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

Paper GitHub repo

Yiming Ai, Zhiwei He, Kai Yu, Rui Wang

Shanghai Jiao Tong University {aiyiming, zwhe.cs, kai.yu, wangrui12}@sjtu.edu.cn

SHANGHAI JIAO TONG UNIVERSITY

1. Summary

- Background
  - 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.
- Contributions
  - 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.

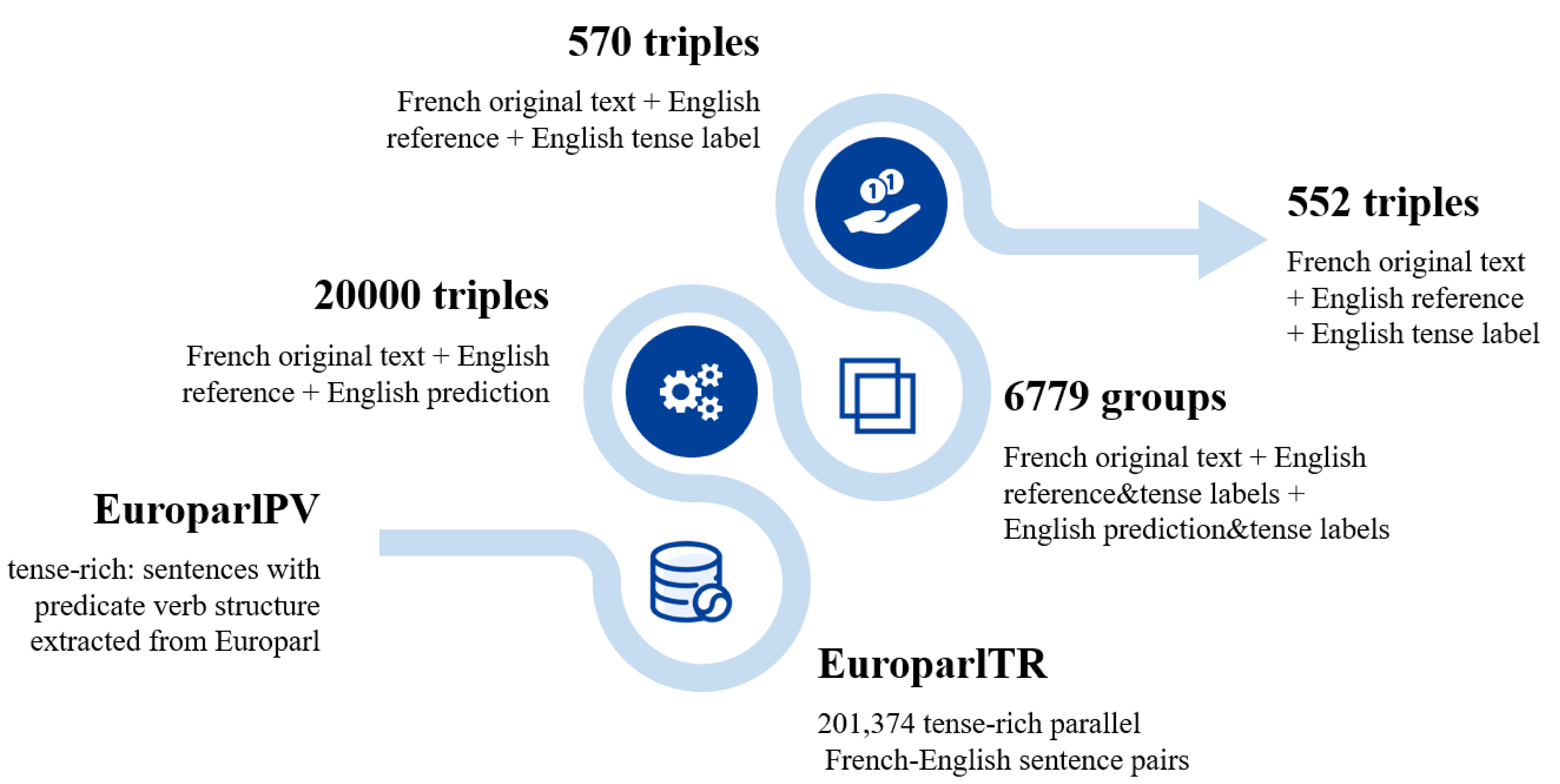
2. Annotation Rules

- **Macro-temporal interval** (present, past and future tenses) * **State of the action** (general, progressive and perfect aspects)
- As there is **no progressive tense** in French, we do not distinguish the progressive tense in English but rather merge the progressive tense into its corresponding base tense.
- Considering the moods, we add another category – **statements containing *modal* verbs** that correspond to the French *subjonctif* and *conditionnel* tenses.

