Computational Modelling of Classifier Choice in Mandarin

Guanyi Chen

g.chen@uu.nl → *g.chen@ccnu.edu.cn*

Central China Normal University

(Joint work with Jani Järnfors, Amber de Bruijn, Kees van Deemter, Muyun Yang, Tiejun Zhao, Rint Sybesma)

What is classifier?

(1) a. san di you
'three drops of oil'
b. san zhi songshu
three CL squirrel
'three squirrels'
c. *san songshu
'three squirrel'

A short history of this project

1. Jani's Master thesis about using BERT for Classifier Choice in Mandarin $\rightarrow$ a short paper in INLG
2. A small chapter in my thesis about a speaker experiment
3. Amber's Master thesis about two reader experiments
4. We now attempt to summarise this whole project into a journal paper for CL (?)

A short history of this project

1. Jani's Master thesis about using BERT for Classifier Choice in Mandarin → a short paper in INLG
2. A small chapter in my thesis about a speaker experiment
3. Amber's Master thesis about two reader experiments
4. We now attempt to summarise this whole project into a journal paper for CL (?)

A short history of this project

1. Jani's Master thesis about using BERT for Classifier Choice in Mandarin $\rightarrow$ a short paper in INLG
2. A small chapter in my thesis about a speaker experiment
3. Amber's Master thesis about two reader experiments
4. We now attempt to summarise this whole project into a journal paper for CL (?)

A short history of this project

1. Jani's Master thesis about using BERT for Classifier Choice in Mandarin $\rightarrow$ a short paper in INLG
2. A small chapter in my thesis about a speaker experiment
3. Amber's Master thesis about two reader experiments
4. We now attempt to summarise this whole project into a journal paper for CL (?)

The General Research Question

We built **Computational Models** as well as **Human Experiments** to investigate the question of

What classifier suits a particular position in Mandarin discourse?

(2) yi ⟨CL⟩ jingcai de ⟨h⟩qiusai⟨/h⟩
'a wonderful ball game'

Particularly, the issues include:

- What algorithms model classifier choice most adequately?
- What factors influence classifier choice?
- How much does the choice of classifier matter for readers?

The Classifier Choice is not trivial

Most classifier choice model are rule-based. BUT ...

(3)
a. yi ge diannao / yi tai diannao
   'a computer'
b. yi ge qiu / yi chang qiu
   'a ball' / 'a (ball) game'
c. yi ge laoshi / yi wei laoshi
   'a teacher'
d. yi ge ren / yi qun ren
   'a person / a bunch of people'
e. yi bei kafei / yi ting kafei
   'a cup/can of coffee'

Study 1: Construct and Evaluate Computational Models

- Input:

(4) yi ⟨CL⟩ jingcai de ⟨h⟩qiusai⟨/h⟩
'a wonderful ball game'

- Data: ChineseClassifierDataset (CCD)
- Models: Rule-based, LSTM, BERT, and BERT as an MLM
- Expectations:
  1. BERT performs the best;
  2. BERT is not good at handling classifiers that add information, e.g., plurality, politeness, and measure.

Study 1: Corpus Evaluation Results

		Macro-averaged			Weighted-averaged		
Model	Accuracy	Precision	Recall	F1	Precision	Recall	F1
`Rule`	61.73	33.24	21.01	23.66	57.85	61.73	58.31
`LSTM`	<u>73.86</u>	45.67	32.24	36.07	72.39	<u>73.86</u>	<u>72.69</u>
`MLM`	62.22	<u>51.91</u>	<u>33.40</u>	<u>37.68</u>	<u>77.28</u>	62.23	68.21
`BERT`	**81.71**	**52.86**	**38.10**	**40.77**	**80.70**	**81.71**	**80.77**

- BERT performs the best;
- BERT has significantly lower accuracy in predicting classifiers that add information;
- MLM predicted more identical classifiers than other models and is good at rarely seen classifiers.

