LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

cnrs université PARIS-SACLAY

Les enjeux de la compréhension des langues

Sahar Ghannay[1]

6 décembre 2022

[1]Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France,
sahar.ghannay@lisn.upsaclay.fr

Plan

Compréhension des langues

Compréhension des langues

Objectif : **extraire** des éléments de sens de la requête de l'utilisateur pour **construire** une représentation symbolique manipulable par le système de dialogue

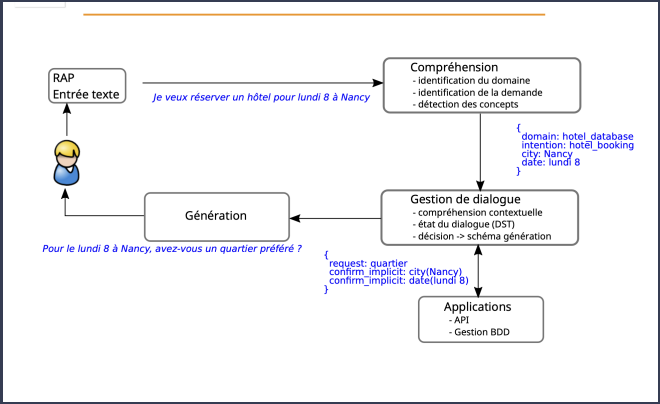


Figure 1: Architecture d'un système de dialogue orienté tâche [Rosset, 2018]

- Tâches de dialogue : Détection de domaine, Détection d'intention, **Détection de concept** (une tâche de remplissage de formulaire (Slot filling))

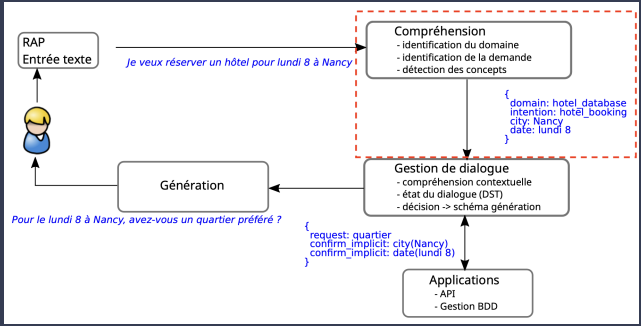


Figure 2: Architecture d'un système de dialogue orienté tâche [Rosset, 2018]

Compréhension de langues : état de l'art

- Modèles simples $\Longrightarrow$ pas gourmands en calcul (CPU suffisant)
  - FST (Finite State Machine) [Hahn et al., 2010]
  - champs aléatoires conditionnels (CRF) [Lafferty et al., 2001]
  - SVM (Support Vector Machine)[Hahn et al., 2010]
  - ...
- Modèles complexes, $\Longrightarrow$ gourmands en calcul (besoin de GPUs)
  - BiLSTM-CNN-CRF [Ma and Hovy, 2016]
  - Réseaux de neurones hybrides [Dinarelli and Grobol, 2019]
  - Transformeurs [Vaswani et al., 2017]
  - modèles auto-supervisés entrainé à partir du texte (BERT) [Devlin et al., 2019] ou de signal audio wav2vec [Baevski et al., 2020] $\Longrightarrow$ Modèle état de l'art

La question de l'utilisabilité du transformeur

La question de l'utilisabilité du transformeur

Modèles actuels :

Modèles plus grands

- $\text{BERT}_{\text{base}}$ possède 110 millions de paramètres et a été pré-entrainé sur 16 gigaoctets(Go) de texte sur **4 TPUs** pendant **4 jours** (TPU de 15 fois à 30 fois plus rapide qu'un GPU) equivalent à **1890 jours** sur **1 GPU**.
- GPT-3 [Brown et al., 2020] a été pré-entraîné sur 45 téraoctets (To) de textes et compte 175 milliards de paramètres sur **512 GPUs** pendant **10 jours**.

