L'EXTRACTION DES CONNAISSANCES À L'HEURE DES MODÈLES DE LANGUE

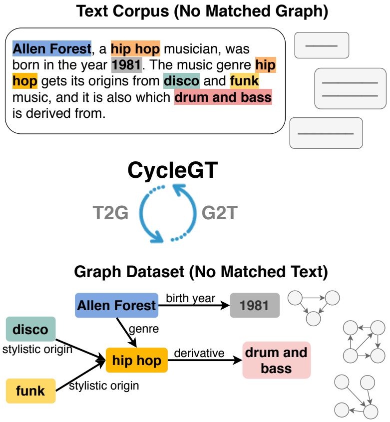
23 Octobre 2023
Séminaire MAIAGE

Vincent Guigue
vincent.guigue@agroparistech.fr
https://vguigue.github.io

MIA PARIS-SACLAY EKINOCS

AGROPARISTECH Institut des Sciences et Industries du Vivant et de l'Environnement

Langage humain *vs* Langage machine

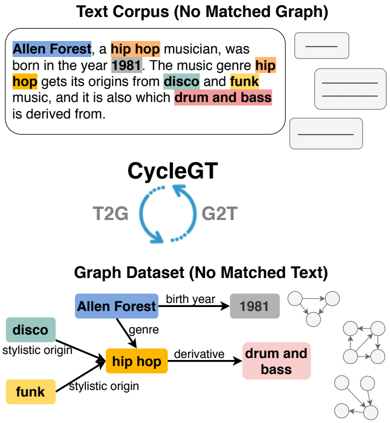


[1]

[1] Qipeng Guo et al. (2020). "CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training". In: CoRR

Langage humain *vs* Langage machine

- Indexation / Recherche d'Information

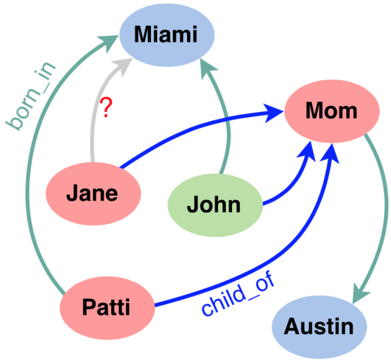


[1]

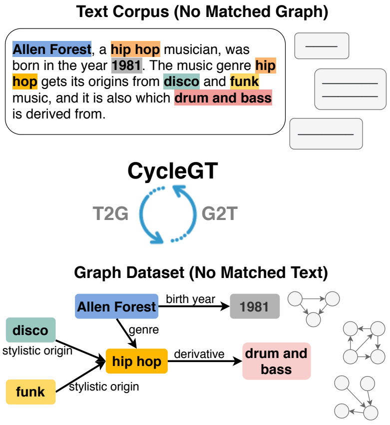
[1] Qipeng Guo et al. (2020). "CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training". In: CoRR

Langage humain *vs* Langage machine

- Indexation / Recherche d'Information
- Raisonnement / Complétion des connaissances



1



2

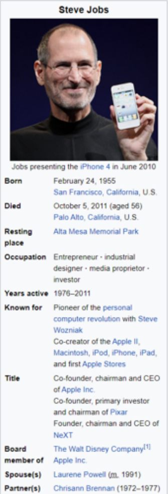
---

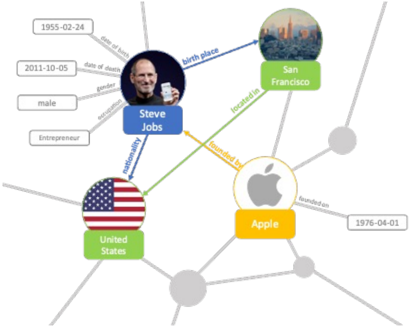
[1] Dat Quoc Nguyen

[2] Qipeng Guo et al. (2020). "CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training". In: CoRR

Enjeux autour des bases de connaissances

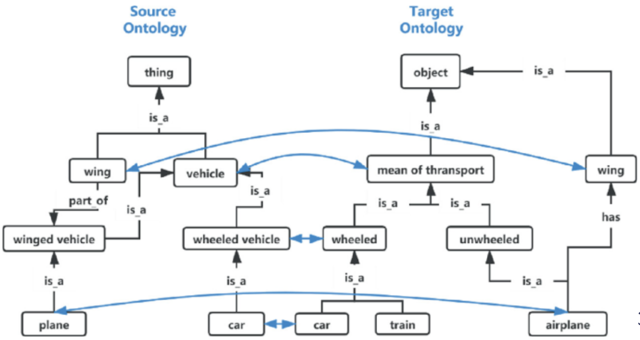
- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts





Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts
- Connexions w/ Machine Learning
  - Alignement / fusion
  - Plongement / TransE
  - Modèle de langue



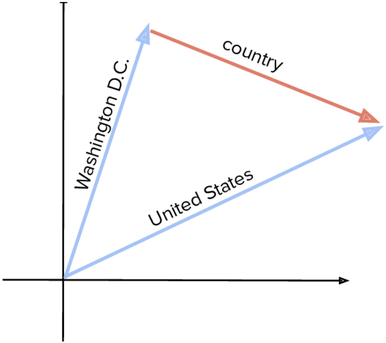
3

[3] Xiaojing Wu, Xingsi Xue, and Wenyu Hu (2021). "Argumentation Based Ontology Alignment Extraction". In: Advanced Machine Learning Technologies and Applications. ISBN: 978-3-030-69717-4

Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts

- Connexions w/ Machine Learning
  - Alignement / fusion
  - Plongement / TransE
  - Modèle de langue



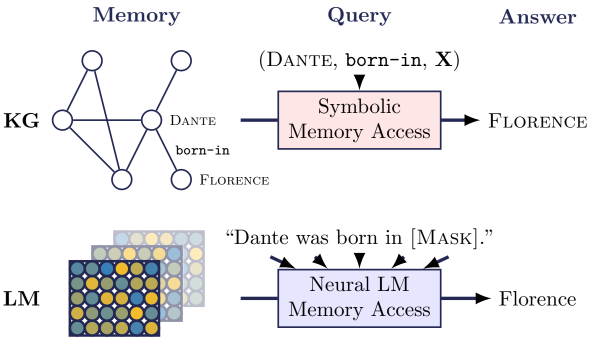
[3]

[3] Antoine Bordes et al. (2013). "Translating embeddings for modeling multi-relational data". In: NeurIPS

Enjeux autour des bases de connaissances

- Construire des bases de connaissances
- Raisonner: règles + inférence logique, ontologies, systèmes experts

- Connexions w/ Machine Learning
  - Alignement / fusion
  - Plongement / TransE
  - Modèle de langue

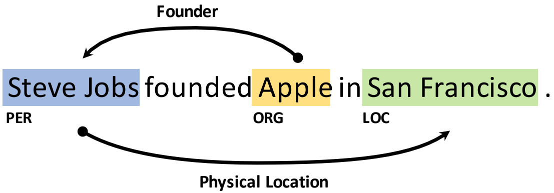


3

- Exhaustivité?
- Fiabilité?

[3] Fabio Petroni et al. (2019). "Language Models as Knowledge Bases?" In: EMNLP. Association for Computational Linguistics

Challenges autour de l'extraction d'information



- Segmenter les entités
- Identifier et/ou typer les entités
- Identifier + classer les liens

- Défi de la segmentation: e.g. New York Times
- Polysémie
- Fautes d'orthographe

$\Rightarrow$ Morphologie + sémantique + contexte

Challenges autour de l'extraction d'information

Founder

Steve Jobs founded Apple in San Francisco .

PER ORG LOC

Physical Location

- Segmenter les entités
- Identifier et/ou typer les entités
- Identifier + classer les liens

- Défi de la segmentation: e.g. New York Times
- Polysémie
- Fautes d'orthographe $\Rightarrow$ Morphologie + sémantique + contexte


Boston


Washington


Philadelphia

Bilan préliminaire

Base de connaissances

+ Efficace / passage à l'échelle
+ Garanties sur les résultats

– Cout de construction / maintenance
– Manque de robustesse aux erreurs

Extraction d'information

– Cout d'exploitation (LLM)
– Manque de garanties/fiabilité

+ Cout de construction/MAJ
+ Robustesse aux erreurs

Des outils aux caractéristiques complémentaires

Named Entity Recognition
Relation Extraction
& Généralisation

[Thèse de Bruno Taillé]

Extraction des entités nommées

Steve	Jobs	founded	Apple	in	San	Francisco	.
B-PER	**E-PER**	O	**S-ORG**	O	**B-LOC**	**E-LOC**	O

IOBES : O = Other (not in an entity)
B = Beginning I = Inside E = End S = Single [4]

Approches historiques:

- Extraction de caractéristiques (majuscules, terminaisons, lexiques, ...)
- Modélisation des probabilités dans la séquence
  - Chaines de Markov Cachées –*HMM*–
  - Champs Aléatoires Conditionnels –*CRF*–

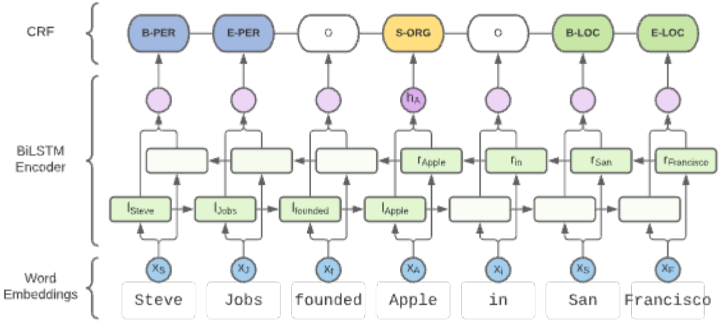
[4] Bruno Taillé (2022). "Contextualization and Generalization in Entity and Relation Extraction". PhD thesis. Sorbonne Université

Extraction des entités nommées

Steve	Jobs	founded	Apple	in	San	Francisco	.
B-PER	**E-PER**	O	**S-ORG**	O	**B-LOC**	**E-LOC**	O

IOBES : **O = Other (not in an entity)**
B = Beginning **I = Inside** **E = End** **S = Single** [4]

