

# Extraction de relation via la validation de relation

**Jose G. Moreno\***, Rashedur Rahman, Charlotte Rudnik, Cong Wang, Brigitte Grau

\*IRIT et LIMSI

02/12/2019

# Bases de connaissance en recherche d'information

[Tous](#) [Images](#) [Actualités](#) [Vidéos](#) [Maps](#) [Plus](#) [Paramètres](#) [Outils](#)

Environ 181 000 000 résultats (0,54 secondes)

## Barack Obama — Wikipédia

[https://fr.wikipedia.org/wiki/Barack\\_Obama](https://fr.wikipedia.org/wiki/Barack_Obama)

Barack Hussein Obama II [baˈʁɑk huˈsɛm oʊˈbɑːmə], né le 4 août 1961 à Honolulu (Hawaï), est un homme d'État américain. Il est le 44<sup>e</sup> président ...

Date de naissance: 4 août 1961 (57 ans) Nom de naissance: Barack Hussein Obama II Religion: Protestantisme Vice-président: Joe Biden

Vous avez consulté cette page de nombreuses fois. Date de la dernière visite : 16/01/19

### Barack Obama, Sr.

Tous les demi-frères et sœurs sont les enfants ... Barack Obama Sr ...

### Ann Dunham

Stanley Ann Dunham, née le 29 novembre 1942 à Wichita et ...

### Michelle Obama

Michelle LaVaughn Obama, née Michelle LaVaughn Robinson le ...

### Malia Obama

Malia Ann Obama, née le 4 juillet 1996 à Chicago (Illinois), est la ...

[Autres résultats sur wikipedia.org](#)

## The Office of Barack and Michelle Obama

<https://barackobama.com/> Traduire cette page

Welcome to the Office of Barack and Michelle Obama. We Love You Back. Play video. The Office of Barack and Michelle Obama. © 2017 | Legal & Privacy.

### À la une

**Barack Obama : son ex-conseiller Alan Krueger se suicide à 58 ans**  
[La Nouvelle Tribune](#) - il y a 2 jours

**L'économiste Alan Krueger, ancien conseiller de Clinton et d'Obama, est mort**  
[Les Échos](#) - il y a 1 jour

**Michelle Obama en admiration devant ses deux filles : "Je leur tire mon chapeau"**  
[La Libre.be](#) - il y a 1 jour

➔ [Plus de résultats pour "barack obama"](#)

## Barack Obama

44<sup>e</sup> président des États-Unis

Barack Hussein Obama II [baˈʁɑk huˈsɛm oʊˈbɑːmə], né le 4 août 1961 à Honolulu, est un homme d'État américain. Il est le 44<sup>e</sup> président des États-Unis, en fonction du 20 janvier 2009 au 20 janvier 2017. [Wikipédia](#)

**Mandat présidentiel** : 20 janvier 2009 – 20 janvier 2017 [Tendances](#)

**Date et lieu de naissance** : 4 août 1961 (Âge: 57 ans), centre médical pour les femmes et les enfants de Kapiolani, Honolulu, Hawaï, États-Unis

**Taille** : 1,85 m

**Épouse** : Michelle Robinson-Obama (m. 1992)

**Parents** : Ann Dunham, Barack Obama, Sr.

### Livres

Voir d'autres éléments (plus de 45)

<b>Les Rêves de mon père</b> 1995	<b>L'Audace d'Espérer</b> 2006	<b>Lettre à mes filles</b> 2010	<b>Le changement... nous pou...</b> 2008	<b>Barack Obama</b> 2011

### Recherches associées

Voir d'autres éléments (plus de 15)

<b>Donald Trump</b>	<b>Michelle Robinson...</b> Épouse	<b>Hillary Clinton</b>	<b>George W. Bush</b>	<b>Recep Tayyip Erdoğan</b> Tendances

[Revenir à cette fiche info](#) [Commentaires](#)

Plus de 50% du contenu affiché provient d'une base de connaissance !

# Agenda

Contexte

Un modèle pour la validation de relation

Extraction de relation via la validation de relation

Extraction de relation avec BERT

Conclusion

# Agenda pour la section 1

## Contexte

- Extraction d'information
- Extraction de relation
- Validation de relation
- Représentation d'entités

## Un modèle pour la validation de relation

- Modèle basé sur les CNNs
- Mécanisme d'attention
- Notre modèle pour la validation de relation
- Expériences et résultats

