Are Neural Networks Extracting Linguistic Properties or Memorizing Training Data? An Observation with a Multilingual Probe for Predicting Tense

Bingzhi Li and **Guillaume Wisniewski**
Université de Paris, LLF, CNRS
75 013 Paris, France
bingzhi.li@etu.u-paris.fr
guillaume.wisniewski@u-paris.fr

Abstract

We evaluate the ability of BERT embeddings to represent tense information, taking French and Chinese as a case study. In French, the tense information is expressed by verb morphology and can be captured by simple surface information. On the contrary, tense interpretation in Chinese is driven by abstract, lexical, syntactic and even pragmatic information. We show that while French tenses can easily be predicted from sentence representations, results drop sharply for Chinese, which suggests that BERT is more likely to memorize shallow patterns from the training data rather than uncover abstract properties.

1 Introduction

The success of deep learning in many NLP tasks is often attributed to the ability of neural networks to learn, without any supervision, relevant linguistic representations. Many works have tried to identify which linguistic properties are encoded in the words and phrases embeddings uncovered during training. For instance, Belinkov et al. (2017) and Peters et al. (2018) studies the capacity of neural networks to uncover morphological information and Linzen et al. (2016), Tenney et al. (2019) or Hewitt and Manning (2019) (among many other) syntactic information. These works are based on the definition and study of *linguistic probes*: a *probe* (Alain and Bengio, 2017) is trained to predict linguistic properties from the representation of language; achieving high accuracy at this task implies these properties were encoded in the representation.

However, as pointed out by Hewitt and Manning (2019), these approaches suffer from a major drawback: there is no guarantee that the probes' good performances result from the ability of the representations to capture relevant linguistic properties rather than to memorize a large number of labeling decisions. Indeed, in most of the tasks considered so far, labels could be deduced directly from surface information, namely the word form (its morphology) and the word position in the sentence. Given the huge number of parameters of current models, there is a high risk that they are only able to extract and memorize lexically-marked patterns from training data with low (if any) generalization power.

To shed new light on this question, we consider, in this work, a multilingual linguistic probe, the goal of which is to predict the tense of a sentence (i.e. the location in time of an utterance). We compare the performance of this probe on two languages, French, which expresses tense by the verb morphology, and Chinese, in which, as explained in §2, the tense is expressed by a combination of lexical, syntactic and pragmatic cues. If intuitively the tense can be predicted from simple surface patterns in French, predicting the tense of a Chinese sentence requires capturing the interaction of all sentence-level factors related to time, and sometimes even the contextual information from the previous utterances. Contrasting the performance achieved by the probe on several languages ensures that the linguistic properties we detect are actually captured by the representation learned by BERT and not by the probe, and thus to avoid a common pitfall of this kind of approaches (Belinkov and Glass, 2019; Barrett et al., 2019).

This work has two main contributions: first, we highlight the interest of contrasting linguistic probes on different languages; second, our experiments (§3-4) show that BERT has a preference for learning lexically marked features

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3080–3089
April 19 - 23, 2021.

over lexically unmarked features and, consequently, is not able to extract an abstract representation that can be applied to (groups of) words that have not been seen at training time.

2 Tense and Aspect

Languages can roughly be classified under two categories: tense language and aspect language, depending on how they denote time relations (Xiao and McEnery, 2004). Tense indicates the temporal location of a situation, while aspect, according to Comrie (1998), refers to "the different ways of viewing the internal temporal constituency of a situation" and denotes how speakers view it in terms of its duration, frequency and whether or not it is completed. In tense language, like English or French, the tense and aspect are often encoded simultaneously in verb morphology. For example, the simple past in English locates a situation before the speech time and is often indicated by the *-ed* inflection(e.g. *worked*). Similarly, the French past tense *imparfait* is marked by the *-ait* inflection in *il travaillait* (he worked). Contrary to these two tense languages, Mandarin Chinese does not have a grammatical category of tense (Smith, 1997) and the verb morphology never changes. Figure 1 presents five sentences with the Chinese verb 加班 (*jiaban*, work overtime): while these sentences have different tenses (simple past, present progressive, habitual present, past progressive and future), the form of the verb is always the same.

The notion of tense in Mandarin Chinese is lexicalized and is often indicated by content words like adverbs of time; aspectual information is systematically conveyed by aspect markers. As an aspect language, the temporal interpretation of a verb is tightly related to the notion of *aspect*. For instance, as noticed by Lin (2006), a verb marked by a perfective aspect particle such as 了 *(le)* often gets a past interpretation, such as in the first sentence of Figure 1. And imperfective aspect privileges a present interpretation: in the example 1.2, the same verb is marked by the imperfective aspect particle 在 *(zai)*, which explains why the sentence gets a present interpretation.

