Detecting Syntactic Features of Translated Chinese

Hai Hu, Wen Li, Sandra Kübler
Indiana University
{huhai,wl9,skuebler}@indiana.edu

Abstract

We present a machine learning approach to distinguish texts translated to Chinese (by humans) from texts originally written in Chinese, with a focus on a wide range of syntactic features. Using Support Vector Machines (SVMs) as classifier on a genre-balanced corpus in translation studies of Chinese, we find that constituent parse trees and dependency triples as features without lexical information perform very well on the task, with an F-measure above 90%, close to the results of lexical *n*-gram features, without the risk of learning topic information rather than translation features. Thus, we claim syntactic features alone can accurately distinguish translated from original Chinese. Translated Chinese exhibits an increased use of determiners, subject position pronouns, NP + "的" as NP modifiers, multiple NPs or VPs conjoined by "、", among other structures. We also interpret the syntactic features with reference to previous translation studies in Chinese, particularly the usage of pronouns.

1 Introduction

Work in translation studies has shown that translated texts differ significantly in subtle and not so subtle ways from original, non-translated texts. For example, Volansky et al. (2013) show that the prefix *mono-* is more frequent in Greek-to-English translations because epistemologically it originates from Greek. Also, the structure of modal verb, infinitive, and past participle (e.g. *must be taken*) is more prevalent in translated English from 10 source languages.

We also know that a machine learning based approach can distinguish translated from original texts with high accuracy for Indo-European languages such as Italian (Baroni and Bernardini, 2005), Spanish (Ilisei et al., 2010), and English (Volansky et al., 2013; Lembersky et al., 2012;

Koppel and Ordan, 2011). Features used in those studies include common bag-of-words features, such as word *n*-grams, as well as part-of-speech (POS) *n*-grams, function words, etc. Although such surface features yield very high accuracy (in the high nineties), they do not contain much deeper syntactic information, which is key in interpreting textual styles. Furthermore, despite the large amount of research on Indo-European languages, few studies have quantitatively investigated either lexical or syntactic features of translated Chinese, and to our knowledge, no automatic classification experiments have been conducted for this language.

Thus the purpose of this paper is two-fold: First, we perform translated vs. original text classification on a balanced corpus of Chinese, in order to verify whether translationese in Chinese is as real as it is in Indo-European languages, and to discover which structures are prominent in translated but not original Chinese texts. Second, we show that using only syntactic features without any lexical information, such as context-free grammar (CFG), subtrees of constituent parses, and dependency triples, perform almost as well as lexical *n*-gram features, confirming the translationese hypothesis from a purely syntactic point of view. These features are also easily interpretable for linguists interested in syntactic styles of translated Chinese. We analyze the top syntactic features ranked by a common feature selection algorithm, and interpret them with reference to previous studies on translationese features in Chinese.

2 Related Work

2.1 Translated vs. Original Classification

The pioneering work of Baroni and Bernardini (2005) is one of the first to use machine learning methods to distinguish translated and orig-

inal (Italian) texts. They experimented with word/lemma/POS *n*-grams and mixed representations and reached an F-measure of 86% using recall maximizing combinations of SVM classifiers. In the mixed *n*-gram representation, they used inflected wordforms for function words, but replaced content words with their POS tags. The high F-measure (85.2%) with such features shows that "function word distributions and shallow syntactic patterns" without any lexical information can already account for much of the characteristics of translated text.

Volansky et al. (2013) is a very comprehensive study that investigated translationese in English by looking at original and translated English from 10 source languages, in a European parliament corpus. While they mainly aimed to test translational universals, e.g. simplification, explicitation, etc., the classification accuracy with SVMs using features such as POS trigrams (98%), function words (96%), function word *n*-grams (100%) provided more evidence that function words and surface syntactic structures may be enough for the identification of translated text.

