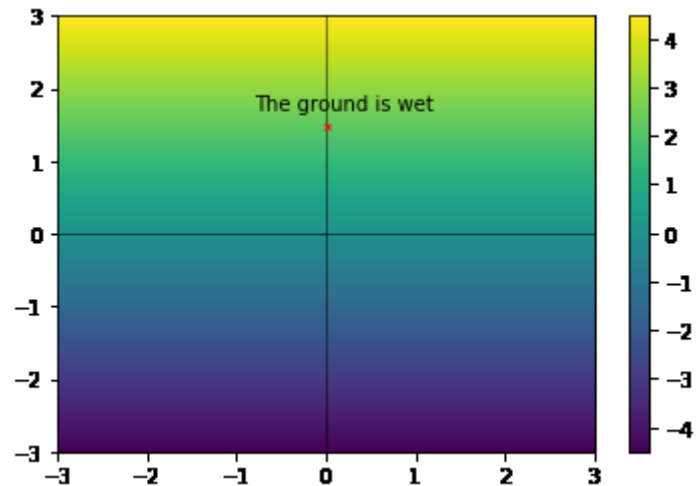


Représentations vectorielles de phrases : signification, compositionnalité et apprentissage

Damien Sileo
Philippe Muller
Tim Van de Cruys
Camille Pradel

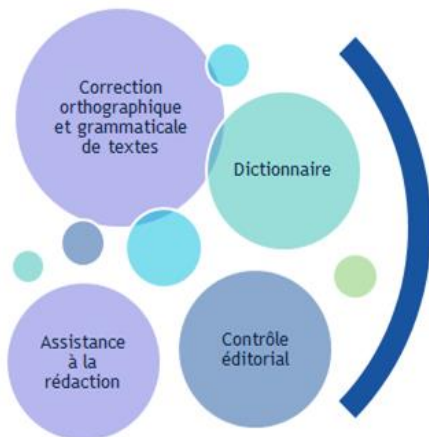


Thèse CIFRE Synapse Développement

similarité textuelle

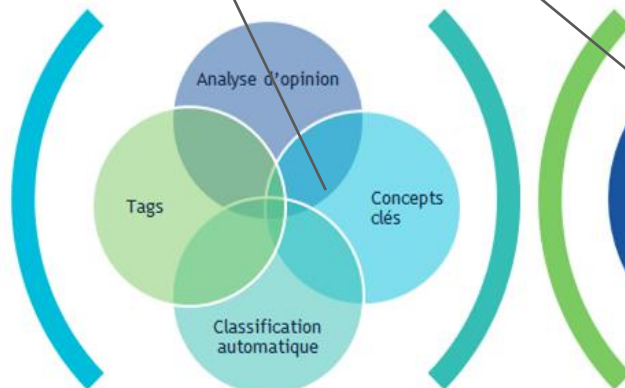
Cordial Business

Pour développer l'excellence rédactionnelle



Synapse Semantic Services

Pour donner de la valeur aux contenus textuels par l'analyse sémantique



Chatbot by Synapse

Pour optimiser la relation client en développant l'interactivité



Plan

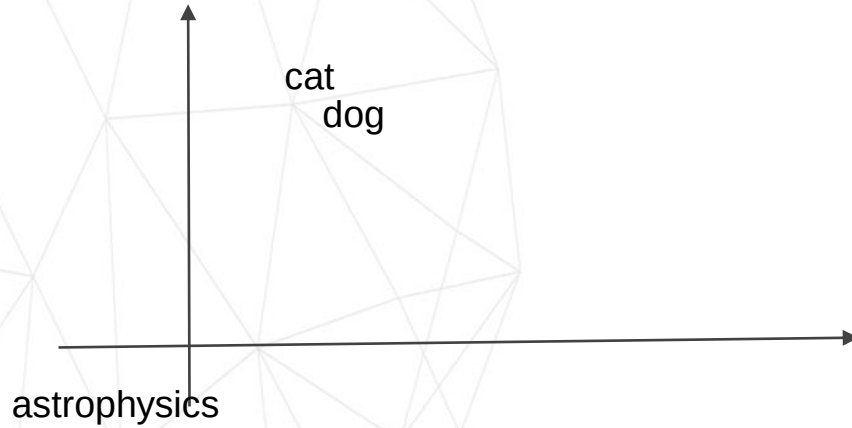
- Représentation vectorielle
- Modes de supervision
- Prédiction de relation entre les textes
- Apprentissage relationnel
- Nouvelles compositions
- Résultats
- Prédiction de connecteurs discursifs
- Résultats



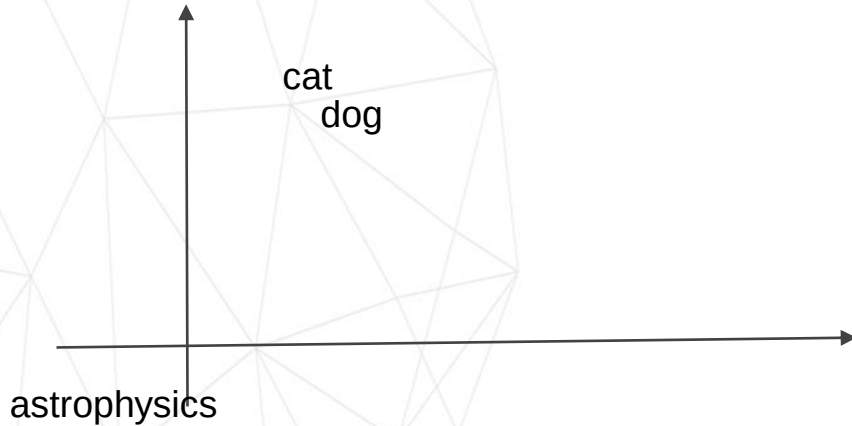
Représentations vectorielles



Représentation vectorielle de mots



Représentation vectorielle de mots



- cf tuto Bénédicte
- hypothèse distributionnelle + embeddings
- résultats surhumains dans certains contextes (vocabulaire non-spécialisé)

→ et pour les phrases ?



Distributional hypothesis (exercice...)

The functional interplay of philosophy and
...and among works of dystopian

The rapid advance in
...calculus, which are more popular in

But because
...the value of opinions formed in

...if
...certainty which every new discovery in

...if the new technology of computer
...frightened by the powers of destruction

...but there is also specialization in

should, as a minimum, guarantee..
fiction...

today suggests...

-oriented schools.

is based on mathematics...

as well as in the religions...

can discover the laws of human nature....

either replaces or reshapes.

is to grow significantly He got a

has given

and technology...

Distributional hypothesis (exercice...)

