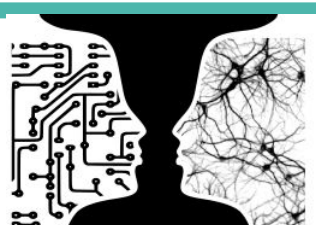
Spécialisation sémantique des méthodes de représentations distributionnelles

Apport des Lexiques Sémantiques pour les word embeddings

Roxane Debruyker, GRAIL



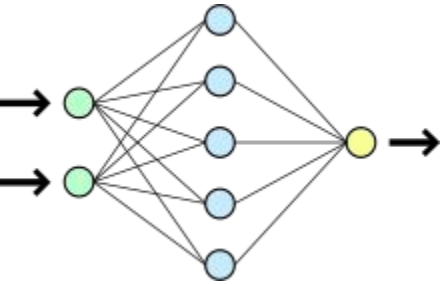
Group for Research in Artificial Intelligence of Laval University

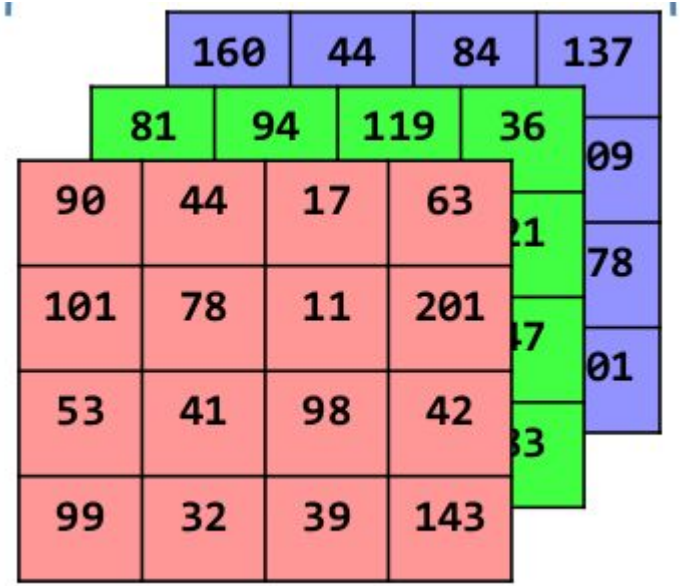
Plan

1. Représentation
2. Lexiques sémantiques
3. Modèles de spécialisation sémantique

1. Méthodes de représentations

a . Raison

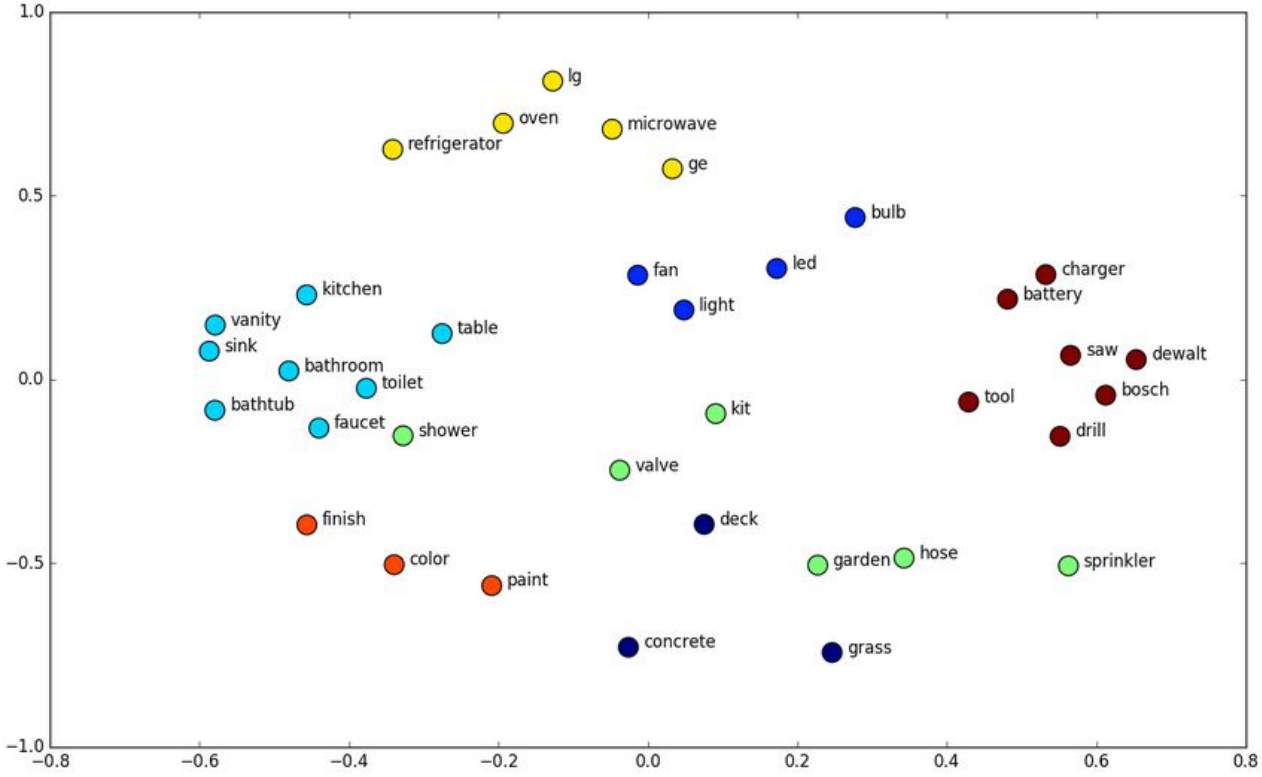




images de Wikipedia

b . Objectif

Proximité sémantique se reflète dans la proximité dans l'espace vectoriel