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.

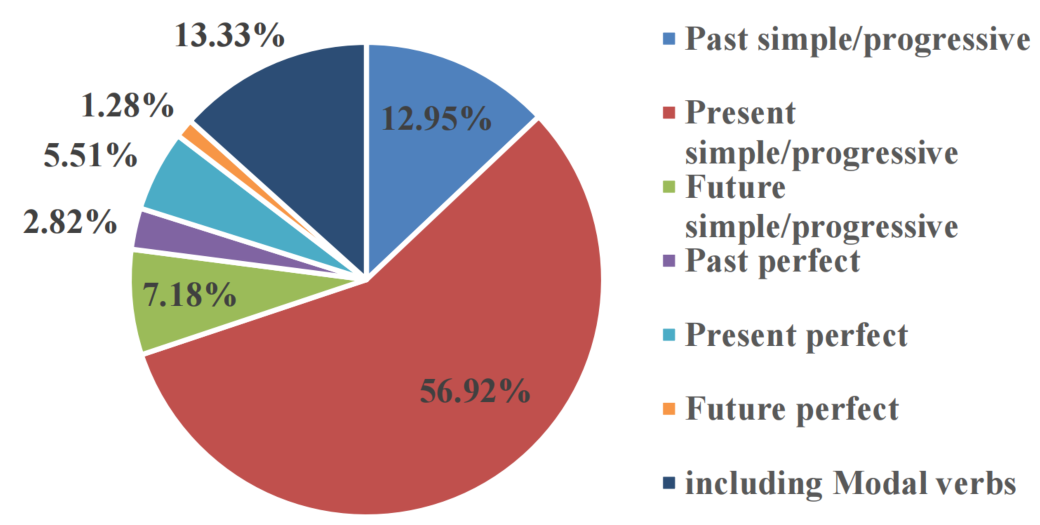
3. Corpus Design

- Tense-rich Europarl, namely EuroparlPV, stems from Loaciaga et al.'s article.
- After data cleaning, we obtain the EuroparlTR.
- We randomly divided EuroparlTR into a training set, a validation set and a test set in the ratio of 8:1:1, and trained a transformer baseline based on this using fairseq with a BLEU value of 33.41.
- With automatic tense annotation, we filtered 6,779 parallel French-English sentence triples with different tense labels for English originals and predictions.
- We manually screened out the representative error-prone French-English sentence triples.
- Human check: two other reviewers at CEFR C1 level, reviewed the tense test set for semantic and tense correspondence, and the tense labels marked by the automatic annotation code.



4. Corpus Characteristics

- **Tense distribution**. The corpus consists of 780 tense structures in 552 sentences, and the distribution of tense classifications is shown in the following table.
- **Elimination of gender effect**. We controlled for the gender variable of French by defaulting all pronouns, which do not indicate explicitly their genders, as masculine.



5. Benchmark

To measure the tense consistency performance of different systems, we introduce a benchmark called **tense (prediction) accuracy**, as shown below:

$$\text{Accuracy} = \frac{N_c}{N_t} \tag{1}$$

where N_c is the number of predicted utterances with the same tense as its reference and N_t is the total number of utterances in the tense set.

6. Experiments

Evaluation Summarization based on 3 test sets

System	Tense set		Europarl testset		WMT15 testset		Tense Accuracy
	BLEU	COMET	BLEU	COMET	BLEU	COMET	
Transformer (tense-rich)	47.71	0.631	27.38	0.269	14.17	-0.429	66.30%
Transformer (tense-poor)	43.24	0.588	27.28	0.264	14.68	-0.444	58.33%
LSTM (tense-rich)	44.21	0.558	25.53	0.126	12.04	-0.590	67.75%
LSTM (tense-poor)	41.92	0.483	26.17	0.147	12.27	-0.598	58.70%
CNN (tense-rich)	47.10	0.567	26.83	0.147	15.30	-0.512	68.48%
CNN (tense-poor)	43.23	0.502	26.95	0.144	14.96	-0.525	57.97%
Bi-Transformer (tense-rich)	47.10	0.632	28.17	0.295	14.72	-0.392	64.13%
Bi-Transformer (tense-poor)	43.87	0.578	28.30	0.298	14.39	-0.428	55.25%
Bing Translator	61.72	0.895	-	-	-	-	77.36%
DeepL Translator	59.50	0.904	-	-	-	-	79.02%
Google Translator	57.00	0.878	-	-	-	-	81.70%

Training Process and Results

We separately extract 100,000 parallel utterances from EuroparlTR and Europarl as tense-rich and tense-poor train sets. We then trained four pairs of French-English systems with different architectures, differing only in the train set. Results are as follows:

- 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.