For Chinese, however, there are very few quantitative studies on translationese (apart from Xiao and Hu, 2015; Hu, 2010, etc.). Xiao and Hu (2015) built a comparable corpus containing 500 original and translated Chinese texts respectively, from four genres. They used statistical tests (log-likelihood tests) to find statistical differences between translated and original Chinese with regard to the frequency of mostly lexical features. They discovered, for example, that translated text use significantly more pronouns than the original texts, across all genres. But they were unable to investigate the syntactic contexts in which those overused pronouns occur most often.

For them, syntactic features were examined through word *n*-grams, similar to previous studies on Indo-European languages, but no text classification task was carried out.

2.2 Syntactic Features in Text Classification

Although *n*-gram features are more prevalent in text-classification tasks, deep syntactic features have been found useful as well. In the Native Language Identification (NLI) literature, which in many respects is similar to the task of detecting translations, various forms of context-free grammar (CFG) rules are often used as features (Bykh

# texts	news	general prose	science	fiction	total
LCMC	88	206	80	111	485
ZCTC	88	206	80	111	485

Table 1: Distribution of texts across genres

and Meurers, 2014; Wong and Dras, 2011). Bykh and Meurers (2014) showed that using a form of normalized counts of *lexicalized* CFG rules plus *n*-grams as features in an ensemble model performed better than all other previous systems. Wong and Dras (2011) reported that using unlexicalized CFG rules (except for function words) from two parsers yielded statistically higher accuracy than simple lexical features (function words, character and POS *n*-grams).

Other approaches have used rules of tree substitution grammar (TSG) (Post and Bergsma, 2013; Swanson and Charniak, 2012) in NLI. Swanson and Charniak (2012) compared the results of CFG rules and two variants of TSG rules and showed that TSG rules obtained through Bayesian methods reached the best results.

Nevertheless, such deep syntactic features are rarely used, if at all, in the identification of translated texts. This is the gap that we hope to fill.

3 Experimental Setup

3.1 Dataset

We use the comparable corpus by Xiao and Hu (2015), which is composed of 500 original Chinese texts from the Lancaster Corpus of Modern Chinese (LCMC), and another 500 human translated Chinese texts from the Zhejiang-University Corpus of Translated Chinese (ZCTC). All texts are of similar lengths (~2000 words), and from different genres. There are four broad genres: news, general prose, science, and fiction (see Table 1), and 15 second-level categories. We exclude texts from the second-level categories "science fiction" and "humor" (both under fiction) since they only have 6 and 9 texts respectively, which is not enough for a classification task.

LCMC (McEnery and Xiao, 2004) was originally designed for "synchronic studies of Chinese and the contrastive studies of Chinese and English" (see Xiao and Hu, 2015, chapter 4.2). It includes written Chinese sampled from 1989 to 1993, amounting to about one million words. ZCTC was created specifically for translation

studies "as a comparable counterpart of translated Chinese" to LCMC (Xiao and Hu, 2015, pp. 48), with the same genre distribution and also one million words in total. The texts in ZCTC are sampled in 2001, all translated by human translators, with 99% originally written in English (pp. 50).

Both corpora contain texts that are segmented and POS tagged, processed by the corpus developers using the 2008 version of ICTCLAS (Zhang et al., 2003), a common tagger used in Chinese NLP research. However, only the segmentation is used in this study since our parser uses a different POS tagset.

In this study, we perform 5-fold cross validation on the whole dataset and then evaluate on the full set of 970 texts.

3.2 Pre-Processing and Parser

We remove URLs and headlines, normalize irregular ellipsis (e.g. "。 。 。", "....") to "......", change all half-width punctuations to full-width, so that our text is compatible with the Chinese Penn Treebank (Xue et al., 2005), which is the training data for the `Stanford CoreNLP` parser (Manning et al., 2014) used in our study.

3.3 Features

Character and word *n*-gram features can be considered upper bound and baseline. On the one hand, they have been used extensively (see Section 2), but on the other hand, they partially encode topic information rather than stylistic differences because of their lexical nature. Consequently, while they are very informative in the current setup, they may not be useful if we want to use the trained model on other texts.

For syntactic features, we use various forms of constituent and dependency parses of the sentences. We extract the following features based on either type of parse using the `CoreNLP` parser with its pre-trained parsing model.

