UN MODÈLE HIERARCHIQUE POUR LA TRANSDUCTION DATA, TEXTE.

Clément Rebuffel[1,2], Laure Soulier[1], Geoffrey Scoutheeten[2] and Patrick Gallinari[1,3]

Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, F-75005 Paris, France
BNP Paribas, CIB, Analytics Consulting
Criteo AI Lab, Paris

DESCRIPTION DE LA TÂCHE ET ETAT DE L'ART

Formalisation

- Données structurées : tableaux, bases de connaissances, etc.
- La donnée source s est caractérisée comme :
  - un ensemble d'entités $s := \{e_i\}_{i=1}^{I}$
  - Une entité e_i est un ensemble d'enregistrements $\{r_{i,1}, ..., r_{i,j}, ..., r_{i,J_i}\}$; où l'enregistrement $r_{i,j}$ est défini comme une paire *clef*, *valeur*: $k_{i,j}$ $v_{i,j}$.
- Chaque donnée source est associée à une description en langage naturel de T mots $y_{1:T} = (y_1, \ldots, y_T)$.
- Le *dataset* $\mathcal{D}$ est une collection de N paires (donnée, description) (s, y).

Les modèles sont entraînés pour maximiser la log-vraisemblance sur N exemples :

$$\arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \sum_{(s,y)\in\mathcal{D}} \log P(\hat{y} = y \mid s; \theta)$$

où θ représente l'ensemble des paramètres du modèle et $P(\hat{y} = y \mid s; \theta)$ la probabilité de générer la bonne séquence y pour le tableau s.

Principales approches

- Approches génératives de type encodeur-décodeur
  - Encodeur-décodeur : Un Bi-RNN encode les éléments linéarisés de la structure. Un RNN génère mot à mot une description basée sur cet encodage.
    e.g. [(Hawks, H/V, H), (Hawks, WINS, 46), ..., (Magic, H/V, V), ...]
  - Mécanisme d'attention : le réseau décodeur apprend à pondérer différemment les éléments du tableau à chaque pas de temps [2].
  - Mécanisme de Copie : le réseau décodeur apprend à alterner entre génération libre et copie d'éléments présents dans le tableau [3].

Exemple - jeu de données WikiBIO [1]

Frederick Parker-Rhodes

Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
	Scientific career
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, Computer Science
Author abbrev. (botany)	Park.-Rhodes

Description attendue :

"Frederick Parker-Rhodes (21 November 1914 – 2 March 1987) was an English linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist."

Est-ce pertinent dans le cas de données complexes ?

- Plusieurs entités qui sont caractérisés par plusieurs enregistrements

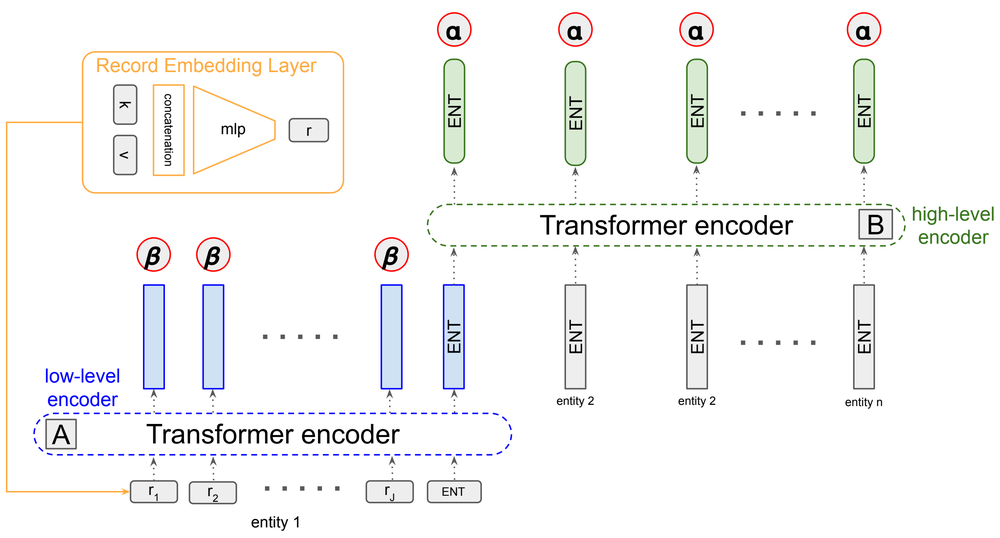
TEAM	H/V	WINS	LOSSES	PTS	REB	AST	...
Hawks	H	46	12	95	42	27	...
Magic	V	19	41	88	40	22	...