https://medium.com/opla/how-to-train-word-embeddings-using-small-datasets-9ced58b58fde

c . Méthodes distributionnelles

Zellig Harris (1954): *If A and B have almost identical environments we say that they are synonyms.*

What does ongchoi mean?

Suppose you see these sentences:

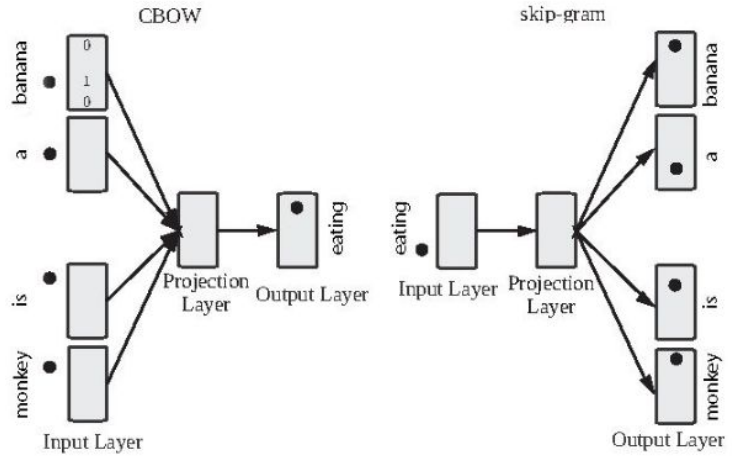
- Ong choi is delicious **sautéed with garlic**.
- Ong choi is superb **over rice**
- Ong choi **leaves** with salty sauces

And you've also seen these:

- …spinach **sautéed with garlic over rice**
- Chard stems and **leaves** are **delicious**
- Collard greens and other **salty** leafy greens

Conclusion:

