ADDICTE :
Analyse distributionnelle en domaine de spécialité

Cécile Fabre, Nabil Hathout, Ludovic Tanguy

CLLE/ERSS, CNRS & UT2J

Séminaire CARTEL / Master LITL
19 novembre 2018

Vue d'ensemble

1. **Le projet ADDICTE : sémantique distributionnelle en domaine spécialisé**
2. Expérimentations en corpus spécialisé : la tâche exploratoire de SemDis
3. Compositionnalité des représentations vectorielles des termes

Le projet ADDICTE

- ADDICTE = *Analyse distributionnelle en domaine de spécialité*
- Projet financé par l'Agence Nationale de la Recherche (ANR).
  Avril 2018 – Septembre 2021
- Partenaires = CEA, CLLE, LIMSI, LS2N.
  Porteur = Emmanuel Morin (LS2N)
- Les objectifs du projet :
  - proposer une solution opérationnelle à l'analyse sémantique distributionnelle en domaine de spécialité ;
  - construire des représentations distributionnelles (*embeddings*) de lexiques spécialisés exploitables dans des tâches du TAL
  - et en particulier, pour construire des représentations sémantico-conceptuelles du domaine (ontologies de domaine, thésaurus, ressources terminologiques) utilisables ;
- Domaines choisis : médecine, traitement automatique des langues

Caractéristiques des langues de spécialité

- sous-langages (sublanguages), langues de spécialité, langues spécialisées
- Définition : « Nous rattachons à l'ensemble « langue de spécialité » toute production langagière réalisée par un spécialiste dans un milieu professionnel, au sujet de sa spécialité » Humbley (1994)
- Des propriétés favorables à l'analyse distributionnelle : « A sublanguage is caracterized by distinctive specialisation of syntax and the occurrence of domain-specific word subclasses in particular syntactic combinations. » Sager (1986)
- Redondance, lexique limité, schémas de cooccurrences contraints, restrictions de sélection tranchées, moins d'ambiguïté

Exemple (Bourigault 2004)

- Comparaison des arguments d'un verbe dans un corpus de langue générale et de langue spécialisée

Sujets de *montrer*		Objets de *montrer*	
Le Monde	**Réa. chirurgicale**	**Le Monde**	**Réa. chirurgicale**
étude	scanner cérébral	exemple	épanchement
enquête	radiographie	limite	hématome
sondage	bilan	signe	persistance
expérience	échographie cardiaque	image	absence
film	ECG	voie	contusion
résultat	doppler	volonté	présence
image	fibroscopie	visage	fracture

Réanimation chirurgicale : EXAMEN montrer PROBLEME

Les sous-langages, premier terrain d'expérimentation de l'analyse distributionnelle

- (Harris, 1954, 1991) : mettre au jour les classes de concepts et les relations d'un sous-langage lié à un domaine d'activité (ex : articles scientifiques dans le domaine de l'immunologie) = *semantic grammar*. (Habert and Zweigenbaum, 2002)
- Hypothèse que seuls les corpus de domaines restreints peuvent garantir la définition de classes sémantiques bien circonscrites.
- Premiers travaux visent la mise au jour de thesaurus de domaines : (Grefenstette, 1994; Habert and Fabre, 1999; Bouaud et al., 1997)
- Automatisation de la procédure harrissienne : corpus annotés, factorisation et sélection des contextes

Une deuxième direction : l'analyse distributionnelle de grands corpus généralistes

- Hypothèse cognitive (Lenci, 2008; Miller and Charles, 1991; McDonald, 1997) :
  - "How words are used within language itself, and not just how they are used to denote perceptible objects and events in the world, may be a rich source of information by which humans, whether as children or adults, learn and develop their semantic knowledge" (Vinson et al., 2014)
  - Des modèles distributionnels développés pour la recherche en cognition et en psychologie (LSA (Landauer and Dumais, 1997)) : idée de plausibilité cognitive
  - Mesures de similarité contextuelle corrélées aux jugements de similarité sémantique de sujets humains
- Priorité au traitement de grands corpus de données non spécifiques
  - Maximiser le volume pour réduire la dispersion (*data sparseness*)
  - Jeux de données génériques