The functional interplay of philosophy and	science	should, as a minimum, guarantee..
...and among works of dystopian	science	fiction...
The rapid advance in	science	today suggests...
...calculus, which are more popular in	science	-oriented schools.
But because	science	is based on mathematics...
...the value of opinions formed in	science	as well as in the religions...
...if	science	can discover the laws of human nature....
...certainty which every new discovery in	science	either replaces or reshapes.
...if the new technology of computer	science	is to grow significantly He got a
...frightened by the powers of destruction	science	has given
...but there is also specialization in	science	and technology...

Distributional hypothesis: sentences

C'est pas facile de généraliser
word2vec aux phrases.

(phrase)

Et leurs contextes aussi.



Distributional hypothesis: sentences

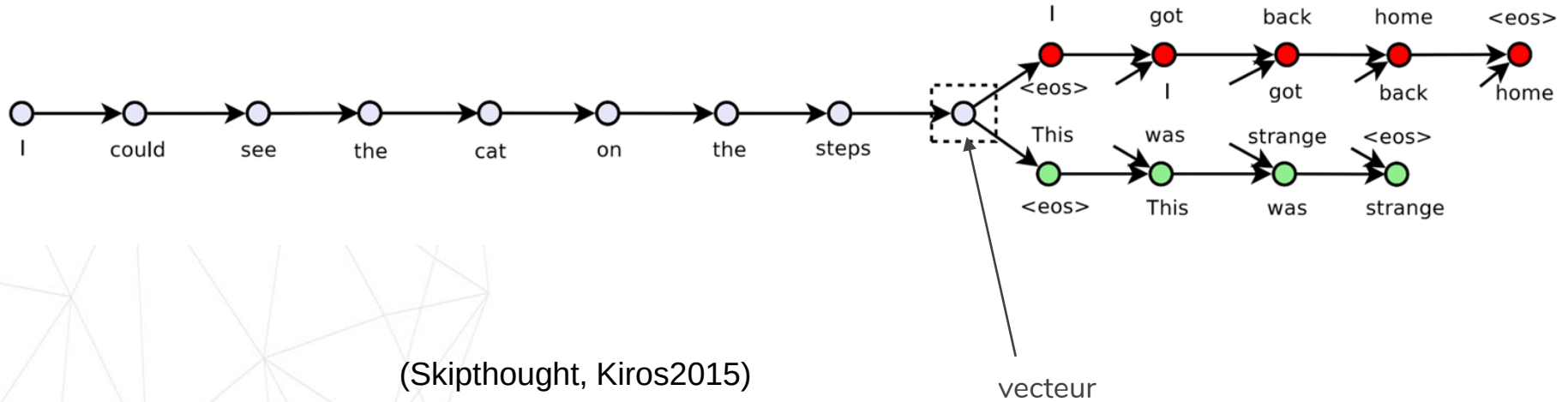
C'est pas facile de généraliser
word2vec aux phrases.

Parce que souvent, les phrases
n'apparaissent qu'une seule fois.

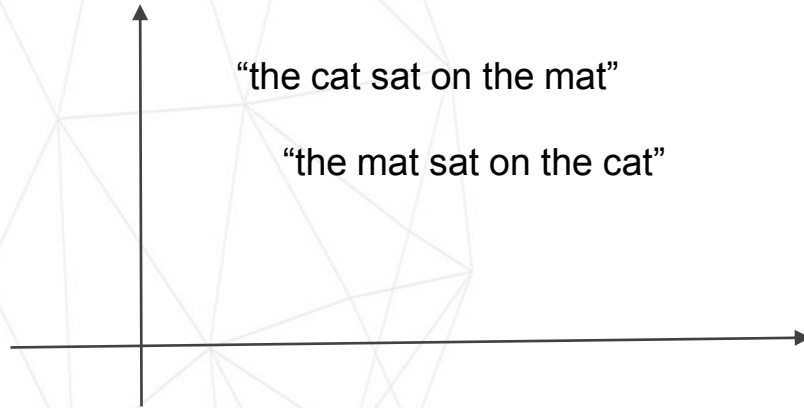
Et leurs contextes aussi.



Représentation vectorielle de phrases



Représentation vectorielle de phrase



Propriétés souhaitables

- transférable
- comparaisons possibles (cosinus, etc)

Que voudrait-on représenter ?

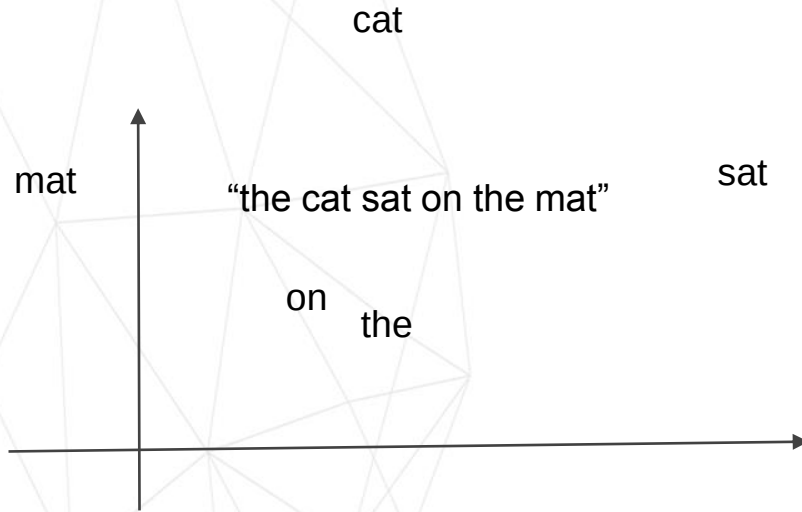
- formule logique ? relations ?
- objectif ?
- aspects ? (qui/quoi/quand/pourquoi...)

Alternatives :

- sous-représentations (mots)
- représentations gaussiennes



Exemple 1 : moyenne des mots



Avantages

- marche (trop ?) bien
- simple
- capture des aspects lexicaux
- mais pas que

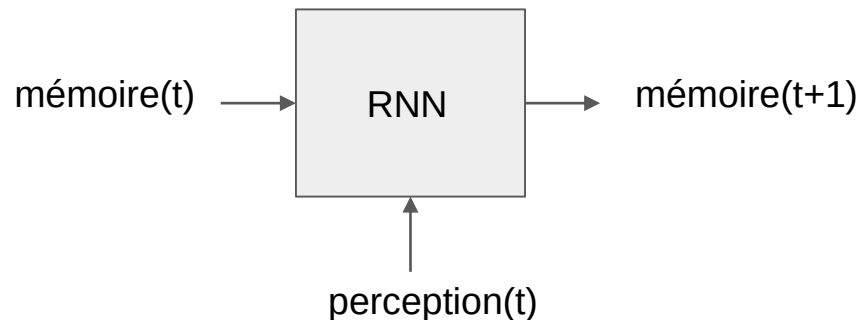
Inconvénients

- simpliste



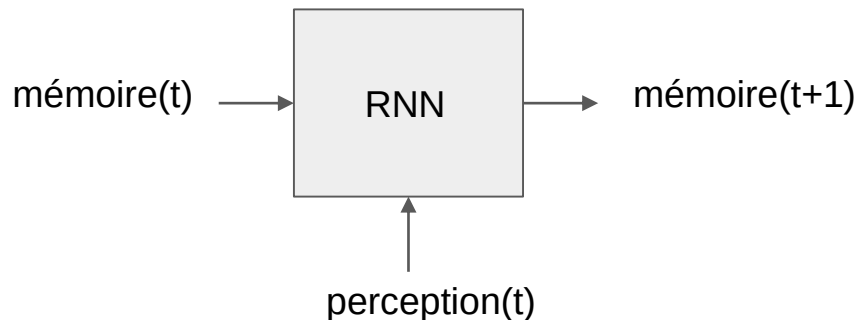
Exemple 2 : réseaux de neurones récurrents