Modèle compact :

- ALBERT [Lan et al., 2020] : les plus petits modèles pré-entraînés, 12 millions (M) de paramètres, taille de modèle $<$50 mégaoctet (Mo) sur 1 GPU pendant 36.7 heurs
- factorisation de la matrice d'attention et partage des paramètres entre les couches
  ☛ **réduire la complexité des calculs, accélérer les phases d'apprentissage et d'inférence**

FrALBERT : un nouveau modèle compact en français pré-entraîné sur Wikipédia

Motivations

Motivations

Pratiques adoptées lors du finetuning des modèles de langue basés sur des transformeurs : inappropriées dans des conditions de ressources limitées
[Zhang et al., 2021, Mosbach et al., 2021]

☛ *Quel est le comportement d'un modèle basé sur de transformeur dans le cadre d'une tâche de compréhension de langues en français, et quel est l'impact écologique des modèles ?*

Notre protocole : optimisation des hyperparamètres, une tâche de conpréhension de la langue en français,
13 modèles transformeurs : 8 modèles monolingues en français & 5 modèles multilingues

Optimisation des hyperparamètres

Les modèles fonctionnent mieux avec des hyperparamètres différents !

Approche : Population-Based Training [Jaderberg et al., 2017] :

- optimiser conjointement une population de modèles et leurs hyperparamètres pour maximiser les performances SLU
- optimisation asynchrone rapide
- ...

☛ de meilleures solutions en moins de temps (contraintes de ressources)

Données

Benchmark MEDIA [Bonneau-Maynard et al., 2006] :
Un corpus en français pour la tâche de comptéhension de langues (NLU/SLU)

- réservation d'hôtel et informations
- 76 étiquettes sémantiques
  - train : 13k énoncés, 10h45 audio
  - dev : 1.3k énoncés, 1h13 audio
  - test : 3.5k énoncés, 2h59 audio

☛ *une des tâches les plus difficiles* [Béchet and Raymond, 2019]

Modèles transformeurs à évaluer

- modèles français monolingues CamemBERT [Martin et al., 2020] & FlauBERT [Le et al., 2020]
- modèles multilingues : XLM-R [Conneau et al., 2020] et mBERT [Devlin et al., 2019]
- deux modèles multilingues compacts : distil-mBERT [Sanh et al., 2019] et small-mBERT [Abdaoui et al., 2020]
- FrALBERT [Cattan et al., 2021] version pré-entraînée du modèle compact ALBERT [Lan et al., 2020]

Modèle	Objectifs	Données	taille du vocabulaire	Tokenization	# paramètres	taille de Modèle
FlauBERT$_{\text{base}}$	MLM	24 Sous-corpus français (71 Gb of text)	68,729	BPE	138 M	553 Mb
CamemBERT$_{\text{base}}$	MLM	OSCAR français(138 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	CCNet français (135 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	OSCAR français(4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{base}}$	MLM	CCNet français (4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{\text{large}}$	MLM	CCNet français (135 Gb of text)	32,005	SentencePiece	335 M	1.35 Gb
CamemBERT$_{\text{base}}$	MLM	Wikipedia français(4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
FrALBERT$_{\text{base}}$	MLM and SOP	Wikipedia français(4 Gb of text)	32,005	SentencePiece	12 M	50 Mb
XLM-R$_{\text{base}}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	278 M	1.12 Gb
XLM-R$_{\text{large}}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	559 M	1.24 Gb
mBERT$_{\text{base}}$	MLM and NSP	Wiki-100	119,547	WordPiece	177 M	714 Mb
small-mBERT$_{\text{base}}$ FR	MLM and NSP	Wiki-100	24,495	WordPiece	104 M	420 Mb
distil-mBERT$_{\text{base}}$	MLM and NSP	Wiki-100	119,547	WordPiece	134 M	542 Mb

Modèles transformeurs à évaluer

- modèles français monolingues CamemBERT [Martin et al., 2020] & FlauBERT [Le et al., 2020]
- modèles multilingues : XLM-R [Conneau et al., 2020] et mBERT [Devlin et al., 2019]
- deux modèles multilingues compacts : distil-mBERT [Sanh et al., 2019] et small-mBERT [Abdaoui et al., 2020]
- FrALBERT [Cattan et al., 2021] version pré-entraînée du modèle compact ALBERT [Lan et al., 2020]

