Équipe TTGV

DEFT 2023

Défi Fouille de Textes@TALN 2023

Tal à Très Grande Vitesse

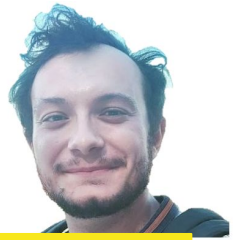
Qui sommes nous ?

Vanessa

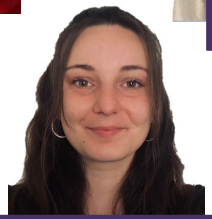
Solène

Aurélien

Christophe



Cyrille



Andréa

Hélène

Tom

Matthieu

Barbara

Gaël

Stratégie générale

Simple/Multiple

Multi-classes

Ajout de contextes

Galactica

Expressions régulières

Annotation d'entités

Approches génératives

Bloomz

Régression Logistique

LSA/LDI

Approche contrastive

Multi-étiquettes

Génération de contextes

Dr-BERT

Transformers

MiniBERT

Simple / multiple

Tâche : prédire si une question admet une ou plusieurs réponses (classification binaire)

Modèles explorés :

- Expressions régulières : extraire des indices dans le libellé des *questions*
- LSI / LDA : regrouper les *réponses* en thèmes disjoints. 1 thème ↔ 1 réponse
- Régression logistique (RL) (Accuracy & Macro-F1 : 0,94)

Résultats :

- Aucun de ces modèles ne permet de résoudre complètement la tâche annexe, mais :
  - ➜ Scores encourageants avec la régression logistique
  - ➜ Prédictions binaires à utiliser comme entrée d'un modèle plus fin pour la tâche principale

Tâche annexe

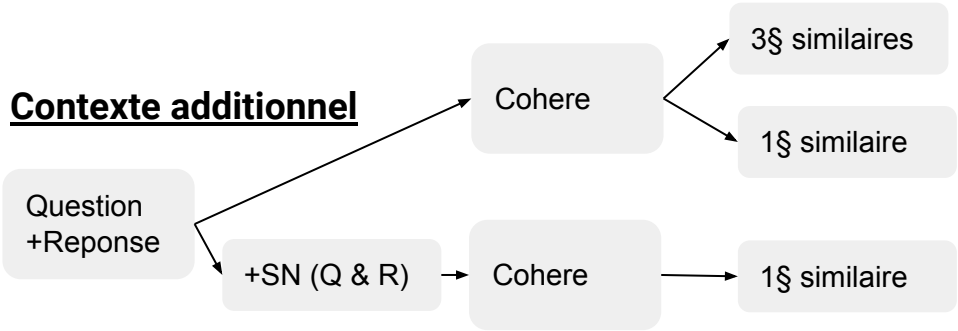
Modèle	macro F1	Accuracy
LDA	13,26	19,13
Régression logistique	27,98	**62,54**

Approche multiclass

Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.

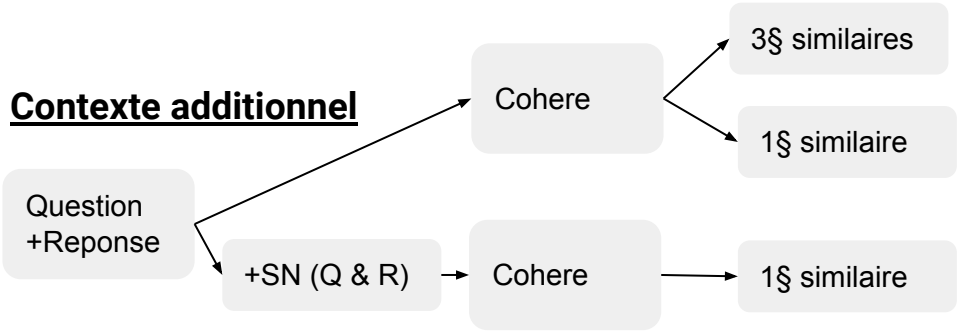
Approche multiclass

Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.



Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.

