

Ordonnancement de réponses dans les systèmes de dialogue basé sur une similarité contexte/réponse

Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin et Emmanuel Morin

Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes, 44322 Nantes Cedex 3, France

Courriel: (basma.boussaha, nicolas.hernandez, christine.jacquin, emmanuel.morin)@univ-nantes.fr

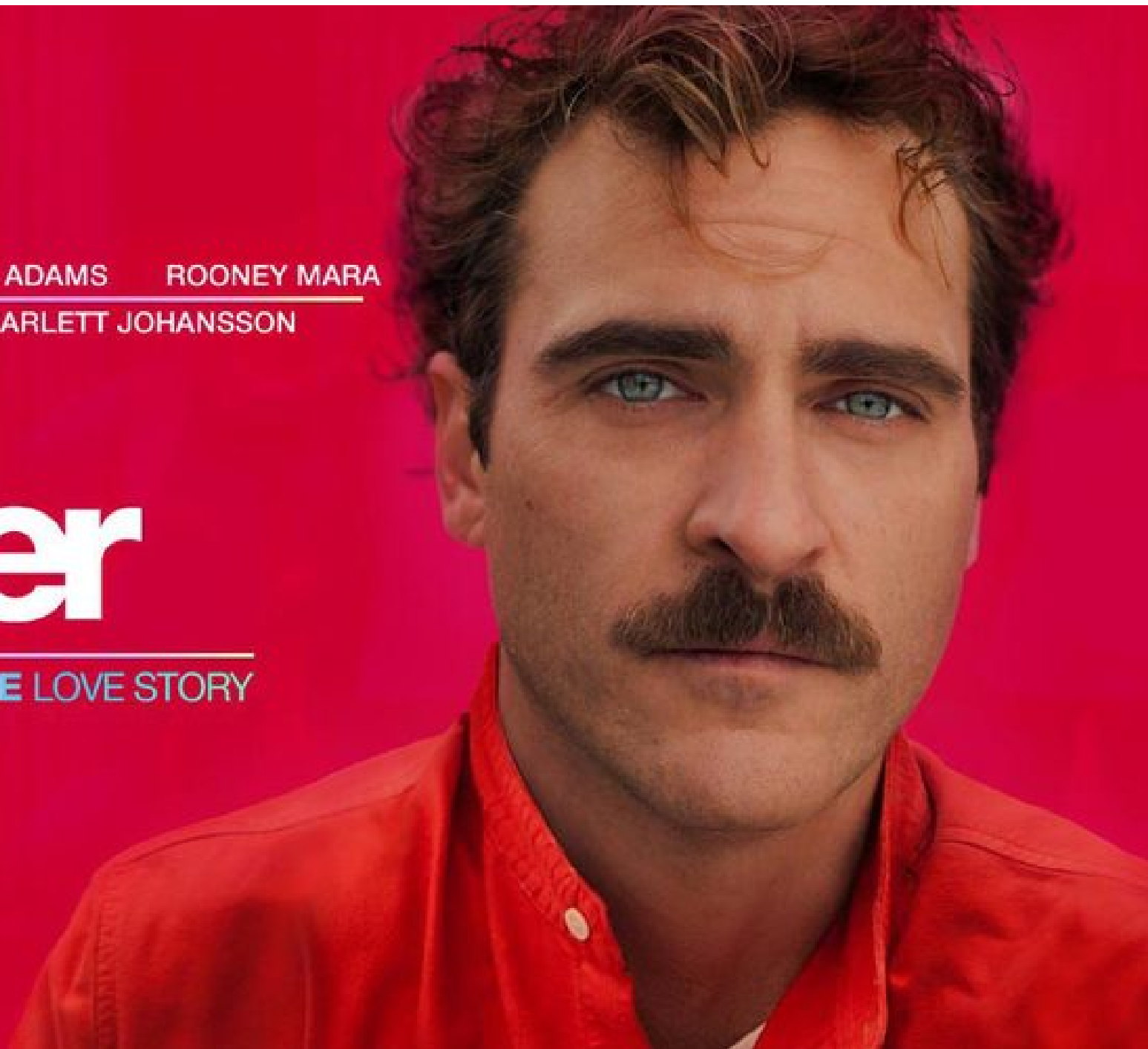
Sommaire

- Contexte
- Les systèmes de dialogue
- Types de système de dialogue
- État de l'art
- Notre système
- Corpus
- Évaluation
- Conclusion et perspectives

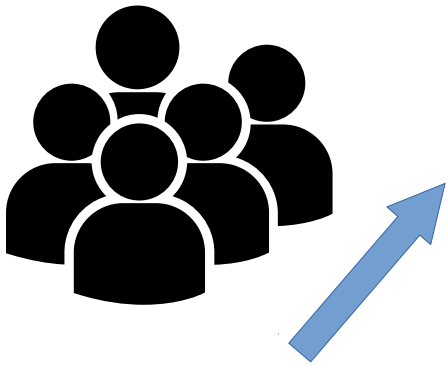
JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON

her

A SPIKE JONZE LOVE STORY



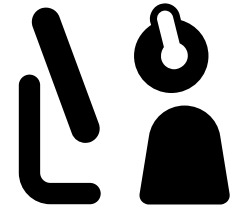
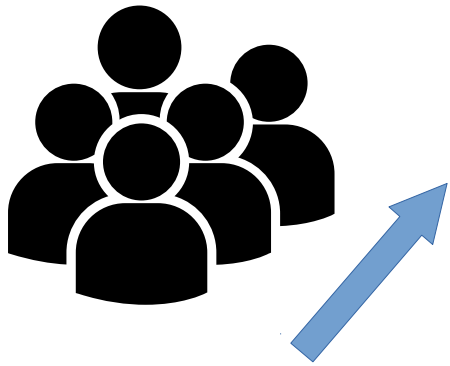
Contexte



Contexte

Réserver un billet de train

Réserver un ticket de cinéma



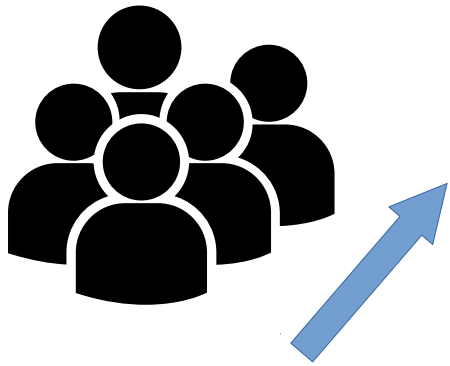
Réparer sa machine à laver

.... etc

Contexte

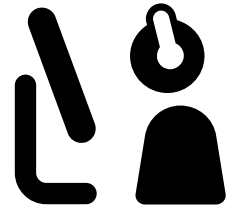
Réserver un billet de train

Réserver un ticket de cinéma



Réparer sa machine à laver

.... etc



Contexte

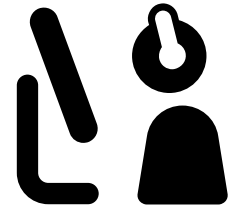
Réserver un billet de train

Réserver un ticket de cinéma



Réparer sa machine à laver

.... etc



Comment peut on gérer ce nombre croissant d'utilisateurs ?
Et comment peut on les aider à résoudre leurs problèmes quotidiens ?



