Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille

Cécile Fabre Nabil Hathout Franck Sajous Ludovic Tanguy

CLLE-ERSS
CNRS & Université de Toulouse

Thématiques actuelles de la recherche en TAL
15 décembre 2014

Vue d'ensemble

1. Introduction
2. Évaluation
3. 720 configurations
4. Analyse des résultats
5. Conclusion

Contexte

- Réponse à la tâche 2 de l'atelier SemDis-2014.
  - Tâche exploratoire.
  - Utilisation d'un corpus commun.
  - Mise en œuvre de méthodes d'analyse distributionnelle.
  - Choix de mots à privilégier pour les observations et les comparaisons.

Corpus TALN

- Corpus constitué d'une sélection d'articles en français issus des conférences TALN et RECITAL sur la période 2007 à 2013.
- 584 articles ; environ 2 millions de mots.
- Disponible et utilisable librement à des fins de recherche.
- http://redac.univ-tlse2.fr/corpus/taln.html

Contexte (2)

TALN est un petit corpus

	corpus	taille
SemDis	TALN	2 M
Ferret (2010)	AQUAINT 2	380 M
Baroni and Lenci (2010)	ukWaC	2 000 M

- Nous défendons l'idée qu'il est possible d'appliquer les méthodes et les outils de l'analyse distributionnelle à un petit corpus spécialisé.
- Recherche à mi-chemin entre :
  - **les méthodes d'extraction et de structuration de terminologie :**
    - opèrent sur des petits corpus ;
    - visent la mise au jour de relations conceptuelles spécifiques.
  - **l'analyse distributionnelle « standard » :**
    - utilise des corpus volumineux tout venant (de tous types) ;
    - identifie des relations de similarité sémantique au sens large.

Historique

- L'analyse distributionnelle basée sur l'analyse syntaxique est une thématique de recherche bien établie de l'axe TAL (CARTEL) de CLLE-ERSS.
  - Travaux initiés au début des années 2000.
  - Voisins de Le Monde ; Voisins de Wikipedia; Voisin d'En Face (Frantext/Le Monde).
- À l'origine, les bases ont été construites au moyen de l'analyseur Syntex et du système d'analyse distributionnelle Upery (Bourigault, 2007).
- Les nouvelles bases sont construites en utilisant l'analyseur Talismane (Urieli, 2013) et un système d'AD générique et paramétrable : BOutAD (CLLE-ERSS) + Librairie R *Wordspace* Evert (2014).

Historique (2)

Retour aux sources de l'analyse distributionnelle

Démarche similaire à celle de Harris et al. (1989).

domaine du TAL	domaine de l'immunologie
584 articles de conférence	25 articles et rapports
2007 – 2013	1935 – 1966
français	anglais et français
analyse syntaxique automatique + extraction des contextes	analyse manuelle
triplets syntaxiques (gouverneur, relation, dépendant)	phrases élémentaires
relations de similarité sémantique	classes de mots (noms) et classes d'opérateurs
données d'évaluation construites par des experts	interprétation des classes par des experts

Objectifs de notre étude

- Comprendre plus finement les mécanismes distributionnels
  - Identifier les principaux paramètres qui déterminent la construction d'une ressource distributionnelle
  - Comprendre les effets de ces différents paramètres
- Plus précisément :
  - Mettre l'accent sur la mise au point des paramètres situés en amont du calcul du similarité.
  - Mieux contrôler les conditions d'utilisation des contextes linguistiques.
- Travailler sur des textes en langue de spécialité impliquant des relations sémantiques spécifiques
- Vérifier que l'analyse distributionnelle reste opérationnelle sur des petits corpus
- Proposer un référentiel pour l'évaluation des modèles distributionnels

Vue d'ensemble

1. Introduction
2. **Évaluation**
3. 720 configurations
4. Analyse des résultats
5. Conclusion

Données pour l'évaluation

L'évaluation des systèmes d'AD est notoirement difficile

- La notion de similarité sémantique ne peut pas être définie de façon précise.

L'évaluation est classiquement indirecte (au sens de Baroni and Lenci (2011))

- Comparaison à des ressources externes de type différent : réseaux lexicaux ; dictionnaires de synonymes ; thésaurus.
- Évaluation dans le cadre de tâches (jugements de synonymie ou d'analogie etc.)

Données pour l'évaluation (2)

Évaluation directe (intrinsèque)

- Construction d'un jeu de données spécifique pour l'évaluation d'une ressource distributionnelle.
- *Pooling method* : examen par des annotateurs experts des réponses proposées par différentes méthodes

Exemple : BLESS *(Baroni & Lenci Evaluation of Semantic Spaces)*

- Jeu de données d'évaluation généraliste, pour l'anglais.
- Comprend les voisins de 200 noms concrets.
- Les voisinages sont restreint à 5 relations sémantiques lexicales : hyperonymie, hyponymie, méronymie, adjectif décrivant un attribut du concept, verbe qui décrit une action, une activité, un événement dans lequel le concept est impliqué.

Constitution du référentiel l'évaluation

- Le descriptif de la tâche proposait initialement 5 noms, 1 verbe et 2 adjectifs.
  Passage à **5 noms**, **5 verbes** et **5 adjectifs**
  - Jeu de donnée équilibré pour les 3 catégories.
  - Les mots-cibles ont des fréquences comparables (600 occurrences en moyenne).
- Pour chaque méthode de calcul des voisins, tous les seuils ont été réglés au plus bas pour obtenir la liste de voisins la plus large
- **4 juges.** Chaque juge choisit pour chaque mot-cible, parmi les voisins proposés par l'ensemble des systèmes, 10 voisins jugés les plus proches sémantiquement
  - La liste des voisins de chaque mot-cible est l'union des réponses des 4 juges.
  - On conserve le nombre d'annotateurs ayant choisi chaque voisin.
  - La situation est idéale : les 4 auteurs sont des spécialistes de TAL

Constitution du référentiel l'évaluation (2)

- **Approche « pragmatique »**
  - Considérer la similarité sémantique *pour ce qu'elle est*
  - On ne cherche pas à retrouver l'inventaire des relations lexicales classiques.
- Accord inter-annotateurs : score moyen de F-mesure = 0,59
- Le jeu de données SemDis2014 est disponible à l'adresse : `http://redac.univ-tlse2.fr/corpus/semdis2014/`