- Ongchoi is a leafy green like spinach, chard, or collard greens

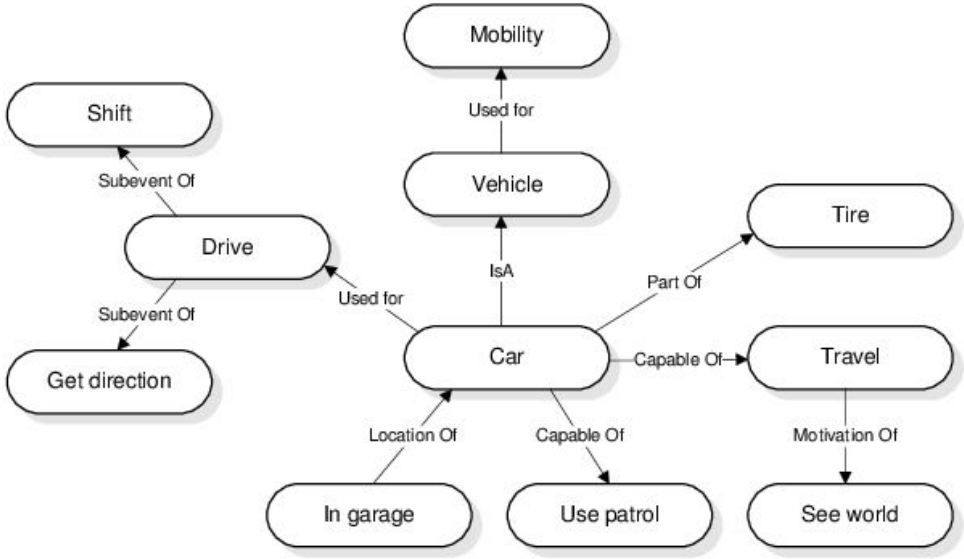


Jurafsky, D., & Manning, C.. Natural language processing. *Instructor*, *212*(998), 3482

2 . Lexiques sémantiques

Lexiques sémantiques

- Annoté par des experts
- Nombre de relations limitées (18 pour WN18 et 36 pour ConceptNet 5.5)
- Informations représentées sous forme de triplets (h, r, t)
- Pas utilisables seuls mais peuvent apporter une spécialisation sémantique



Organization Workshop Organizers - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Concepts-related-to-car-in-ConceptNet_fig1_277286109 [accessed 8 Jan, 2020]

BabelNet

Paraphrase.org

ConceptNet
An open, multilingual knowledge graph

WORDNET

3. Modèles de spécialisation sémantique

a. Modèles joints

- Change directement la fonction objectif : en plus de CBOW ou SGNS, on ajoute les informations de KG.
- Influence tout le vocabulaire
- Ne permet de spécialiser qu'un modèle vu que se greffe à la fonction objectif

$$J = \frac{1}{T}\underbrace{\sum_{t=1}^{T} \log P(w_t|w_{t-c}^{t+c})}_{\text{Distributional: CBOW}} + \underbrace{\frac{C}{N}\sum_{i=1}^{N}\sum_{w\in\mathbf{R}_{w_i}} \log P(w|w_i)}_{\text{Knowledge Resource}}$$

Yu, M., & Dredze, M. (2014, June). Improving lexical embed
Meeting of the Association for Computational Linguistics (Vo

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

b. Post-Processing : Retrofitting

- MAJ les représentations de vecteurs pré-entraînés
- Rapproche les vecteurs pour les concepts reliés
- Combinaison linéaire entre la position originale du vecteur et la moyenne des voisins
- Valeurs de α et β

$$L(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - q_i^0\|^2 + \sum_{(i,j) \in E} \beta_{i,j} \|q_i - q_j^0\|^2 \right]$$

Seulement deux matrices en mémoire ; converge rapidement

Mais très dépendant de l'ordre d'apparition

c. ConceptNet Numberbatch

Speer, R., Chin, J., & Havasi, C. (2017, February). Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

- MAJ simultanée pour tout le monde grâce à multiplication matricielle
- Les mots dans le lexiques qui ne sont pas pré-entraînés s'accumulent autour de 0

$$W^{k+1} = \text{normalize}\left[\left(SW^k + AW^0\right)(I + A)^{-1}\right]$$

Et les relations?

Les modèles précédents permettent d'utiliser des nouvelles connexions entre les concepts mais la nature des relations n'est pas utilisée

- Synonymes et antonymes
- Hyperonymie/hyponymie et relations asymétriques : Lexical Entailment
- Les 36 de ConceptNet

Objectif : avoir un *unique* modèle qui prend tout ça en compte

Synonyme/Antonyme

Counter-fitting

- Se rapproche des synonymes
- S'éloigne des antonymes
- Préserve l'espace vectoriel

Explicit retrofitting:

