Analyses *post-hoc* d'une tâche de substitution lexicale

Cécile Fabre et Ludovic Tanguy (CLLE-ERSS)

avec l'aide précieuse de :
Camille Mercier et Laura Rivière

et aussi celle de :
Nabil Hathout, Lydia-Mai Ho-Dac, François Morlane-Hondère,
Philippe Muller, Franck Sajous et Tim Van de Cruys

6 Novembre 2017

Plan de la présentation

Les tâches relatives au traitement de la polysémie (Navigli 2009)

- **Tâche explicite :** *Word Sense Disambiguation*.
  Nécessite l'attribution explicite d'un sens au mot cible.
  Tâche supervisée, fondée sur l'inventaire préalable des différents sens d'un mot.
- **Tâches implicites :**
  Le traitement réalisé s'appuie sur une désambiguisation implicite.
  - *Word Sense Induction / Discrimination*
    Une tâche non supervisée, recourant généralement à des méthodes de clustering. L'évaluation fait appel à des inventaires existants ou des pseudo-mots.
  - *Lexical Substitution*
    Etant donné un mot-cible apparaissant dans une phrase, le but est de proposer un ou plusieurs substituts qui n'altèrent pas le sens global de la phrase. Le choix du substitut est libre. Il est confronté aux données d'un gold généralement constituées par les réponses fournies par des annotateurs humains.

La substitution lexicale - Exemple

Trouver des substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place
écart

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place
écart
distance

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place
écart
distance
intervalle

La substitution lexicale - Exemple

Identifier les substituts pour le mot *espace* dans la phrase suivante, de manière à conserver le sens de la phrase :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place
écart
distance
intervalle
(...)

Intérêts de la tâche de substitution lexicale

- Impliquée dans de nombreuses applications en TAL (aide à la rédaction, simplification de texte, questions-réponses...)
- Offre une alternative aux modèles classiques de *Word Sense Disambiguation* :
  - Se passer de l'identification explicite et exhaustive des sens des mots : problème de couverture et de granularité
  - Se passer de la phase d'annotation sémantique de chaque sens des lemmes cibles
- Combine 2 opérations :
  1. Identification de substituts potentiellement similaires à la cible considérée hors contexte
  2. Sélection des substituts pertinents en fonction du contexte du mot-cible dans la phrase

Notre objectif avec la tâche compétitive de l'atelier SemDis2014 :

- Moins la phase de désambiguïsation en elle-même
- Que la procédure d'identification de mots similaires :
  - La tâche fournit une évaluation externe pour des modèles de sémantique distributionnelle
  - Et permet de les confronter à des modèles fondés sur des ressources lexicales

Quelques jalons

- Tâche originelle : English Lexical Substitution Task
  - Proposée aux ateliers SemEval-2007 (McCarthy & Navigli 2009)
  - 201 mots cibles (N, V, Adj, Adv), 10 phrases par cible
  - Annotation par 5 sujets
  - 8 systèmes participants
- Depuis :
  - Autres langues : allemand (Chokalov et al. 2014), (Miller et al. 2015), français (Fabre et al. 2014)
  - Changement d'échelle : crowd sourcing et substitution de tous les mots (Kremer et al. 2014) (30000 mots-cibles)

Le jeu d'évaluation – les mots

30 mots cibles : 10 N, 10 V, 10 A

- Mots polysémiques, fréquents et se prêtant à la substitution
- Critères de choix :
  - Présence dans *Le Robert*
  - Au moins 2 sens bien identifiables
  - Plus de 500 occ. dans le corpus frWaC
  - Des synonymes suffisamment nombreux et eux-mêmes fréquents

Mots sélectionnés

Noms
affection, capacité, couverture, débit, direction, don, espace, intérêt, montée, vaisseau

Verbes
arrêter, commander, entraîner, éplucher, essuyer, faucher, fonder, interpréter, maintenir, taper

Adjectifs
aisé, compris, grossier, hermétique, incorrect, mince, modeste, obscur, riche, vaseux

