivalence	Non-literal translation of proverbs, idioms, or fixed expressions: ***A friend in need is a friend indeed***. → **患难 见 真情**。 ("Misfortune tests the sincerity of friends.") No change in meaning and point of view, a word-for-word translation makes sense but the translator has produced a different translation: *protect all locations* ***at all times*** → **日夜** ("day and night") 保护 所有 的 地点
Transposition	Change grammatical categories without changing the meaning: *She* ***was careful*** *not to question him,* ***fearful*** *that he might leave them.* → 她 也 **小心 地** ("carefully") 从不 问起， **生怕** ("to fear") 他 走 了。
Modulation	Change the point of view, can be encountered both at lexical and syntactic level: *I* ***like*** *the dreams of the future* ***better than*** *the history of the past* . → 我 **不** ("don't") 缅怀 ("recall") 过 去 的 历史，**而** ("but") 致力于 ("devote myself to") 未来 的 梦想。 Slight meaning change at lexical level according to the context: *he had* ***rudely*** *bellowed across the supper table to her* → 他 隔 着 餐桌 对 她 **大声** ("loudly") 吼叫
Mod+Trans	Combine the transformations in *Modulation* and *Transposition*, which causes transformations in both grammatical categories and syntactic structures: *One by one the other elders now timidly rise* ***with*** *innocuous requests.* → 其他 的 长老 一个 接 一个 怯生生 地 站起来，**提出 了** ("put forward") 一些 不关痛痒 的 要求
Particularization	The source segment could be translated into several target segments with more specific meaning, and the translator has chosen one of them according to the context: *"Yes,* ***put*** *you to bed"* → "是的，**服侍** *("serve")* 你 上床 睡觉" Specify the meaning of a segment in context: *On his* ***best days***, *Gomes is a very nice, solid bench player.* → 当 他 **打 得 好 的 时候** ("play well")，戈麦斯 是 很 优秀、很 得力 的 板凳 球员。 Translate a pronoun by the thing(s) it references: *He then requested her to stay where* ***she*** *was* → 他 先 让 **苔丝** ("Tess") 在 外面 等 着
Generalization	Several source words or expressions could be translated into a more general target word or expression, and the translator used the latter to translate: *a research that will be embraced by* ***millions of*** *bleary-eyed Britons* → 一 项 即将 被 **广大** ("numerous") 睡眼惺忪 的 英国人 所 知道 的 研究 The translation of an idiom by a non-fixed expression: ***Every man has a fool in his sleeve.*** → **人人 都有 糊涂 的 时候**。("Every man is a fool sometimes.") The removal of a metaphorical image: *But should clouds gather over the Atlantic, or* ***tempers rise*** *in the Middle East [...]* → 如果 大西洋 风云 再 起，中东 **战火 重 燃** ("war resumes") 的 话 [...]
Figurative translation	Introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor: *He gave the required information,* ***in words as suitable as he could find***. → 他 **字斟句酌 地** ("weigh one's words") 作 了 回答。 Use personification to translate: *For Joanne, new opportunities* ***are opening.*** → 对 乔安娜 而 言，新 的 机遇 **现 已 向 她 招手。** ("are waving to her")

Table 3: Definition and important rules of eight translation techniques for aligned segments

This tool is available for research purpose under the license *GNU Affero General Public License v3.0*. *Yawat* allows us to align words and segments (continuous and discontinuous) in a parallel corpus[12], then to attribute the categories adapted to our task on bilingual units or monolingual units (which means unaligned units) (see an example in Figure 2). The annotation is conducted by the first author and two other Chinese students, both holders of master's degrees in translation studies. Their native language is Chinese.

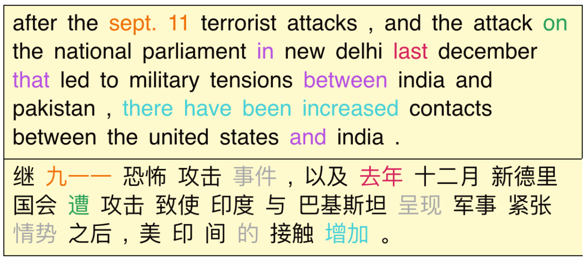
The annotation guidelines are established in an iterative way during the annotation process.[13] In the guidelines, for each category, typical examples, counter-examples as well as difficult borderline examples are systematically provided. We use tables to recapitulate essential information to better guide annotators in making decisions. Annotators are

[12] The boundary of translation units is not fixed in advance, annotators should decide it by themselves.