## Extraction de relation via la validation de relation

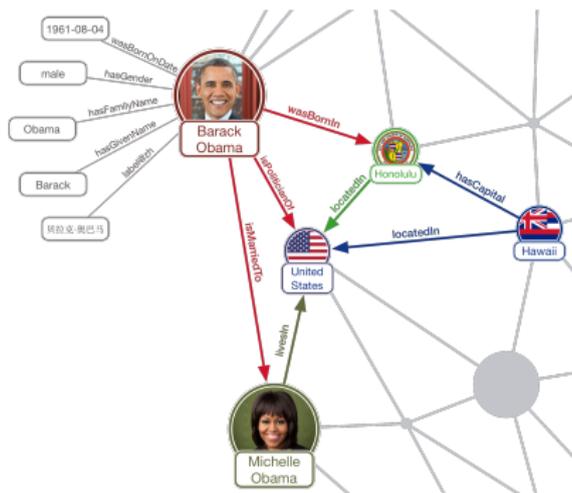
- Extraction de relation avec un CNN
- Expériences et résultats

## Extraction de relation avec BERT

- BERT
- Extraction de relation avec BERT
- Extraction de relation avec BERT via la validation de relation
- Expériences et résultats

## Conclusion

# Extraction d'information

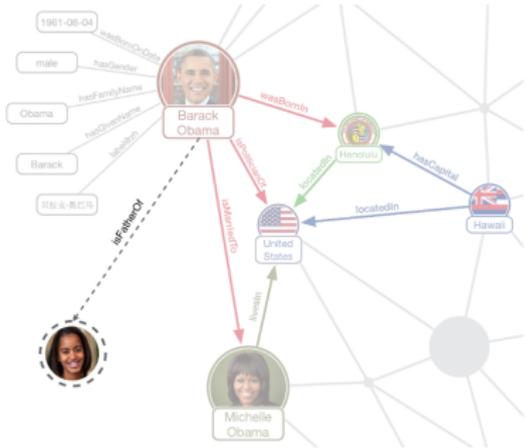


## Population de bases de connaissances

- ▶ Reconnaissance d'entités nommées
- ▶ Disambiguation/Normalisation d'entités
- ▶ **Extraction de relations**
- ▶ **Validation de relations**

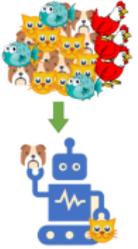
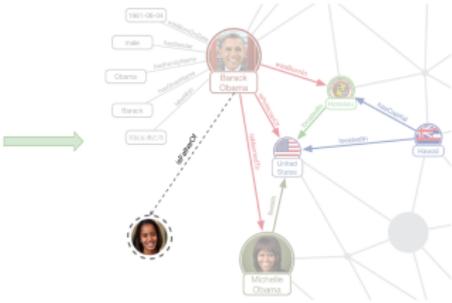
# Extraction de relation

The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and **Sasha**, 17, through IVF.

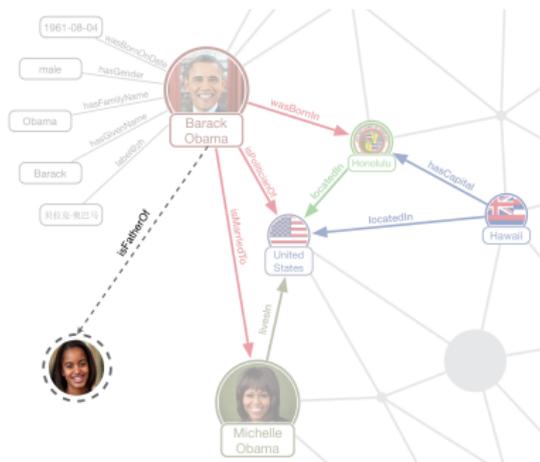


# Extraction de relation

The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and **Sasha**, 17, through IVF.



# Validation de relation



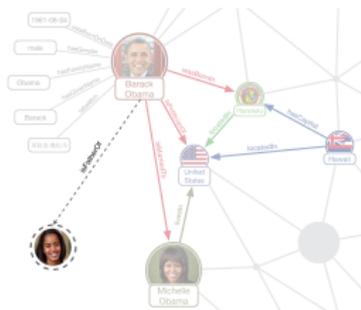
Since arriving on the island, **Barack**, Michelle, **Malia** and Sasha have gone on a hike at the Makiki Loop Hawaii Nature Center, seen President Obama's childhood school and dined at Morimoto restaurant.



The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and Sasha, 17, through IVF.



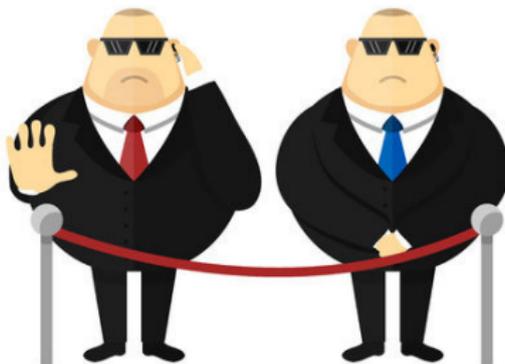
# Validation de relation



Since arriving on the island, **Barack**, **Michelle**, **Malia** and **Sasha** have gone on a hike at the Makiki Loop Hawaii Nature Center, seen President Obama's childhood school and dined at Morimoto restaurant.