Le jeu d'évaluation – les phrases

300 phrases, 10 pour chaque mot cible

- Chaque phrase doit clairement illustrer un des sens du mot
- Elle doit être bien formée, pas trop longue
- Différents sens identifiés mais sans viser nécessairement l'équilibre

Exemples

sens	n°	phrase
tuer,	1	La guerre franco-prussienne **faucha** le jeune artiste à l'âge de 29 ans.
renverser	2	**Fauchée** par une voiture, une promeneuse de 57 ans décède sur le coup, sa belle-soeur est grièvement blessée.
	4	Sur une première offensive italienne, la France récupère le ballon et Zambrotta **fauche** Vieira.
couper	6	C'est pourquoi dans les marais, certaines parcelles sont **fauchées** tardivement l'été.
	7	Il y croit, même s'il reste sous le coup d'une condamnation à quatre mois de prison pour avoir **fauché** un champ de maïs transgénique en 2004.
voler	9	Louis XV est un mauvais roi parce qu'il s'est laissé **faucher** l'Inde et le Canada par les Anglais.
	10	On picolait un peu - une bouteille d'alcool **fauchée** chez Ceron.

L'annotation

Annotateurs francophones, étudiants et chercheurs en linguistique

- 7 annotateurs par phrase
- 3 substituts au maximum
- Extraits des consignes :
  - Votre tâche est de trouver des mots qui peuvent se substituer à ce mot en rouge tout en préservant au maximum le sens de la phrase
  - Vous pourrez proposer jusqu'à 3 substituts, mais si aucun ne vous vient à l'esprit n'insistez pas
  - Les mots simples sont à privilégier
  - La phrase résultante doit être correcte, mais des modifications syntaxiques légères sont tolérées :
    - *Le gros garçon s'amuse* / *Le garçon obèse*
    - *Paul a échoué dans sa tentative* / *Paul a râté sa tentative*

L'interface d'annotation

Outil de gestion d'enquêtes LimeSurvey

Groupe 3 / 6 (5 phrases à annoter)

L' expérience montre , que la montée en température de ce que l' on doit cuire dans le faitout s' opère en 4 phases

	Substituer le mot en rouge (ou laissez les champs vides si aucun substitut ne vous vient à l'esprit)
Proposition 1	
Proposition 2	
Proposition 3	

L' activité volcanique est en baisse singulière au cours du mois , sans aucune éruption phréatique significative , des débits de vapeurs à plus faible pression

	Substituer le mot en rouge (ou laissez les champs vides si aucun substitut ne vous vient à l'esprit)
Proposition 1	
Proposition 2	
Proposition 3	

Résultats

4014 propositions

- Moyenne par phrase :
  - 13 propositions
  - 7 substituts différents

Exemple (id = 208) :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
place (4), *distance* (3), *espacement* (2), *écart* (1), *d'écarts* (1), *de distance* (1), *volume* (1), *c'est plus étroit* (1)

Nettoyage des données

Filtrage, normalisation, lemmatisation

- Validation automatique de 88% des propositions
- Vérification manuelle des propositions restantes
  - Suppression des mot outils
  - Lemmatisation des formes ambigües
  - Exclusion de propositions mal formées
- Au final, 1771 substituts différents retenus
- *place* (4), *distance* (4), *espacement* (2), *écart* (2), *volume* (1)

Traitement des soumissions

- Verbes pronominaux
- Infinitif / Adj $\rightarrow$ ppé

Accord inter-annotateurs

Deux mesures de l'accord

accord par paire proportion moyenne de réponses identiques pour chaque phrase et chaque paire d'annotateurs

accord avec le mode proportion moyenne d'annotateurs qui ont inclus dans leurs réponses le mode = la réponse la plus fréquente (calculable pour 77% des phrases)