Fonctions récurrentes et paramétriques



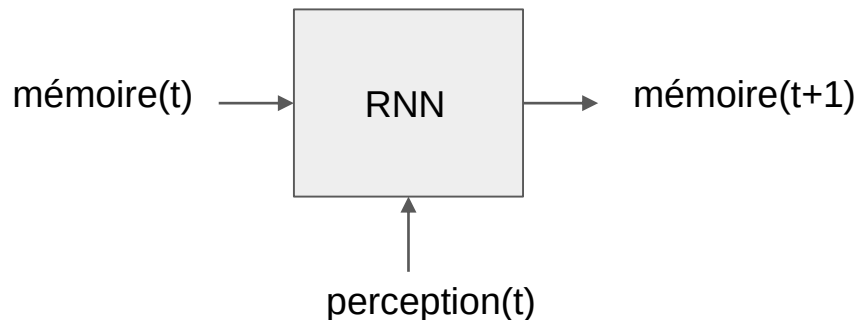
Exemple 2 : réseaux de neurones récurrents

Fonctions récurrentes et paramétriques



Exemple 2 : réseaux de neurones récurrents

Fonctions récurrentes et paramétriques



$$h_t = \tanh \left(W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \right)$$

$$h = f(x, h)$$

$f+x$ = intelligence

Comment évaluer ?

Mooney's (alleged) famous quote:

“you cannot cram the meaning of a whole sentence into a single vector”

Evaluation of potential of representations to help learn content-related tasks:

- Conneau et al., *SentEval: An Evaluation Toolkit for Universal Sentence Representations*, LREC 2018.
- Baroni et al. *What you can cram into a single vector: Probing sentence embeddings for linguistic properties*. ACL 2018

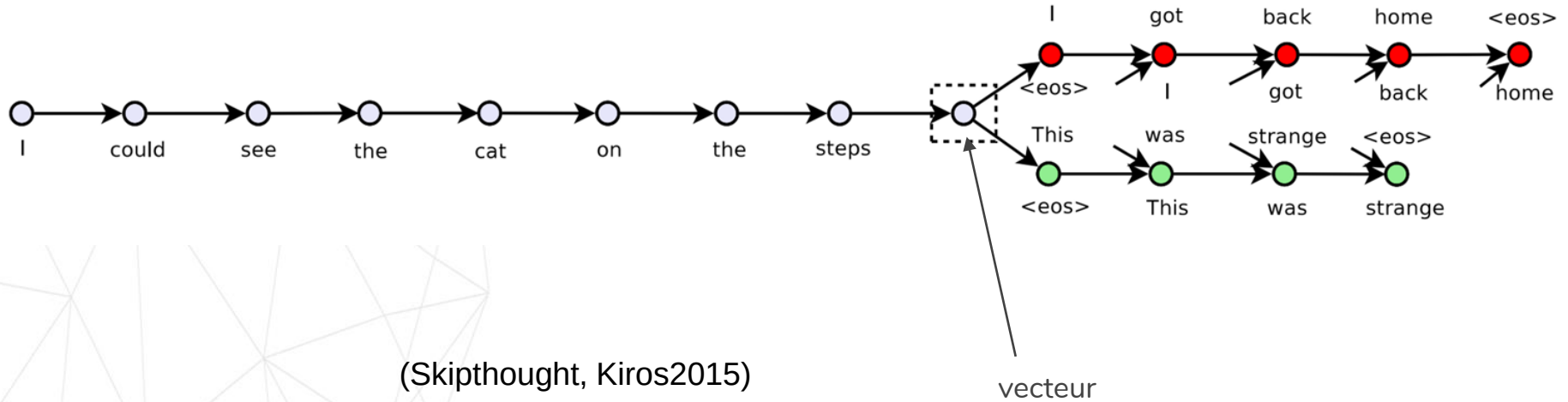
Evaluation sur les propriétés

	SentLen	WC	TreeDepth	TopConst	BShift	Tense	SubjNum	ObjNum	SOMO	CoordInv
Majority vote	20.0	0.5	17.9	5.0	50.0	50.0	50.0	50.0	50.0	50.0
Hum. Eval.	100	100	84.0	84.0	98.0	85.0	88.0	86.5	81.2	85.0
<i>BiLSTM-max encoder</i>										
Untrained	66.2	88.8	43.1	68.8	70.3	88.7	84.6	81.7	73.0	69.1
AutoEncoder	98.5	17.5	42.3	71.0	69.5	85.7	85.0	80.9	73.0	70.9
NMT En-Fr	79.3	58.3	45.7	80.5	71.2	87.8	88.1	86.3	69.9	71.8
NMT En-De	78.2	56.0	46.9	81.0	69.8	89.1	89.7	87.9	71.3	73.5
NMT En-Fi	77.5	58.3	45.8	80.5	69.7	88.2	88.9	86.1	71.9	72.8
Seq2Tree	91.8	10.3	54.6	88.7	80.0	89.5	91.8	90.7	68.6	69.8
SkipThought	59.6	35.7	42.7	70.5	73.4	90.1	83.3	79.0	70.3	70.1
NLI	65.1	87.3	38.5	67.9	63.8	86.0	78.9	78.5	59.5	64.9

