# Detecting Syntactic Features of Translated Chinese[1]

Hai Hu, Wen Li, Sandra Kübler
Indiana University

2nd Workshop on Stylistic Variation at NAACL-HLT
June 2018

[1] https://www.youtube.com/watch?v=Q1WgnwIvVZE

# Detecting Translationese

## Translated texts differ from original texts: translationese

- prefix *mono-* more frequent in Greek-to-English translations
- "modal verb + infinitive + past participle" more frequent in translated English (e.g. *must be taken*)

# Detecting Translationese

## Translated texts differ from original texts: translationese

- prefix *mono-* more frequent in Greek-to-English translations
- "modal verb + infinitive + past participle" more frequent in translated English (e.g. *must be taken*)

## ML-based classifiers can distinguish them

- SVMs: 98% accuracy w/ POS trigrams (Volansky et al. 2013), but:
- mostly lexical or shallow syntactic features
- few studies in Chinese

# This Work

A classification task: translated vs. original

- in Chinese
- using syntactic features
  $\rightarrow$ capture deeper translationese
- interpret the features

# Dataset

Genre-balanced corpus (4 genres; 15 sub-genres) (Xiao and Hu 2015)

- LCMC: Lancaster Corpus of Mandarin Chinese
- ZCTC: Zhejiang Corpus of Translated Chinese

# Dataset

Genre-balanced corpus (4 genres; 15 sub-genres) (Xiao and Hu 2015)

- LCMC: Lancaster Corpus of Mandarin Chinese
- ZCTC: Zhejiang Corpus of Translated Chinese

| # texts | news | prose | science | fiction | total |
|---|---|---|---|---|---|
| LCMC: original | 88 | 206 | 80 | 111 | 485 |
| ZCTC: translation | 88 | 206 | 80 | 111 | 485 |

# Dataset

Genre-balanced corpus (4 genres; 15 sub-genres) (Xiao and Hu 2015)

- LCMC: Lancaster Corpus of Mandarin Chinese
- ZCTC: Zhejiang Corpus of Translated Chinese

| # texts | news | prose | science | fiction | total |
|---|---|---|---|---|---|
| LCMC: original | 88 | 206 | 80 | 111 | 485 |
| ZCTC: translation | 88 | 206 | 80 | 111 | 485 |

- Each text: around 2000 words
- Segmented and POS-tagged (Zhang et al. 2003)
- We removed urls, normalized punctuations, etc.

# Features

3 types:

- *n-gram features*: upper bound
- *Constituency treelets*: CFG rules, CFG subtrees
- *Dependency graphs*: variants of dependency graphs

# *n*-gram features

With lexical information:

- character 1-3 grams
- word 1-3 grams

Without lexical information:

- POS 1-3 grams

# Syntactic features

Parses from Stanford CoreNLP (Manning et al. 2014)

## CFGR

Count of CFG rules:

