SCIENCES SORBONNE UNIVERSITÉ

CONCEPTUAL GROUNDING FOR TEXT REPRESENTATION LEARNING

Journée IA - RI

Monday 2nd December, 2019

Laure Soulier

LIP6

MLIA Machine Learning & Deep Learning for Information Access

1 Text grounding

2 Enhancing text representation with knowledge resources

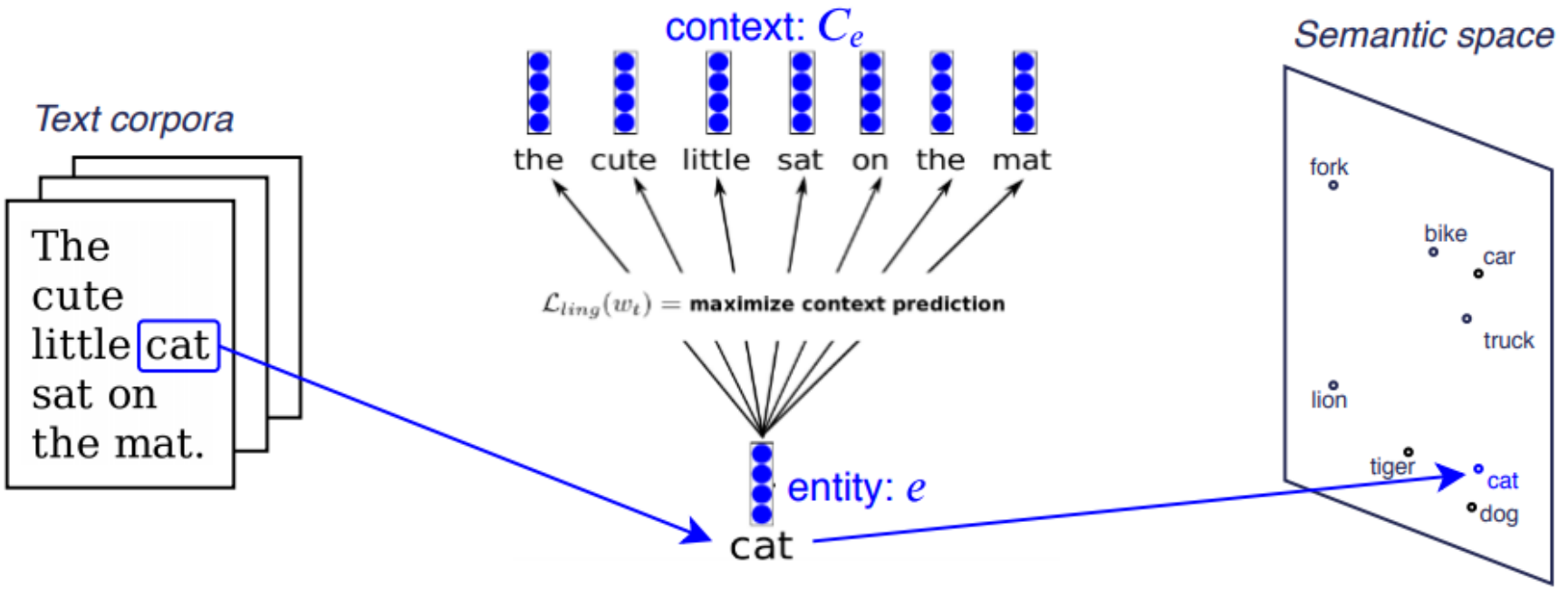
3 Learning Multi-Modal Word Representation Grounded in Visual Context

4 Conclusion and Perspectives

Text grounding

Representation learning

Goal: Represent the *semantic* of a word in a vector space



Distributional Hypothesis:
"Words that occur in similar contexts should have similar meanings"

(Harris 1954, Firth 1957)

Models:

→ `Word2Vec` (Mikolov et al. 2013)

→ `Glove` (Pennington et al. 2014)

→ `ELMo` (Peters et al. 2018)

Property:
Semantic similarity is quantified by spatial proximity

Representation learning

SCIENCES SORBONNE UNIVERSITÉ

Encoding word semantics —Applications

- → Information retrieval
- → Language understanding (QA, summarization, NER, POS Tagging, sentiment analysis)
- → Machine translation
- → Statistical language modeling (speech recognition, dialog systems)
- → Zero-Shot Learning
- → ...

Why learning text embeddings is a difficult task?

SCIENCES SORBONNE UNIVERSITÉ

→ Human reporting bias

→ We are more likely to report **unusual facts** and **facts with values**

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Figure 1: N-gram frequencies for various verbal events and the number of times Knext learns that "A person may <x> ..."

Human Reporting Bias [Gordon, Van Durme, 2013]

The frequency at which objects, relations, or events occur in natural language are significantly different from their real-world frequency.

Why learning text embeddings is a difficult task?

- Human reporting bias
  - We are more likely to report **unusual facts** and **facts with values**...
  - ... while unlikely to mention **something expected** or **trivial facts**

Example

- When we say:

"Hand me the salt"

Why learning text embeddings is a difficult task?

SCIENCES SORBONNE UNIVERSITÉ

- Human reporting bias
  - We are more likely to report **unusual facts** and **facts with values**...
  - ... while unlikely to mention **something expected** or **trivial facts**

Example

- When we say:

  "Hand me the salt"

- We mean:

  "Hand me the salt ... or rather the receptacle that contains it. It has a cylindrical shape and is about 3 inches high. This object does not float in the air and lies on the table, in other words, in direct contact with it. If I ask you this favor, it means that it is closer to you than it is to me. Besides, when you give it to me, the aperture should be on the top so that the salt is not spoiled on the table because of gravity, which makes that dropped objects fall down. When you hand it to me, I expect that you give it to by making contact with my hand, located at the end of my arm, that I am going to bring closer to you. Salt is an ionic compound that can be formed by the neutralization reaction of an acid and a base. It is usually extracted from

Why learning text embeddings is a difficult task?

SCIENCES SORBONNE UNIVERSITÉ

- → Word co-occurences do not capture all grammatical peculiarities
  - → Confusion between the notions of semantic similarity and conceptual association [**Hill2015a**]

Example

- → [*car*, *bike*]: similar because common physical features, common function, or same category
- → [*car*, *petrol*]: functional relationship, associated because they frequently occur together in space and language.

Why learning text embeddings is a difficult task?

SCIENCES SORBONNE UNIVERSITÉ

- → Word co-occurences do not capture all grammatical peculiarities
  - → Confusion between the notions of semantic similarity and conceptual association [**Hill2015a**]
  - → Embeddings fail to detect synonyms/antonyms [**mrksic2016**, **Mohammad:2008**]

east	expensive	British
west	pricey	American
north	cheaper	Australian
south	costly	Britain
southeast	overpriced	European
northeast	inexpensive	England

Critical in certain domain applications

E.g., dialog tracking for restaurant booking (expensive vs. cheap)

Why learning text embeddings is a difficult task?

SCIENCES SORBONNE UNIVERSITÉ

- → Word co-occurrences do not capture all grammatical peculiarities
  - → Confusion between the notions of semantic similarity and conceptual association [**Hill2015a**]
  - → Embeddings fail to detect synonyms/antonyms [**mrksic2016**, **Mohammad:2008**]
  - → Embeddings conflate the contextual evidence of different meanings of a word into a single vector [**IacobacciPN15**]

bank$^n_1$ (geographical)	***bank***$^n_2$ (financial)	***number***$^n_4$ (phone)	***number***$^n_3$ (acting)	***hood***$^n_1$ (gang)	***hood***$^n_{12}$ (convertible car)
upstream$^r_1$	commercial_bank$^n_1$	calls$^n_1$	appearing$^v_6$	tortures$^n_5$	taillights$^n_1$
downstream$^r_1$	financial_institution$^n_1$	dialled$^v_1$	minor_roles$^n_1$	vengeance$^n_1$	grille$^n_2$
runs$^v_6$	national_bank$^n_1$	operator$^n_{20}$	stage_production$^n_1$	badguy$^n_1$	bumper$^n_2$
confluence$^n_1$	trust_company$^n_1$	telephone_network$^n_1$	supporting_roles$^n_1$	brutal$^a_1$	fascia$^n_2$
river$^n_1$	savings_bank$^n_1$	telephony$^n_1$	leading_roles$^n_1$	execution$^n_1$	rear_window$^n_1$
stream$^n_1$	banking$^n_1$	subscriber$^n_2$	stage_shows$^n_1$	murders$^n_1$	headlights$^n_1$

Text grounding

SCIENCES SORBONNE UNIVERSITÉ

Knowledge resources

PRINCETON UNIVERSITY
WordNet
A lexical database for English

MeSH

DBpedia

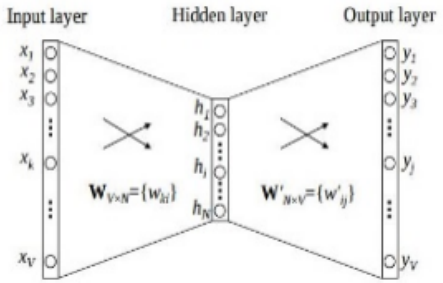
Visual databases

COCO
Common Objects in Context

VISUALGENOME

Many others...

reddit



Grounded text representations

Raw text data

The man near the **bank** is grabbing his **apple**. He would like to **send an email before eating** to alert his colleagues about the coming storm.