Revenir aux domaines de spécialité

- Les embeddings "généraux" construits à partir de grands corpus ne captent pas le sens des termes des domaines de spécialité
- Des besoins applicatifs importants :
  - Construction de bases de connaissances
  - Recherche de documents, extraction d'information (ex : chatbots)
  - Besoins spécifiques : fouille de textes pour l'exploitation des retours d'expérience, classification de gènes, extraction de connaissances à partir de la littérature scientifique, etc. Cohen and Widdows (2009).
- Des caractéristiques particulières à prendre en compte :
  - Taille des corpus spécialisés : efficacité des méthodes distributionnelles corrélée à la quantité de données
  - Les unités de sens sont des termes : généralement des unités polylexicales
    - ex : *analyseur syntaxique en dépendances*, *insuffisance cardiaque congestive*
    - Aggrave la dispersion des données
  - Des modalités d'évaluation à définir

ADDICTE : Améliorer les représentations distributionnelles issues de corpus de spécialité

- Amélioration **exogène** des contextes distributionnels
  - Compléter l'information disponible dans les corpus spécialisés par l'apport de ressources externes
  - Corpus de spécialité proches, de discours vulgarisé, de langue générale
  - Ressources terminologiques et ontologies spécifiques

ADDICTE : Améliorer les représentations distributionnelles issues de corpus de spécialité

- Amélioration **endogène** des contextes distributionnels
  - Enrichir l'analyse linguistique pour compenser la rareté informationnelle
    - Exploiter les propriétés structurelles des textes
    - Enrichir les contextes par des informations de nature syntaxique
  - Exploiter la compositionnalité des termes complexes
    - Combiner l'information des composants pour inférer celle du terme (morphologiquement ou syntaxiquement) complexe
    - Généraliser l'information distributionnelle d'un des composants (ex : la tête du terme)

Vue d'ensemble

1. Le projet ADDICTE : sémantique distributionnelle en domaine spécialisé
2. **Expérimentations en corpus spécialisé : la tâche exploratoire de SemDis**
3. Compositionnalité des représentations vectorielles des termes

Semdis : un atelier sur la sémantique distributionnelle en français

Première édition en 2013

Dans le cadre de TALN, un atelier à large couverture.
Grande variété de méthodes, d'objectifs et de données expérimentales.
Table ronde : volonté de proposer des activités structurantes (tâches collaboratives)

Seconde édition en 2014

- Première tâche : substitution lexicale en français
- Seconde tâche : tâche exploratoire sur le corpus TALN

Tâche exploratoire : consignes

Cette deuxième tâche est une tâche exploratoire qui permettra d'examiner plus en détail les résultats de méthodes distributionnelles sur un ***corpus spécialisé de petite taille****. Pour cela, nous proposons aux participants d'utiliser un corpus commun : il s'agit d'un corpus constitué d'une sélection d'articles en français issus des conférences TALN et RECITAL[...].*
Nous invitons les participants à déployer une ou plusieurs techniques d'analyse distributionnelle sur ce corpus, avec les prétraitements et annotations de leur choix. Chacun pourra analyser ce corpus selon ses objectifs propres, et étudier les phénomènes sémantiques qui lui paraissent les plus pertinents (mise au jour de la polysémie, d'une organisation terminologique, étude de relations sémantiques spécifiques, compositionnalité, etc.). Nous demandons, pour illustrer la démarche et les résultats, de privilégier un ensemble de mots que nous avons sélectionnés dans le but de faciliter les échanges.

Trois contributions avec des techniques et des angles différents.

Tâche exploratoire : mots-cibles

Les mots que nous avons sélectionnés sont les suivants (fréquence dans le corpus de 2 millions de mots) :
1 verbe : **calculer** (1235), 2 adjectifs : **complexe** (766), **précis** (376),
5 noms : **fréquence** (647), **graphe** (1116), **méthode** (3808), **sémantique** (413), **trait** (1806)
Les critères retenus sont :

- une fréquence minimale pour permettre de déployer confortablement des méthodes distributionnelles classiques ;
- un lien clair avec le domaine du corpus (le TAL) ;
- un potentiel (intuitif) à illustrer un panel de phénomènes sémantiques ; certains mots sont très spécifiques (graphe), d'autres ont a priori de nombreux synonymes (méthode), d'autres sont polysémiques (trait, précis), certains ont des acceptions particulières dans ce domaine par rapport à un discours plus général (trait, fréquence), quant à sémantique, il est notoirement difficile à cerner.

