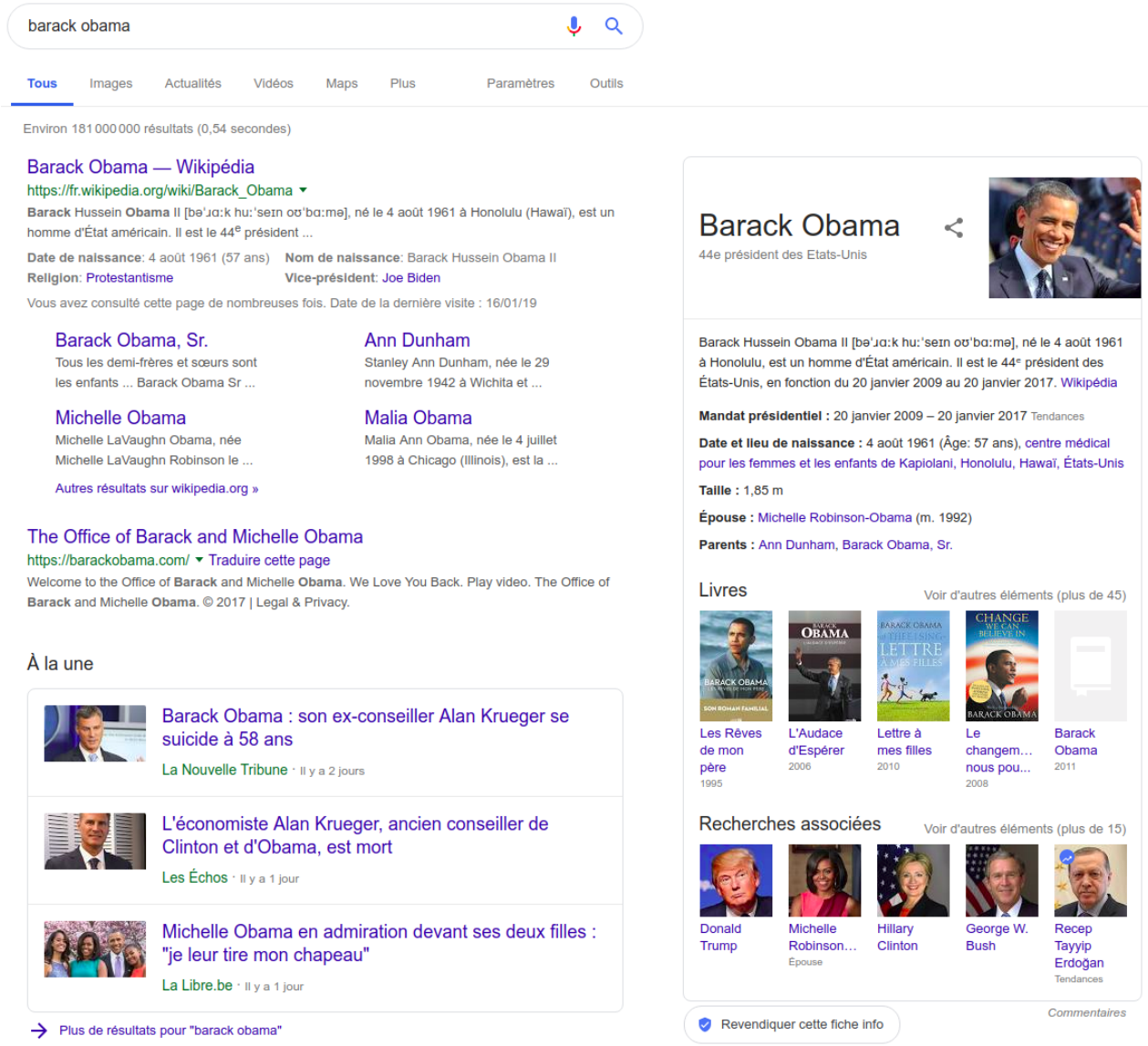
Extraction de relation via la validation de relation

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

*IRIT et LIMSI

02/12/2019

Bases de connaissance en recherche d'information

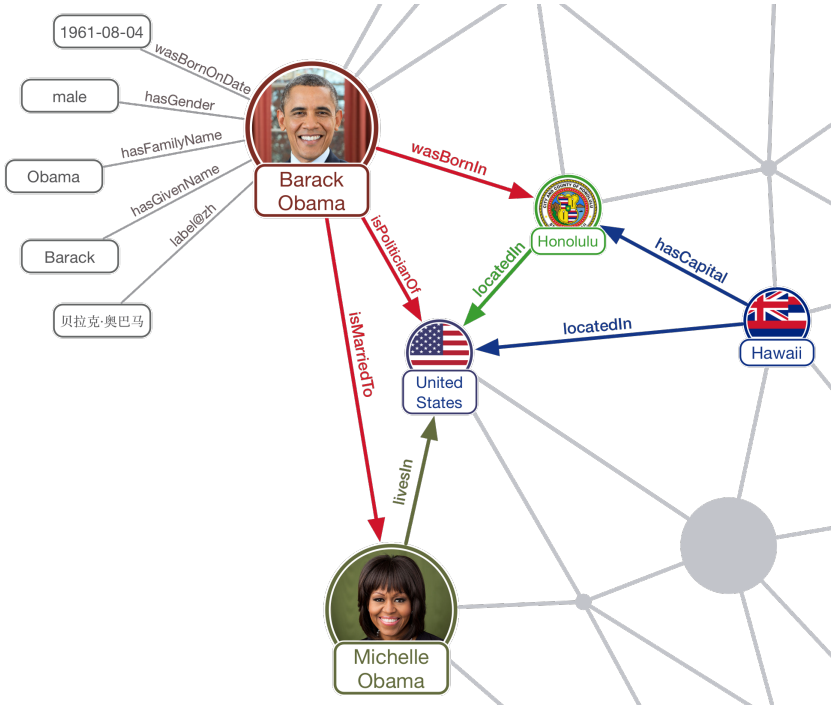


Plus de 50% du contenue affiché provient d'une base de connaissance !