Tâche en français

- Accord par paire : 25,8%
- Accord avec le mode : 73%

Tâche en anglais

- Accord par paire : 27,75%
- Accord avec le mode : 50,67%

Participants

Règles du jeu

Chaque équipe participante pouvait soumettre jusqu'à 5 *runs* (ensemble de résultats)
Chaque *run* contient les réponses pour chaque phrase, maximum 10 propositions, la première proposition étant supposée être la meilleure

3 équipes, 9 *runs*, 1 baseline

- *Proxteam* (Yann Desalle, Emmanuel Navarro, Yannick Chudy, Pierre Magistry et Bruno Gaume) : 3 soumissions
- *CEA* (Olivier Ferret) : 5 soumissions
- *Alpage* (Kata Gábor) : 1 soumission

Proxteam

Principes de l'approche : balades aléatoires dans des graphes lexicaux

Ressources :

- base lexicale *Jeux de Mots*, relation *synonyme* et *idée associée*
- dictionnaire de synonymes *DicoSyn*
- corpus *Le Monde (1991-2000)* (analyse distributionnelle)

Variantes : combinaisons de ressources

- Proxteam_JDM_Syn : (*Jeux de Mots* + *Dicosyn*)
- Proxteam_AxeParaProx_JDM_Syn : idem, mais autre façon de combiner les deux ressources
- Proxteam_LM : corpus *Le Monde (1991-2000)*

CEA

Principes de l'approche : réseaux neuronaux (embeddings)

Liste de synonymes à partir d'un dictionnaire (dictionnaire de synonymes de Word XP, DicoSyn) ou d'une ressource distributionnelle (FreDist, à partir de Wikipedia et d'un corpus journalistique)
Sélection des synonymes les plus proches sémantiquement des mots pleins de la phrase

Variantes : ressources et mesures de similarité

cea_list-isc_l2_sent (euclidienne, Dicosyn, sélection contextuelle)
cea_list-isc_cos_sent (cosinus, Dicosyn, sélection contextuelle)
cea_list-isc_cos_w2 (cosinusn Dicosyn, sélection non contextuelle)
cea_list-fredist_cos_sent (cosinus, FreDist, sélection contextuelle)
cea_list-word_cos_sent (cosinus,Word XP,sélection contextuelle)

Alpage

Principes de l'approche : Wolf et analyse distributionnelle

Recherche des synonymes de la cible dans Wolf (WOrdnet Libre du Français), en sélectionnant le bon synset en se basant sur les mots de la phrase.
Si le mot n'est pas dans Wolf, on se rabat sur un thesaurus distributionnel (construit sur la Wikipedia).
1 seule variante

Baseline

Point de comparaison

Pas de sytème existant hérité d'une génération précédente : besoin de créer une *baseline*
Pour avoir une idée du gain en performance des approches plus sophistiquées des participants

Baseline sur le Robert des Synonymes

Principe : prendre pour chaque mot-cible les synonymes de plus haute fréquence (mesurée sur FrWac)
Réponse indépendante de la phrase

Exemple

espace → place, lieu, zone, course, champ, univers, distance, écart, ciel, surface

Mesures d'évaluation

Principes

On attribue un score à un substitut trouvé dans un *run* en fonction du nombre de juges qui ont proposé ce substitut (de 0 à 7).

Best et *OOT*

- *Best* : score de la première réponse du système (en gros, la précision)
- *OOT* (Out of Ten) : score cumulé sur les 10 réponses du système (en gros, le rappel)

On normalise chacun de ces scores par le nombre total de propositions des juges pour la phrase.

Détails du calcul

Best et *OOT*

Pour la phrase i, G_i est l'ensemble des substituts du Gold, $score_i()$ indiquant le nombre de juges qui les ont proposés P_i est l'ensemble des propositions du système dont $best_i$ est celle estimée être la meilleure.