Révolutions successives: représentation des mots & contextualisation



- **Pretrained word embeddings**
  (Huang 2015) SENNA
- **Character-level word embeddings**
  (Lampe 2016) SENNA + char-BiLSTM
- **Contextualized embeddings**
  (Peters 2018) ELMo
  (Akbik 2018) Flair
  (Devlin 2019) BERT

4 Bruno Taillé (2022). "Contextualization and Generalization in Entity and Relation Extraction".

Extraction des entités nommées

ELMo (Peters 2018)
- **char-CNN** word representation **(ELMo[0])**
- **BiLSTM** LM at a **word** level
- Weighted sum fusion (learned weights)

Flair (Akbik 2018)
- **BiLSTM** LM at a **character** level
- Word represented with the concatenation of its ends

BERT (Devlin 2019)
- **Transformer** LM at a **subword** level (WordPiece)
- Masked LM and Next Sentence Prediction
- **$\text{BERT}_{\text{LARGE}}$ feature-based = frozen LM**

CoNLL03 Test Set (F1)		
$\text{BERT}_{\text{LARGE}}$	(Devlin 2019)	92.8
ELMo	(Peters 2018)	92.2
Flair	(Akbik 2018)	92.0*
TagLM (SENNA + LM)	(Peters 2017)	91.9
SENNA + char BiLSTM	(Lample 2016)	90.9
SENNA	(Huang 2015)	88.8

(Peters 2018) Deep contextualized word representations, NAACL-HLT 2018
(Akbik 2018) Contextual String Embeddings for Sequence Labeling, COLING 2018
(Devlin 2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL-HLT 2019

Superposition lexicale: apprentissage *vs* test

Proportion of mentions in test set are seen during training.

3 types of mentions :

Exact match	Mention seen with the same type
Partial match	At least one non stop-word seen in a mention of same type
New	All non stop-words are new

Train :	Georges Washington	(PER)
	Barack Obama	(PER)
Test :	**Donald Trump**	**(PER)**
	Barack Obama	**(PER)**
	Georges Bush	**(PER)**
	Washington DC.	**(LOC)**
	Obama	**(PER)**

(Augenstein 2017) Generalisation in named entity recognition: A quantitative analysis, CSL 2017
(Moosavi 2017) Lexical Features in Coreference Resolution: To be Used With Caution, ACL 2017

Superposition lexicale: apprentissage *vs* test

CoNLL 2003
- News articles
- 4 languages (**English**, German, Dutch, Spanish)
- 4 types (PER, ORG, LOC, MISC)

OntoNotes 5.0
- 6 genres (news, conversations, web…)
- 3 languages (**English**, Arab, Chinese)
- 18 types (11 entities + 7 values)

WNUT 17 (Workshop on Noisy User-generated Text)
- Web Text (Twitter, Reddit, Youtube, Stack Overflow)
- **English**
- 6 types (PER, LOC, Corporation, Group, Creative Work, Product)

2.6 Entity Names Annotation

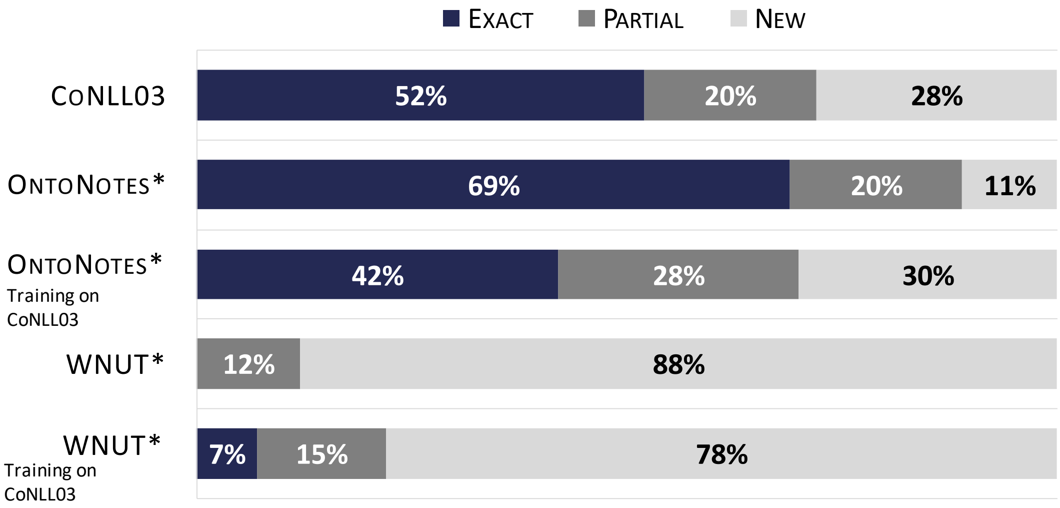
Names (often referred to as "Named Entities") are annotated according to the following set of types:

PERSON	People, including fictional
NORP	Nationalities or religious or political groups
FACILITY	Buildings, airports, highways, bridges, etc.
ORGANIZATION	Companies, agencies, institutions, etc.
GPE	Countries, cities, states
LOCATION	Non-GPE locations, mountain ranges, bodies of water
PRODUCT	Vehicles, weapons, foods, etc. (Not services)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK OF ART	Titles of books, songs, etc.
LAW	Named documents made into laws
LANGUAGE	Any named language

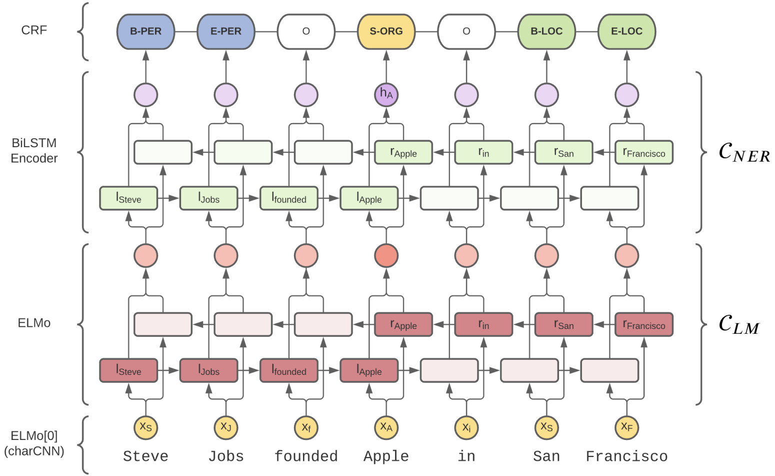
The following values are also annotated in a style similar to names:

DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
PERCENT	Percentage (including "%")
MONEY	Monetary values, including unit
QUANTITY	Measurements, as of weight or distance
ORDINAL	"first", "second"
CARDINAL	Numerals that do not fall under another type

Superposition lexicale: apprentissage *vs* test



Séparation des performances: les modèles

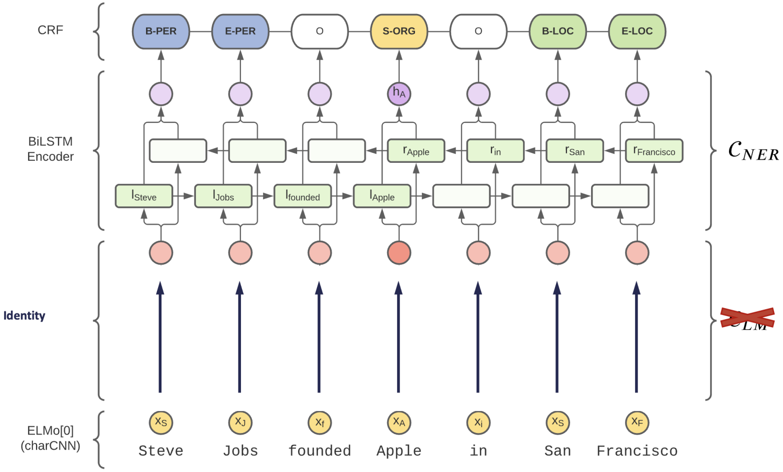


BiLSTM + ELMo

$C_{NER} + C_{LM}$

Transformer = contextualisation globale (*vs* LSTM = locale)

Séparation des performances: les modèles

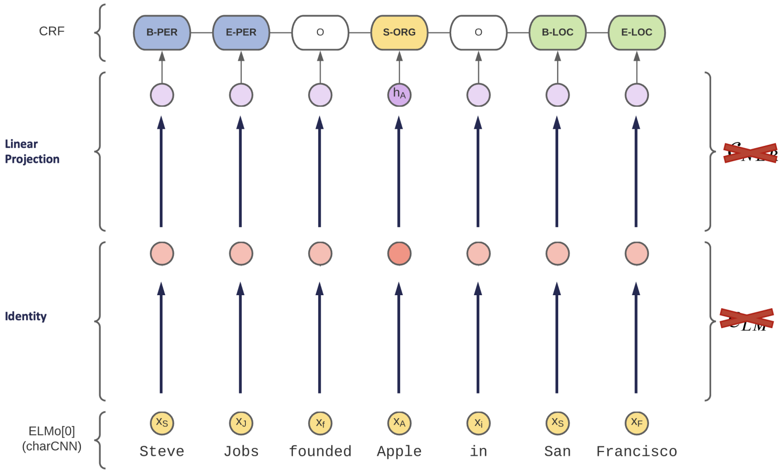


BiLSTM + ELMo $C_{NER} + C_{LM}$

BiLSTM + ELMo[0] C_{NER}

Transformer = contextualisation globale (*vs* LSTM = locale)

Séparation des performances: les modèles



BiLSTM + ELMo	$C_{NER} + C_{LM}$
BiLSTM + ELMo[0]	C_{NER}
Map + ELMo	C_{LM}
Map + ELMo[0]	no context

Transformer = contextualisation globale (*vs* LSTM = locale)