Agenda

Agenda pour la section 1

Extraction d'information

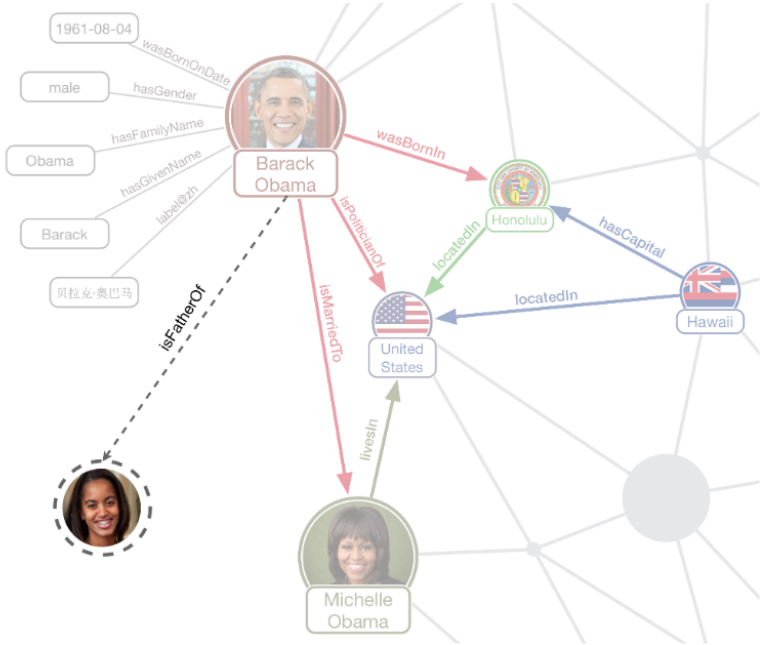


Population de bases de connaissances

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

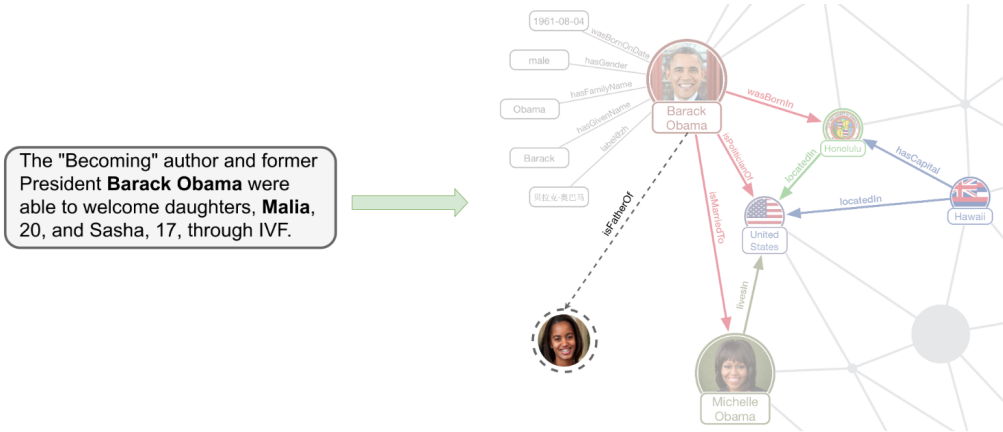
Extraction de relation

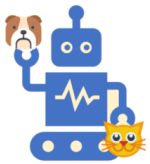
The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and Sasha, 17, through IVF.



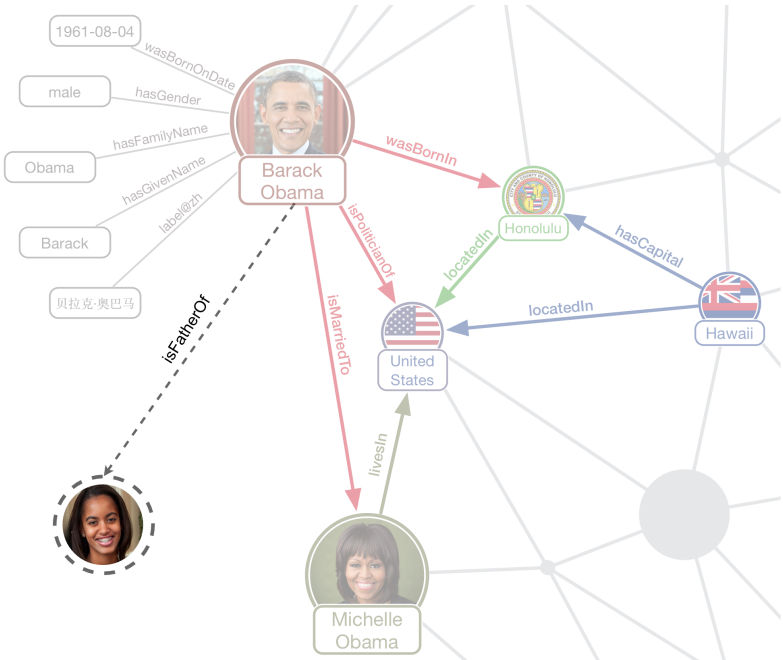
Extraction de relation

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





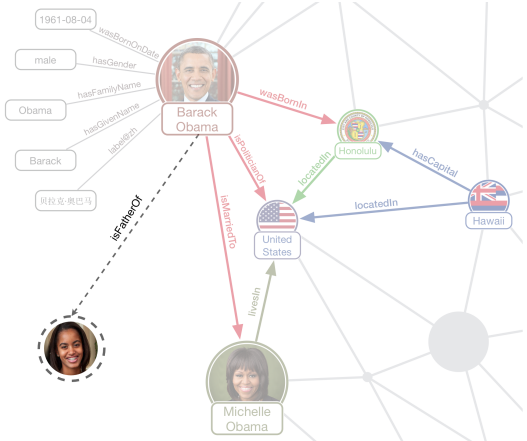
Validation de relation



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.

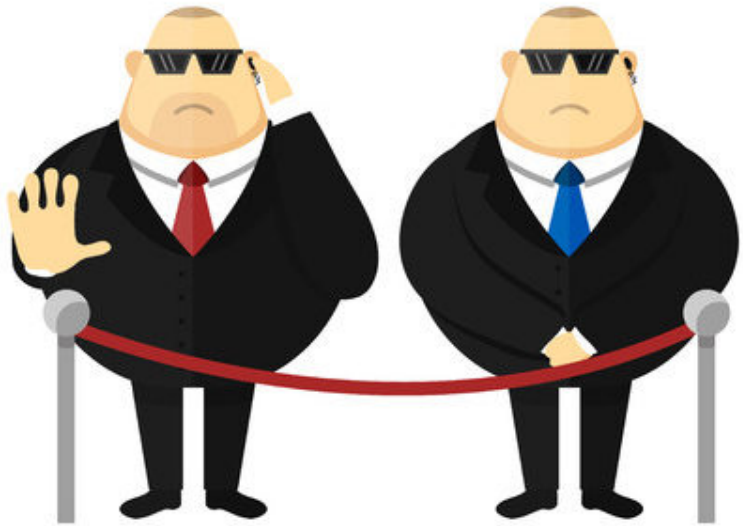
The "Becoming" author and former President **Barack Obama** were able to welcome daughters, **Malia**, 20, and Sasha, 17, through IVF.