Contributions

General objectives

Incorporating common-sense and word knowledge in text representations

→ Exploring the potential of grounded on NLP & IR tasks, not to compete with all representation learning baselines

→ Knowledge-empowered text representations
  → Incorporating word senses through concepts in knowledge resources
  → Leveraging word association through concept relations

IRIT Collaboration with IRIT: Gia-Hung Nguyen, Lynda Tamine, Nathalie Souf - ESWC 2018 & ACM TOIS 2019

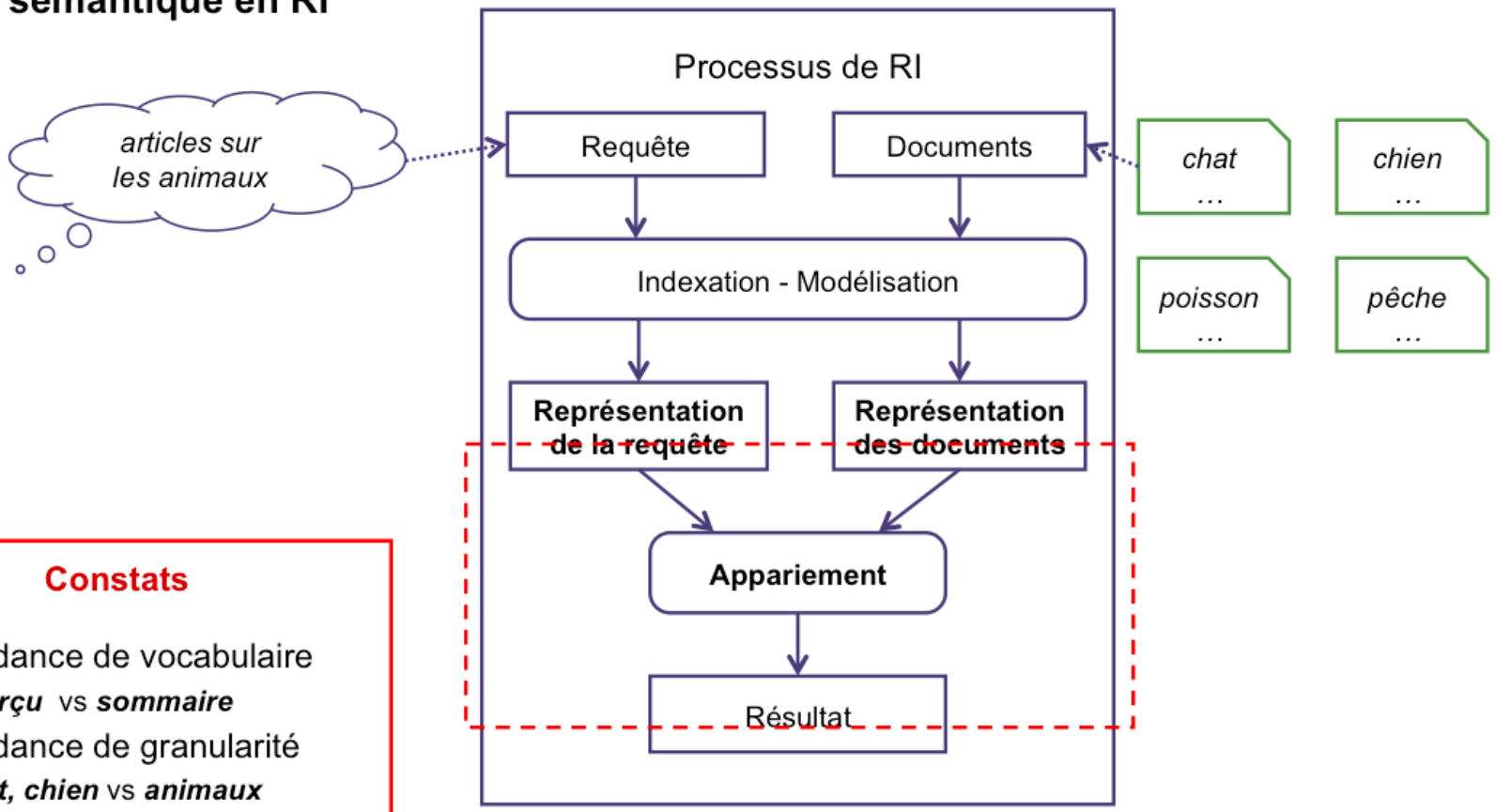
→ Visual-grounded word representations
  → Grounding words in images to improve word description
  → Leveraging the visual context to incorporate word functionalities
  → Extension to sentence representations

MLIA Work with Eloi Zablocki, Patrick Bordes, Benjamin Piwowarski, Patrick Gallinari (CHIST-ERA MUSTER project) - AAAI 2018 & EMNLP 2019

Enhancing text representation with knowledge resources

Motivations

❑ **Fossé sémantique en RI**



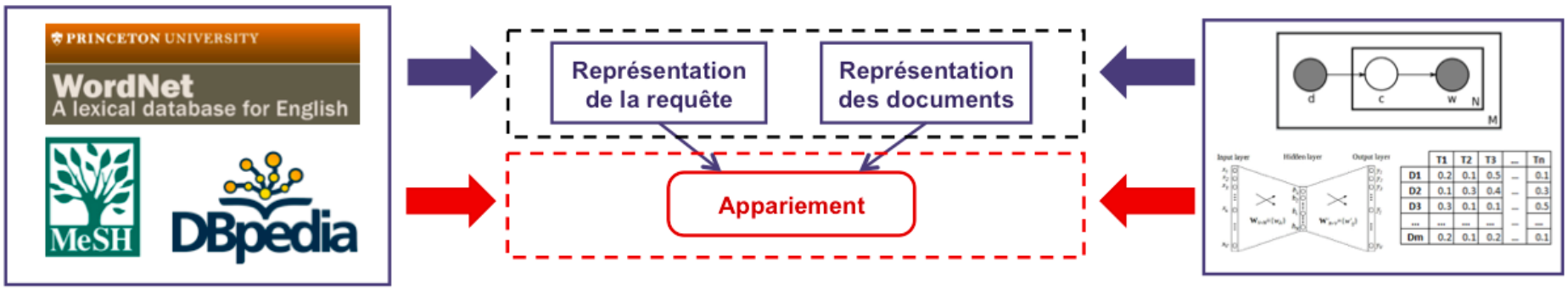
Constats

Ω Discordance de vocabulaire
aperçu vs ***sommaire***

Ω Discordance de granularité
chat, chien vs ***animaux***

Ω Polysémie
pêche (*fruit* vs *pêcher*)

Motivations



❖ **Sémantique relationnelle**

- **Niveau représentation** : expansion de représentations lexicales *(Navigli and Velardi 2003; Pal et al. 2014 ; Xiong and Callan 2015a)*
- **Niveau appariement** : inférence logique sur les graphes de concepts *(Koopman et al. 2016 ; Xiong and Callan, 2015b)*

❖ **Sémantique distributionnelle**

- **Niveau représentation** : représentation distribuée LSA *(Deerwester et al., 1990)*, PLSA *(Hofmann, 1990)*, WordEmbedding *(Mikolov et al., 2013 ; Pennington et al., 2014)*
- **Niveau appariement** : réseau de neurones siamois, convolutif *(Huang et al., 2013 ; Guo et al., 2016 ; Mitra et al., 2017)*

Offline vs. online representation learning

Offline learning

Retrofitting text embeddings [**Faruqui2015**, **mrksic2016**]