Une tâche amusante, éducative et utile

- Choisissez les 3 meilleurs voisins de *Complexe* parmi :
  polarisé, *aisé*, *biomédical*, *polysémique*, *facile*, *long*, *étendu*, *divers*, *souple*, *simulé*, *hétérogène*, *ardu*, *haut*, *précis*

Une tâche amusante, éducative et utile

- Choisissez les 3 meilleurs voisins de *Complexe* parmi :
  polarisé, *aisé*, *biomédical*, *polysémique*, *facile*, *long*, *étendu*, *divers*, *souple*, *simulé*, *hétérogène*, *ardu*, *haut*, *précis*
- Choisissez les 3 meilleurs voisins de *Graphe* parmi :
  vecteur, *prétraitement*, *noeud*, *règle*, *format*, *arbre*, *champ*, *réseau*, *statut*, *adjonction*, *cascade*, *flux*, *regroupement*, *liste*

Une tâche amusante, éducative et utile

- Choisissez les 3 meilleurs voisins de *Complexe* parmi :
  polarisé, *aisé*, *biomédical*, *polysémique*, *facile*, *long*, *étendu*, *divers*, *souple*, *simulé*, *hétérogène*, *ardu*, *haut*, *précis*
- Choisissez les 3 meilleurs voisins de *Graphe* parmi :
  vecteur, *prétraitement*, *noeud*, *règle*, *format*, *arbre*, *champ*, *réseau*, *statut*, *adjonction*, *cascade*, *flux*, *regroupement*, *liste*
- Choisissez les 3 meilleurs voisins de *Annoter* parmi :
  découper, *inventer*, *procurer*, *contenir*, *formaliser*, *interrogeabler*, *catégoriser*, *concevoir*, *juger*, *calculer*, *dire*, *traiter*, *aligner*, *reconnaître*

Mots-cibles et voisins sémantiques

mot-cible	**vois.**	**acc.**	**exemples** (nombre annotateurs)
adjectifs			
complexe	19	0,58	*compliqué* (4), *composé* (3), *simple* (3)
correct	19	0,55	*bon* (4), *pertinent* (4), *valide* (4)
important	17	0,65	*grand* (4), *majeur* (4), *principal* (4)
précis	16	0,72	*détaillé* (4), *exhaustif* (4), *fin* (3)
spécialisé	15	0,73	*juridique* (4), *médical* (4), *spécifique* (3)
noms			
fréquence	19	0,58	*nombre* (4), *poids* (4), *probabilité* (4)
graphe	20	0,55	*réseau* (4), *structure* (4), *treillis* (4)
méthode	17	**0,75**	*algorithme* (4), *approche* (4), *procédure* (3)
trait	19	0,57	*attribut* (4), *caractéristique* (3), *propriété* (3)
sémantique	23	**0,40**	*définition* (4), *contenu* (3), *sens* (3)
verbes			
annoter	20	0,50	*classer* (4), *étiqueter* (4), *baliser* (3)
calculer	23	0,47	*construire* (4), *estimer* (4), *évaluer* (4)
décrire	19	0,57	*détailler* (4), *présenter* (4), *représenter* (4)
évaluer	18	0,65	*mesurer* (4), *tester* (4), *valider* (4)
extraire	20	0,58	*acquérir* (4), *identifier* (3), *sélectionner* (3)

Petit bilan

- Des accords variés
  - Facile : *méthode*, *spécialisé*, *précis*, *évaluer*, les adjectifs en général
  - Difficile : *sémantique*, *calculer*, les noms en général
- Des relations qu'on ne trouve nulle part ailleurs :
  - trait-attribut
  - fréquence-poids
  - spécialisé-juridique
  - annoter-aligner

Vue d'ensemble

1. Introduction
2. Évaluation
3. **720 configurations**
4. Analyse des résultats
5. Conclusion

Méthodes d'analyse distributionnelle

- L'analyse distributionnelle établit une relation de similarité sémantique entre les unités qui apparaissent fréquemment dans les mêmes contextes.
- 2 niveaux de variation :
  1. Les contextes
  2. La manière de mesurer la similarité entre les contextes

Contextes

- Cooccurrents graphiques dans une fenêtre données
- **Cooccurrents syntaxiques**
  (relativement aux relations de dépendance syntaxiques)
  - Analyse syntaxique en dépendances du corpus TALN effectuée par Talismane.
  - L'analyse syntaxique permet de spécifier les caractéristiques linguistiques des contextes.

Expérience

- Variations selon 6 facteurs / paramètres.
- Les paramètres ont entre 2 et 10 valeurs chacun.
- La combinaison des différentes valeurs des paramètres forme

720 configurations

Extraction des triplets

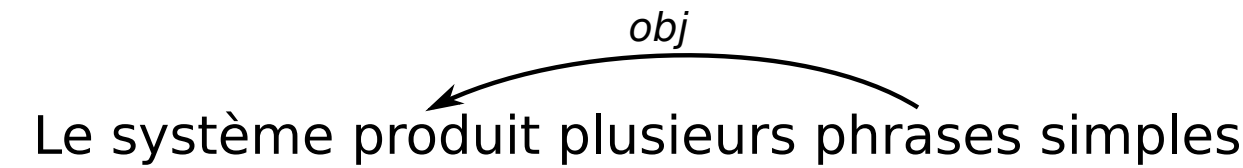
- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)

Extraction des triplets

- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)



Extraction des triplets

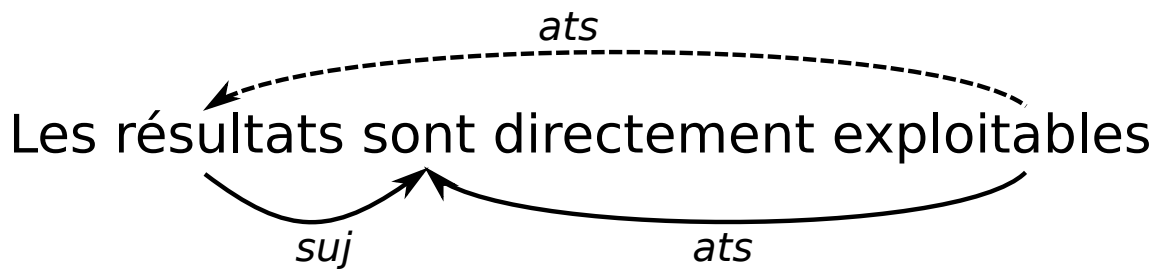
- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)