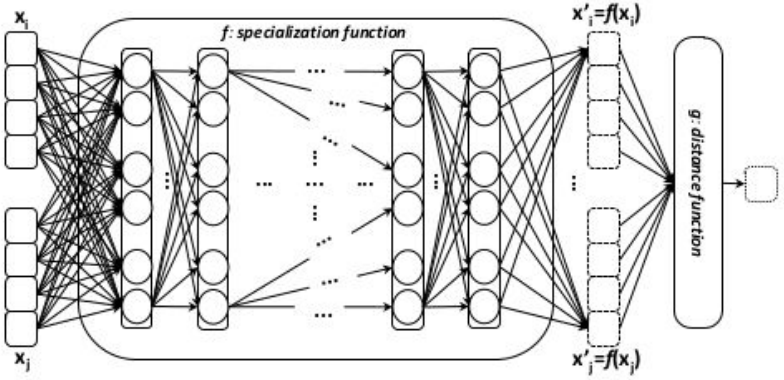
- Prend toutes les paires de mots (w_i, w_j, r)
- Prends les k plus proches voisins de i et j
- Rapproche les syn, éloigne les ant, laisse les autres à la même place
- Réseau de neurone profond

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

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

$$\mathrm{VSP}(V, V') = \sum_{i=1}^{N} \sum_{j\in N(i)} \tau\left(d(\mathbf{v}'_i, \mathbf{v}'_j) - d(\mathbf{v}_i, \mathbf{v}_j)\right)$$

Mrkšić, N., Séaghdha, D. O., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P. H., ... & Young, S. (2016). Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.



Glavaš, G., & Vulić, I. (2018). Explicit retrofitting of distributional word vectors.

Lexical Entailment Attract-Repel

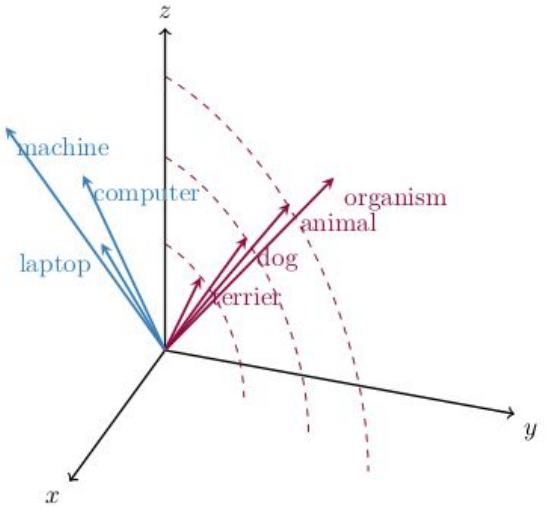
Vulić, I., & Mrkšić, N. (2017). Specialising word vectors for lexical entailment. *arXiv preprint arXiv:1710.06371*.

Joue sur la norme des vecteurs

La magnitude de la différence entre deux normes détermine l'intensité de la relation de hiérarchie

Les concepts reliés par relation hyper/hyponyme ont la même direction

$$C(\mathcal{B}_A, T_A, \mathcal{B}_R, T_R, \mathcal{B}_L, T_L) = Att(\mathcal{B}_S, T_S) + \dots + Rep(\mathcal{B}_A, T_A) + Reg(\mathcal{B}_A, \mathcal{B}_R, \mathcal{B}_L) + \dots + Att(\mathcal{B}_L, T_L) + LE_j(B_L)$$



$$D_1(\mathbf{x}, \mathbf{y}) = |\mathbf{x}| - |\mathbf{y}|$$

$$D_2(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x}| - |\mathbf{y}|}{|\mathbf{x}| + |\mathbf{y}|}$$

$$D_3(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x}| - |\mathbf{y}|}{\max(|\mathbf{x}|, |\mathbf{y}|)}$$

Mais si on veut représenter toutes les relations ?

Link prediction & Graph prediction

Contient beaucoup d'informations sur le monde en général, donc bien sûr incomplet ⇒ on a besoin d'outils pour prédire de nouveaux liens sachant les anciens

Transforme les entités et relations en embeddings

Link prediction : calcule la probabilité d'être un triplet valide ou non.