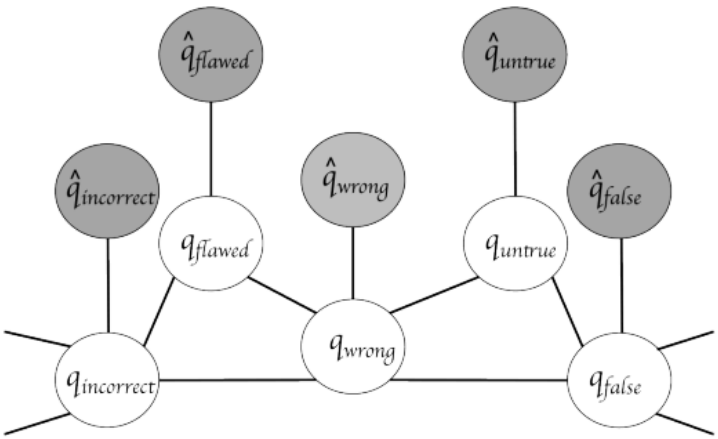
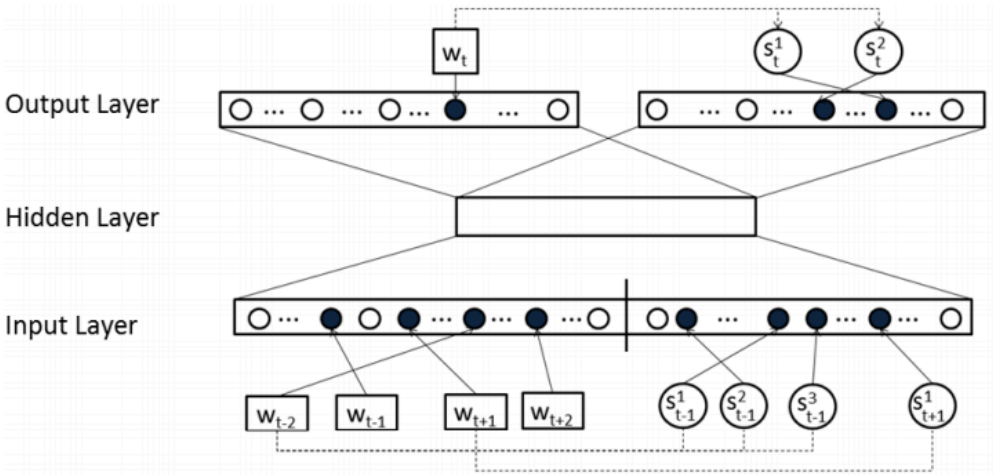


Figure 1: Word graph with edges between related words showing the observed (grey) and the inferred (white) word vector representations.

$$\mathcal{L} = \sum_{i=1}^{n} \left[\alpha_i \left\| q_i - \hat{q}_i \right\|^2 + \sum_{(i,j) \in E} \beta_{ij} \left\| q_i - q_j \right\|^2 \right]$$

Online learning

Joint learning of word, concept, text embeddings [**Liu2016b**, **DBLP:journals/corr/ManciniCIN16**]



$$\mathcal{L} = -\log\left(p\left(w_t | W^t, S^t\right)\right) - \sum_{s \in S_t} \log\left(p\left(s | W^t, S^t\right)\right)$$

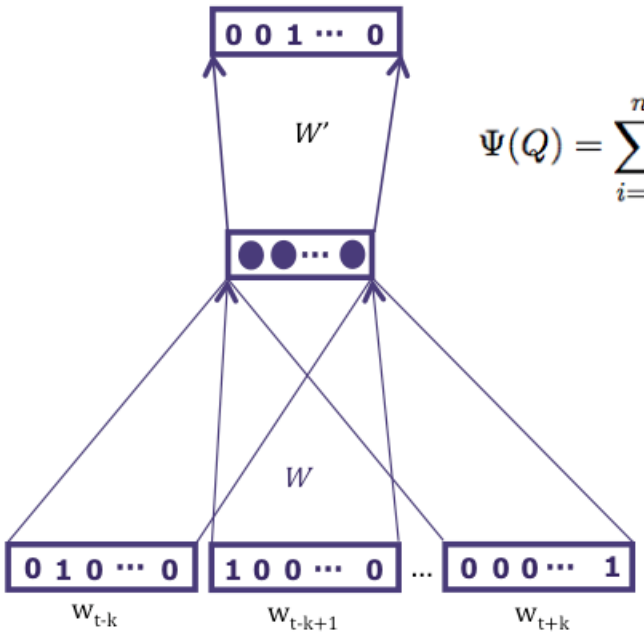
Previous work: Offline models

→ Learning word representations

→ Updating word vectors with constraints in the knowledge base

❑ ***« Retrofitting Word Vectors to Semantic Lexicons »***

[Faruqui et al., 2015]



$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j)\in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

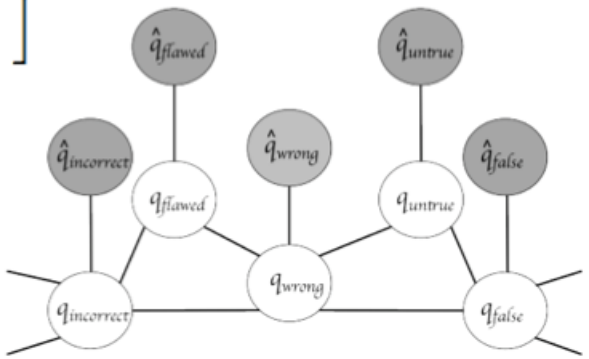


Figure 1: Word graph with edges between related words showing the observed (grey) and the inferred (white) word vector representations.

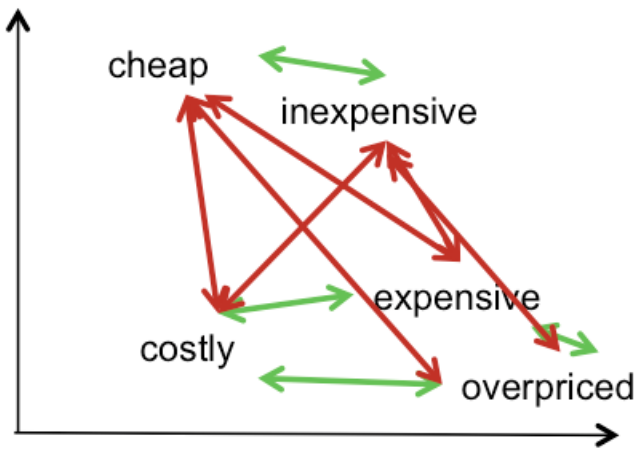
Previous work: Offline models

→ Learning word representations

→ Updating word vectors with constraints in the knowledge base

❑ ***« Counter-fitting Word Vectors to Linguistic constraints »***

[Mrksic et al., 2016]



- **Attraction des synonymes**

$$\mathrm{SA}(V') = \sum_{(u,w)\in S} \tau\left(d(\mathbf{v}'_u, \mathbf{v}'_w) - \gamma\right)$$

- **Abrogation des antonymes**

$$\mathrm{AR}(V') = \sum_{(u,w)\in A} \tau\left(\delta - d(\mathbf{v}'_u, \mathbf{v}'_w)\right)$$

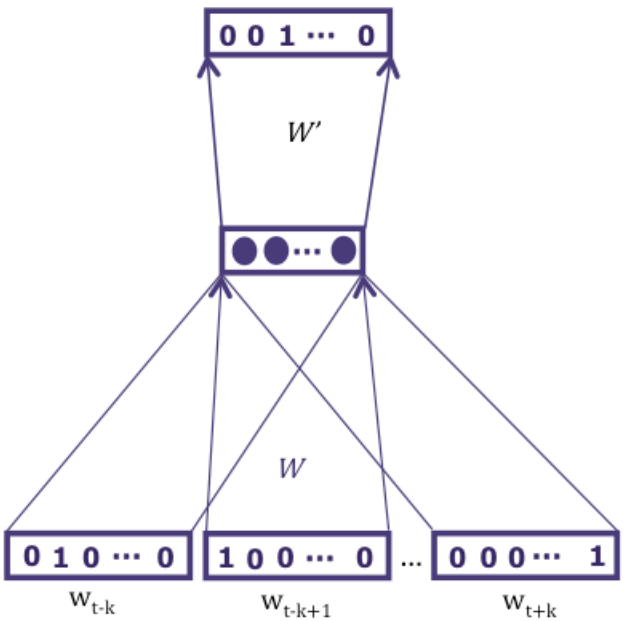
Previous work: Online models

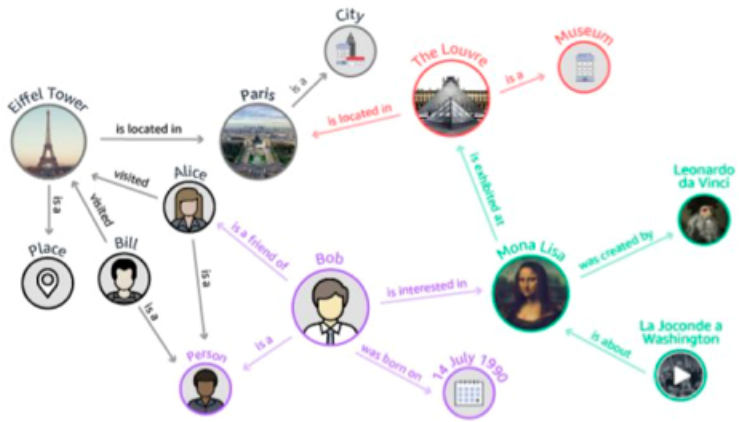
→ Joint learning of word and concept representations

→ revisited PV-DM

❑ ***« Improving Lexical Embeddings with Semantic Knowledge »***

[Yu and Dredze, 2014]



