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

Yiming Ai, Zhiwei He, Kai Yu, and Rui Wang*
Shanghai Jiao Tong University
{aiyiming, zwhe.cs, kai.yu, wangrui12}@sjtu.edu.cn

Abstract

Tense inconsistency frequently occurs in machine translation. However, there are few criteria to assess the model's mastery of tense prediction from a linguistic perspective. In this paper, we present a parallel tense test set, containing French-English 552 utterances[1]. We also introduce a corresponding benchmark, tense prediction accuracy. With the tense test set and the benchmark, researchers are able to measure the tense consistency performance of machine translation systems for the first time.

1 Introduction

Translation tools are often found in a variety of social situations to enable cross-linguistic communication. Tenses are used to express time relative to the moment of speaking. Human translators frequently pay close attention to tense correspondence (Gagne and Wilton-Godberfforde, 2020). Similarly, machine translation (MT) systems are supposed to maintain temporal consistency between the original text and the predicted text to avoid misunderstandings by users. However, accurately keeping the tense consistency is undoubtedly difficult. Taking French-English (one of the most classic language pairs for MT) as an example in Table 1, the original text is in *plus-que-parfait de l'indicatif* of French, corresponding to the *past perfect* tense in English, while the English prediction provided by Google Translator is in the *past simple* tense.

Sentence	Tense
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*

Table 1: An example of tense corrspondence in machine translation

In fact, this is not an isolated case. You can also find several examples in Appendix B. Besides. the translation mechanics may not the only reason leading to tense inconsistency. The corpora matter as well. For example, we have extracted 20,000 pairs English-French parellel sentences from the widely used dataset Europarl (Koehn, 2005), and we have observed all groups of parallel utterances where the original French texts are in the *plus-que-parfait de l'indicatif* tense, examining the tenses of their English counterparts. As a sentence may include several tenses, there are 195 occurences of *plus-que-parfait* tense in total. Among them, only 35.28% English sentences are in the correct *past perfect* tense, as shown in Table 2. Although, compared to other tense correspondences, the pair of *plus-que-parfait* and *past-perfect* is prone to error in datasets and there are only 0.94% of sentences in Europarl are in plus-que-parfait, we cannot easily ignore this issue. Like Europarl, tense correspondences are generally credible but unreasonable for certain tenses in several common datasets.

Tense of Counterpart	Proportion
Past perfect (correct)	35.28%
Past simple	54.46%
Present perfect	8.21%
Present	2.05%

Table 2: Preliminary statistics of translation tense

In addition to the train set, the difficulty of remaining tense consistency also stems from the lack of metrics on measuring the model's mastery of tense information. The research of Marie et al.

[*] Corresponding author

[1] The following updates will be shown at: https://github.com/rutilel/TeCS-A-Dataset-and-Benchmark-for-Tense-Consistency.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1930–1941
July 9-14, 2023 ©2023 Association for Computational Linguistics

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.

Table 3: French-English tense pairs, annotation format of English tenses and corresponding example sentences *(Where the modal verb contains can, may, shall, must, could, might, should and would.)*

(2021) shows that 98.8% of *ACL papers[2] in the field of MT from 2010 to 2020 used BLEU (Papineni et al., 2002) scores to evaluate their models. However, the reliability of BLEU has been questioned in the era of neural machine translation (NMT) as its variants only assess surface linguistic features (Shterionov et al., 2018), and many studies have shown that BLEU has difficulty in portraying the degree of semantic information mastered by the model, i.e. its score does not necessarily improve when more semantic information is mastered (Mathur et al., 2020; He et al., 2023), not to mention specific tense information. We have also applied BLEU to measure various baselines on our tense test set in Section 4, and the results explicitly support the above statement. In addition, reviewing the evaluation criteria related to MT tasks over the past ten years, we are surprised to find that there are no criteria to assess the model's mastery of tense prediction from a linguistic perspective.

Therefore, our paper is devoted to the study of NMT based on semantic understanding in terms of tense. We construct a tense parallel corpus test set consisting of 552 pairs of tense-rich, error-prone parallel utterances for NMT systems, and then propose a new task for evaluating the effectiveness of model translations from the perspective of tense consistency. This paper makes three contributions: (1) the presentation of the construction of the tense test set, including its tense labels; (2) the proposal of a feasible and reproducible benchmark for measuring the tense consistency performance of NMT systems; and (3) the various experiments for different baselines with the above test set and corresponding benchmark.