`NP` $\rightarrow$ `DP NP`

`IP` $\rightarrow$ `NP VP PU`

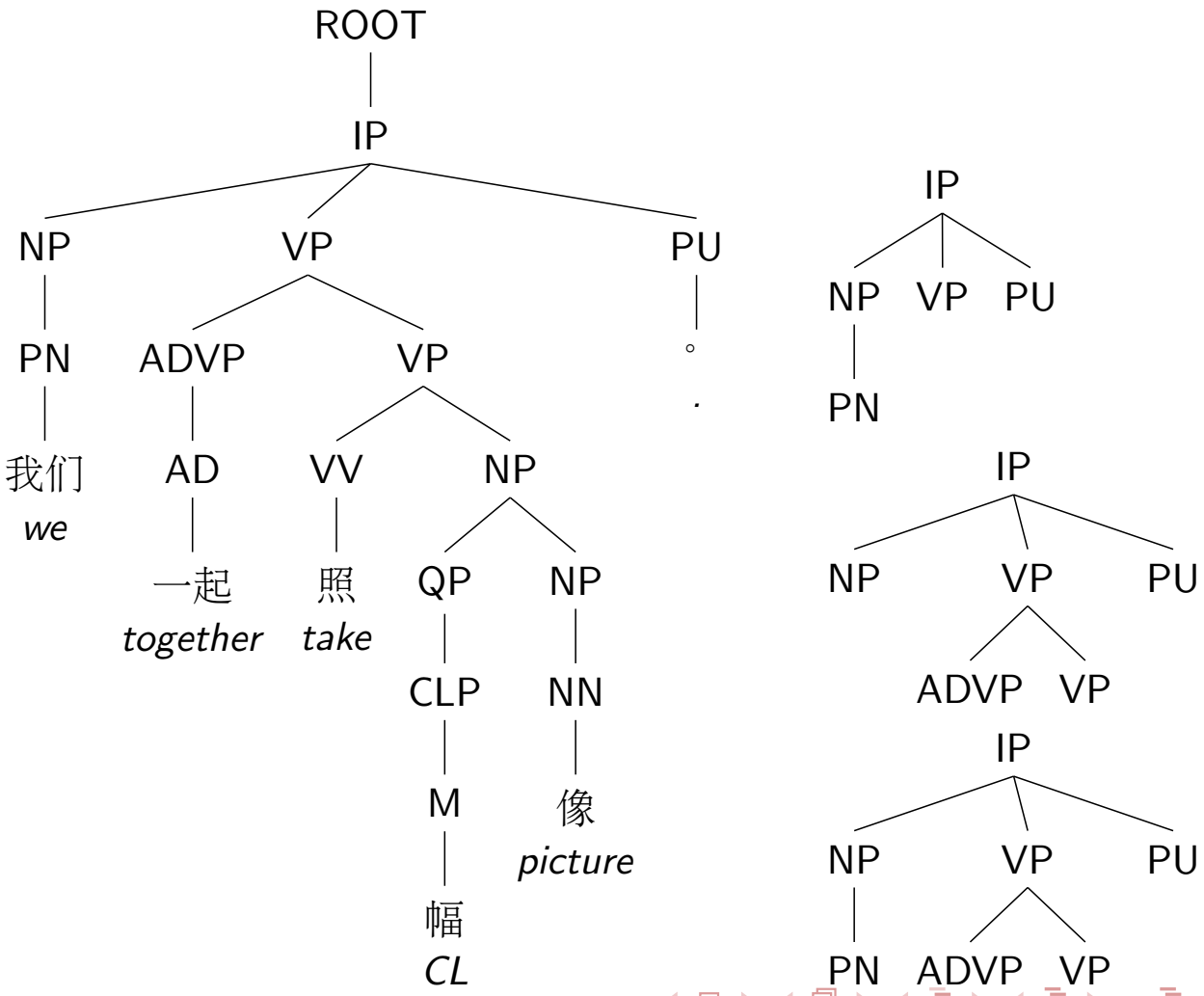
etc.

## Subtrees

Part of **unlexicalized** constituent tree of depth 2/3, following data-oriented parsing paradigm (Bod et al. 2003).

# Syntactic features

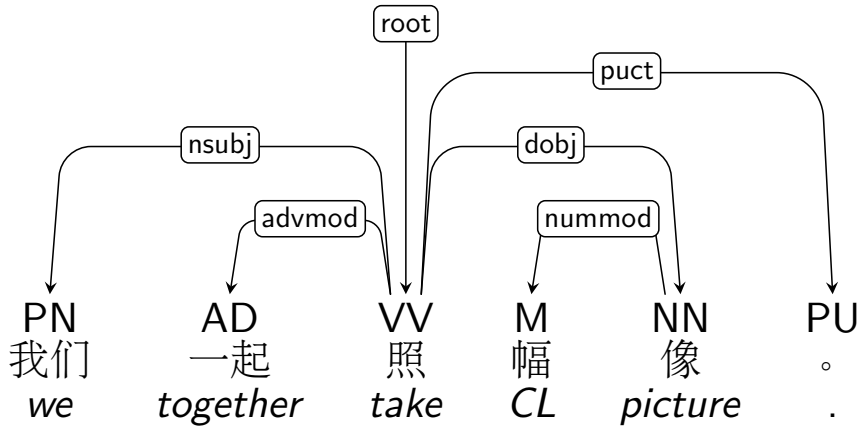
Left: Example tree Right: All subtrees of depth 2 with IP as root



# Syntactic features

## Variants of Dependency Graphs

- **depTriple**. [POS_head, relation, POS_dependent], e.g., [VV, nsubj, PN]
- **depPOS**. [POS_head, POS_dependent], e.g., [VV, PN].
- **depLabel**. Only the dependency relation, e.g., [nsubj].
- **depTripleFuncLex**. Same as depTriple; replace POS with lexical item when it's function word. e.g. [VV, nsubj, 我们(we)]



# Classifier and Feature Selection

- **Support Vector Machines** from scikit-learn (Pedregosa et al. 2011) and
- **Information Gain** for feature selection (Liu et al. 2016; Wong and Dras 2011).