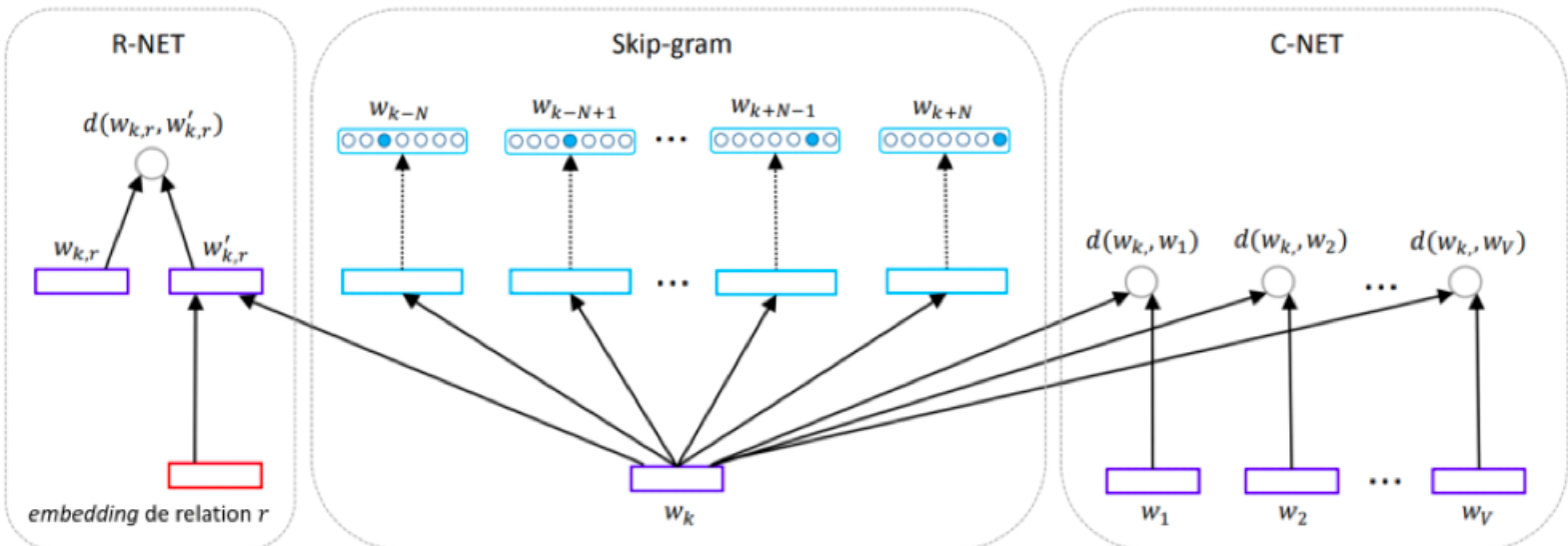


$$L(\theta) = \underbrace{\alpha L_{RCM}}_{\text{Relational loss (related words=similar rep.)}} + \underbrace{J_{CBOW}}_{\text{Textual loss}}$$

Previous work: Online models

→ Joint learning of word and concept representations
→ revisited PV-DM

❑ ***« RC-Net: a general framework for incorporating knowledge into word representations »*** *[Xu et al., 2014]*

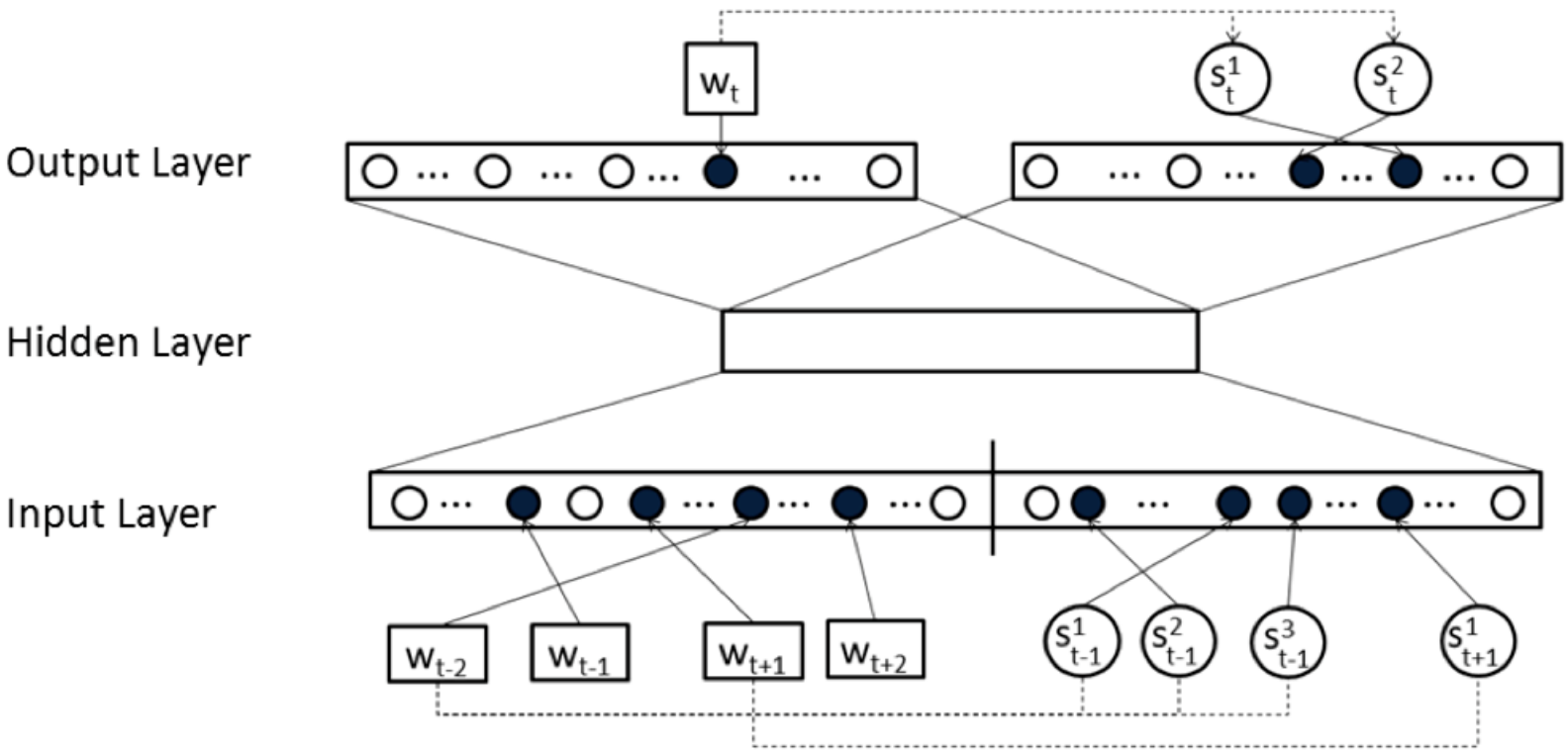


$$L(\theta) = \underbrace{\alpha J_r}_{\text{Relational loss (transE)}} + \underbrace{J_{skipgram}}_{\text{Textual loss}} + \underbrace{\beta J_c}_{\text{Categorial loss}}$$

Previous work: Online models

→ Joint learning of word and concept representations
→ revisited PV-DM

❑ ***« Embedding Words and Senses Together via Joint Knowledge-Enhanced training »*** *[Mancini et al., 2017]*



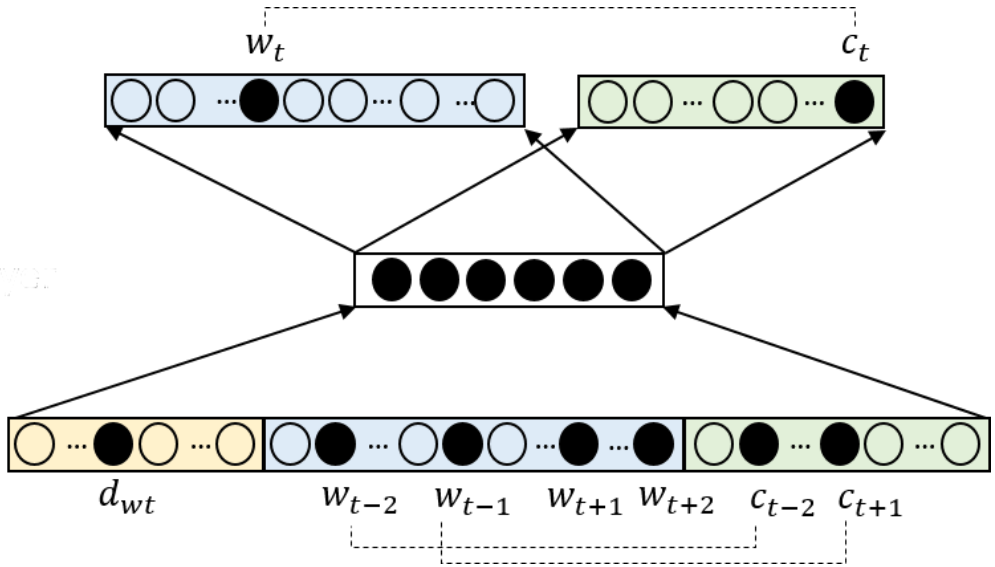
Previous work

Knowledge-enhanced representation learning				
Joint learning of embeddings (Online learning)				
Embeddings learning	**word embedding**	**concept embedding**	**document embedding**	**with relational constraints**
[Liu et al., 2016]	X			X
[Yu and Dredze, 2014]	X			X
[Jauhar et al., 2015]	X			X
[Liu et al., 2018]	X	X		X
[Mancini et al., 2016]	X	X		
[Cheng et al., 2015]	X	X		
[Yamada et al., 2016]	X	X		
Our model	**X**	**X**	**X**	**X**
Retrofitting of embeddings (Offline learning)				
[Faruqui et al., 2014]	X			X
[Glavas et Vulic, 2018]	X			X
[Mrksic et al., 2016]	X			X
[Jauhar et al., 2015]	X			X