- *Best* : score de la première réponse du système

$$best(i) = \frac{score_i(best_i)}{\sum_{a \in G_i} score_i(a)} \quad (1)$$

- *OOT* (Out of Ten) : score cumulé sur les 10 réponses du système

$$oot(i) = \frac{\sum_{a \in P_i} score_i(a)}{\sum_{a \in G_i} score_i(a)} \quad (2)$$

Exemple

Phrase 17

À l'endroit le plus ***mince*** *, sa largeur est de 3,35 mètres et son épaisseur de seulement 1,80 mètre .*

Gold : étroit (5), fin (4), petit (2)

Système X : petit ; étroit ; sec ; maigre ; léger ; délicat ; faible ; plat ; menu

Résultats

$\text{BEST} = 2/(5+4+2) = 2/11 = 0.18$

$\text{OOT} = (2+5+0+0+0+0+0+0+0)/11 = 0.64$

Remarques

OOT est forcément égal ou supérieur à Best

Best ne peut a priori pas atteindre la valeur 1

Les deux scores dépendent fortement de la dispersion des substituts du Gold

Résultats

Run	**best** ↓	**oot**
Proxteam_JDM_Syn	.097	.402
CEA_list-word_cos_sent	.075	.236
Proxteam_AxeParaProx_JDM_Syn	.065	.357
Alpage_WoDiS	.063	.205
Proxteam_LM	.051	.212
baseline	.045	.325
CEA_list-fredist_cos_sent	.040	.236
CEA_list-isc_cos_w2	.037	.284
CEA_list-isc_cos_sent	.033	.287
CEA_list-isc_l2_sent	.010	.231

Résultats par POS

		best			**oot**	
Run	*Nom*	*Adj.*	*Ver.*	*Nom*	*Adj.*	*Ver.*
Proxteam_JDM_Syn	**.110**	**.106**	.075	**.398**	**.429**	**.379**
CEA_list-word_cos_sent	.075	.074	.076	.195	.245	.268
Proxteam_AxeParaProx_JDM_Syn	.055	.054	**.087**	.311	.396	.363
Alpage_WoDiS	.054	.072	.061	.191	.211	.213
Proxteam_LM	.052	.040	.061	.233	.166	.237
baseline	.044	.040	.052	.294	.336	.344
CEA_list-fredist_cos_sent	.032	.028	.060	.181	.225	.303
CEA_list-isc_cos_w2	.030	.041	.041	.243	.281	.329
CEA_list-isc_cos_sent	.025	.034	.040	.233	.287	.340
CEA_list-isc_l2_sent	.004	.012	.015	.163	.230	.300
Moyenne	.048	.050	.057	.244	.281	.308

Vue d'ensemble

Questions de recherche

- Quels sont les items qui ont posé le plus de problème aux juges ?
- Quels sont les items qui ont posé le plus de problème aux systèmes ?
- Sont-ce les mêmes ?
- Quelles sont les caractéristiques des items faciles/difficiles (pour chaque cas) ?
- Peut-on prédire la difficulté de la tâche ?

Méthodologie

Mesure de la difficuté

- Pour les systèmes : le score OOT (moyenne sur les systèmes)
- Pour les humains : la dispersion des propositions de substitut (entropie normalisée)

Mesure des caractéristiques des items

- Ensemble de traits susceptibles d'influencer la difficulté
- 3 niveaux : mot, sens, phrase

Analyse

- Corrélation entre caractéristiques et mesures de la difficulté
- Modèle prédictif global

Dispersion des propositions des juges

Entropie normalisée

$H = -\sum_i p_i log(p_i)/log(N)$ Entre 0 (1 cas) et 1 (équiprobabilité)

Phrase "Facile" : un substitut évident, peu d'alternatives

*J'aime toucher et sentir la matière et ne pas laisser d'**espace** entre mes mains et ce que je crée.*

vide (7), *distance* (3), *place* (2), *interstice* (1), *intervalle* (1), *séparation* (1) → H=0,55

Phrase "Difficile" : pas de substitut évident, des propositions dispersées

*Notre vocation est de mettre en valeur vos **espaces** extérieurs, leur donner un style qui vous corresponde.*

lieu (2), *zone* (2), *emplacement* (1), *endroit* (1), *place* (1), *superficie* (1), *environnement* (1) → H=0,86

