Analyse sémantique automatique

Pascal Amsili / Marie Candito
M2 Linguistique Informatique
Paris Diderot
13-14

1

Contenu

- **Sémantique lexicale (en bref)**
  - unité lexicale
  - signifiants et ambiguïté
  - relations lexicales
  - représentation du sens d'une UL
  - rôles sémantiques
- **Sémantique lexicale computationnelle**
  - ressources pour l'anglais
    - réseau lexical : Wordnet
    - classes sémantiques : VerbNet
    - rôles sémantiques : FrameNet, PropBank
  - pour le français?
  - tâches
    - Calcul de similarité lexicale
    - **WSD** : désambiguisation lexicale (word sense disambiguation)
    - **SRL** : étiquetage de rôles sémantiques (semantic role labeling)
    - **RTE** : Reconnaissance d'inférence textuelle (recognizing textual entailment)
    - Lexical Acquisition
      - acquisition de cadres de sous-catégorisation
      - acquisition de préférences lexicales
      - reconnaissance de composés (multi-word units)

2

Sources

- Boleda et Evert :
  - cours ESSLLI 2009 computational lexical semantics
- Diana MacCarthy (Univ Melbourne)
  - intro Comp Lexical Semantics
  - http://lct-master.org/index.php?id=teaching_material
- Jurafsky & Martin
  - chapitres 19 et 20
- cours Pascal Denis

3

Retour sur la sém. formelle

- vous avez vu comment calculer et interpréter des représentations sémantiques

  (1) *A man bought a donkey to John*

  (2) ∃e, x, y, t [man′(x) ∧ donkey′(y) ∧ buy′(e, t, x, y, John) ∧ t < now]

- Mais à partir de (1) nous sommes capables de répondre à :

  *Does the man **own** a donkey?*

  Did John receive money from the man?

  *Who **sold** the donkey to the man?*

4

Ce qu'il manque

- Pour répondre à ces questions à partir de (2), il nous faut
  - les correspondances entre vocabulaire linguistique (*to buy, to sell, donkey*) et vocabulaire des représentations sémantiques
    - a-t-on besoin de buy' et sell' tels que

      buy′(e1, t1, x, y, z) ⟷ sell'(e1, t1, z, y, x)

      ou bien utilise-t-on un seul prédicat?
  - le sens de man', donkey' , buy'...
  - ou en tous cas les inférences possibles

    buy′(e1, t1, x, y, z) → own'(e2, t2, x, y)

=> = sémantique lexicale

5

Domaines de la sémantique

- découpage traditionnel (aux frontières parfois floues):
- **sémantique lexicale**
  - le sens des **unités lexicales (=mots)**
- **sémantique compositionnelle**
  - ou sémantique de la phrase / clause / énoncé :
  - comment le sens des mots se combine pour une phrase donnée
- **sémantique du discours**
  - comment se combine le sens des phrases / clauses
  - rem: est également compositionnelle...
- **pragmatique**
  - comment la connaissance du monde intervient dans l'interprétation
    - d'une phrase
    - d'un discours (texte)
    - d'un dialogue

6

Sémantique lexicale

- unité lexicale
- signifiants et ambiguïté
- relations lexicales
- représentation du sens d'une UL
- rôles sémantiques

7

Unité lexicale

- mot-forme = forme fléchie = plus petite unité de sens qui soit autonome
- mot-lemme = regroupement de mots-formes, qui ne varient que par la flexion
  - dénotation (ou référence) stable
  - est associé à un sens précis (= signifiant + signifié)
  - traditionnellement réprésenté par une des formes
    - infinitif, masculin sing ...
- c'est le mot-lemme qui est utile en sémantique lexicale
- ds ce cours: unité lexicale = lexème = mot-lemme

8

Signifiants

- il existe un autre usage courant du mot « mot » :
  - en parlant des « différents sens d'un mot »
  - => on fait référence au signifiant uniquement
- mot-signifiant
  - forme acoustique ou graphique d'un lexème
- toujours bien préciser/comprendre si « mot »
  - inclut la flexion ou pas
  - inclut le sens ou pas