Extraction des triplets

- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

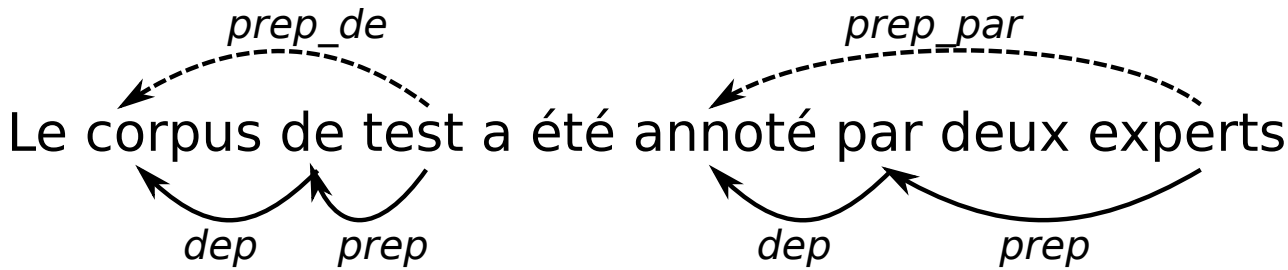
  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)



Extraction des triplets

- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)



Extraction des triplets

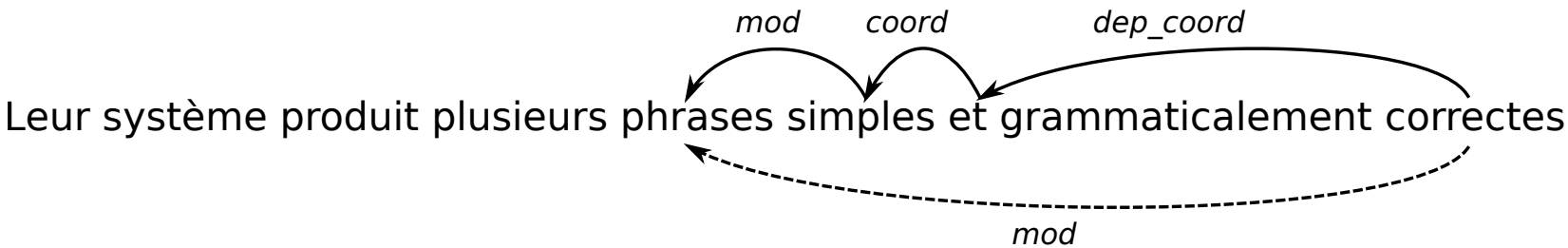
- L'analyse syntaxique fournit pour chaque mot 4 informations : lemme, catégorie, gouverneur et type de la relation de dépendance
- La première étape de l'AD est l'extraction de triplets:

  `<gouverneur; relation; dépendant>`
- Jeu de relations initiales (directes ou indirectes)
  - sujet (*suj*)
  - objet (`obj`)
  - modifieur (`mod`)
  - attribut du sujet (`ats`)
  - préposition (`prép`)

- Talismane fournit un score de confiance pour chaque dépendance.
- Paramètre 1 : On ne conserve que les dépendances dont le score de confiance dépasse 0%, 70%, 80%, 90%, 95% ou 98%

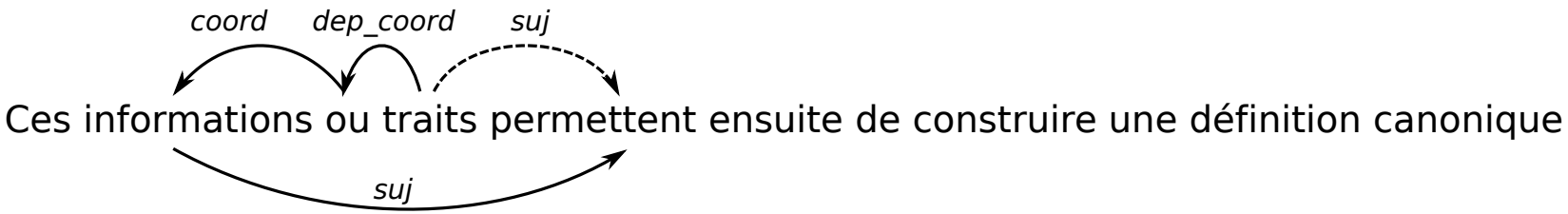
Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



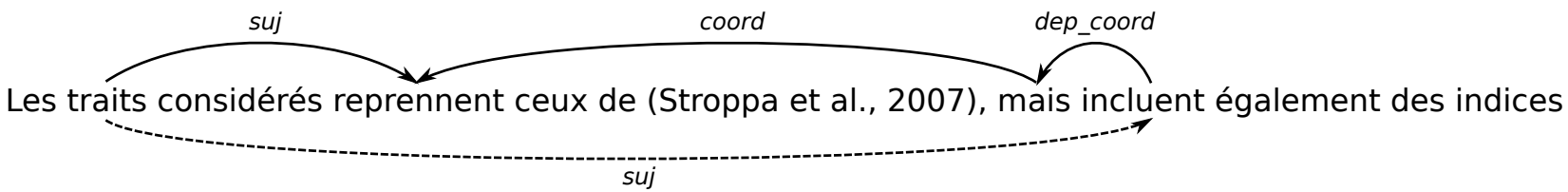
Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



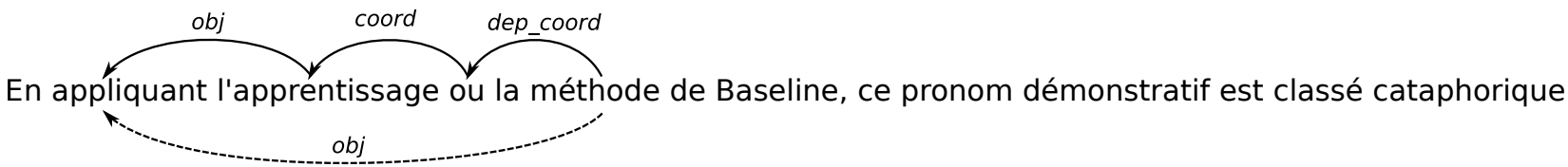
Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



