

DL@IRIS

Activités en deep learning de l'équipe IRIS
José G Moreno
01/07/2021



Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Projets en cours

- ANR

- CoST (Porteur - 2 WP)
- MEERQAT (1 WP)
- LawBot (1WP)

- Contrats CIFRE

- Renault
- Synapse

- Participation projets EU

COST - ANR

Modelling Complex Search tasks

HOME PROJECT DATASETS


1. **Modeling patterns of search behaviour from user interactions:** understanding and modelling user behaviour within a complex search task is a key step for modelling the task itself. Unlike prior query-driven predictive models which rely on elementary user interaction facets (e.g., click models), we plan to mine high-level user behaviour patterns by jointly relating multiple observable user interactions (e.g., query reformulation, clicks) to both subtasks and task attributes (e.g., level of cognitive complexity) and user's cognitive context (e.g., domain knowledge). At the subtask level, we hypothesize that similar subtasks are performed.
2. **Learning representations of complex search tasks.** By analogy with the importance of query and document representation in traditional IR models, this step is fundamental for designing task-based information access models. Recent work tackled the problem of task representation from the perspective of discovering coherent successive queries in search sessions. Our perspective in the CoST project is radically different since we attempt to build the representations of tasks that support their completion based on system-driven assistance.
3. **Designing task-driven information access models.** We consider here the problem of matching information relevance with task completion. Only a very few and recent work tackled this challenge in the context of specific tasks. Our aim in the CoST project is to provide solutions to generic complex search tasks by relying on their learned representations and understanding of cognitive user's search abilities.

Partners



The Multimedia Entity Representation and Question Answering Tasks (MEERQAT) project is a collaborative project (2020-2024) funded by the [French National Research Agency \(ANR\)](#).

The project proposes to tackle the problem of **analyzing ambiguous visual and textual content by learning and combining their representations and by taking into account the existing knowledge about entities.**



[Learn more](#)

LawBot DESCRIPTION CONSORTIUM PUBLICATIONS ANNCES

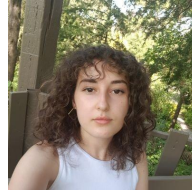
LAWBOT est un projet interdisciplinaire d'étude de la jurisprudence des tribunaux français, à l'interface entre les sciences juridiques, économiques et informatiques. Le projet vise des modèles computationnels prédictifs de la jurisprudence des tribunaux français. Ces modèles ont l'objectif de répondre à des requêtes en langage naturel décrivant des situations factuelles en formulant une prédiction sur des résultats judiciaires probables et en générant automatiquement un texte argumentatif motivant ces résultats du point de vue du droit.

Synapse



Thèses

- Lila Boualili (IR)
- Luis Lugo (User IR)
- Raphaël Sourty (KB)
- Nicolas Bizzozzero (User IR)
- Jesus Lovon Melgarejo (KB)
- Maxime Arens (Conversational IR)
- Alexis Dusart (Social IR)
- Nazish Hina (IR)
- Encadrants:
 - 2 PR et 4 MCF (75% de l'équipe)

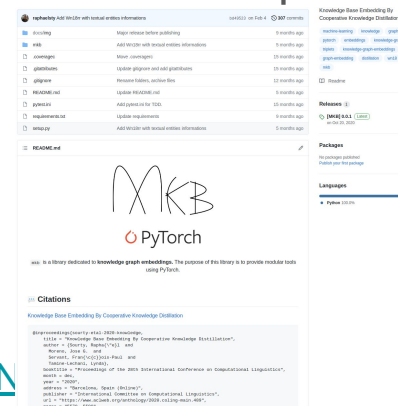


Sujets

- DL pour la **recherche d'information**
 - Modèles de recherche d'information
 - Représentation des comportements utilisateurs des moteurs de recherche
- DL pour la **représentation des informations structurées**
 - Représentation de bases des connaissances
 - Intégration des informations textuelles dans des bases de connaissances
- DL pour l'**extraction d'information**
 - Détection d'événements dans des documents textuelles
 - Identification des entités et liaison vers une base de connaissances