Graph prediction : Il ne faut pas juste prédire si les deux mots sont aussi reliés mais aussi savoir par quelle type de relations ils sont reliés

Calculer les graphs embeddings

Entités et relations dans des espaces vectoriels différents

Model	Score Function	Symmetry	Antisymmetry	Inversion	Composition
TransE	$\|\|h + r - t\|\|$	x	✓	✓	✓
DistMult	$< h, r, t >$	✓	x	x	x
ComplEx	$Re(< h, r, t >)$	✓	✓	✓	x
RotatE	$\|\|h \odot r - t\|\|$	✓	✓	✓	✓

Table 1: The pattern modeling and inference abilities of several models [Sun et al., 2019b].

$$f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) = \sigma(\boldsymbol{q_i}^T \boldsymbol{A_r} \boldsymbol{q_j})$$

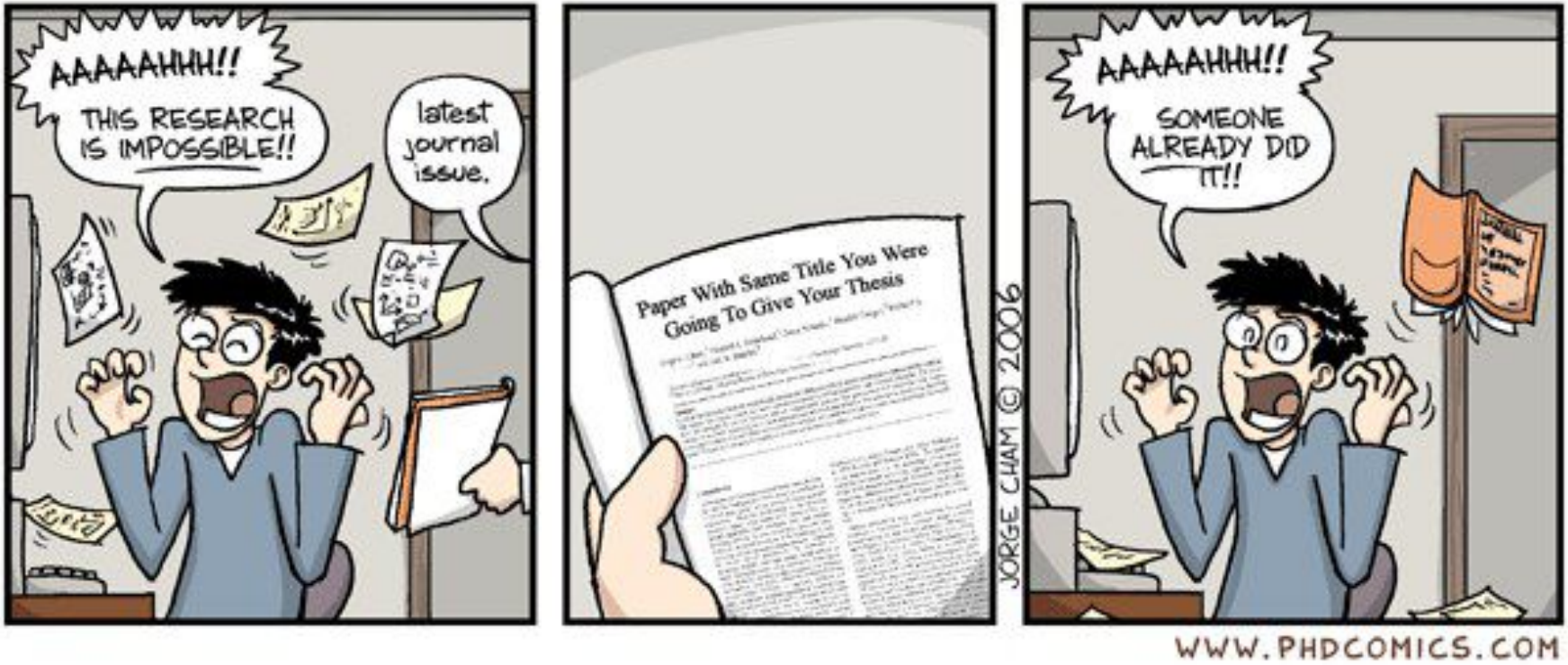
SimplE

Idée

Entraîner des 36 relations embeddings de ConceptNet avec SimplE (Mr). Fine-tuner mes embeddings ensuite, mais toujours sur ConceptNet… Redondant… mais apprendre tout d'une shot est biaisé aussi.