Voisinage distributionnel de ces mots dans le corpus et dans FRWAC (Word2vec CBOW)

TALN

Trait : propriété, variable, attribut, contrainte, tuple...
Graphe : treillis, sous-graphe, diagramme, marche, réseau...
Fréquence : productivité, apparition, cooccurrent, effectif, proportion...
Sémantique : syntaxe, dualité, représentation, phonologie
Précis : souple, délicat, approximatif, ambigu, ardu...

FRWAC

Trait : contour, hachure, tracer, zébrure, pointillé...
Graphe : diagramme, topologique, algorithme, polynôme, histogramme...
Fréquence : Hz, kHz, MHz, impédance, oscillateur...
Sémantique : syntaxe, ontologie, grammaire, linguistique...
Précis : concis, détaillé, rigoureux, exact, explicite...

Problématiques : paramétrage et évaluation

TALN avec modèle 1

Trait : propriété, variable, attribut, contrainte, tuple...
Graphe : treillis, sous-graphe, diagramme, marche, réseau...
Fréquence : productivité, apparition, cooccurrent, effectif, proportion...
Sémantique : syntaxe, dualité, représentation, phonologie
Précis : souple, délicat, approximatif, ambigu, ardu...

TALN avec modèle 2

Trait : propriété, caractéristique, ressource, contrainte, unité
Graphe : arbre, base, liste, lexique, représentation
Fréquence : probabilité, similarité, précision, contexte, valeur, score
Sémantique : finance, ambiguïté, signification, définition, polysémie
Précis : détaillé, adéquat, correct, fiable, opérationnel, général

Grande variation dans les résultats : comment sélectionner la meilleure méthode ?

Vue d'ensemble (Fabre et al., 2014; Tanguy et al., 2015)

Objectif

Comparer les différentes stratégies pour construire un modèle distributionnel sur un corpus spécialisé
Identifier les paramètres critiques et le "meilleur" modèle

Méthode

- Définition d'un jeu de test (extension des mots-cibles précédents)
- Sélection d'un ensemble de modèles distributionnels en faisant varier les modes de calcul et les paramètres
- *Pooling Method* (étape 1) : calcul des n plus proches voisins distributionnels des mots cibles
- *Pooling Method* (étape 2) : annotation manuelle de ces seuls voisins pour construire un gold standard
- Évaluation des performances de chaque modèle

Terrain de jeu : extension du jeu de test

Limites des 5 mots proposés

- pas équilibrés en termes de catégorie (une seul verbe)
- peu variés en termes de fréquence
- pas assez nombreux pour garantir la couverture de notre évaluation

30 mots au total

- 10 adjectifs : *sémantique important complexe temporel correct précis spécialisé significatif empirique computationnel*
- 10 noms : *méthode trait élément performance graphe fréquence contrainte sémantique dépendant signification*
- 10 verbes : *décrire évaluer extraire calculer annoter valider caractériser conduire indexer apparier*

Annotation

Tâche

Pensez-vous que ces deux mots sont liés sémantiquement dans le domaine du TAL ?
Oui/Non

Accord

Kappa de Fleiss moyen : 0.55
Plus haut pour les adjectifs, plus bas pour les verbes

Regroupement des réponses

Pour chaque paire, le nombre d'annotateurs l'ayant validée.
Exemple :
contrainte : *condition* (4), *règle* (4), *critère* (2), *inégalité* (1)

Quelques observations

- Diversité des relations sémantiques :
  - Nombreux synonymes (*complexe/compliqué*, *computationnel/algorithmique*)
  - antonymes (*correct/erroné*, *empirique/théorique*)
  - autres (*sémantique/syntaxique*, *spécialisé/biomédical*).
- Nombreux termes du domaines
  fréquence/TF, *méthode/algorithme*, *performance/f-score*
- Nombreuses relations propres au domaine
  fréquence/probabilité, *graphe/matrice dépendant/gouverneur*.
- Des mots difficiles à annoter: *indexer* : peu de voisins à 4 points (*classifier*, *référencer*, *lister*)
- Le Kappa n'est que faiblement corrélé à la fréquence des mots

Mesure d'évaluation

Contraintes

- Chaque mot est évalué par un score (0 à 4)
- Limitation (arbitraire) aux 50 plus proches voisins de chaque mot-cible
- Prise en compte du rang

Normalized Discounted Cumulative Gain (nDCG)

Järvelin and Kekäläinen (2002)

$$DCG = \sum_{i=1}^{50} annot_i / log_2(i+1)$$

Puis normalisée par la plus haute valeur possible pour chaque cas. Le score nDCG est maximal quand les voisins les plus pertinents (mieux notés) apparaissent dans les premiers rangs, sans bruit intersticiel.