[13] The guidelines are available: `https://yumingzhai.github.io/files/Annotation_guide_EN_ZH.pdf`

Translation technique	Definition and important rules
Unaligned segments	
Explicitation	Introduce clarifications that are implicit in the source text: *the building blocks of the universe* → 宇宙 **形成**("form") 的 **最** ("most") 基本 单位 Add Chinese-specific words: *the knife* → 这 **把** 刀 (Chinese measure word) *I will bring it to China.* → 我 可以 **把** 它 带到 中国 来。 (necessary addition due to syntactic order change in translation)
Reduction	Deliberately remove certain words in translation (including content words): Removal of preposition: *A spokesman* ***from*** *the Ministry of National Defense* → 国防部 发言人 Removal of copula: *Peter* ***is*** *six years old.* → 彼得 六岁 。 Removal of the anticipatory « it »: ***It*** *was a pleasant surprise to learn of her marriage.* → 得知 她 结婚 是 件 令 人 惊喜 的 事。
No Type	Function words necessary in English but not in Chinese: *The tragedy of the world is* ***that*** *those who are imaginative have but slight experience.* → 世界 的 悲剧 就 在于 有 想象力 的 人 缺乏 经验。 Segments not translated but which do not impact the meaning: *The present state, application and development of coal mine hydraulic drill rig are described* ***in this paper***. → 介绍 了 煤矿 用 液压 钻 车 现状，使用 情况 及 发展 。 Target segments added without reason, which do not correspond to any source segment.

Table 4: Definition and important rules for unaligned segments



after the sept. 11 terrorist attacks , and the attack on the national parliament in new delhi last december that led to military tensions between india and pakistan , there have been increased contacts between the united states and india .

继 九一一 恐怖 攻击 事件 , 以及 去年 十二月 新德里 国会 遭 攻击 致使 印度 与 巴基斯坦 呈现 军事 紧张 情势 之后 , 美 印 间 的 接触 增加 。

Figure 2: The interface of *Yawat* for the annotation task. Black tokens are literally translated. Tokens in purple are unaligned English tokens, and those in grey are unaligned Chinese tokens. The other tokens in different colors are translated using different non-literal translation techniques

encouraged to consult language resources in case of doubt, for example, Cambridge dictionary, Chinese dictionary of idioms, etc. We established annotation conventions concerning the annotation of punctuation, unaligned segments and linguistic anaphora. The guidelines also contain a tutorial about using *Yawat*.

To calculate the inter-annotator agreement, two annotators have independently annotated a control-corpus (100 sentence pairs containing 2k English tokens and 4k Chinese characters). The Kappa of Cohen (Cohen, 1960) is 0.58 for the bilingual segments whose boundaries are the same[14] (which cover 56% of 2k English tokens). Compared to the agreement on our previous English-French control-corpus of TED Talks, which was 0.67 (covering 73% of 3k English tokens), we can see that the annotation is much more difficult on this new language pair.

Since the agreement is moderate, we adopt an annotation scheme where one sub-corpus receives three passes of successive annotation (Zhai et al., 2018), in order to eliminate the disagreement on categories and on segment boundaries. This phase of annotation by three passes is still ongoing.

The statistics of the annotated corpus are presented in the table 5 (including 100 sentence pairs for the control-corpus).

number of sentence pairs	654
number of English tokens	15 739
number of Chinese characters	25 000
average number of EN tokens	24
average number of ZH characters	38

Table 5: Statistics of the annotated corpus

6. Corpus Analysis

In Table 6, we compare the annotation statistics per category of our previously annotated English-French corpus of TED Talks (Zhai et al., 2018), with those of our English-Chinese multi-genre corpus presented in this work. The number of English tokens annotated in each category and their corresponding percentage show that, unsurprisingly, literal translations represent the largest part. Meanwhile, the seven non-literal categories (cf. Figure 1) are also not negligible, they occupy 16.13% (EN-ZH) and 18.32% (EN-FR), respectively.