3.3.1 Context-Free Grammar

Context-free grammar rules (CFGR) We use the count of each CFG rule extracted from the parse trees.

Subtrees Subtrees are defined as any part of the constituent tree of any depth, closely following the data-oriented parsing (DOP) paradigm (Bod et al., 2003; Goodman, 1998). Our features differ from the DOP model as well as TSG (Post and Gildea,
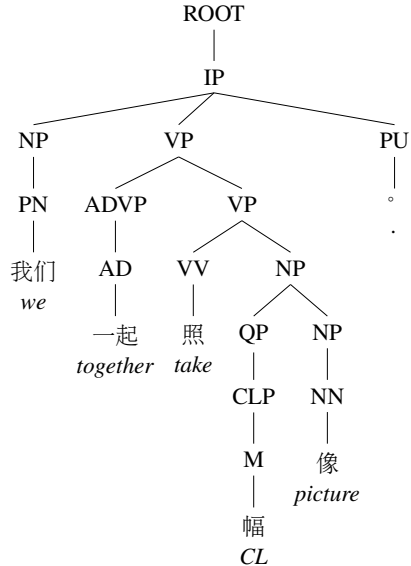

Figure 1: Example constituent tree of the Chinese sentence meaning *We take a picture together*
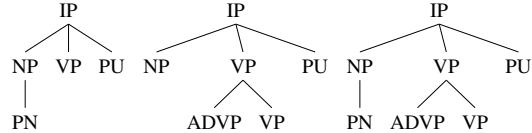

Figure 2: All subtrees of depth 2 with root IP in the tree from Figure 1

2009; Sangati and Zuidema, 2011; Swanson and Charniak, 2012) in that we do not include any lexical information in order to exclude topical influence from content words. Thus no lexical rules are considered, and POS tags are considered to be the leaf nodes (Figure 2).

We experiment with subtrees of depth up to 3 since the number of subtrees grows exponentially as the depth increases. With depth 3, we are already facing more than 1 billion features. Performing subtree extraction and feature selection becomes difficult and time consuming. Also note that CFGRs are essentially subtrees of depth 1. So with increasing maximum depth of subtrees, we test fewer local relations in constituent parses. In the future, we plan to use Bayesian methods (Post and Gildea, 2009) to sample from all the subtrees.

We also conduct separate experiments using subtrees headed by a specific label (we only look at NP, VP, IP, and CP, since they are the most frequent types of subtrees). For example, using NP subtrees as features will inform us how important the noun phrase structure is in identifying translationese.
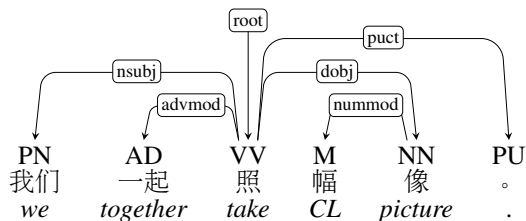
Figure 3: Example dependency graph

3.3.2 Dependency Graphs

Dependency relations, as well as the head and dependent are extracted to construct the following features.

depTriple We combine the POS of a head and its dependent along with the dependency relation, e.g., [VV, nsubj, PN] describes a dependency relation of a nominal subject (nsubj) between a verb (VV) and a pronoun (PN).

depPOS Here only the POS tags of the head and dependent are used, e.g., [VV, PN].

depLabel Only the dependency relation, e.g., [nsubj].

depTripleFuncLex Same as depTriple, except when the word is a function word, we use the lexical item instead of the POS. e.g. [VV, nsubj, 我们] where "我们" (*we*) is a function word (Figure 3).

It should be noted that no lexical information are included in our syntactic features, except for the function words in depTripleFuncLex.

3.3.3 Combination of Features

If combined feature sets work significantly better than one feature set alone, we can draw the conclusion that they model different characteristics of translationese. We experiment with combination of CFGR/subtree and depTriple features.