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 an 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*

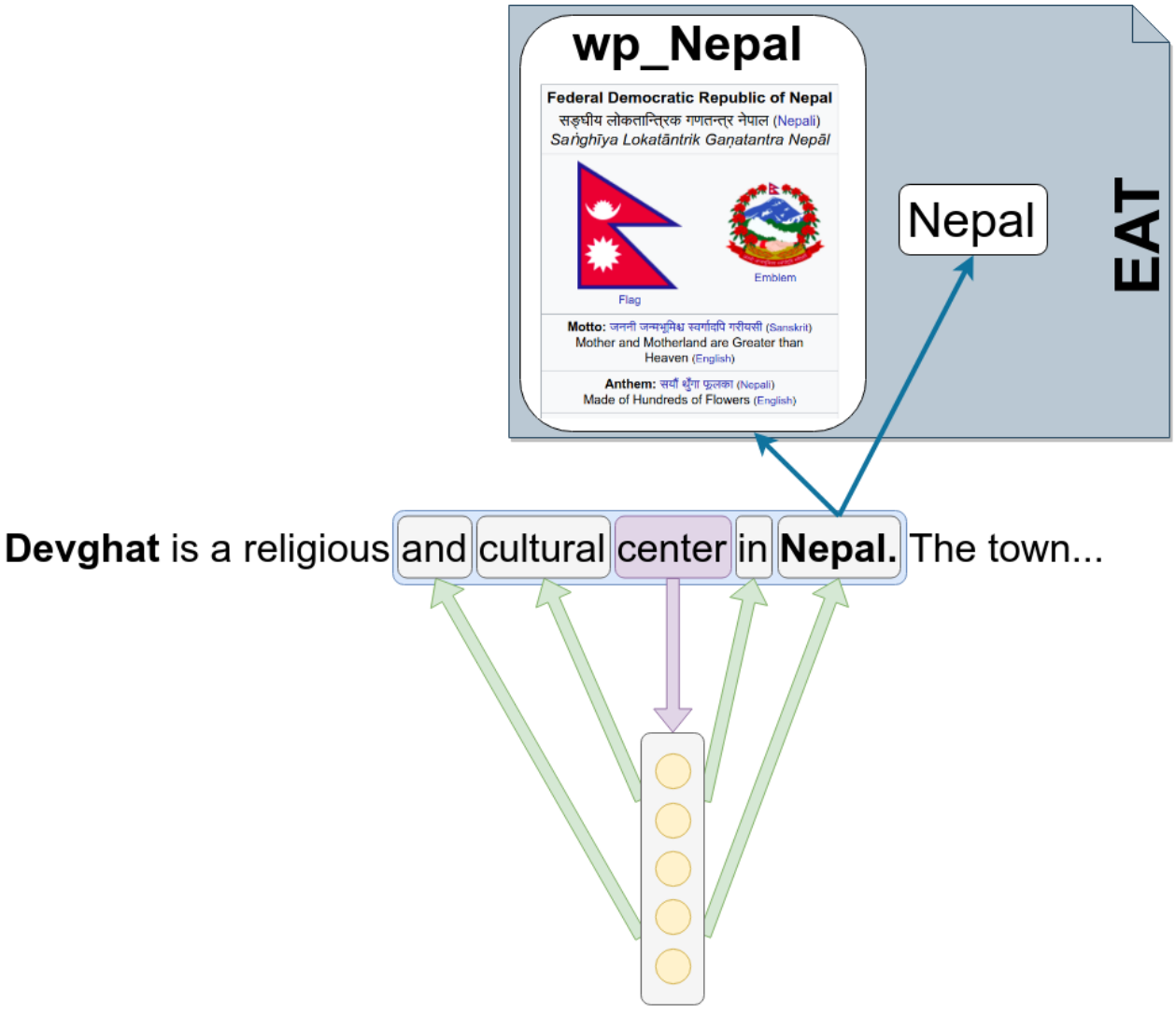
Modèle Extended Anchor Text (EAT)

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

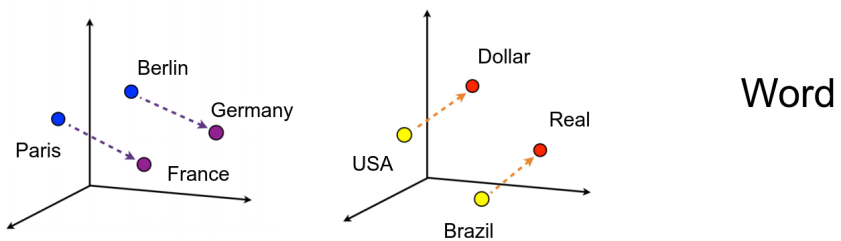


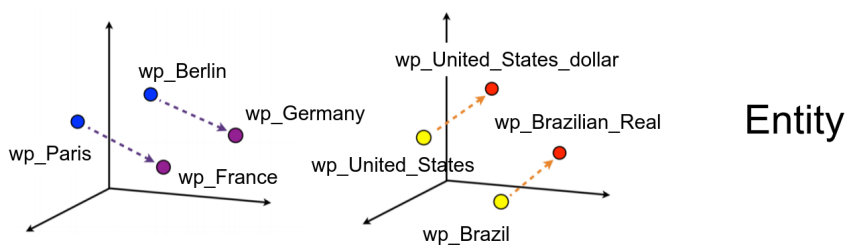
Caractéristiques techniques du modèle EAT