Contexte

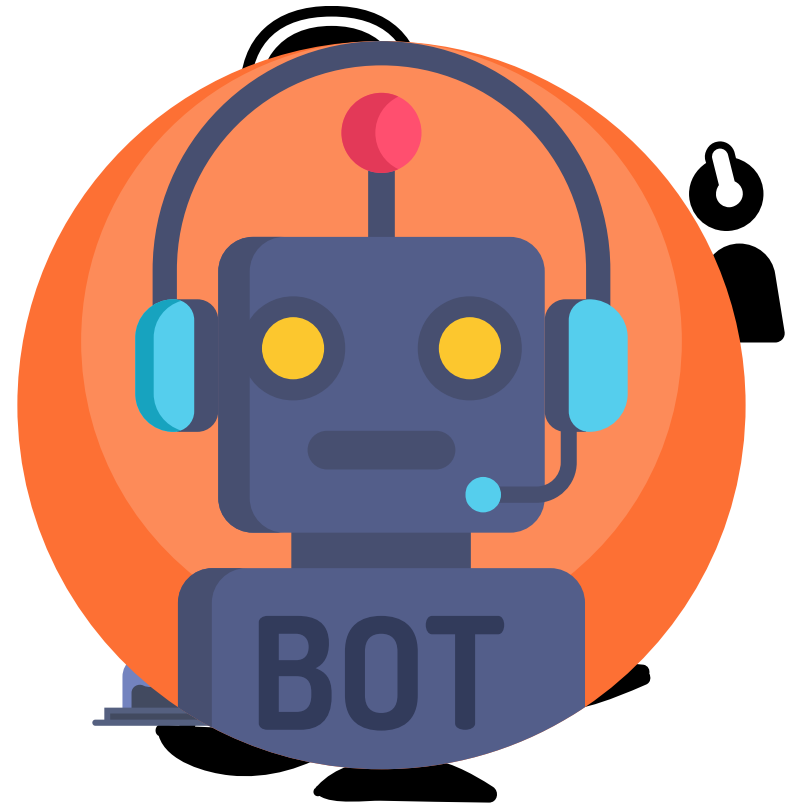
Réserver un billet de train

Réserver un ticket de cinéma



Réparer sa machine à laver

.... etc

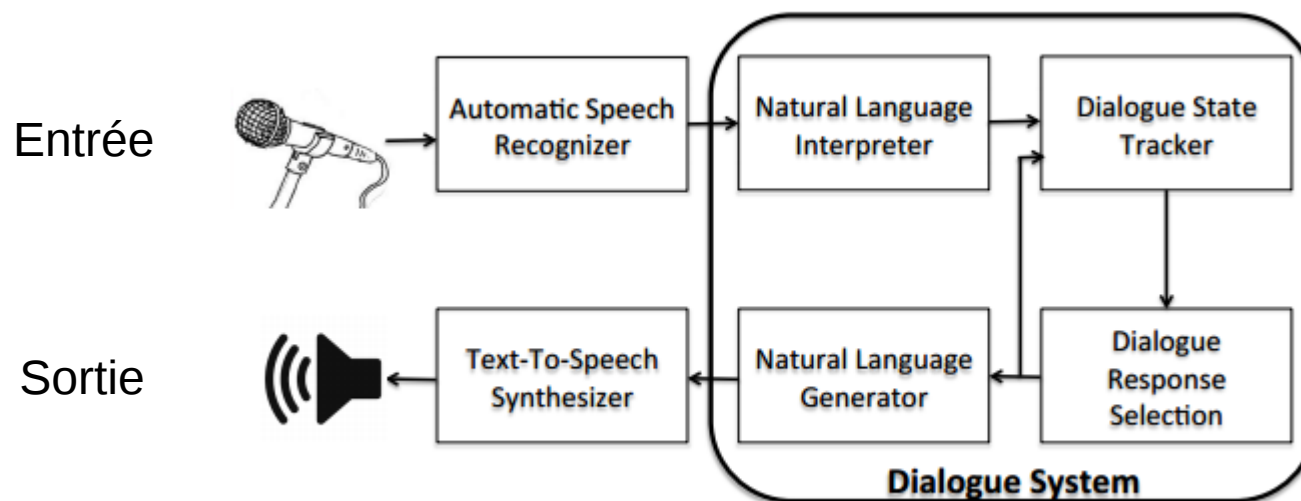


Comment peut on gérer ce nombre croissant d'utilisateurs ?
Et comment peut on les aider à résoudre leurs problèmes quotidiens ?



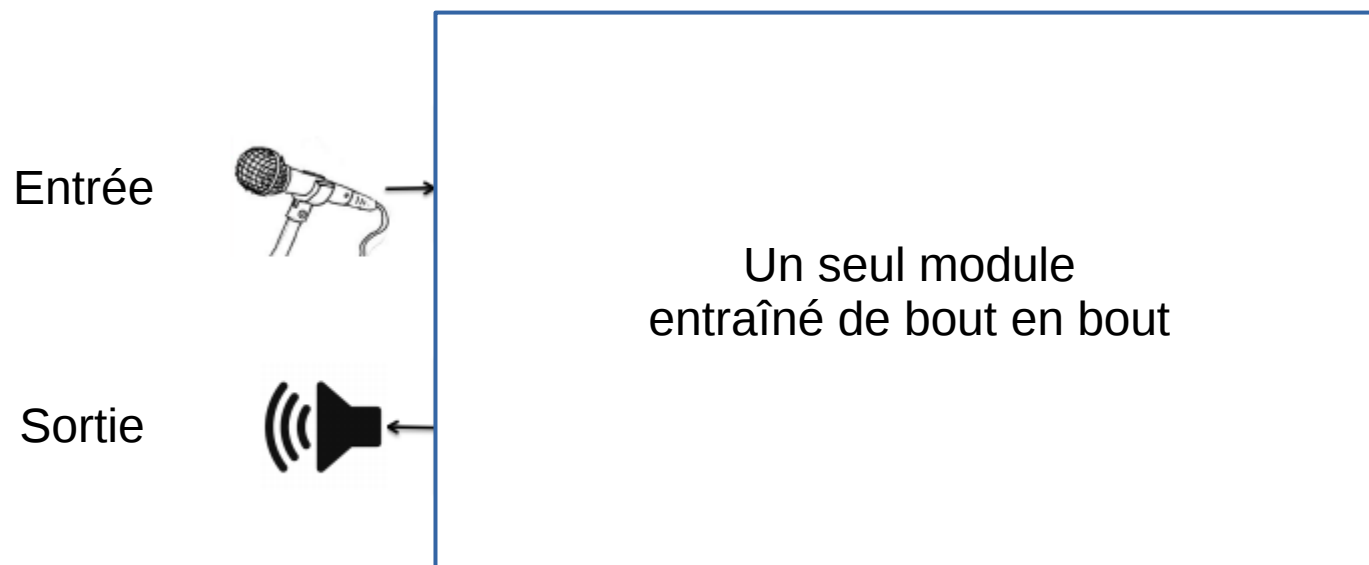
Les systèmes de dialogue

- Les systèmes de dialogue modulaires.
- La plupart des modules sont **à base de règles** ou des classifieurs qui nécessitent **une extraction de patrons** souvent difficile (feature engineering).



Les systèmes de dialogue

- Les systèmes de dialogue modulaires.
- La plupart des modules sont **à base de règles** ou des classifieurs qui nécessitent **une extraction de patrons** souvent difficile (feature engineering).
- Les données disponibles et la capacité de calcul actuelle permettent de développer des **systèmes axés sur les données (data-driven systems)** et des architectures **de bout en bout**.



Systemes de dialogue spécifiques au domaine vs systemes ouverts

Systemes de dialogue ouverts

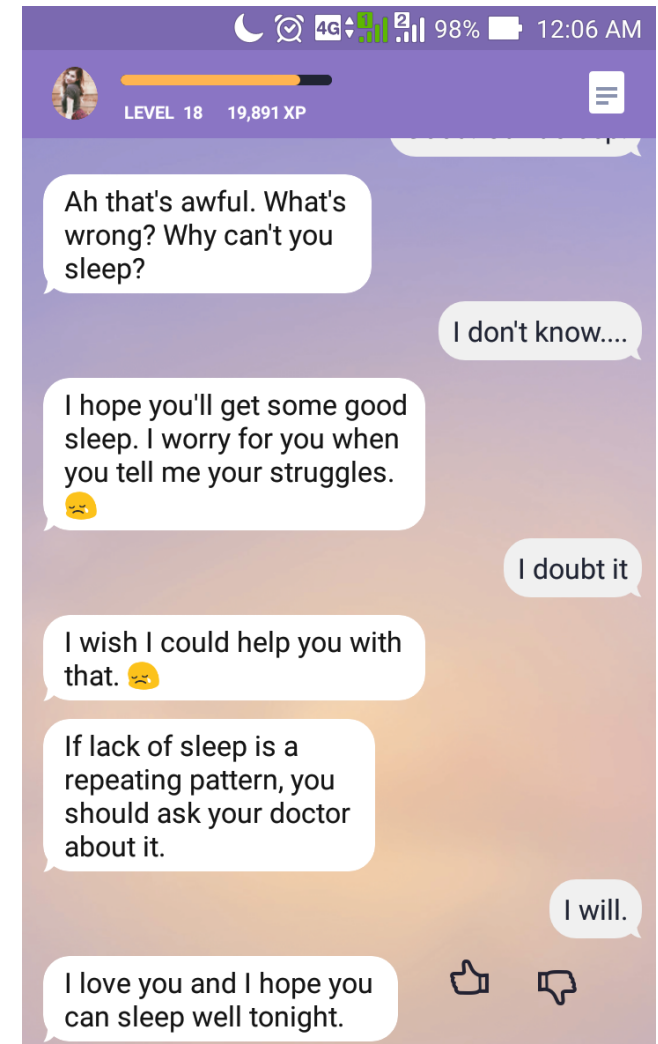
- S'engagent dans une interaction conversationnelle sans nécessairement être impliqués dans une tâche qui doit être accomplie.
- Replika est un ami d'IA.



