TeCS: A Dataset and Benchmark for Tense Consistency of Machine Translation

Yiming Ai, Zhiwei He, Kai Yu, Rui Wang

{aiyiming, zwhe.cs, kai.yu, wangrui12}@sjtu.edu.cn

Background

- **FR:** Mais on les avait votés lors de la dernière période de session.

  Plus-que-parfait

- **EN:** But we voted on them during the last part-session.

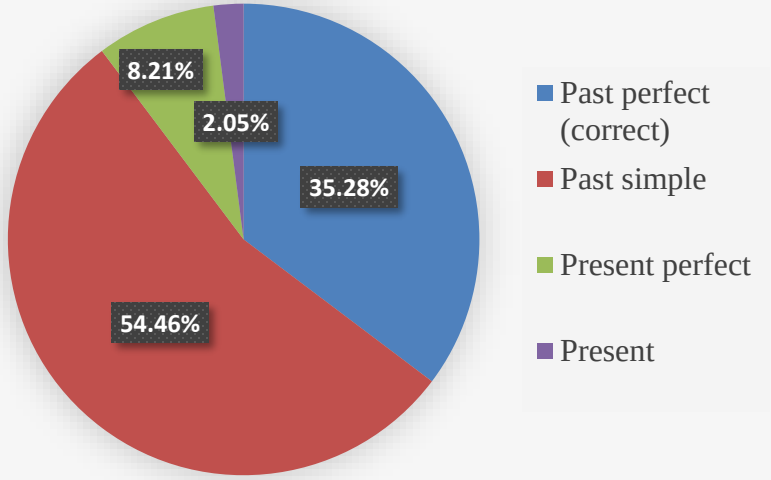
  Past simple

- **Correction:** But we had voted on them during the last part-session.

  Past perfect √

- There are several **tense consistency errors** in the common corpora, for instance, Europarl.
- **Lack of metrics** on measuring the model's mastery of tense information.

Distribution of English counterparts of French sentences in plus-que-parfait



Contributions

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

- Presentation of the construction of the **tense test set**, including its tense labels
- Proposal of a feasible and reproducible **benchmark** for measuring the tense consistency performance of NMT systems
- **Various experiments** for different baselines with the above test set and corresponding benchmark.

Annotation Rules

- Macro-temporal interval (**present**, **past** and **future** tenses) ✕ State of the action (**general**, **progressive** and **perfect** aspects)
- Progressive tense ⟹ corresponding base tense
- Add a new category -- statements containing modal verbs

French Tenses	**English Tense**	**Format**	**Example**
Imparfait, Passé composé, Passé simple, Passé récent	Past simple / progressive	*Past*	That **was** the third point.
Présent, Future proche	Present simple / progressive	*Present*	The world **is changing**.
Future simple, Future proche	Future simple / progressive	*Future*	I **will communicate** it to the Council.
Plus-que-parfait	Past perfect	*PasPerfect*	His participation **had been notified**.
Passé composé	Present perfect	*Preperfect*	This phenomenon **has become** a major threat.
Future antérieur	Future perfect	*Futperfect*	We **will have finished** it at that time.
Subjonctif, Conditionnel	including Modal verbs	*Modal*	We **should be** less rigid.