Modèle	Objectifs	Données	taille du vocabulaire	Tokenization	# paramètres	taille de Modèle
FlauBERT$_{base}$	MLM	24 Sous-corpus français (71 Gb of text)	68,729	BPE	138 M	553 Mb
CamemBERT$_{base}$	MLM	OSCAR français(138 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet français (135 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	OSCAR français(4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{base}$	MLM	CCNet français (4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
CamemBERT$_{large}$	MLM	CCNet français (135 Gb of text)	32,005	SentencePiece	335 M	1.35 Gb
CamemBERT$_{base}$	MLM	Wikipedia français(4 Gb of text)	32,005	SentencePiece	110 M	445 Mb
FrALBERT$_{base}$	MLM and SOP	Wikipedia français(4 Gb of text)	32,005	SentencePiece	12 M	50 Mb
XLM-R$_{base}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	278 M	1.12 Gb
XLM-R$_{large}$	MLM	CC-100 (2.5 Tb of text)	250,002	BPE	559 M	1.24 Gb
mBERT$_{base}$	MLM and NSP	Wiki-100	119,547	WordPiece	177 M	714 Mb
small-mBERT$_{base}$ FR	MLM and NSP	Wiki-100	24,495	WordPiece	104 M	420 Mb
distil-mBERT$_{base}$	MLM and NSP	Wiki-100	119,547	WordPiece	134 M	542 Mb

Expériences

Expériences

- Optimisation des hyperparamètres [Jaderberg et al., 2017] : nb. epochs de fineTuning varie de 5 à 100, la taille du batch varie de 8 à 32 et le taux d'apprentissage dans la plage de 1 à 5
- benchmark MEDIA
- métriques d'évaluation
  - F1
  - Concept Error Rate (CER) :

$$CER = \frac{\#Insertion \ + \ \#Substitution \ + \ \#Deletions}{\#\ Concepts}$$

- Métriques écologiques :
  - Temps (secondes)
  - Énergie (kWh)
  - CO_2 (g)

Résultats en NLU et coûts écologiques sur lab-IA

Étape	Fine-tuning (1 époch)			Inférence				
Modèles	Temps (s)	Énergie (kWh)	CO_2 (g)	Temps (s)	Énergie (kWh)	CO_2 (g)	F1	CER
$\text{FlauBERT}_{\text{base}}$	121.89	765.24	554.04	9.44	26.52	19.20	89.0	8.1
$\text{CamemBERT}_{\text{large, CCNet 135 Gb}}$	144.31	659.34	477.36	23.67	66.58	48.20	89.2	7.8
$\text{CamemBERT}_{\text{base, OSCAR 138 Gb}}$	130.06	789.69	571.74	9.46	26.58	19.24	89.9	*__7.5__
$\text{CamemBERT}_{\text{base, CCNet 135 Gb}}$	118.58	671.42	486.11	7.28	20.44	14.80	89.3	7.9
$\text{CamemBERT}_{\text{base, OSCAR 4 Gb}}$	116.59	623.44	451.37	7.43	20.87	15.11	89.7	8.3
$\text{CamemBERT}_{\text{base, CCNet 4 Gb}}$	115.54	662.78	479.86	7.31	20.53	14.86	89.7	8.3
$\text{CamemBERT}_{\text{base, Wiki 4 Gb}}$	109.57	645.96	467.68	7.19	20.19	14.62	__90.0__	8.4
$\text{FrALBERT}_{\text{base, Wiki 4 Gb}}$	72.65	474.69	343.67	4.26	11.95	8.65	89.8	8.6
$\text{XLM-R}_{\text{base}}$	125.26	549.48	397.82	8.04	22.59	16.35	89.5	8.5
$\text{XLM-R}_{\text{large}}$	196.74	1 155.15	836.33	26.03	73.24	53.02	89.9	8.0
$\text{mBERT}_{\text{base}}$	119.36	673.56	487.66	8.39	23.56	17.06	88.9	8.7
$\text{distill-mBERT}_{\text{base}}$	80.10	545.46	394.91	7.04	19.75	14.30	*87.5	*10.1
$\text{small-mBERT}_{\text{base-fr}}$	112.08	589.84	427.04	7.69	21.59	15.63	*88.8	*8.1