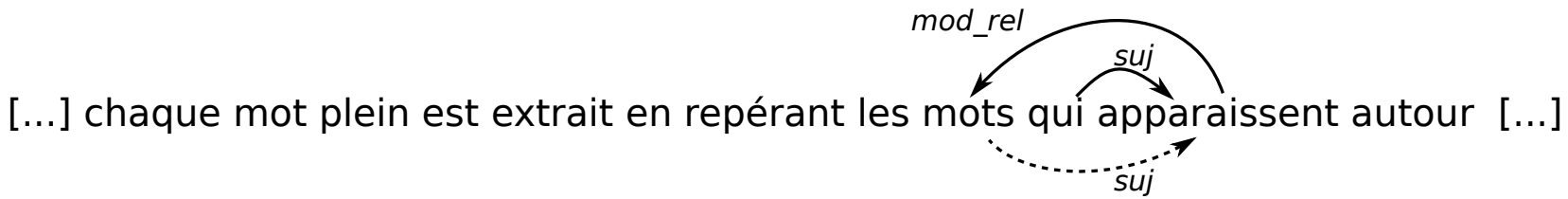
Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



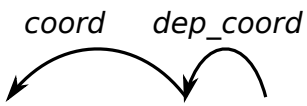
Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



Paramètre 2 : Normalisation des triplets

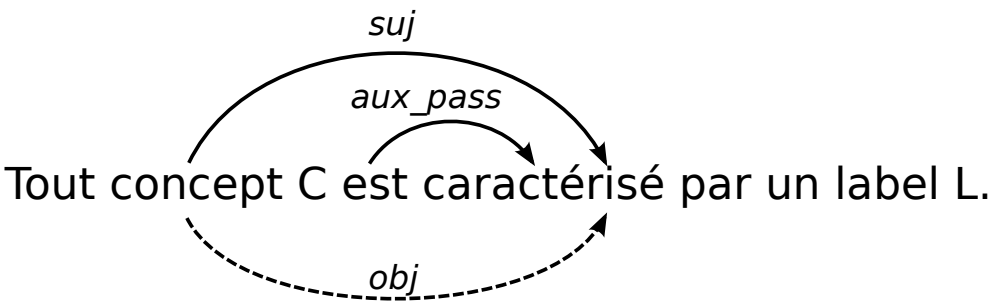
1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



Ces informations ou traits permettent ensuite de construire une définition canonique

Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.



Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.

Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.

Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.

Paramètre 2 : Normalisation des triplets

1. Distribution des relations sur les éléments coordonnés, en position de dépendants ou de gouverneurs.
2. Récupération de l'antécédent des pronoms relatifs sujet ou objet.
3. Ajout de la relation de coordination (*coord*)
4. Transformation de la relation *suj* en *obj* lorsque le gouverneur de cette relation est un passif
5. Conversion de la relation *ats* en *mod*
6. Regroupement des relations *prép*.

1 à 4 ajoutent des triplets

5 et 6 réduisent la dispersion des informations

Paramètre 3 : filtrage sur les fréquences

- On ne conserve que les triplets dont la fréquence dépasse :
  - 2 occurrences
  - 5 occurrences

Pondération

Association entre mots-cibles et contextes syntaxiques

- Un triplet <gouverneur; relation; dépendant> induit 2 contextes syntaxiques :
  - <gouverneur; relation> comme contexte de dépendant
  - <relation; dépendant> comme contexte de gouverneur

  Exemple : le triplet <NC:résultat; mod; ADJ:exploitable> induit les 2 contextes :
  - <NC:résultat; mod>
  - <mod; ADJ:exploitable>
- Chaque mot-cible est représenté par un vecteur de contextes syntaxiques

Paramètre 4 : pondération des associations

- Relativiser la fréquence d'un couple mot/contexte par rapport à la fréquence du mot et du contexte.
- Refléter la spécificité de chaque contexte pour la cible considérée.

- information mutuelle :

$$IM(l,c) = \log_2 \left(\frac{N \times f(l,c)}{f(l) f(c)} \right)$$

- t-score :

$$t\text{-}score(l,c) = \frac{f(l,c) - \frac{f(l)f(c)}{N}}{\sqrt{f(l,c)}}$$

$f(l)$ = nombre d'occurrences du lemme l
$f(c)$ = nombre d'occurrences du contexte syntaxique c
$f(l,c)$ = nombre d'occurrences de l dans le contexte c
N = nombre total d'occurrences de triplets

Paramètre 5 : mesure de similarité

- Cosinus :

$$cos(l_1, l_2) = \frac{\sum_i p_{1i} p_{2i}}{\sqrt{\sum_i {p_{1i}}^2 \sum_i {p_{2i}}^2}}$$

p_{ji} = pondération (IM ou t-score) du contexte c_i pour le lemme l_j

- Jaccard :

$$jacc(l_1, l_2) = \frac{|C(l_1) \cap C(l_2)|}{|C(l_1) \cup C(l_2)|}$$

$C(l_i)$ = ensemble des contextes dans lesquels le lemme l_i apparaît

Croisement similarité / pondération

- Jaccard n'utilise pas la pondération
- Les paramètres 4 et 5 définissent 3 mesures de similarité :
  - cosIM = cosinus sur des vecteurs d'IM
  - cosTS = cosinus sur des vecteurs de t-score
  - Jacc = Jaccard ; basé sur la productivité des lemmes

Paramètre 6 : filtrage sur le nombre de contextes partagés

- Les contextes marginaux tendent à rapprocher des couples de mots non pertinents
- On élimine les voisins dont le nombre de contextes partagés avec le mot-cible est inférieur à un seuil donné.
- 10 seuils $= 1 \ldots 10$

Paramètres et configurations obtenues

paramètre	nb	valeurs
seuil sur le score de confiance des dépendances syntaxiques	6	{0, 70, 80, 90, 95, 98}
normalisation des relations	2	{norm, nonorm}
seuil sur le nombre d'occurrences des triplets	2	{2, 5}
mesure de similarité	3	{cosIM, cosTS, Jaccard}
seuil sur le nombre de contextes partagés	10	[1, 10]

0_norm_2_*_1

- 1,4 million de couples
- 10 963 lemmes ont des voisins

98_norm_5_*_10

- 3 638 couples
- 279 lemmes ont des voisins

Vue d'ensemble

1. Introduction
2. Évaluation
3. 720 configurations
4. **Analyse des résultats**
5. Conclusion

Méthode de comparaison

Pour chacune des 720 configurations,
pour chacun des 15 mot-cibles du jeu d'évaluation,
on compare les 20 premiers voisins distributionnels avec les voisins de référence.
La comparaison prend en compte :