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

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

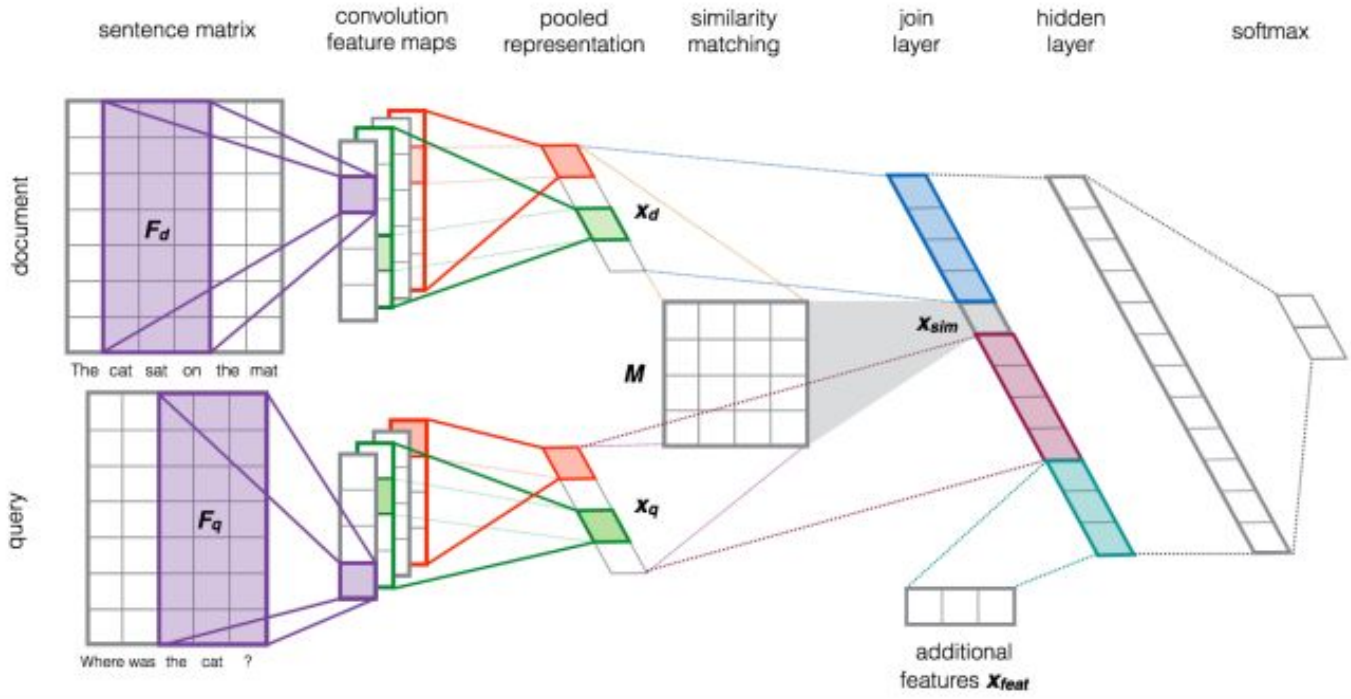
Ent Acc.	EAT.NOE	EAT	word2vec	glove
$ALL_{Sem} - Family$	0.5870	**0.6688**	0.6503	0.6649





Agenda pour la section 2

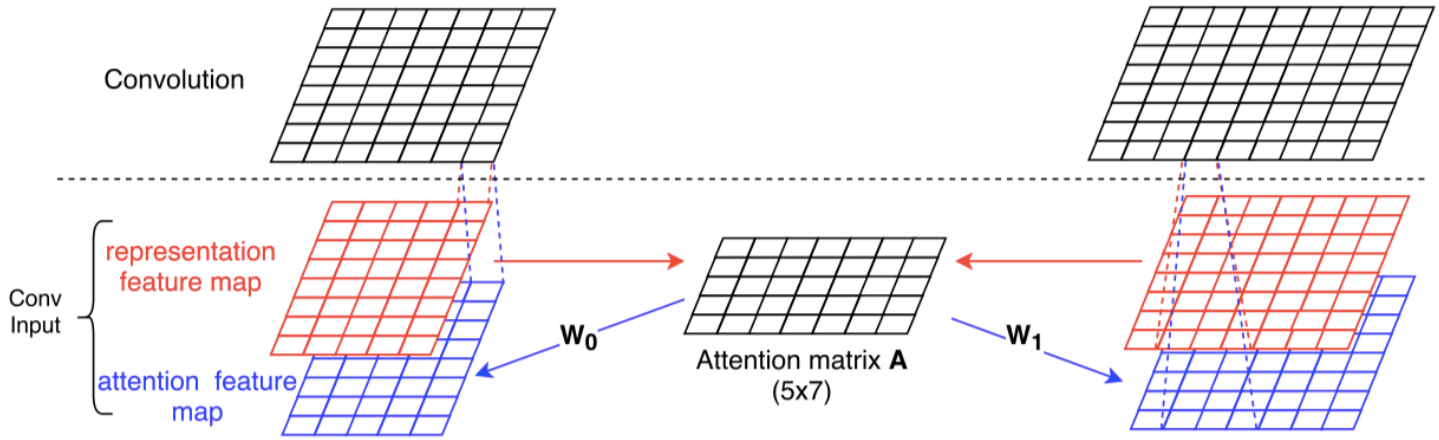
Short-text Matching CNN (SMCNN)