Online representation learning with knowledge resources

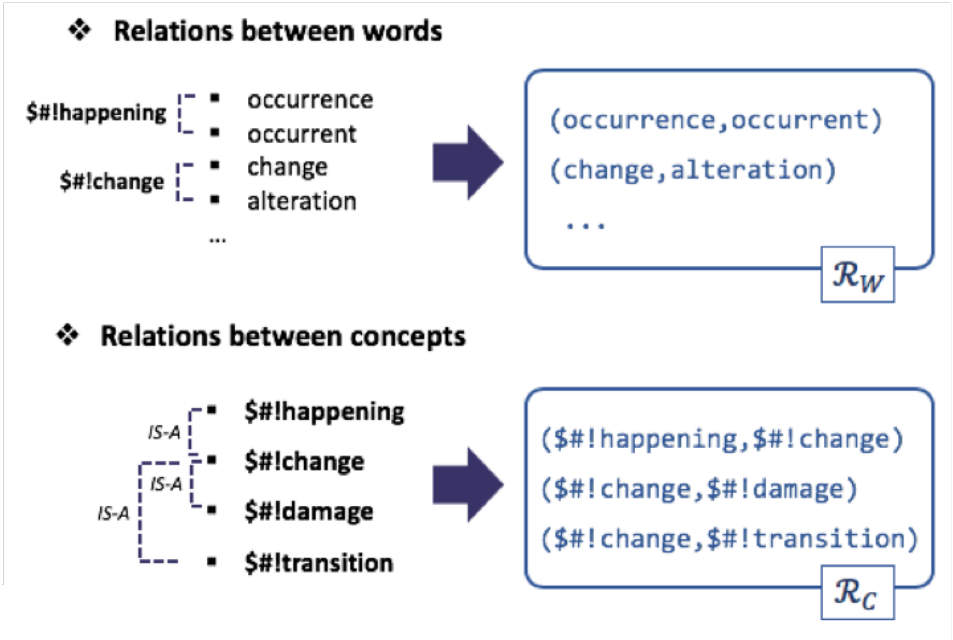
SCIENCES SORBONNE UNIVERSITÉ

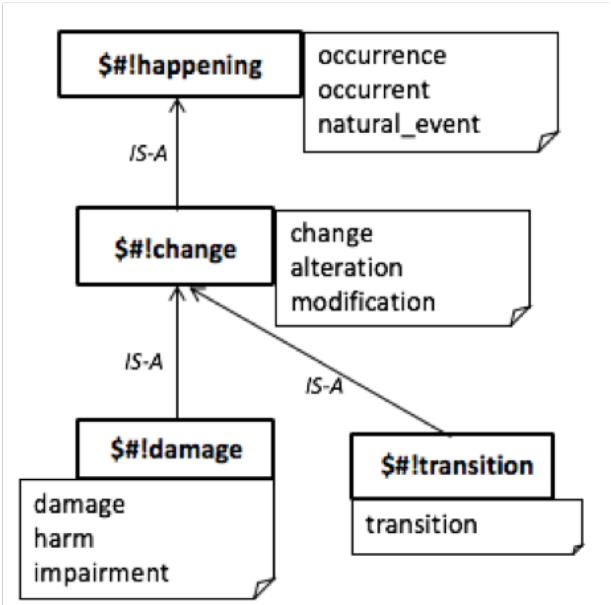
- → Learning documents embeddings using PV-DM model
- → Associating words with concepts/synsets



$$\Psi(D) = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\frac{1}{|d_{wc}|}\sum_{w_i\in d_{wc}}[\log p(w_i|w_{i\pm w}, c_{i\pm w}, d) + \log p(c_i|w_{i\pm w}, c_{i\pm w}, d) - \frac{\gamma}{|d_{wc}|}||\hat{d_{wc}}||^2] \quad (1)$$

Online representation learning with knowledge resources





$$\sum_{d\in D}\sum_{w_t\in d}\left[\log P\left(w_t|w_{t\pm k},c_{t\pm k},d\right)+\log P\left(c_t|w_{t\pm k},c_{t\pm k},d\right)-\frac{\gamma}{|d|}\|\vec{d}\|^2\right]$$
$$+\alpha_W\sum_{(w_iw_j)\in\mathcal{R}_W}\text{sim}\left(\overrightarrow{w_i},\overrightarrow{w_j}\right)$$
$$+\alpha_C\sum_{(c_k,c_l)\in\mathcal{R}_W}\text{sim}\left(\overrightarrow{c_k},\overrightarrow{c_l}\right)$$

Evaluation protocol

Data sets	ROBUST	OHSUMED	TREC Med
Type	site de nouvelles	articles médicaux	rapport de visites médicales
# documents	~528 000	~348 000	~17 000
# queries	250	63	35
Sample of query	"Best Retirement Country"	"adult respiratory distress syndrome"	"patients with gastroesophageal reflux disease"

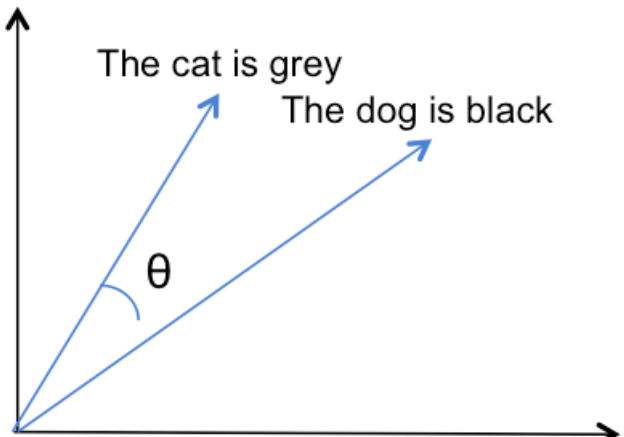
Resources	DBpedia	MeSH
# extracted concepts	250 000	18 000
used relations	"gold:hypernym"	"IS-A"

Evaluation protocol

Modèles de référence	$TF - IDF$	représentation classique en RI
	AWE	moyenne des représentations de mots du document
	AWE_R	moyenne des représentations de mots ajustées par une relation [Faruqui et al. 2015]
	$D2V$	représentations de document apprises par ParagraphVector [Le and Mikolov., 2014]

Tâches	*Tâches TALN*	Similarité des phrases/ documents	Classification des phrases
	Tâches RI	Réordonnancement des documents	Expansion de la requête

- Similarité de phrases



Corrélation

Note entre 1 et 10

Evaluation protocol

Modèles de référence	$TF - IDF$	représentation classique en RI
	AWE	moyenne des représentations de mots du document
	AWE_R	moyenne des représentations de mots ajustées par une relation [Faruqui et al. 2015]
	$D2V$	représentations de document apprises par ParagraphVector [Le and Mikolov., 2014]

Tâches	*Tâches TALN*	Similarité des phrases/ documents	Classification des phrases
	Tâches RI	Réordonnancement des documents	Expansion de la requête

- Ré-ordonnancement des documents

$$RSV(q,d) = \alpha \cdot IRScore(q,d) + \underbrace{(1-\alpha) \cdot NeuralScore_{KB}(q,d)}$$

Représentations distribuées de la requête et du document
→ Même niveau de granularité

Evaluation protocol

Modèles de référence	$TF - IDF$	représentation classique en RI
	AWE	moyenne des représentations de mots du document
	AWE_R	moyenne des représentations de mots ajustées par une relation [Faruqui et al. 2015]
	$D2V$	représentations de document apprises par ParagraphVector [Le and Mikolov., 2014]

Tâches	*Tâches TALN*	Similarité des phrases/ documents	Classification des phrases
	Tâches RI	Réordonnancement des documents	Expansion de la requête

- Expansion de la requête

m : mot ou concept candidat

$$p(m|q*) = \alpha p_{mle}(m|q) + (1-\alpha) p_{emb}(m|q)$$

$$p_{emb}(m|q) = \frac{\sigma(\vec{e}_m, \vec{q})}{\sum_{m' \in V} \sigma(\vec{e}_{m'}, \vec{q})}$$