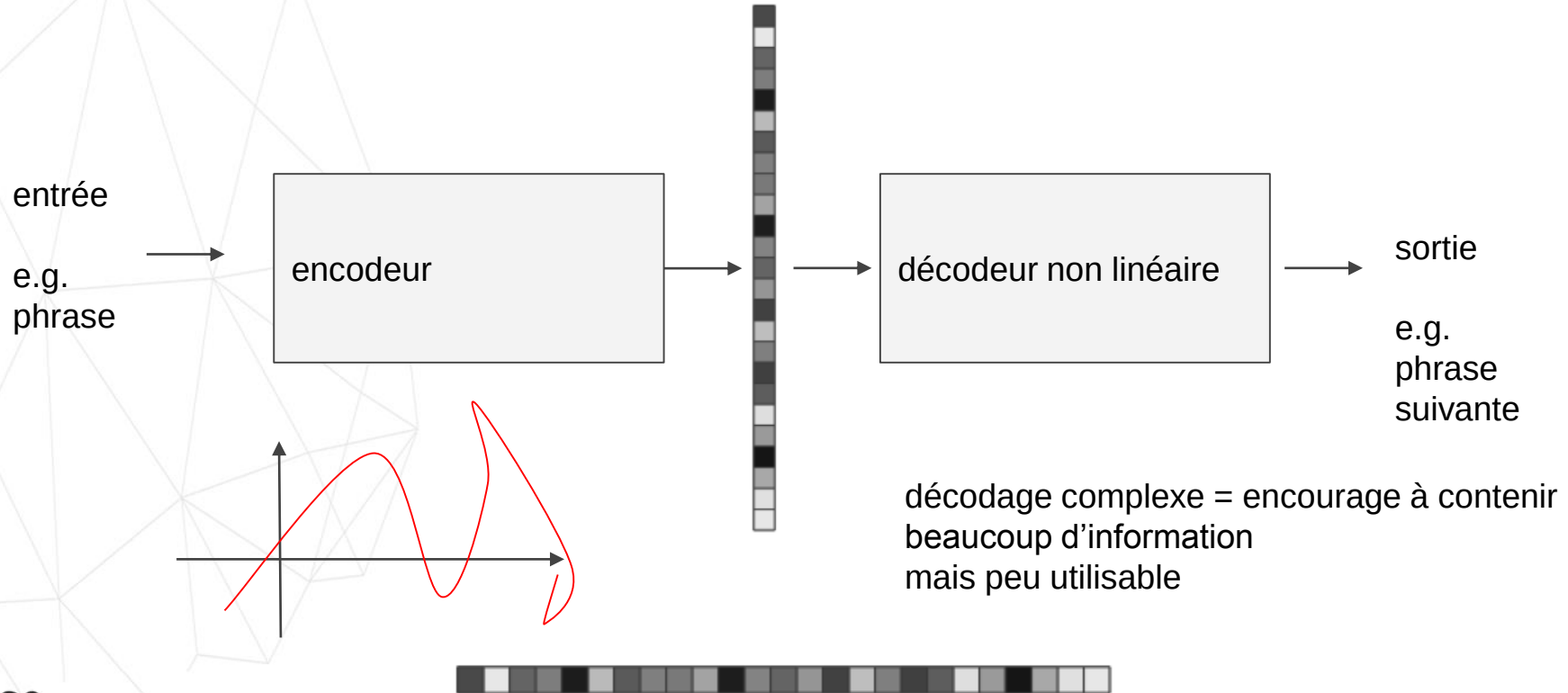
Evaluation en transfert

Model	MR	CR	SUBJ	MPQA	SST-2	SST-5	TREC	MRPC	SICK-E	SICK-R	STSB
<i>Baseline representations</i>											
Chance	50.0	63.8	50.0	68.8	50.0	28.6	21.2	66.5	56.7	0.0	0.0
<i>BiLSTM-max encoder</i>											
Untrained	75.6	78.2	90.0	88.1	79.9	39.1	80.6	72.2	80.8	83.3	70.2
AutoEncoder	68.3	74.0	87.2	84.6	70.8	34.0	85.0	71.0	75.3	70.4	55.1
NMT En-Fr	76.5	81.1	91.4	89.7	77.7	42.2	89.6	75.1	79.3	78.8	68.8
NMT En-De	77.7	81.2	92.0	89.7	79.3	41.0	88.2	76.2	81.0	80.0	68.7
NMT En-Fi	77.0	81.1	91.5	90.0	80.3	43.4	87.2	75.0	81.7	80.3	69.5
Seq2Tree	65.2	74.4	88.3	80.2	66.5	31.6	85.0	72.0	74.8	65.1	36.1
SkipThought	78.0	82.8	93.0	87.3	81.5	41.9	86.8	73.2	80.0	82.0	71.5
NLI	79.2	86.7	90.0	89.8	83.5	46.4	86.0	74.5	84.5	87.5	76.6