9

Ambiguïté des signifiants

- un même signifiant
  - peut avoir plusieurs sens
  - (i.e. correspondre à plusieurs lexèmes)
- différents cas :
  - homographie : ambiguïté graphique de mots-formes
    - convient / convient, couvent / couvent
  - homophonie : ambiguïté acoustique de mots-formes
    - vert / ver
  - homonymie : ambiguïté graphique et acoustique de mots-lemmes
    - avocat / avocat

10

Sémantique lexicale

- unité lexicale
- signifiants et ambiguïté
- relations lexicales
- représentation du sens d'une UL
- rôles sémantiques

11

Homonymie / Polysémie

- homonymie
  = 2 lexèmes distincts ayant même signifiant
  - avocat$_1$ / avocat$_2$
- polysémie
  = 2 lexèmes ayant même signifiant, mais sémantiquement proches
  - polysémie régulière
    - contenant / contenu
    - bâtiment / organisation / personnes y travaillant
    - ressenti d'un sentiment / évocation d'un sentiment
      - je suis triste / ce film est triste
  - ou pas
    - connaître qqun / connaître un fait

12

Tests de polysémie

- interprétations distinctes:
  - le signifiant peut-il être utilisé dans une phrase avec plusieurs interprétations concurrentes?
    - *Pauline porte une jupe*
- zeugme :
  - caractère marqué d'énoncé avec évocation simultanée d'interprétations concurrentes
    - *elle porte une jupe et un paquet*
    - *il vit à Paris et une période difficile*
    - => si zeugme alors sens distincts
- en pratique: souvent difficiles de trancher entre 1 sens / plusieurs sens

13

Polysémie?

- sens de l'adjectif « rapide » ?
  - une voiture rapide = qui peut rouler vite
  - une décision rapide = que l'on prend rapidement
  - un scribe rapide = qui écrit rapidement
- les tests précédents sont inopérants
  - donneraient plutôt plusieurs sens
    - ?un scribe et une voiture rapide
  - => pourtant noyau de sens commun
  - => et nb de « sens » serait énorme, cf. dépendant des actions typiques faites avec l'objet / réalisées par l'agent
  - => la représentation de « rapide » et « voiture » et « décision » ... devrait capturer ces nuances

14

Synonymie

- lexèmes qui ont le même sens
  - dans tout ou partie de leurs contextes
    - *voiture / automobile*
    - *aspirine / acide acétylsalicilique*
    - *vomir / dégueuler*
- synonymie stricte existe peu (pas ?)
  - registre de langue
  - préférences lexicales
    - *grièvement / gravement blessé*
    - *??grièvement / gravement hypothéqué*

15

Antonymie

- lexèmes dont les sens respectifs diffèrent par deux caractéristiques/traits opposés
  - *sombre / clair*
  - *chaud / froid*
  - *dedans / dehors*
- opposition
  - binaire (*beau / laid, juste/injuste*)
  - d'extrémités sur une échelle
  - de sens d'évolution (*augmenter / décroître*)
  - ...

16

Hyponymie/Hyperonymie

- lexème1 de sens plus spécifique que lexème2
  - lexème1 hyponyme de lexème2
  - lexème2 hyperonyme de lexème1
- formellement:
  - caractérisation extensionnelle : référents de l'hyponyme inclus dans référents de l'hyperonyme
  - implication
    - A hyponyme de B <-> A(x) -> B(x)

17

Autres

- Meronymie/holonymie
  - lexèmes dont les sens sont en relation partie-tout
    - *roue* est méronyme de *voiture*

18

Sémantique lexicale

- unité lexicale
- signifiants et ambiguïté
- relations lexicales
- représentation du sens d'une UL
- rôles sémantiques

19

Représentation du sens d'un lexème

- inadéquation d'une représentation en extension
  - on peut connaître le sens du mot chien
  - sans connaître l'ensemble des chiens
  - d'autant que cet ensemble évolue

20

Représentation du sens d'un lexème

- définition en intension
  - conditions nécessaires et suffisantes
    - oiseau(x) <-> animal(x) ^ a-des-ailes(x) ^ ...
- via traits sémantiques atomiques
    - oiseau +ANIMAL +A-DES-AILES
- ou primitives sémantiques
    - KILL(x,y) <-> CAUSE(x, (BECOME(NOT(ALIVE(y)))
- rem: problèmes
    - comment délimiter l'ensemble des conditions nécessaires et suffisantes pour un mot?
      - quid d'un oiseau ayant perdu ses ailes ?
        - définition prototypique / instance s'approchant de la définition
    - traits = trop simplistes
    - primitives = d'où viennent-elles? comment les lister?