The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and **Sasha**, 17, through IVF.



# Extraction vs Validation de Relation

## Extraction de relation

- ▶ Entrée : un texte
  - ▶ ***Steve Jobs lived in California***
- ▶ Sortie : Relation dans le texte parmi un ensemble fermé de candidats
  - ▶ *Relation "inhabit"*

## Validation de relation

- ▶ Entrée : Un triplet donné et un texte
  - ▶ *<Steve\_Jobs, death\_place, California>*
  - ▶ ***Steve Jobs lived in California***
- ▶ Sortie : Vérification sur la pertinence du texte pour le triplet
  - ▶ *Non validé*

# Modèle Extended Anchor Text (EAT)



## Caractéristiques

- ▶ EAT permet une représentation jointe des mots et des entités avec des techniques classiques de plongement sémantique
- ▶ EAT utilise un corpus annoté comme Wikipédia

Moreno et al.,(ESWC 2017)

# Représentations mots vs entités

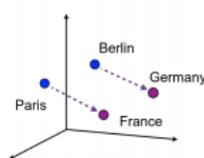


# Caractéristiques techniques du modèle EAT

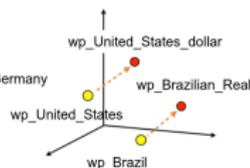
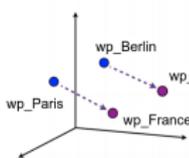
## Vecteurs des mots et d'entités pré-entraînés

- ▶ 200 dimensions
- ▶ Vocab. de 5.2M = 1.8M d'entités et 3.4M de mots (tokens)

Ent Acc.	EAT.NOE	EAT	word2vec	glove
<i>ALL<sub>Sem</sub> - Family</i>	0.5870	<b>0.6688</b>	0.6503	0.6649



Word



Entity

# Agenda pour la section 2

## Contexte

- Extraction d'information
- Extraction de relation
- Validation de relation
- Représentation d'entités

## Un modèle pour la validation de relation

- Modèle basé sur les CNNs
- Mécanisme d'attention
- Notre modèle pour la validation de relation
- Expériences et résultats

## Extraction de relation via la validation de relation

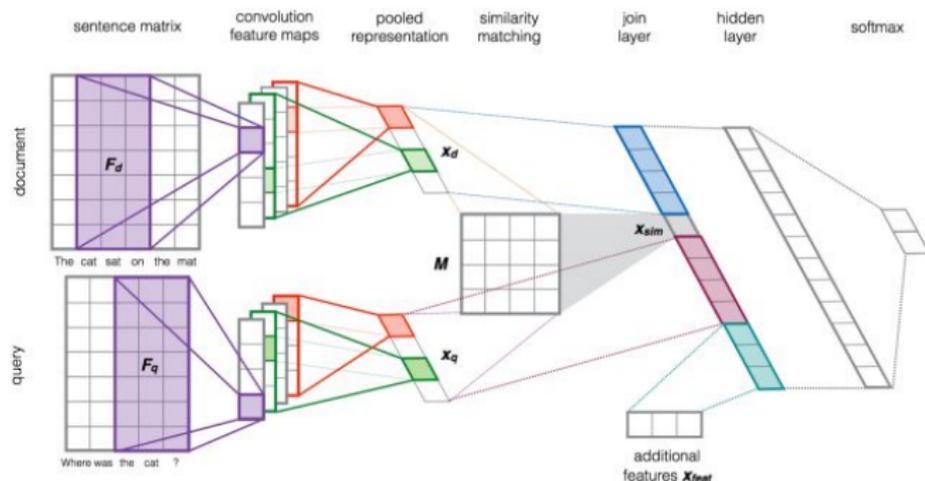
- Extraction de relation avec un CNN
- Expériences et résultats

## Extraction de relation avec BERT

- BERT
- Extraction de relation avec BERT
- Extraction de relation avec BERT via la validation de relation
- Expériences et résultats