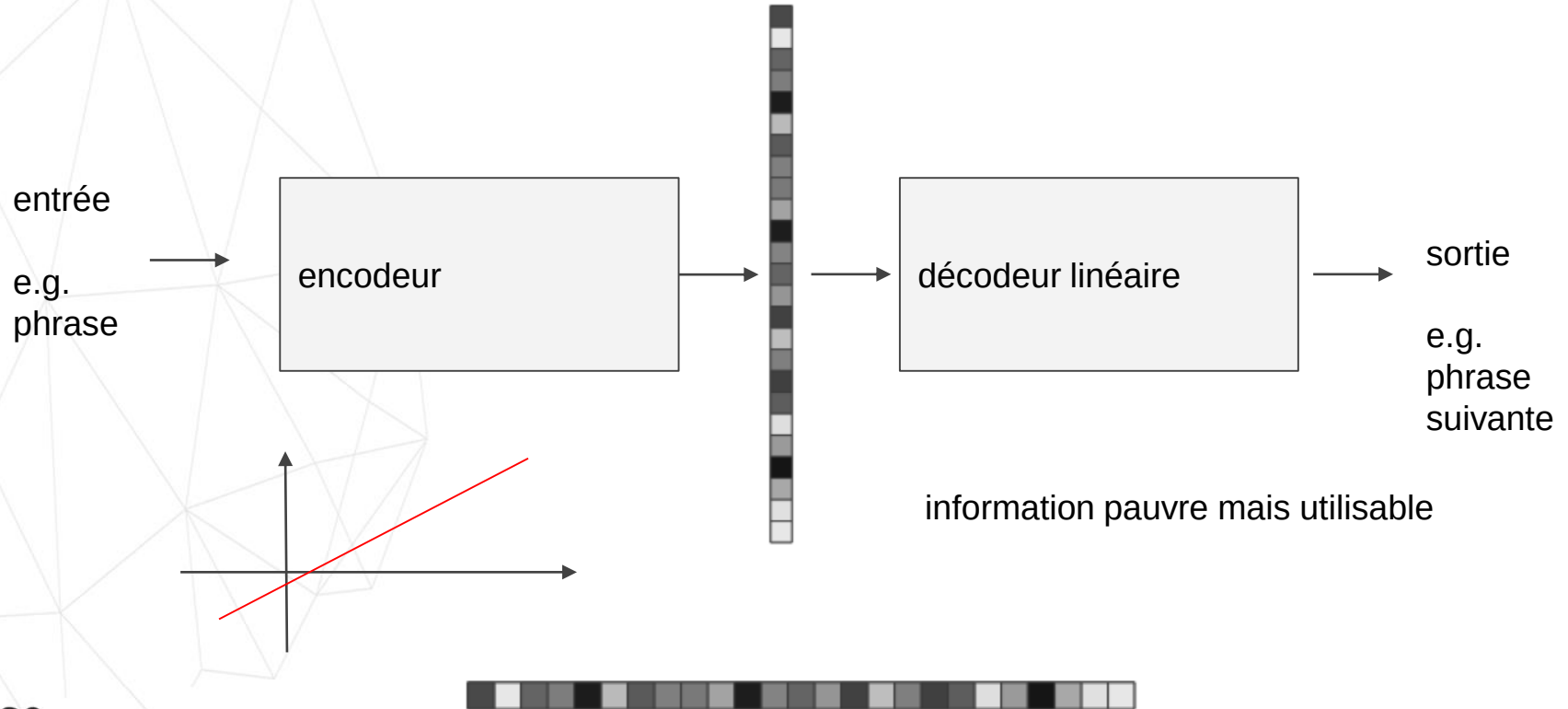
Représentation vectorielle de phrases



Compromis information/information utilisable



Compromis information/information utilisable



Comment apprendre ?

données	exemples	transférabilité
(x,x)	auto-encodage, modèle de langue	oui si décodeur simple
(x,y)	catégorisation, sentiments	oui mais par aspects
(x,x')	traduction, dialogue	oui si décodeur simple
$((x,x'),y)$	similarité, nli, discours	oui?

x, x' = phrases
 y = label



Plusieurs tâches (x,x',y) déjà établies:

- Natural language inference (SNLI, SICK, RTE)
(*"a man is typing"*, *"there is a man"*, **entailment**)
- Similarité sémantique (SICK, STS)
(*"a man is typing"*, *"a guy types"*, **similar**)
- Relations discursives (PDTB)
(*"a man is typing"*, *"he is fast"*, **expansion**)
- Answer selection (TREC), text ordering, etc....

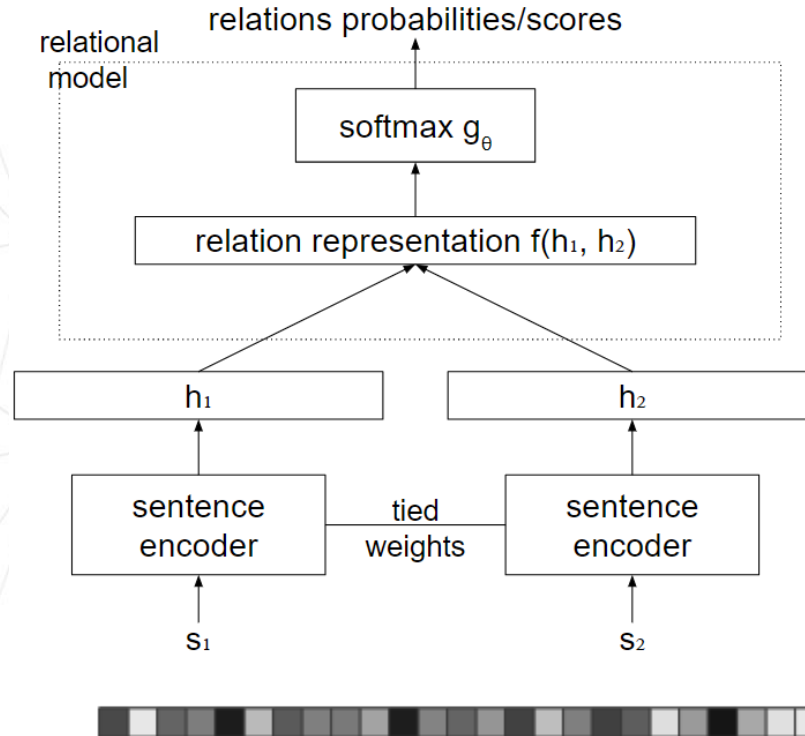
Sentence Pair Scoring (<https://arxiv.org/pdf/1603.06127.pdf>) unifie ces tâches



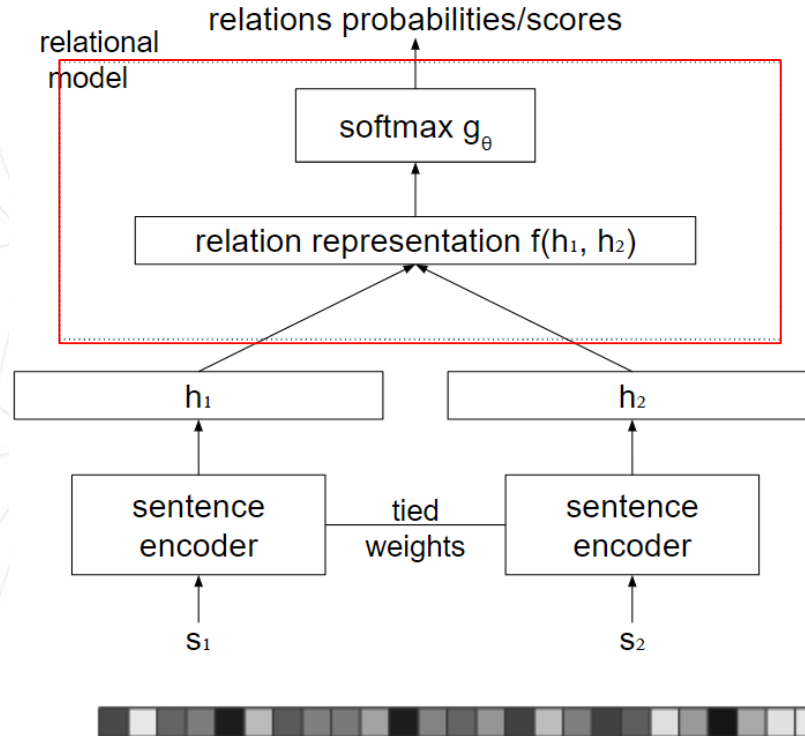
Composition de vecteurs pour le raisonnement



Modèle d'apprentissage relationnel textuel



Modèle d'apprentissage relationnel textuel



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
- ???