Mais pas sur tout, pas très neural, dur à dériver, …. Mais surtout ….

$$
\begin{aligned}
a &= || \, q_i^k - q_i^0 \, ||^2 \\
b &= || \, q_i^k - < t_j^{-1}, M_r >^T ||^2 \\
c &= || \, q_i^k - < M_{r-1}^{-1}, h_j^{T-1} > ||^2 \\
L(Q) &= \sum_{i \in Q} \alpha_i \cdot a + \beta_i \sum_{(j,r):(i,j,r) \in E} \beta_{i,j} \frac{b+c}{2}
\end{aligned}
\tag{5}
$$


AAAAAHHH!!
THIS RESEARCH IS IMPOSSIBLE!!
latest journal issue.
Paper With Same Title You Were Going To Give Your Thesis
JORGE CHAM © 2006
AAAAAHHH!!
SOMEONE ALREADY DID IT!!
WWW.PHDCOMICS.COM

Functional retrofitting

Lengerich, B. J., Maas, A. L., & Potts, C. (2017). Retrofitting distributional embeddings to knowledge graphs with functional relations. *arXiv preprint arXiv:1708.00112*.

$$\Psi_{\mathcal{G}}(\mathcal{Q};\mathcal{F}) = \sum_{i\in\mathcal{Q}} \alpha_i \|\boldsymbol{q_i} - \hat{\boldsymbol{q}}_{\boldsymbol{i}}\|^2 + \sum_{(i,j,r)\in\mathcal{E}} \beta_{i,j,r} f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) - \sum_{(i,j,r)\in\mathcal{E}^-} \beta_{i,j,r} f_r(\boldsymbol{q_i}, \boldsymbol{q_j}) + \sum_{r\in\mathcal{R}} \rho_\lambda(f_r)$$

La fonction de score entre les concepts est testée en étant :

- Linéaire
- Non-linéaire

Cette solution permet de fine-tuner des embeddings pour un domaine spécifique grâce à des KG spécialisés (ici médecine)

Facile à dériver si la fonction de distance est additionnelle (TransE) mais pas multiplicative (DisMult) et encore moins si elle implique des inverses

Et encore un autre apparaît... RA-Retrofit

$$\mathcal{L}_1(u,u',v,v',r) = \mathcal{W}(u,r,v)\max\{m+\phi(u,r,v)-\phi(u',r,v'),0\},$$

Trop d'information en une step

$$\mathcal{L}(u,u',v,v',r) = \mathcal{L}_1(u,u',v,v',r) + \mu \sum_{s\in\{u,u',v,v'\}} \mathcal{L}_2(s) + \eta_1 \parallel \Theta_1 \parallel + \eta_2 \parallel \Theta_2 \parallel,$$

Pas adapté pour réseau de neurones

- Ne présente pas les vraies et les fausses paires ensemble
- Trouver comment gérer l'apprentissage des relation embeddings
  - Désolidariser l'entraînement?
- Tester plusieurs

Fang, L., Luo, Y., Feng, K., Zhao, K., & Hu, A. (2019, May). Knowledge-Enhanced Ensemble Learning for Word Embeddings. In *The World Wide Web Conference* (pp. 427-437).

Conclusion

- Possibilité d'améliorations des systèmes actuels
  - Fusionner les idées, les domaines
  - Les rendre robustes et scalables
- Pas de code propre et réutilisable
  - Réimplémenter les algos existants
  - Les transformer en problèmes plus neural network
  - Travailler sur la représentation de relation
- Utilisations pour le transfert de représentation pour des langues sous représentées
- Adaptation avec les word embeddings contextuels

4. Éléments connexes

A. Negative Sampling
B. Poids des relations
C. Pré-traitement

a. Poids des relations

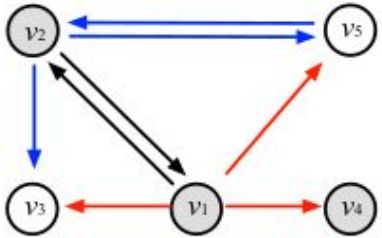
ConceptNet:

- Si on retrouve plusieurs fois un triple, posté par plusieurs personne différente ou dans plusieurs graphes de connaissances, on lui met un poids élevé.