3.4 Classifier and Feature Selection

For the machine learning experiments, we use support vector machines, in the implementation of the svm.SVC classifier in scikit-learn (Pedregosa et al., 2011). We perform 5-fold cross validation and average over the results. When extracting the folds, we perform stratified sampling across genres so that both training and test data are balanced. Since the number of CFGR/subtree features is much greater than the number of training texts, we perform feature selection by filtering using information gain (Liu et al., 2016; Wong and

Features	F-measure (%)
char *n*-grams(1-3)	95.3
word *n*-grams(1-3)	94.3
POS *n*-grams(1-3)	93.9

Table 2: Results for the lexical and POS features

Dras, 2011) to choose the most discriminative features. Information gain has been shown to select highly discriminative, frequent features for similar tasks (Liu et al., 2014). We experiment with different numbers of features, ranging between the values of 100, 1 000, 10 000, and 50 000.

4 Results

4.1 Empirical Evaluation

First we report the results based on lexical and POS features in Table 2 (F-measure).

Character *n*-grams perform the best, achieving an F-measure of 95.3%, followed by word *n*-grams with an F-measure of 94.3%. Both settings include content words that indicate the source language. In fact, out of the top 30 character *n*-gram features that predict translations, 4 are punctuations, e.g., the first and family name delimiter "·" in the translations of English names and parentheses "（）"; 11 are function words, e.g. "的" (particle), "可能" (*maybe*), "在" (*in/at*), and many pronouns (*he, I, it, she, they*); all others are content words, where "斯" (*s*) and "尔" (*r*) are at the very top, mainly because they are common transliterations of foreign names involving "s" and "r", followed by "公司" (*company*), "美国" (*US*), "英国" (*UK*), etc. Lexical features have been extensively analyzed in Xiao and Hu (2015), and they reveal little concerning syntactic styles of translated text; thus we will refrain from analyzing them here.

POS *n*-grams also produce good results (F-measure of 93.9%), confirming previous research on Indo-European languages (Baroni and Bernardini, 2005; Koppel and Ordan, 2011). Since they are not lexicalized and thus avoid a topical bias, they provide a better comparison to syntactic features.

Syntactic features: Table 3 presents the result for the syntactic features described in Section 3.3. The best performing unlexicalized syntactic features can reliably classify texts into "original" and "translated", with F-measures greater than 90%,

Features	F (%)
Unlexicalized syntactic features	
CFGR	90.2
subtrees: depth 2	90.9
subtrees: depth 3	92.2
depTriple	91.2
depPOS	89.9
depLabel	89.5
depTripleFuncLex	93.8
Combinations of syntactic features	
CFGR + depTriple	90.5
subtree_d2 + depTriple	91.0
POS n-grams + unlex syn features	
POS + subtree_d2	93.6
POS + depTriple	93.4
POS + subtree_d2 + depTriple	93.8
Char n-grams + unlex syn features	
char + subtree + depTriple	94.4
char + pos + subtree + depTriple	95.5

Table 3: Classification based on syntactic features

Features	F (%)
CFGR NP	86.4
CFGR VP	85.6
CFGR IP	86.6
CFGR CP	68.4
subtrees NP d2	86.0
subtrees VP d2	85.6
subtrees IP d2	89.0
subtrees CP d2	71.6
subtrees NP d3	83.6
subtrees VP d3	86.7
subtrees IP d3	86.9
subtrees CP d3	77.7

Table 4: Results for individual subtrees

which are close to the performance of the purely lexicalized features in Table 2. This suggests that although lexical features do achieve slightly better results, syntactic features alone can capture most of the differences between original and translated texts.

Note that when we increase the depth of constituent parses from 1 (CFGR) to subtrees of depth 3, the F-measure increases by 2 percent, which is a highly significant difference (McNemar (McNemar, 1947) on the 0.001 level). Thus, including deeper constituent information proves helpful in detecting the syntactic styles of texts.