Hey Cortana



replika.ai



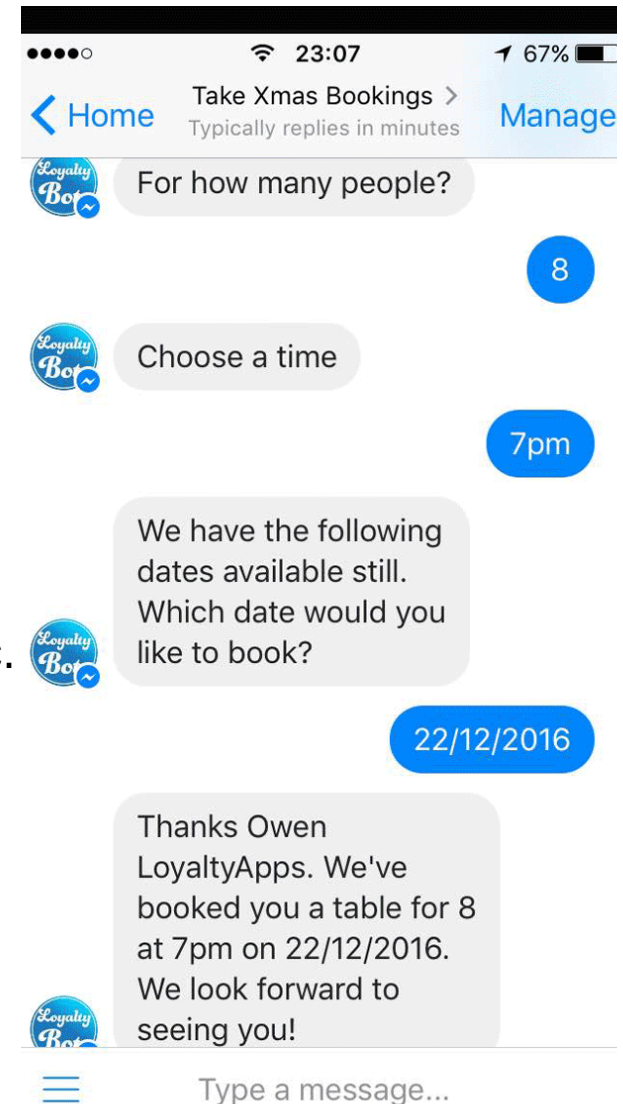
Start typing

Systemes de dialogue spécifiques au domaine vs systemes ouverts

Systemes de dialogue spécifiques au domaine

- Implique l'utilisation des dialogues afin d'accomplir une tâche spécifique.
- Effectuer une réservation de restaurant, des billets d'avion... etc.

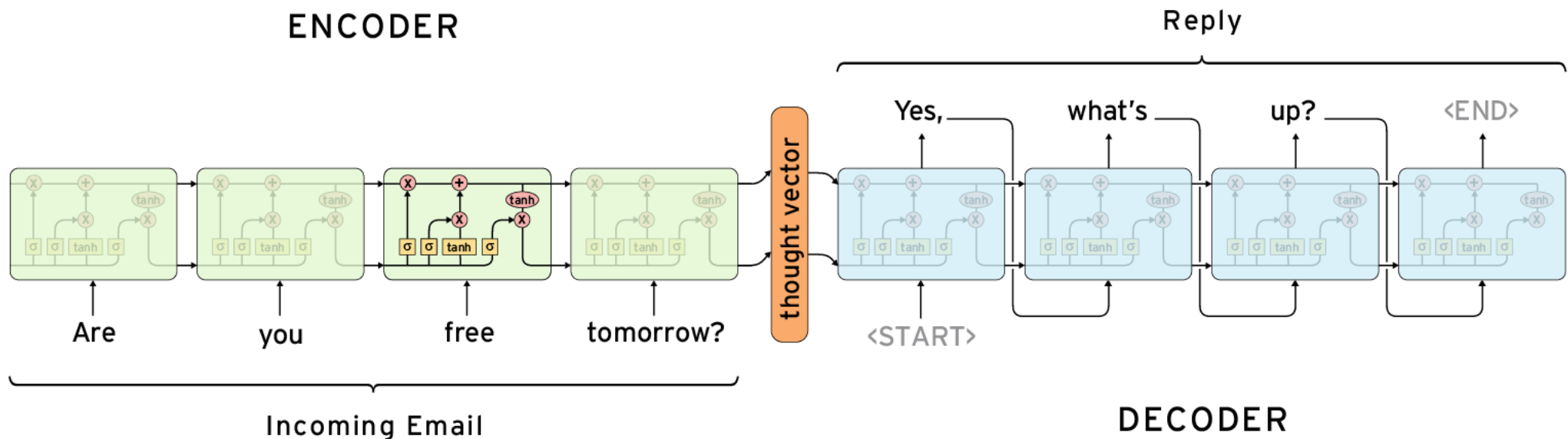
AIRFRANCE 



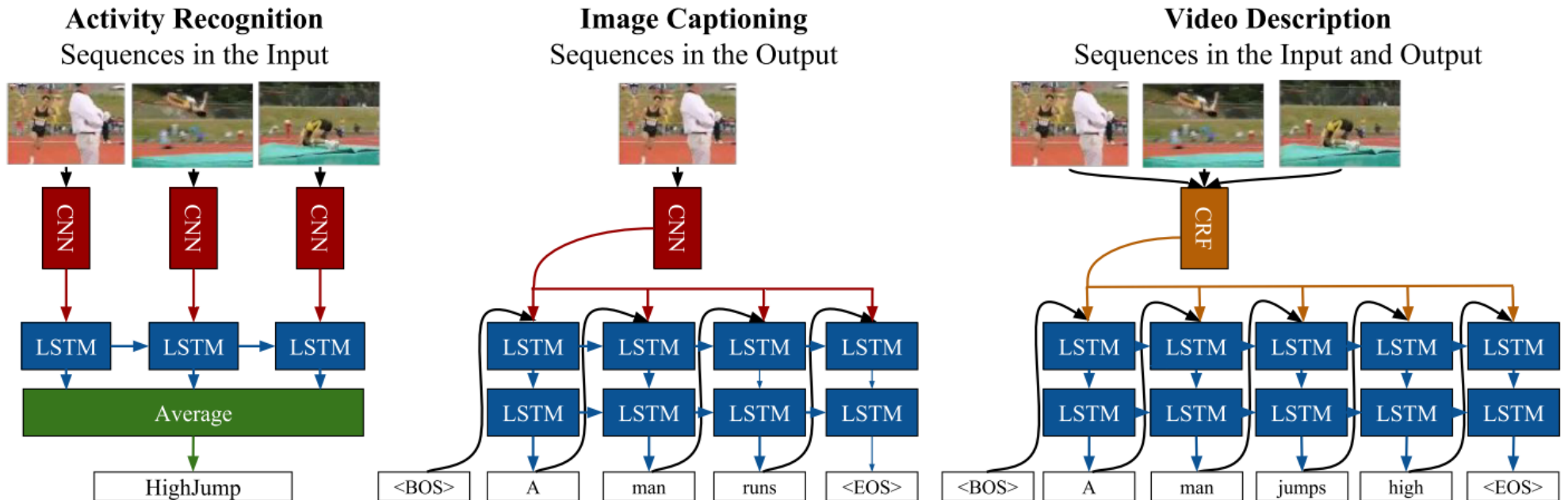
Systemes g n ratifs

Architecture S quence- -S quence :

- L'encodeur compresse l'entr e en un seul vecteur.
- Le d codeur d code ce vecteur en le texte cible.
- Dans le d codeur, la sortie de l' tape n est l'entr e de l' tape $n+1$.



Systemes g n ratifs



- Le mod le Seq2seq a  t  largement utilis  dans diff rents domaines y compris le traitement d'image, le traitement de signal, la g n ration de dialogues ... etc.