## Conclusion

# Short-text Matching CNN (SMCNN)

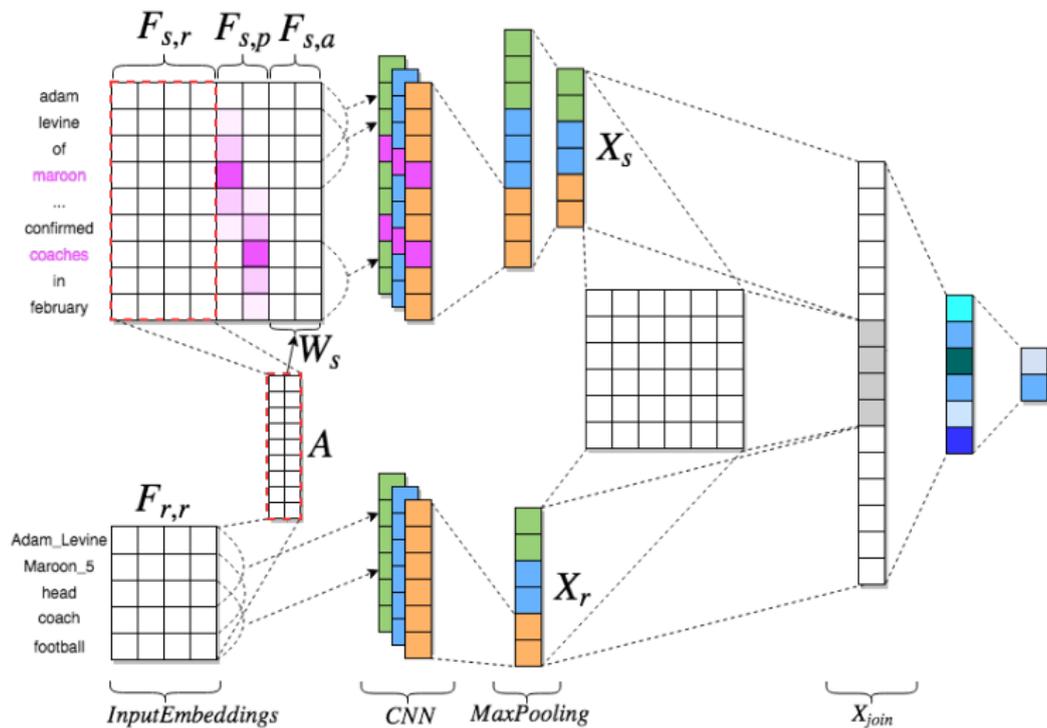


SMCNN - Severyn et Moschitti, (SIGIR 2015)

- ▶ Utilisé en Question-Réponse et Sélection de réponses
- ▶ La similarité  $sim(x_q, x_d) = x_q^T M x_d$ , où  $M$  est une matrice de similarité
- ▶ Adaptation :
  - ▶ Document : un texte à vérifier
  - ▶ Query : un triplet de relation

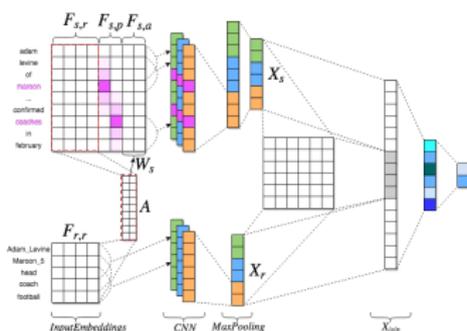


# ABSMCNN



Notre modèle

Moreno et al., (CORIA 2019)



Notre modèle

Moreno et al., (CORIA 2019)

- ▶ ABSMCMNN-POS et ABSMCMNN : correspondent au modèle décrit, mais ce dernier n'inclut pas  $F_{s,p}$ .
- ▶ SMCNN-POS et SMCNN : similaires aux modèles décrit précédemment (avec et sans POS), mais les deux n'incluent pas la matrice  $A$  et par conséquent  $F_{s,a}$ .

# Jeux de données

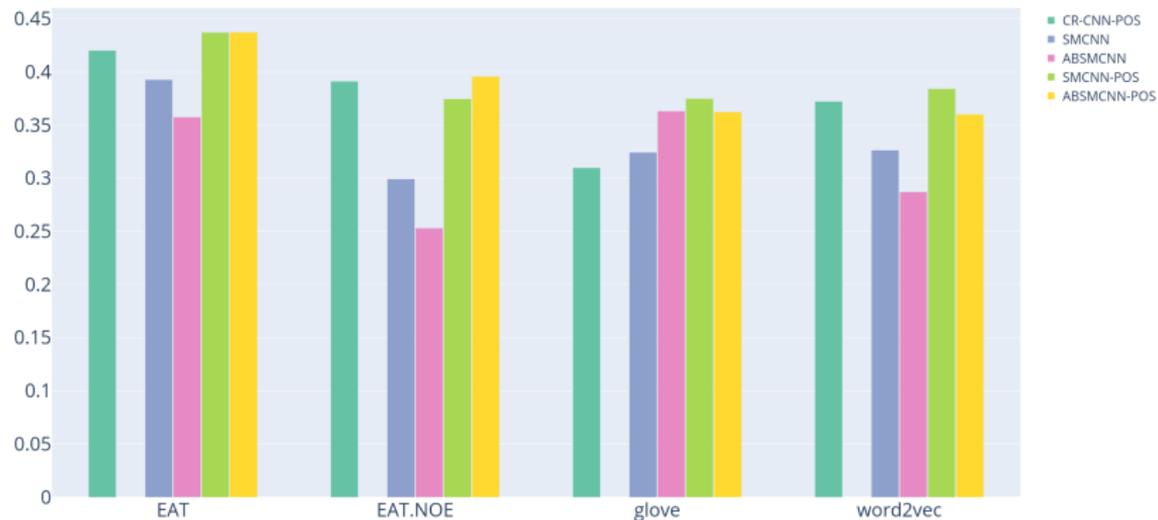
- ▶ dataKBP : un sous ensemble du corpus de TAC-KBP CSSF (Cold Start Slot Filling) des années 2015 et 2016.
- ▶ dataQA : jeu de données construit à partir de WebQuestions