Représentations distribuées de la requête et des éléments candidats
→ Niveaux de granularité différents

Results

Sentence relatedness / classification tasks

→ Improvements for out-of-domain datasets

Training dataset	Robust				Ohsumed				TREC Med			
Eval. benchmark	SUBJ	MPQA	TREC	MRPC	SUBJ	MPQA	TREC	MRPC	SUBJ	MPQA	TREC	MRPC
TF-IDF	72.13	68.45	79.98	69.12	**33.13**	25.35	31.48	30.32	22.55	21.99	21.48	19.75
AWE	73.10	68.04	79.52	68.05	32.50	25.74	32.12	29.35	21.92	21.87	20.51	19.25
AWE_{KB}	75.71	69.08	**81.91**	68.75	35.61	**26.63**	**34.18**	**31.20**	**22.63**	**22.45**	**21.24**	**20.60**
D2V	73.52	69.35	79.30	70.81	31.56	25.01	32.07	28.23	21.55	21.58	20.68	18.56
$SD2V_{off}$	**76.15**$^{\bullet}$	**72.11**$^{\Delta}_{\bullet}$	79.50	71.90$^{\Delta}_{\bullet}$	32.65	25.40	32.15$^{\bullet}$	27.88$^{\bullet}$	22.00	21.18$^{\bullet}$	20.92	18.21$^{\bullet}$
$SD2V_{on}$	75.44$^{\bullet}$	70.89$^{\bullet}$	79.56	**72.04**$^{\bullet}$	32.99	25.53	32.57$^{\bullet}$	28.80$^{\bullet}$	21.81$^{\bullet}$	21.38	21.00	18.15$^{\bullet}$

Results

	Document re-ranking						Query expansion					
	Robust		Ohsumed		TREC Med		Robust		Ohsumed		TREC Med	
	MAP	%Chg	MAP	%Chg	MAP	%Chg	MAP	%Chg	MAP	%Chg	MAP	%Chg
BM25	**0.251**		0.2147		0.312		**0.251**		0.2147		0.312	
AWE	0.250	-0.40%	0.201	-2.24%	0.349$^{\bullet}$	+11.83%	0.250	-0.40 %	0.252$^{\bullet}$	+17.51%	0.289$^{\bullet}$	-7.08%
AWE_{KB}	0.251	+0.00%	0.301$^{\bullet}$	+40.20%	0.350$^{\bullet}$	+12.24%	0.251	0.00%	0.254$^{\bullet}$	+18.30%	0.2901$^{\bullet}$	-7.02%
D2V	0.2505	-0.20%	0.300$^{\bullet}$	+39.78%	0.356$^{\bullet}$	+14.07%	0.2511	+0.04%	0.255$^{\bullet}$	+19.19%	0.291$^{\bullet}$	-6.67%
$SD2V_{off}$	**0.251**	0.00%	0.3018$^{\bullet}$	+40.57%	**0.3591**$^{\Delta}$	+15.10%	0.2464	-1.83%	0.258$^{\bullet}$	+20.17%	**0.3205**$^{\Delta}$	+2.72%
$SD2V_{on}$	0.2507	-0.12%	**0.302**$^{\bullet}$	+40.66%	0.3554$^{\bullet}$	+13.91%	0.2443$^{\bullet}$	-2.67%	**0.2599**$^{\Delta}_{\bullet}$	+21.05%	0.2889$^{\bullet}$	-7.40%

Table 7. Comparison of online/online learning approaches on IR evaluation tasks: re-ranking and query expansion. Metric: MAP. *%Chg* denotes the effectiveness improvement of models w.r.t. BM25. Δ: significance test of offline vs. online models. $\bullet$: significance test of BM25 vs. our model.

IR tasks

- → Improvements vs. BM25 are important in the medical field (between 15% and 40%)
- → Higher growth rates on re-ranking task (40.6% vs. 21.16% for query expansion)
- → Better performance compared to neural models (+3%)

Results: Validating relational constraints

Relational constraints

→ C1: word-word relations.

→ C2: concept-concept relations.

Training dataset	Robust	Ohsumed	TREC Med
$SD2V_{on}$	0.14	0.02	0.02
$SD2VReg_{on}$	**0.18**	**0.14**	0.14

Table 1: Validation of the relational constraints C1 and C2 on pivotal words. Metric: P@10.

Results

Training dataset	Robust				Ohsumed				TREC Med			
Eval. benchmark	SUBJ	MPQA	TREC	MRPC	SUBJ	MPQA	TREC	MRPC	SUBJ	MPQA	TREC	MRPC
$SD2V_{off}$	76.15	72.11	79.50	71.90	32.65	25.40	32.15	27.88	22.00	21.18	20.92	18.21
$SD2VReg_{off}$	**77.02**	**72.78**	**80.89**$^{\bullet}$	73.37$^{\bullet}$	33.83	26.34	33.12	28.77	**23.02**	21.99	**22.48**$^{\bullet}$	18.93
$SD2VIns_{off}$	76.41	72.00	80.20	**73.77**$^{\bullet}$	**34.37**$^{\bullet}$	**27.42**$^{\bullet}$	**34.18**$^{\bullet}$	29.31$^{\bullet}$	22.94	20.94	21.32	19.44$^{\bullet}$
$SD2V_{on}$	75.44	70.89	79.56	72.04	32.99	25.53	32.57	28.80	21.81	21.38	21.00	18.15
$SD2VReg_{on}$	75.69	71.09	80.01	73.62$^{\bullet}$	34.33$^{\bullet}$	27.28$^{\bullet}$	33.21	**29.44**	22.78	**22.74**$^{\bullet}$	22.03	**19.61**$^{\bullet}$
$SD2VIns_{on}$	76.78$^{\bullet}$	71.13	79.43	72.27	33.31	26.08	32.76	28.96	22.40	22.05	20.34	19.28$^{\bullet}$

Table 14. Comparison of the relation constraint strategies on the on classification tasks (SentEval benchmark). Metric: Accuracy (%). *%Chg* denotes the effectiveness improvement of models w.r.t. BM25. Δ: significance test of regularization vs. training instance. $\bullet$: significance test w.r.t model scenario without relation.

→ Constraining the learning with relational knowledge is effective in both NLP and IR tasks.

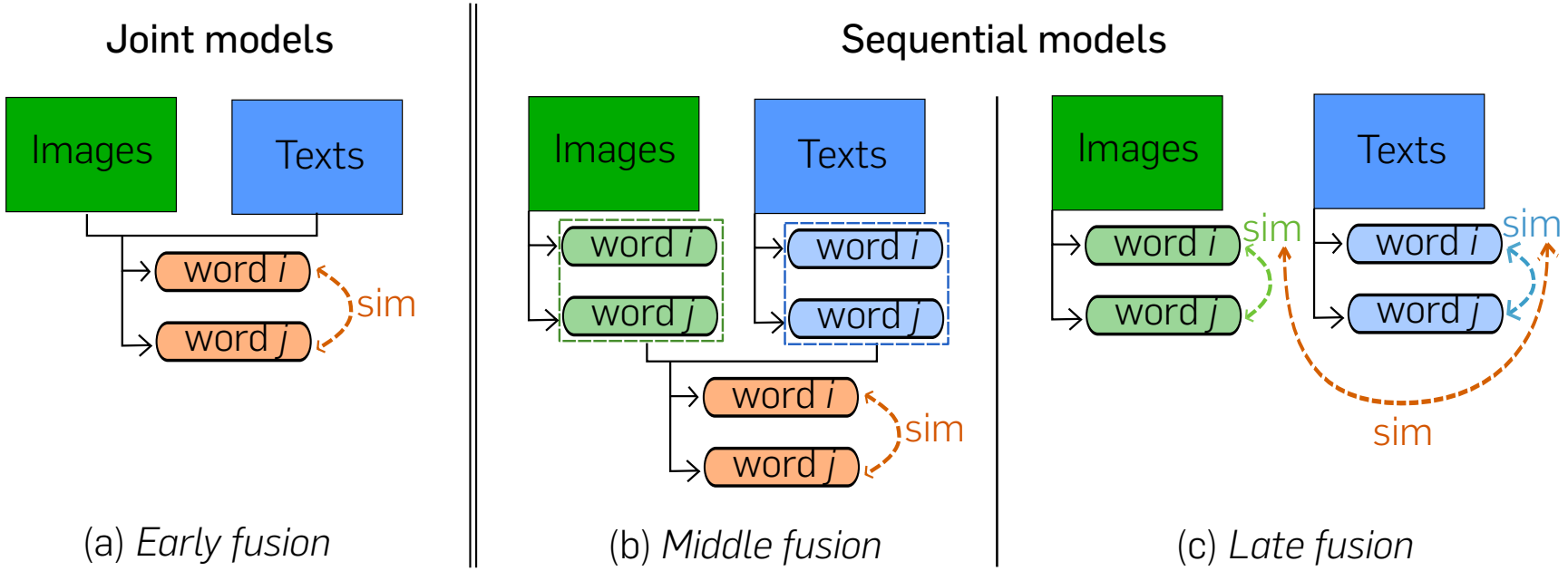
→ The learning leverages from both word-word relations and concept-concept relations