However, combination of different types of syntactic features does not increase the accuracy over the dependency results. Adding syntactic features to POS *n*-gram or character *n*-gram features decreases the POS *n*-gram results slightly, thus indicating that both types of features cover the same information, and POS *n*-grams are a good approximation of shallow syntax. The lack of improvement when adding syntactic features may also be attributed to their unlexicalized nature in this study. Our syntactic features are completely unlexicalized, whereas research in NLI has shown that CFGR features need to include at least the function words to give higher accuracy (Wong and Dras, 2011). Although this suggests that in terms of classification accuracy, unlexicalized syntactic features cannot provide more information than *n*-gram features, we can still draw some very interesting observations about styles of translated and original texts, many of which are not possible with simple *n*-gram features. We will discuss those in the following sections.

4.2 Constituency Features

The top ranking CFG features are shown in Table 5. The top three features in translated section (bottom half) of the table tell us that pronouns (PN) and determiners (DT) are indicative of translated text. We will discuss pronouns in Section 5; as for determiners, dependency graph features in Table 7 further show that among them, "该" (*this*), "这些" (*these*) and "那些" (*those*) are the most prominent. The parenthesis rule (PRN) captures another common feature of translation, i.e., giving the original English form of proper nouns ("加州大学洛杉矶分校（UCLA）") or putting translator's notes in parentheses. Furthermore, the prominence of the two rules NP → DNP NP and DNP → NP DEG in translation indicates that when an NP is modified by another NP, translators tend to add the particle "的" (DE; DEG for DE Genitive) between the two NPs, for example:

- (NP (DNP (NP 美国) (DEG 的)) (NP 政治)). Gloss: "US DE politics", i.e. US politics

- (NP (DNP (NP 舆论) (DEG 的)) (NP 谴责)). Gloss: "media DE criticism", i.e. criticism from the media

- (NP (DNP (NP 脑) (DEG 的)) (NP 供血)). Gloss: "brain DE blood supply", i.e. cerebral circulation

In all three cases above, "的" can be dropped, and the phrases remain grammatical. But there are

Rank	CFGR	Predicts
2.0	VP → VP PU VP	original
5.0	VP → VP PU VP PU VP	original
10.0	NP → NN	original
10.2	NP → NN PU NN	original
13.6	IP → NP PU VP	original
14.8	NP → NN NN	original
15	NP → ADJP NP	original
16.6	IP → NP PU VP PU	original
18.2	VP → VV	original
19.6	VP → VV NP	original
1.0	NP → PN	translated
4.0	NP → DP NP	translated
6.2	DP → DT	translated
6.6	IP → NP VP PU	translated
6.8	PRN → PU NP PU	translated
6.8	NP → NR	translated
10.0	CP → ADVP IP	translated
10.6	NP → DNP NP	translated
16.4	ADVP → CS	translated
16.8	DNP → NP DEG	translated

Table 5: Top 20 CFGR features; rank averaged across 5-fold CV

Rank	NP CFGR	Predicts
2.0	NP → NN	original
4.0	NP → NN NN	original
5.4	NP → NN **PU** NN	original
6.2	NP → ADJP NP	original
9.8	NP → NN **PU** NN **PU** NN	original
9.8	NP → NP ADJP NP	original
12.2	NP → NP PU NP	original
12.6	NP → NN NN NN	original
14.6	NP → NP NP	original
17.0	NP → NP QP NP	original
18.4	NP → QP NP	original
1.0	NP → PN	translated
4.2	NP → DP NP	translated
6.0	NP → NR	translated
7.2	NP → DNP NP	translated
14.4	NP → QP DNP NP	translated
16.2	NP → NP PRN	translated
16.2	NP → NR **CC** NR	translated
18.2	NP → NP **CC** NP	translated

Table 6: Top 20 NP features (PN: pronoun; NR: proper N; CC: coordinating conjunction)

many cases where "的" is mandatory in the "NP modifying NP" structure. Thus, it is easier to use "的", since it is almost always grammatical, but decisions when to drop "的" are much more subtle. Translators seem to make the safer decision by always using the particle after the NP modifiers, thus making the structure more frequent.

Now we turn to features of subtrees rooted in specific syntactic categories. The classification results are shown in Table 4. Using only NP-headed rules gives us an F-measure of 86.4%. Larger subtrees fare slightly worse, probably indicating data sparsity. However, these results mean that noun phrases alone often provide enough information whether the text is translated.