In total, the percentage of literal translations is higher in the English-French corpus. Chinese translations use much less the complicated category *Modulation+Transposition*, but *Particularization* seems to be more employed. Concerning *Explicitation*, 1 924 Chinese characters have been annotated by this category, which occupy 7.70% of 25 000

[14] Since the segment boundaries are independently decided by two annotators, they can differ.

characters in total; compared to 364 French words in this category, which occupy only 1.02% of 35 588 French words in total. The proportions of *Reduction* and *No type* are also higher in the English-Chinese corpus.

Translation technique	Nb EN tokens		Percentage	
	-ZH	-FR	-ZH	-FR
Literal	9 091	23 733	57.76%	69.49%
Equivalence	752	1 685	4.78%	4.93%
Transposition	624	1 141	3.96%	3.34%
Modulation	510	1 247	3.24%	3.65%
Mod+Trans	56	1 171	0.36%	3.43%
Particularization	361	555	2.29%	1.63%
Generalization	175	401	1.11%	1.17%
Figurative	60	57	0.38%	0.17%
Total non-literal	**2 538**	**6 257**	**16.13%**	**18.32%**
Explicitation	0	0	0.00%	0.00%
Reduction	855	797	5.43%	2.33%
No type	2 416	1 939	15.35%	5.68%
Lexical shift	574	1 049	3.65%	3.07%
Translation error	24	71	0.15%	0.21%
Uncertain	241	306	1.53%	0.90%
Total	15 739	34 152	100.00%	100.00%

Table 6: Comparing the annotation statistics per category of our previous EN-FR corpus of TED Talks and of the EN-ZH corpus presented in this work

Qualitatively, we present below two characteristics of Chinese translations. Structural changes at sentence level tend to be more important in Chinese than in French, and Chinese language prefers to use short and compact clauses. Hence English conjunction words are often replaced by a comma to break a long and complicated sentence to several shorter clauses, for example:

She had little luck as an actress but worked as a model before moving to Hollywood in 1933 for a part in the chorus of Roman Scandals .

她在演艺方面很不走运，便转而以当模特为业。1933年，她搬到好莱坞，在《罗马丑闻》的歌舞队中担任一个角色。

For English-Chinese translations, the word alignments are sometimes much less diagonal than English-French word alignments. For example, Figure 3 shows the word alignment matrix of this pair of sentences:

Many circumstances could be imagined under which he would pass London Bridge.

我们可以设想 ("We can assume")，他过伦敦桥时的情形也许是多种多样的。("when he passes London Bridge the circumstances could be varied.")

Because of syntactic differences, this phenomenon of segment reordering is more evident when translating long and complicated English sentences. It could also occur even though all English words are literally translated.

7. Evaluation

7.1. Compare human and machine translation

During the annotation, we observed that the distance could be large between good human non-literal translations and machine translations provided by online MT services. Humans can recognize these non-literal translations as good quality (Schaeffer and Carl, 2014), but would automatic MT evaluation metrics, such as BLEU (Papineni et al., 2002), penalize them?

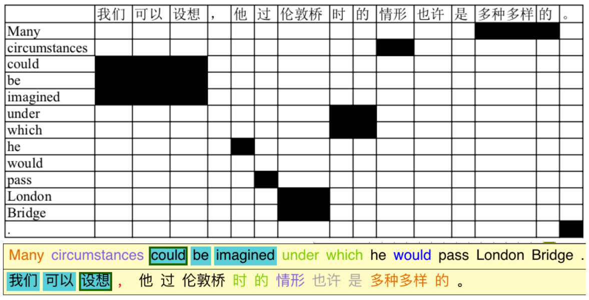

我们
可以
设想
，
他
过
伦敦桥
时
的
情形
也许
是
多种多样
的
。
Many
circumstances
could
be
imagined
under
which
he
would
pass
London
Bridge
.
Many circumstances could be imagined under which he would pass London Bridge .
我们 可以 设想 ， 他 过 伦敦桥 时 的 情形 也许 是 多种多样 的 。

Figure 3: Example of less diagonal word alignment of an English-Chinese sentence pair

In order to study this question, we conducted an experiment to investigate the correlation between the proportion of literally translated English tokens and the BLEU score of the corresponding human translation compared to four machine translations. Four principal MT engines' API have been used during this experiment: Google[15], Microsoft[16], Baidu[17] and Tencent[18].