2 Annotation Rules and Tools

As the first work of the MT tense study, we choose English-French, one of the most classic language pairs of MT, to construct the dataset[3].

TENSE, the dominant topic of our research, is a combination of tense and aspect. In the modern grammar system of English, "a tense system is a system associated with the verb where the basic contrasts in meaning have to do with the location in time of the situation, or the part of it under consideration" (Huddleston et al., 2021). The modern grammatical system divides tense into present and preterit based on the inflections added to the end of verbs, and the aspect into perfective and progressive on the state where an action is (Kamp, 1991). While this tense classification system is too crude for daily life, we therefore apply the following classification methods. On the one hand, we classify the tenses according to the macro-temporal interval of the action into three major time intervals, namely present, past and future tenses; on the other hand, we classify the tenses according to the state of the action into general, progressive and perfect aspects. Hence, 9 kinds of tenses are born through combining the three tenses and the three aspects.

French and English belong to the same Indo-European language family and share many similarities in various respects. The main difference is that in French there is another grammatical point called *mode*, part of which is like the *aspect* in English. In terms of tenses, we will generally discuss the tenses in the indicative mode of French and will describe the others later in this section. In the following, if there is no mode qualifier before a tense, it is by default in the indicative mode. Careful identification and comparison of the subdivided tenses in the three main tense intervals, English and French, reveals a very similar usage of the tenses, as sum-

[2] The papers only includes *ACL main conferences, namely ACL, NAACL, EACL, EMNLP, CoNLL, and AACL.

[3] Please refer to the Limitations for more details.

marised in Table 3. 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, e.g. the present perfect progressive tense into the category of the present perfect tense.

When discussing tenses from a semantic point of view, the modes also need to be taken into account. The grammatical correlations between French and English modes are quite complicated. Considering the corresponding grammatical expressions of 2 modes strongly related to tense, *conditionnel* and *subjonctif*, in French rely on the usage of modal verbs, we introduce *modal verbs* to simplify the distinguishment of the modes.

Based on these grammatical rules, we merge the nine common tenses in English into seven categories that correspond reasonably and rigorously to French, namely the 6 tense categories of **past/present/future + simple/perfect** and statements containing *modal* verbs that correspond to the French *subjonctif* and *conditionnel* tenses. We construct an automatic annotation method based on the spaCy package (Honnibal et al., 2020). First, we label the grammatical components of each word in the sentence based on the spaCy package, and then we define and compare the grammatical structures of the verb phrases with the structures of each tense classification to derive the sentence tense labels. During this process, to simplify the annotation process and better correspond with French *futur proche* tense, we classify the expression '*be going to do*', grammatically in Future tense, into the Present tense, just like expressions '*be about to do*' and '*be + verb progressive*', whose stucture are in *Present* tense but the real meaning is about the close future. Also, a sentence may have several tense structures, in this case, the tense label consists several tenses. For example, the label of the sentence '*So it is in that spirit that we have made this change.*' is '*Present+PrePerfect*'.

3 Corpus Design and Characteristics

3.1 Corpus Design

We choose the tense-rich Europarl, namely EuroparlPV, processed by Loáiciga et al. (2014) as the source corpus, for it contains all the sentences with predicate verb structures in the original Europarl dataset (Koehn, 2005). First, we cleaned the source corpus, including deleting sentences without counterparts, English sentences in the French part and vice versa. After this, we obtain 201,374 tense-rich parallel French-English sentence pairs, namely EuroparlTR. We randomly divided them 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 (Ott et al., 2019) with a BLEU value of 33.41. Then we compared a total of 20,000 parallel sentences' triples (*original Europarl French text, original Europarl English text, transformer English prediction*).