Terrain de jeu : méthodes distributionnelles envisagées

(Méthodes fréquentielles uniquement : avant la vague Word2vec)

Principaux paramètres considérés

Contexte d'un mot : cooccurrence ou relation syntaxique ?

> La *structure* que nous appelons prédicative (qu'on aurait pu aussi appeler argumentale) est un **graphe** de *relations* prédicat-argument

Pas de supériorité claire relevée dans des expériences sur des grands corpus et des jeux d'évaluation standard (Kiela and Clark, 2014), (Padó and Lapata, 2007), (Peirsman et al., 2007)...
Mais plusieurs questions subsistent :

- Quid des petits corpus spécialisés ?
- Comment utiliser au mieux les informations syntaxiques ?

Variations des modèles

Contextes à base de fenêtre

- Taille de la fenêtre (distance de 1, 3 ou 5 tokens)
- Direction de la fenêtre (gauche, droite, les deux)

Contextes syntaxique

- Quelles relations à prendre en compte (suj, obj, mod, etc.)
- Spécifier la relation dans la définition du contexte ou pas
- Résoudre les configurations complexes ou pas (préposition, coordination, passif, etc.)
- Au final, 4 niveaux de sophistication

Variations des modèles

Autres paramètres (deux types de modèles)

- Sélection des contextes : seuil de fréquence, productivité
- Mesure de la similarité entre mot et contexte (Information mutuelle, t-score, z-score, log-likelihood)
- Mesure de la similarité entre vecteurs (cosinus, jaccard)

Au total

- 2300 modèles différents
- Pour chacun des 30 mots-cibles, le score NCDG de chaque modèle

Meilleures configurations

Scores moyens

	Adjectifs		Noms	
Contextes	**Config**	**NDCG**	**Config**	**NDCG**
Fenêtre	Win1LR_10-zscore-log-0	0,539	Win3LR_60-ll-log-5	0.615
Syntaxe	Synt3_4-MI-none-0	**0,558**	Synt3_5-ll-root-0	**0.666**

	Verbes		Total	
Contextes	**Config**	**NDCG**	**Config**	**NDCG**
Fenêtre	Win3LR_30-ll-log-10	**0,554**	Win3LR_30-ll-log-0	0,559
Syntaxe	Synt3_2-ll-root-0	0,525	Synt4_4-ll-root-0	**0,561**

- Léger avantage aux modèles syntaxiques, non significatif (Wilcoxon paired test, $p>0.05$).
- Noms > ajdectifs > verbes

Analyse globale

Calcul des facteurs les plus déterminants

Principal : fréquence du mot-cible (scores plus élevés pour les mots fréquents, ρ=0.6)
Autres facteurs:

- type de contexte (dependency > window *en moyenne*),
- mesure d'association (quelques mesures très mauvaises)
- catégorie du mot-cible (noms > adjectifs > verbes)

Zoom sur les modèles syntaxiques

Calcul des facteurs les plus déterminants

Principal : niveau de normalisation : $3 > 4 > 2 > 1$
(limite : coordination et passif OK, mais pas les relations indirectes complexes ou les relations à 3 mots)
Autres facteurs : identiques aux tendances globales

Quelques différences significatives pour les relations syntaxiques

- *Modifieur de nom* est vital pour les adjectifs, et très important pour les noms
- *Objet* est vital pour les verbes, pas pour les noms
- supprimer *sujet* ne diminue pas significativement la performance
- *Prep* est important pour les noms et les verbes
- *Adverbe* est utile pour les adjectifs, inutile pour les verbes
- toutes les autres sophistications n'ont pas d'effet mesurable

Analyse qualitative 1 - contextes mieux traités par les modèles syntaxiques

Tête des syntagmes nominaux

- Les grands SN sujets ont leur tête loin du verbe
- Example: constructions avec des quantifieurs (moitié, totalité, ensemble, partie, etc.), qui sont informatifs pour des verbes comme *valider*, *annoter*, etc.:
  - *La **moitié** d'entre elles ont ensuite été **validées** manuellement.*
  - *la **moitié** des entrées de S L X sont **jugées** incorrectes.*

Coordinations

- Les coordinations éloignent les mots hors de la fenêtre
- Exemple :
  - *pour **annoter** les expressions référentielles (les markables) **et** les **relations***
  - *la capacité qu'a le système cognitif de **sélectionner** les unités **et** les **relations***