Limitations and Perspectives

- (Main) Pending issues and Perspectives
  - **Robustness of the models:** significant performance variation depending on multiple factors (knowledge resource, task, annotation quality, etc.)
  - **Exploiting contextual representation learning models (Elmo, BERT, ...)**
  - **Considering the relation types** in the learning objective

Learning Multi-Modal Word Representation Grounded in Visual Context

Multi-modal fusion techniques



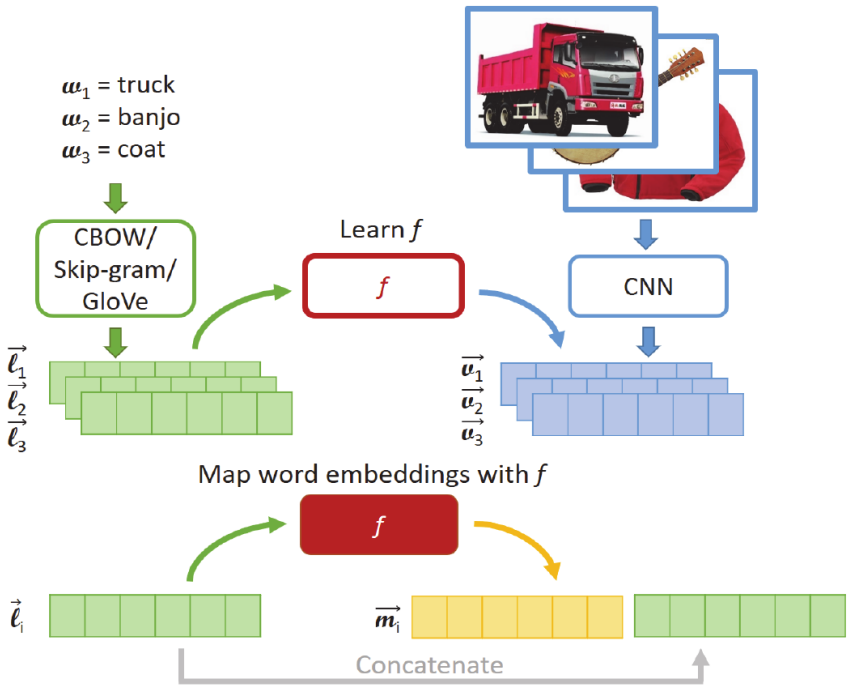
(a) *Early fusion* (b) *Middle fusion* (c) *Late fusion*

→ Joint models: aligned/non aligned text and images (skip-gram extension, grounded models)

→ Sequential models: combination of text representations (e.g., GloVe or Word2Vec) and image representations (pre-trained CNN):

- → **Middle fusion**: form multi-modal representations (e.g. concatenation, CCA)
- → **Late fusion**: interaction in the downstream task (e.g. linear combination of scores)

Example of sequential technique (middle fusion)



Sequential model (Collell et al. 2017)

Observation

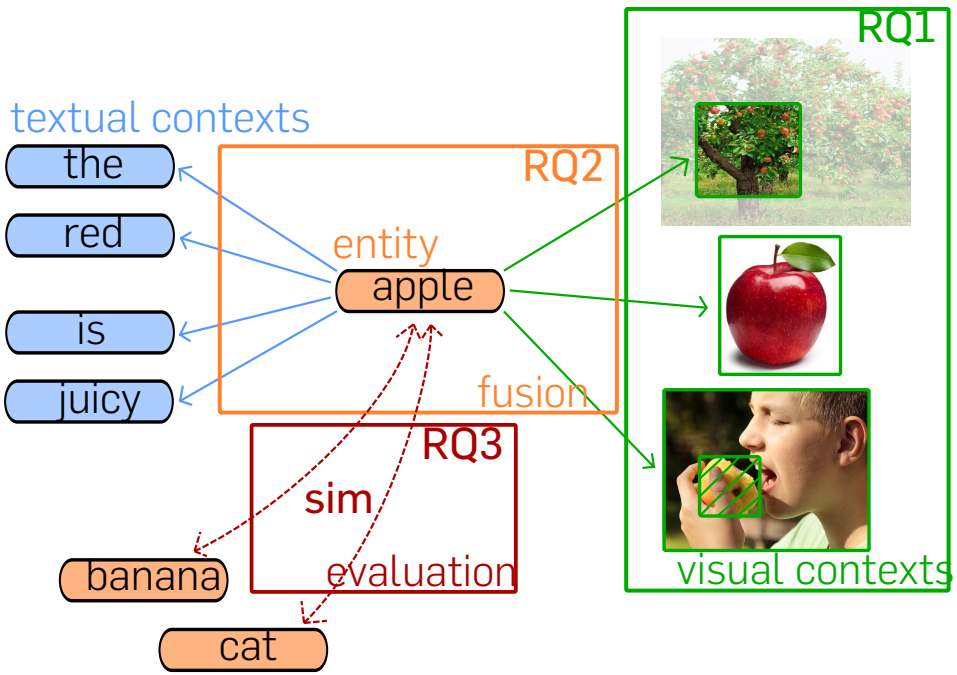
→ Multimodal models have shown **complementarity** of text and language ...

→ ...but use **direct features** from objects and ignore **visual context**

Research questions

Research Questions

- → **RQ1**: What is a visual context and how can we model it?
- → **RQ2**: How can we learn representations jointly from texts and images using contexts?
- → **RQ3**: How can we evaluate the contribution of the visual modality to the final embeddings?



Visual skip-gram: model

Recall: Text skip-gram

$$\mathcal{L}_{\text{text}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot u_c) + \sum_{c^-} \log \sigma(-t_e^\top \cdot u_{c^-}) \right]$$

T, U embedding tables, σ sigmoid function, $\mathcal{C}_e$ set of contexts of entity e

Visual skip-gram: model

Recall: Text skip-gram

$$\mathcal{L}_{\text{text}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot u_c) + \sum_{c^-} \log \sigma(-t_e^\top \cdot u_{c^-}) \right]$$

T, U embedding tables, σ sigmoid function, $\mathcal{C}_e$ set of contexts of entity e

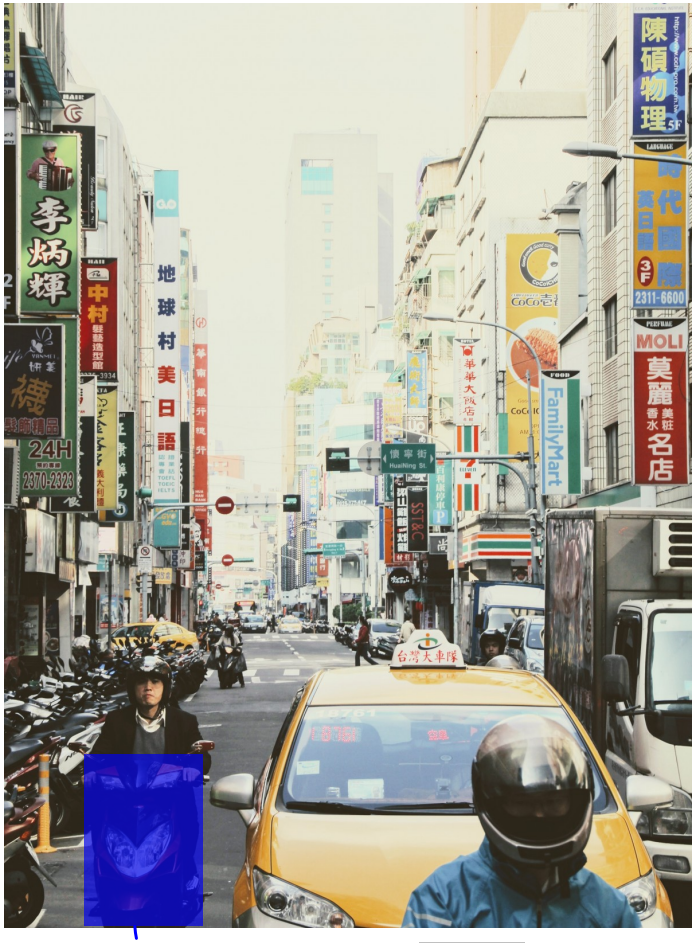
Idea: Use a skip-gram objective with visual contexts

$$\mathcal{L}_{\text{image}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot f_\theta(c)) + \sum_{c^-} \log \sigma(-t_e^\top \cdot f_\theta(c^-)) \right]$$