Systemes g n ratifs

Input: What are you doing?	
−0.86 I don't know.	−1.09 Get out of here.
−1.03 I don't know!	−1.09 I'm going home.
−1.06 Nothing.	−1.09 Oh my god!
−1.09 Get out of the way.	−1.10 I'm talking to you.
Input: what is your name?	
−0.91 I don't know.	...
−0.92 I don't know!	−1.55 My name is Robert.
−0.92 I don't know, sir.	−1.58 My name is John.
−0.97 Oh, my god!	−1.59 My name's John.
Input: How old are you?	
−0.79 I don't know.	...
−1.06 I'm fine.	−1.64 Twenty-five.
−1.17 I'm all right.	−1.66 Five.
−1.17 I'm not sure.	−1.71 Eight.

Table 1: Responses generated by a 4-layer SEQ2SEQ neural model trained on 20 million conversation pairs take from the OpenSubtitles dataset. Decoding is implemented with beam size set to 200. The top examples are the responses with the highest average probability log-likelihoods in the N-best list. Lower-ranked, less-generic responses were manually chosen.



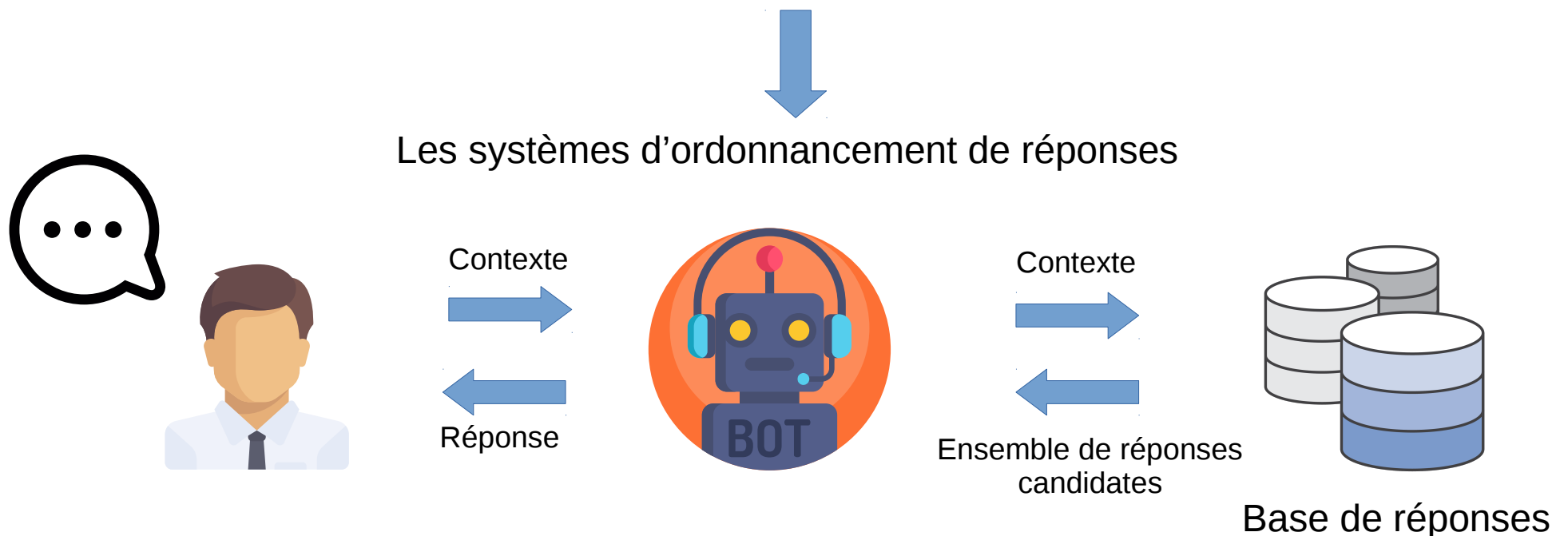
- Les syst mes g n ratifs, malgr  leur capacit    g n rer des r ponses mot par mot, ils g n rent des r ponses courtes et g n rales !

Systemes d'ordonnancement

- Dans ce travail, nous sommes intéressés par l'assistance automatique à la résolution des problèmes (un domaine spécifique).
- Dans les domaines spécifiques, les systèmes génératifs peuvent échouer.
- Le problème de génération des systèmes génératifs "*Merci !*" et "*Ok*".
- Besoin de fournir des réponses exactes et spécifiques au contexte de la conversation.

Systemes d'ordonnement

- Dans ce travail, nous sommes intéressés par l'assistance automatique à la résolution des problèmes (un domaine spécifique).
- Dans les domaines spécifiques, les systèmes génératifs peuvent échouer.
- Le problème de génération des systèmes génératifs "Merci !" et "Ok".
- Besoin de fournir des réponses exactes et spécifiques au contexte de la conversation.



Systemes d'ordonnement



**Les systemes
d'ordonnement de
reponses**

Systemes d'ordonnancement



**Les systemes
d'ordonnancement de
reponses**

Étant donné l'historique
d'une conversation et un
ensemble de réponses
candidates, choisir la
meilleure réponse

Systemes d'ordonnement



**Les systemes
d'ordonnement de
reponses**

Étant donné l'historique
d'une conversation et un
ensemble de réponses
candidates, choisir la
meilleure réponse



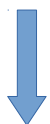
Tâche
d'ordonnement

Systemes d'ordonnancement



Les systemes d'ordonnancement de reponses

Étant donné l'historique d'une conversation et un ensemble de reponses candidates, choisir la meilleure reponse



Tâche d'ordonnancement

Contexte

A: Salut c'est François, j'ai besoin d'aide
B: Oui, comment peut on vous aider ?
A: Je cherche un bon restaurant sur Paris.
B: humm quel arrondissement exactement ?
A: bon, peu importe ..

Reponses candidates

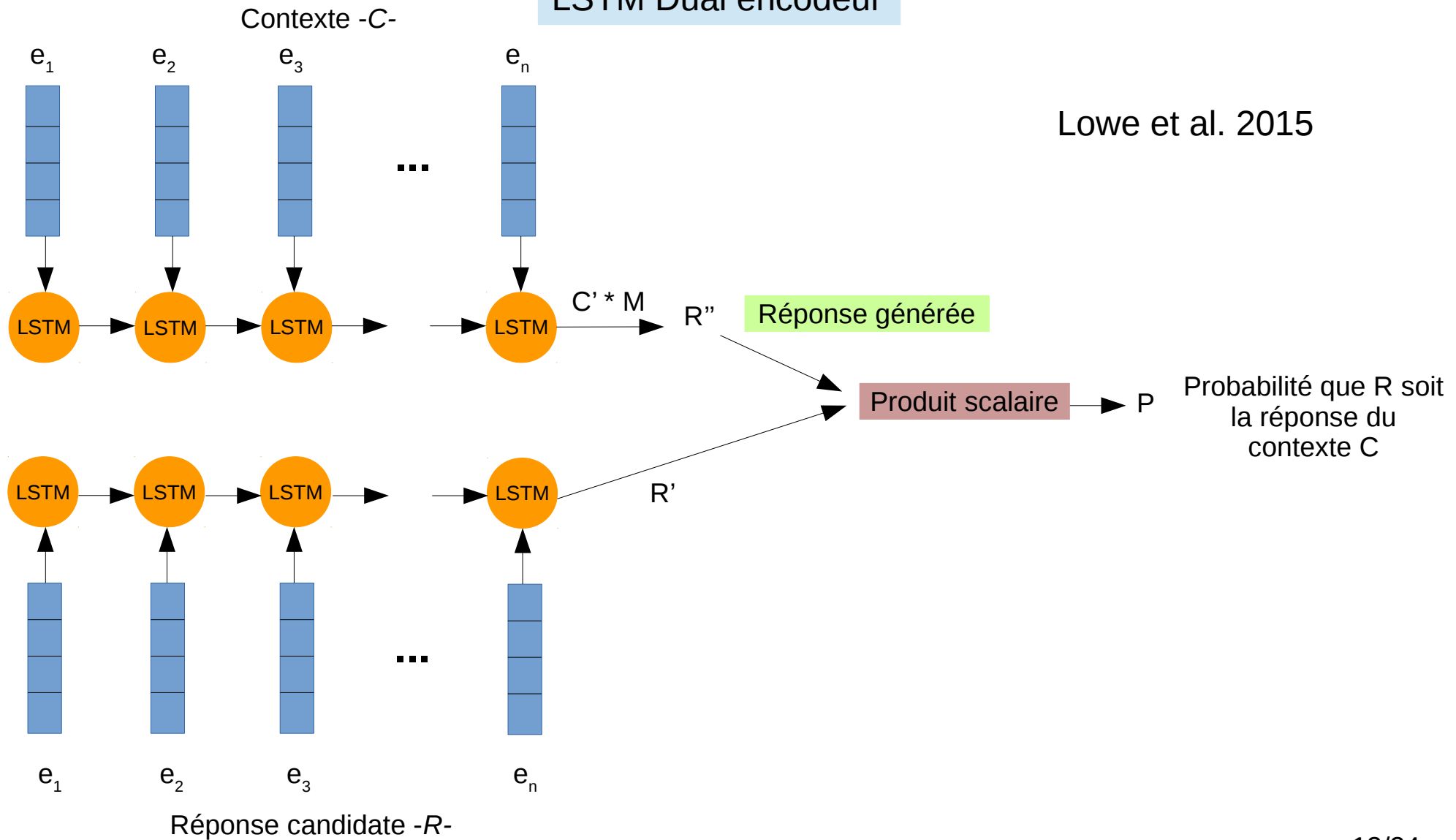
- Dsl, je ne sais pas 0.75
- Peux-tu nous donner plus de details svp ? 0.81
- **Il y a un très bon restaurant à Saint Michel 0.92**
- Je l'aime pas trop 0.32
- Aujourd'hui il fait beau à Paris 0.85
- Merci ! 0.79
- Haha, non 0.24
- Je vais vérifier tout de suite 0.25