Classfication	Times	Proportion
Past	101	12.95%
Present	444	56.92%
Future	56	7.18%
Past perfect	22	2.82%
Present perfect	43	5.52%
Future perfect	10	1.28%
Modal	104	13.33%

Table 4: Distribution of 780 tense structures in 552 annotated sentences of the corpus

In the construction process, with the code mentioned in Section 2, we first automatically annotated the original English text and English prediction in the 20,000 pairs of parallel utterances, given the corresponding tense labels. Then, we filtered **6,779** parallel French-English sentence triples with different tense labels for English originals and predictions. On the basis of the automatic selection, we manually screened out the representative parallel French-English sentence pairs with a certain degree of translation difficulty and a complex grammatical structure. We also corrected the reference translations that did not justify the tense or semantics. It is worth noting that the author has a level of English and French that meets the C1 standard of The Common European Framework of Reference for Languages (CEFR), representing the ability to express herself effectively and flexibly in English and French in social, academic and work situations. A total of **570** parallel pairs of statements were selected at this stage.

Following this, 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. The tense test set was further refined. The final test set contains **552** parallel French-English sentence pairs. You can see more details in Appendix D.

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%

Table 5: Experimental results of various baselines and common business translators

3.2 Corpus Characteristics

In the following paragraphs, we describe the statistical features of our corpus and the elimination of gender coordination influence.

Tense distribution. The corpus consists of 780 tense structures in 552 sentences, and the distribution of tense classifications is shown in Table 4. In the test set, sentences in present tense are the most, corresponding the situation of the reality: we use present tense most frequently and future perfect sense least frequently.

Elimination of gender effect. Unlike English, gender coordination exists in French. For example, the French sentences '*Nous nous sommes donc* ***abstenus***.' and '*Nous nous sommes donc* ***abstenues***.' both correspond to the English '*We therefore abstained.*'. That is, the MT system's ability to learn gender coordination affects its ability to recognize tense structures, which in consequence affects the maintenance of tense consistency between original French text and predicted English sentence. Therefore, to better measure the tense-predicting capability of different MT systems, rather than their ability to recognize pronominal gender, we controlled for the gender variable by defaulting all pronouns, which do not indicate explicitly their genders, as masculine. These pronouns consists of 167 *je* (I), 114 *nous* (we, us) and 28 *vous* (you).

4 Experimental Results

To measure the tense consistency performance of different systems, we introduce a benchmark called **tense (prediction) accuracy**, as shown in Eq. (1).

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

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.

To verify the validity of our tense corpus, the following approach was adopted: To begin with, $100,000$ parallel utterance pairs from the EuroparlTR (containing $201,374$ pairs) mentioned in Section 3.1 were extracted as the tense-rich train set, and $100,000$ parallel utterance pairs from the Europarl corpus (Koehn, 2005) were extracted as the tense-poor train set. There were no overlapping utterances between the latter and the former. We performed the same preprocessing procedure, including data cleaning, tokenization and BPE coding. We then trained four pairs of French-English NMT systems with different architectures based on fairseq (Ott et al., 2019), where two systems in each pair differed only in the train set. After this, we summarized the scores evaluated by SacreBLEU (Post, 2018) and COMET (Rei et al., 2020) and tense prediction accuracies of the eight systems on different test sets. We have applied three types of test sets: our tense set, the Europarl test set and the WMT15 test set. The Europarl test set contains 3,000 parallel utterance pairs drawn from the Europarl corpus, the exact same field of train set, while the WMT15 is a test set for the WMT15 (Bojar et al., 2015), deriving from data in the different field of train set. Besides, we also apply our approach to mesure the tense consistency performance of several business translators, including Bing Translator, DeepL Translator and Google Translator. The results are listed in Table 5:

1) The BLEU and COMET scores based on the Europarl set and the WMT15 set are quite similar for each system pair, which indicates that the translation capabilities of the two systems are similar in the general evaluation dimension. This suggests that 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.

2) However, there are large differences in our tense set. The tense consistency performance of systems trained on the tense-rich train set was significantly better than that of systems trained on the tense-poor train set. This indicates that our tense set can capture the tense consistency performance.

3) Further investigation of the BLEU or COMET) scores and tense prediction accuracy for each system reveals their positive correlation for the same architecture, but not across architectures. To measure the tense consistency performance across different architectures, we should focus more on tense accuracy rather than BLEU scores only.