However, according to the genre of the text, only 2% to 12% of verbs in Chinese have aspect markers (McEnery and Xiao, 2010)[1] and the tenses should often be inferred from contextual cues like lexical and syntactic features when there is no explicit aspect marker(Saillard, 2014). For instance, in the absence of aspect marker in sentence 1.3, the adverb 常常 (*changchang*, often) leads to a habitual present interpretation. These contextual cues can even invalidate the default correlation of time and aspect we have previously described, as in example 1.4, in which the verb group gets a past interpretation even if it is marked by an imperfective aspect particle because of the past temporal context introduced by the adverbial clause. Finally, in example 1.5, the modal auxiliary 会 *(hui)* and temporal expression 晚上 (*wanshang*, tonight/in the evening) lead to a future-tense interpretation.

Thus, unlike French and English, the time of a Chinese sentence can only rarely be deduced from a surface analysis of the sentence (i.e. from the characters composing its words) and in order to determine the tense, it is necessary to take into account both syntactic and even pragmatic information.

3 Creating a Corpus Annotated with Tense Information[2]

The tense prediction task we consider in this work requires corpora in which the tense of each verb is identified. To the best of our knowledge, there is no such publicly available corpus. For languages such as French or English, in which tenses are described by verb morphology, it is possible to easily build such corpus from morpho-syntactically labeled treebanks (e.g. from the UD project (Zeman et al., 2020)). However, this approach cannot be readily generalized to languages such as Chinese, in which, time is not explicitly marked.

We propose to leverage parallel French-Chinese corpora to obtain tense annotations for Chinese sentences automatically. Our approach relies on two hypotheses. First, we assume that the tense of a translated sentence (target sentence) is the same as the tense of

[1] Aspect markers occur more frequently in narrative texts than in expository texts

[2] **The code of all our experiments as well as the corpora we used in this work can be downloaded from** `https://github.com/bingzhilee/tense_Representation_Bert`.

(1) 他 加班 了。
Ta jiaban le
PRON.3SG work_overtime PFV
He worked overtime.

(2) 他 在 加班。
Ta zai jiaban
PRON.3SG IPFV work_overtime
He is working overtime.

(3) 他 常常 加班。
Ta changchang jiaban
PRON.3SG often work_overtime
He often works overtime.

(4) 我 去 找 他 时, 他 在 加班。
Wo qu zhao ta shi, ta zai jiaban
PRON.1SG go find PRON.3SG time, PRON.3SG IPFV work_overtime
When I went to see him, he was working overtime.

(5) 晚上 他 会 不 会 还 在 加班?
Wanshang ta hui bu hui hai zai jiaban
Tonight PRON.3SG MOD NEG MOD still IPFV work_overtime
Will he still work overtime tonight?

Figure 1: Examples of different ways to express the tense in Chinese: tenses are indicated by both aspect markers and a combination of lexical and syntactic cues and not by the verb morphology as in French or English. IPFV describes one of the two imperfective aspect markers and PFV one of the two perfective aspect markers

its original (source) sentence ignoring translationese effects (Toury, 2012; Baker et al., 1993). Second, we decided to associate each sentence with the tense of its main clause and not label each verb tense. This assumption allows us to mitigate errors related to the verbal structures identifications and misalignments between Chinese and French verbs as labels are defined at the sentence level.

We considered, in this work, the French-Chinese NEWSCOMMENTARY[3] parallel corpus (Barrault et al., 2019). To extract tense information, we use the `Stanza` pipeline (Qi et al., 2020) with its pretrained French model to find the root of each sentence and extract its tense and its mode from its (automatic) morphological analysis. We also use the dependency analysis to identify periphrastic catena expressing future (*aller* + INFINITIVE) or past (*venir de* + INFINITIVE). With these information, we can define the tense of each sentence easily by mapping the tense of the root verb to one of the three labels PRESENT, PAST or FUTURE[4] and the tense of a Chinese sentence is defined as the tense of its French translation.

We evaluate our tense extraction procedure on the PUD corpus[5]. It appears that `Stanza` is able to correctly predict the tense of 95% of the sentences (i.e. identify the root of the sentence and correctly predict its morphological information). Therefore, we consider that the tense labels are predicted with sufficient quality to measure a model's ability to capture time information. However, most of the prediction errors are due to the same construction: the auxiliary *être* followed by the past participle of the verb, that can be used to form either the passive voice or the *passé composé* tense of a verb. As a result, most of our corpus' passive sentence is labeled as PAST while they are at the present tense.