- l'ordre dans lequel les voisins sont classés
- le nombre d'annotateurs qui ont choisi ce mot comme voisin

nDCG Normalised Discounted Cumulated Gain

- Méthode de RI pour les référentiels valués.
- *nDCG* donne un score élevé aux systèmes qui renvoient en premier les voisins pertinents pour le plus grand nombre d'annotateurs.
- La normalisation permet la comparaison de voisinages de tailles différentes.
- Les scores moyens de *nDCG* permettent de comparer les paramétrages.

$$nDCG = \frac{DCG}{DCGI} \qquad \text{où} \qquad DCG = \sum_{i=1}^{20} \frac{annot_i}{\log_2(i+1)}$$

- $annot_i$ = nombre d'annotateurs qui ont sélectionné le voisin numéro i du système comme un bon voisin
- $DCGI$ = coefficient de normalisation
  = valeur maximale de DCG, obtenue par un système qui renverrait tous les mots dans l'ordre décroissant de pertinence.

nDCG moyens pour les valeur des paramètres

Paramètre	Moyenne	Écart-type
Score global	0,446	0,234
Score de confiance		
0%	**0,473**	0,233
70%	0,466	0,228
80%	0,464	0,231
90%	0,453	0,231
95%	0,428	0,230
98%	0,391	0,239
Normalisation		
Avec	**0,448**	0,234
Sans	0,443	0,233
Seuil de fréquence des triplets		
2	**0,500**	0,211
5	0,391	0,242

nDCG moyens pour les valeur des paramètres (2)

Paramètre	Moyenne	Écart-type
Mesure de similarité		
Cosinus IM	**0,521**	0,245
Cosinus t-score	0,389	0,233
Jaccard	0,427	0,202
Seuil sur les contextes partagés		
1	0,385	0,251
2	0,438	0,216
3	0,466	0,204
4	**0,474**	0,206
5	0,467	0,224
6	0,464	0,232
7	0,456	0,238
8	0,448	0,245
9	0,434	0,250
10	0,426	0,251

Configuration optimale

0_norm_2_cosIM_3

- pas de filtrage sur les relations de dépendance
- normalisation des contextes
- élimination des triplets de fréquence inférieure à 2
- pondération = IM, similarité = cosinus
- élimination des voisins qui ont moins de 3 contextes syntaxiques partagés avec la cible.

$$nDCG \text{ moyen} = 0{,}659$$

Variation par catégorie

nDCG pour les 720 configurations

nDCG	Adjectifs	Noms	Verbes
Maximum	0,827	0,917	0,872
Moyenne	0,311	0,533	0,493
Écart-type	0,212	0,221	0,206

Configurations optimales

verbes	`0_norm_2_cosIM_3`
noms	`80_norm_2_cosIM_7`
adjectifs	`0_norm_2_cosIM_1`

Comparaison *nDCG* / accord inter-annotateurs

Mot-cible	Maximum	Moyenne	Accord
complexe	0,620	0,194	0,58
correct	0,773	0,343	0,55
important	0,827	0,527	0,65
précis	0,748	0,285	0,72
spécialisé	0,454	0,208	0,73
Tous les adjectifs	0,591	0,311	0,65
fréquence	0,776	0,587	0,58
graphe	0,760	0,547	0,55
méthode	0,917	0,729	0,75
sémantique	0,649	0,237	0,40
trait	0,802	0,565	0,57
Tous les noms	0,733	0,533	0,57
annoter	0,607	0,355	0,50
calculer	0,815	0,545	0,47
décrire	0,816	0,504	0,57
évaluer	0,872	0,677	0,65
extraire	0,793	0,383	0,58
Tous les verbes	0,761	0,493	0,55

Pas de corrélation : globalement, les systèmes n'obtiennent pas de bons scores pour les mots que les humains analysent facilement ($r = 0,13$) ; sauf pour les noms ($r = 0,96$) et les verbes ($r = 0,47$).

Un cas facile : *Méthode* ($nDCG = 0,89$)

Rang	Mot	Pertinence
1	approche	4
2	technique	4
3	système	1
4	algorithme	4
5	stratégie	4
6	modèle	1
7	outil	1
8	méthodologie	4
9	processus	3
10	module	0
11	procédure	4
12	mesure	0
13	étape	1
14	analyseur	0
15	classifieur	1
16	règle	1
17	ressource	0
18	travail	0
19	critère	0
20	résultat	0

- Les voisins pertinents sont retrouvés ; la plupart des voisins « non pertinents » sont acceptables.
- Variété des relations syntaxiques et unité sémantique.
- Principaux contextes syntaxiques :

```
<proposer; obj>
<permettre; suj>
<présenter; obj>
<prep(de); apprentissage>
```

Un cas difficile : *Sémantique* (nom) ($nDCG = 0,30$)

Rang	Mot	Pertinence
1	propriété	0
2	signification	2
3	ambiguïté	1
4	nature	1
5	polysémie	0
6	syntaxe	3
7	aspect	0
8	définition	4
9	idée	0
10	diversité	0
11	notion	3
12	comportement	0
13	représentation	2
14	diacritique	0
15	caractéristique	1
16	distribution	1
17	délimitation	0
18	fermeture	0
19	structure	0
20	spécificité	1

- Le sens de *sémantique* est difficile à cerner (faible taux d'accord inter-annotateurs)
- Principaux contextes syntaxiques :

```
<mod; lexical>
<mod; compositionnel>
<prep(de); Montague>
<prep(de); mot>
<prep(de); phrase>
<prep(de); texte>
```

Complexe ($nDCG = 0,38$)

Rang	Mot	Pertinence
1	distinct	0
2	fastidieux	0
3	multimots	2
4	particulier	0
5	simple	3
6	composé	3
7	incomplet	0
8	typique	0
9	long	0
10	considéré	0
11	trivial	3
12	extrait	0
13	délicat	4
14	monosémique	0
15	spécifique	1
16	compliqué	4
17	classique	0
18	coûteux	3
19	fondamental	0
20	visé	0

- Les voisins sont des adjectifs généralement utilisés sur le plan rhétorique pour structurer le discours
  distinct, *particulier*, *visé*, *considéré*
- Les contextes partagés par les voisins expriment des notions très générales
  problème, *format*, *besoin*, *genre*, *tâche*, *modèle*, *configuration*, *phénomène*
- Les voisins sont des adjectifs sous-spécifiés pouvant modifier une large gamme de noms.
- `<NC:terme; mod>` est le contexte prédominant dans le corpus ($\rightarrow$ *terme complexe*)