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - **asymétrique**, **interaction faible**



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - **asymétrique**, **interaction faible**
- hadamard $h_1 \odot h_2$



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - asymétrique, interaction faible
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - asymétrique, interaction faible
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte
- absolute difference $|h_1 - h_2|$



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - asymétrique, interaction faible
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte
- absolute difference $|h_1 - h_2|$
 - symétrique, interaction forte



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - asymétrique, interaction faible
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte
- absolute difference $|h_1 - h_2|$
 - symétrique, interaction forte
- outer $h_1 \otimes h_2$

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{A} = \begin{bmatrix} u_1 v_1 & u_1 v_2 & \dots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \dots & u_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_m v_1 & u_m v_2 & \dots & u_m v_n \end{bmatrix}$$



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - **asymétrique**, **interaction faible**
- hadamard $h_1 \odot h_2$
 - **symétrique**, **interaction forte**
- absolute difference $|h_1 - h_2|$
 - **symétrique**, **interaction forte**
- outer $h_1 \otimes h_2$
 - **asymétrique**, **interaction forte**

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{A} = \begin{bmatrix} u_1 v_1 & u_1 v_2 & \dots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \dots & u_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_m v_1 & u_m v_2 & \dots & u_m v_n \end{bmatrix}$$



Compositions existantes

features $f(h_1, h_2)$

- concatenation $[h_1, h_2]$
 - **asymétrique**, **interaction faible**
- hadamard $h_1 \odot h_2$
 - **symétrique**, **interaction forte**
- absolute difference $|h_1 - h_2|$
 - **symétrique**, **interaction forte**
- outer $h_1 \otimes h_2$
 - **asymétrique**, **interaction forte**, **dimension d^2**

$$\mathbf{u} \otimes \mathbf{v} = \mathbf{A} = \begin{bmatrix} u_1 v_1 & u_1 v_2 & \dots & u_1 v_n \\ u_2 v_1 & u_2 v_2 & \dots & u_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_m v_1 & u_m v_2 & \dots & u_m v_n \end{bmatrix}$$



Apprentissage relationnel

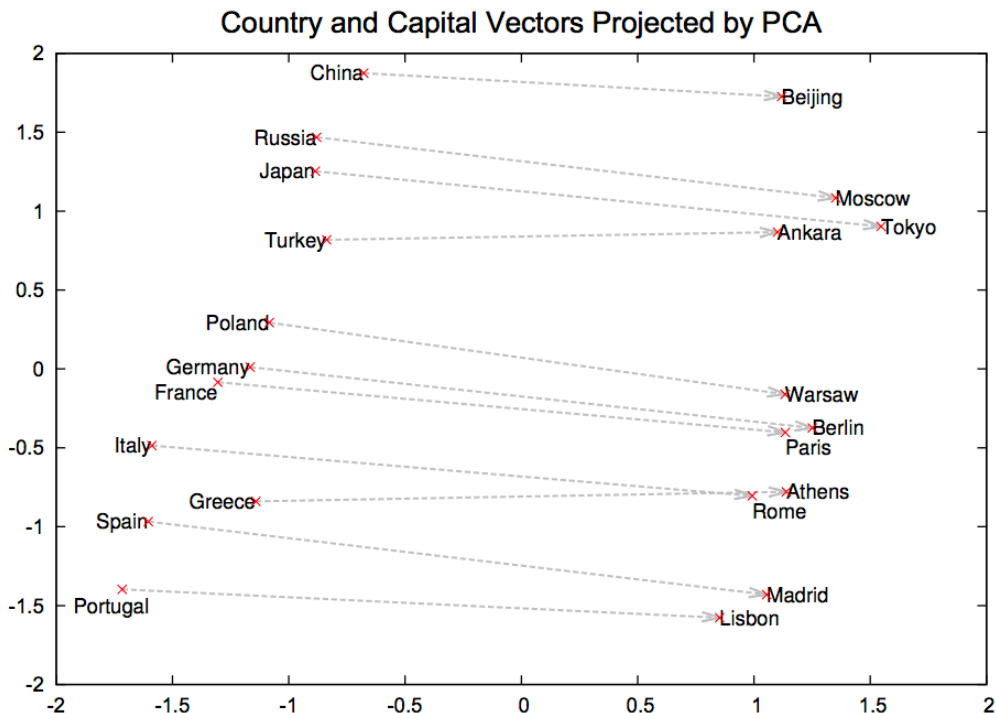
- un autre domaine du Machine Learning
- Un peu lié au NLP
- données (e_i, r, e_o)
- KB type Yago, Freebase...
- entités = embeddings
- relations = fonctions
- score = $f_r(e_i, e_o)$

Exemples

TransE : $f_r(e_i, e_o) = |e_i + r - e_o|$

DistMult : $f_r(e_i, e_o) = |e_i \odot r \odot e_o|$

~30 articles sur les compositions



Modèles d'apprentissage relationnel

Model	Scoring function	Relation parameters
Unstructured (Bordes et al., 2013a)	$\ e_1 - e_2\ _p$	-
TransE (Bordes et al., 2013b)	$\ e_1 + w_r - e_2\ _p$	$w_r \in \mathbb{R}^d$
RESCAL (Nickel et al., 2011)	$e_1^T W_r e_2$	$W_r \in \mathbb{R}^{d^2}$
DistMult (Yang et al., 2015)	$\langle e_1, w_r, e_2 \rangle$	$w_r \in \mathbb{R}^d$
ComplEx (Trouillon et al., 2016)	$\text{Re} \langle e_1, w_r, \overline{e_2} \rangle$	$w_r \in \mathbb{C}^d$

- absolute difference $|h_1 - h_2|$
 - symétrique, interaction forte
- outer $h_1 \otimes h_2$
 - asymétrique, interaction forte, dimension d^2
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte



Modèles d'apprentissage relationnel

Model	Scoring function	Relation parameters
Unstructured (Bordes et al., 2013a)	$\ e_1 - e_2\ _p$	-
TransE (Bordes et al., 2013b)	$\ e_1 + w_r - e_2\ _p$	$w_r \in \mathbb{R}^d$
RESCAL (Nickel et al., 2011)	$e_1^T W_r e_2$	$W_r \in \mathbb{R}^{d^2}$
DistMult (Yang et al., 2015)	$\langle e_1, w_r, e_2 \rangle$	$w_r \in \mathbb{R}^d$
ComplEx (Trouillon et al., 2016)	$\text{Re} \langle e_1, w_r, \overline{e_2} \rangle$	$w_r \in \mathbb{C}^d$

- absolute difference $|h_1 - h_2|$
 - symétrique, interaction forte
- outer $h_1 \otimes h_2$
 - asymétrique, interaction forte, dimension d^2
- hadamard $h_1 \odot h_2$
 - symétrique, interaction forte

Modèles d'apprentissage relationnel

Model	Scoring function	Relation parameters
Unstructured (Bordes et al., 2013a)	$\ e_1 - e_2\ _p$	-
TransE (Bordes et al., 2013b)	$\ e_1 + w_r - e_2\ _p$	$w_r \in \mathbb{R}^d$
RESCAL (Nickel et al., 2011)	$e_1^T W_r e_2$	$W_r \in \mathbb{R}^{d^2}$
DistMult (Yang et al., 2015)	$\langle e_1, w_r, e_2 \rangle$	$w_r \in \mathbb{R}^d$
Complex (Trouillon et al., 2016)	$\text{Re} \langle e_1, w_r, \bar{e}_2 \rangle$	$w_r \in \mathbb{C}^d$

- absolute difference $|h_1 - h_2|$
 - symétrique, interaction forte
 - outer $h_1 \otimes h_2$
 - asymétrique, interaction forte, dimension d^2
 - hadamard $h_1 \odot h_2$
 - symétrique, interaction forte
- ???
???