In the end, this procedure results in a corpus of 4,764 documents containing 174,347 Chinese and French sentences annotated with tense information. As expected, most of the sentences are in the present tense and the corpus is highly unbalanced: 67% of the examples are labeled PRESENT, 27% PAST and only 6% FUTURE. Our corpus also confirms the observations reported in section 1 on the limited use of temporal markers in Chinese: only 16% of the sentences have an explicit temporal marker.[6] We consider 80% of the data for

[3]We consider the 15th version of the corpus as, according to our preliminary experiments, most other parallel corpora (e.g. OpenSubtitles) contains almost exclusively sentences in the present tense.

[4]This mapping is more precisely defined in Appendix A Table 2

[5]PUD is a UD corpus that has not been used to train the `Stanza` models.

[6]More precisely: 75% of sentences in the past, 88% of sentences in the present and 85% of sentences in the

	Chinese		French	
	micro prec.	macro prec.	micro prec.	macro prec.
featSVM	73%	60%	–	–
BertSVM	71%	57%	82%	75%
fineBert	77%	68%	94%	95%

Table 1: Results achieved by our different models in the tense prediction task.

training, 10% for testing and 10% for the validation set.

4 Experimental Results

Models The task of tense prediction consists in finding, given a representation $\mathbf{x} \in \mathbb{R}^n$ of a sentence, a label describing its tense (see §3 for a definition of these labels). We consider three models. The first one, denoted featSVM, uses a SVM and a set of handcrafted features designed to capture the information identified as relevant for determining the tense of a Chinese sentence. We rely on the theoretical study of Smith and Erbaugh (2005) to define the features:[7] indicators to describe the presence of aspectual markers (e.g. 了 or 过) or temporal adverbs (e.g. 昨天 *(yesterday)* or 明天 *(tomorrow)*), the sentence root verb, modal auxiliaries (e.g. 会 *(will (probably))*, ...

Our second model, denoted BertSVM is a simple SVM that uses, as sole sentence representation, the embeddings generated by pretrained multilingual Bert. We used the second-to-last hidden layer of all tokens in the sentence and did average pooling. The embeddings on [CLS] and [SEP] were masked to zero before pooling, so they are not included(Xiao, 2018). These representations are kept unchanged. Finally, we consider fineBert a neural network in which Bert pretrained language representation are fine-tuned on the tense prediction task. More precisely, we stack a *softmax* layer on top of the pre-trained Bert model and estimate the weights of this layer while updating Bert parameters by minimizing the cross-entropy loss.

We used Google's pre-trained Multilingual cased Bert in our experiments. It was trained on the entire Wikipedia dump for each language. The French and Chinese training sets have comparable sizes. The performance of mBert on XNLI cross-lingual classification is similar for these two languages: 76.9 for French and 76.6 for Chinese (Martin et al., 2019; Devlin et al., 2018).

In our experiments, we used the SVM implementation provided by the sklearn library (Pedregosa et al., 2018) and TensorFlow in our fine-tuning experiment. Hyperparameters of the SVM have been chosen by 5-folds cross-validation.

Results We evaluate the results of the tense prediction task using both micro and macro precision to account for the imbalance between classes.[8] Results are reported in Table 1.[9] As expected, the best results both for French and Chinese are achieved by fineBert, the model in which the word and sentence representations are tailored to the tense prediction task. The relatively good performance of the featSVM shows the relevance of the considered features and validates the theoretical analysis of Smith and Erbaugh (2005). It also highlights the difficulty of defining handcrafted features generic enough to capture time information in all conditions.

Comparing performances achieved on Chinese and French is particularly interesting since it shows that our very simple architecture is able to predict almost perfectly the tense of French sentences (which can be deduced directly from the morphology of the verb and therefore from a surface analysis) but that its performance drops drastically when applied to Chinese sentences, the tense of which has to be inferred from a wide array of both lexical and syntactic cues. This observation suggests that the model is only memorizing patterns from

future have no explicit markings

[7] See Appendix B for a full description of the features considered.

[8] The macro-precision calculates the precision independently for each class, then takes the average (so all classes are treated equally), while the micro-precision aggregates the all classes' contributions to calculate the average metric.

[9] Table 3 in Appendix C provides the precision for each class.

the training set rather than inferring a meaningful representation of the sentence.