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

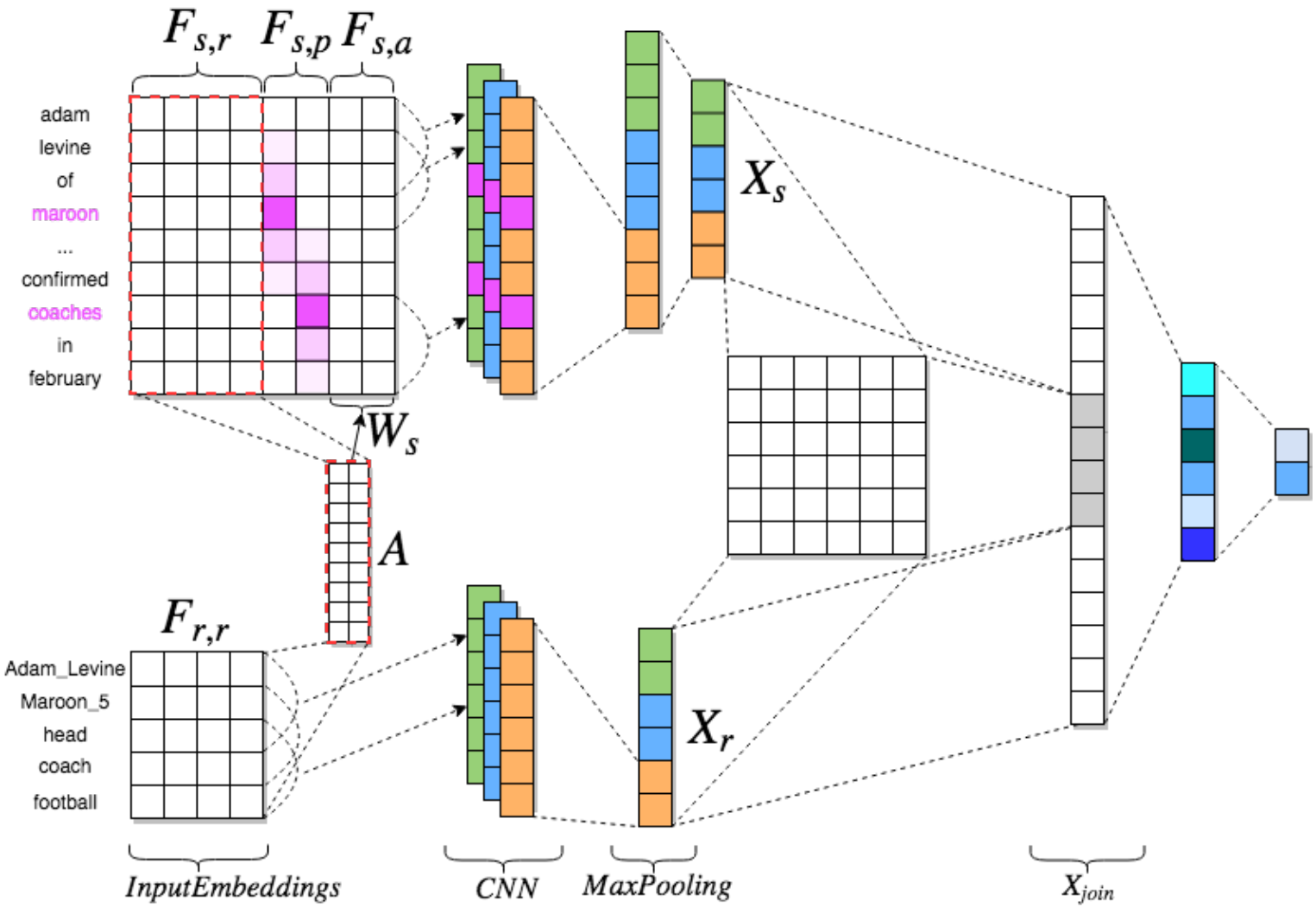
Mécanisme d'attention



ABCNN : Attention-Based CNN - Yin et al., (TACL 2015)

- Adaptation :
  - Matrice d'attention entre le texte à vérifier ($F_{s,r}$) et le triplet de relation ($F_{r,r}$)
  - $A_{i,j} = match\_score(F_{s,r}[i,:], F_{r,r}[j,:])$
  - $match\_score$: similarité cosinus
  - $F_{s,a} = W_s \cdot A^T$

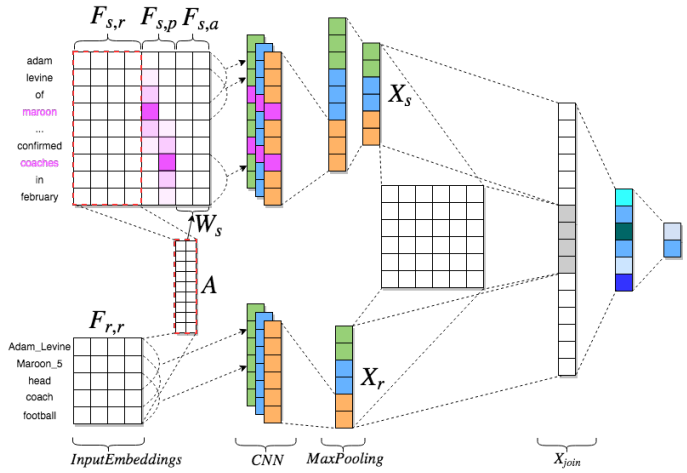
ABSMCNN



Notre modèle

Moreno et al., (CORIA 2019)

ABSMCNN



Notre modèle

Moreno et al., (CORIA 2019)

- ABSMCNNN-POS et ABSMCNNN : correspondent au modèle décrit, mais ce dernier n'inclut pas $F_{s,p}$.
- SMCNN-POS et SMCNNN : 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

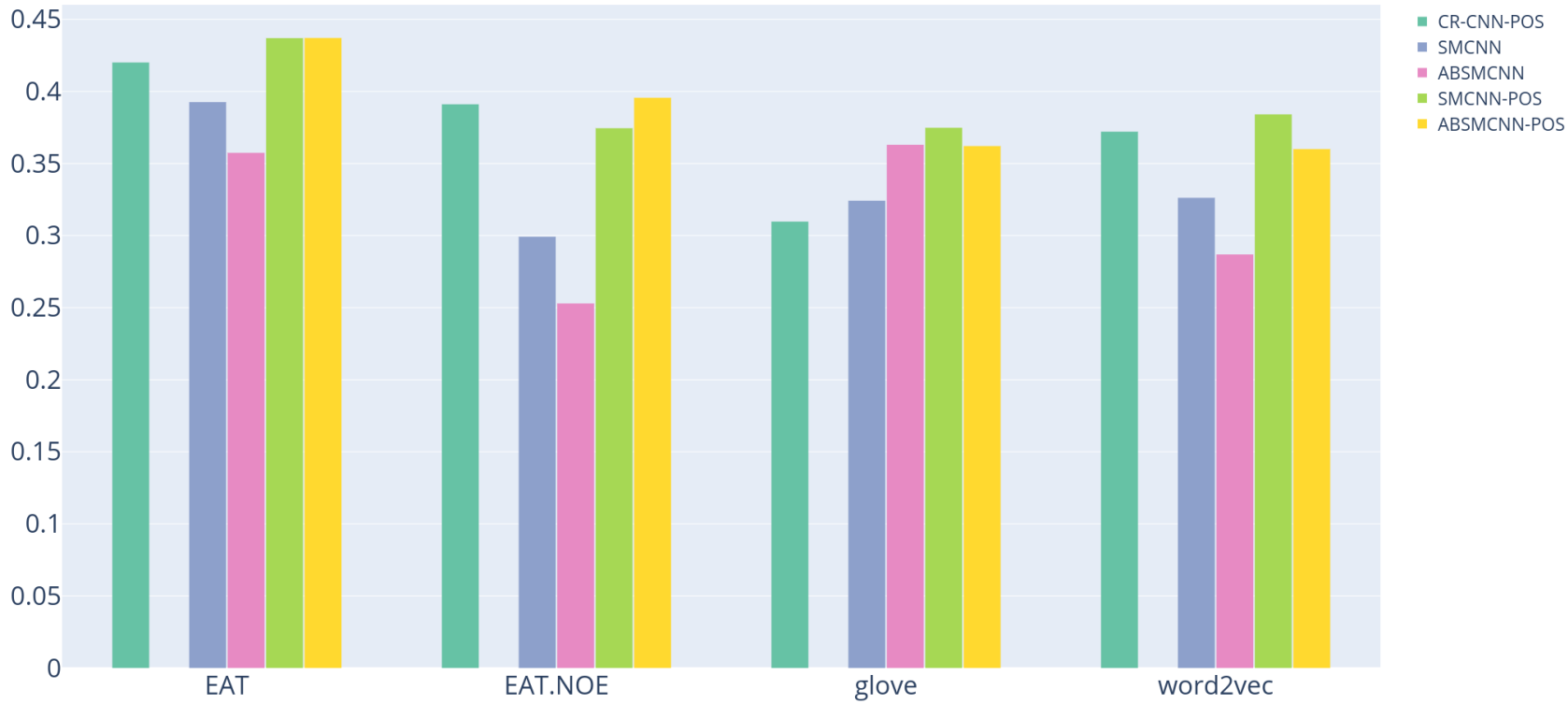
- dataKBP : un sous ensemble du corpus de TAC-KBP CSSF (Cold Start Slot Filling) des années 2015 et 2016.
- dataQA : jeu de données construit à partir de WebQuestions