Approche multiclass



Approche contrastive

1) DrBERT Finetuning → QUAERO (NER) → annotation → freeze

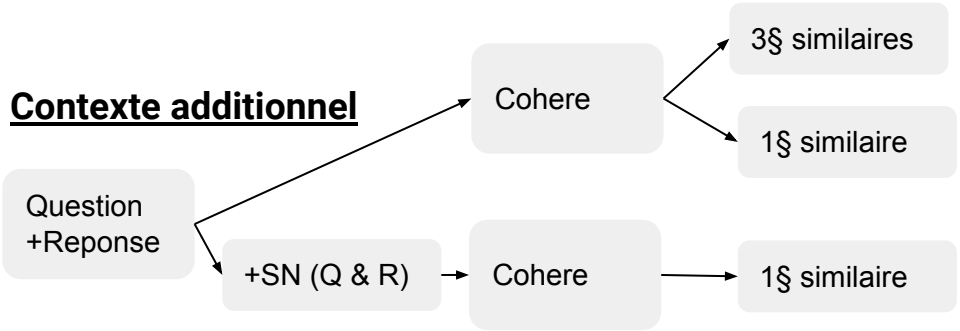
2) Reconnaître entités ayant substituées des entités réelles

$$\mathcal{L}_{\text{contrastive}} = \mathbb{1}_{e \in E^+} \log P(e|C) + (1 - \mathbb{1}_{e \in E^+}) \log(1 - P(e|C))$$

$$\mathcal{L} = \mathcal{L}_{\text{classif}} + \lambda \mathcal{L}_{\text{contrastive}}$$

Approche multiclass

Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.
Infirmée



Approche contrastive

1) DrBERT Finetuning → QUAERO (NER) → annotation → freeze

2) Reconnaître entités ayant substituées des entités réelles

$$\mathcal{L}_{\text{contrastive}} = \mathbb{1}_{e \in E^+} \log P(e|C) + (1 - \mathbb{1}_{e \in E^+}) \log(1 - P(e|C))$$

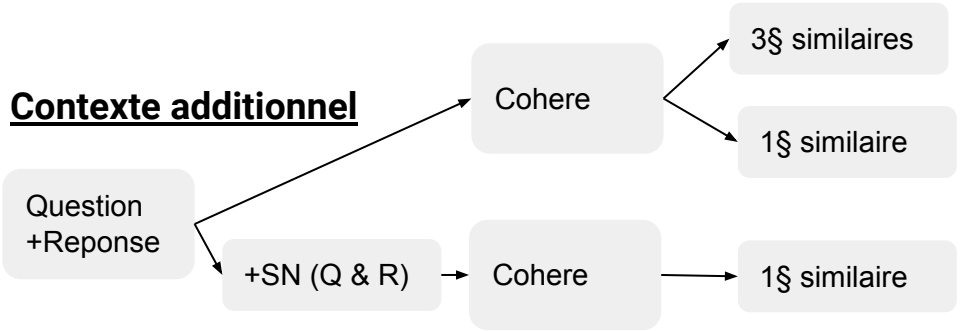
$$\mathcal{L} = \mathcal{L}_{\text{classif}} + \lambda \mathcal{L}_{\text{contrastive}}$$

Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43

Approche multiclass

Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.
Infirmée



Modèle	Hamming	EMR
Camembert-base	33,80	14,31
Dr-bert-7GB	**39,08**	**17,68**
Dr-bert-7GB$_{\text{contexte}}$	35,31	15,27
Dr-bert-7GB$_{\text{contrast}}$	36,06	16,40

Approche contrastive

1) DrBERT Finetuning → QUAERO (NER) → annotation → freeze

2) Reconnaître entités ayant substituées des entités réelles

$$\mathcal{L}_{\text{contrastive}} = \mathbb{1}_{e \in E^+} \log P(e|C) + (1 - \mathbb{1}_{e \in E^+}) \log(1 - P(e|C))$$

$$\mathcal{L} = \mathcal{L}_{\text{classif}} + \lambda \mathcal{L}_{\text{contrastive}}$$

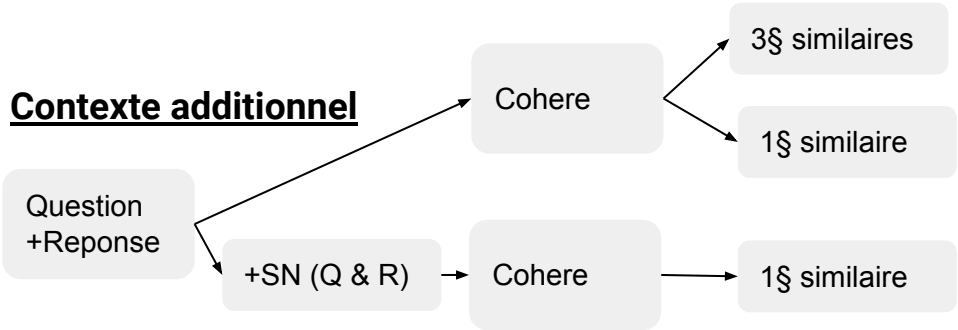
Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43

Approche multiclass

Hyp : Un contexte affiné et une approche contrastive peuvent améliorer le résultat.
Infirmée

Contexte additionnel



Modèle	Hamming	EMR
Camembert-base	33,80	14,31
Dr-bert-7GB	**39,08**	**17,68**
Dr-bert-7GB$_{\text{contexte}}$	35,31	15,27
Dr-bert-7GB$_{\text{contrast}}$	36,06	16,40

Approche contrastive

1) DrBERT Finetuning → QUAERO (NER) → annotation → freeze

2) Reconnaître entités ayant substituées des entités réelles

$$\mathcal{L}_{\text{contrastive}} = \mathbb{1}_{e \in E^+} \log P(e|C) + (1 - \mathbb{1}_{e \in E^+}) \log(1 - P(e|C))$$

Tâche annexe

Modèle	macro F1	Accuracy
LDA	13,26	19,13
DrBERT Multiclass contrastive	**31,51**	60,45
Régression logistique	27,98	**62,54**

$$\mathcal{L} = \mathcal{L}_{\text{classif}} + \lambda \mathcal{L}_{\text{contrastive}}$$

Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43

Approche multilabel

Hyp1 : Chaque classe/réponse doit être indépendante. (logique humaine)
Hyp2 : Le type peut contrôler l'approche multilabel.

Représentation :

CLS SEP Question SEP Reps

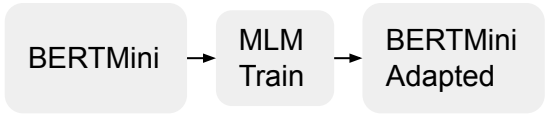
Approche multilabel

Hyp1 : Chaque classe/réponse doit être indépendante. (logique humaine)
Hyp2 : Le type peut contrôler l'approche multilabel.

Représentation :

CLS SEP Question SEP Reps

1) Adaptation du LM



Approche multilabel

Hyp1 : Chaque classe/réponse doit être indépendante. (logique humaine)
Hyp2 : Le type peut contrôler l'approche multilabel.

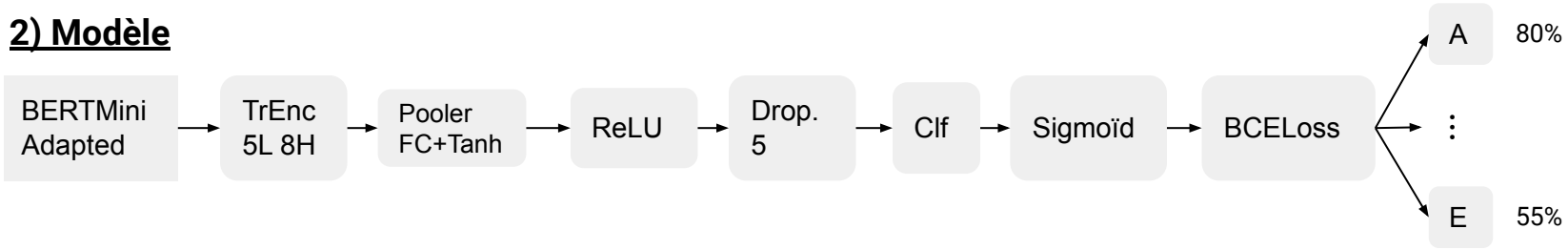
Représentation :

CLS | SEP | Question | SEP | Reps

1) Adaptation du LM

BERTMini → MLM Train → BERTMini Adapted

2) Modèle



Approche multilabel

Hyp1 : Chaque classe/réponse doit être indépendante. (logique humaine) **Infirmée**

Hyp2 : Le type peut contrôler l'approche multilabel. **Confirmée**

Représentation :

CLS SEP Question SEP Reps

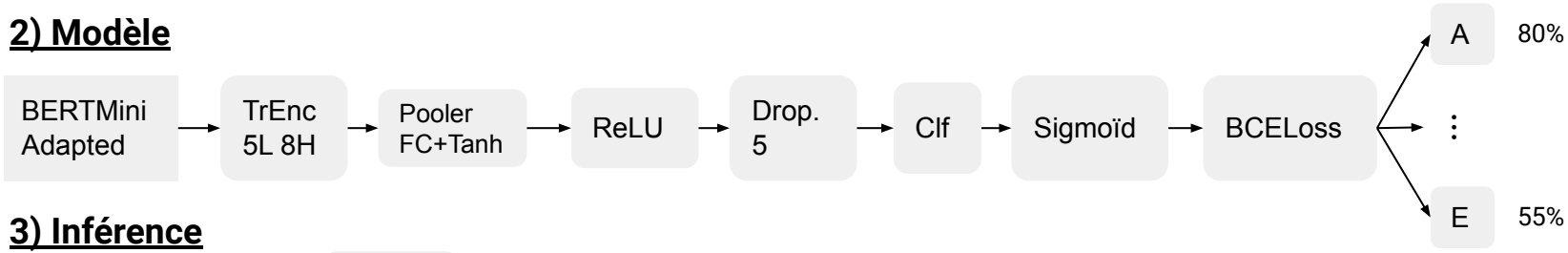
Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43
DAPT Multilabel avec type prédit (LR)	39,15	11,58

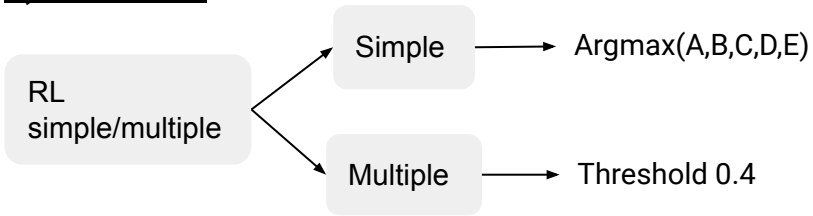
1) Adaptation du LM

BERTMini → MLM Train → BERTMini Adapted

2) Modèle



3) Inférence



Pour maximiser l'EMR !

Approche multilabel

Autres approches :

- Représentation Tf-IDF
- LightGBM, RandomForest, SGDClassifier, Regression Logistique…
- Stratégie OneVSRest

Scores très très bas (max ~0.27 Hamming Score)

Approches génératives

Hyp1: l'aléatoire comme score de confiance

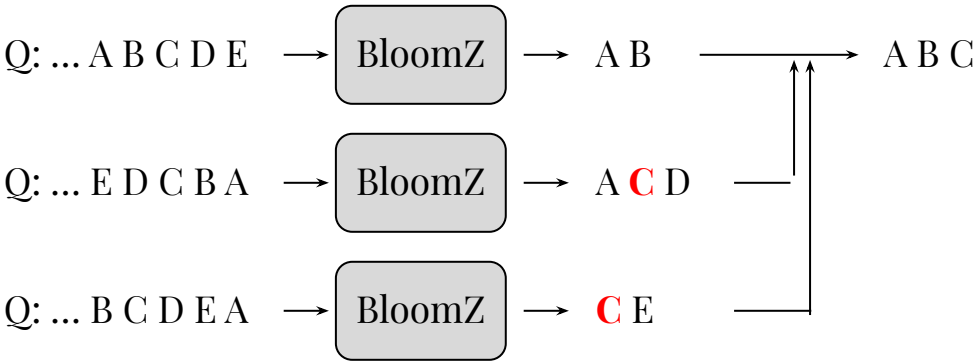
Hyp2: l'ajout d'info par prompt tuning améliore la prédiction

copie de la question

Question : ... Choix : (A) ... (B) ... (C) ... (D) ... (E) ... Réponses

ZeroShot:
1) générer les 20 tokens les + probables
2) Si une lettre majuscule apparaît elle devient une étiquette

3 exécutions pour pallier les réponses uniques :



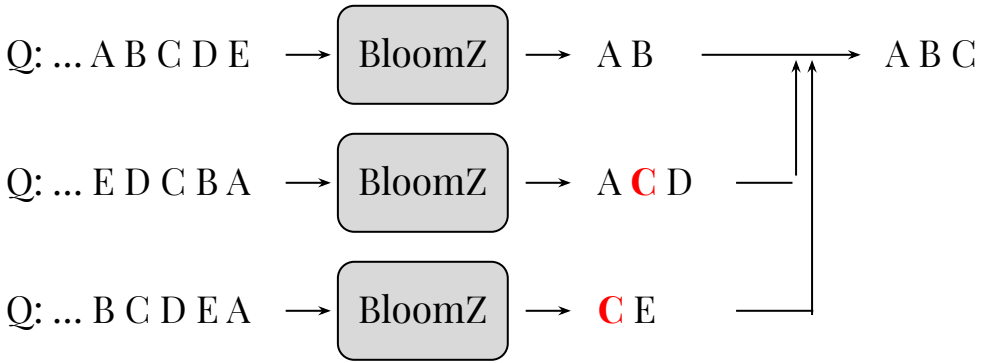
Approches génératives

copie de la question

Question : ... Choix : (A) ... (B) ... (C) ... (D) ... (E) ... Réponses

ZeroShot:
1) générer les 20 tokens les + probables
2) Si une lettre majuscule apparaît elle devient une étiquette

3 exécutions pour pallier les réponses uniques :



Hyp1: l'aléatoire comme score de confiance confirmée
Hyp2: l'ajout d'info par prompt tuning améliore la prédiction infirmée

Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43
DAPT Multilabel avec type prédit (LR)	39,15	11,58
BloomZ zero-shot	**41,54**	**23,95**

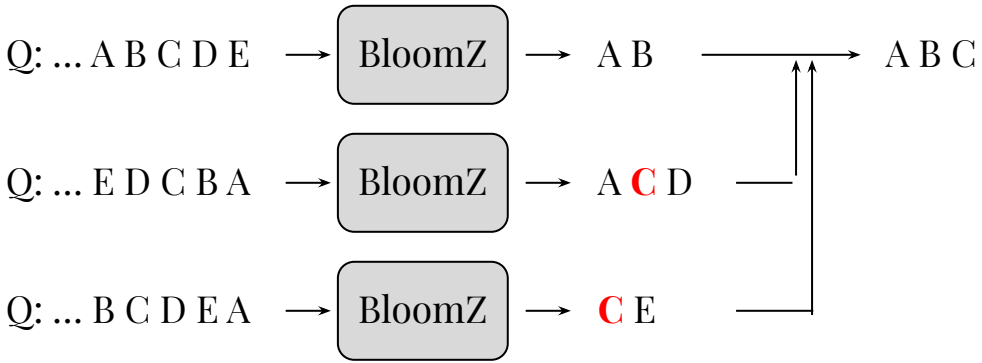
Approches génératives

copie de la question

Question : ... Choix : (A) ... (B) ... (C) ... (D) ... (E) ... Réponses

ZeroShot:
1) générer les 20 tokens les + probables
2) Si une lettre majuscule apparaît elle devient une étiquette

3 exécutions pour pallier les réponses uniques :



Hyp1: l'aléatoire comme score de confiance confirmée
Hyp2: l'ajout d'info par prompt tuning améliore la prédiction infirmée

Autres essais :
- in-context few-shot
- sans instruction fine-tuned: bloom7b1, vicuna-13b, ...
- variantes de prompts
- 1 à 5 phrases de contexte (cohere)
- variation du prompt, de langue, etc.
- soft-prompt tuning: 1 vecteur de params

Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43
DAPT Multilabel avec type prédit (LR)	39,15	11,58
BloomZ zero-shot	**41,54**	**23,95**

Limites

- **Aucune approche ne correspond** réellement à la tâche
  - Il faut une nouvelle loss
- **BloomZ zero-shot reste meilleur**
  - Alors qu'il prédit une étiquette
- **Threshold perfectible** (multilabel) (Abhishek & Hamarneh, 2021)
  - devrait être continu et appris
- **Pas de ranking**

Tâche principale

Modèle	Hamming	EMR
DrBERT Multiclass contrastive	37,22	15,43
DAPT Multilabel avec type prédit (LR)	39,15	11,58
BloomZ zero-shot	**41,54**	**23,95**

Tâche annexe

Modèle	macro F1	Accuracy
LDA	13,26	19,13
DrBERT Multiclass contrastive	**31,51**	60,45
Régression logistique	27,98	**62,54**

Merci

DEFT 2023

Défi Fouille de Textes@TALN 2023

Tal à Très Grande Vitesse

Contexte

Récupération des termes spécifiques :

- Lemmatisation et POS-tagging
- Conserver les syntagmes nominaux dont la freq < la freq moyenne (12)

Récupération des contextes :

Utilisation de Cohere pour récupérer des paragraphes de pages Wikipédia vectorisées (calcul de similarité) :

- Concaténation de la question et de la réponse pour retrouver les paragraphes les plus similaires pour chacune des cinq entrées ;
- Concaténation des **syntagmes nominaux** de la question et de la réponse pour retrouver le paragraphe le plus similaire pour chacune des cinq entrées.

Remarques : Les contextes restent trop longs.

→ Mieux cibler les paragraphes Wikipédia (contrôle du domaine)
→ Mieux cibler les extraits pertinents