Nouvelles compositions

features $f(h_1, h_2)$

- ComplEx $[h_1 \odot h_2, h_1^r \odot h_2^c - h_1^c \odot h_2^r]$
 - asymétrique, interaction forte
- translation $|h_1 - h_2 + t|$
 - asymétrique, interaction forte



Nouvelles compositions

features $f(h_1, h_2)$

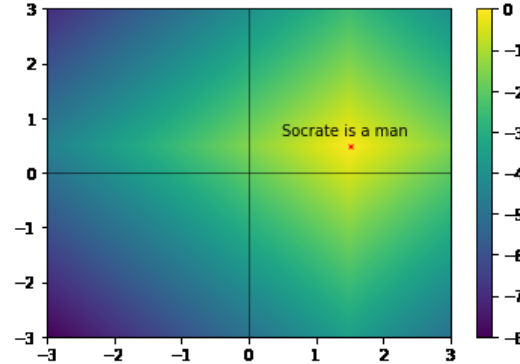
- ComplEx $[h_1 \odot h_2, h_1^r \odot h_2^c - h_1^c \odot h_2^r]$
 - asymétrique, interaction forte [□](#)
- translation $|h_1 - h_2 + t|$
 - asymétrique, interaction forte [□](#)



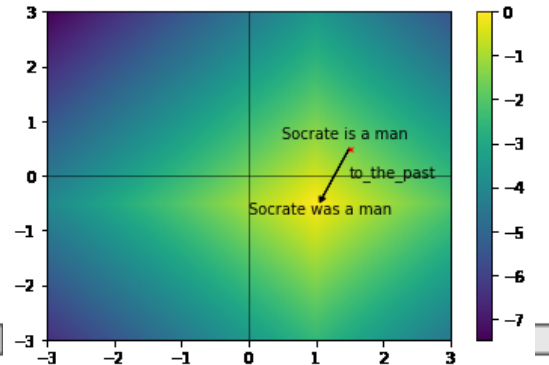
Visualisation - Heatmap (to_the_past, s1)

Absolute
 $|h_1 - h_2|$

to the past
↙



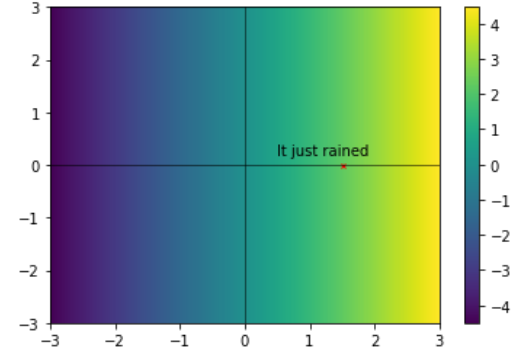
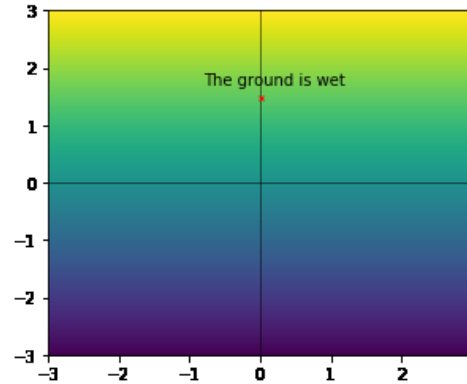
Translation
 $|h_1 - h_2 + t|$



Visualisation - Heatmap (entailment, s1)

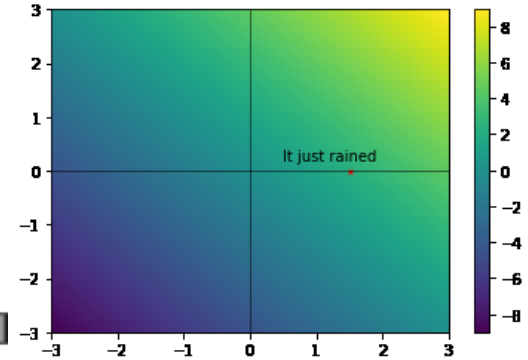
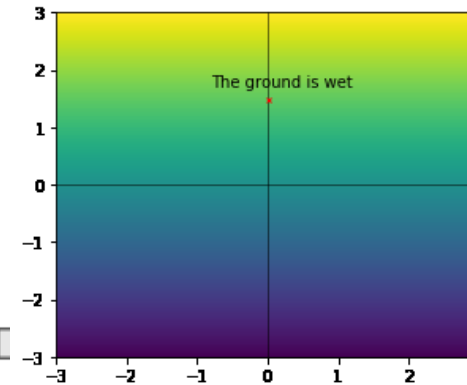
ground
moisture

Hadamard
 $[h_1 \odot h_2]$



nothing → recent rain
occurrence

Complex
 $[h_1 \odot h_2,$
 $hr_1 \odot hc_2 -$
 $hc_1 \odot hr_2]$



Results - Setup

Encodage bi-LSTM dimension 2*2048

Embedding pré-entraînés

Optimisation SGD

6 runs par modèle

Deux tâches:

- Inférence

le chat est sur la table

un félin est sur la table,

implication

- Prédiction connecteurs de discours

il fait beau

je suis content

donc



Results - NLI

	subjectivité			type de	paraphrase		discours		inférence			
	sentiment		polarité	question								
m,s	MR	SUBJ	MPQA	TREC	MRPC [⊖]	MRPC ^β	PDTB [⊖]	PDTB ^β	SICK [⊖]	SICK ^β	STS14	$\mathcal{T}$
$\mathcal{T}=\text{NLI}$												
$\odot, -^{Infer}_{Sent}$	81.1	92.4	90.2	88.2	76.2	-	-	-	86.3	-	70	84.5
$\odot, -$	81.2	92.7	90.4	89.6	76.1	76.2	46.7	47.2	86.6	86.9	69.5	84.2
$\alpha, -$	81.4	92.8	90.5	89.6	75.4	75.9	46.6	47.1	86.7	86.9	69.5	84.3
$\beta, -$	81.2	92.6	90.5	89.6	76	75.8	46.5	47	86.6	87	69.5	84.2
$\odot, t$	81.1	92.7	90.5	89.7	76.5	76.2	46.4	47.8	86.5	86.8	70	84.8
α, t	81.3	92.6	90.6	89.2	76.2	76.2	47.2	47.6	86.5	87.3	70	84.6
β, t	81.2	92.7	90.4	88.5	75.8	76.2	47.3	47.8	86.8	87	69.8	84.2