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)

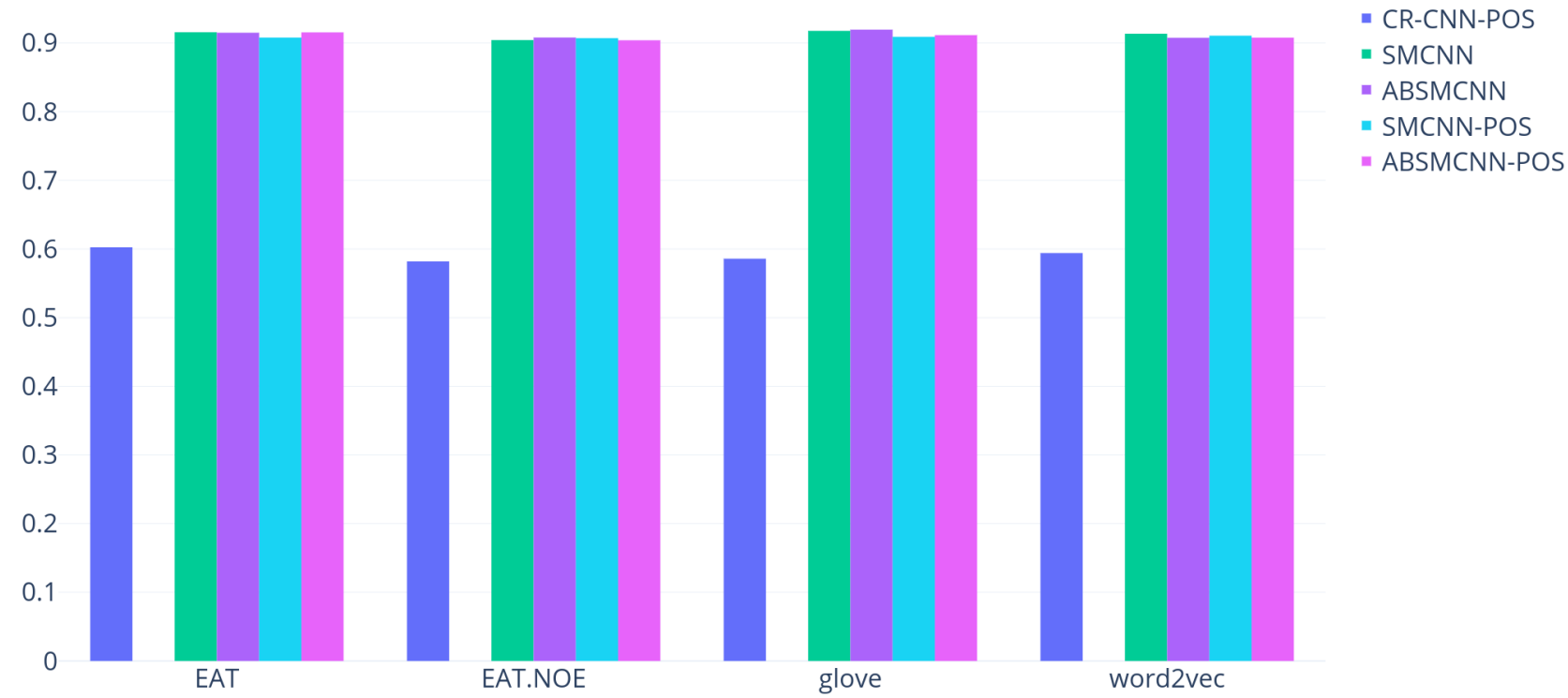
		# Positif	**# Négatif**	**# Total**	**# Relation**
dataQA	Entraînement	4364	4364	8,728	501
	Test	925	925	1,850	311
dataKBP	Entraînement	5,884	8,795	14,679	12
	Test	1,109	4,827	5,936	12

Résultats sur dataKBP



Résultats en termes de F_1 pour le jeu de données dataKBP

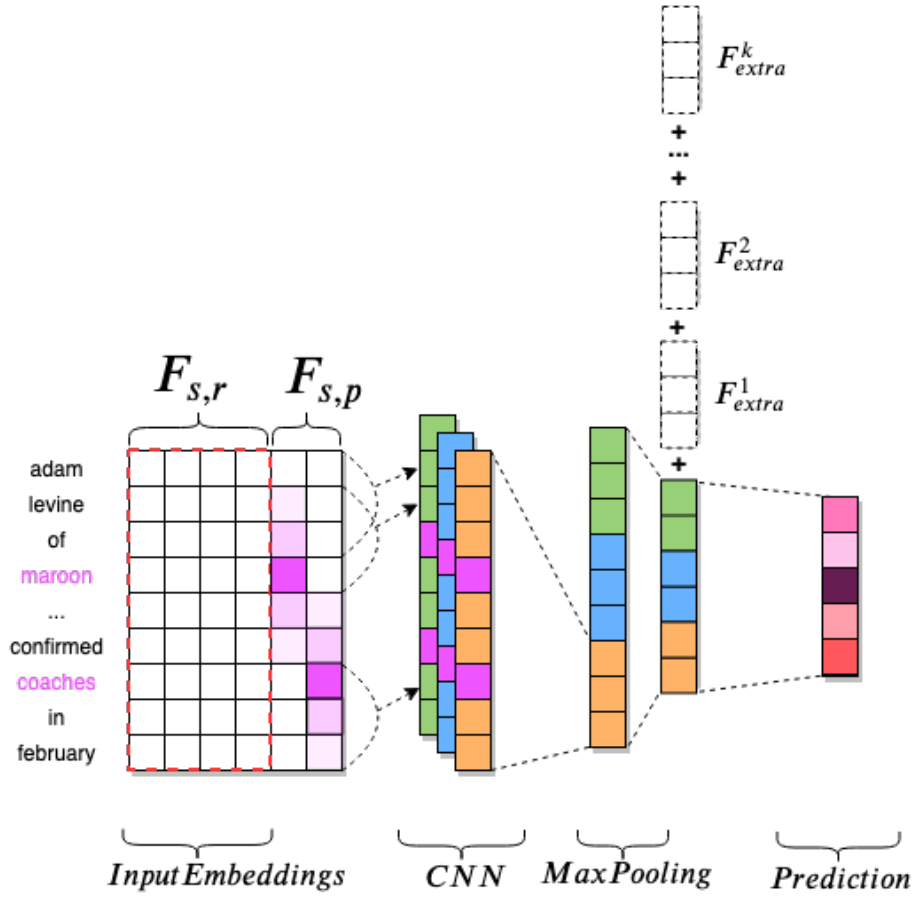
Résultats sur dataQA



Résultats en termes de F_1 pour le jeu de données dataQA

Agenda pour la section 3

CNN extraction de relation



CNN pour l'extraction de relation

- CNN adapté de Zeng et al., (COLING 2014)
- Adaptation : Utilisation de F^i_{extra} 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 **bowl**
- (content-container, pears, bowl)