Different numbers of features, ranging from 100 to 50, 000, reporting best results.

# Results

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

# Results for individual subtrees

| 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 |

Only CFG rules headed by NP (or VP, IP): fairly accurate!

# Linguistic Interpretation of Features

## Top 20 CFGR features

| 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 |

# Linguistic Interpretation of Features

**Top 20 CFGR features**

More prominent in translated Chinese:

NP $\rightarrow$ PN: pronouns

NP $\rightarrow$ DP NP

DP $\rightarrow$ DT: "该" (this), "这些" (these), "那些" (those)

PRN $\rightarrow$ PU NP PU: "加州大学洛杉矶分校(UCLA)"

NP $\rightarrow$ DNP NP

DNP $\rightarrow$ NP DEG: $NP_1$ as the modifier of $NP_2$

# Linguistic Interpretation of Features

**Top 20 CFGR features**

More prominent in translated Chinese:

NP → PN: pronouns

NP → DP NP

DP → DT: "该" (this), "这些" (these), "那些" (those)

PRN → PU NP PU: "加州大学洛杉矶分校(UCLA)"

NP → DNP NP

DNP → NP DEG: $NP_1$ as the modifier of $NP_2$

- (NP (DNP (NP 美国) (DEG 的)) (NP 政治)).
  Gloss: "US DEG politics", i.e. US politics
- (NP (DNP (NP 舆论) (DEG 的)) (NP 谴责)).
  Gloss: "media DEG criticism", i.e. criticism from the media
- (NP (DNP (NP 脑) (DEG 的)) (NP 供血)).
  Gloss: "brain DEG blood supply", i.e. cerebral circulation

# Linguistic Interpretation of Features





**Top 20 CFGR features**

More prominent in translated Chinese:

NP → PN: pronouns

NP → DP NP

DP → DT: "该" (this), "这些" (these), "那些" (those)

PRN → PU NP PU: "加州大学洛杉矶分校(UCLA)"

NP → DNP NP

DNP → NP DEG: $NP_1$ as the modifier of $NP_2$

- DEG 的 is optional in all three cases, but sometimes it is required.
- Translaters seem to make the safer decision by always using DEG 的 after the NP modifiers.

# Linguistic Interpretation of Features

## CFGR features headed by NP

| 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 |

# Linguistic Interpretation of Features

**CFGR features headed by NP**

- in original:
NP $\rightarrow$ NN **PU** NN
e.g. "全院 医生、护士 最先挖掘的..."
*doctors, nurses from the hospital first dug out...*

- in translated:
NP $\rightarrow$ NR **CC** NR
NP $\rightarrow$ NP **CC** NP
e.g. "对经济和股市非常敏感"
*very sensitive to the economy* **and** *the stock market.*

**"、": Chinese specific punctuation**

# Linguistic Interpretation of Features

**Pronouns (PN)**:

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Previous studies have identified the overuse of pronouns in translation (He 2008; Xiao and Hu 2015).

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Previous studies have identified the overuse of pronouns in translation (He 2008; Xiao and Hu 2015).

But subject pronouns? Object pronouns?

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Previous studies have identified the overuse of pronouns in translation (He 2008; Xiao and Hu 2015).

But subject pronouns? Object pronouns?
Can be explored w/ syntactic structures.

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Top subtree (depth=2) features involving pronouns (PN)

| 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 |

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Mostly subject pronouns.

Only 1 object of preposition.

Only 1 object of verb, but:
(VP VV (NP **PN**) IP) = "make + pronoun + V."
e.g. "让 他们 懂得 ..." (*make* **them** *understand ...*)
**them** = object of "make" + subject of "understand".