Séparation des performances: les résultats

OntoNotes*

C_{LM} +9.6 %

C_{NER} +7.3 %

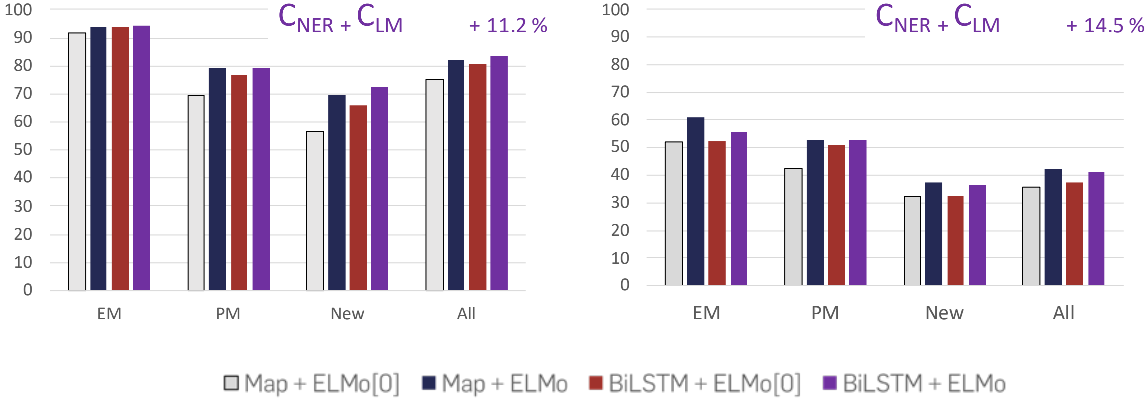
$C_{NER} + C_{LM}$ + 11.2 %

WNUT*

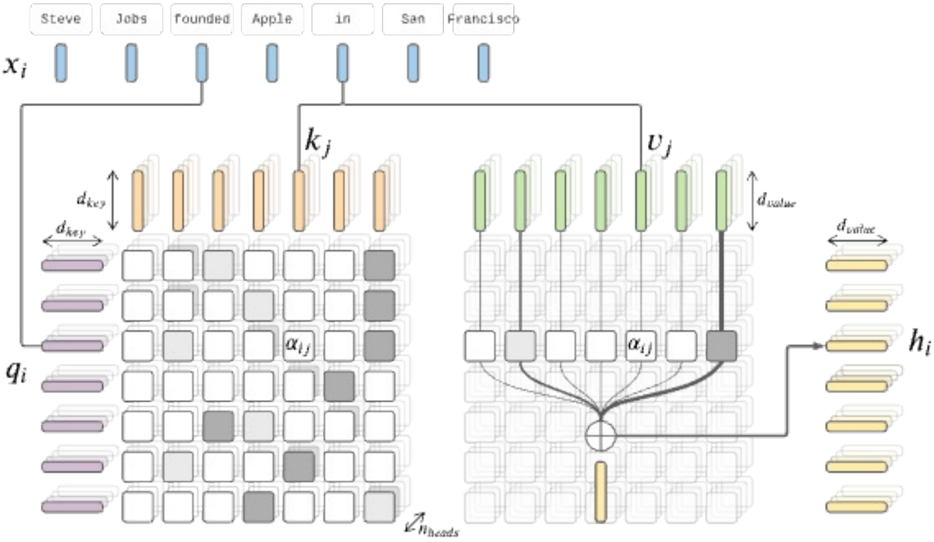
C_{LM} +18.4 %

C_{NER} +5.0 %

$C_{NER} + C_{LM}$ + 14.5 %



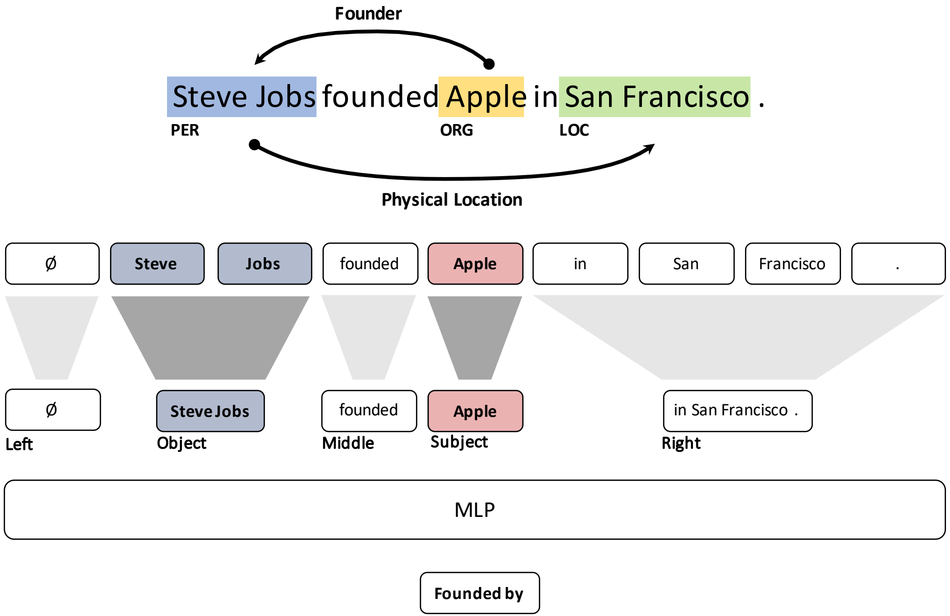
Transformer & NER



$$s_{ij} = \frac{q_i^T k_j}{\sqrt{d_k}} \qquad \alpha_{ij} = \frac{exp(s_{ij})}{\sum_j exp(s_{ij})} \qquad h_i = \sum_j \alpha_{ij} v_j$$

(Vaswani 2017) Attention is all you need, NeurIPS 2017

Extraction de relation: pipeline & piecewise pooling



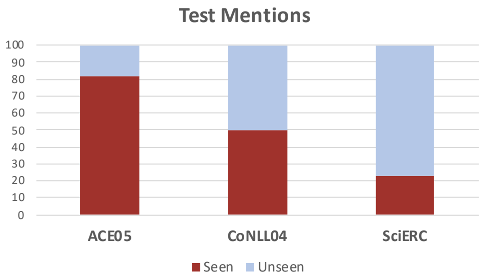
Superposition des ensembles d'apprentissage & test

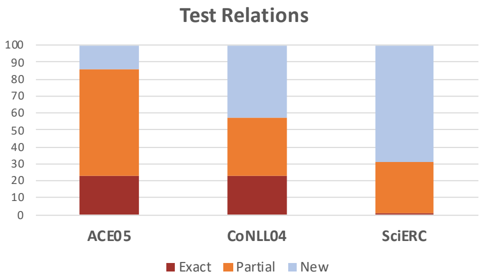
NER
(Augenstein 2017, Taillé 2020)

Seen — Exact Match with the same type
Unseen

RE

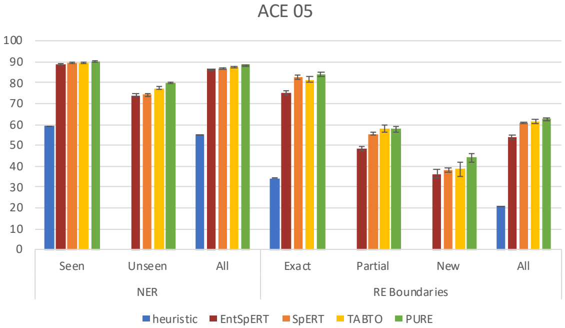
Exact Match — Triple **(head, predicate, tail)** exactly seen during training
Partial Match — **(head, predicate, ...)** or **(..., predicate, tail)** seen during training
New — Otherwise

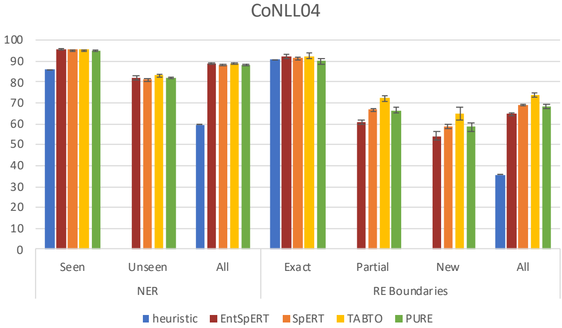




Relation Extraction *vs* End-to-end Relation Extraction

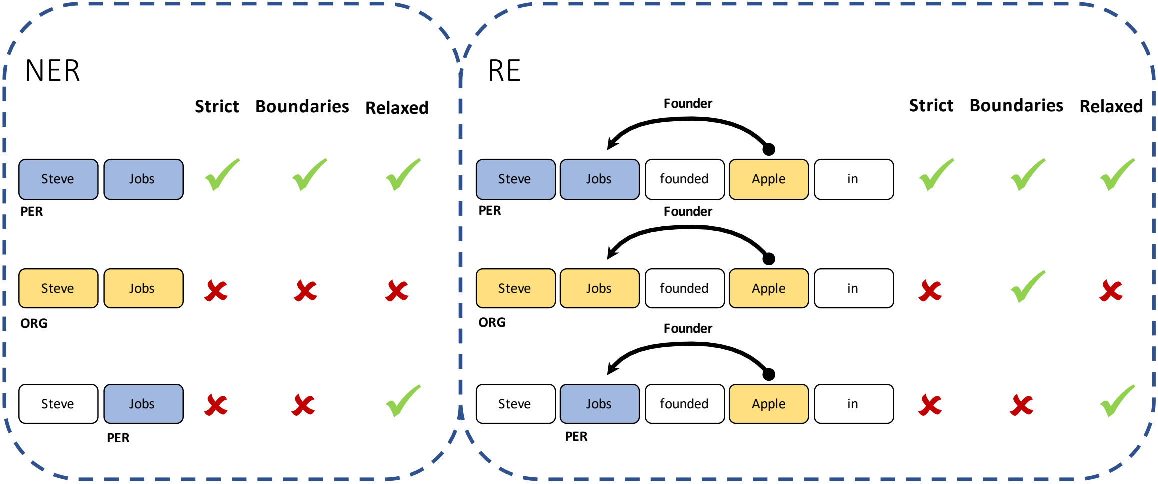
Superposition des ensembles d'apprentissage & test





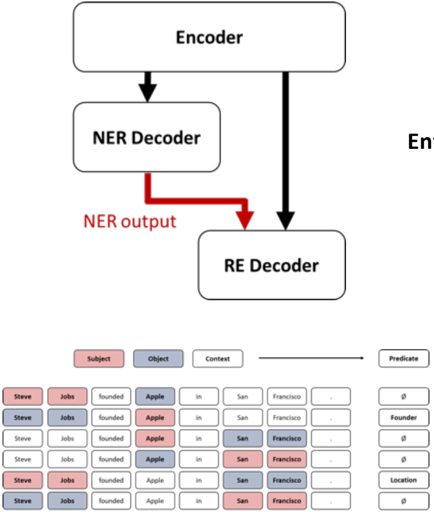
Relation Extraction *vs* End-to-end Relation Extraction

Problème supplémentaire: la définition des métriques



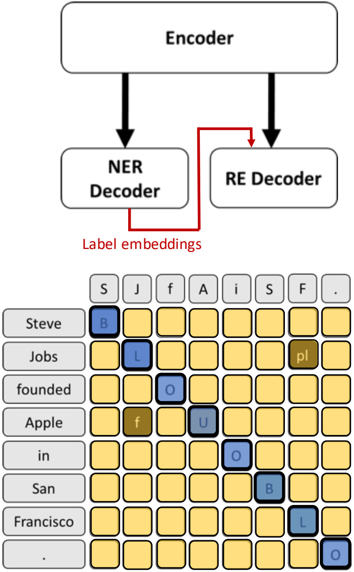
(Bekoulis 2018a) Joint Entity Recognition and Relation Extraction as a Multi-head Selection Problem, Expert Systems with Applications 2018

Dépasser les approches *pipelines*... Un challenge!