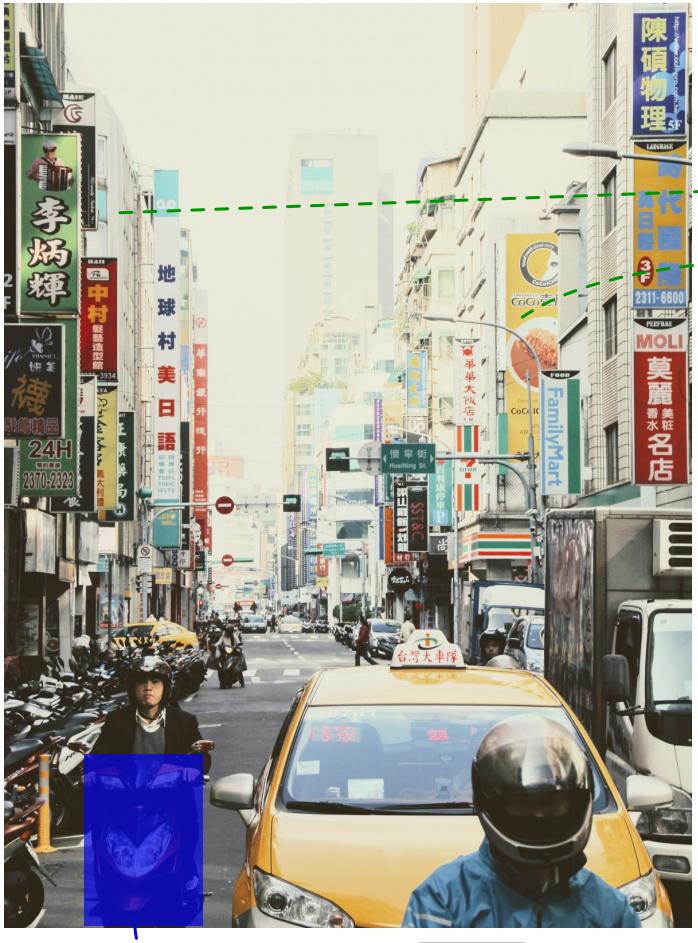
→ What is the context $c \in \mathcal{C}_e$?

→ What is the context modeling function $f_\theta(c)$?

Visual skip-gram: instantiation



Visual skip-gram: instantiation



CONTEXT

high-level: *other objects in image*

building

advertising board

$\in \mathbb{R}^d$

Embedding table

ENTITY

scooter

$\in \mathbb{R}^d$

Embedding table

Visual skip-gram: instantiation

CONTEXT

high-level: *other objects in image*

building

advertising board

$\in \mathbb{R}^d$

Embedding table

low-level:

- *full-image where entity is masked out*
- *random patches around entity*

Visual context

CNN

$\in \mathbb{R}^{2048}$

N

$\in \mathbb{R}^d$

projection

extract visual features

ENTITY

scooter

$\in \mathbb{R}^d$

Embedding table

Multi-modal skip-gram

Model

$$\mathcal{L}(T, U, \theta) = \mathcal{L}_{\text{text}}(T, U) + \alpha \mathcal{L}_{\text{image}}(T, \theta), \text{ where } \alpha \in [0, 1]$$

- → T **shared** multi-modal word embeddings
- → U **textual** context parameters
- → θ **visual** context parameters
- → T, U and θ learned with SGD ; α found with cross-validation

Grounding words in visual context: Experiments

Evaluation

- → **Word Similarity**: correlation between cosine similarity and human judgement. e.g. *sim*('cat', 'dog') = 0.8 ; *sim*('cloud', 'book') = 0.1 ;
- → **Feature-Norm Prediction**: predict objects' attributes from embedding with a linear SVM. e.g. *has_legs*('cat') = True, *is_red*('dog') = False
- → **Concreteness Prediction**: predict words' concreteness with a linear SVM. e.g. *conc*('dog')=0.9, *conc*('life')=0.1

Baselines

- → Text only
- → Text + direct visual features from objects
- → Text + visual contexts (sequential CCA)

Data

- → Wikipedia (4.5 million articles)
- → Visual Genome (108k images)

Multi-modal skip-gram: results

Results

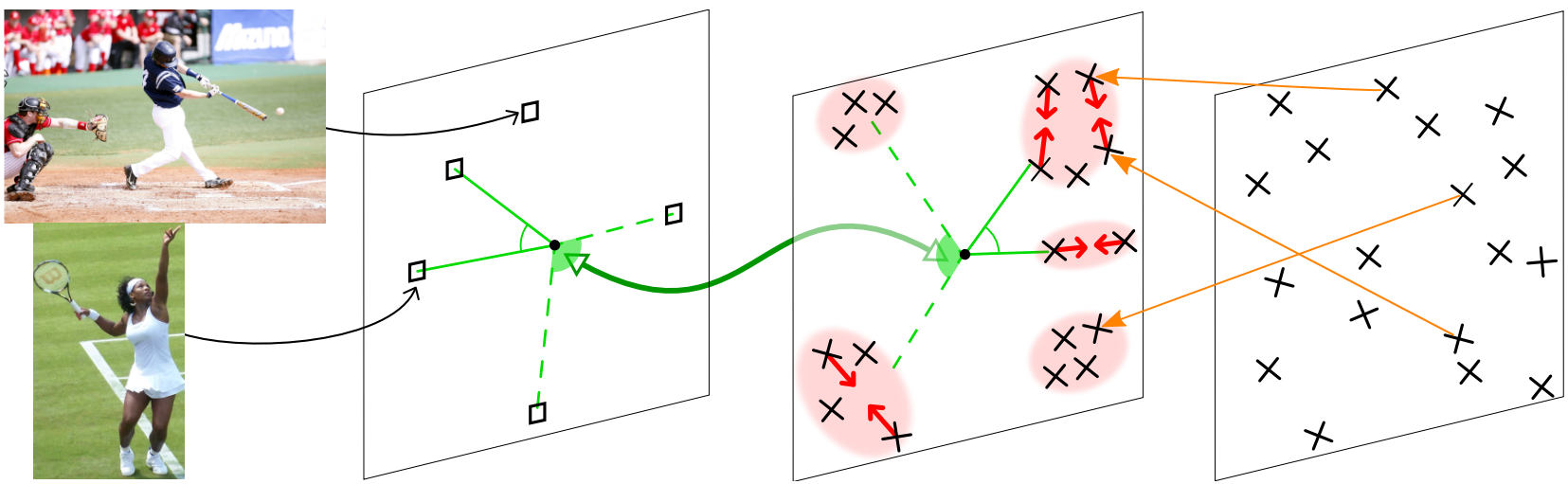
- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks
- → Visual context is **complementary** to visual features from objects
  - → ensemble model performs 6% better
- → **High-level context** > low-level context
  - → 1% average improvement on all tasks
- → **Joint model** > Sequential CCA
  - → 5% average improvement

Ongoing work: grounded sentences representations

We distinguish two sources of information

- → **Cluster information**: implicit knowledge that sentences associated with the same video refer to the same underlying reality
- → **Perceptual information**: high-level information extracted from a video using a pre-trained CNN

To preserve textual semantics and to avoid an over-constrained textual space: grounded space which partially transfer the structure of visual space in the textual one.



Conclusion and Perspectives

Conclusion and Perspectives

External resources can help NLP/IR tasks

- → ...for richer semantics/common sense captured in text representations
- → Investigating the the semantic gap/reporting bias between language and images/knowledge resources
- → Evaluating on real tasks (open-domain QA, machine translation, ...)

Language can help building external resources

- → Knowledge base completion
- → Commonsense mining

Language can help Computer Vision

- → using a semantic space (build with NLP techniques)
- → evaluating NLP models (captioning, VQA)
- → low/few supervision (e.g. zero-shot learning for object detection)

Open question

SCIENCES SORBONNE UNIVERSITÉ

Language Models as Knowledge Bases?

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, Sebastian Riedel

(Submitted on 3 Sep 2019 (v1), last revised 4 Sep 2019 (this version, v2))

Recent progress in pretraining language models on large textual corpora led to a surge of improvements for downstream NLP tasks. W
these models may also be storing relational knowledge present in the training data, and may be able to answer queries structured as "
Language models have many advantages over structured knowledge bases: they require no schema engineering, allow practitioners to
relations, are easy to extend to more data, and require no human supervision to train. We present an in-depth analysis of the relationa
(without fine-tuning) in a wide range of state-of-the-art pretrained language models. We find that (i) without fine-tuning, BERT contai
competitive with traditional NLP methods that have some access to oracle knowledge, (ii) BERT also does remarkably well on open-don
supervised baseline, and (iii) certain types of factual knowledge are learned much more readily than others by standard language mode
surprisingly strong ability of these models to recall factual knowledge without any fine-tuning demonstrates their potential as unsuper
The code to reproduce our analysis is available at this https URL.

Comments: accepted at EMNLP 2019
Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:1909.01066 [cs.CL]
(or arXiv:1909.01066v2 [cs.CL] for this version)

Questions?

SCIENCES SORBONNE UNIVERSITÉ

—Thanks for your attention —

Questions ?