Résultats en NLU et coûts écologiques sur lab-IA

Étape / Modèles	Fine-tuning (1 époch) Temps (s)	Énergie (kWh)	CO_2 (g)	Inférence Temps (s)	Énergie (kWh)	CO_2 (g)	F1	CER
$FlauBERT_{base}$	121.89	765.24	554.04	9.44	26.52	19.20	89.0	8.1
$CamemBERT_{large,\ CCNet\ 135\ Gb}$	144.31	659.34	477.36	23.67	66.58	48.20	89.2	7.8
$CamemBERT_{base,\ OSCAR\ 138\ Gb}$	130.06	789.69	571.74	9.46	26.58	19.24	89.9	*__7.5__
$CamemBERT_{base,\ CCNet\ 135\ Gb}$	118.58	671.42	486.11	7.28	20.44	14.80	89.3	7.9
$CamemBERT_{base,\ OSCAR\ 4\ Gb}$	116.59	623.44	451.37	7.43	20.87	15.11	89.7	8.3
$CamemBERT_{base,\ CCNet\ 4\ Gb}$	115.54	662.78	479.86	7.31	20.53	14.86	89.7	8.3
$CamemBERT_{base,\ Wiki\ 4\ Gb}$	109.57	645.96	467.68	7.19	20.19	14.62	__90.0__	8.4
$FrALBERT_{base,\ Wiki\ 4\ Gb}$	72.65	474.69	343.67	4.26	11.95	8.65	89.8	8.6
$XLM\text{-}R_{base}$	125.26	549.48	397.82	8.04	22.59	16.35	89.5	8.5
$XLM\text{-}R_{large}$	196.74	1 155.15	836.33	26.03	73.24	53.02	89.9	8.0
$mBERT_{base}$	119.36	673.56	487.66	8.39	23.56	17.06	88.9	8.7
$distill\text{-}mBERT_{base}$	80.10	545.46	394.91	7.04	19.75	14.30	*87.5	*10.1
$small\text{-}mBERT_{base\text{-}fr}$	112.08	589.84	427.04	7.69	21.59	15.63	*88.8	*8.1

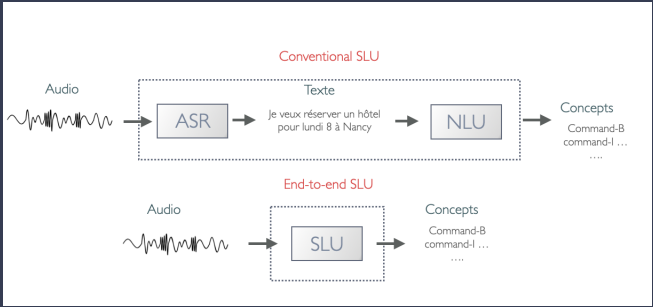
☛ F1/CER : résultats comparables entre $CamemBERT_{base,\ Wiki\ 4\ Go}$ & $FrALBERT_{base,\ Wiki\ 4\ Go}$

☛ Coûts : de 26 % à 57 % de moins dans la phase de finetuning & de 41 % à 45 % de moins dans la phase d'inférence

Compréhension de la parole

Compréhension de la parole

Objectif : **extraire** de l'infromation sémantique à partir du signal audio.



- **wav2vec** [Evain et al., 2021] : possède **317 millions** de paramètres et a été entrainé sur **3000 heures** de parole en français pendant **2 semaines sur 64 GPUs**.