RA-Retrofit :

- plus deux mots ont des voisins en communs, plus le poids de leur relation.
- Exemple (Internet, RelatedTo, Computer) et (Multipart, RelatedTo, Computer)
- s(v1, r, v2) > s(v1, r, v4)

$$\mathcal{W}(u, r, v) = \epsilon_1 + \frac{|u^N \cap v^N|}{|u^N \cup v^N| + \epsilon_2},$$



b. Negative sampling

- Si (i,r,j), générer (i,r, j')

$$P((u,r,v)) = 1 - \sqrt{\frac{s}{g(u,v)}},$$

RA-Retrofit :

- Pour le negative sampling, prendre deux mots qui ne sont pas reliés avec une forte probabilité

Explicit-Retrofitting :

- Rapproche les synonymes, éloigne les antonymes, s'assure que les "negative examples" gardent la même distance
- Les examples négatifs se font en regardant les PPV de i.

Quelle proportion de negative sampling?

c. Pré-traitement des données

- Faire correspondre le vocabulaire d'un modèle pré-entrainé et d'un KG
- Et de plusieurs modèles pré-trainés + plusieurs KG
- Ambiguïté des termes
- Lemmatiser?
- Le sring matching est un peu archaïque, peut entraîner des erreurs
- Peut être que si les méthodes marchent bien, on peut voir comment augmenter les données

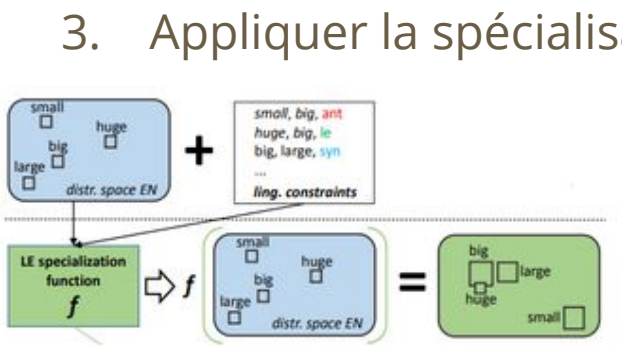
Contenu supplémentaire

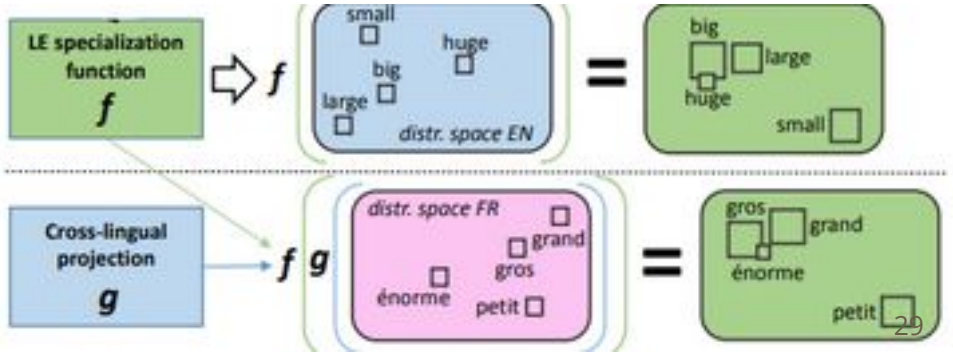
Transfert de connaissance entre langues

Un avantage d'utiliser des lexiques sémantiques, c'est qu'on peut aussi faciliter le transfert d'information d'une langue à une autre.

Langue source : X, langue visée : Y

1. Aligner les vocabulaires de X et Y
2. Spécialisation sémantique de X grâce à fonction f
3. Appliquer la spécialisation pour Y





Évaluation des word embeddings - ANNEXE???

- Souvent similarité et relatedness ensemble (WS, MEN, ...)
- Mais certains corpus isolent une relation (SimLex, TOEFL synonym question, GRE antonym)
- Analogy (Google Analogy)
- Lexical Entailment : Hyperlex