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	$19_{(8*2+1)}$
	Test	2717	
dataQA	Train	8728	284
	Test	1850	

Exemples SemEval10 - Message-Topic(e1,e2)

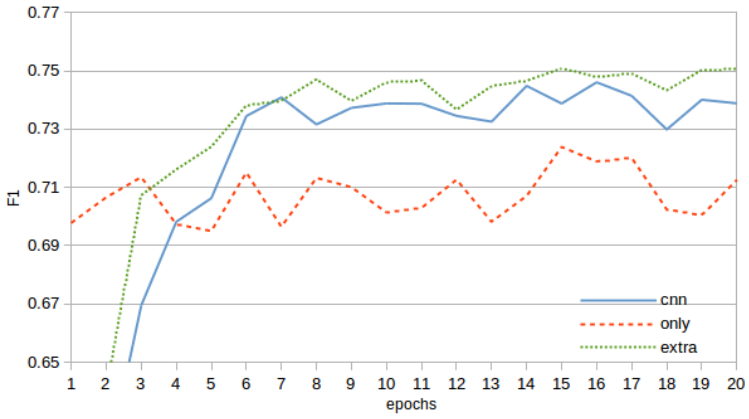
1. The final **[programme]**$_{e1}$ detailed the **[history]**$_{e2}$ of Russborough House
2. The **[letter]**$_{e1}$ contains a description of the **[demolition]**$_{e2}$ of the old synagogue
3. On 17 May 2005, the committee held a **[hearing]**$_{e1}$ concerning specific **[allegations]**$_{e2}$
4.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

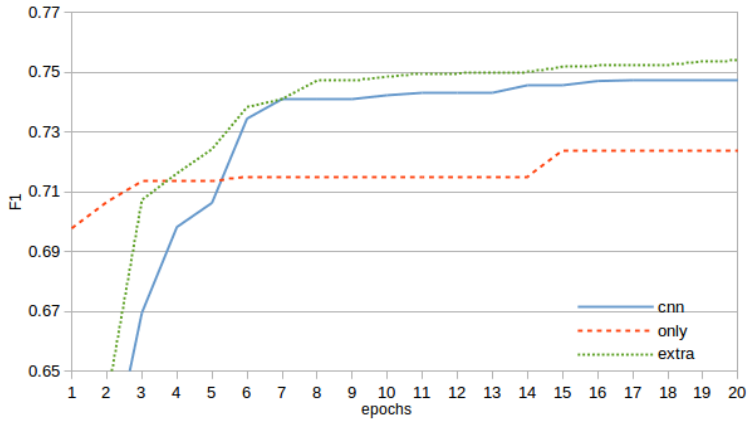
Résultats sur SemEval10 et dataQA

	SemEval10		dataQA	
	F_1	*Accuracy*	F_1	*Accuracy*
F^i_{extra}	0.7239±0.0002	0.7536±0.0002	0.0000±0.0000	0.5000±0.0000
CNN	0.7474±0.0034	0.7777±0.0023	0.0524±0.0019	0.5049±0.0015
EXTRA	**0.7539±0.0027**	**0.7825±0.0019**	**0.0579±0.0029**	**0.5063±0.0014**