Scores des systèmes

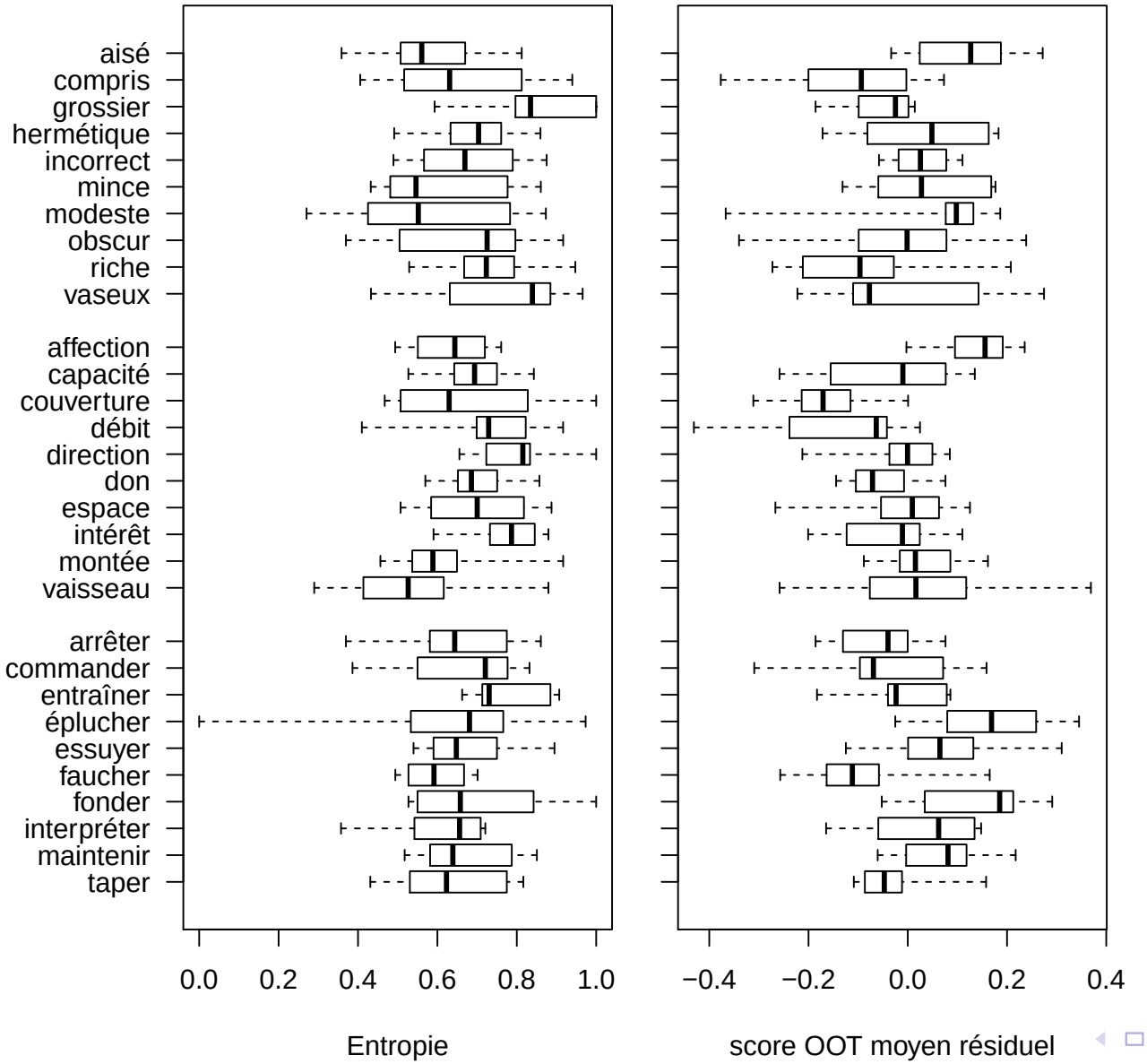
Mesure de la difficulté : OOT moyen sur les 9 systèmes participants