21

Représentation du sens d'un lexème

- le lexique génératif (Pustejovky, 94)
  - critique de la gestion de la polysémie par inventaire de sens
  - entrées lexicales **structurées**
  - polysémie régulière capturée par règles sur ces structures

22

Sémantique lexicale

- unité lexicale
- signifiants et ambiguïté
- relations lexicales
- représentation du sens d'une UL
- rôles sémantiques : voir cours sur SRL

23

Sémantique lexicale computationnelle

- ressources pour l'anglais
  - réseau lexical : Wordnet
  - classes sémantiques : VerbNet
  - rôles sémantiques : FrameNet, PropBank
- pour le français?
- tâches
  - Calcul de similarité lexicale
  - **WSD** : désambiguisation lexicale (word sense disambiguation)
  - **SRL** : étiquetage de rôles sémantiques (semantic role labeling)
  - **RTE** : Reconnaissance d'inférence textuelle (recognizing textual entailment)

24

Wordnet

- LA ressource lexicale la plus utilisée en TAL
  - développée à Princeton depuis 1985
  - pour l'anglais
  - téléchargeable ici : http://wordnet.princeton.edu
  - voir Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670
  - utilitaire de recherche : wn
- WordNet =
  - **synsets** (= groupe de sens, presque synonymes)
  - reliés par relations lexicales et relations sémantico-conceptuelles

25

Couverture

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Polysemy information

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

26

Outils

- visualisation en ligne
- exécutable **`wn`** (installation locale)
- Wordnet dans NLTK
  - voir prochain TP

27

Synset

- sens =
  - lemme + cat
  - numéro de sens
  - caractérisé par son appartenance à un synset
- synset =
  - liste de sens
  - définition (+ exemple)
  - relations avec d'autres synsets
- NB:
  - la définition du sens est l'appartenance au synset, et donc les sens se définissent mutuellement

28

Exemple (`wn bass -s -over`)

The noun "bass" has 8 senses in WordNet.
1. bass[1] - (the lowest part of the musical range)
2. bass[2], bass part[1] - (the lowest part in polyphonic music)
3. bass[3], basso[1] - (an adult male singer with the lowest voice)
4. sea bass[1], bass[4] - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass[1], bass[5] - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass[6], bass voice[1], basso[2] - (the lowest adult male singing voice)
7. bass[7] - (the member with the lowest range of a family of musical instruments)
8. bass[8] - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

The adjective "bass" has 1 sense in WordNet.
1. bass[1], deep[6] - (having or denoting a low vocal or instrumental range)
"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

29

Exemple (en ligne)

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- S: (n) **bass**, basso (an adult male singer with the lowest voice)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

30

Relations entre noms

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Has-Instance		From concepts to instances of the concept	*composer*^{1} → *Bach*^{1}
Instance		From instances to their concepts	*Austen*^{1} → *author*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Antonym		Opposites	*leader*^{1} → *follower*^{1}

31

Relations entre verbes

Relation	Definition	Example
Hypernym	From events to superordinate events	*fly*^{9} → *travel*^{5}
Troponym	From a verb (event) to a specific manner elaboration of that verb	*walk*^{1} → *stroll*^{1}
Entails	From verbs (events) to the verbs (events) they entail	*snore*^{1} → *sleep*^{1}
Antonym	Opposites	*increase*^{1} ⟺ *decrease*^{1}