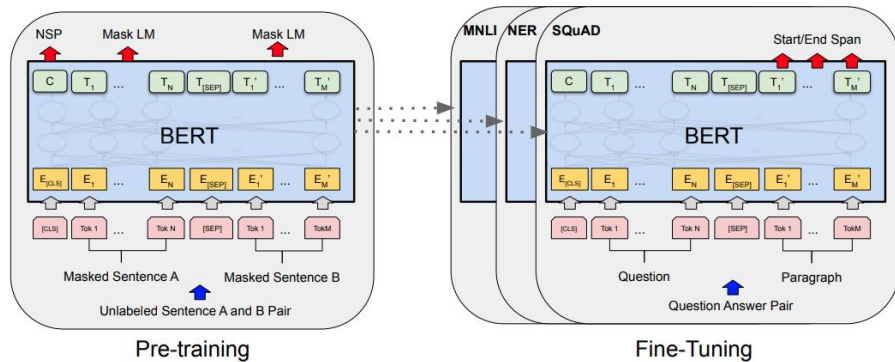
Publications et pépites

- Des publications (depuis 2019) dans des conférences/revues phares en RI et TAL (**A+**, **A**)
 - **SIGIR, ACL, ACMTOIS, ECIR, COLING, CoNLL, SAC,**
- 2 papiers longs + 1 papier court présentés à ECIR2021 sur des thématiques DL (7% de papiers longs à ECIR avec 1 auteur IRIS)
- Publication en code source ouvert
 - <https://github.com/raphaelsty/mkb> (43 étoiles)
 - https://github.com/BOUALILILila/markers_bert (16 étoiles)
 - <https://github.com/BOUALILILila/ExactMatchMarking> (2 étoiles)
 - <https://github.com/lelugom/search>
 - <https://github.com/Houssam93/Feature-Focus-in-Multi-Task-Learning-N>



Background (two works)

BERT



- One model to rule them all!

- Original
 - base - 110M
 - large - 330M
 - multilingual
- Language specific modes
 - FR - CamemBERT, FleuBERT
 - DE - dbmdz, deepset
- Large and public available models
- Fine-tuning is a classical transfer knowledge strategy

Addressed problems

- Information retrieval (reranking)
 - Passage retrieval
 - Ad-hoc retrieval



- Information extraction
 - Named entity recognition

Luke Rawlence **PERSON** joined Aiimi **ORG** as a data scientist in Milton Keynes **PLACE**, after finishing his computer science degree at the University of Lincoln. **ORG**

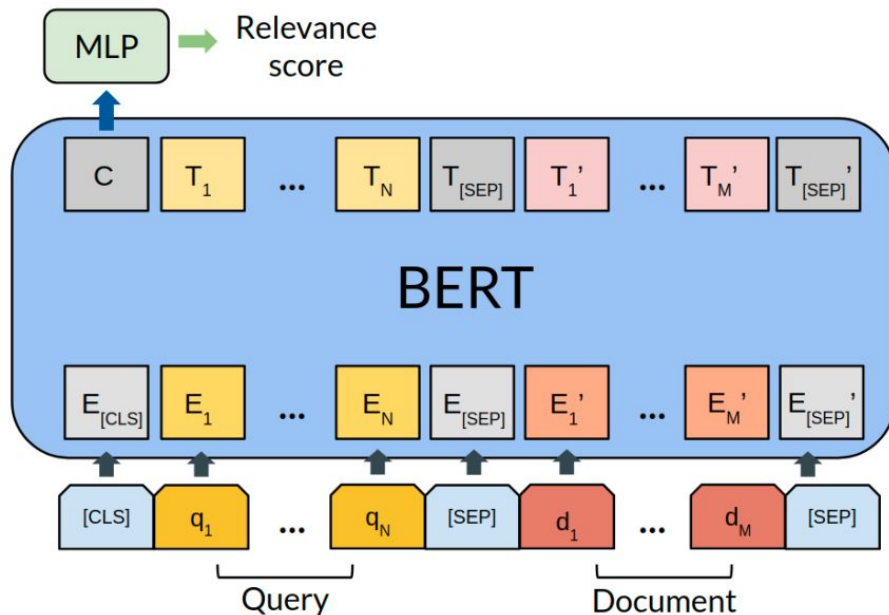
DL pour la recherche d'information

Context

- Context
 - Passage ranking task
 - BERT Pre-trained Language Model
 - Exact term-matching signals integration
- BERT manages to “correctly” model the semantic similarity, but **search is not only semantic matching**
 - Exact terms is a strong signal in IR but BERT fails to model it
 - Research Question
 - Can we **introduce exact term-matching signals** on BERT? May it help to improve its performance?