Problème : forte dépendance à la dispersion des réponses ($\rho = -0,58$ sur les 300 phrases)
Explication : pas de score élevé si peu de "bonnes" réponses possibles

Solution : les résidus

Principe : on fait une régression linéaire pour prédire OOT en fonction de l'entropie
On compare le score OOT réel et le score prédit, on prend la différence
Le résidu obtenu est la partie de la variation de OOT qui n'est pas "expliquée" par la dispersion des réponses

Variations


aisé
compris
grossier
hermétique
incorrect
mince
modeste
obscur
riche
vaseux
affection
capacité
couverture
débit
direction
don
espace
intérêt
montée
vaisseau
arrêter
commander
entraîner
éplucher
essuyer
faucher
fonder
interpréter
maintenir
taper
0.0
0.2
0.4
0.6
0.8
1.0
Entropie
−0.4
−0.2
0.0
0.2
0.4
score OOT moyen résiduel

Niveaux des caractéristiques linguistiques

3 niveaux

- Le mot-cible lui-même (N=30)
- Le sens dans lequel le mot-cible est utilisé dans la phrase (N=81)
- La phrase (N=300)

Annotation du sens

- Découpage proposé par le Larousse en ligne (supposé refléter la fréquence d'emploi)
- Double annotation + négociation + adjudication
- Annotation par le synonyme le plus net
  affection → /amour/, /maladie/
  vaseux → /boueux/, /fatigué/, /douteux/

Caractéristiques (1)

Niveau du mot

- Fréquence du mot-cible (dans GLAFF / FrWac, log) :
  min. 759 (*vaseux*) max. 427 900 (*espace*)
- Nombre de synonymes (renvois analogiques du Robert, cf. Dicosyn) :
  min. 5 (*vaseux*) max. 60 (*grossier*)
- Nombre de sens du mot-cible (cf supra) :
  min. 2 (12 mots-cibles) max. 5 (*couverture*)

Niveau du sens

- Rang du sens, normalisé entre 0 (+ fréquent) et 1 (+ rare)
- Nombre de synonymes du sens (après alignement avec DicoSyn) :
  min. 0 (10 cas, e.g. sens */diplôme/* de *capacité*) max. 43 (sens */énigmatique/* de *obscur*)

Caractéristiques (2)

Niveau de la phrase

- Fréquence moyenne des mots pleins de la phrase (log)
  2,5 : *La prolifération de bactéries et de champignons dans la bouche* ***entraîne*** *une coloration et une transformation des papilles gustatives.*
  6,7 : *Ce n'est plus l'État mais les marchés qui* ***commandent*** *l'économie...*
- Complexité syntaxique (distance moyenne entre dépendant et gouverneur)
  1,1 : ***Éplucher*** *et émincer finement les champignons, les arroser avec la moitié du jus de citron.*
  3,2 : *Et surtout, pour finir, quel homme,* ***obscur*** *ou célèbre, a jamais vaincu le tombeau ?*
- Position du mot dans la phrase (0=début, 1=fin)

Corrélations

ρ de Spearman, p $<$ 0,05

Mesure	**Facteurs significativement corrélés à la difficulté**
Entropie des annotations	rang du sens (0,33) fréquence moyenne des mots de la phrase (0,20) nombre de synonymes (0,19) fréquence du mot-cible (0,13)
Score *OOT* moyen résiduel	nombre de sens (0,37) nombre de synonymes (0,23) rang du sens (0,22) fréquence du mot-cible (0,12)

Corrélations, par POS du mot-cible

Mesure	**Adjectifs**	**Noms**	**Verbes**
Entropie	rang du sens (0,43) fréq. mot-cible (0,22)	fréq. mot-cible (0,38) nb. de sens (0,27) fréq. mots phrase (0,25)	rang du sens (0,39) fréq. mots phrase (0,26) fréq. mot-cible (0,23)
OOT	rang du sens (0,43) nb. de synonymes (0,27)	nombre de sens (0,40) complexité phrase (0,23) nb. de synonymes (0,22)	nombre de sens (0,52) nb. de synonymes (0,29) rang du sens (0,22) complexité phrase (0,21)