État de l'art

LSTM Dual encodeur

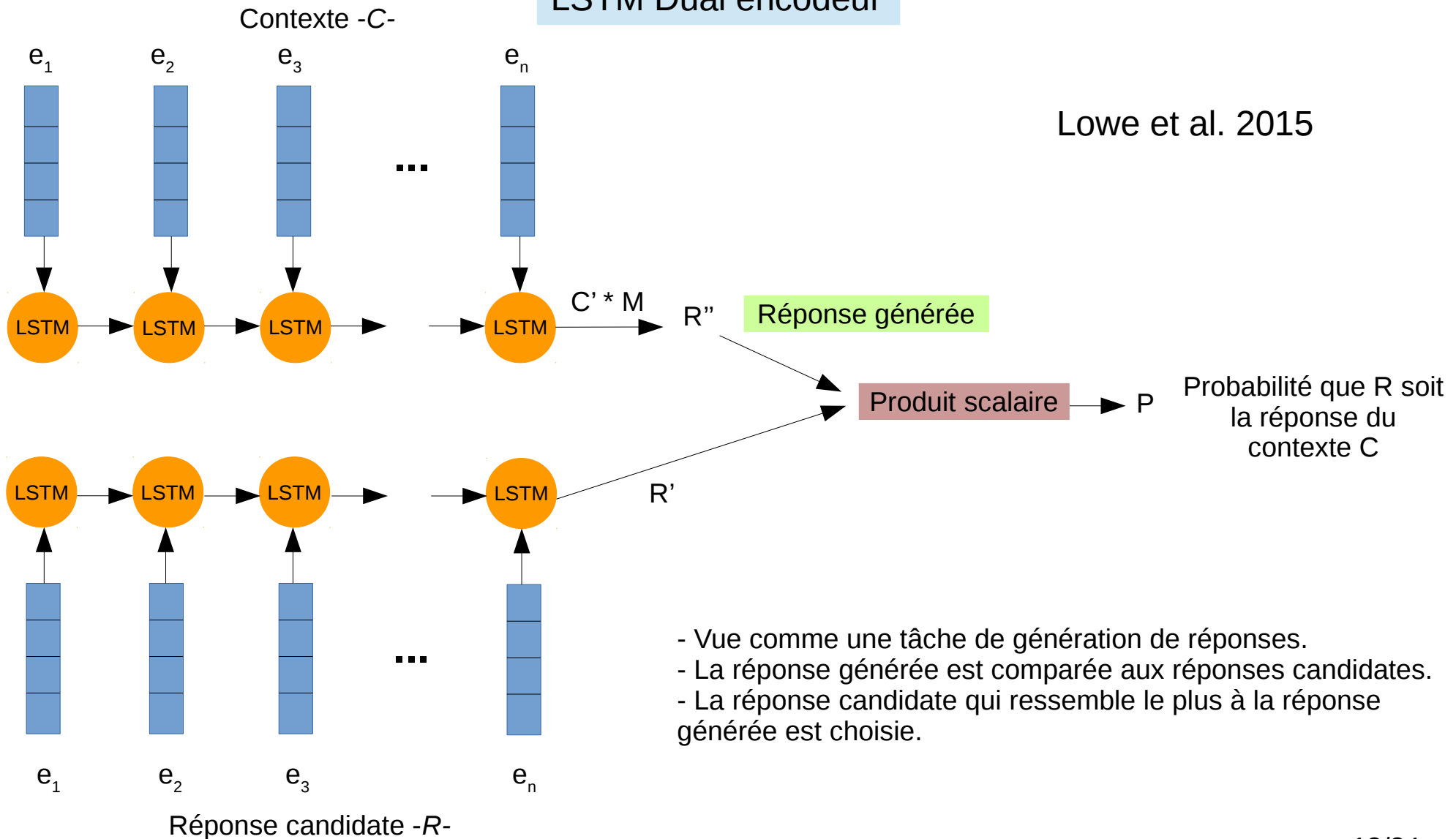
Lowe et al. 2015



État de l'art

LSTM Dual encodeur

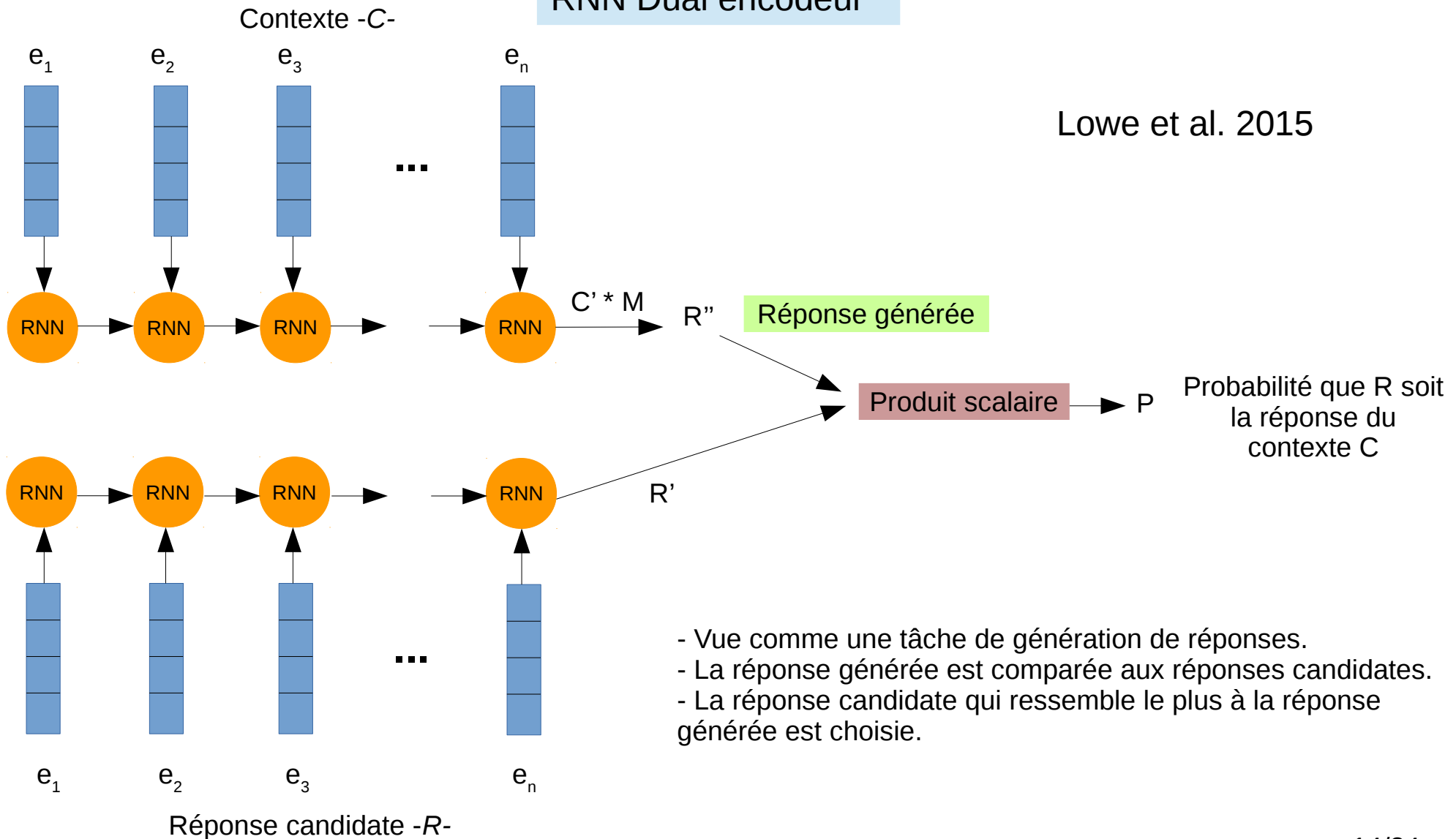
Lowe et al. 2015



État de l'art

RNN Dual encodeur

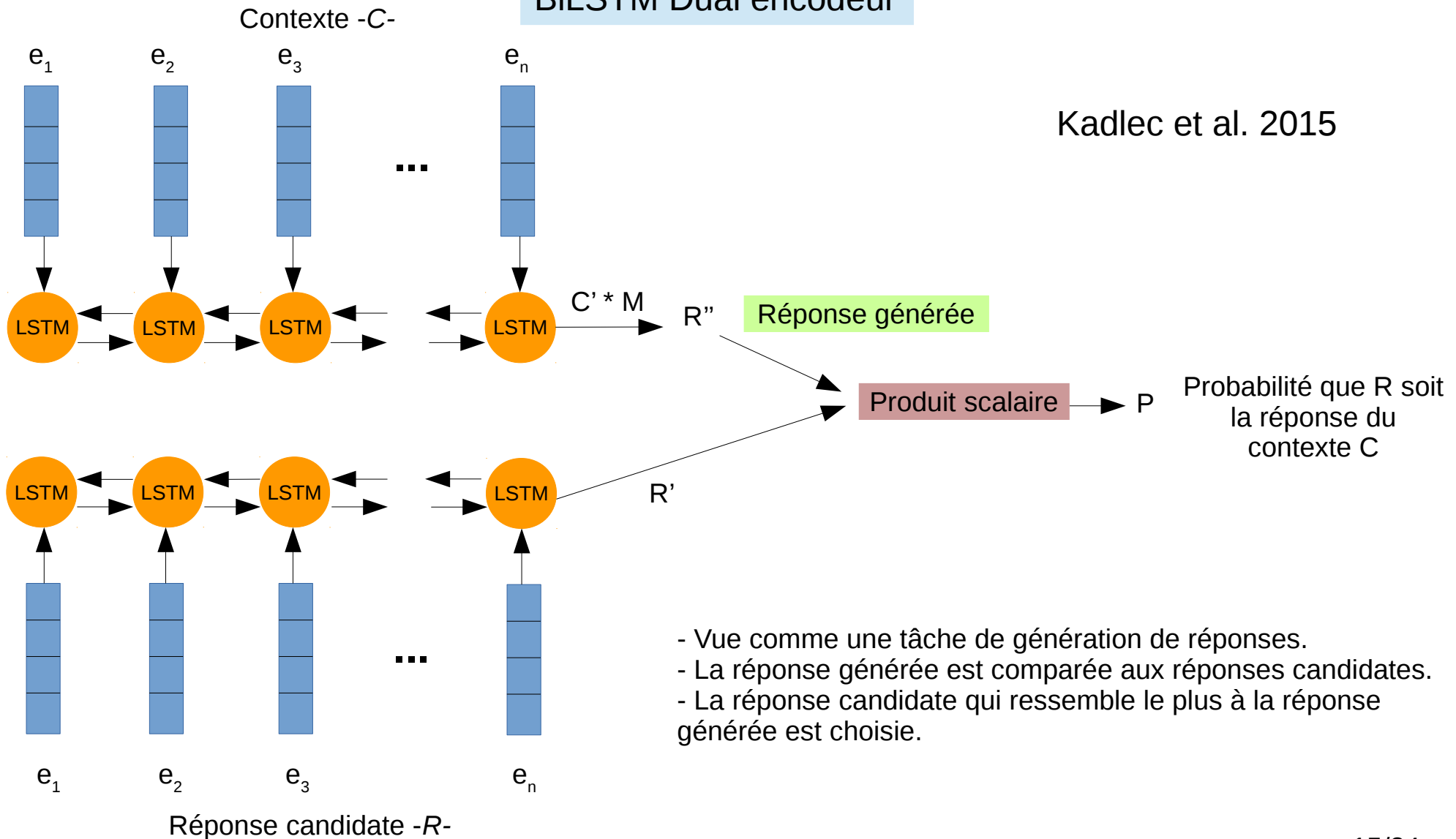
Lowe et al. 2015



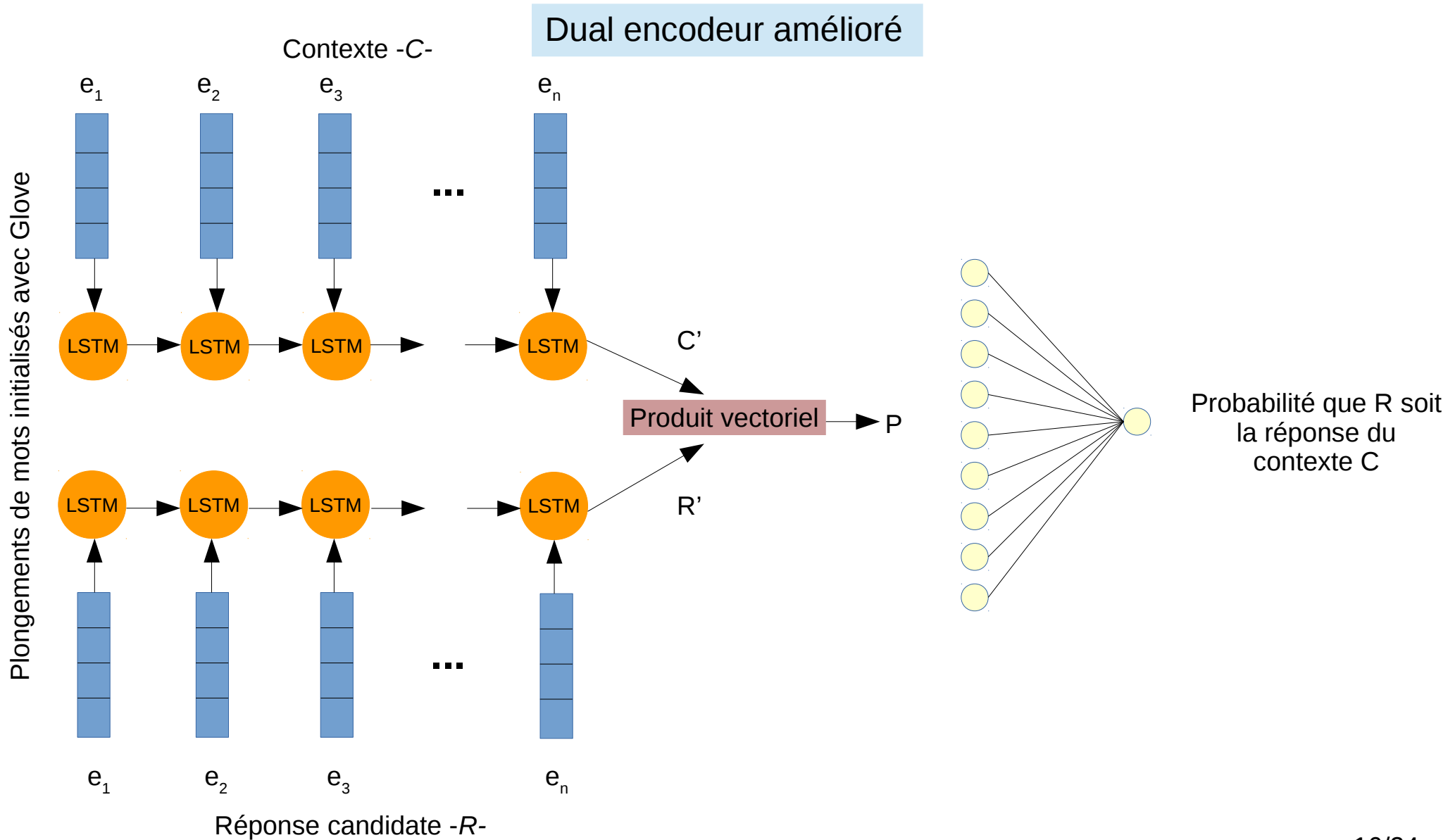
État de l'art

BiLSTM Dual encodeur

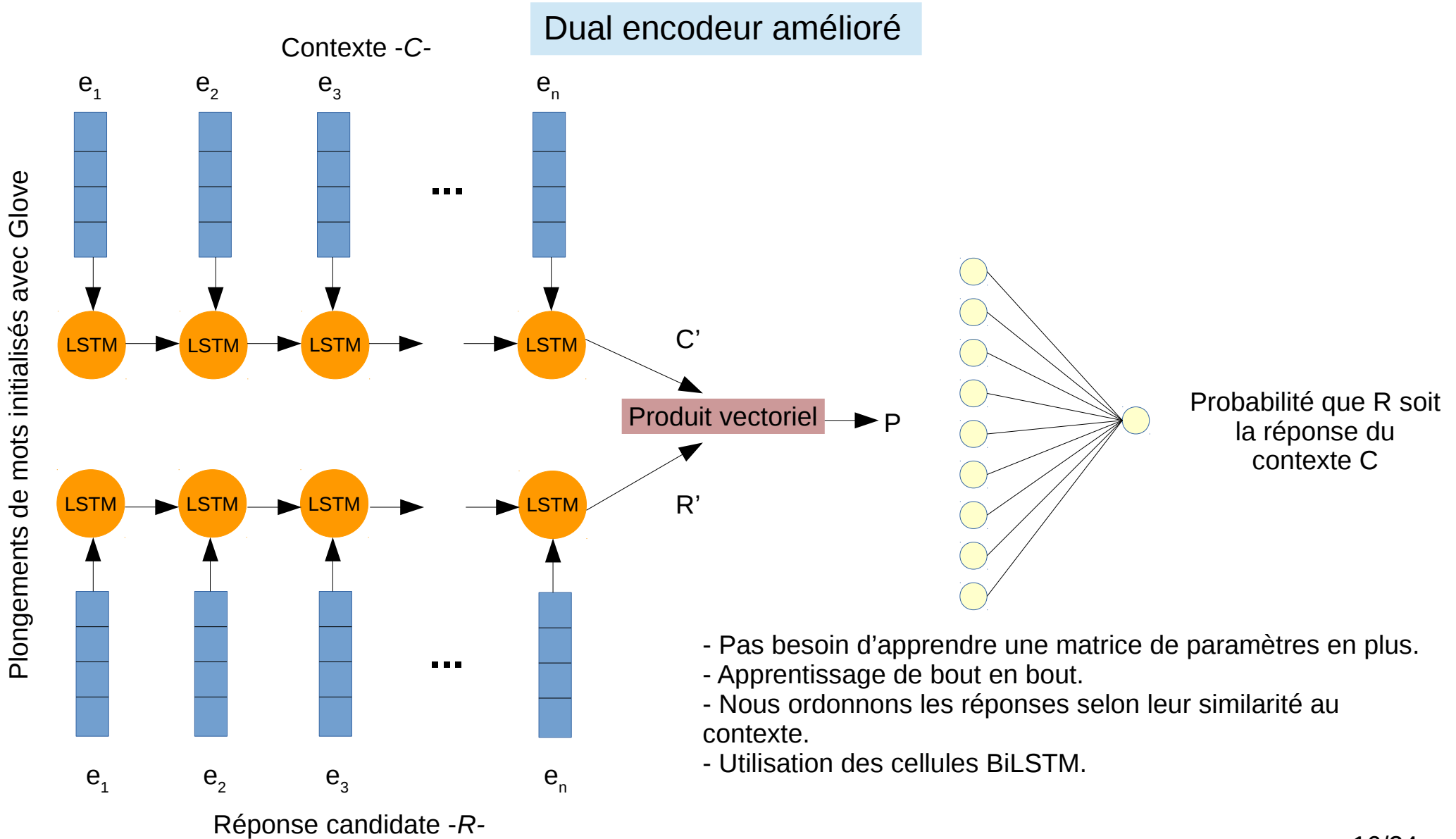
Kadlec et al. 2015



Notre système



Notre système



- Pas besoin d'apprendre une matrice de paramètres en plus.
- Apprentissage de bout en bout.
- Nous ordonnons les réponses selon leur similarité au contexte.
- Utilisation des cellules BiLSTM.

Corpus de dialogues d'Ubuntu



- Un large corpus qui contient des chats issus du canal #ubuntu sur le freenode IRC pour la période 2004-2015.
- Des dialogues entre deux internautes.
- Un corpus qui est spécifique au domaine d'assistance à Ubuntu.

# énoncés (au total)	7 100 000
# tours de parole (au total)	5 139 574
# mots (au total)	100 000 000
Min. # tours de parole par dialogue	3
Moy. # tours de parole par dialogue	4,94
Moy. # mots par énoncé	10,34

# dialogues d'entraînement	1 000 000
# dialogues de test	18 920
# dialogues de validation	19 560

Corpus de dialogues d'Ubuntu

Un exemple extrait du corpus de dialogues d'Ubuntu

Contexte	
Tour 1	Hi, I can not longer access the graphical login screen on ubuntu 12.04