32

Voir la hiérarchie (`wn bass -hypen`)

```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

33

Wordnets pour d'autres langues

- Il existe des wordnets à moindre échelle pour bcp de langues
- diverses techniques de projection automatique du WN anglais vers d'autres langues
- pour le français
  - EuroWordnet comprend une partie FR
  - WOLF (Benoît Sagot) : wordnet **libre** du français

34

VerbNet

- classes sémantiques de verbes anglais
  - Kipper-Schuler, 2006, Univ Colorado
- avec une classification guidée par la **syntaxe**
  - plus précisément les **alternances syntaxiques**
    - cf. Beth Levin, 1993, English verb classes and alternations: a preliminary investigation, Chicago, The University of Chicago Press
  - les verbes admettant les mêmes ensembles d'alternances sont sémantiquement proches
- exemple :
  - « causative / inchoative alternation »
    John broke the window => The window broke
  - caractéristique des verbes de changement d'état ou de position

35

VerbNet (suite)

- rôles sémantiques
- sémantique compositionnelle avec primitives
- (voir cours sur SRL)

36

Sémantique lexicale computationnelle

- ressources pour l'anglais
  - réseau lexical : Wordnet
  - classes sémantiques : VerbNet
  - rôles sémantiques : FrameNet, PropBank
- pour le français?
- tâches
  - **Calcul de similarité lexicale**
  - **WSD** : désambiguisation lexicale (word sense disambiguation)
  - **SRL** : étiquetage de rôles sémantiques (semantic role labeling)
  - **RTE** : Reconnaissance d'inférence textuelle (recognizing textual entailment)

37

Similarité lexicale

- Motivations
- mesures de similarité d'après thésaurus
- mesures de similarité distributionnelle
- Evaluation

38

Motivations

- La synonymie est une relation forte, binaire
- or intuitivement, 2 mots non synonymes, peuvent avoir des sens **plus ou moins reliés**
  - exemples
    - voiture / automobile
    - gentillesse / générosité
    - gentillesse / automobile
- => la « similarité lexicale » quantifie ce lien
  - rem: selon la valeur de similarité, on capture en fait si 2 mots sont
    - synonymes, similaires, reliés (même champ sémantique), n'ont rien à voir…

39

Motivations

- similarité lexicale
  - quantifiée par un réel sim(mot1,mot2) comme score $\in[0, +\infty]$
- utilisable pour toute application de TAL, pour capturer la variation lexicale
  - ex: j'ai vu dans mon corpus que « banane » peut être l'objet de « manger »,
  - « banane » sémantiquement proche de « poire » => a priori « poire » peut être objet de « manger »

40

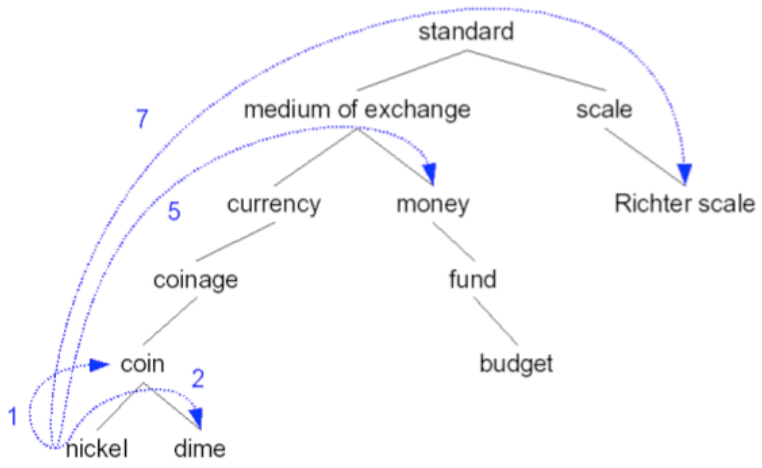
Similarité d'après thésaurus

- thésaurus = ressource lexicale avec liens entre entrées, en particulier liens d'hyperhonymie
- similarité calculable d'après la position des nœuds dans le thésaurus

41

Similarité via chemins

- similarité identifiée à la longueur du chemin



42

Similarité via chemins

- algorithme de base :
  - pour 2 **synsets** ou **concepts** ou **nœuds** s1 et s2 :