PLAYER	PTS	REB	AST	STL	BLK	CITY	...
Al Horford	17	13	4	2	0	Atlanta	...
Kyle Korver	8	3	2	1	2	Atlanta	...
Jeff Teague	17	0	7	2	0	Atlanta	...
N. Vucevic	21	15	3	1	1	Orlando	...
Tobias Harris	15	4	1	2	1	Orlando	...
...	...	...	...	...	...	...	...

H/V: home or visiting; PTS: points; REB: rebounds; AST: assists; STL: steals; BLK: blocks

The **Atlanta Hawks** (**46-12**) beat the **Orlando Magic** (**19-41**) **95-88** on Friday. **Al Horford** had a good all-around game, putting up **17 points, 13 rebounds, four assists and two steals** in a tough matchup against **Nikola Vucevic**. **Kyle Korver** was the lone Atlanta starter not to reach double figures in points. **Jeff Teague** bounced back from an illness, he scored **17 points** to go along with **seven assists and two steals**. After a rough start to the month, the **Hawks** have won three straight and sit atop the Eastern Conference with a nine game lead on the second place Toronto Raptors. The **Magic** lost in devastating fashion to the Miami Heat in overtime Wednesday. They blew a seven point lead with 43 seconds remaining and they might have carried that with them into Friday's contest against the **Hawks**. **Vucevic** led the **Magic** with **21 points and 15 rebounds**. **Aaron Gordon** (ankle) and **Evan Fournier** (hip) were unable to play due to injury. The **Magic** have four teams between them and the eighth and final playoff spot in the Eastern Conference. The **Magic** will host the Charlotte Hornets on Sunday, and the **Hawks** with take on the Heat in Miami on Saturday.

MODÈLE HIÉRARCHIQUE POUR LE DATA-TO-TEXT



Originalité du modèle : éléments considérés comme non ordonnés et exploitation de l'architecture Transformer [4].

- Record Embedding Layer: $\mathbf{r}_{i,j} = \text{ReLU}(\mathbf{W}_r[\mathbf{k}_{i,j}; \mathbf{v}_{i,j}] + \mathbf{b}_r)$
- Low-level encodeur : encoder la collection d'enregistrement pour une même entité.
- High-level encodeur : encode la collection d'entités pour une
- Attention hiérarchique : identifier l'entité importante puis son enregistrement important

$$\mathbf{c_t} = \sum_{i=1}^{I}(\alpha_{i,t}(\sum_{j}\beta_{i,j,t}\mathbf{r}_{i,j})) \quad (1)$$

$$\text{où} \quad \alpha_{i,t} \propto exp(\mathbf{d}_t\mathbf{W}_\alpha\mathbf{e}_i) \quad et \quad \beta_{i,j,t} \propto exp(\mathbf{d}_t\mathbf{W}_\beta\mathbf{h}_{i,j}) \quad (2)$$