5 Conclusion

We presented the French-English parallel tense test set and introduced the corresponding benchmark *tense prediction accuracy*, providing a brand-new approach to measure the tense consistency performance of machine translation systems. This test set firstly focuses on the tense prediction ability, posing a new dimension to improve the MT quality.

In the future, we will endeavour to generalize the test set to other languages. Considering there are statements like "the use of tense A in language X is equivalent or similar to the use of tense B in English" in grammar books of other languages(Durrell et al., 2015), even across language families(Gadalla, 2017) and human translators also apply such rules(Santos, 2016), we are confident in taking this forward.

Limitations

In this work, we focus on creating the English-French tense corpus. These two languages are among the most frequently and widely used languages in the world. In addition, they have several similarities in tenses, which are pretty helpful for research on tense consistency through machine translation. Thanks to the distinctive tense structures, the study of these two languages makes it possible to examine many common tense issues, but there are also some tense issues in other languages that are not covered by this study. For example, the implicit tense expressions in Chinese are difficult to correspond to the explicit tense expressions in English (Jun, 2020). Hence, our next step will be to extend the tense test set to other language families and even cross-language families to further study tense consistency. Also, as for future work, we will optimize both the tense annotation method and the tense prediction accuracy calculation. Besides, we did not propose a new method to improve the tense prediction accuracy. To be further, we will endeavour to improve the existing machine translation systems according to tense consistency.

Acknowledgements

Yiming, Zhiwei, and Rui are with MT-Lab, Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200204, China. Rui is supported by the General Program of National Natural Science Foundation of China (6217020129), Shanghai Pujiang Program (21PJ1406800), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Beijing Academy of Artificial Intelligence (BAAI) (No. 4), CCF-Baidu Open Fund (No. CCF-BAIDU OF2022018, and the Alibaba-AIR Program (22088682). We also thank the computational resource from the SJTU student innovation center.

Ethics Statement

Our tense test set is based on the widely used public corpus Europarl in the field of machine translation. In creating this test set, we only corrected tense and description errors of some English references and did not change the original semantics, so there are no ethical issues arising.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical*

Machine Translation, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Martin Durrell, Katrin Kohl, Gudrun Loftus, and Claudia Kaiser. 2015. *Essential German Grammar*. Routledge.

Hassan Abdel-Shafik Hassan Gadalla. 2017. *Translating tenses in Arabic-English and English-Arabic contexts*. Cambridge Scholars Publishing.

Christophe Gagne and Emilia Wilton-Godberfforde. 2020. *English-French Translation: A Practical Manual*. Routledge.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Rodney Huddleston, Geoffrey K Pullum, and Brett Reynolds. 2021. *A student's introduction to English grammar*. Cambridge University Press.

Guo Jun. 2020. Translation principles of tense problem in machine translation in process of chinese-english translation. *Solid State Technology*, 63(4):5678–5687.

Hans Kamp. 1991. Tense and aspect in english and french. *Edinburgh: DYANA*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Sharid Loáiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 674–681, Reykjavik, Iceland. European Language Resources Association (ELRA).

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Diana Santos. 2016. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems*. Brill.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Dimitar Shterionov, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32(3):217–235.

A Online Translation

- Google Translator: `https://translate.google.com/` as of December of 2022.

- Bing Translator: https://www.bing.com/translator as of December of 2022.

- DeepL Translator: https://www.deepl.com/translator as of December of 2022.

B Examples of Translators' Errors

Table 6 shows several translating errors of common business translators. The display form is a group of five sentences: original French sentence, corresponding English reference, Bing translation, DeepL translation, and Google translation.

C Examples of Baseline Prediction Errors and Corresponding Annotations

Table 7 shows several examples of predictions and corresponding annotations of baselines in Section 4. Each group consists ten sentences, which are original French sentence, corresponding English reference, Transformer(tense-rich) prediction, Transformer(tense-poor) prediction, LSTM(tense-rich) prediction, LSTM(tense-poor) prediction, CNN(tense-rich) prediction, CNN(tense-poor) prediction, Bi-Transformer(tense-rich) prediction and Bi-Transformer(tense-poor) prediction.

D Additional Notes On Human Review

D.1 Recruitment of Human Reviewers

We recruited reviewers from students majoring in French. Taking Diplôme Approfondi de Langue Française(DALF) C1 French exam results, International English Language Testing System(IELTS) exam results, and their GPA in French courses into account, we recruited 2 reviewers in the same country of the authors' at last.