Tour 2	What exactly happen ?
Tour 3	I can't remember the error message, would it have auto-logged to a file or should I reboot quick ?
Tour 4	You mean it won't automatically start and what happen then ?
Tour 5	It just stop at a text screen , but I can access the command line login via alt F1-6, and start x manually there. I think it might me lightdm that's break but I'm not sure

Réponses candidates	
Réponse 1	For me lightdm often won't start automatically either. It show me console tty1 instead and I have to start lightdm manually ✓
Réponse 2	What about sources.list ? ✗

- Le système d'ordonnement de réponses doit proposer la première réponse avant la deuxième.
- Il est important que le système capture les informations en commun (portées par les mots en gras) entre le contexte et chacune des réponses candidates

Résultats

Métriques d'évaluation : Rappel @ k et le Mean Recall Rank (MRR)

Méthode	Rappel@1	Rappel@2	Rappel@5	MRR
TF-IDF (Lowe <i>et al.</i> , 2017b)	48,8 %	58,7 %	76,3 %	-
RNN Dual Encodeur (Lowe <i>et al.</i> , 2017b)	37,9 %	56,1 %	83,6 %	-
LSTM Dual Encodeur (Lowe <i>et al.</i> , 2017b)	55,2 %	72,1 %	92,4 %	-
BiLSTM Dual Encodeur (Kadlec <i>et al.</i> , 2015)	54,2 %	71,6 %	91,9 %	-
Similarité LSTM Dual Encodeur (<i>Sim LSTM DE</i>)	62,9[†] %	78,5[†] %	95,2[†] %	76,1[†] %
Similarité BiLSTM Dual Encodeur (<i>Sim BiLSTM DE</i>)	63,7[†] %	79,1[†] %	95,2[†] %	76,7[†] %