Analyse qualitative 2 - contextes mieux traités par les modèles à fenêtre

Relations syntaxiques indirectes

- Exemple 1 : adjectifs pertinents modifiant les arguments du verbe
  - *Une question est donc construite pour **valider** la réponse **candidate**.*
  - *Plus un terme **candidat** est **jugé** important ...*
- Example 2 : expansions de SN parfois plus pertinentes que les têtes
  - *Sinon nous **sélectionnons** un **échantillon** de **documents** que nous analysons à la main.*
  - *Un de nos objectifs est d'**annoter** un **ensemble** de **documents***

Contextes filtrés

Exemple : pronoms personnels sujets

- ***Nous présentons** dans cet article une nouvelle manière d'aborder le problème*
- *Dans cette section, **nous** allons **décrire** les deux principales approches*

Retour au projet ADDICTE

Contextes

- Les contextes syntaxiques semblent un léger atout, notamment pour pallier aux faibles fréquences
- L'exploitation des informations syntaxiques doivent être faite avec attention.
- Marge de progression importante en termes d'exploitation du contexte

Pistes

- Besoin d'explorer les contextes lexico-syntaxiques des occurrences, et d'observer quelles caractéristiques influencent les représentations.
- Explorer les caractéristiques structurelles des textes (parties des articles, titres, bibliographie, etc.) avec des structures et emplois potentiellement différents.

Limites de l'approche (1)

Type de relation sémantique

Nous n'avons pas distingué les relations sémantiques entre les mots (synonymie, antonymie, spécifique/générique, morphologie, etc.)
Or, celles-ci sont très importantes dans les ressources teminologiques
Comment ajuster le paramétrage en fonction des relations visées ?

Exemple : (Bernier-Colborne and Drouin, 2016)

- Corpus et terminologie riche du domaine de l'environnement, comparaison de plusieurs modèles en fonction des relations visées
  - 2 groupes de relations : les relations de quasi-synonymie, antonymie et hypo/hypéronymie s'établissent entre des termes qui se trouvent dans des configurations syntaxiques similaires
  - la relation de dérivation s'établit entre des termes qui cooccurrent avec les mêmes mots dans des contextes syntaxiques variés, de grande taille

Limites de l'approche (2)

Termes complexes

Nous n'avons utilisé que des termes simples.
Quid de la représentation dans les modèles des composants de termes complexes ?

Exemples

- **point** : vue, repère, embarras, plan, blogosphère
- **base** : projection, réécriture, création, liste
- **référence** : entraînement, test, mention,
- **moteur** : portail, logiciel, prototype, outil
- **tour** : message, parole, client, émission, enregistrement

Vue d'ensemble

1. Le projet ADDICTE : sémantique distributionnelle en domaine spécialisé
2. Expérimentations en corpus spécialisé : la tâche exploratoire de SemDis
3. **Compositionnalité des représentations vectorielles des termes**

Compositionnalité et langues de spécialité

La compositionnalité joue un rôle important dans la sémantique des langues de spécialité :

- La plupart des termes sont complexes
  - sur le plan syntaxique (multi-termes)
    système d'exploitation
  - sur le plan morphologique (compositions savantes)
    angioblaste = vaisseau + germe
- Les terminologies sont des collections de termes structurées par différentes relations : hyperonymie / hyponymie, relations d'équivalence, relations d'association
  - ***angioblaste*** est un hyponyme de ***cellule souche***
  - ***curage axillaire*** est un équivalent de ***curage ganglionnaire axillaire***

Compositionnalité et langues de spécialité (2)

Contrainte

Les représentations sémantiques doivent être compatibles

- avec la structure linguistique des termes
- avec les relations qui structurent les terminologies

Besoin

Acquérir des termes et des relations terminologiques à partir d'espaces sémantiques distributionnels construits à partir de corpus spécialisés