EXPÉRIMENTATIONS

PROTOCOLE

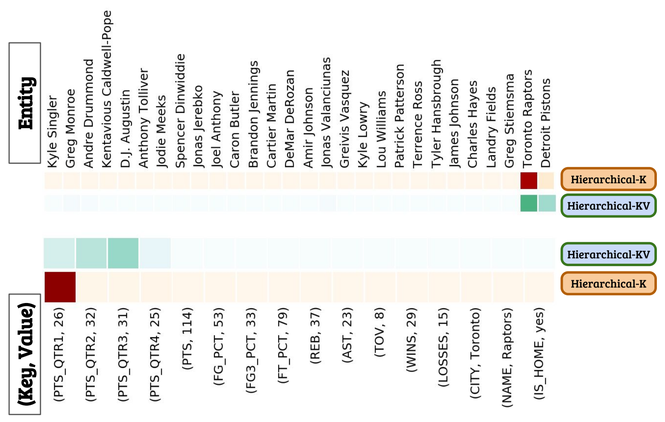
- Jeu de données : Rotowire. 4000 tableaux de statistiques associés à des descriptions écrites par des professionnels. Descriptions: 337 mots, vocabulaire de 11,000 mots uniques. Tableaux: 39 noms de colonnes, 628 cellules, 28 entités.
- Métriques :
  - BLEU : ressemblance par rapport au résumé attendu
  - RG : quantité d'éléments factuels dans le résumé généré
  - CS : quantité d'éléments factuels par rapport au résumé attendu
  - CO : ordre des éléments factuels par rapport au résumé attendu
- Modèles de référence :
  - *Wiseman* [5] standard encodeur-decodeur avec mécanisme de copie.
  - *Li* [6] standard encodeur-decodeur avec mécanisme de copie retardé : texte à trou généré et rempli par des éléments des données avec un réseau à pointeur.
  - *Puduppully-plan* [7] modèle de planification des enregistrements qui génère ensuite le texte associé.
  - *Puduppully-updt* [8] standard encodeur-decodeur avec un mécanisme d'attention qui met à jour les représentations des éléments.
  - *Flat*: notre modèle avec Transformer sans hiérarchie

RÉSULTATS

Efficacité du modèle.

	BLEU	RG		CS			CO	Nb
		P%	#	P%	R%	F1		Params
Gold descriptions	100	96.11	17.31	100	100	1	100	
Wiseman	14.5	75.62	**16.83**	32.80	39.93	36.2	15.62	45M
Li	16.19	84.86	19.31	30.81	38.79	34.34	16.34	-
Pudupully-plan	16.5	87.47	34.28	34.18	51.22	41	18.58	35M
Puduppully-updt	16.2	**92.69**	30.11	38.64	48.51	43.01	**20.17**	23M
Flat	$16.7_{.2}$	76.62_{1}	$18.54_{.6}$	$31.67_{.7}$	42.9_{1}	$36.42_{.4}$	$14.64_{.3}$	14M
Hierarchical-kv	$17_{.3}$	89.04_{1}	$21.46_{.9}$	$38.57_{1.2}$	$51.50_{.9}$	$44.19_{.7}$	$18.70_{.7}$	14M
Hierarchical-k	$\mathbf{17.5}_{.3}$	$89.46_{1.4}$	$21.17_{1.4}$	$\mathbf{39.47}_{1.4}$	$\mathbf{51.64}_{1}$	$\mathbf{44.7}_{.6}$	$18.90_{.7}$	14M

Analyse des modules d'attention.



Hierarchical-K
The Raptors got off to a quick start in this one, out-scoring the Pistons 26 - 25 in the first quarter alone. [...]

Hierarchical-KV
The Raptors got off to a quick start in this game, out-scoring the Pistons 31 - 25 right away in the first quarter. [...]

RÉFÉRENCES

[1] Lebret R., Grangier D., Auli M., Neural Text Generation from Structured Data with Application to the Biography Domain. EMNLP, 2016
[2] Bahdanau D., Cho K., Bengio Y., Neural Machine ICLR 2015
[3] Gulcehre C., Ahn S., Nallapati R., Zhou B., Bengio Y., Pointing the Unknown Words. ACL 2016
[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Kaiser, L., Polosukhin, I., Attention is all you need. NIPS, 2017
[5] Wiseman S., Shieber S. M., Rush A. M., Challenges in Data-to-Document Generation. EMNLP, 2017
[6] Li, L., Wan, X., Point Precisely: Towards Ensuring the Precision of Data in Generated Texts Using Delayed Copy Mechanism. ACL, 2018
[7] Puduppully R., Dong L., Lapata M., Data-to-Text Generation with Content Selection and Planning. AAAI, 2018
[8] Puduppully R., Dong L., Lapata M., Data-to-text Generation with Entity Modeling. ACL, 2019