    - longchem(s1,s2) = le nb d'arêtes dans le chemin le plus court entre s1 et s2, en suivant des liens d'hyper/ d'hyponymie
    - D = profondeur maximale

$$simsens(s1,s2) = \frac{1}{longchem(s1,s2)}$$

$$sim_{Leacock/Chodorow\_98}(s1,s2) = -\log \frac{longchem(s1,s2)}{2 * D}$$

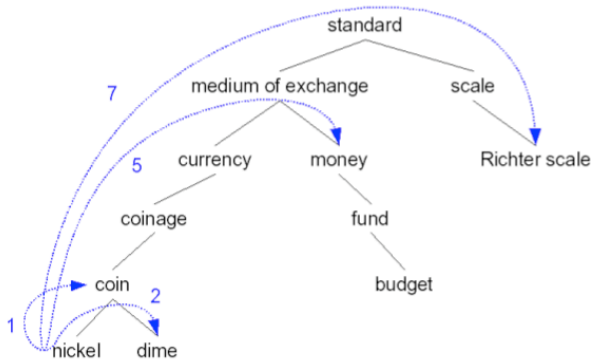
  - approximation, pour 2 **lemmes** w1 et w2 :

$$simlem(w1,w2) = \max_{s1 \in sens(w1), s2 \in sens(w2)} simsens(s1,s2)$$

43

Problème avec ce type de métrique

- chaque instance de relation IS-A ne représente par forcément le même écart de similarité
  - nickel/money plus proches que nickel/standard



44

Introduction de comptes sur corpus

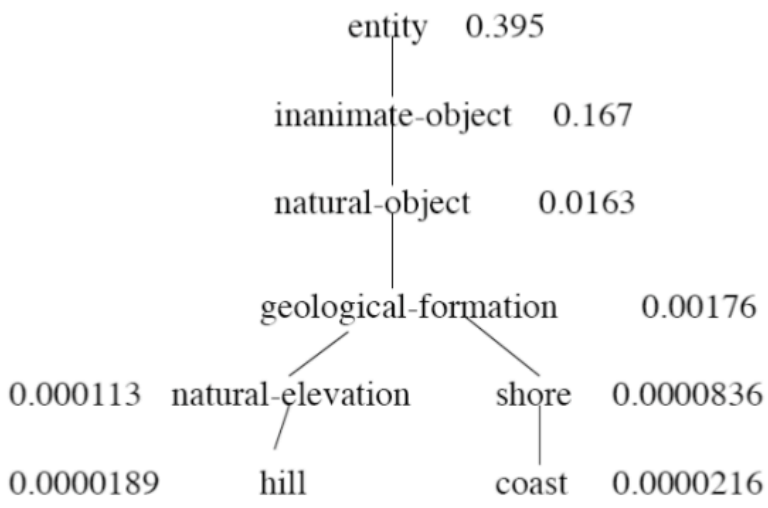
- Idée : ajouter des probabilités aux nœuds du thésaurus
  - si mots(s) = les mots des nœuds hyponymes de s
  - C un corpus de taille N
  - alors Resnik (95) définit P(s), la proba qu'un mot choisi au hasard dans C soit une instance de s ou d'un de ses hyponymes

$$P(s) = \frac{\sum_{w \in mots(s)} occ(w)}{N}$$

  - plus un nœud est bas dans la hiérarchie, moins grand est son ensemble de mots, et donc plus petite est P(s)

45

=> Wordnet augmenté de probas sur corpus



46

Information Content (IC) similarity

- contenu informationnel = information content
  - IC(s) = -log(P(s))
    - plus un concept est précis / spécifique, plus son IC est grand
    - inversement, un concept général aura un IC petit
- Plus petit ancêtre commun = lower common subsumer (LCS)
  - LCS(s1,s2) = le nœud le plus bas qui soit hyperhonyme de s1 et s2

47

Mesures de similarité dérivées de l'IC

Resnik (1995)'s similarity:

$$sim_{resnik}(c_1, c_2) = -\log P(LCS(c_1, c_2))$$

Lin (1998)'s similarity:

$$sim_{lin}(c_1, c_2) = \frac{2P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

Jiang and Conrath (1997)'s similarity:

$$sim_{JC}(c_1, c_2) = \frac{1}{2\log P(LCS((c_1, c_2)) - (\log P(c_1) + \log P(c_2))}$$

48

Similarité via dictionnaires: Lesk étendu

- idée de base : plus 2 mots sont similaires, plus leurs définitions vont avoir du vocabulaire commun
  - *planche : **Morceau de bois** plat et **allongé**, destiné à la construction ou à la fabrication d'objets, à l'aménagement d'éléments de rangement, à des rayonnages, etc.*
  - *poutre : **Morceau de bois**, métal ou béton armé, de forme **allongée**, de section étudiée pour une bonne résistance à la flexion.*
- Pour tout n-gramme de lemmes co-occurrent
  - ajouter n^2 au score
  - pour l'exemple on obtient $3^2 + 1^2 = 10$
- Lesk étendu utilise également les recouvrements entre définitions des hyper/hypo/méro-nymes

49

Evaluation de similarités via thésauri

- Evaluation intrinsèque
  - coefficient de corrélation entre score obtenus par un algorithme et score fourni par humains
    - mesure statistique
- Evaluation extrinsèque
  - (ou orienté tâche)
  - évaluation du gain obtenu en utilisant la mesure de similarité dans une application
    - détection de plagiat
    - notation automatique de devoirs (!)