Methodology

- Base Model
 - We use Nogueira et al. model as our starting point
 - A single classification layer is added on top of BERT to perform pointwise classification



Methodology

- MarkedBERT
 - We use a simple marking technique to augment the model's input sequence with marker tokens to highlight exact term-matches in the query-passage pair
 - Standard input :
 - Query = “ghost meaning urban”
 - Passage = “ghost town an urban area with a fixed boundary that is smaller than a city...”
 - Input sequence:
 - [[CLS], ghost, meaning, urban, [SEP], ghost, town, an, urban, area, with, a, fixed, boundary,that, is, smaller, than, a, city,..., [SEP]]
 - Augmented input sequence:
 - [[CLS], ghost, meaning, urban, [SEP], **[e1]**, ghost, **[e1]**, town, an, **[e2]**, urban, **[e2]**, area, with, a, fixed, boundary,that, is, smaller, than, a, city,..., [SEP]]

Experimental Setup

- Dataset
 - MsMarco (<https://microsoft.github.io/msmarco/>)
 - A large-scale dataset obtained from no less than half a million queries sampled from Bing's search query logs
 - Each query has sparse relevance judgments by human editors
 - The passage ranking set contains 8.8M passages
- BERT configuration
 - BERT-base (12 layers, 768 hidden size, 12 heads, 110M parameters)
- Ranking pipeline
 - We use a two-stage ranking pipeline:
 - BM25 to produce an initial list of the top-1000 candidate passages per query
 - Base Model and MarkedBert are used to re-rank the candidate passages

Results on the dev set

Model	MRR@10 (%)
BM25 [11]	18.4
Doc2query + BM25 [11]	22.1
Base Model	30.2
MarkedBERT	32.8

- Both BERT-based models outperform traditional baselines by large margins
- MarkedBERT significantly outperforms the Base Model by about 9%

More IR datasets

Hybrid runs Model	Robust04					GOV2				
	nDCG@20		P@20		Field	nDCG@20		P@20		Field
Birch (MS-MB)	0.5137	+15.0%	0.4404	+12.0%	Title	0.5608	+9.4%	0.6409	8.1%	Title
BERT-MaxP (MS)	0.5453	+8.7%	0.4522	+9.3%	Desc	0.5600	+9.5%	0.6356	+9.0%	Title
Parade	0.5605	+5.7%	0.4661	+6.0%	Desc	0.5750	+6.7%	0.6530	+6.1%	Title
Parade (ELECTRA)	0.5713	+3.7%	0.4717	+4.8%	Desc	0.5851	+4.8%	0.6678	+3.7%	Title
T5-base	0.5298	+12.0%	-	-	Hybrid	-	-	-	-	-
T5-large	0.5345	+11.0%	-	-	Hybrid	-	-	-	-	-
T5-3B	0.6091	-2.7%	-	-	Hybrid	-	-	-	-	-
Sim-Pair <small>BERT</small>	0.5701	-	0.4815	-	Hybrid	0.5998	-	0.6758	-	Hybrid
Sim-Pair <small>ELECTRA</small>	0.5927	-	0.4942	-	Hybrid	0.6133	-	0.6926	-	Hybrid

- We also successfully demonstrate the power of MarkedBERT (Sim-Pair) on traditional ad-hoc IR datasets (competitive results w.r.t. SoTA)

Take-away messages

- **Exact term matching highlight can benefit state-of-the-art** passage reranking model based on BERT
- Introducing marker tokens to the input sequence **induces more focus on the exact matched terms** considered important for relevance matching
- Augmenting the input sequence with marker tokens can be used to highlight different terms and **provide BERT-like models with additional information** to better accomplish other tasks

https://github.com/BOUALILILila/markers_bert

DL pour l'extraction d'information

NER in Historical Documents

- Named Entity Recognition (NER):
 - Identifies the named entities and classifies them under various predefined classes
- Digitization errors:
 - damaged documents, low quality scans, historic fonts
 - spelling or orthographical variants, inflected forms
- Segmentation format:
 - line-level or article-level
- Multilingualism:
 - English, French, German



ÉCHOS

L

a promotion dans l'ordre de la légion
d'honneur relative aux Expositions de
Milan et de Dusseldorf paraîtra vendredi
à l'Officiel.