Corpus Design

570 triples

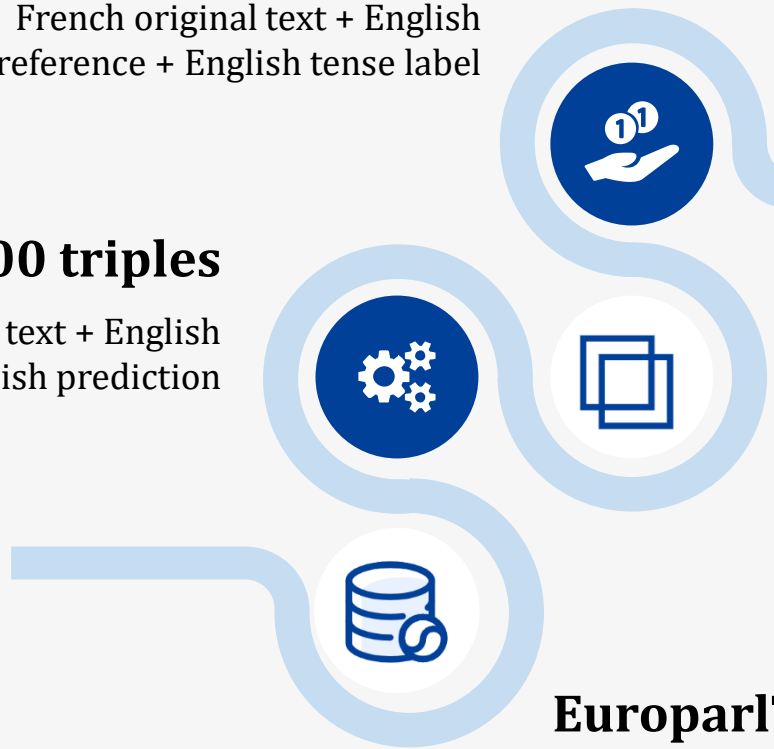
French original text + English reference + English tense label

20000 triples

French original text + English reference + English prediction

EuroparlPV

tense-rich: sentences with predicate verb structure extracted from Europarl



552 triples

French original text
+ English reference
+ English tense label

6779 groups

French original text +
English reference&tense labels +
English prediction&tense labels

EuroparlTR

201,374 tense-rich parallel French-English sentence pairs

Experiments and Conclusions

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

- Tense (Prediction) Accuracy:

$$\text{Accuracy} = \frac{N_c}{N_t}$$

- Experiments:
  - **4 architectures**: Transformer, LSTM, CNN and Bi-transformer
  - **3 test sets**: our tense set, Europarl test set and WMT15 test set
  - **2 metrics**: BLEU and COMET
  - **3 business translators**: Bing, DeepL and Google translator
- Conclusions:
  - By relying solely on the difference in BLEU scores on traditional test sets, we are **unable** to measure the tense prediction ability of the systems
  - Our tense set can capture the **tense consistency performance.**
  - To measure the tense consistency performance across different architectures, we should **focus more on tense accuracy rather than BLEU scores only.**

More Details?

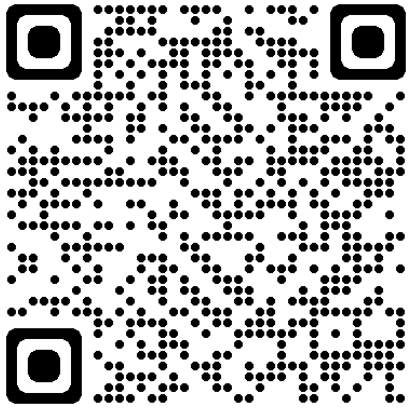
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

- Why adopt such annotation rules?
- More details of the corpus design?
- What are the characteristics of the corpus?
- What specific experiments did we implement?
- How about the results of the experiments?
- How to use our metrics for your own experiments?

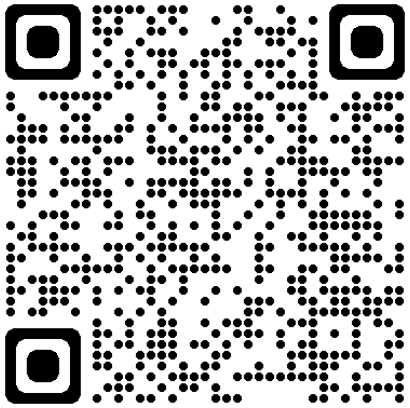
If you would like to know more:

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Paper



GitHub repo



SHANGHAI JIAO TONG UNIVERSITY

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Thank you