    - tâches de TAL: analyse syntaxique, WSD...
- Jiang-Conrath et extended Lesk ont tendance à mieux fonctionner

50

Similarité via thésaurus : outils

- il existe diverses implémentations de similarités lexicales utilisant WordNet
  - module perl WordNet::Similarity (Patwardhan et Pederson, 2003)
  - NLTK …

51

Lacunes des méthodes via thésaurus

- fortement dépendantes de la couverture du thésaurus
  - pour de nombreuses langues : couverture faible
  - même pour l'anglais : il existe toujours des mots non couverts, en particulier spécifiques à un domaine
- s'appuient sur hyper/hypo-nymie
  - bien définie pour noms, lacunaire pour adj, v
- => technique complémentaire : similarité distributionnelle

52

Similarité distributionnelle

- Bloomfield, Harris : analyse distributionnelle
  - regroupements sur base distributionnelle
  - similarité syntaxique
- peut être étendue à la sémantique
  - Firth (57) : « You shall know a word by the company it keeps »
- Nida (75) (cité par Lin 98)
  - *A bottle of tezguino is on the table*
  - *Everybody likes tezguino*
  - *Tezguino makes you drunk*
  - *We make tezguino out of corn*
- Idée :
  - le sens d'un mot inconnu est devinable par le contexte
  - donc le contexte doit aider à caractériser le sens
  - **la similarité de contextes doit aider à caractériser la similarité de sens**

53

Modélisation de la similarité distrib.

- Modèle vectoriel
- fonctionnant en général sur les lemmes ou les paires lemme+cat
  - cf. on utilise en général des corpus non désambiguisés
- un lemme représenté par un vecteur
- espace vectoriel = une dimension par « **contexte** » possible d'un lemme
- similarité(lem1, lem2) = similarité entre les vecteurs représentant lem1, lem2

54

Modélisation de la similarité distrib.

- Quels **contextes** ?
- Quels **fonction de poids** donner aux contextes, pour un lemme donné
  - i.e. quelle valeur donner à la dimension du contexte *c*, pour le lemme *lem* ?
- Quelle **fonction de similarité** ?

55

Contextes linéaires

- un « contexte » peut être décomposé en
  - une relation + un mot
- contextes linéaires :
  - on n'utilise que l'ordre des mots, pas de structure
  - exemple:
    - dans « *les singes cueillent des bananes dans la jungle* »
    - en ignorant les mots outils,
    - on a pour *banane* une occurrence de chacun des contextes suivants:
      - (apparaît_à_une_distance_de_3_mots, *jungle*)
      - (apparaît_à_une_distance_de_-2_mots, *cueillir*)
      - (apparaît_à_une_distance_de_-3_mots, *singe*)

56

Contextes syntaxiques

- mais l'apparition de modifieurs peut modifier les distances
  - *les singes cueillent souvent des bananes dans la jungle*
  - => *2 contextes sont différents*
    - (apparaît_à_une_distance_de_-**3**_mots, cueillir)
    - (apparaît_à_une_distance_de_-**4**_mots, singe)
    - et donc sont comptabilisés à des dimensions différentes ds le vecteur de *banane*
- => contexte syntaxique : chemin dans l'arbre de dépendances syntaxiques
  - contexte syntaxique à un pas : (objet_inverse, *cueillir*)