Résultats d'évaluation en utilisant les métriques **Rappel @ k** et le **MRR**

- ✓ Amélioration significative des résultats sur toutes les métriques par rapport aux systèmes état de l'art.
- ✓ Les cellules BiLSTM permettent d'améliorer les résultats.
- ✓ Grâce à cette amélioration, notre système est devenu capable de mieux apparier le contexte avec la meilleure réponse.

Évaluation de l'impact des plongements de mots

- **Word2Vec** : **Corpus** : ensemble d'entraînement, **outil** : word2vec, **vocabulaire** = 770k.
- **FastText** : **Corpus**: Wikipedia, **outil** : FastText, **vocabulaire** = 2.5M.
- **Glove** : **Corpus** : Common Crawl Corpus , **outil** : Glove, **vocabulaire** = 2.2M.

Systeme	Rappel@1	Rappel@2	Rappel@5	MRR
Word2Vec-gelé	62,2 %	77,8 %	94,6 %	75,5 %
Word2Vec-affiné	63,3 %	78,4 %	94,9 %	76,2 %
FastText-gelé	58,9 %	75,1 %	94,2 %	73,2 %
FastText-affiné	61,7 %	77,8 %	94,7 %	75,2 %
Glove-gelé	62,5 %	78,0 %	94,8 %	75,7 %
Glove-affiné	62,9 %	78,5 %	95,2 %	76,1 %

Évaluation de l'impact des plongements de mots

- **Word2Vec** : **Corpus** : ensemble d'entraînement, **outil** : word2vec, **vocabulaire** = 770k.
- **FastText** : **Corpus**: Wikipedia, **outil** : FastText, **vocabulaire** = 2.5M.
- **Glove** : **Corpus** : Common Crawl Corpus , **outil** : Glove, **vocabulaire** = 2.2M.