Entity Filtering

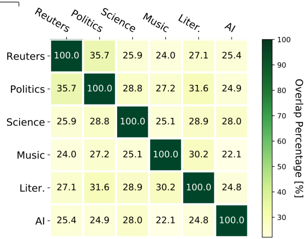
Multi-Head Selection



Conclusions intermédiaires

- Les performances en extraction d'information sont **sur-évaluées**
  - Mémorisation *vs* Généralisation
  - Nouveaux benchmark: WNUT, Cross-NER
- Le problème de **transfert** est critique:
  - En cout d'étiquetage ET en perte de performances
- **Modifier** un modèle de langue en profondeur = mauvaise idée
  - RLHF
  - Modification légère: Prefix tuning, adapter, LoRA, pipeline

Domain	Unlabeled Corpus			Labeled NER			Entity Categories
	# paragraph	# sentence	# tokens	# Train	# Dev	# Test	
Reuters	-	-	-	14,987	3,466	3,684	person, organization, location, miscellaneous
Politics	2.76M	9.07M	176.56M	200	541	651	politician, person, organization, political party, event, election, country, location, miscellaneous
Natural Science	1.72M	5.32M	98.50M	200	450	543	scientist, person, university, organization, country, location, discipli enzyme, protein, chemical compound, chemical element, event, astronomical object, academic journal, award, theory, miscellaneou
Music	3.49M	9.82M	194.62M	100	380	456	music genre, song, band, album, musical artist, musical instrumen award, event, country, location, organization, person, miscellaneou
Literature	2.69M	9.17M	177.33M	100	400	416	book, writer, award, poem, event, magazine, person, location, organization, country, miscellaneous
Artificial Intelligence	97.04K	287.62K	5.20M	100	350	431	field, task, product, algorithm, researcher, metrics, university, country, person, organization, location, miscellaneous



[4]

[4] Zihan Liu et al. (2020). "CrossNER: Evaluating Cross-Domain Named Entity Recognition". In: arXiv

Reconnaissance d'entités nommées: pistes de recherche & exploitation des algorithmes

Auto-supervision (ex-distillation)

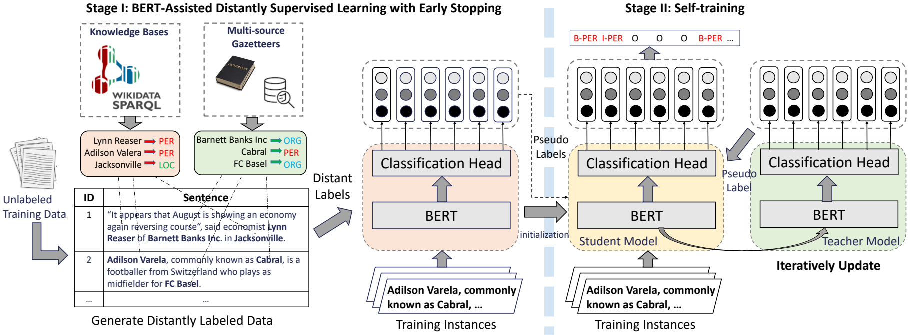
Processus NER standard

1. Liste d'entités
2. Etiquetage automatique du corpus

   regex

3. Inférence sur le test
4. Ré-apprentissage

   Teacher-student



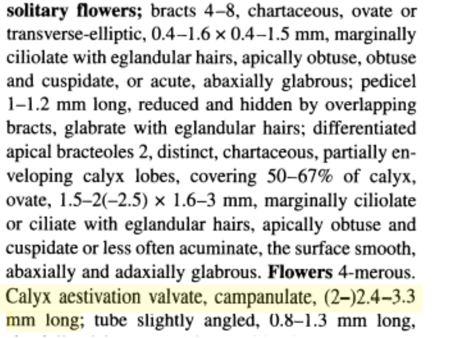
5

[5] Chen Liang et al. (2020). "Bond: Bert-assisted open-domain named entity recognition with distant supervision". In: ACM SIGKDD

Application à l'analyse des descriptions de flores

Extraction d'Information ⇒ Clé-valeur

<ORGAN> Flowers </ORGAN> 4-merous. Calyx aestivation valvate, campanulate, 2-3.6mm long, abaxially **<DESC-SURFACE> glabrous </DESC-SURFACE>**



solitary flowers; bracts 4–8, chartaceous, ovate or transverse-elliptic, 0.4–1.6 × 0.4–1.5 mm, marginally ciliolate with eglandular hairs, apically obtuse, obtuse and cuspidate, or acute, abaxially glabrous; pedicel 1–1.2 mm long, reduced and hidden by overlapping bracts, glabrate with eglandular hairs; differentiated apical bracteoles 2, distinct, chartaceous, partially enveloping calyx lobes, covering 50–67% of calyx, ovate, 1.5–2(–2.5) × 1.6–3 mm, marginally ciliolate or ciliate with eglandular hairs, apically obtuse and cuspidate or less often acuminate, the surface smooth, abaxially and adaxially glabrous. **Flowers** 4-merous. Calyx aestivation valvate, campanulate, (2–)2.4–3.3 mm long; tube slightly angled, 0.8–1.3 mm long,

TABLE I

STATISTICS ON THE DATASET : CLASSES, NUMBER OR DISTINCT WORDS IN EACH CLASS AND NUMBER OF OCCURRENCES IN THE CORPUS.

Set	Class	Occurrences	Number of words
$\mathcal{Y}_0$	Flower	22890	23
	Fruit	4968	10
	Habit	1920	3
	Leaf	4364	5
	Part-of	23849	25
	Stem-root	3296	7
$\mathcal{Y}_1$	Color	18342	15
	Disposition	8405	21
	Form	24816	64
	Position	10936	13
	Surface-texture	18325	23

[6]

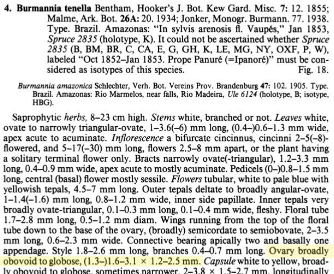
[6] Maya Sahraoui et al. (2022). "NEARSIDE: Structured kNowledge Extraction frAmework from SpecIes DEscriptions". In: Biodiversity Information Science and Standards

Application à l'analyse des descriptions de flores

Models	Precision	Recall	Score F1
Baseline	100/93.83	75.74/70.82	86.19/79.26
Baseline w/ lm	100/95.15	85.28/80.82	92.05/86.54
Baseline w/self-train	100/94.42	84.29/80.15	91.47/86.22

MODEL'S ABILITY TO DETECT AND CLASSIFY NEW ENTITIES, OUT OF THE TRAIN SET'S DISTRIBUTION. (DETECTION/CLASSIFICATION SCORES)

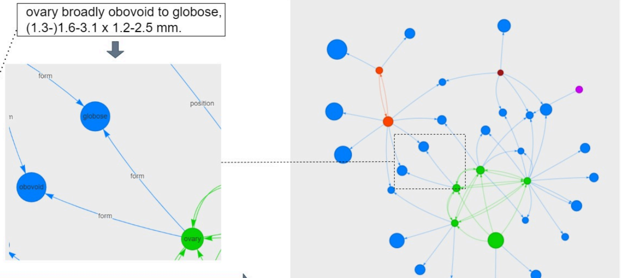
Models	Precision	Recall	Score F1
Baseline	100/92.33	64.78/54.52	78.62/62.76
Baseline w/ lm	100/89.88	69.21/57.73	81.80/65.17
Baseline w/self-train	100/90.76	68.95/57.82	81.62/64.90



4. **Burmannia tenella** Bentham, Hooker's J. Bot. Kew Gard. Misc. **7:** 12. 1855; Malme, Ark. Bot. **26A:** 20. 1934; Jonker, Monogr. Burmann. 77. 1938. Type. Brazil. Amazonas: "In sylvis arenosis fl. Vaupés," Jan 1853, *Spruce 2835* (holotype, K). It could not be ascertained whether *Spruce 2835* (B, BM, BR, C, CA, E, G, GH, K, LE, MG, NY, OXF, P, W), labeled "Oct 1852–Jan 1853. Prope Panuré (=Ipanoré)" must be considered as isotypes of this species. Fig. 18.

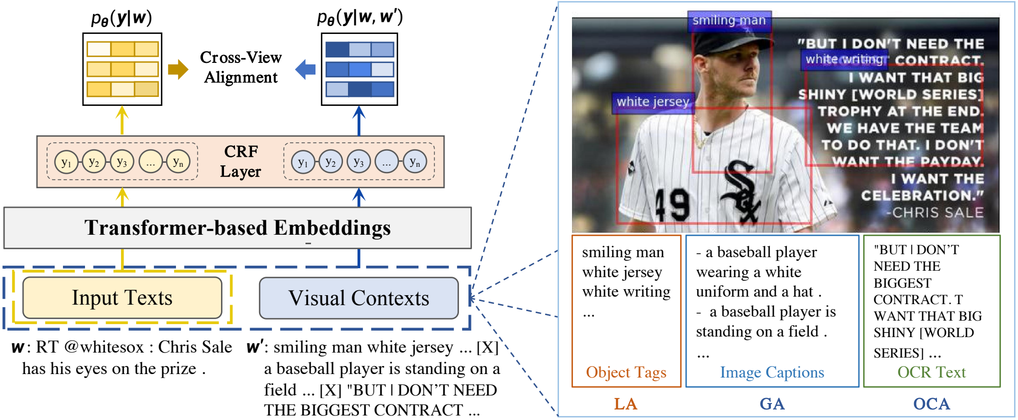
Burmannia amazonica Schlechter, Verh. Bot. Vereins Prov. Brandenburg **47:** 102. 1905. Type. Brazil. Amazonas: Rio Marmelos, near falls, Rio Madeira, *Ule 6124* (holotype, B; isotype, HBG).