  - contexte plus long : (objet_inverse+modifieur(*dans*), jungle)
- remarque : on peut utiliser à la fois les contextes linéaires et les contextes syntaxiques

57

Pondération des contextes

- tous les contextes n'ont pas le même contenu informationnel
  - les lemmes fréquents ont tendance à être très ambigus, et donc les contextes les mettant en jeu sont bruités
  - l'interprétation d'un modifieur peut varier selon le nom tête
    - le contexte (apparaît_à_une_distance_1_mots, *rapide*) est moins discriminant que (apparaît_à_une_distance_1_mots, *torrentiel*)
- => pondération des contextes
  - pondération globale (« rapide » plus vague que « torrentiel »)
  - et relative au mot de départ (« torrentiel » particulièrement pertinent dans le contexte de « pluie »
- => capturée par scores d'association entre mots, ou entre mot et contexte

58

Pondération des contextes

- soit w le lemme dont on construit le vecteur
- soit (r, w') un contexte
- on note occ(w,r,w') le nb d'occurrences de w dans le contexte de (r,w')
- on remplace par * pour noter les comptes sommant sur toutes les valeurs possibles
  - occ(*,r,w') = nb d'occ de n'importe quel lemme dans le contexte de (r,w')

59

Pondération des contextes

- proba
  - P(w | r,w') = occ(w,r,w') / occ(*, r, w')
  - (=estimation par max de vraisemblance)

60

Pointwise Mutual Information:

- rapport entre la probabilité d'occurrence conjointe de 2 évènements, et la probabilité d'occurrence conjointe si ces événements étaient indépendants

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- dans le cas de la sim distributionnelle cela donne

$$I(w,(r,rw') = \log_2 \frac{P(w,rw')}{P(w)P(rw')}$$

$$= \log_2 \left(\frac{occ(w,r,w')}{occ(*,*,*)} \frac{occ(*,*,*)}{occ(w,*,*)} \frac{occ(*,*,*)}{occ(*,r,w')}\right)$$

61

Lin 1998

- raffinement spécifique au cas à trois variables (w, r, w') :
  - on a toujours P(w,r,w')=P(r) P(w|r) P(w'|r,w)
  - l'hypothèse d'indépendance est de calculer l'occurrence conjointe de w,r,w' comme si w et w' étaient indépendants sachant r
  - sous cette hypothèse : P(w,r,w')=P(r) P(w|r) P(w'|r)

$$assoc_{Lin98} = \log_2 \frac{P(w,r,w')}{P(r)P(w \mid r)P(w' \mid r)}$$

$$= \log_2 \frac{occ(w,r,w')occ(*,r,*)}{occ(*,r,w')occ(w,r,*)}$$

62

Fonctions de similarité

- soit le lemme $\mathbf{l_j}$, on note v_j le vecteur associé à $\mathbf{l_j}$ et $\mathbf{p_{ij}}$ le poids du contexte c_i pour le lemme l_j

$$\text{rappel : } \vec{v_j}.\vec{v_k} = \sum_{i=1}^{C} p_{ij} * p_{ik} \text{ et } \left\|\vec{v_j}\right\| = \sqrt{\vec{v_j}.\vec{v_j}}$$

- alors on définit les similarités

$$\cos(\vec{v_j},\vec{v_k}) = \frac{\vec{v_j}.\vec{v_k}}{\left\|\vec{v_j}\right\|\left\|\vec{v_k}\right\|} \qquad dice(\vec{v_j},\vec{v_k}) = \frac{\vec{v_j}.\vec{v_k}}{\left\|\vec{v_j}\right\|^2 + \left\|\vec{v_k}\right\|^2 \Big/ 2}$$

$$jaccard(\vec{v_j},\vec{v_k}) = \frac{\vec{v_j}.\vec{v_k}}{\left\|\vec{v_j}\right\|^2 + \left\|\vec{v_k}\right\|^2 - \vec{v_j}.\vec{v_k}}$$

63