Model	CER	temps d'inférence
end-2-end (SLU : wav2vec)	14.5	180
cascade (ASR (wav2vec) + NLU (Camembert))	11.2	9.46

Table 1: Résultats SLU sur MEDIA utilisant les deux approches [Ghannay et al., 2021]

Conclusions

Conclusions

- FrALBERT, le premier modèle compact disponible gratuitement pour le français [1]
  - Performance comparable sur le tâche SLU française aux plus grands modèles
- Les modèles compacts sont une alternative aux modèles à forte consommation d'énergie, des performances comparables tout en réduisant leur taille et leur complexité de calcul

☛ On a toujours besoin d'utiliser des GPU même avec des petits modèles

1. Disponible sur HuggingFace ici :`https://huggingface.co/qwant/fralbert-base`

Merci pour votre attention !

Questions ?

Références i

Abdaoui, A., Pradel, C., and Sigel, G. (2020).
Load what you need : Smaller versions of mutililingual BERT.
In Proceedings of SustaiNLP : Workshop on Simple and Efficient Natural Language Processing, pages 119–123, Online. Association for Computational Linguistics.

Baevski, A., Zhou, Y., Abdelrahman, M., and Auli, M. (2020).
wav2vec 2.0 : A framework for self-supervised learning of speech representations.
Advances in Neural Information Processing Systems.

Béchet, F. and Raymond, C. (2019).
Benchmarking benchmarks : introducing new automatic indicators for benchmarking spoken language understanding corpora.
In InterSpeech.

Références ii

Bonneau-Maynard, H., Ayache, C., Bechet, F., Denis, A., Kuhn, A., Lefevre, F., Mostefa, D., Quignard, M., Rosset, S., Servan, C., and Villaneau, J. (2006).
Results of the French Evalda-Media evaluation campaign for literal understanding.
In LREC.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
Language models are few-shot learners.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Références iii

Cattan, O., Servan, C., and Rosset, S. (2021).
On the Usability of Transformers-based models for a French Question-Answering task.
In Recent Advances in Natural Language Processing, RANLP 2021, Online.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm'an, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020).
Unsupervised cross-lingual representation learning at scale.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT : Pre-training of deep bidirectional transformers for language understanding.

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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dinarelli, M. and Grobol, L. (2019).
Modèles neuronaux hybrides pour la modélisation de séquences : le meilleur de trois mondes.
In
TALN-RECITAL 2019 - 26ème Conférence sur le Traitement Automatique des Langues Naturelles, Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL), Toulouse, France. ATALA.

Evain, S., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021).
Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark.
In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021), NeurIPS 2021 Datasets and Benchmarks Track, on-line, United States.

Ghannay, S., Caubrière, A., Mdhaffar, S., Laperrière, G., Jabaian, B., and Estève, Y. (2021).
Where are we in semantic concept extraction for Spoken Language Understanding ? ⋆.
In SPECOM 2021 23rd International Conference on Speech and Computer, Saint Petersburg, Russia.


Hahn, S., Dinarelli, M., Raymond, C., Lefevre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., and Riccardi, G. (2010).
Comparing stochastic approaches to spoken language understanding in multiple languages.
IEEE Transactions on Audio, Speech, and Language Processing, 19(6) :1569–1583.


Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017).
Population based training of neural networks.
arXiv preprint arXiv :1711.09846.


Lafferty, J., McCallum, A., and Pereira, F. C. (2001).
Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020).
ALBERT : A lite BERT for self-supervised learning of language representations.
In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020).
Flaubert : Unsupervised language model pre-training for french.
In Proceedings of The 12th Language Resources and Evaluation Conference, pages 2479–2490, Marseille, France. European Language Resources Association.

Ma, X. and Hovy, E. (2016).
End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.


Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020).
CamemBERT : a tasty French language model.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7203–7219, Online. Association for Computational Linguistics.


Mosbach, M., Andriushchenko, M., and Klakow, D. (2021).
On the stability of fine-tuning {bert} : Misconceptions, explanations, and strong baselines.
In International Conference on Learning Representations.


Rosset, S. (2018).
Dialogue humain-machine.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019).
Distilbert, a distilled version of BERT : smaller, faster, cheaper and lighter.
CoRR, abs/1910.01108.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017).
Attention is all you need.
In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2021).
Revisiting few-sample {bert} fine-tuning.
In International Conference on Learning Representations.