Une fête au Collège de France.

Sous la présidence de M. Doumergue,
ministre de l'instruction publique, et du
comte Tomielli, ambassadeur d'Italie, dans
la grande salle du Collège de France aura
lieu le 15 mars prochain, à deux heures pré-
cises, une fête en l'honneur de Carducci, le
grand poète italien.

Major Challenge

- Despite the adoption of BERT-based models, these models struggle to adapt to new content (historical documents)

Gold Standard

Allemagne que pas iJKeaz Schwietz le bour,-ijeu de Br eslasi. _Piappi les pliante fameux de Reindel_ figurjal la femme _JVie & e, l'horrible mégère de Hambourg, qui assassina une vingtaine d'enfents confiés à _j ses soins mercenaires.

BERT

Allemagne que pas iJKeaz Schwietz le bour,-ijeu de Br eslasi. _Piappi les pliante fameux de Reindel_ figurjal la femme _JVie & e, l'horrible mégère de Hambourg, qui assassina une vingtaine d'enfents confiés à _j ses soins mercenaires.

Gold Standard

Amiens werde zwar im Augenblick noch gehalten, aber der Entwiclungsangriff von Lille aus, also vo^i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vor auszusehen.

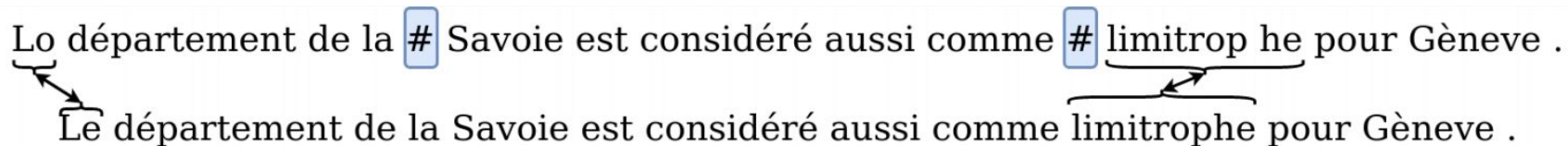
BERT

Amiens werde zwar im Augenblick noch gehalten, aber der Entwiclungsangriff von Lille aus, also vo^i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vor auszusehen.

A straightforward solution: Preprocessing

- An example of a French instance from the training data

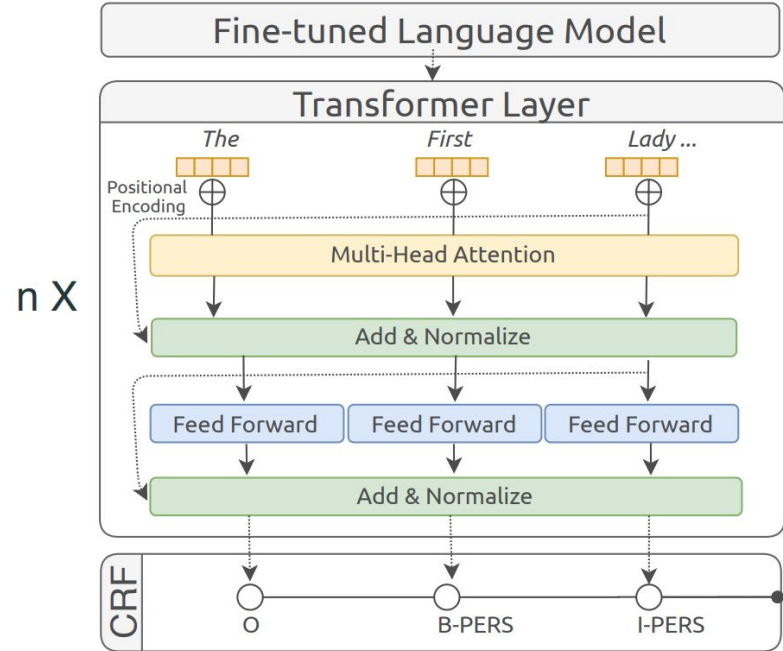
Lo département de la # Savoie est considéré aussi comme # limitrop he pour Genève .
Le département de la Savoie est considéré aussi comme limitrophe pour Genève .