To corroborate this interpretation, we have evaluated the performance of FINEBERT in terms of the similarity between the test and train data: for each language, we have trained a 5-gram language model with Kneser-Ney smoothing using `KenLM` (Heafield et al., 2013) and divided the test set sentences into 3 groups of equal size according to the probability that the sentence was generated by the language model.[10] Because of the way tense is expressed in Chinese, ensuring that the verb of the test sentences are not in the train set is not enough. It appears, as expected, that performance drops significantly when the similarity of train and test sentences decreases: for Chinese (resp. French), the macro precision drops from 70% (resp. 96%) for the test phrases that are the most similar to the train set to 66% (resp. 93%) for the test phrases that are the most different from the train set, while their similarity with the train set (measured as the mean of $\log_{10} p(x)$ over the test test) drops from -30.62 to -93.93 (resp. -70.4 to -231.18). Detailed results are presented in Appendix D.

These results clearly show that the higher the similarity with the train sentences is, the more accurate the model is. Again, this observation questions the capacity of the model to capture relevant linguistic properties rather than simply memorizing the training data.

Discussion Our experiments clearly show that BERT prefers learning lexically marked features rather than lexically unmarked features. These results indicate that, even if several confounders exist, neural networks tend to memorize shallow patterns from the training data rather than uncover abstract linguistic properties.

There is a first well-known confounding factor when interpreting probing results: high classification accuracy could result from the fact that the probe has learned the linguistic task and not from the properties captured by the representation. In our work, we avoid this pitfall by considering a multilingual probe setting: our conclusions are not based on an absolute score but on the comparison of the performance achieved in French and Chinese by the same probe.

The difference in performance between the French and Chinese models is a second confounder. This difference can result from either the training set size or the model's architecture tailored to extract only lexically-marked information. In recent work, Warstadt et al. (2020) suggests that it may be possible that, if more data were available, BERT could eventually learn to predict Chinese tense. However,it must be pointed out that the French model achieves a precision of 76.9% and the Chinese model a precision of 76.6% on the XLNI cross-lingual classification task, the standard evaluation benchmark of sentence representation models. Therefore, we believe that our conclusions are not biased by the language modeling performance.

There is a third and last possible confounder: it is possible that, as explained in §2, sometimes, in Chinese the cues to tense may appear in an earlier sentence. Gong et al. (2012) showed that the tense of previous sentences has a close relation to the current sentence. Considering contextual features in our feature-based SVM classifier only increases the accuracy by 2%, therefore, we believe that this factor has only a moderate impact.

5 Conclusion

We have shown that the performance of a tense prediction model varies dramatically depending on how the language expresses time, a result that suggests that BERT is more likely to memorize shallow patterns from the train set rather than uncover abstract properties. Our work also highlights the interest of comparing linguistic probes across languages which opens up a new avenue for research.

Acknowledgments

We sincerely thank the reviewers and Program Chairs for their careful reviews and insightful comments, which are of great help in improving the manuscript. We would like to express special gratitude to Professor Claire Saillard, this work would not have been possible without her helpful comments and advice.

[10] More precisely, we have ordered the sentences of the test set according to their probability estimated by the language model and considered the 5,814 first (resp. last) sentences as the most different (resp. similar) from the train set.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *Text and Technology: In honour of John Sinclair*. John Benjamins Publishing. Google-Books-ID: dTlIAAAAQBAJ.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong, China. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bernard Comrie. 1998. *Aspect: an introduction to the study of verbal aspect and related problems*, reprinted edition. Number 2 in Cambridge textbooks in linguistics. Cambridge Univ. Press, Cambridge. OCLC: 247564008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. 2012. N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 276–285, Jeju Island, Korea. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jo-Wang Lin. 2006. Time in a Language Without Tense: The Case of Chinese. *Journal of Semantics*, 23(1):1–53.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *CoRR*, abs/1911.03894.

Tony McEnery and Richard Xiao. 2010. *Corpus-based contrastive studies of English and Chinese*. Routledge, New York. OCLC: 1086452482.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. Scikit-learn: Machine Learning in Python. *arXiv:1201.0490 [cs]*. ArXiv: 1201.0490.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv:2003.07082 [cs]*. ArXiv: 2003.07082.

Claire Saillard. 2014. De l'aspect au temps. Expression de la temporalité et de l'aspectualité en français L2 par des apprenants sinophones . In *The expression of temporality by L2 learners of French and English*, Montpellier : Université Paul Valéry, Unknown Region.

Carlota S Smith. 1997. *The parameter of aspect*. Kluwer, Dordrecht; Boston. OCLC: 1012457859.