## Exemple

- ▶ Phrase : The district gave 56% of its vote to democratic party nominee Barack Obama and 43% to republican party nominee John McCain
- ▶ Triplet : (Barack\_Obama, /government/politician/party, Democratic\_Party) -> (ETA\_Barack\_Obama, "government politician party", EAT\_Democratic\_Party)

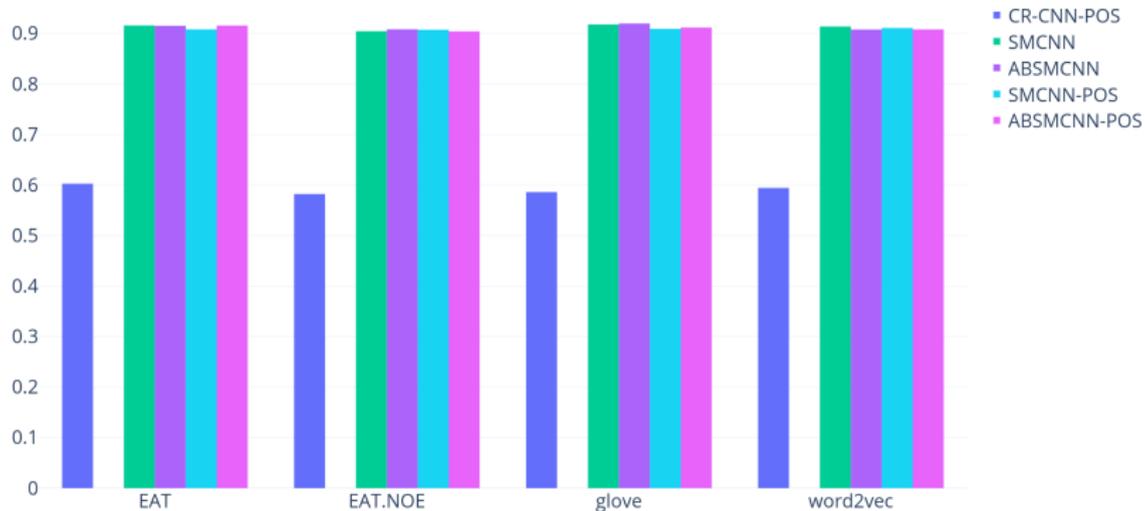
		# Positif	# Négatif	# Total	# Relation
dataQA	Entraînement	4364	4364	8,728	501
	Test	925	925	1,850	311
dataKBP	Entraînement	5,884	8,795	14,679	12
	Test	1,109	4,827	5,936	12

# Résultats sur dataKBP



Résultats en termes de  $F_1$  pour le jeu de données dataKBP

# Résultats sur dataQA



Résultats en termes de  $F_1$  pour le jeu de données dataQA

# Agenda pour la section 3

## Contexte

- Extraction d'information
- Extraction de relation
- Validation de relation
- Représentation d'entités

## Un modèle pour la validation de relation

- Modèle basé sur les CNNs
- Mécanisme d'attention
- Notre modèle pour la validation de relation
- Expériences et résultats