D.2 Instructions Given to Reviewers

We offer the annotation rules in Section 2, and require the reviewers to accomplish the following tasks:

- Determine whether the tense of the English translation is accurate and reasonable. If not, give an English translation that you consider reasonable.

- Determine whether the meaning of the English translation is correct. If not, give an English translation that you consider reasonable.

- Determine whether the corresponding tense label of the English translation is correct according to the natural language understanding.

E Experimental Setup

E.1 Model

Table 8 provides the number of parameters, training budget, and hyperparameters of each model. All experiments were performed on a single V100 GPU and the hyperparameters are by default. We report the result of a single run for each experiment.

E.2 Data

Table 9 shows the data statistics we used in this paper.

E.3 Packages

Table 10 shows the packages we used for preprocessing, model training, evaluation and tense labeling.

Sentence	**Tense**
Origin: On avait fait des comparaisons.	
Reference: We had made comparisons.	*Past perfect*
Bing: Comparisons were made.	*Past simple*
DeepL: Comparisons were made.	*Past simple*
Google: We made comparisons.	*Past simple*
Origin: Qui avait cru qu ’il serait facile de réunir l’ Europe ?	
Reference: Who had thought that it would be easy to reunite Europe?	*Past perfect+Modal*
Bing: Who thought it would be easy to bring Europe together?	*Past simple+Modal*
DeepL: Who thought it would be easy to reunite Europe?	*Past simple+Modal*
Google: Who thought it would be easy to reunite Europe?	*Past simple+Modal*
Origin: Je pensais avoir été assez clair.	
Reference: I thought I had been quite clear.	*Past simple+Past perfect*
Bing: I thought I was pretty clear.	*Past simple+Past simple*
DeepL: I thought I had made myself clear.	*Past simple+Past perfect*
Google: I thought I was clear enough.	*Past simple+Past simple*
Origin: Un versement similaire avait eu lieu l ’année précédente.	
Reference: A similar payment had taken place in the previous year.	*Past perfect*
Bing: A similar payment had taken place the previous year.	*Past perfect*
DeepL: A similar payment was made the previous year.	*Past simple*
Google: A similar payment had taken place the previous year.	*Past perfect*
Origin: C ’est pour cela que la voie avait été tracée à Helsinki.	
Reference: That’s why the way had been paved in Helsinki.	*Present simple+Past perfect*
Bing: That is why the path was paved out in Helsinki.	*Present simple+Past simple*
DeepL: That is why the way was paved in Helsinki.	*Present simple+Past simple*
Google: This is why the way had been traced in Helsinki.	*Present simple+Past perfect*
Origin: Je citerai pour exemple le vote à la majorité qualifiée.	
Reference: I will cite qualified majority voting as an example.	*Future simple*
Bing: One example is qualified majority voting.	*Present simple*
DeepL: An example is qualified majority voting.	*Present simple*
Google: I will cite as an example qualified majority voting.	*Future simple*
Origin: Nous espérons tous qu ’elle finira.	
Reference: We all hope that it will come to an end.	*Present simple+Future simple*
Bing: We all hope that it will end.	*Present simple+Future simple*
DeepL: We all hope it will end.	*Present simple+Future simple*
Google: We all hope it ends.	*Present simple+Present simple*
Origin: Que se passera-t-il si une nouvelle crise survient l ’année prochaine ?	
Reference: What will happen if a new crisis occurs next year?	*Future simple+Present simple*
Bing: What will happen if a new crisis occurs next year?	*Future simple+Present simple*
DeepL: What happens if there is another crisis next year?	*Present simple+Present simple*
Google: What will happen if a new crisis occurs next year?	*Future simple+Present simple*
Origin: Nous en avons terminé avec les explications de vote.	
Reference: We have finished with the explanations of vote.	*Present perfect*
Bing: That concludes the explanations of vote.	*Present simple*
DeepL: That concludes the explanations of vote.	*Present simple*
Google: We have finished with the explanations of vote.	*Present perfect*
Origin: Le fait est que le génie Internet est sorti de sa bouteille.	
Reference: The fact is that Internet genius has gone out of its bottle.	*Present simple+Present perfect*
Bing: The fact is that the Internet genie is out of the bottle.	*Present simple+Present simple*
DeepL: The fact is that the Internet genie is out of the bottle.	*Present simple+Present simple*
Google: The thing is, the internet genius is out of the bottle.	*Present simple+Present simple*
Origin: Je voulais simplement le mentionner puisqu ’on a cité certains pays.	
Reference: I just wanted to mention that because some countries have been mentioned.	*Past simple+Present perfect*
Bing: I just wanted to mention this because some countries have been mentioned.	*Past simple+Present perfect*
DeepL: I just wanted to mention it because some countries were mentioned.	*Past simple+Past simple*
Google: I simply wanted to mention it since certain countries have been mentioned.	*Past simple+Present perfect*
Origin: La dynamique de croissance et de création d ’emplois est évacuée.	
Reference: The dynamic of growth and job creation has run its course.	*Present perfect*
Bing: The momentum for growth and job creation has been removed.	*Present perfect*
DeepL: The dynamics of growth and job creation are evacuated.	*Present simple*
Google: The dynamic of growth and job creation is evacuated.	*Present simple*