Table 6 shows the top 20 CFGR features headed by an NP. This gives us an idea of the distinctive structures of noun phrases in original and translated texts. Apart from the obvious over-use of pronouns (PN) and determiner phrases (DP) for NPs in translated text, there are other very interesting patterns: For original Chinese, nouns inside a complex noun phrase tend to be conjoined by a Chinese specific punctuation "、"(similar to the comma in "I like apples, oranges, bananas, etc."), indicated by the high ranking of NP rules involving PU. This punctuation is most often used to separate elements in a list, and a check using Tregex

(Levy and Andrew, 2006) for the parsed sentences retrieves many phrases like the following from the LCMC corpus: "全院医生、护士最先挖掘的..." (*doctors, nurses from the hospital first dug out...*). In contrast, in translated Chinese, those nouns are more likely to be conjoined by a conjunction (CC), exemplified by the following example from the ZCTC corpus: "对经济**和**股市非常敏感" (*very sensitive to the economy **and** the stock market.*). Here, to conjoin doctors and nurses, or the economy and the stock market, either "、" or "*and*" is grammatical, but original texts favor the former while the translated text, probably influenced by English, prefers the conjunction.

4.3 Dependency Features

Features based on dependency parses have similar F-measures, but should be easier to obtain than subtrees of depth greater than 1. Using the lexical items for function words (depTripleFuncLex) can further improve the results, showing that the choice of function words is indeed very indicative of translationese. A selection of top ranking depTripleFuncLex features is shown in Table 7.

Chinese-specific punctuations such as "、" predicts original Chinese text, as we have already seen, but notice that it is also often used to conjoin verbs (VV_PUNCT_、). Translated texts, in contrast, use more determiners (*these*, *such*, *those*,

25

Rank	Feature	Predicts	Gloss
1.0	VV_CONJ_VV	original	
2.4	VV_PUNCT_，	original	
2.6	NN_PUNCT_、	original	
4.8	VV_PUNCT_、	original	
11.0	NN_CONJ_NN	original	
18.0	NN_DET_各	original	each
21.4	VA_PUNCT_，	original	
25.0	NN_ETC_等	original	etc.
28.2	VV_PUNCT_：	original	
33.2	VV_PUNCT_！	original	
39.0	NN_DEP_三	original	three
41.2	NN_DET_全	original	all
42.6	VA_**NSUBJ**_NN	original	
77.2	VV_DOBJ_NN	original	
94.8	VV_**NSUBJ**_NN	original	
5.4	VV_**NSUBJ**_我	translated	I
8.2	VV_ADVMOD_将	translated	will
10.0	VV_**NSUBJ**_他	translated	he
10.2	NN_DET_该	translated	this
11.6	NN_DET_这些	translated	these
14.0	NR_CASE_的	translated	DE
17.0	VV_**NSUBJ**_他们	translated	they
24.0	VV_**NSUBJ**_她	translated	she
27.6	他_CASE_的	translated	his
29.6	NN_NMOD:ASSMOD_他	translated	he
31.0	VV_PUNCT_。	translated	period
33.6	VV_ADVMOD_但是	translated	but
35.6	VV_**NSUBJ**_你	translated	you
35.8	VV_ADVMOD_如果	translated	if
37.6	VV_MARK_的	translated	DE
37.8	NN_DET_任何	translated	any
40.6	VV_CASE_因为	translated	because
41.2	NR_CC_和	translated	and
44	NN_DET_那些	translated	those
47.2	VV_NSUBJ_它	translated	it
191.0	VV_DOBJ_它	translated	it