Cumulative 4-gram BLEU scores with uniform weights are calculated for this experiment.[19] All Chinese translations are tokenized at character level, since Chinese words are formed by combining characters, which are the minimal building blocks of meaning.

The box plot below (see Figure 4) shows the distributions of proportion of literally translated English tokens for each corpus genre.[20] The median value (represented by the orange mark) is the smallest for the genre *Literature*, and the largest for the genre *News*. Indeed, many more non-literal translations are employed when translating literary texts, and many fewer are found in the corpus of news. Besides, Table 7 presents the average number of English tokens and Chinese characters per sentence and per corpus genre. These information reveal the differences across different genres.

For each corpus genre, Table 8 presents Pearson and Spearman correlation coefficient (Benesty et al., 2009; Hauke and Kossowski, 2011) between the proportion of literally translated English tokens and the cumulative 4-gram BLEU score (comparing one human translation to four machine translations). Figures 5 and 6 show the relationship between these two variables for the sub-corpus of official doc-

[15] `https://cloud.google.com/translate/docs/`

[16] `https://azure.microsoft.com/fr-fr/services/cognitive-services/translator-text-api/`

[17] `https://api.fanyi.baidu.com/api/trans/product/index`

[18] `https://ai.qq.com/product/nlptrans.shtml#text`

[19] We compute BLEU scores with the NLP toolkit NLTK (Bird and Loper, 2004). For scoring sentences, we use the sentence_bleu() function with a smoothing function (method 4).

[20] The genre *Scientific article* is ignored for this experiment, since the translation direction is from Chinese to English.

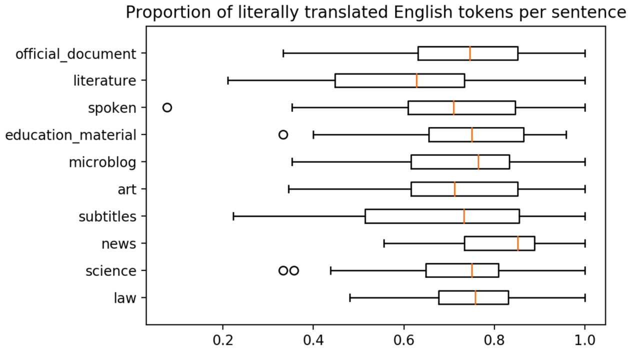


Figure 4: Distributions of proportion of literally translated English tokens per sentence in each genre of corpus

Corpus genre	English	Chinese
official document	34	53
literature	26	40
spoken	14	21
education material	20	31
microblog	21	33
art	27	48
subtitles	12	15
news	31	47
science	20	30
law	37	51

Table 7: Average number of English tokens and Chinese characters per sentence and per corpus genre

ument (UnitedNations-corpus) and of law. The proportion of non-literal translations is the highest for the genre *Literature*, and the Spearman correlation 0.58 (close to strong correlation threshold 0.60) shows significant evidence that non-literal translations get lower BLEU scores. There exist fewer non-literal translations for the genre *Official document*, however, their presence has also been reflected by the strong Spearman correlation (0.66). Only the genre *Law* shows a weak negative correlation, which is rather surprising, since the textual style is close to *Official document*. We obtain weak or even very weak correlation for the other genres, which deserves a more in-depth study.

This experiment is conducted based on 500 sentence pairs annotated (50 pairs for each of ten genres). To confirm our hypothesis that BLEU metric does penalize non-literal human translations, we need to continue the annotation while assuring the annotation quality and the characteristics of each corpus genre. Besides BLEU, we could further investigate other MT metrics, such as METEOR (Banerjee and Lavie, 2005) and TER-plus (Snover et al., 2009), which use paraphrases during the evaluation.

