Projet Science des données

Contents

Remerciements

Analyse du problème

Ce projet s'inscrit dans le cadre du Défi IA, organisé par l'INSA Toulouse. Cette édition du Défi IA porte sur le traitement automatique du language.

La tâche est simple : attribuer la bonne catégorie d'emploi à une description de poste. Il s'agit donc d'une tâche de classification multi-classes avec 28 classes.

Les données ont été extraites de CommonCrawl. Les données sont donc représentatives de ce que l'on peut trouver sur la partie anglophone de l'Internet, et contiennent donc un certain biais. L'un des objectifs de ce concours est de concevoir une solution qui soit à la fois précise et équitable.

Classification

La précision sera mesurée par le macro F1 score :

$$F_{1_{marco}} = \frac{1}{C} \sum_{k=1}^{C} 2 \cdot \frac{\text{precision}_k \cdot \text{recall}_k}{\text{precision}_k + \text{recall}_k}$$

$$$$

où la présicion est définie tel quel :

$$\text{precision}_k = \frac{tp}{tp + fp}$$

et le recall est défini tel quel :

$$\text{recall}_k = \frac{tp}{tp + fn}$$

L'équité sera mesurée par le Disparate Impact:

$$DI_k = \frac{\mathbb{P}(\text{Job} = k | \text{Gender} = F)}{\mathbb{P}(\text{Job} = k | \text{Gender} = M)}$$

Fairness vis à vis de `Gender`

Analyse du jeu de données :

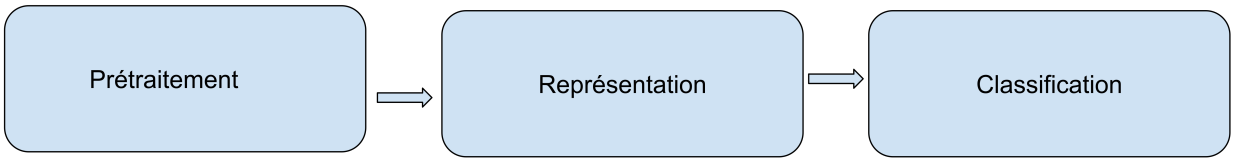
Structure & Complexité

Extrait de descritpions : - Diversite des descriptions : - cas normaux representatifs - cas pathologiques (classes ou il y a peu de descriptions, fautes d'orthographes)

Réalisation du projet

Méthodologie et découpage du projet :

Nous avons separe notre projet en 3 parties : la peparation et le pretraitment, la representation et la classification.



Prétraitement

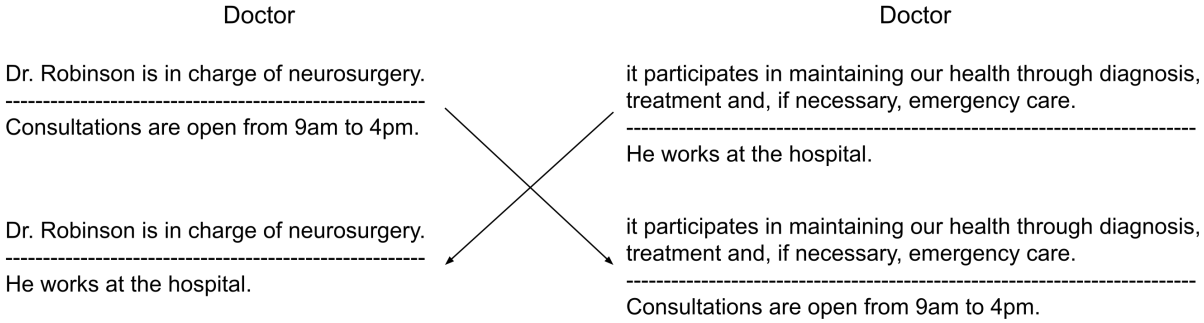
Nous avons utiliser plusieurs pretraitement : lesquels ?

Augmentation des donnees

Pour rendre notre modele de classification plus robuste, nous avons augmenter le jeu de donnees. Nous avons donc utilise plusieurs techniques :
(Ces techniques sont inspiree de *L'AUGMENTATION DE DONNEES EN NLP*[1])

- Augmentation par crossover :

Nous sélectionnons des descriptions faisant reference au meme metier. Nous decoupons les descriptions en phrases et nous creons des exemples en melangant les phrases des deux descriptions.



- Augmentation par substitution/insertion
- Augmentation par rétro-traduction : Nous traduisons une description dans une langue etrangère (en chinois, par exemple) puis nous retraduisons dans la langue originale.