Table 7: Top depTripleFuncLex features

Rank	Feature	Function
1.0	(NP **PN**)	NA
2.2	(IP (NP **PN**) VP)	Subj.
5.2	(DNP (NP **PN**) DEG)	Genitive
6.6	(IP (NP **PN**) VP PU)	Subj.
38.0	(IP (NP **PN**) (VP VV VP))	Subj.
56.0	(IP (NP **PN**) (VP ADVP VP))	Subj.
77.0	(IP (ADVP) (NP **PN**) VP)	Subj.
81.0	(IP (NP **PN**) (VP ADVP VP) PU)	Subj.
81.0	(IP (ADVP AD) (NP **PN**) (VP))	Subj.
93.5	(PP P (NP **PN**)	Obj. of prep.
93.5	(IP (NP **PN**) (VP (VV IP))	Subj.
93.6	(VP VV (NP **PN**) IP)	Obj. of verb

Table 8: Top subtree (depth=2) features involving pronouns (PN)

each, etc.) and pronouns (*he*, *they*, etc.), which will be discussed in more detail in the following section. These results are in accordance with previous research on translationese in Chinese (He, 2008; Xiao and Hu, 2015).

5 Analyzing Features: Pronouns

In this section, we discuss one example where syntactic features provide unique information about the stylistic differences between original and translated Chinese that cannot be extracted from lexical sequences, yielding new insights into translationese in Chinese: We have a closer look at the use of pronouns. For this investigation, we examine the top 100 subtrees with depth 2, selected by information gain.

Our results not only confirm the previous finding that pronoun usage is more prominent in translated Chinese (He, 2008; Xiao and Hu, 2015, among others, see Section 2.1), but also provide more insights on the details of pronoun usage in translated Chinese, by looking at the syntactic structures that involve a pronoun (PN) and their ranking after applying the feature ranking algorithm (see Table 8).

The high ranking of pronoun-related features (4 out of the top 10 features involve pronouns) confirms the distinguishing power of pronoun usage. Crucially, it appears that pronouns in subject position or as a genitive (as part of DNP phrase such as 他的书, *his book*), are more prominent than pronoun in the object position in translated texts. In fact, pronouns as the object of a preposition (captured by subtree "(PP P (NP PN))") ranked only about 93rd among all features. Also, pronouns as the object of a verb only shows up once in the top 100 features, and they are of the structure "(VP VV (NP PN) IP)". When searching for sentences with such structures (using Tregex), we almost always encounter phrases similar to "make + pronoun + V.", e.g. "让 他们 懂得 ..." (*make them understand ...*), where the pronoun is both the object of "make", and the subject of "understand". All this shows that the over-usage of pronouns in translated texts is more likely to occur in subject positions, or in a genitive complement, rather than as the direct object of a verb. Even when it appears in the object position, it appears to play both the roles of subject and object. To our knowledge, this characteristic has not been discussed in previous studies in translationese.

If we examine the dependency features, we see the same pattern. Pronouns serving as the subject of verbs rank very high (5.4, 10, 17, 24, 35.6, see Table 7), whereas pronouns as the object of verbs

are not in the top 100 features (the highest ranking 191, VV_DOBJ_它 *it*). Thus we see the two types of syntactic features (constituent trees and dependency trees) converging to the same conclusion. If we look at the pronoun issue from the opposite side, a reasonable consequence would be that in original texts, more common nouns should serve as the subject, which is indeed what we find. VV_NSUBJ_NN predicts "original" and ranks 94.8.

The conclusion concerning pronoun usage drawn from the ranking of syntactic features coincides with observation of (non-)pro-drop in English and Chinese. I.e., Chinese is pro-drop while Enlgish is not. Thus, the overuse of pronouns in Chinese texts translated from English is an example of the interference effect (Toury, 1979), where translators are likely to carry over linguistic features in the source language to the target language. A further observation is that, in Chinese, subject pro-drop seems to be more frequent. The reason is that subject pro-drop does not require much context, while object-drop generally requires the dropped object to be discourse old (c.f. Li and Thompson, 1981). This explains why pronoun overuse occurs more often in subject position in translated text, because object pro-drop in Chinese itself is less common in original Chinese text.

We are not trying to imply that lexical features should not be used. Rather, we want to stress that syntactic features offer a more in-depth and comprehensive picture to linguists interested in the style of translated text. The pronoun analysis presented above is only one such example. We can perform such analyses for any feature of interest and gain a deeper understanding of how they occur in both types of text.