However, preliminary results support our hypothesis for the corpus of genre *Official document* and *Literature*. BLEU scores are lower when human translations are more non-literal than machine translations; and gradually higher when human and machine translations are both more literal and similar. Since the algorithm of BLEU compares the matching n-grams between translations, it could penalize human translations with non-literal but correct expressions.

Corpus genre	Avg. literal proportion	Correlation coefficient	
		Pearson	Spearman
official_document	0.74 ± 0.14	0.67	0.66
literature	0.60 ± 0.18	0.54	0.58
spoken	0.72 ± 0.20	0.40	0.38
education_material	0.74 ± 0.15	0.37	0.37
microblog	0.73 ± 0.17	0.38	0.37
art	0.72 ± 0.16	0.36	0.29
subtitles	0.68 ± 0.23	0.16	0.26
news	0.81 ± 0.11	0.16	0.19
science	0.73 ± 0.16	0.12	0.14
law	0.75 ± 0.11	-0.20	-0.17

Table 8: Correlation coefficient between the proportion of literally translated English tokens in each sentence, and the BLEU-4 score calculated by comparing one human translation with four machine translations. The average literal proportions are presented with their standard deviation

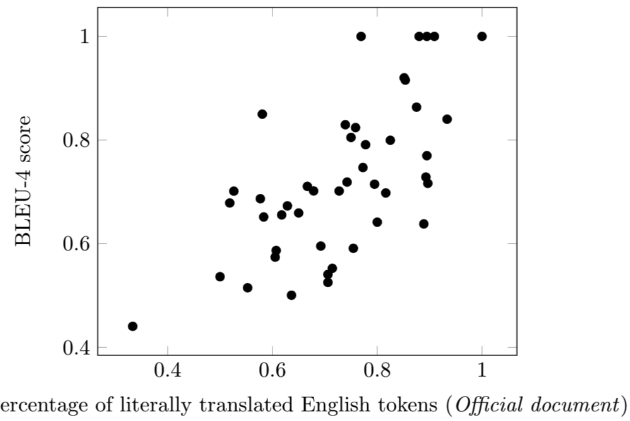


Figure 5: Strong positive correlation between the proportion of literally translated English tokens and the BLEU-4 score on the sub-corpus of United Nations

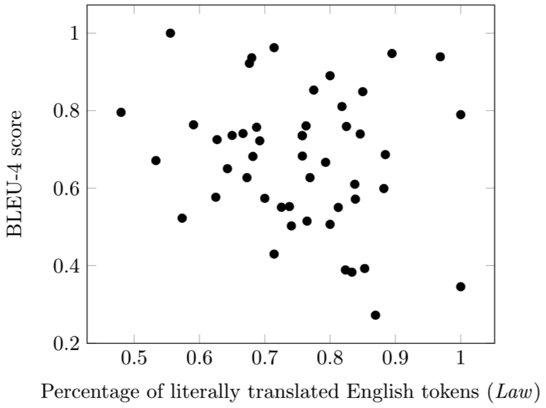


Figure 6: Weak negative correlation between the proportion of literally translated English tokens and the BLEU-4 score on the sub-corpus of law

7.2. Automatic binary classification of translation techniques

The goal of constructing this annotated corpus is to verify the hypothesis that it is possible to automatically recognize different sub-sentential translation techniques. After the deduplication of our annotated instances across corpus genres, the distribution of different categories are shown in table 9. Since the amount of different non-literal instances is still limited, in this work we conduct an experiment of binary classification (literal versus non-literal).[21] We obtain 4 316 literal translations (including the instances of category *Lexical_shift*[22]) and 1 244 non-literal translations by combining all the other categories. Following our previous work (Zhai et al., 2019), the experiment is conducted in a simplified scenario, where the classifier predicts the translation technique of a pair of translations whose boundaries are provided by the annotators.