## Extraction de relation via la validation de relation

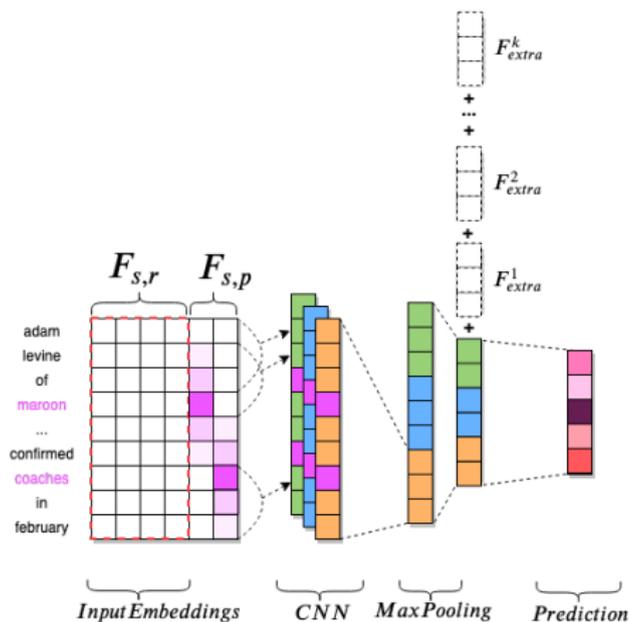
- Extraction de relation avec un CNN
- Expériences et résultats

## Extraction de relation avec BERT

- BERT
- Extraction de relation avec BERT
- Extraction de relation avec BERT via la validation de relation
- Expériences et résultats

## Conclusion

# CNN extraction de relation



CNN pour l'extraction de relation

- ▶ CNN adapté de Zeng et al., (COLING 2014)
- ▶ Adaptation : Utilisation de  $F_{extra}^i$  en ajoutant les dernières couches de la validation de relation.

# Jeux de données

Le corpus **SemEval10** tâche 8 qui est largement utilisé pour l'extraction de relation

## Exemple

- ▶ There were apples, **pears** and oranges in the **owl**
- ▶ (content-container, pears, owl)

Le corpus **dataQA** transformé pour l'extraction de relation (relations regroupés par la valeur de  $r$ , les exemples négatives sont regroupés dans la classe "other").

Dataset	# Examples	# Relations
SemEval10	Train 8000	19 <sub>(8*2+1)</sub>
	Test 2717	
dataQA	Train 8728	284
	Test 1850	

# Exemples SemEval10 - Message-Topic(e1,e2)

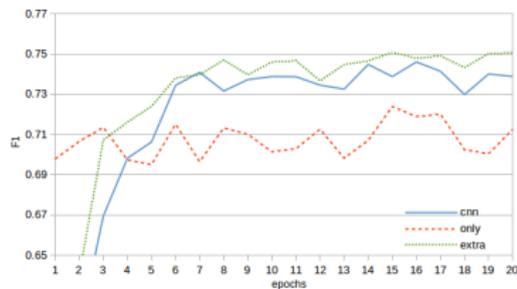
1. The final **[programme]<sub>e1</sub>** **detailed** the **[history]<sub>e2</sub>** of Russborough House
2. The **[letter]<sub>e1</sub>** **contains** a description of the **[demolition]<sub>e2</sub>** of the old synagogue
3. On 17 May 2005, the committee held a **[hearing]<sub>e1</sub>** **concerning** specific **[allegations]<sub>e2</sub>**
4. The **[newsletter]<sub>e1</sub>** **tells** of practical **[projects]<sub>e2</sub>** developed to help those affected by the pandemic

# Résultats sur SemEval10 et dataQA

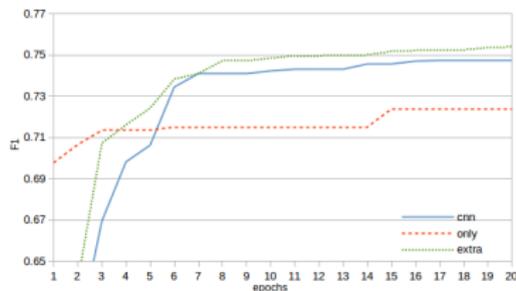
	SemEval10		dataQA	
	$F_1$	Accuracy	$F_1$	Accuracy
$F_{extra}^i$	0.7239±0.0002	0.7536±0.0002	0.0000±0.0000	0.5000±0.0000
<i>CNN</i>	0.7474±0.0034	0.7777±0.0023	0.0524±0.0019	0.5049±0.0015
<i>EXTRA</i>	<b>0.7539±0.0027</b>	<b>0.7825±0.0019</b>	<b>0.0579±0.0029</b>	<b>0.5063±0.0014</b>

- ▶  $F_{extra}^i$  est un NN simple qui utilise uniquement les caractéristiques du modèle de validation
- ▶ *CNN* est un NN basée les convolutions utilisé comme baseline
- ▶ *EXTRA* est un NN basée les convolutions auquel nous ajoutons les caractéristiques du modèle de validation

# Résultats sur SemEval10 par époques



$avg(F_1)$



$avg(F_1^{max})$

Notre architecture **EXTRA** obtient  $avg(F_1^{(9+1)}) = 0.8242$  valeur qui dépasse l'architecture **CNN**  $avg(F_1^{(9+1)}) = 0.8190$