Saprophytic *herbs*, 8–23 cm high. *Stems* white, branched or not. *Leaves* white, ovate to narrowly triangular-ovate, 1–3.6(–6) mm long, (0.4–)0.6–1.3 mm wide, apex acute to acuminate. *Inflorescence* a bifurcate cincinnus, cincinni 2–5(–8)-flowered, and 5–17(–30) mm long, flowers 2.5–8 mm apart, or the plant having a solitary terminal flower only. Bracts narrowly ovate(-triangular), 1.2–3.3 mm long, 0.4–0.9 mm wide, apex acute to mostly acuminate. Pedicels (0–)0.8–1.5 mm long, central (basal) flower mostly sessile. *Flowers* tubular, white to pale blue with yellowish tepals, 4.5–7 mm long. Outer tepals deltate to broadly angular-ovate, 1–1.4(–1.6) mm long, 0.8–1.2 mm wide, inner side papillate. Inner tepals very broadly ovate-triangular, 0.1–0.3 mm long, 0.1–0.4 mm wide, fleshy. Floral tube 1.7–2.8 mm long, 0.5–1.2 mm diam. Wings running from the top of the floral tube down to the base of the ovary, (broadly) semicordate to semiobovate, 2–3.5 mm long, 0.6–2.3 mm wide. Connective bearing apically two and basally one appendage. Style 1.8–2.6 mm long, branches 0.4–0.7 mm long. Ovary broadly obovoid to globose, (1.3–)1.6–3.1 × 1.2–2.5 mm. *Capsule* white to yellow, broadly obovoid to globose, sometimes narrower, 2–3.8 × 1.5–2.7 mm, longitudinally



Perspective: extension vers la multimodalité

⇒ Retrouver les **entités dans les images** à partir d'approche texte/image

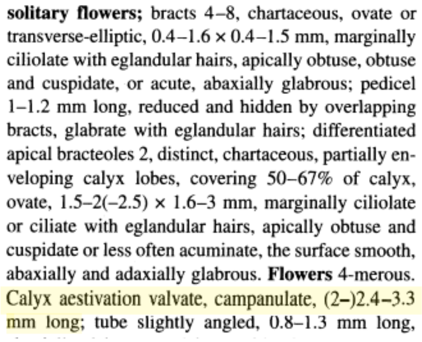


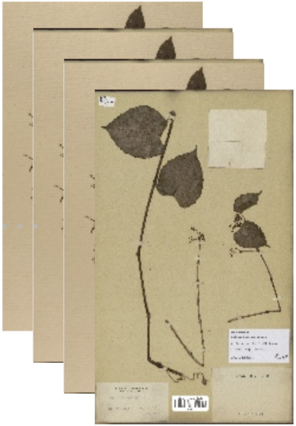
[6]

6 Xinyu Wang et al. (2022). "ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition". In: NAACL

Perspective: extension vers la multimodalité

⇒ Retrouver les **entités dans les images** à partir d'approche texte/image


solitary flowers; bracts 4–8, chartaceous, ovate or transverse-elliptic, 0.4–1.6 × 0.4–1.5 mm, marginally ciliolate with eglandular hairs, apically obtuse, obtuse and cuspidate, or acute, abaxially glabrous; pedicel 1–1.2 mm long, reduced and hidden by overlapping bracts, glabrate with eglandular hairs; differentiated apical bracteoles 2, distinct, chartaceous, partially enveloping calyx lobes, covering 50–67% of calyx, ovate, 1.5–2(–2.5) × 1.6–3 mm, marginally ciliolate or ciliate with eglandular hairs, apically obtuse and cuspidate or less often acuminate, the surface smooth, abaxially and adaxially glabrous. **Flowers** 4-merous. Calyx aestivation valvate, campanulate, (2–)2.4–3.3 mm long; tube slightly angled, 0.8–1.3 mm long,



⇒ Construire des systèmes pédagogiques pour l'identification de taxons

Dynamic NER

Cas extrême où les entités changent de type tout le temps!

Exemple: détecter les joueurs de NBA... Avec le résultat du match: victoire/défaite

A trio of 20 - point - plus efforts and a 17 - rebound night helped hand the Cavs a surprising home loss , their first defeat of the season overall . Dennis Schroder ' s season - high 28 points led the way , while Kent Bazemore put together a stellar 25 - point tally while often going up against LeBron James ' typically stingy defense . Dwight Howard dominated down low with 17 boards , 15 of them on the defensive glass . Atlanta managed a strong 51 percent success rate from the field , helping to key the victory . Kyrie Irving posted 29 points , which came on a season - high 27 shot attempts . Kevin Love ' s 24 - point , 12 - rebound double - double was next , while LeBron James posted 23 points . Poor shooting was Cleveland ' s undoing , as they posted a 37 percent success rate from the field , and 26 percent on 42 shot attempts from beyond the arc .

Same entity
Different context
Different Label

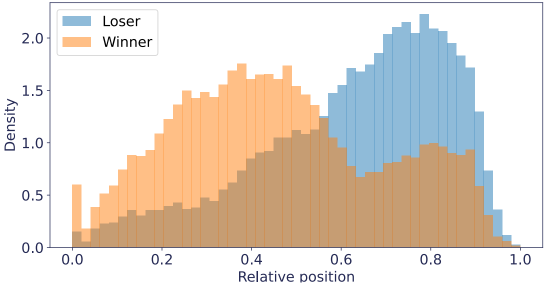
LeBron James and Kyrie Irving stepped up for a second straight night in Kevin Love ' s absence , combining for 60 points on 23 - of - 41 shooting . Irving added a career - high 13 assists , six rebounds and a steal , while James posted nine rebounds and six assists . Richard Jefferson supplied 10 points in Love ' s stead , and Tristan Thompson hauled in 15 rebounds . A pair of 10 - point efforts from Channing Frye and Iman Shumpert paced the second unit . Giannis Antetokounmpo ' s 28 points led Milwaukee , and Jabari Parker was right behind him with 27 points , as the duo tried to keep pace with Cleveland ' s Big Two . However , John Henson , Tony Snell , and Matthew Dellavedova , the remaining members of the first unit , could only combine for nine points between them . Malcolm Brodgon supplied 11 points off the bench as the only other double - digit scorer . [6]

[6] Tristan Luiggi et al. (2023). "Dynamic Named Entity Recognition". In: ACM SAC

Discussion D-NER

- Proposition de nouvelle ressource

Models	Set	RotoWire DNET	RotoWire DNER	RotoWire Entity
BERT-Linear	Seen	0.81	0.66	0.86
	Seen/Unseen	0.81	0.65	0.85
	Unseen	0.80	0.63	0.81
BERT-CLS	Seen	0.81	**0.67**	0.88
	Seen/Unseen	0.81	**0.68**	0.87
	Unseen	0.80	**0.67**	0.85
BERT-CRF	Seen	-	0.67	**0.90**
	Seen/Unseen	-	0.67	**0.88**
	Unseen	-	0.66	**0.87**
BERT-CLS-CRF	Seen	-	0.61	0.82
	Seen/Unseen	-	0.61	0.81
	Unseen	-	0.60	0.79

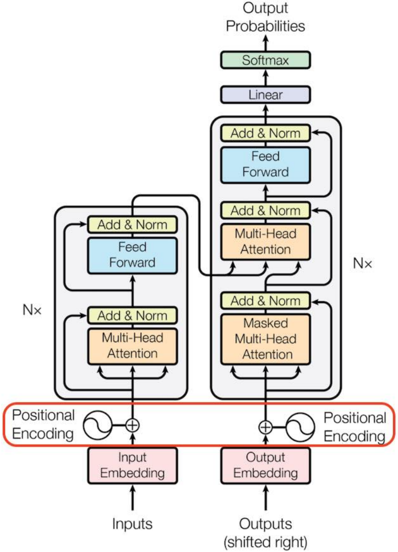


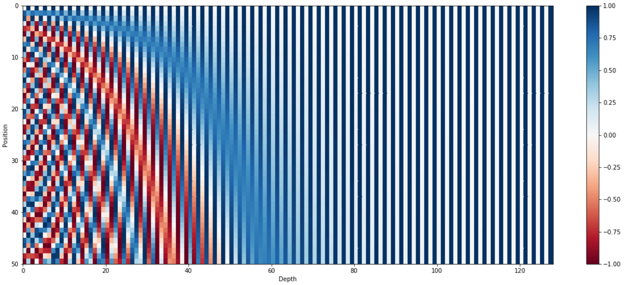
- Analyse de l'intérêt de la couche CRF
- Distinction in-domain / ood

⇒ Des perspectives vers l'encodage de la position des mots

Détection des entités dans un document structuré

Encodage de la position dans les documents:

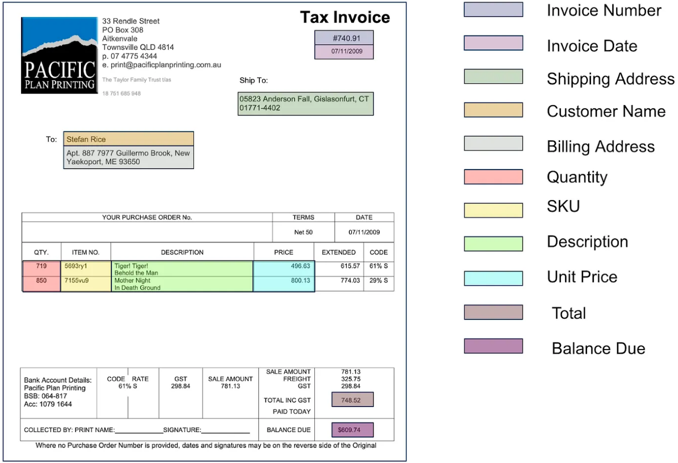




[7]

[7] A. Kazemnejad (2019). Transformer Architecture: The Positional Encoding.