Opportunité

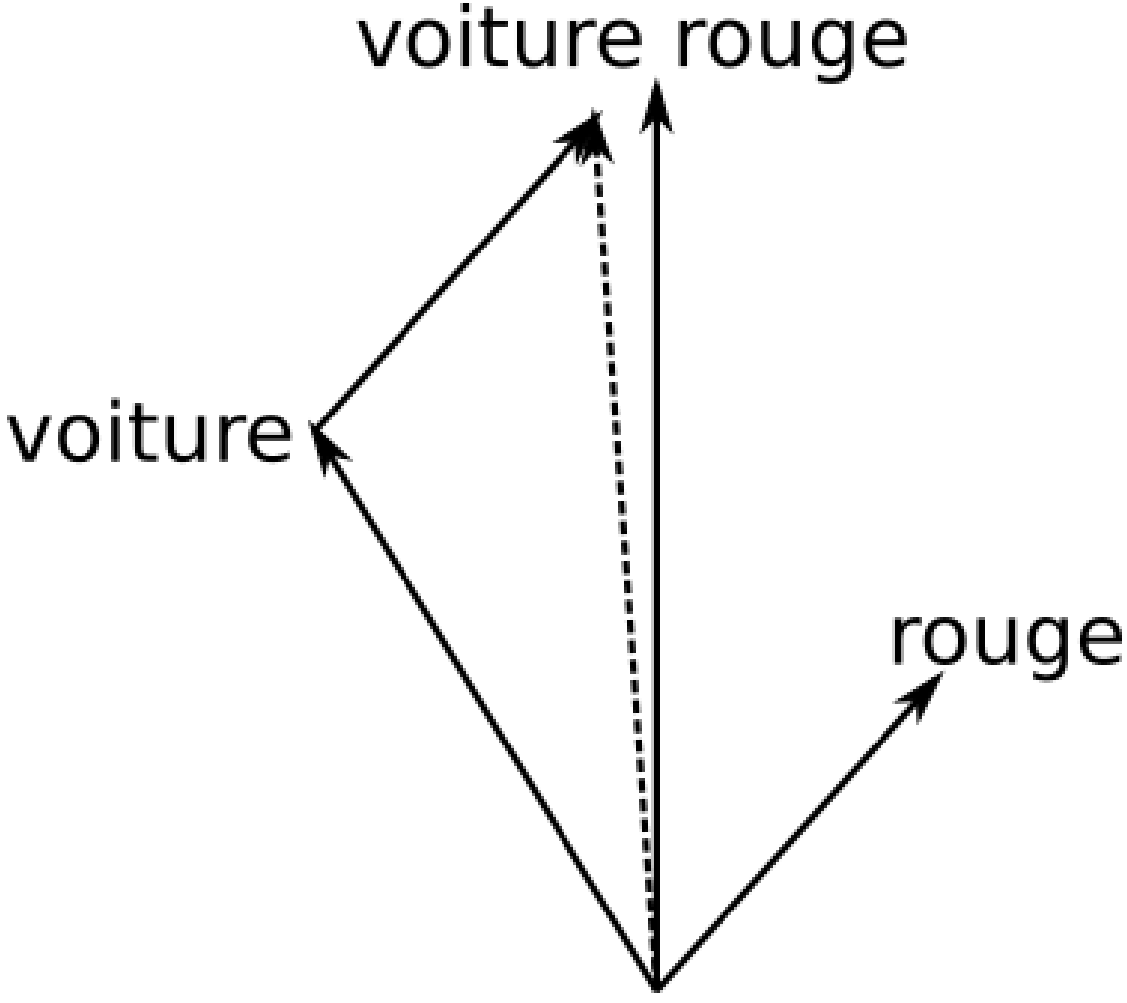
À partir d'un espace sémantique distributionnel construit à partir d'un corpus spécialisé et d'une terminologie,
utiliser la structure des termes et les relations terminologiques pour construire des représentations pour les termes qui n'en ont pas.

Modèles additifs

Comment représenter le sens d'un terme complexe ?

- La représentation de ***voiture rouge*** est définie à partir de toutes les occurrences de ce SN dans le corpus
- Le sens du SN est compositionnel :
  une entité est une ***voiture rouge*** si
  1. l'entité est une ***voiture*** et
  2. l'entité est ***rouge***
- La représentation de ***voiture*** dans ***voiture rouge*** est la même que celle de ***voiture*** dans les autres contextes. Idem pour ***rouge***.
- La représentation distributionnelle sont des vecteurs sur lesquels il possible de réaliser des additions.
- La représentation de ***voiture rouge*** peut être calculée en additionnant les représentations de ***voiture*** et de ***rouge***
- Le sens de $A \wedge B$ est représenté par le vecteur $v_A + v_B$

Modèles additifs



Modèles additifs

petit + corpus (corpus TALN)

sous-corpus, **échantillon**, vocabulaire, grand, bitexte, volume, ...

Modèles additifs

petit + corpus (corpus TALN)

sous-corpus, **échantillon**, vocabulaire, grand, bitexte, volume, ...