Spécialisé ($nDCG = 0,39$)

Rang	Mot	Pertinence
1	cible	0
2	biomédical	4
3	généraliste	3
4	juridique	4
5	considéré	0
6	multilingue	0
7	analysé	0
8	médical	4
9	anglais	0
10	bilingue	0
11	technique	4
12	structuré	0
13	monolingue	0
14	volumineux	0
15	japonais	0
16	orthographié	0
17	existant	0
18	annoté	0
19	vietnamien	0
20	source	0

- Les contextes qui rapprochent les voisins distributionnels sont des noms de types de données langagières : *langue*, *terme*, *document*, *domaine*, *texte*, *corpus*, *discours*, *lexique*
- Moins d'adjectifs sous-spécifiés que pour *complexe*.

Vue d'ensemble

1. Introduction
2. Évaluation
3. 720 configurations
4. Analyse des résultats
5. **Conclusion**

Conclusion

Résultats atteints

1. Nous disposons d'un paramétrage « optimal »
   - *optimal* : pour ce corpus, les paramètres considérés, et le jeu d'évaluation
   - Préférence pour le cosinus et l'information mutuelle
2. Nous avons dégagé de nouvelles questions et hypothèses
   - Pourquoi l'AD a-t-elle plus de difficulté à capter la similarité entre les adjectifs qu'entre les mots des autres catégories ? (alors que pour les experts c'est l'inverse)
   - Cette sous-performance est-elle liée à une faible variété dans les contextes ? à la présence de beaucoup d'adjectifs « sous-spécifiés » ?

Perspectives

- Elargir la référence
  - à 30 mots, 10 de chaque catégorie
  - en faisant varier la fréquence (notamment avec de basses fréquences)
  - en considérant des phénomènes différents (spécialisé/générique)
  - en utilisant une méthode d'annotation plus adaptée (jugement binaire par *pooling method*)
  - 30 mots, 50 voisins considérés par mot
    $\rightarrow$ 1500 jugements par annotateur !
- Étudier plus finement la contribution des différentes relations syntaxiques
  - en multipliant les règles d'extraction des contextes syntaxiques
  - en envisageant plus de configurations différentes (avec/sans une relation spécifique, etc.)
- Comparer les méthodes syntaxiques à celles basées sur des fenêtres graphiques.
- Plusieurs milliers de configurations prévues, et donc des heures d'amusement en 2015.

Syntaxe vs. méthodes par fenêtre graphique

Une question obsolète ?

- La question n'est pas nouvelle, plusieurs travaux déjà réalisés. De la lassitude ? (cf. atelier SemDis 2014)
- Nous considérons la question encore ouverte :
  - finesse des contextes syntaxiques utilisés
  - multiplication des études sur les mesures de similarité (pour les 2 types de méthodes)
  - pertinence en fonction de la taille des corpus (// avec la tâche d'attribution d'auteurs : n-grammes vs. traits linguistiques)

Syntaxe vs. méthodes par fenêtre graphique (2)

« Dans les applications de commande en langue naturelle [...] »

Fenêtres graphiques

P:dans	-	1	#	NC:application	-	1	DET:les
P:dans	+	1	DET:les	NC:application	+	1	P:de
P:dans	+	2	NC:application	NC:application	-	2	P:dans
P:dans	+	3	P:de	NC:application	+	2	NC:commande
P:dans	+	4	NC:commande	NC:application	-	3	#
P:dans	+	5	P:en	NC:application	+	3	P:en
DET:les	-	1	P:dans	NC:application	+	4	NC:langue
DET:les	+	1	NC:application	NC:application	+	5	ADJ:naturel
DET:les	-	2	#	P:de	-	1	NC:application
DET:les	+	2	P:de	P:de	+	1	NC:commande
DET:les	+	3	NC:commande	P:de	-	2	DET:les
DET:les	+	4	P:en	P:de	+	2	P:en
DET:les	+	5	NC:langue	...			

Contextes syntaxiques

<application; prep_de; commande>
<application; prep_en; langue>
<langue; mod; naturel>

Syntaxe vs. méthodes par fenêtre graphique (3)

Couples mots/contextes

- contextes syntaxiques : entre 125 000 et 840 000 couples différents
- fenêtres graphiques [-5;+5] : 3,4 millions de couples différents
- fenêtres graphiques [-100;+100] : 11,9 millions de couples différents

Nettoyer les données en amont

Traitements post-analyse syntaxique

- corriger l'étiquetage (et la lemmatisation) VPP/ADJ

  « le Satellite d'une relation subordonnante modifie son Nucleus par l'adjonction d'un arbre auxiliaire ancré par la relation subordonnante. »

```
52   arbre        arbre        NC    nc    g=m|n=s          50   prep
53   auxiliaire   auxiliaire   ADJ   adj   n=s              52   mod
54   ancré        ancrer       VPP   v     g=m|n=s|t=past   52   mod
```

  « On doit être capable de déterminer quand déclarer un énoncé ancré. »

```
13   énoncé   énoncé   NC    nc    g=m|n=s          11   obj
14   ancré    ancrer   ADJ   adj   g=m|n=s|t=past   13   mod
```

- lemmatiser les mots inconnus

  (e.g. 4 flexions pour un ADJ inconnu $\rightarrow$ 4 contextes syntaxiques différents pour un nom modifié par cet ADJ)

- détection de langue phrase par phrase
  (abstracts en anglais, exemples linguistiques, etc.)