~0.3% de gain 📈



Results - Disc

	subjectivité			type de	paraphrase	discours		inférence				
	sentiment		polarité	question								
m,s	MR	SUBJ	MPQA	TREC	MRPC _⊖	MRPC _β	PDTB _⊖	PDTB _β	SICK _⊖	SICK _β	STS14	$\mathcal{T}$
$\mathcal{T}=\text{Disc}$												
$\odot, -$	80.4	92.7	90.2	89.5	74.5	74.8	47.3	48.2	83.2	83.6	57.9	35.7
$\alpha, -$	80.4	92.9	90.2	90.2	75	74.9	47.9	49.3	83.3	83.8	57.8	35.9
$\beta, -$	80.2	92.8	90.2	88.4	74.9	75	47.5	48.8	82.9	83.4	57.7	35.9
$\odot, t$	80.2	92.8	90.2	90.4	74.6	74.9	48.5	48.7	83.4	83.6	58.6	36.1
α, t	80.2	92.9	90.3	90.3	75.1	75.2	47.8	48.6	83.2	83.5	58.3	36.1
β, t	80.2	92.8	90.3	89.7	74.4	74.6	47.9	48.9	83.7	83.9	58.2	35.7

~0.3% de gain 📈



Additionally,

Prédiction de marqueurs de discours



Prédiction de marqueurs

Pourquoi est-ce une bonne tâche ?

- Données “gratuites”
- Sens = Utilisation
 - Wittgenstein's *Philosophical Investigations*: "For a large class of cases—though not for all—in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language."
- Marqueurs = labels bruités pour plusieurs tâches
 - entailment (c = therefore),
 - subjectivity analysis (c = factually)
 - sentiment analysis (c = sadly/happily/fortunately)
 - etc.

Problèmes:

- Travail linguistique nécessaire
- Ambiguïté, échelles non “phrasistique”



Découverte automatique de connecteurs de discours



Découverte automatique de connecteurs de discours

- 1) Collecte de milliards de paires de phrases s_1, s_2
- 2) Sélection des paires où $s_2[0]$ est un adverbe suivi par virgule
- 3) Equilibrage des fréquences des $s_2[0]$ (Certains connecteurs sont très fréquents, e.g. *But*, par rapport à *Logically*)



Découverte automatique de connecteurs de discours

- 1) Collecte de milliards de paires de phrases s_1, s_2
- 2) Sélection des paires où $s_2[0]$ est un adverbe suivi par virgule
- 3) Equilibrage des fréquences des $s_2[0]$ (Certains connecteurs sont très fréquents, e.g. *But*, par rapport à *Logically*)
- 4) Faire le tri



Découverte automatique de connecteurs de discours

- 1) Collecte de milliards de paires de phrases s_1, s_2
- 2) Selection des paires où $s_2[0]$ est un adverbe suivi par virgule
- 3) Equilibrage des fréquences des $s_2[0]$ (Certains connecteurs sont très fréquents, e.g. *But*, par rapport à *Logically*)
- 4) Faire le tri
 - Comment ?



Découverte automatique de connecteurs de discours

- 1) Collecte de milliards de paires de phrases s_1 , s_2
- 2) Selection des paires où $s_2[0]$ est un adverbe suivi par virgule
- 3) Equilibrage des fréquences des $s_2[0]$ (Certains connecteurs sont très fréquents, e.g. *But*, par rapport à *Logically*)
- 4) Faire le tri
 - Comment ?

$s_2[1:]$ doit être grammatical

$s_2[0]$ doit dépendre de l'interaction entre s_1 et $s_2[1:]$



Découverte automatique de marqueurs de discours

Idée: Prédire le marqueur à partir de s_1 et $s_2[1:]$ indépendamment

- Utilisation d'un classifieur linéaire + modèle trigramme (FastText)

candidate marker	accuracy (%)
evidently,	0.0
frequently,	0.0
meantime,	0.0
truthfully,	0.0
supposedly,	0.1

Table 4: Candidate discourse markers that are the most difficult to predict from shallow features

candidate marker	accuracy (%)
defensively,	65.5
afterward	71.1
preferably,	71.9
this,	72.7
very,	90.7



Table 5: Candidate discourse markers that are the easiest to predict from shallow features. This shows candidates that are unlikely to be interesting discourse cues.



Comparaison au précédent

author	discourse markers / classes	classes	markers
Jernite et al. (2017)	ADDITION, CONTRAST, TIME, RESULT, SPECIFIC, COMPARE, STRENGTH, RETURN, RECOGNIZE	9	40
Nie et al. (2017)	<i>and, but, because, if, when, before, though, so, as, while, after, still, also, then, although</i>	15	15
current work	<i>later, often, understandably, gradually, or, ironically, namely, ...</i>	174	174

Table 2: Discourse markers or classes used by previous work on unsupervised representation learning



[illegible]

[illegible]

Résultats

	<i>N</i>	MR	CR	SUBJ	MPQA	SST2	TREC	SICK-R	SICK-E	MRPC	AVG
InferSent	1.0	81.1	86.3	92.4	90.2	84.6	88.2	88.4	86.1	76.2	85.9
MTL	124	82.5	87.7	94	90.9	83.2	93	88.8	87.8	78.6	87.4
SkipThought	74	76.5	80.1	93.6	87.1	82	92.2	85.8	82.3	73	83.6
QuickThought	174	81.3	84.5	94.6	89.5	-	92.4	<u>87.1</u>	-	75.9	-
DisSent	4.7	80.1	84.9	93.6	90.1	84.1	93.6	84.9	83.7	75	85.6
DiscoveryBase	1.7	82.5	86.3	94.2	90.5	85.2	91.8	85.7	84	75.8	86.2
DiscoveryHard	1.7	81.6	86.5	93.9	90.5	84.8	90	85.4	83.2	76.5	85.8
Discovery10	1.7	81.2	85.1	93.7	90.2	83	90	85.9	83.8	75.8	85.4
DiscoveryAdv	1.4	81.4	85.8	93.8	90.5	83.4	92	86	84.3	75.7	85.9
DiscoveryShuffled	1.7	81.4	86.1	94.1	90.9	85.3	90.4	85.6	83.6	75.4	85.9
DiscoveryBig	3.4	82.6	<u>87.4</u>	94.5	91.0	85.2	93.4	86.4	<u>84.8</u>	<u>76.6</u>	<u>86.9</u>





SYNAPSE DÉVELOPPEMENT

Les experts de l'intelligence artificielle appliquée
au traitement du langage

Merci

Synapse Développement
7, Boulevard de la Gare
F-31500 Toulouse – France
+33(0)5 61 63 03 74

www.synapse-developpement.fr

 info@synapse-fr.com

 synapse_dev

 Synapse Développement