Effet d'optique

- voisins de ***corpus*** = **sous-corpus**, **échantillon**, donnée, bitexte, lexique, jeu, site, dictionnaire
- voisins de ***petit*** = réduit, court, élevé, restreint, grand, limité, représentatif, homogène

Modèles additifs

système + exploitation

enrichissement, application, indexation, outil, intégration, acquisition, apport, élaboration, filtrage, évaluation

système

programme, **module**, **processus**, modèle, classifieur, **logiciel**, **prototype**, correcteur, décodage, sqr

exploitation

enrichissement, exploration, **intégration**, **indexation**, acquisition, élaboration, **utilisation**, extraction, identification, **adaptation**

Modèles additifs

partie + discours

proposition, définition, corps, citation, style, structure, repris, ftb, thème, description

partie

ligne, version, définition, **majorité**, thématique, famille, proposition, norme, **taille**, **fenêtre**

discours

citation, corps, **connecteur**, **discursif**, **style**, **ftb**, **marqueur**, **rhétorique**, **thème**, **énoncé**

Modèles additifs

Beaucoup de relations sémantiques lexicales ne sont pas préservées par word2vec

traitement $\leftrightarrow$ outil (classification : classifieur = étiquetage : ?)

identification, détection, analyse, segmentation, reconnaissance, catégorisation, correction, acquisition, sélection, simplification

➜ ***classification*** et ***étiquetage*** sont situés dans des zones denses de noms de traitements

➜ ***classifieur*** ou ***étiqueteur*** ne permettent un « déplacement » de ces représentations suffisant pour la relation.

Densité dans les espaces vectoriels distributionnels

- Le modèle additif dans les embeddings word2vec est sensible aux variations dans la densité des espaces vectoriels distributionnels

➜ Problème de robustesse et de fiabilité

Modèles multiplicatifs

- Modèles inspirés des grammaires catégorielles
- ***voiture rouge***
  - ***voiture*** représente une entité
  - ***rouge*** est un opérateur : $X \rightarrow X$ rouge

- ***rouge*** est représenté par une matrice R
  qui transforme le vecteur v qui représente ***voiture***
  en un vecteur v' qui représente ***voiture rouge***.
- La transformation est une multiplication

$$v' \approx Rv$$

- on utilise une approximation de R :
  - déterminer les représentations v'_i de tous les ***X rouge*** du corpus d'apprentissage
  - déterminer les représentations v_i de tous les ***X*** correspondants
  - calculer $\hat{R}$ par la méthode des moindres carrés (régression linéaire) entre les vecteurs prédits $\hat{v'_i}$ et les vecteurs effectifs v'_i.

Réseaux de neurones

- On entraîne un réseau de neurones (récurrents RNN, convolutionnels CNN, etc.)
  - les entrées sont les composants
    petit chat, ***soldat à pied***
  - les sorties sont les représentations des composés
    chaton, ***fantassin***
- Les composants peuvent aussi être des syntagmes produits par un analyseur syntaxique
  - les entrées sont les constituants des syntagmes
  - les sorties sont les représentations des syntagmes considérés comme des unités

Composition morphologique : Thèse de Marine Wauquier

- Représentation des constructions morphologiques par les barycentres des mots construits par le procédé
- Représentation du sens d'un affixe par la moyenne des différences entre les représentations des dérivés et de leurs bases.

Composition morphologique : fastText

- Modèle des représentations des mots et des n-grammes de caractères qu'ils contiennent

Modèle additif

La représentation d'un mot est obtenu en additionnant le vecteur de la forme complète avec les vecteurs de tous les n-grammes qu'il contient.

Objectif

Prédire de meilleures représentations pour les mots inconnus en se basant sur les représentations des n-grammes de caractères qu'ils contiennent.

Composition morphologique : fastText (2)

- Les représentations des mots longs (***morphosyntaxique***) sont « meilleures » que celles des mots courts (***lemme***)

lemme

femme, lexème, stemme, flemm, morphème, cbme, gamme, chiasme, alarme, axiome, phonème, arme, tlfnome, légume, tokens

morphosyntaxique

morpho-syntaxique, macro-syntaxique, micro-syntaxique, syntaxique, morphosyntaxe, morpho-syntaxe, lexico-syntaxique, syntaxiques, morphosyntaxiquement, lexico-syntaxiques, morpho-syntaxiquement, morphologique, morpho-phonologique

- La similarité formelle l'emporte sur la similarité sémantique
  - ***catégorie***, ***grammatical*** sont absents des premiers voisins de ***morphosyntaxique***

Thèse Yizhe Wang

- *Analyse sémantique et son exploitation en domaines de spécialité*.
- Étude et exploitation de la compositionnalité morphologique et syntaxique dans les espaces distributionnels construits à partir de corpus spécialisés.
- Direction : Béatrice Daille (LS2N) et Nabil Hathout (CLLE).