  Note : on peut espérer que la machinerie statistique travaille à l'élimination de ce bruit... sauf si ce bruit est régulier. Dans tous les cas : allègement des calculs

  Note (2) : traitement possible en amont de l'analyse syntaxique

Étendre le jeu dévaluation

Adjectif	**Fréquence**	**Spécialisé/Générique**
sémantique (+)	3074	S
important	1287	G
complexe	741	S
temporel (+)	698	S
correct	622	G
précis	383	G
spécialisé	377	S
significatif (+)	351	S+G
empirique (+)	86	G
computationnel (+)	60	S

Étendre le jeu dévaluation (2)

Nom	Fréquence	Spécialisé/Générique
méthode	3816	G
trait	1814	S
élément (+)	1576	G
performance (+)	1315	S
graphe	1119	S
contrainte (+)	947	S+G
fréquence	952	S
sémantique	398	S
dépendant (+)	96	S
signification (+)	76	S

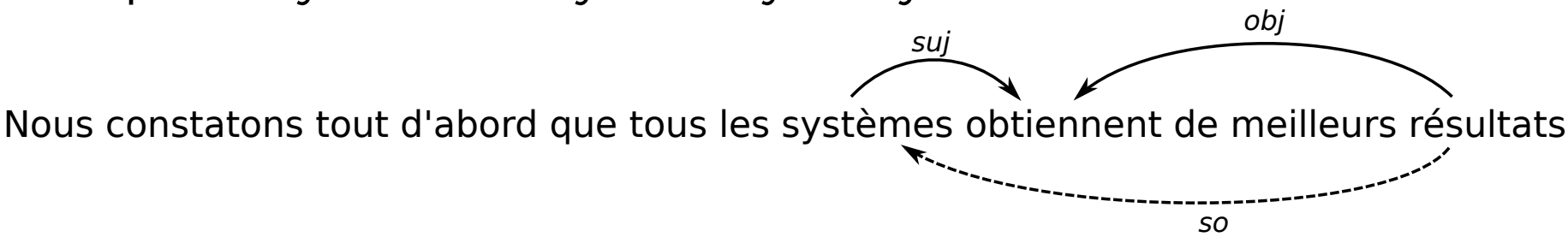
Étendre le jeu dévaluation (3)

Verbe	Fréquence	Spécialisé/Générique
décrire	1458	G
évaluer	1302	S
extraire	1165	S
calculer	1014	S
annoter	790	S
valider (+)	379	G
caractériser (+)	374	G
conduire (+)	366	G
indexer (+)	66	S
apparier (+)	54	S

Multiplier les règles d'extraction des triplets syntaxiques

Ajout de nouvelles relations

Exemples : *Sujet-Verbe-Objet* et *Sujet-Objet*



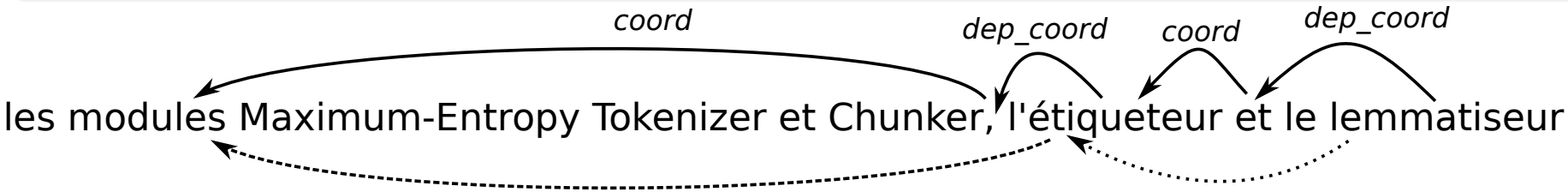

```
<NC:système; SVO/obtenir; NC:résultat>
<NC:système; SO; NC:résultat>
```

Multiplier les règles d'extraction des triplets syntaxiques

Ajout de règles de normalisation

Exemple : la coordination

« Nous utilisons le module Maximum-Entropy Tokenizer et Chunker, l'étiqueteur et le lemmatiseur basé sur WordNet. »



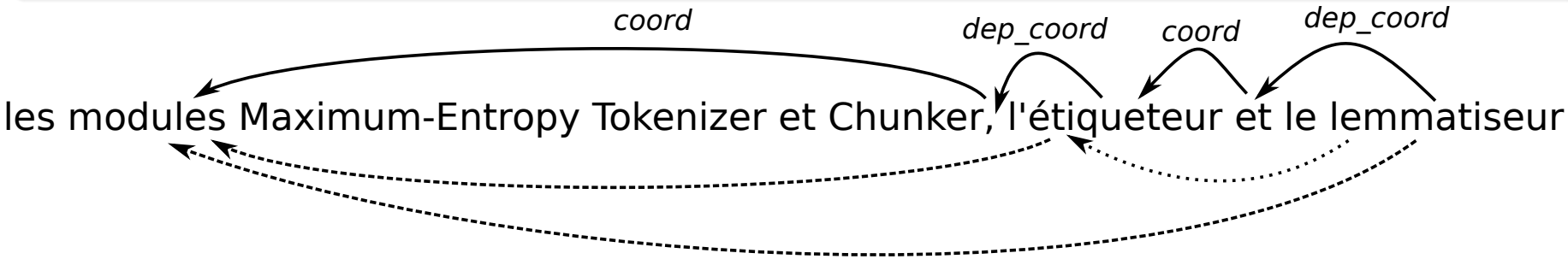
- Coordination : `<module; coord; étiqueteur>`, `<étiqueteur; coord; lemmatiseur>`

Multiplier les règles d'extraction des triplets syntaxiques

Ajout de règles de normalisation

Exemple : la coordination

« Nous utilisons le module Maximum-Entropy Tokenizer et Chunker, l'étiqueteur et le lemmatiseur basé sur WordNet. »



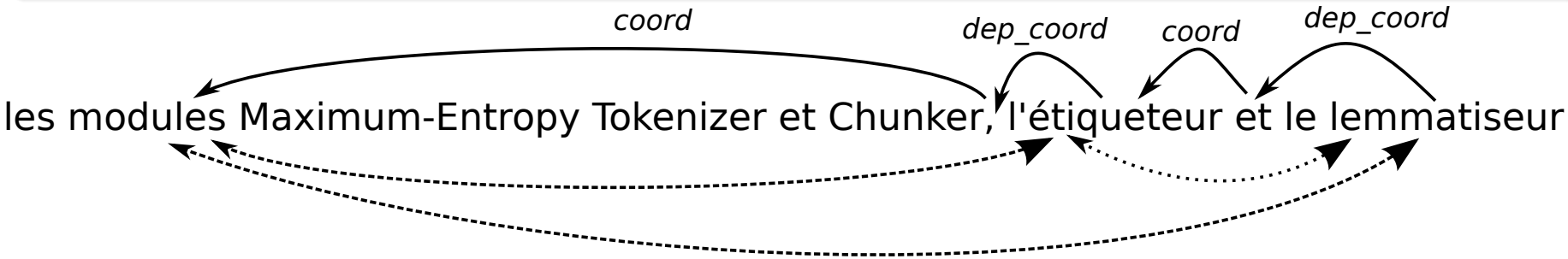
- Coordination : `<module; coord; étiqueteur>`, `<étiqueteur; coord; lemmatiseur>`
- Fermeture transitive : + `<module; coord; lemmatiseur>`