Conclusions

Des caractéristiques communes aux annotateurs et aux systèmes

- Les caractéristiques inhérentes au mot-cible (niveau de la forme et du sens) jouent un rôle central
- En particulier : difficulté plus grande pour les mots-cibles fréquents, et pour les sens rares (ex : *arrêter* $\rightarrow$ plus facile pour le sens /stopper/ que /décider/ ou /interpeller/)

Traitement différencié par POS

- Adjectifs semblent moins dépendants des facteurs contextuels
- Noms et Verbes : certaines caractéristiques des phrases ont un impact :
  - Les annotateurs ont plus de difficulté à traiter les phrases dont les mots sont très fréquents
  - Les systèmes sont sensibles à la complexité des phrases

Modélisation

Régression linéaire multiple à partir de toutes les caractéristiques

- OOT : $R^2 = 0,37$ (37% de la variance expliquée par les traits et leur interaction)
- Entropie : $R^2 = 0,19$
- Moralité :
  - de nombreux autres facteurs interviennent
  - le système est plus prédictible que l'humain
  - mais ce type de modèle est très complexe à interpréter...

Motivations de l'évaluation post-hoc

- Les sytèmes proposent-ils des candidats pertinents que les juges n'ont pas identifiés ?
- Question légitime étant donnée la tâche :
  - Tâche de production difficile pour les sujets, choix non tranchés, variation
  - Capacité des systèmes à collecter de nombreux substituts à partir de ressources textuelles ou lexicales

Première expérience menée par Mc Carthy & Navigli

Evaluation du décalage entre le *gold* et les propositions des systèmes

- Regroupement des réponses des sujets et des systèmes
- Réannotation par 3 sujets pour 100 phrases
- Des décalages observés, dans des proportions jugées minimes :
  - 18 % des réponses des systèmes seuls sont jugées correctes
  - 28 % des réponses des sujets seuls sont jugées mauvaises
- Pas de mesure de l'impact sur les performances des systèmes

Constitution du deuxième jeu de données

- Regroupement des réponses des sujets et des systèmes, pour l'ensemble des phrases

Sujets seuls	**Systèmes seuls**	**Sujets et systèmes**	**total**
788	11318	983	13089
6%	86.4%	7,5%	100%

- Annotation des 13.089 substituts par des juges
  - 3 à 7 juges par substitut (4,2 en moyenne)
  - substituts regroupés par blocs de 90
  - valeurs de 0 (mauvais) à 3 (bon)

Deuxième annotation avec LimeSurvey

Comment [interpréter] la montée des incivilités et de la délinquance qui met en cause le 'vouloir-vivre ensemble' ?

	3	2	1	0	Sans réponse
rendre	○	○	○	●	○
entonner	○	○	○	●	○
composer	○	○	○	●	○
entendre	○	●	○	○	○
exécuter	○	○	○	●	○
manifester	○	○	○	●	○
comprendre	●	○	○	○	○
deviner	○	○	○	●	○
translater	○	○	○	●	○
analyser	●	○	○	○	○
lire	○	○	○	●	○
apparaître	○	○	○	○	●
danser	○	○	○	○	●

Résultats

- 6.034 substituts (46%) ont un score positif
- score moyen : 0,51

Exemple (id = 208) :

Les sièges sont plus étroits, il y a moins d'**espace** entre les rangées.
Rappel : 5 substituts proposés par les sujets
Ici : 22 substituts reçoivent des scores positifs (en caractères romains, ceux issus du premier jeu)
distance (3), place (3), espacement (3), écart (2.75), *écartement* (2.5), *intervalle* (2.5), *éloignement* (2), *interstice* (2), *marge* (1), *surface* (1), volume (0.75), *étendue* (0.75), *ouverture* (0.75), *air* (0.5), [...], *zone* (0.25), *an* (0), *atelier* (0), *attribut* (0), [...], *blanc* (0), *centre* (0), [...], *interligne* (0), *jardin* (0), [...] *univers* (0), *visa* (0)