Objectifs

- Exploiter les espaces vectoriels sémantiques pour améliorer l'extraction terminologiques à partir de corpus spécialisés
- Identifier les relations qui existent entre les termes extraits par extracteurs terminologiques. Les extracteurs proposent des listes de termes généralement non structurées.

Thèse Yizhe Wang (2)

Termes TALN Yatea

analyse syntaxique, point de vue, traduction automatique, corpus de test, Association for, traitement automatique, Traitement Automatique, In Actes, état de l' art, noms propres, analyseur syntaxique, base de données, nombre de mots, relations sémantiques, unités lexicales, corpus de référence, Thèse de doctorat, modèle de langage, moteur de recherche, structures de traits, représentation sémantique, reconnaissance de la parole, arbre de dérivation, ressources lexicales, mesure de similarité, langue arabe, meilleurs résultats, paires de phrases, ressources linguistiques, Université Paris, ...

In Actes n'est pas un terme
corpus de référence est-un ***ressource linguistique***

Références

Bernier-Colborne, G. and P. Drouin (2016). Évaluation des modèles sémantiques distributionnels : le cas de la dérivation syntaxique. In *Actes de la 23e conférence sur le traitement automatique des langues naturelles (TALN-2016)*, Paris, pp. 125–138.

Bouaud, J., B. Habert, A. Nazarenko, and P. Zweigenbaum (1997). Regroupements issus de dépendances syntaxiques en corpus: catégorisation et confrontation à deux modélisations conceptuelles. In *Actes de 1re journées Ingénierie des Connaissances*, Roskoff, pp. 207–223.

Cohen, T. and D. Widdows (2009). Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics 42*(2), 390 – 405.

Fabre, C., N. Hathout, F. Sajous, and L. Tanguy (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France, pp. 266–279.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Références (2)

Habert, B. and C. Fabre (1999). Elementary dependency trees for identifying corpus-specific semantic classes. *Computers and the Humanities 33*(3), 207–219.

Habert, B. and P. Zweigenbaum (2002). Contextual aquisition of information categories. what has been done and what can be done automatically? In B. Nevin and S. Johnson (Eds.), *The legacy of Zellig Harris. Language and Information into the 21st century*, Volume 2: computability of language and computer applications, Chapter 8, pp. 203–231. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Harris, Z. (1954). Distributional structure. *Word 10*(2-3), 146–162. Traduction française dans *Langages* (20) 1970.

Harris, Z. S. (1991). *A theory of language and information: a mathematical approach*. Oxford: Clarendon Press.

Humbley, J. (1994). Oralisation de sigles en aéronautique. *Linx 30*(1), 133–152.

Références (3)

Järvelin, K. and J. Kekäläinen (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS) 20*(4), 422–446.

Kiela, D. and S. Clark (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pp. 21–30. Association for Computational Linguistics.

Landauer, T. K. and S. T. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review 104*(2), 211.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics 20*(1), 1–31.

McDonald, S. (1997). A context-based model of semantic similarity. *ACM Transactions on Programming Languages and Systems (TOPLAS) 15*(5), 795–825.

Références (4)

Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and cognitive processes 6*(1), 1–28.

Padó, S. and M. Lapata (2007). Dependency-based construction of semantic space models. *Computational Linguistics 33*(2), 161–199.

Peirsman, Y., K. Heylen, and D. Speelman (2007). Finding semantically related words in dutch. co-occurrences versus syntactic contexts. In *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Documents.*, Roskilde, Denmark, pp. 9–16.

Sager, N. (1986). Sublanguage: Linguistic phenomenon, computational tool. *Analyzing language in restricted domains: sublanguage description and processing*, 1–17.

Tanguy, L., F. Sajous, and N. Hathout (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *Traitement Automatique des Langues 56*(2).

Vinson, D., M. Andrews, and G. Vigliocco (2014). Giving words meaning: why better models of semantics are needed in language production research. *The Oxford Handbook of Language Production*, 134.