Table 6: French-English utterances and corresponding translations by Bing, DeepL, Google translators. The words in orange indicate the translated verbs. The tenses in blue indicate the wrong predictions.

Sentence	Tense
Origin: J ’avais considéré que Mme Lulling était une Luxembourgeoise.	
Reference: I had assumedthat Mrs Lulling was a Luxembourgoise.	*PasPerfect+Past*
Transformer1: I believed that Mrs Lulling was a Luxembourgois.	*Past+Past*
Transformer2: I considered that Mrs Lulling was a daughter.	*Past+Past*
LSTM1: I thought that Mrs Lulling was a Luxembourgoof.	*Past+Past*
LSTM2: I considered that Mrs Lulling was a stranglehold.	*Past+Past*
CNN1: I considered that Mrs Lulling was a Luxembourgo.	*Past+Past*
CNN2: In my view, Mrs Lulling was a Luxembourger.	*Past+Past*
Bi-Transformer1: I thought that Mrs Lulling was a Luxembourgois.	*Past+Past*
Bi-Transformer2: I thought that Mrs Lulling was a sort of Greens.	*Past+Past*
Origin: Mais on les avait votés lors de la dernière période de session.	
Reference: However, they had been voted on at the last part-session.	*PasPerfect*
Transformer1: But we voted for them at the last part-session.	*Past*
Transformer2: But we voted for them at the last part-session.	*Past*
LSTM1: However, we had voted in favour of the last part-session.	*PasPerfect*
LSTM2: However, we had voted in the last part-session.	*PasPerfect*
CNN1: But we voted in the last part-session.	*Past*
CNN2: However, we voted in the last part-session.	*Past*
Bi-Transformer1: But we were voting on them at the last part-session.	*Past*
Bi-Transformer2: We, though, voted on them at the last part-session.	*Past*
Origin: Il avait été averti par l ’association des employeurs irlandais.	
Reference: He had been alerted by the Irish employers’ association.	*PasPerfect*
Transformer1: He was told it by the Irish employers’ association.	*Past*
Transformer2: The Irish employers’ association had warned.	*PasPerfect*
LSTM1: He was told it by the Irish employers’ association.	*Past*
LSTM2: It was warned by the association of the Irish employers.	*Past*
CNN1: He was told by the Irish employers’ association.	*Past*
CNN2: It was warned by the association of the Irish employers.	*Past*
Bi-Transformer1: He was told it by the Irish employers’ association.	*Past*
Bi-Transformer2: The Irish employers’ association had been notified by the Irish employers’ association.	*PasPerfect*
Origin: Je suis très curieux de voir ce que nous allons faire.	
Reference: I am very curious to see what we are going to do.	*Present*
Transformer1: I am very curious to see what we are going to do.	*Present*
Transformer2: I am very curious about what we are going to do.	*Present*
LSTM1: I am very curious to see what we will do.	*Present+Future*
LSTM2: I am very keen to see what we are going to do.	*Present*
CNN1: I am very curious to see what we are going to do.	*Present*
CNN2: I am very curious to see what we are going to do.	*Present*
Bi-Transformer1: I am very curious to see what we are going to do.	*Present*
Bi-Transformer2: I am very interested to see what we are going to do.	*Present*
Origin: Nous espérons maintenant qu ’il va agir de façon énergique.	
Reference: We now hope that he is going to act decisively.	*Present*
Transformer1: We now hope that it will act decisively.	*Present+Future*
Transformer2: Let us now hope that it will act energetically.	*Present+Future*
LSTM1: We now hope that it will act vigorously.	*Present+Future*
LSTM2: Let us hope now that it will act energetically.	*Present+Future*
CNN1: We now hope that it is going to act energetically.	*Present*
CNN2: Let us hope that it is going to act vigorously.	*Present*
Bi-Transformer1: We now hope that it will act vigorously.	*Present+Future*
Bi-Transformer2: Let us now hope that this will take a strong stand.	*Present+Future*
Origin: D’ici là, je suis sûr que nous serons passés à au moins 27 États membres.	
Reference: By then, I am sure we will have enlarged to at least 27 Member States.	*Present+FutPerfect*
Transformer1: That is why I am sure that we will be left to at least 27 Member States.	*Present+Future*
Transformer2: In this connection, I am sure we will have had at least 27 Member States.	*Present+FutPerfect*
LSTM1: I am sure that we will be at least 27 Member States.	*Present+Future*
LSTM2: That is why I am sure we will be at least 27 Member States.	*Present+Future*
CNN1: I am sure that we will be at least 27 Member States.	*Present+Future*
CNN2: That is why I am sure we will be able to pass on at least 27 Member States.	*Present+Future*
Bi-Transformer1: I am sure that we will be doing so at least 27 Member States.	*Present+Future*
Bi-Transformer2: I am sure that we will have at least 27 Member States.	*Present+Future*