# Agenda pour la section 4

## Contexte

- Extraction d'information
- Extraction de relation
- Validation de relation
- Représentation d'entités

## Un modèle pour la validation de relation

- Modèle basé sur les CNNs
- Mécanisme d'attention
- Notre modèle pour la validation de relation
- Expériences et résultats

## Extraction de relation via la validation de relation

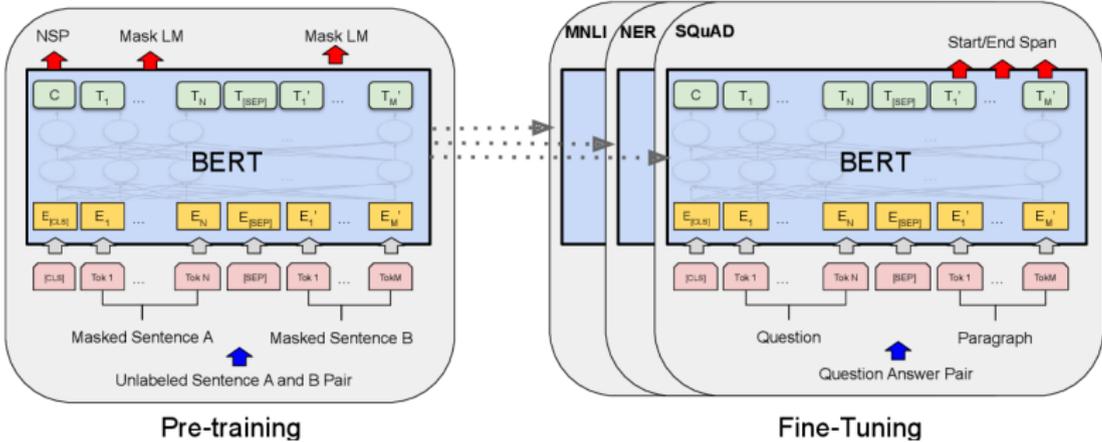
- Extraction de relation avec un CNN
- Expériences et résultats

## Extraction de relation avec BERT

- BERT
- Extraction de relation avec BERT
- Extraction de relation avec BERT via la validation de relation
- Expériences et résultats

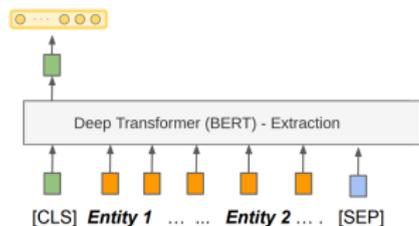
## Conclusion

# BERT : une révolution en TAL



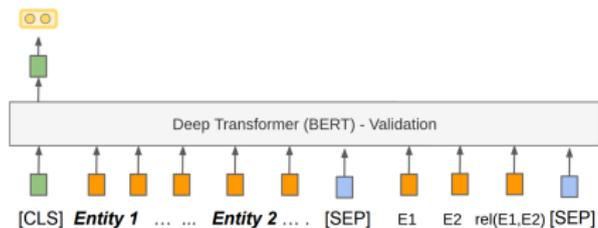
Devlin et al., (NAACL 2019)

# BERT pour la extraction et validation de relation



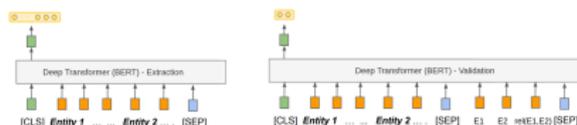
Extraction de relation

Baldini Soares et al., (ACL 2019)



Validation de relation

# BERT pour la extraction et validation de relation



## Entraînement

- ▶ 5 exemples négatives
- ▶ Toutes les classes sont traitées de la même façon (même "Other")
- ▶ Le nom de chaque relation est utilisé comme  $rel(E1,E2)$ . Par exemple,  $content-container(E1,E2) \rightarrow content\ container$

## Prédiction

- ▶ Chaque candidat est évalué et on garde celui avec la probabilité de prédiction la plus élevée

# Résultats sur SemEval10

	SemEval10		
	$F_1$	Accuracy	$F_1^{(9+1)}$
$F_{extra}^i$	0.7239	0.7536	-
CNN	0.7474	0.7777	0.8190
EXTRA (nous)	0.7539	0.7825	0.8242
BERT-EM+MTB* (ACL,2019)	-	0.8334 (0.8425)	0.8703 (0.8770)
BERT-er+vr (nous)	-	<b>0.8472</b>	<b>0.8831</b>

\*Notre implémentation de Baldini Soares et al. (ACL 2019)

# Résultats détaillés

		<i>Accuracy</i>	$F_1^{(9+1)}$
Total exemples dans le teste	2717	-	
Exemples avec une seule prédiction	1925	0.7085	~0.74
Exemples avec la bonne réponse dans le pool	2512	0.9245	~0.96

Avec une meilleur validation les résultats peuvent être encore améliorés.