Translation technique	Nb instances	
Literal	3 982	Literal (4 316)
Lexical shift	334	
Equivalence	356	Non-literal (1 244)
Transposition	315	
Modulation	197	
Mod+Trans	27	
Particularization	242	
Generalization	77	
Figurative	30	

Table 9: The number of annotated instances per category (after deduplication)

We randomly take 1 244 literal translations in order to build a balanced data set. The toolkit *Scikit-Learn* (Pedregosa et al., 2011) is used to train a large variety of statistical supervised classifiers, which are based on different classification algorithms. Default values of their hyperparameters have been used.

The evaluation is based on five-fold cross-validation (with *StratifiedKFold*), using the average accuracy over five folds as metric. The *DummyClassifier* is used as a baseline, which generates random predictions by respecting the distribution of training classes.

For the moment, we have adapted the basic features exploited in our previous work for the language pair English-French (Zhai et al., 2019), by comparing the segment length (number of tokens and characters, the ratio between them) and the difference of Part-of-Speech (PoS) tags (English and Chinese PoS tag sets are mapped to a universal tag set (Petrov et al., 2012)). Other more complicated features will be adapted in our future work, such as exploiting syntactic parsing structures, external linguistic resources ConceptNet (Speer et al., 2017) and information from automatic word alignment.

In Table 10, we compare the binary classification results of two data sets: EN-ZH (1 244 instances for both literal and non-literal class) and EN-FR (1 127 instances likewise). For the EN-FR pair, all above-mentioned features have been used, and the hyperparameters have been tuned by holding out 10% of data as test, and conducting a three-fold cross validation on the remaining data. The best classifiers for the two language pairs are *Multi-layer Perceptron* and *Random Forest*, respectively, and both obtain significantly better results than the baseline of *DummyClassifier*. The experiment on the English-Chinese pair remains to be improved, nonetheless, the preliminary results are favorable to our hypothesis.

Classifier	Average accuracy over five folds (with standard deviation)	
	EN-ZH	EN-FR
Dummy (baseline)	52.21% ± 0.00%	53.19% ± 0.10%
MLP	**70.10% ± 1.06%**	84.65% ± 2.27%
GradientBoosting	69.58% ± 1.30%	86.20% ± 2.03%
Adaboost	69.21% ± 1.27%	83.41% ± 1.53%
LogisticRegression	69.05% ± 1.10%	84.78% ± 1.92%
RandomForest	68.89% ± 0.64%	**87.22% ± 1.92%**
MultinomialNB	68.33% ± 0.78%	80.83% ± 2.78%
DecisionTree	68.13% ± 1.75%	79.68% ± 1.90%
BernoulliNB	66.32% ± 1.33%	81.50% ± 1.51%
SVM	66.28% ± 0.93%	85.14% ± 2.08%
KNN	65.48% ± 4.84%	83.50% ± 0.67%
GaussianNB	59.29% ± 5.36%	64.15% ± 2.03%

Table 10: Classification results of distinguishing literal and non-literal sub-sentential translations. Comparison of performance on the EN-ZH and EN-FR data sets

8. Conclusion

Human non-literal translation techniques have been widely examined in translation studies and in contrastive linguistics. However, they receive less attention in developing NLP applications. One of our long-term objectives is to leverage the automatic classification of sub-sentential translation techniques to improve the quality of paraphrase resources. In this work, we extend our previous studies of manual annotation and of automatic classification for the English-French pair to the more distant English-Chinese pair. We have presented our multi-genre corpus, the details of manual annotation and the characteristics of Chinese translation. Fifty sentence pairs for each genre have been annotated in order to consolidate our annotation guidelines. We conducted two experiments to verify our hypothesis, which are supported by our preliminary results: 1) BLEU scores could penalize non-literal human translations when they are more different from machine translations; 2) it is possible to automatically distinguish literal and non-literal translations. We will continue our effort on annotating high-quality parallel corpus and on fine-grained multi-class classification. We hope that this annotated corpus will be useful for linguistic contrastive studies and for fine-grained evaluation of NLP tasks, such as automatic word alignment and machine translation.