Study 1: Human Evaluation

Model	Fluency	Clarity
Corpus	4.96 (2.01)	5.10 (1.99)
Rule	4.41 (2.17)	4.56 (2.16)
LSTM	4.68 (2.09)	4.81 (2.09)
BERT	4.92 (2.02)	5.02 (2.02)

- We compared models using **Wilcoxon's Signed-Rank test with Bonferroni Correction** and reported both p-values and effect sizes.
  - Corpus and BERT outperformed Rule and LSTM in terms of fluency and clarity;
  - No clear difference between BERT and Corpus.
- Fluency and Clarity scores are highly correlated (WHY?; Spearman's Correlation)
- Corpus evaluation and Human evaluation seem to be consistent (**Mood's Median Test**), but the conclusions are slightly different.

Study 2: How well can Human Speakers Choose Classifiers?

- Though we expected the task setting could mimic the environment when humans select classifiers, they have major differences;
- We asked human participants to do the same task to shed light on how good our models are compared to humans.
- We conducted two speaker experiments:
  1. Randomly sampled data, almost all of which are frequently used classifiers;
  2. Breath-first sampled data, where we first sampled 100 distinct classifiers and sampled data that use these classifiers accordingly.

Study 2: Results

	Accuracy (SD)	Percent Agreement
Experiment A	70.97 (2.28)	67.92
Experiment B	41.82 (2.16)	47.22

- Both LSTM and BERT perform better than Humans
- But for infrequent classifiers, Humans are slightly better (compared to the macro-averaged Recall of BERT)
- Are we right that it is impossible to compute Kappa in this case?

Study 3: How does the Choice of True Classifiers Matter to Human Readers?

- In many cases, different uses of classifiers result in similar meanings, especially “true” classifiers (i.e., not measure words)
- esp. the choice between the general purpose classifier and the specific classifier

(5) yi ge diannao / yi tai diannao
'a computer'

- Maybe for readers, these choices do not matter.

Therefore ...

- Focusing on true classifiers, we conducted a larger-scale reader experiment (compared to the human evaluation)
- We compared BERT and Rule-based models to Corpus as well as GE (which always selects the general purpose classifier ge)
- Similar to study 2, we used two sets of data: a randomly sampled one and a breath-first sampled one.

Study 3: Results

- Corpus, BERT and RULE are all significantly better than GE;
- BERT and Corpus are still indistinguishable;
- BERT and RULE are indistinguishable in terms of fluency on the use of fluently used classifiers. BERT is the clear winner in terms of clarity and infrequently used classifiers.
- BERT and Corpus were rated with no significant difference on frequently used classifiers and frequently used classifiers.
  - Human readers have higher tolerance on incorrect choice of infrequent classifiers;
  - OR infrequent classifiers often have equally good frequent alternatives;
  - Though we haven't tried LLMs, BERT is perfect enough for this classifier choice.

Something Personal 1

- National Language Resources Monitoring and Research Center
  - w/ Institute of Linguistics
  - w/ Many other U. in China, e.g., Tsinghua University, Beijing Foreign Language University, and Beijing Language and Culture University
- Laboratory of Artificial Intelligence and Smart Learning
  - NLP + Education
  - w/ Faculty of Artificial Intelligence in Education

Something Personal 2: (NLG) Corpus in Chinese

- I have a small project (25k Euro) for constructing NLG corpus in Chinese;
- Chinese is not considered a low-resource language;
- BUT (seemingly) there is no gold standard NLG corpus;
- Is it still meaningful to collect a WebNLG-like corpus in Chinese?
- Any other options?

Something Personal: Three-Modality LLM Evaluation (working proposal)

- to Noah's Ark Lab of Huawei
- Vision + Language + Speech
- With proper evaluation, can we know:
  - Can representations from these three modalities be mapped into a single space?
  - What we can benefit from additionally modelling speech?