	Nombre de candidats					
	2		3		4	
	True	False	True	False	True	False
BERT-re+rv	338	154	37	52	2	4
	68.69%	31.30%	41.57%	58.42%	33.33%	66.66%

Notre modèle a plus de mal quand le nombre de candidats augmente.

# Agenda pour la section 5

## Contexte

- Extraction d'information
- Extraction de relation
- Validation de relation
- Représentation d'entités

## Un modèle pour la validation de relation

- Modèle basé sur les CNNs
- Mécanisme d'attention
- Notre modèle pour la validation de relation
- Expériences et résultats

## Extraction de relation via la validation de relation

- Extraction de relation avec un CNN
- Expériences et résultats

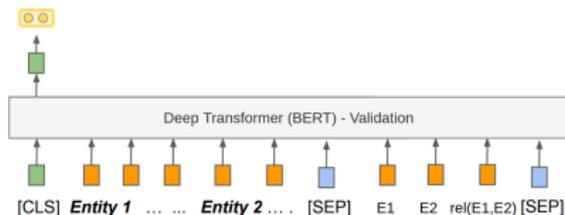
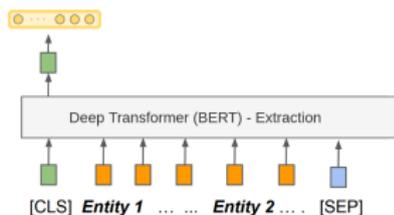
## Extraction de relation avec BERT

- BERT
- Extraction de relation avec BERT
- Extraction de relation avec BERT via la validation de relation
- Expériences et résultats

## Conclusion

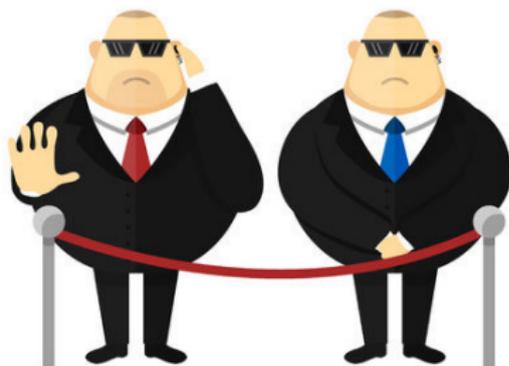
## Conclusion -> Messages pour SemEval10

- ▶ L'extraction de relation avec BERT (même simple) donne de très bons résultats  $F_1^{(9+1)} = 0.8770$
- ▶ Après nos expériences avec 5 runs, une bonne partie des bonnes réponses sont dans le pool  $\sim 0.96$



## Conclusion -> Messages pour SemEval10

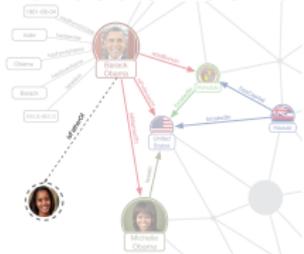
- ▶ L'extraction de relation avec BERT (même simple) donne de très bons résultats  $F_1^{(9+1)} = 0.8770$
- ▶ Après nos expériences avec 5 runs, une bonne partie des bonnes réponses sont dans le pool  $\sim 0.96$
- ▶ Sachant que le modèle le plus performant est à 0.902, la validation de relation semble être plus importante que l'extraction ! mais très peu de personnes s'y intéressent :(



## Conclusion -> Messages pour SemEval10

- ▶ L'extraction de relation avec BERT (même simple) donne de très bons résultats  $F_1^{(9+1)} = 0.8770$
- ▶ Après nos expériences avec 5 runs, une bonne partie des bonnes réponses sont dans le pool  $\sim 0.96$
- ▶ Sachant que le modèle le plus performant est à 0.902, la validation de relation semble être plus importante que l'extraction ! mais très peu de personnes s'y intéressent :(
- ▶ Notre modèle d'extraction de relation via la validation de relation obtient la troisième place de l'état de l'art (0.8927) sur une collection très utilisée (+300 citations) avec la possibilité d'être facilement encastrable dans de nouveaux modèles :)

## Validation de relation



Since arriving on the island, Barack, Michelle, Malia and Sasha have gone on a hike at the Makiki Loop Hawaii Nature Center, seen President Obama's childhood school and dined at Mormoto restaurant.



The "Becoming" author and former President Barack Obama was able to welcome daughters, Malia, 20, and Sasha, 17, through IVF.



## Extraction de relation