Détection des entités dans un document structuré



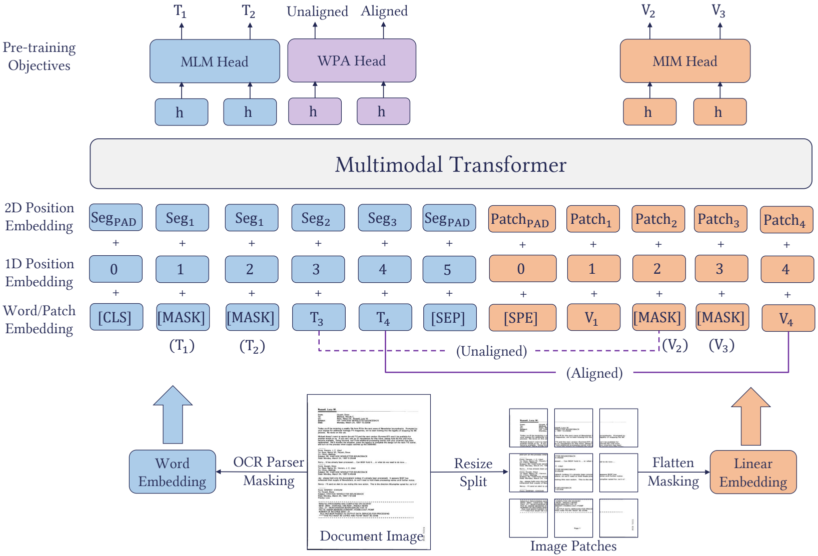
7

- Texte
- Image
- Coordonnées des mots

Puiser dans les modalités pour améliorer les performances

[7] Yiheng Xu et al. (2020). "Layoutlm: Pre-training of text and layout for document image understanding". In: ACM SIGKDD

Détection des entités dans un document structuré



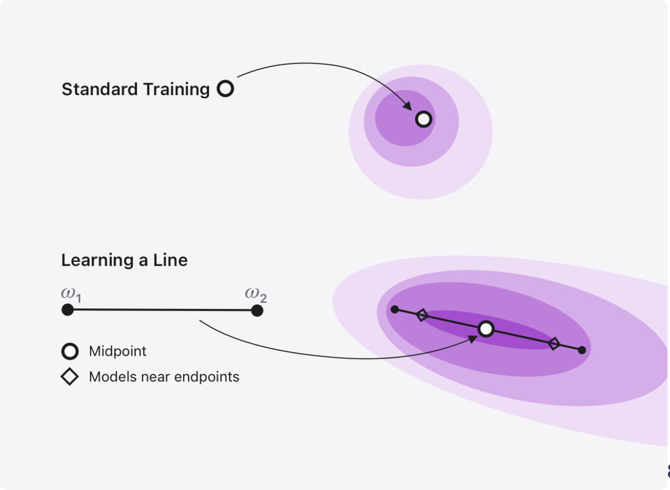
[7]

⇒ A quel moment souhaite-t-on mélanger les modalités?

[7] Yupan Huang et al. (2022). "LayoutLMv3: Pre-training for Document AI with Unified Text and Image

Optimisation robuste pour la généralisation

Optimisation de sous-espaces



[8]

Création de *régions homogènes* dans l'espace de représentation

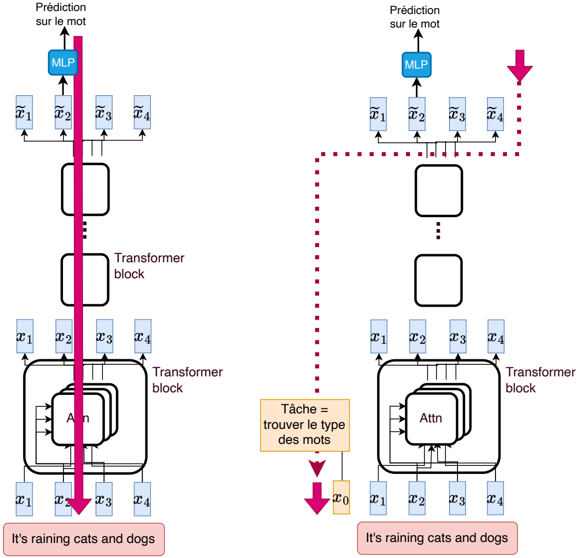
$\Rightarrow$ Améliorer l'espace de représentation

[8] Mitchell Wortsman et al. (2021). Learning Neural Network Subspaces.

Prefix-tuning & optimisation

- Impossible de maintenir plusieurs versions des paramètres d'un LLM
- Possible de travailler sur des approches parcimonieuses

$\Rightarrow$ Amélioration dans diverses tâches GLUE... Mais pas encore en NER[9]



[9] Louis Falissard, Vincent Guigue, and Laure Soulier (2023). "Improving generalization in large language models by learning prefix subspaces". In: EMNLP

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

Comment analyser la phrase suivante?

`Azawad reprend les armes`

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

En allant chercher du contexte sur internet (ou ailleurs):

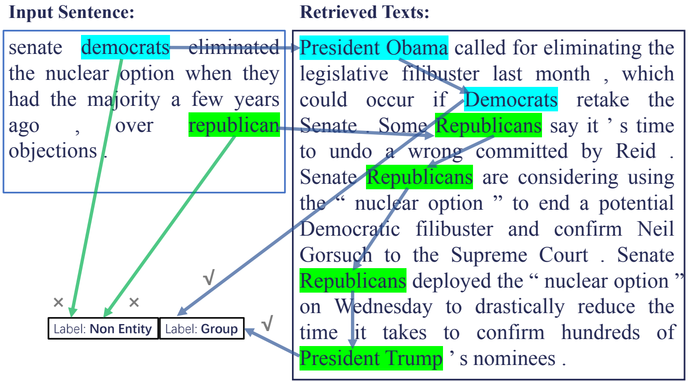
`Azawad reprend les armes`

`Le Mouvement national de l'Azawad (MNA), créé en novembre 2010`

`Le secrétaire général du mouvement est Ahmed Ould Sidi Mohamed`

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?



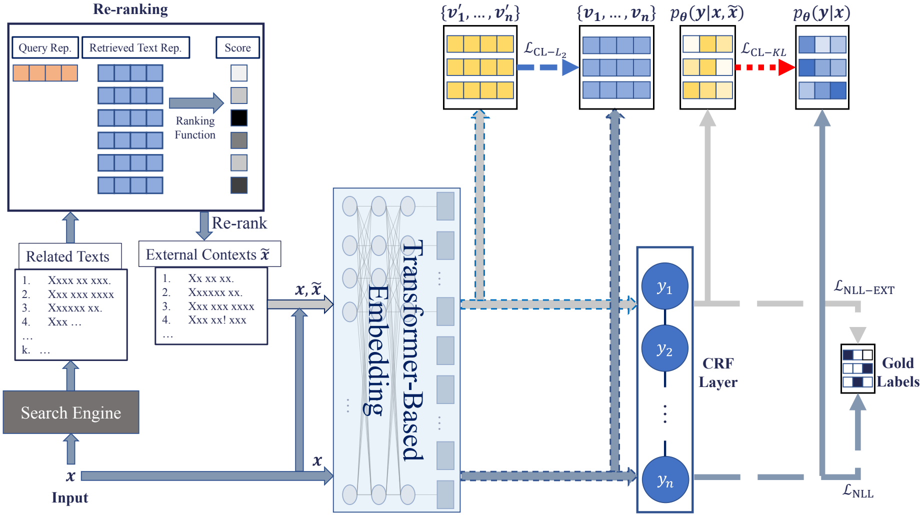
[10]

---

[10] Xinyu Wang et al. (2021). "Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning". In: ACL

Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?



Contextualisation des phrases à analyser

Erreurs en NER = problème de contextualisation?

Amélioration des performances: significatives... Mais décevantes

	Social Media		News		Biomedical		E-commerce
	WNUT-16	WNUT-17	CoNLL-03	CoNLL++	BC5CDR	NCBI	
Evaluation: **W/ CONTEXT**							
W/ CONTEXT	$57.43^\dagger$	$60.20^\dagger$	$93.27^\dagger$	$94.56^\dagger$	$90.76^\dagger$	$89.01^\dagger$	$83.15^\dagger$
CL-L_2	$58.61^\dagger$	$60.26^\dagger$	$93.47^\dagger$	$94.62^\dagger$	$\mathbf{90.99}^\dagger$	$89.22^\dagger$	$83.87^\dagger$
CL-KL	$\mathbf{58.98}^\dagger$	$\mathbf{60.45}^\dagger$	$\mathbf{93.56}^\dagger$	$\mathbf{94.81}^\dagger$	$90.93^\dagger$	$88.96^\dagger$	$\mathbf{83.99}^\dagger$

[10]

[10] Xinyu Wang et al. (2021). “Improving Named Entity Recognition by External Context Retrieving and Cooperative Learning”. In: ACL

Contextualisation et modèles de langue

- Modèle de langue = sélection des documents du contexte (BERT-Score)
- Contextualisation directe possible avec un modèle de langue
- Ouverture: reformulation de phrase
- ... Voire recherche directe des entités

Expériences préliminaires: recherche de prompts

Example of prompt	Persona	Reflection pattern	Answer format
Could you provide more information about the entities in the provided text.			
Act as an expert linguist. Could you provide more information about the entities in the provided text. **Provide outputs that an expert linguist would create.**	✓		
Could you provide more information about the entities in the provided text. Moreover, **Please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response.**		✓	
Could you provide more information about the entities in the provided text. **You should enumerate your answers as a list of propositions prefixed by a number.**			✓
You act as an expert linguist, Could you provide more information about the entities in the provided text. **Provide outputs that an expert linguist would create.** Moreover, **Please address any potential ambiguities or limitations in your answer in order to provide a more complete and accurate response. Provide outputs that an expert linguist would create.**	✓	✓	✓

⇒ Dépassement des résultats de CL-NER[11]

[11] Herserant et al. 2024. En soumission

Contextualisation et modèles de langue

- Modèle de langue = sélection des documents du contexte (BERT-Score)
- Contextualisation directe possible avec un modèle de langue
- Ouverture: reformulation de phrase
- ... Voire recherche directe des entités