9. Bibliographical References

Ahrenberg, L. (2017). Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria, September. Association for Computational Linguistics, Shoumen, Bulgaria.

[21] All the code used in this work is available: `https://github.com/YumingZHAI/human_vs_machine_translation`.

[22] In our data set, *Lexical_shift* instances are very close to literal translations after lemmatizing the English segment.

Ahrens, K., Zeng, H., and Wong, S.-h. R. (2018). Using a Corpus of English and Chinese Political Speeches for Metaphor Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Chang, B. and Bai, X. (2003). The Markup Guidelines for the Chinese-English Parallel Corpus of Peking University. *Journal of Chinese Language and Computing*, 13(2):195–214.

Chen, Q., Kwong, O. Y., and Zhu, J. (2018). Detecting free translation in parallel corpora from attention scores. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December. Association for Computational Linguistics.

Chuquet, H. and Paillard, M. (1989). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Deng, D. and Xue, N. (2017). Translation Divergences in Chinese–English Machine Translation: An Empirical Investigation. *Computational Linguistics*, 43(3):521–565.

Dorr, B. J., Pearl, L., Hwa, R., and Habash, N. (2002). Duster: A method for unraveling cross-language divergences for statistical word-level alignment. In *Conference of the Association for Machine Translation in the Americas*, pages 31–43. Springer.

Fadaee, E. (2011). Translation techniques of figures of speech: A case study of George Orwell's "1984 and Animal Farm". *Journal of English and Literature*, 2(8):174–181.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Germann, U. (2008). Yawat: Yet Another Word Alignment Tool. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Demo Papers*, pages 20–23. The Association for Computer Linguistics.

Hauke, J. and Kossowski, T. (2011). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Larson, M. L. (1984). *Meaning-based translation: A guide to cross-language equivalence*, volume 366. University press of America Lanham, MD.

Li, Z. and Sun, M. (2009). Punctuation As Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, 35(4):505–512, December.

Liu, Y. and Sun, M. (2015). Contrastive unsupervised word alignment with non-local features. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2295–2301. AAAI Press.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 881–893.

Melamed, I. D. (1998a). Annotation style guide for the blinker project, version 1.0. *Technical Report IRCS TR*.

Melamed, I. D. (1998b). Manual annotation of translational equivalence: The blinker project. *Technical Report IRCS TR*.

Molina, L. and Hurtado Albir, A. (2002). Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta*, 47(4):498–512.

Monti, J., Sangati, F., and Arcan, M. (2015). TED-MWE: a bilingual parallel corpus with MWE annotation. Towards a methodology for annotating MWEs in parallel multilingual corpora. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 193–197, Trento.

Newmark, P. (1981). *Approaches to Translation (Language Teaching Methodology Senes)*. Oxford: Pergamon Press.

Newmark, P. (1988). *A textbook of translation*, volume 66. Prentice Hall New York.

Đorđević, J. (2017). Translation techniques revisited: the applicability of existing solutions in non-literary translation. *FACTA UNIVERSITATIS-Linguistics and Literature*, 15(1):35–47.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pavlick, E., Bos, J., Nissim, M., Beller, C., Van Durme, B., and Callison-Burch, C. (2015). Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1512–1522.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petrov, S., Das, D., and McDonald, R. T. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096. European Language Resources Association (ELRA).

Schaeffer, M. and Carl, M. (2014). Measuring the cognitive effort of literal translation processes. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 29–37.

Snover, M. G., Madnani, N., Dorr, B., and Schwartz, R. (2009). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., Lu, Y., Li, S., Wang, Y., and Wang, L. (2014). UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Vinay, J.-P. and Darbelnet, J. (1958). *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

Xu, Y. and Yvon, F. (2016). Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 628–635, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Zhai, Y., Max, A., and Vilnat, A. (2018). Construction of a Multilingual Corpus Annotated with Translation Relations. In *First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111.

Zhai, Y., Safari, P., Illouz, G., Allauzen, A., and Vilnat, A. (2019). Towards Recognizing Phrase Translation Processes: Experiments on English-French. *CoRR*, abs/1904.12213.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Portorož, Slovenia, May. European Language Resources Association (ELRA).