6   Conclusion and Future Work

To our knowledge, the current study is the first machine learning experiment on translated vs. original Chinese. We find that translationese can be identified with roughly the same high accuracy using either lexical *n*-gram features or syntactic features. More importantly, we show how syntactic features can yield linguistically meaningful features that can help decipher differences in styles of translated and original texts. For example, translated Chinese features more determiners, subject-position pronouns, NP modifiers involving "的", and multiple NPs or VPs conjoined by the Chinese-specific punctuation "、". Our methodology can, in principle, be applied to any stylistic comparisons in the digital humanities, and can yield stylistic insights much deeper than the pioneering work of Mosteller and Wallace (1963).

In future work, we will investigate tree substitution grammar (TSG), which extracts even deeper constituent trees (c.f. Post and Gildea, 2009), and detailed feature interpretation for phrases headed by other tags (ADJP, PP, etc.) and for specific genres. It is also desirable to improve the accuracy of constituent parsers for Chinese, along the lines of (Wang et al., 2013; Wang and Xue, 2014; Hu et al., 2017), since accurate syntactic trees are the prerequisite for accurate feature interpretation. While the parser in this study works well, better parsers will undoubtedly be a plus.

Acknowledgement

References

Marco Baroni and Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing* 21(3):259–274.

Rens Bod, Remko Scha, Khalil Sima'an, et al. 2003. *Data-oriented parsing*. CSLI Publications.

Serhiy Bykh and Detmar Meurers. 2014. Exploring syntactic features for native language identification: A variationist perspective on feature encoding and ensemble optimization. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. pages 1962–1973.

Joshua Goodman. 1998. *Parsing inside-out*. Ph.D. thesis, Harvard University.

Yang He. 2008. *A Study of Grammatical Features in Europeanized Chinese*. Commercial Press.

Hai Hu, Daniel Dakota, and Sandra Kübler. 2017. Non-deterministic segmentation for chinese lattice parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. pages 316–324.

Xianyao Hu. 2010. A corpus-based multi-dimensional analysis of the stylistic features of translated chinese (in chinese). *Foreign Language Teaching and Research* 42(6):451–458.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *CICLing*. Springer, volume 6008, pages 503–511.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the ACL: HLT*. pages 1318–1326.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. *Computational Linguistics* 38(4):799–825.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. pages 2231–2234.

Charles Li and Sandra Thompson. 1981. *A functional reference grammar of Mandarin Chinese*. Berkeley: University of California Press.

Can Liu, Sandra Kübler, and Ning Yu. 2014. Feature selection for highly skewed sentiment analysis tasks. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*. Dublin, Ireland, pages 2–11.

Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, et al. 2016. IUCL at SemEval-2016 task 6: An ensemble model for stance detection in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*. pages 394–400.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David Mc-Closky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. pages 55–60.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. *Religion* 17:3–4.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association* 58(302):275–309.

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

Matt Post and Shane Bergsma. 2013. Explicit and implicit syntactic features for text classification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 866–872.

Matt Post and Daniel Gildea. 2009. Bayesian learning of a tree substitution grammar. In *Proceedings of the ACL-IJCNLP 2009 Conference*. pages 45–48.

Federico Sangati and Willem Zuidema. 2011. Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 84–95.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the ACL*. pages 193–197.

Gideon Toury. 1979. Interlanguage and its manifestations in translation. *Meta: Journal des traducteurs/Meta: Translators' Journal* 24(2):223–231.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities* 30(1):98–118.

Zhiguo Wang and Nianwen Xue. 2014. Joint POS tagging and transition-based constituent parsing in Chinese with non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. volume 1, pages 733–742.

Zhiguo Wang, Chengqing Zong, and Nianwen Xue. 2013. A lattice-based framework for joint Chinese word segmentation, POS tagging and parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 623–627.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 1600–1610.

Richard Xiao and Xianyao Hu. 2015. *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Springer.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11(2):207–238.

Huaping Zhang, Hongkui Yu, Deyi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. pages 184–187.