Expériences préliminaires: premiers problèmes

Task	Variation	*Empty*	*Denied*	*Fail*	*Correct*
Reformulation	Classic	214 (6.31%)	374 (11.02%)	441 (12.99%)	2365 (69.68%)
	Persona	215 (6.33%)	257 (7.57%)	262 (7.72%)	2660 (78.37%)
	Reflexion pattern	209 (6.16%)	433 (12.76%)	216 (6.36%)	2536 (74.72%)
	Answer format	-	-	-	-
	All	118 (3.48%)	310 (9.13%)	103 (3.03%)	2863 (84.35%)
Named Entity Recognition	Classic	214 (6.31%)	313 (9.22%)	484 (14.26%)	2383 (70.21%)
	Persona	225 (6.63%)	222 (6.54%)	320 (9.43%)	2627 (77.40%)
	Reflexion pattern	221 (6.51%)	328 (9.66%)	273 (8.04%)	2572 (75.78%)
	Answer format	-	-	-	-
	All	134 (3.95%	258 (7.60%)	109 (3.21%)	2893 (85.24%)
Context Variation	Classic	237 (6.98%)	347 (10.22%)	415 (12.23%)	2395 (70.57%)
	Persona	221 (6.51%)	285 (8.40%)	256 (7.54%)	2632 (77.55%)
	Reflexion pattern	209 (6.16%)	338 (9.96%)	215 (6.33%)	2632 (77.55%)
	Answer format	-	-	-	-
	All	136 (4.01%)	292 (8.60%)	91 (2.68%)	2875 (84.71%)

Conclusion

- Auto-supervision
- Multi-modalité
- Dynamicité + encodage de la position
- Technique d'optimisation
- Contextualisation

- Gangner en performances en NER est difficile

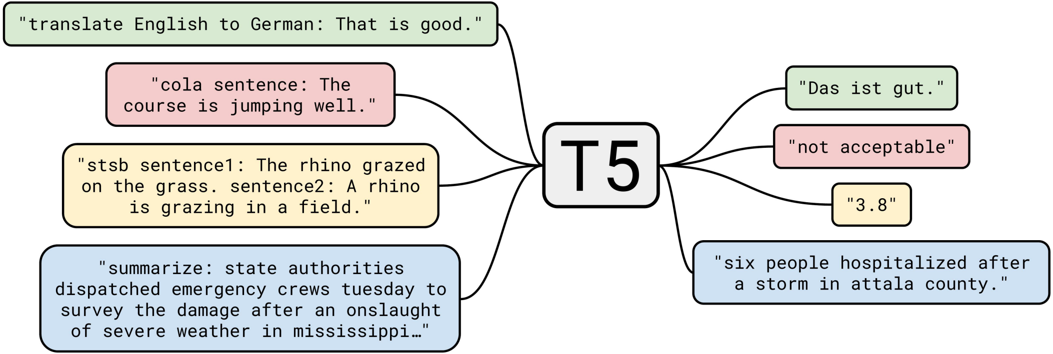
  Et publier en NER est encore plus difficile!
- 100% de performance n'est pas un objectif réaliste

Extraction d'Information & IA générative

Vers un détécteur d'entité génératif

S'il s'agit d'une traduction humain/machine... Autant partir d'un traducteur!

Une idée pas si récente:

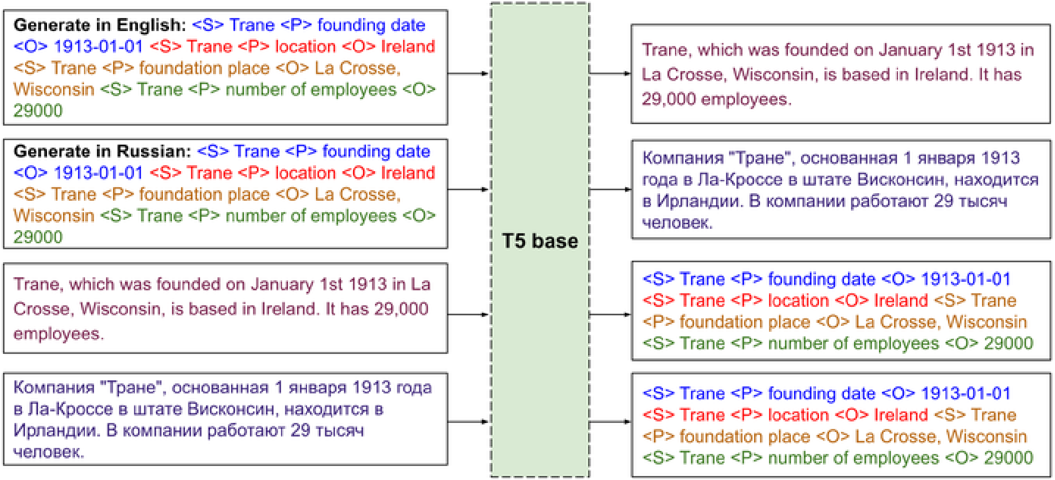


[11]

[11] Colin Raffel et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: The Journal of Machine Learning Research 1

Vers un détécteur d'entité génératif

S'il s'agit d'une traduction humain/machine... Autant partir d'un traducteur!



Extraction d'information & question answering

Résoudre l'extraction en posant des questions:

Steve	Jobs	founded	Apple	in	San	Francisco	.

Multi Turn Question Templates :

Which company is mentioned in the text ?

Who founded ____________ ?

[11]

[11] Xiaoya Li et al. (2019). "Entity-Relation Extraction as Multi-Turn Question Answering". In: ACL

Extraction d'information & question answering

Résoudre l'extraction en posant des questions:

Steve	Jobs	founded	Apple	in	San	Francisco	.

Multi Turn Question Templates :

Which company is mentioned in the text ? **Apple**

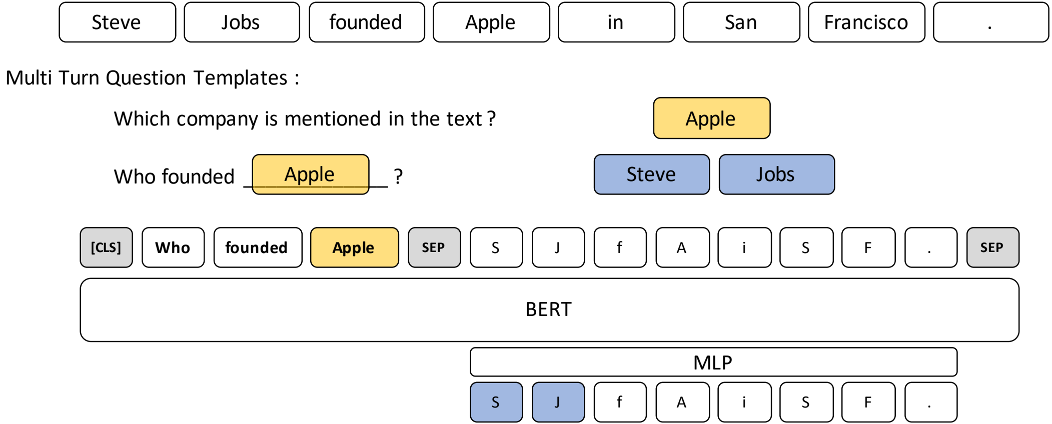
Who founded ___Apple___ ?

11

[11] Xiaoya Li et al. (2019). "Entity-Relation Extraction as Multi-Turn Question Answering". In: ACL

Extraction d'information & question answering

Résoudre l'extraction en posant des questions:



[11]

[11] Xiaoya Li et al. (2019). "Entity-Relation Extraction as Multi-Turn Question Answering". In: ACL

Extraction d'information & reformulation

En posant les bonnes questions:

V

Soit la phrase complexe suivante:
Le chat de la maison se balade dans le jardin, il est tigré et joue souvent avec celui des voisins.
Peux-tu décomposer la phrase en propositions élémentaires?

Bien sûr, voici la décomposition de la phrase complexe en propositions élémentaires :

1. Le chat de la maison se balade dans le jardin.
2. Il est tigré.
3. Il joue souvent avec celui des voisins.

Chacune de ces propositions exprime une idée distincte de la phrase complexe d'origine.

Le problème devient trivial

$\Rightarrow$ Ouverture vers un sentence-BERT-Score

Extraction d'information & reformulation

Ou en posant la question directement:

V

Soit la phrase complexe suivante:
Le chat de la maison se balade dans le jardin, il est tigré et joue souvent avec celui des voisins.
Peux-tu extraire la liste des entités et leurs relations?

Bien sûr ! Voici la liste des entités et de leurs relations dans la phrase complexe que vous avez fournie :

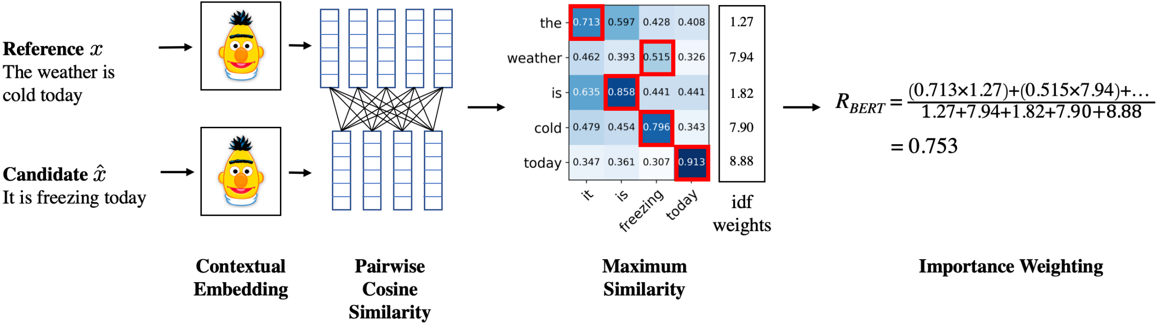
1. Entités :
   - Le chat de la maison
   - Le jardin
   - Le chat des voisins
2. Relations :
   - Le chat de la maison se balade dans le jardin.
   - Le chat de la maison est tigré.
   - Le chat de la maison joue souvent avec le chat des voisins.