Multiplier les règles d'extraction des triplets syntaxiques

Ajout de règles de normalisation

Exemple : la coordination

« Nous utilisons le module Maximum-Entropy Tokenizer et Chunker, l'étiqueteur et le lemmatiseur basé sur WordNet. »



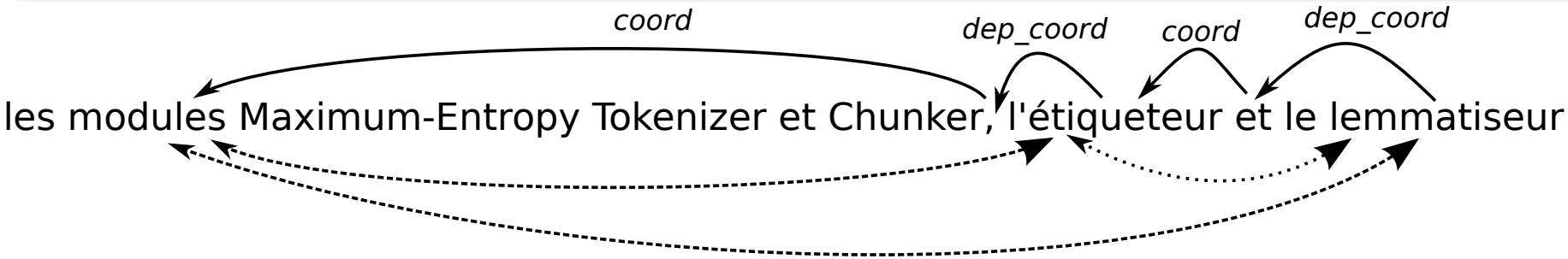
- Coordination : `<module; coord; étiqueteur>`, `<étiqueteur; coord; lemmatiseur>`
- Fermeture transitive : + `<module; coord; lemmatiseur>`
- Symétrisation : + `<étiqueteur; coord; module>`, `<lemmatiseur; coord; étiqueteur>` et `<lemmatiseur; coord; module>`

Multiplier les règles d'extraction des triplets syntaxiques

Ajout de règles de normalisation

Exemple : la coordination

« Nous utilisons le module Maximum-Entropy Tokenizer et Chunker, l'étiqueteur et le lemmatiseur basé sur WordNet. »



- Coordination : `<module; coord; étiqueteur>`, `<étiqueteur; coord; lemmatiseur>`
- Fermeture transitive : + `<module; coord; lemmatiseur>`
- Symétrisation : + `<étiqueteur; coord; module>`, `<lemmatiseur; coord; étiqueteur>` et `<lemmatiseur; coord; module>`
- Distribution des relations : ici, `<utiliser; obj>`

Multiplier les règles d'extraction des triplets syntaxiques...

Et observer finement la contribution de chaque relation...

- 65536 configurations possibles pour l'extraction des triplets
- 324 configurations pour les mesures de similarité

Globalement, quel est l'apport de la conversion de la relation *ats* en *mod* ?

Isoler les facteurs

- Fixer une mesure de similarité
- Choisir quelques configurations d'extraction de triplets
- Pour chacune d'elle, ajouter ou supprimer une règle

Isoler les facteurs

Configurations de base

1. dépendances directes de Talismane (*suj*, *obj* et *mod*)
2. dépendances directes + relations *prép*, *coord* et *ats*

Variations

- Cfg2 - *mod*, Cfg2 - *suj*, Cfg2 - *obj*, etc.
  (permet de quantifier l'apport d'une relation donnée /rap à une config. « de base »)
- Cfg2 + antécédance du ProRel, Cfg2 + normalisation des coordinations, Cfg2 + prise en compte des passifs, etc.

Isoler les facteurs (2)

Choisir une configuration « intelligente »

Prendre toutes les dépendances (directes et indirectes) + les normalisations correspondantes (sym. + trans. + distrib. de la coordination, pro. rel., passifs, etc.) et tester :

- l'apport d'une nouvelle relation (SVO)
- une variante de cette nouvelle relation (SO)
- des détails pour lesquels on a peu d'intuition (e.g. dissocier les verbes pronominaux réflexifs de leur équivalent non pronominal)

Isoler les facteurs (3)

Finalement...

Que conclure si l'impact d'une relation/normalisation n'est pas probant ?

- la relation/normalisation ne sert à rien ?
- le jeu d'évaluation est trop petit ?
  $\rightarrow$ étudier aussi l'impact sur le classement des voisins
- le corpus est trop petit ?

Répéter les manipulations

Autres corpus

- Corpus spécialisé de taille réduite
  Autre spécialité → quels juges pour l'évaluation ?
- Corpus spécialisé plus volumineux
  Spécialisé *et* volumineux ?
- Corpus plus « général », plus « diversifié » ?

Soyez réaliste, demandez l'impossible

- Idéalement : déterminer une configuration optimale par catégorie syntaxique, taille et type de corpus ?
  → e.g. pertinence du regroupement des prépositions/taille du corpus
- Se poser la question (et y répondre) du compromis effort/bénéfice. Exemple (pour finir sur une note bassement matérielle), le parsing de FrWaC :
  - Maxent, **beam=2**, 3 machines personnelles « sérieuses » : 3 mois
  - SVM, **beam=5**, un cluster de calcul « moyen » : moins d'une semaine

Références

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 673–721.

Baroni, M. and A. Lenci (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 1–10. Association for Computational Linguistics.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Habilitation à diriger des recherches, Université Toulouse II-Le Mirail, Toulouse.

Evert, S. (2014). Distributional semantics in R with the wordspace package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, Dublin, Ireland, pp. 110–114. Association for Computational Linguistics and Dublin City University.

Références (2)

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

Ferret, O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *7th International Conference on Language Resources and Evaluation (LREC'10)*, Malta, pp. 3338–3343.

Harris, Z. S., M. Gottfried, T. Ryckman, P. Mattick, A. Daladier, T. N. Harris, and S. Harris (1989). *The form of information in science: analysis of an immunology sublanguage*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat, Université de Toulouse II le Mirail.