Carlota S. Smith and Mary S. Erbaugh. 2005. Temporal interpretation in Mandarin Chinese. *Linguistics*, 43(4). Number: 4.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Gideon Toury. 2012. *Descriptive Translation Studies –and beyond: Revised edition*. John Benjamins Publishing. Google-Books-ID: jX4weQT63rIC.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Richard Xiao and Tony McEnery. 2004. *Aspect in Mandarin Chinese: a corpus-based study*. Number v. 73 in Studies in language companion series. J. Benjamins Pub, Amsterdam ; Philadelphia. OCLC: ocm56526702.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agić, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Angelina Aquino, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Ethan Chi, Jinho Choi, Yongseok Cho, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin

Mititelu, Maria Mitrofan, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Mapping of French Tense

When building our corpus, the tense of each sentence was deduced from the automatic morphological analysis of its root verb using the mapping defined in Table 2.

French passé composé describes a situation that was completed in the past and emphasizes its results in the present. Smith (1997)considers that the Passé composé presents two tense values. When used to present a given state of things, passé composé is temporally present. When used to denote past facts, passé composé, called a preterit by Smith, is temporally past. Which tense value of passé composé should we take into account, the perfect present or past? Concerning the translation of perfect present into Chinese, it' s interesting to read the corpus study of Xiao and McEnery (2004), according to which the perfect present tense of English is most frequently (71%) translated into Chinese by perfective aspect. Whereas in Chinese, Lin (2006) contends a preferential correlation between perfective aspect and past tense. We have thus decided to classify the passé composé into the Past category.

B Features used in featSVM

1. **Root verb**: The root verb may denote some intrinsic features related to tense(Xiao and McEnery, 2004) In the automatic morphological analysis generated by Stanza, the root word is the verb with a dependency label of *root*, or the Chinese adjectives (Chinese adjectives can be used as verbs) governed by root directly.

2. **Aspect markers**: Perfective aspect marker (了 le, 过 guo) and the imperfective aspect marker (着 zhe).We didn't consider another imperfective aspect marker (在 zai) because `Stanza` didn't annotate this marker.

3. **Temporal nouns**: We have extracted a list of temporal nouns. These words have been annotated by `Stanza` with the dependency label *nmod:tmod*. A complex sentence could contain multiple *nmod:tmod* words. We only consider the one that is governed by its root verb or by the verb governed directly by the root word. This list mainly contains words like 现在 *(now)*, 明天 *(tomorrow)*, 刚才 *(just now)*.

4. **Temporal adverbs**: We have determined a list of adverbs with temporal connotation. For example, 已经 *(already)*, 一直 *(always)*, 曾 *(once)*. These adverbs have been annotated by `Stanza` with the dependency label *advmod*. Like the temporal nouns, we only take into account the temporal adverb directly governed by the root word.

5. **Modal auxiliaries**: Words that express necessity, expectation, possibility of the action described by the verb. The bounded situation in the future in Chinese is often expressed by modal auxiliaries 要 *(is going to)* or 会 *(it is probable that)*.

6. **Contextual tense**: We consider the tense of the previous sentence as contextual tense, which provides contextual temporal cues for some Chinese sentences that have no temporal words or aspect markers at all.

7. **Words and POS patterns**: Combinations of word and POS tag for each word in the whole sentence. These features are expected to capture some special syntactic structure. For example, the structure 就要 + predicate + 了 *(...is going to happen)* indicates a near-future situation.

C Performance of different models for each class

Table 3 presents the results of different classifiers for each tense category. It shows that more frequent data are more likely to be better predicted: the Present class (67% of the examples) gets the highest score for all classifiers except French FINEBERT.

D Results of the performance of FineBert

Table 4 details the performance of FineBert and the impact of the similarity between the train and test sentences. The similarity is measured as the mean of $\log_{10} p(x)$ over the test test.

Tense predicted for the root verb	Label of the sentence
présent de l'indicatif	PRESENT
présent du conditionnel	PRESENT
présent de l'impératif	PRESENT
imparfait	PAST
plus-que-parfait	PAST
passé simple	PAST
passé récent	PAST
passé composé	PAST
futur simple	FUTURE
futur proche	FUTURE

Table 2: Mapping between French tenses and tense labels used to build our corpus.

	Chinese			French		
	Future	Past	Present	Future	Past	Present
featSVM	36%	63%	81%	-	-	-
BertSVM	31%	60%	79%	64%	73%	88%
fineBERT	51%	71%	82%	95%	97%	94%

Table 3: Precision achieved by the different classifiers for each class in tense prediction tasks

	Chinese			French		
	sim.	micro prec.	macro prec.	sim.	micro prec.	macro prec.
subgroup1	-93.93	75%	66%	-231.18	92%	93%
subgroup2	-55.92	77%	68%	-135.41	94%	94%
subgroup3	-30.62	79%	70%	-70.74	96%	96%

Table 4: Results based on the similarity of test sentences to the train set