Table 7: French-English utterances and corresponding predictions by baselines mentioned in Section 4. The words in orange indicate the translated verbs. The tenses in blue indicate the wrong predictions.

Model	# Param.	GPU Hours	Hyperparam.	
			learning rate	dropout
Transformer	83M	0.9h	5e-4	0.3
LSTM	58M	0.8h	1e-3	0.2
CNN	30M	0.7h	0.25	0.2
Bi-Transformer	83M	1.7h	5e-4	0.3

Table 8: The number of parameters, training budget (in GPU hours), and hyperparameters of each model.

Split	Name	# Sent.	Domain
Train	Train set from EuroparlTR (tense-rich)	97K	Politics
	Train set from Europarl (tense-poor)	97K	Politics
Test	Tense set	552	Politics
	Europarl test set	2950	Politics
	WMT15 test set	3003	News
Valid	Valid set from EuroparlTR (tense-rich)	717	Politics

Table 9: Data statistics. Training data has been filtered to avoid data leakage.

Usage	Package	License
Preprocessing	mosesdecoder (Koehn et al., 2007)[1]	LGPL-2.1
	subword-nmt (Sennrich et al., 2016)[2]	MIT
Model training	fairseq (Ott et al., 2019)[3]	MIT
Evaluation	SacreBLEU (Post, 2018)[4]	Apache 2.0
	COMET (Rei et al., 2020)[5]	Apache 2.0
Tense labeling	spaCy (Honnibal et al., 2020)[6]	MIT

[1] https://github.com/moses-smt/mosesdecoder

[2] https://github.com/rsennrich/subword-nmt

[3] https://github.com/facebookresearch/fairseq

[4] https://github.com/mjpost/sacrebleu
(nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1)

[5] https://github.com/Unbabel/COMET
(wmt20-comet-da)

[6] https://github.com/explosion/spaCy

Table 10: Packages we used for preprocessing, model training, evaluation and tense labeling.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 1,2,3,4

☑ B1. Did you cite the creators of artifacts you used?
Section 1,2,3,4 and Appendix E

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 2,3,4 and Appendix E

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section Ethics Statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix D, E

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3,4 and Appendix E

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix E

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Appendix E

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2,4 and Appendix E

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3.1

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix D

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix D and Section Ethics statement

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section Ethics statement

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix D