Language Models as Knowledge Bases?

Ben Kabongo
M1 DAC, Sorbonne Université

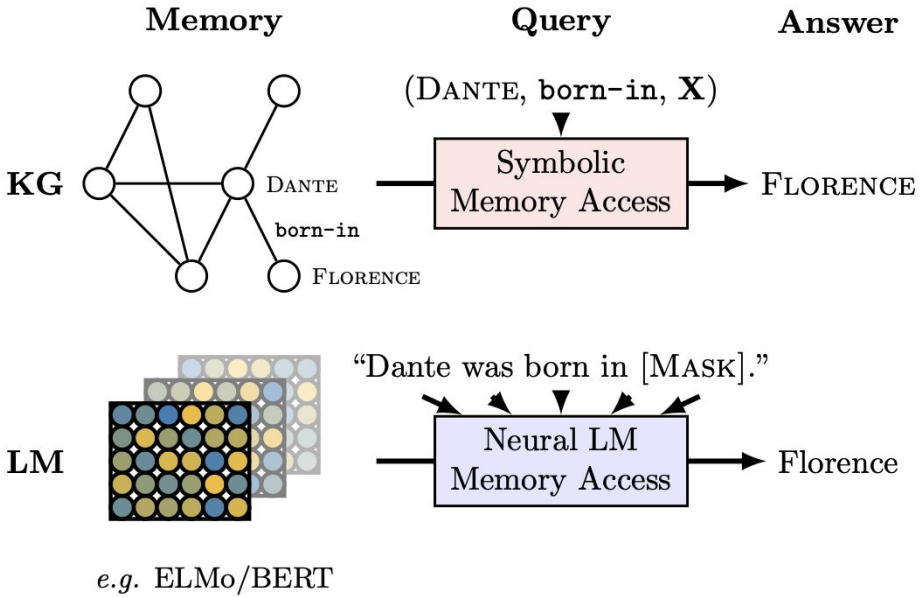
01/03/2022

Sommaire

- Problématique
- Modèles de langage
- LAMA : Language Models Analysis
  - Bases de connaissances
  - Baseline
  - Métriques
  - Choix d'implémentation
- Résultats
- Conclusions

Problématique

- **Avantages des modèles de langage par rapport aux bases de connaissance :**
  - Pas de supervision dans les modèles de langage : *pas de schéma*
  - Classe ouverte de relations
  - Facilement extensible



Les bases de connaissance et les grands modèles de langage peuvent répondre aux mêmes questions.

Problématique

- **Les modèles de langage peuvent-ils donc être utilisés comme des bases de connaissance ?**
  - **Connaissances relationnelles des modèles de langage : ELMo, BERT**
    - Quantité des connaissances relations des modèles
    - Différences avec d'autres types de connaissances : faits, bons sens, réponses aux questions
    - Comparaison des performances avec les bases de connaissance
- **Quel apport dans les tâches faisant intervenir le sens commun ? Apprentissage par renforcement ?**
  - Lien avec le projet :
    - **Les modèles de langage sont-ils capables de générer des instructions pour les robots ?**

Modèles de langages

Model	Based-model	#Param	#Corpus
fairseq-fconv	ConvNet	324M	103M
Transformer-XL	Transformer	257M	103M
ELMo (original)	BiLSTM	93.6M	800M
ELMo (5.5B)	BiLSTM	93.6M	5.5B
BERT (base)	Transformer	110M	3.3B
BERT (large)	Transformer	340M	3.3B

LAMA : LAnguage Model Analysis

- **Corpus de faits (tirés de bases de connaissances)**
  - **Fait :** triplet (*sujet*, *relation*, *objet*) ou couple (*question*, *réponse*)
  - **Transformation des données pour l'étude**
- Test des connaissances factuelles et de sens commun des modèles de langage
  - **Prédiction du jeton manquant**
    - (?, relation, objet) ; (sujet, ?, objet) ; (sujet, relation, ?)
    - (?, réponse) ; (question, réponse)
- **Evaluation des modèles**
  - **Classification du jeton manquant par rapport à un vocabulaire fixe.**

Bases de connaissances de référence

- **Google-RE (Sujet-Relation-Objet)**
  - **Relations considérées :**
    - place of birth, date of birth, place of death
- **T-REX (Sujet-Relation-Objet)**
  - 41 relations considérées
- **ConceptNet (Sujet-Relation-Objet)**
  - La base de connaissance que nous étudions dans ce projet
  - 16 relations considérées
- **SQuAD (Question-Réponses)**
  - Jeu de données de questions-réponses
  - 305 questions contextuelles

LAnguage Model Analysis : Baseline

- **Freq :** (Sujet-Relation-Objet)
  - Pour chaque (*sujet*, *rel*) indique la fréquence des (sujet=?, relation=*rel*, objet=*sujet*)
  - **Modèle qui pour une relation donnée prédit tout le temps les mêmes objets**
  - **Performance de la limite supérieure**
- **RE :** (Sujet-Relation-Objet)
  - **Modèle d'extraction de relations** : extraction des triplets étant donnée une phrase
  - **Entrées :** ensemble de faits -> **Sortie :** graphe de triplets
  - **Usage :** retrouver le sujet dans le graphe, classification des objets en fonction des scores donnés par RE
  - **Variantes :** différentes implémentations des liaisons d'entités
    - **REn :** correspondance exacte des chaînes
    - **REo :** correspondance exacte + oracle
- **DrQA :** (Question-Réponse)
  - **Prédit la réponse aux questions**
    - Recherche TF-IDF des documents pertinents
    - Extraction des réponses
    - **Contraintes sur la réponse des modèles :** un seul token

Métriques

- **Métriques basées sur le rang**
  - **Calcul par relation**
  - **Valeurs moyennes par relation**
- **Précision moyenne à k (P@k)**
  - Valeur à 1 si classification parmi les k premiers résultats
  - 0 sinon

LAMA : considérations et choix d'implémentation

- **Modèles définis manuellement**
  - Définition pour chaque relation d'un modèle d'interrogation de l'emplacement objet de la relation
  - Les bases de connaissance traditionnelles n'ont qu'une seule façon d'interroger les connaissances
    - *working-for* est complètement différent de *is-working-for*
  - **Mesure de la limite inférieure des connaissances des modèles de langage**
- **Token unique**
  - **La prédiction ne porte que sur un seul token**
- **Interrogation sur les emplacements d'objets**
  - Les interrogations portent sur les emplacements d'objets : (sujet, relation, ?)
    - Inclusion des relations inverses : *contient, contenu par*
- **Intersection des vocabulaires des modèles de langage**
  - **Taille du vocabulaire utilisé : 21K**

Résultats

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**
T-REx	1-1	937	2	1.78	-	0.6	10.0	17.0	36.5	10.1	13.1	68.0	**74.5**
	N-1	20006	23	23.85	-	5.4	**33.8**	6.1	18.0	3.6	6.5	32.4	34.2
	N-*M*	13096	16	21.95	-	7.7	**36.7**	12.0	16.5	5.7	7.4	24.7	24.3
	Total	34039	41	22.03	-	6.1	**33.8**	8.9	18.3	4.7	7.1	31.1	32.3
ConceptNet	Total	11458	16	4.8	-	-	-	3.6	5.7	6.1	6.2	15.6	**19.2**
SQuAD	Total	305	-	-	**37.5**	-	-	3.6	3.9	1.6	4.3	14.1	17.4

Conclusions

- Modèle le plus performant : **BERT-large**
- **Les performances d'extraction de relations puissent être difficiles à améliorer avec plus de données**
- **Extraction d’une base de connaissance non trivale**
- **Réponse à la problématique :**
    - **Les modèles de langage formés sur des grands corpus pourraient devenir une alternative viable aux bases de connaissances traditionnelles extraites du texte à l'avenir.**