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Top depTripleFuncLex features involving pronouns (PN)

| Rank | Feature | Predicts | Gloss |
|---|---|---|---|
| 5.4 | VV_**NSUBJ**_我 | translated | I |
| 10.0 | VV_**NSUBJ**_他 | translated | he |
| 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 |
| 35.6 | VV_**NSUBJ**_你 | translated | you |
| 47.2 | VV_**NSUBJ**_它 | translated | it |
| 191.0 | VV_DOBJ_它 | translated | it |

# Linguistic Interpretation of Features

**Pronouns (PN)**:

Top depTripleFuncLex features involving pronouns (PN)

| Rank | Feature | Predicts | Gloss |
|---|---|---|---|
| 5.4 | VV_**NSUBJ**_我 | translated | I |
| 10.0 | VV_**NSUBJ**_他 | translated | he |
| 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 |
| 35.6 | VV_**NSUBJ**_你 | translated | you |
| 47.2 | VV_**NSUBJ**_它 | translated | it |
| 191.0 | VV_DOBJ_它 | translated | it |

The first "DOBJ" feature ranks 191th.

# So what?

- Confirms previous results showing more pronouns in translated texts (He 2008)
- More pronouns in subj. rather than obj. position
- Chinese: pro-drop, English: non-pro-drop
- More importantly, pro-drop seems to happen more often in subject position in Chinese (c.f. Li and Thompson 1981)

# Conclusion and Future Work

## Conclusion

- Syntactic features are good at detecting translations (90%+)
- Linguistically meaningful features are easily interpretable
- Interesting results concerning NPs and pronouns

## Implication

Syntactic features can be applied to study styles of translationese, and allow for analysis of deeper structures.

## Future Work

- More feature analysis
- Theory motivated features to (dis-)confirm previous hypotheses

Thanks!
Questions and comments?

# lexical features

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.

# Related Work

## Detecting Translationese w/ ML Classifiers

**Baroni and Bernardini (2005)**: translated vs. ori. Italian.
Features: wordform, lemma, pos, mixed.
SVMs: 85.2% F-measure $>$ human judgment.

**Koppel and Ordan (2011)**: "Englishes" translated from It., Fr., Es., De., Fin. can be distinguished.
Features: counts of function words $\rightarrow$ 92.7% accuracy

**Volansky et al. (2013)**: translated vs. ori. English.
Features: 33 feature sets based on 4 translation universals.
TTR $\rightarrow$ 76%; mean word length $\rightarrow$ 66%.
Character ngrams, contextual function words $\rightarrow$ 100%.

# Related Work

## Detecting Translationese for Chinese

Europeanized/translational Chinese has been studied for decades, but no text classification task has been done to our knowledge.

- First discussed in Wang (1944); case study of a novel (Kubler 1985)
- Corpus study (He 2008); compared frequencies of mostly lexical features (Xiao and Hu 2015)
- More pronouns, passives, connectives and certain affixes in translated Chinese

# Related Work

## Syntactic Features

Mostly **lexical features** in translation studies.
However, in Native Language Identification, syntactic features are popular:

- CFG rules + *n*-grams improve accuracy (Bykh and Meurers 2014; Wong and Dras 2011)
- TSG rules are also helpful (Post and Bergsma 2013; Swanson and Charniak 2012)

# References I


Baroni, M. and S. 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.

Bod, R., R. Scha, K. Sima'an, et al. (2003). *Data-oriented parsing*. CSLI Publications.

Bykh, S. and D. 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*, pp. 1962–1973.

He, Y. (2008). *A Study of Grammatical Features in Europeanized Chinese*. Commercial Press.

# References II


Koppel, M. and N. Ordan (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the ACL: HLT*, pp. 1318–1326.

Kubler, C. C. (1985). *A study of Europeanized grammar in modern written Chinese*, Volume 10. Student Book Company.

Li, C. and S. Thompson (1981). *A functional reference grammar of Mandarin Chinese*. Berkeley: University of California Press.

Liu, C., W. Li, B. Demarest, Y. Chen, S. Couture, D. Dakota, N. Haduong, N. Kaufman, A. Lamont, M. 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)*, pp. 394–400.

# References III


Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

Pedregosa, F., 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.

Post, M. and S. 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, pp. 866–872.

# References IV


Swanson, B. and E. Charniak (2012). Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the ACL*, pp. 193–197.

Volansky, V., N. Ordan, and S. Wintner (2013). On the features of translationese. *Digital Scholarship in the Humanities 30*(1), 98–118.

Wang, L. (1944). Theory of chinese grammar.

Wong, S.-M. J. and M. Dras (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1600–1610.

Xiao, R. and X. Hu (2015). *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Springer.

# References V

Zhang, H., H. Yu, D. Xiong, and Q. Liu (2003). HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 184–187.