- line-level context would have been too short to grasp
- reconstructed the original text, including hyphenated words
- Freeling 4.1 (Padró and Stanilovsky, 2012)

Our robust NER model

- BERT-based model with additional Transformers
 - stack of identical layers
 - multi-head self-attention mechanism
 - position-wise fully connected feed-forward network
- Our hypothesis is that extra layers will add enough flexibility to the model to adapt to new content



Existing Datasets

- HIPE Dataset
 - high-level entity types, Person, Location, Organization, Product, and Time.
 - CLEF 2020 Evaluation Lab HIPE challenge (Ehrmann et al., 2020)
 - German articles (1790-1940) and French articles (1790-2010)
- Our dataset
 - high-level entity types, Person, Location, Organization, and Product.
 - historical newspapers in French (1814-1944) and German (1845-1945)
 - IAA scores: 0.90 for French and 0.91 for German

<https://zenodo.org/record/4573313#.YN2nMYNfjdE>

Baselines

- German
 - BiLSTM-CNN: FastText embeddings
 - BiLSTM-CNN (transfer learning): FastText embeddings + de-GermEval (German Wikipedia and News Corpora)
 - bert-base-german-europeana
- French
 - BiLSTM-CNN: FastText embeddings
 - BiLSTM-CNN (transfer learning): FastText embeddings + fr-WikiNER (French Wikipedia)
 - camembert-large

Results on two multilingual datasets (FR,DE)

	DE			FR			DE			FR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CNN												
fuzzy	83.3	70.1	76.1	89.9	83.9	86.8	81.2	42.4	55.7	82.2	77.2	79.6
strict	69.4	58.4	63.4	77.7	72.5	75.0	54.8	28.6	37.6	65.5	61.4	63.4
BiLSTM-CNN (transfer learning)[†]												
fuzzy	81.1	75.0	77.9**	87.8	88.8	88.3	76.4	49.4	60.0**	83.6	77.8	80.6*
strict	67.4	62.2	64.7**	77.3	78.2	77.7	48.6	31.4	38.1**	66.9	62.3	64.5*
BERT												
fuzzy	83.4	88.3	85.8**	89.5	91.9	90.7*	60.1	67.0	63.4**	86.1	81.8	83.9**
strict	74.1	78.5	76.2**	81.1	83.3	82.1*	46.8	52.2	49.4**	70.1	66.6	68.3**
BERT+1×Transf												
fuzzy	85.8	87.3	86.5**	91.3	92.9	92.1**	82.3	66.4	73.5**	88.7	82.1	85.3**
strict	77.2	78.6	77.9**	83.5	84.9	84.2**	62.7	50.6	56.0**	74.4	68.9	71.5**
BERT+2×Transf												
fuzzy	87.0	87.2	87.1**	91.5	92.4	91.9**	83.3	64.4	72.6**	89.7	80.1	84.7**
strict	78.6	78.7	78.7**	83.4	84.2	83.8**	64.9	50.2	56.6**	75.0	67.0	70.8**

We ranked top-1 in the CLEF2020 task!

Take-away messages

- $n \times$ Transformer: increase the ability of the architecture to **better model** long-range contexts and historical documents specificities
- **BERT-alone** has **unbalanced attention** to misspelled or corrupted words when the most informative words contain such errors
- The approach **does not degrade** the results for **modern datasets****

Gold Standard	
BERT	
BERT+n x transf	
Gold Standard	
BERT	
BERT+n x transf	

An example of NER predictions

Questions

Per Query Type Analysis

- Classify queries based on the lexical answer type using a rule-based classifier (<https://github.com/superscriptjs/qtypes>)

Type	NUM	HUM	LOC	DESC	ENTY	ABBR
#Queries	1039	500	527	2063	363	9
Base Model	29.1	35.6	42.0	33.2	30.5	49.4
MarkedBERT	33.0	38.4	44.6	35.0	31.5	47.9
Δ MRR@10(%)	13.4	07.8	06.2	05.4	03.2	-0.03

- MarkedBERT performs better than the Base Model across all query answer types
- Exact term-matching highlight benefits the most the numerical answer type queries (NUM)
- Slight deterioration of the abbreviation answer type queries compared to the Base Model (ABBR)