Métriques associées aux IA Génératives

L'IA Générative = beaucoup de **problèmes pour évaluer** les sorties du système!

Approche BLEU / ROUGE ↘↘ ! Besoin de métriques sur les contenus

⇒ BERT-Score



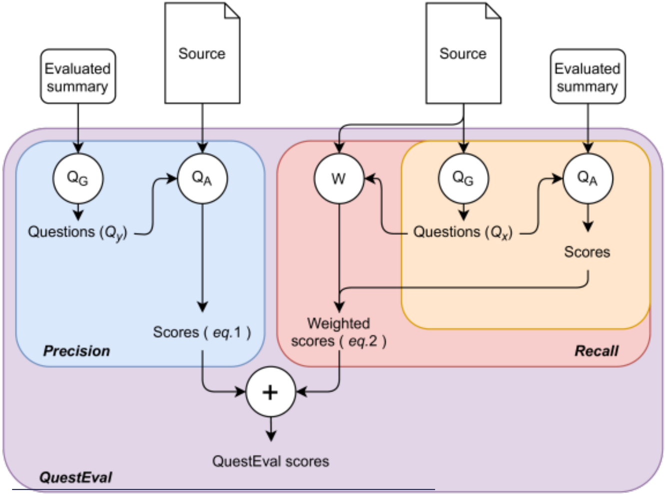
12

[12] Tianyi Zhang et al. (2019). "BERTScore: Evaluating Text Generation with BERT". In: ICLR

Métriques associées aux IA Génératives

L'IA Générative = beaucoup de **problèmes pour évaluer** les sorties du système!

Approche BLEU / ROUGE ↘↘ ! Besoin de métriques sur les contenus



[12]

Générer des questions puis répondre sur la source et la cible

IAGen pour évaluer IAGen

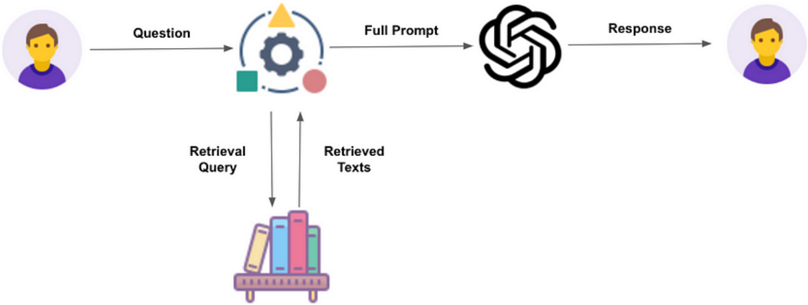
[12] Thomas Scialom et al. (2021). "QuestEval: Summarization Asks for Fact-based Evaluation". In: EMNLP.

A-t-on encore besoin d'extraction d'information?

Enjeux autour du RAG –*Retrieval Augmented Generation*–

Comment répondre à une question complexe?

1. Décomposer la question (avec un LLM)
2. Rechercher les documents pertinents / moteur de recherche
3. Rechercher les paragraphes/phrases pertinents (avec un LLM / BERT-score)
4. Construire une réponse par raisonnement automatique (LLM)
5. *et/ou* Dialoguer avec les documents Q/A (LLM)



Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?

What is 3*4+9*9? Give the answer first and then the explanation.

The answer is 99.

To solve this expression, we follow the order of operations or PEMDAS rule, which stands for Parentheses, Exponents, Multiplication and Division, and Addition and Subtraction.

There are no parentheses or exponents, so we start with multiplication and division.

Multiplying 3 by 4 gives 12, and multiplying 9 by 9 gives 81.

Then, we add these two products together, giving us:

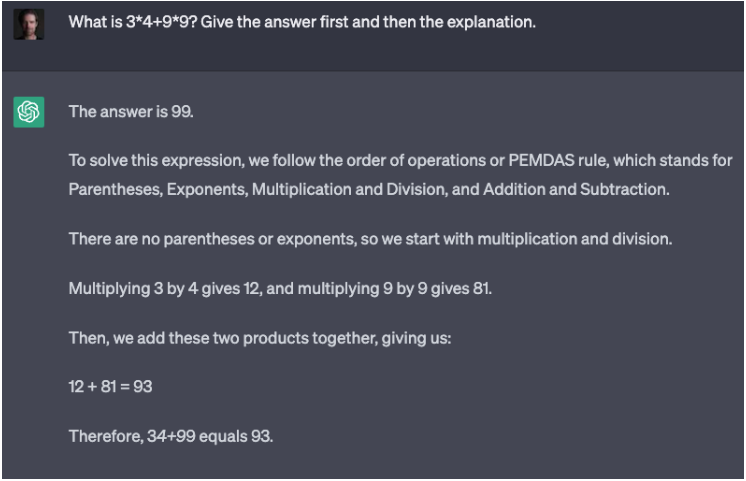
12 + 81 = 93

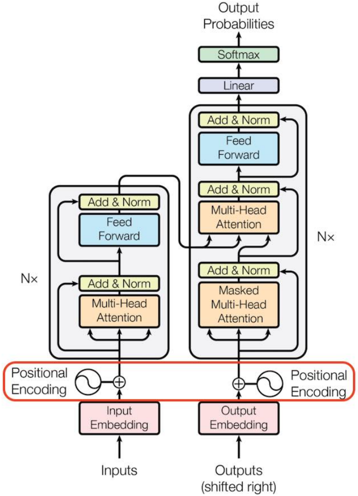
Therefore, 34+99 equals 93.

Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?





Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker

[13]

[13] Dheeru Dua et al. (2019). “DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs”. In: Proceedings of NAACL-HLT

Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?

Task: Basic Math
Problem: Before December, customers buy 1346 ear muffs from the mall. During December, they buy 6444, and there are none. In all, how many ear muffs do the customers buy?

Predicted Answer: 1346.0 ✗
Generated Program:

```
answer = 1346.0 + 6444.0
print(answer)
# Result ==> 7790.0
```

Gold Answer: 7790.0 ✓

Task: Muldiv
Problem: Tickets to the school play cost 6 for students and 8 for adults. If 20 students and 12 adults bought tickets, how many dollars' worth of tickets were sold?

Predicted Answer: 48 ✗
Generated Program:

```
a=20*6
b=12*8
c=a+b
answer=c
print(answer)
# Result ==> 216.0
```

Gold Answer: 216 ✓ [13]

[13] Swaroop Mishra et al. (2022). "LILA: A Unified Benchmark for Mathematical Reasoning". In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?

Exemple d'une question numérique sur un tableau

Nation	Gold Medal	Silver Medal	Bronze Medal
Canada	**3**	**1**	**2**
Mexico	2	3	4
Columbia	1	3	1

Question : How many medals did Canada get in the tournament?

Réponse : 3 + 1 + 2 = 6

1. Comprendre les scénarios complexes décrits dans les données, (contexte du nombre de médailles)
2. Trouver l'enchaînement des opérations nécessaires, (additions)
3. Identifier les variables mathématiques, (gold medal = 3, silver medal = 1, bronze medal = 2)
4. Effectuer les calculs (gold medal + silver medal + bronze medal)

   = 3 + 1 + 2

   = 6

[13]

[13] Fengbin Zhu et al. (2021). "TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance". In: ACL

Raisonnement automatique

Limite des modèles de langue: inaptitude au calcul !

Comment aborder le raisonnement automatique?

- Des jeux de données: *Drop*, *Lila*, *TatQA*...
- Des propositions générales:
  - Approches spécifiques
  - Chain of Thoughs, structuration des réponses
  - Capacité à coder
  - Approches mixtes internes/externes ≈ toolsformer[13]

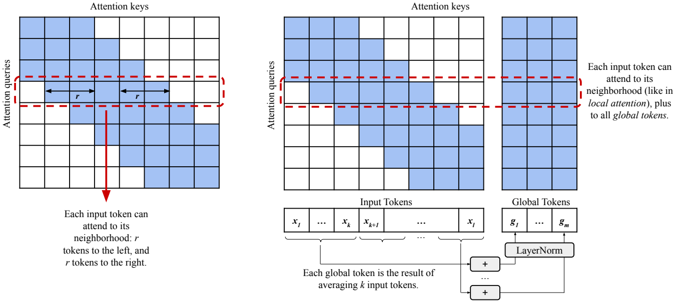
Comment évaluer? Peut-on évaluer les étapes intermédiaires du raisonnement ou seulement le résultat final? [14]

[13] Timo Schick et al. (2023). "Toolformer: Language models can teach themselves to use tools". In: arXiv preprint arXiv:2302.04761

[14] Sarah Abchiche et al. (2023). "Intégration du raisonnement numérique dans les modèles de langue: État de l'art et direction de recherche". In: CORIA. ATALA

Modèle de langue & limite des prompts

- 500 tokens avec BERT
- 16k avec LongFormer[15]
- 2000 avec chatGPT
- 32k avec GPT4
- >50k avec LongT5



a) LongT5 Local Attention b) LongT5 Transient Global (TGlobal) Attention [16]

- Hierarchiser l'attention ⇒ factoriser les calculs
- Mélanger attention locale (rapide) et attention globale clusterisée (+chère)

⇒ Vers des systèmes >300k tokens pour gérer des livres

[15] Iz Beltagy, Matthew E. Peters, and Arman Cohan (2020). "Longformer: The Long-Document Transformer". In: arXiv

[16] Mandy Guo et al. (2022). "LongT5: Efficient Text-To-Text Transformer for Long Sequences". In: NAACL

Conclusion

- Beaucoup de questions ouvertes autour des LLM
- Knowledge bases $\neq$ Extraction d'information...

$\Rightarrow$ beaucoup de questions sur l'extraction d'information

- Mutation des outils, des formulations, des performances, du mode d'interaction
- Limites sur les modèles de langue
  - Exploiter des sorties au format textuel[17]
  - Taille des entrées[18]
  - Risque d'hallucination[19]
  - Taille des modèles de langue: quelle tendance pour le futur?

[17] Herserant et al. 2024. En soumission

[18] Florian Le Bronnec et al. (2024). "LOCOST: Long Contexts with State Space Encoders for Conditional Text Generation". In: En soumission

[19] Pierre Erbacher et al. (2024). "Navigating Uncertainty: Optimizing API dependency for Hallucination Reduction in Closed-Book QA". In: En soumission