[1](BOURDOIS 2020)

- Augmentation par echange

Représentation

Comme representation nous avons utiliser plusieurs vectorizeurs :

- CountVectorizer
- TFIDFVectorizer
- WordEmbeeding

Classification

Deep learning

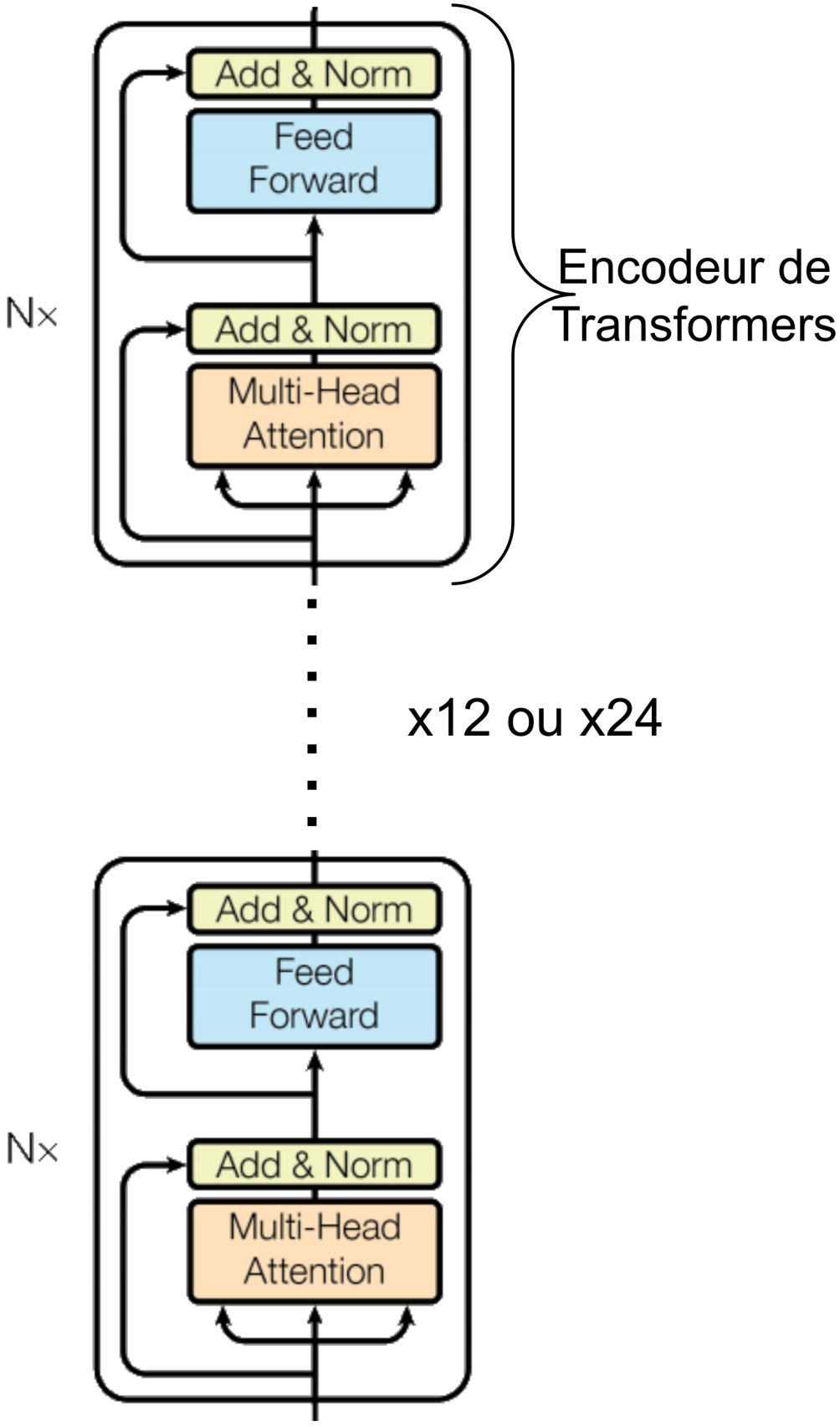
Parmis les modeles de language les plus connu nous retrouvons :

- Modele recurrent : LSTM , BLSTM
- Modeles Transformer Based :
  - BERT
  - GPT2 - GPT3
  - T5

Nos architectures

- Rappel sur BERT

BERT pour **Bidirectional Encoder Representations from Transformers** (Devlin et al. (2019)) est un modèle de language. L'architecture BERT est un empillement d'encodeurs de Transformers (Vaswani et al. (2017)).


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Encodeur de Transformers
x12 ou x24
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx

BERT apprend de façon auto-supervisee, l'entrée se suffit à elle même, pas besoin de labelliser quoi que ce soit. BERT est entrainé sur deux taches : * La prediction de tokens masqués appelle MLM ("Masked language model"). * La prediction de la prochaine phrase (Next)

- Implémentation et fine-tunning de BERT

Differents Modeles

Résultats & Performances

Conclusions

References

BOURDOIS, Loïck, trans. 2020. "L'AUGMENTATION DE DONNEES EN NLP." *Loïck BOURDOIS*. https://lbourdois.github.io/blog/nlp/Data-augmentation-in-NLP/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [Cs]*, May. http://arxiv.org/abs/1810.04805.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *arXiv:1706.03762 [Cs]*, December. http://arxiv.org/abs/1706.03762.