- F^i_{extra} 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



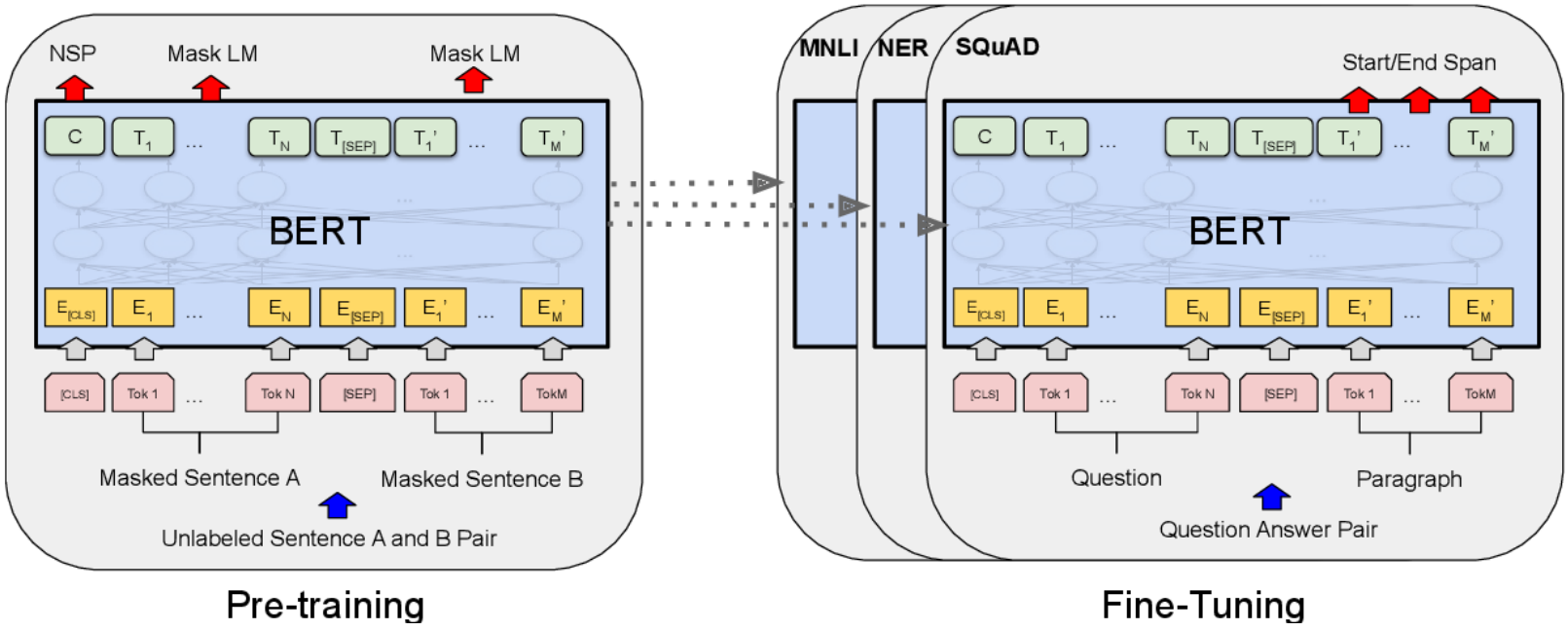


$avg(F_1)$ $avg(F_1^{max})$

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

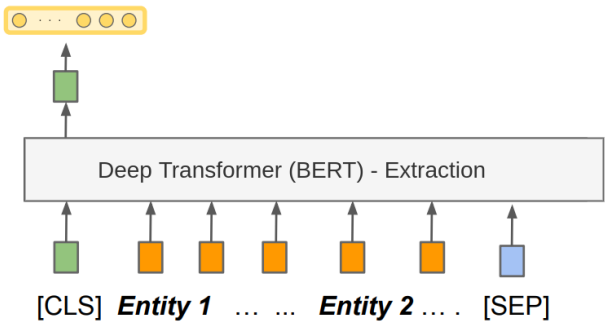
Agenda pour la section 4

BERT : une révolution en TAL



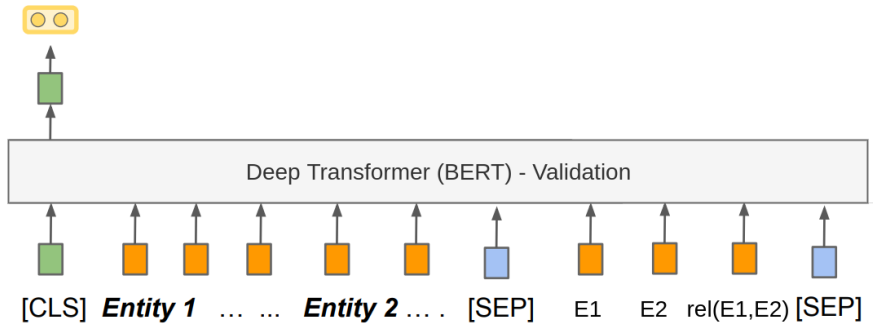
Deviln et al., (NAACL 2019)

BERT pour la extraction et validation de relation



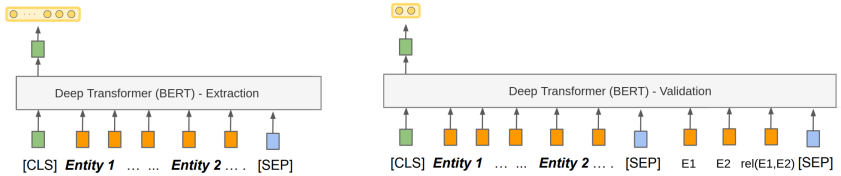
Extraction de relation

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) -> content container

Prédiction

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

Résultats sur SemEval10

	SemEval10		
	F_1	*Accuracy*	$F_1^{(9+1)}$
F^i_{extra}	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)	-	**0.8472**	**0.8831**

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

Résultats détaillés

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

Avec une meilleur validation les résultats peuvent être encore améliorés.

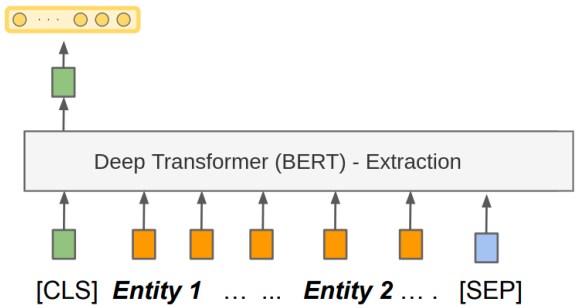
	Nombre de candidants					
	2		3		4	
	True	False	True	False	True	False
BERT-re+rv	338	154	37	52	2	4
	68.69%	31.30%	41.57%	58.42%	33.33%	66.66%

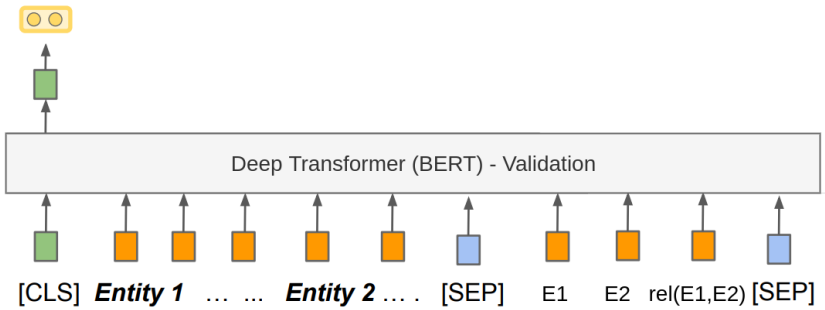
Notre modèle a plus de mal quand le nombre de candidats augmente.

Agenda pour la section 5

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

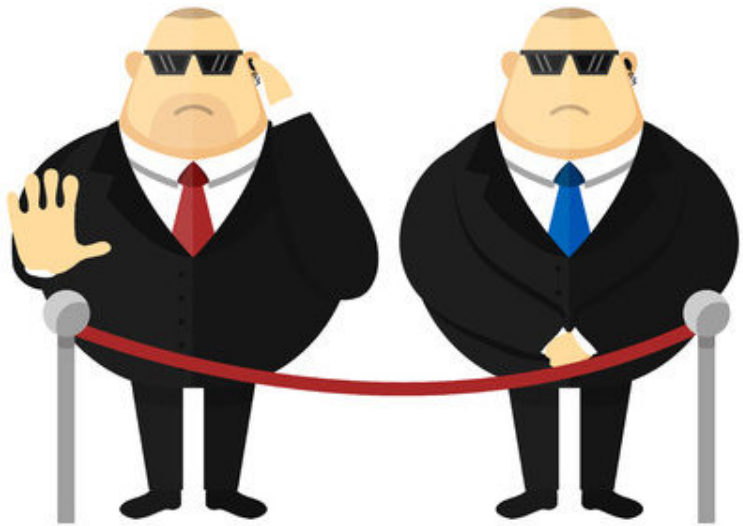




Conclusion -> Messages pour SemEval10

- 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

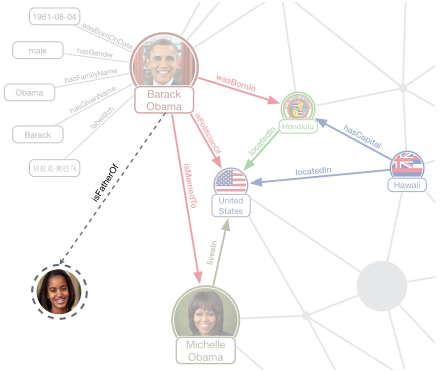
- L'extraction de relation avec BERT (même simple) donne de très bons résultats $F_1^{(9+1)} = 0.8770$
- Après nos expériences avec 5 runs, une bonne partie des bonnes réponses sont dans le pool $\sim$0.96

- 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 utilisé (+300 citations) avec la possibilité d'être facilement encastrable dans de nouveaux modèles :)

Résumé

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 de relation