Système	Rappel@1	Rappel@2	Rappel@5	MRR
Word2Vec-gelé	62,2 %	77,8 %	94,6 %	75,5 %
Word2Vec-affiné	63,3 %	78,4 %	94,9 %	76,2 %
FastText-gelé	58,9 %	75,1 %	94,2 %	73,2 %
FastText-affiné	61,7 %	77,8 %	94,7 %	75,2 %
Glove-gelé	62,5 %	78,0 %	94,8 %	75,7 %
Glove-affiné	62,9 %	78,5 %	95,2 %	76,1 %

- ✓ Pour tous les systèmes, l'affinement des poids de la couche des plongements de mots permet d'améliorer les résultats.

Évaluation de l'impact des plongements de mots

- **Word2Vec** : **Corpus** : ensemble d'entraînement, **outil** : word2vec, **vocabulaire** = 770k.
- **FastText** : **Corpus**: Wikipedia, **outil** : FastText, **vocabulaire** = 2.5M.
- **Glove** : **Corpus** : Common Crawl Corpus , **outil** : Glove, **vocabulaire** = 2.2M.

Systeme	Rappel@1	Rappel@2	Rappel@5	MRR
Word2Vec-gelé	62,2 %	77,8 %	94,6 %	75,5 %
Word2Vec-affiné	63,3 %	78,4 %	94,9 %	76,2 %
FastText-gelé	58,9 %	75,1 %	94,2 %	73,2 %
FastText-affiné	61,7 %	77,8 %	94,7 %	75,2 %
Glove-gelé	62,5 %	78,0 %	94,8 %	75,7 %
Glove-affiné	62,9 %	78,5 %	95,2 %	76,1 %

- ✓ Pour tous les systèmes, l'affinement des poids de la couche des plongements de mots permet d'améliorer les résultats.
- ✓ Les plongements pré-entraînés avec Glove donnent les meilleurs résultats.

Évaluation de l'impact des plongements de mots

- **Word2Vec** : **Corpus** : ensemble d'entraînement, **outil** : word2vec, **vocabulaire** = 770k.
- **FastText** : **Corpus**: Wikipedia, **outil** : FastText, **vocabulaire** = 2.5M.
- **Glove** : **Corpus** : Common Crawl Corpus , **outil** : Glove, **vocabulaire** = 2.2M.

Système	Rappel@1	Rappel@2	Rappel@5	MRR
Word2Vec-gelé	62,2 %	77,8 %	94,6 %	75,5 %
Word2Vec-affiné	63,3 %	78,4 %	94,9 %	76,2 %
FastText-gelé	58,9 %	75,1 %	94,2 %	73,2 %
FastText-affiné	61,7 %	77,8 %	94,7 %	75,2 %
Glove-gelé	62,5 %	78,0 %	94,8 %	75,7 %
Glove-affiné	62,9 %	78,5 %	95,2 %	76,1 %

- ✓ Pour tous les systèmes, l'affinement des poids de la couche des plongements de mots permet d'améliorer les résultats.
- ✓ Les plongements pré-entraînés avec Glove donnent les meilleurs résultats.
- ✓ L'entraînement des plongements de mots au niveau des mots dans le cas de Glove et Word2Vec donnent de meilleurs résultats en comparaison à ceux obtenus au niveau des caractères dans le cas de FastText.

Évaluation qualitative et quantitative des résultats

Contexte	<ul style="list-style-type: none"> - Hello .. Is it possible to disable GPG check for a specific APT repository ? - Why would you ever need to do that - It's for a custom repository in enterprise environment. But that's unimportant isn't it. <p>Was that a statement that it's not possible ?</p>		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,06	0,87	1	3rd party repo ? PPA ? what is it ?
0,29	0,25	0	Find it sticky edge
0,17	0,40	0	That response doesn't help me in the slightest

Notre système retrouve la bonne réponse contrairement au Dual Encodeur de base

Contexte	<ul style="list-style-type: none"> - How can I remount a drive as read/write ? - mount -o rw /dev/whatever /whatever I believe there's a remount option I think - Thanks - I'd say check the mount man page also. I forgot the syntax for the remount option 		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,96	0,49	1	Okay
0,62	0,88	0	That's sound like a good idea find out I'm missing authz_hos somehow
0,14	0,87	0	Thanks I will read that

Le Dual Encodeur de base retrouve la bonne réponse contrairement à notre système

Évaluation qualitative et quantitative des résultats

Contexte	<ul style="list-style-type: none"> - Is there a length limitation on the hostname in SSH? - 255 char for the FQDN - FQDN? 		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,99	0,94	1	Full Qualify Domain Name : mycomputer.kitchen.myhouse.com
0,01	0,27	0	Alright good luck
0,01	0,08	0	You have to do it once the bios hand off to grub

Les deux systèmes retrouvent la bonne réponse

Contexte	<ul style="list-style-type: none"> - Is there a script that can generate a live cd iso of your currently run ubuntu hdd install ? - Remastersys - Have you use it ? 		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,05	0,18	1	I hadn't have much luck with it, but that is a while ago (2 years)
0,88	0,71	0	It can
0,91	0,56	0	Not for me I doubt I could figure it out to be honest, but any theme installations come to mind as a guess

Les deux systèmes ne réussissent pas à trouver la bonne réponse

Évaluation qualitative et quantitative des résultats

Contexte	<ul style="list-style-type: none"> - Is there a length limitation on the hostname in SSH? - 255 char for the FQDN - FQDN? 		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,99	0,94	1	Full Qualify Domain Name : mycomputer.kitchen.myhouse.com
0,01	0,27	0	Alright good luck
0,01			

Boussaha et al. 2018		Réussite	Échec
Lowe et al. 2015			
Réussite		7437 (39.31%)	4476 (23.66%)
Échec		2143 (11.33%)	4864 (25.70%)

Contexte	<ul style="list-style-type: none"> - Remastersys - Have you use it? 		
Lowe et al.	Boussaha et al.	Étiquette	Réponse
0,05	0,18	1	I hadn't have much luck with it, but that is a while ago (2 years)
0,88	0,71	0	It can
0,91	0,56	0	Not for me I doubt I could figure it out to be honest, but any theme installations come to mind as a guess

Les deux systèmes ne réussissent pas à trouver la bonne réponse

Conclusion et perspectives

- Intérêt : assistance automatique à la résolution des problèmes.
- Orientation vers les systèmes d'ordonnement de réponses: conviennent le mieux à notre tâche (à cause du problème de généralisation des systèmes génératifs),
- Nous avons construit un système qui apprend la similarité entre le contexte et la réponse pour apparier le contexte avec la bonne réponse.
- Résultats intéressants que l'on peut améliorer en faisant une analyse des erreurs.
- Futur: utiliser le mécanisme d'attention, les pré-traitements et l'ordonnement par paire (*pairwise ranking*).
- Explorer l'impact des informations discursives tels que les actes de dialogue et les relations rhétoriques.
- Évaluer notre approche sur d'autres corpus et sur d'autres langues (Arabe, Chinois ... etc).

Références

- Lowe, Ryan, Nissan Pow, Iulian Serban, and Joelle Pineau. "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems." arXiv preprint arXiv:1506.08909 (2015).
- Xu, Zhen, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. "Incorporating Loose-Structured Knowledge into LSTM with Recall Gate for Conversation Modeling." arXiv preprint arXiv:1605.05110 (2016).
- Wu, Yu, Wei Wu, Zhoujun Li, and Ming Zhou. "Response Selection with Topic Clues for Retrieval-based Chatbots." arXiv preprint arXiv:1605.00090 (2016).
- Wu, Yu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots." In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 496-505. 2017.
- Lowe, Ryan Thomas, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. "Training end-to-end dialogue systems with the ubuntu dialogue corpus." Dialogue & Discourse 8, no. 1 (2017): 31-65.

- Code implémenté en python avec Keras ayant Tensorflow en backend.
- Code source: https://github.com/basma-b/dual_encoder_udc

Merci !