Accord inter-annotateurs

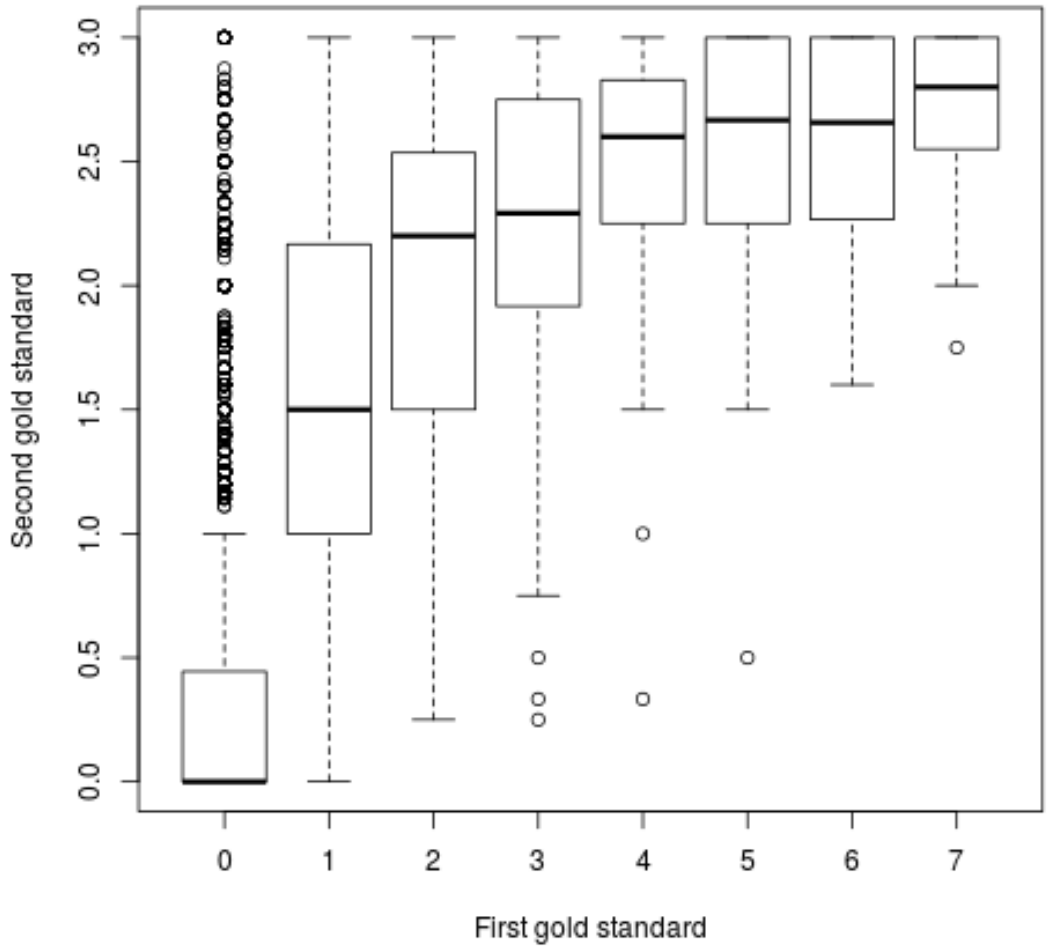
Nouveau mode de calcul du fait du score mesuré sur une échelle de 4 valeurs.

- **écart type moyen** : **0.38** $\pm$ 0.01, (IC 95%). L'écart type est calculé pour chaque substitut
- **taux de décision unanime** : **0.56**. 56% des substituts ont reçu le même score de la part de tous les juges.

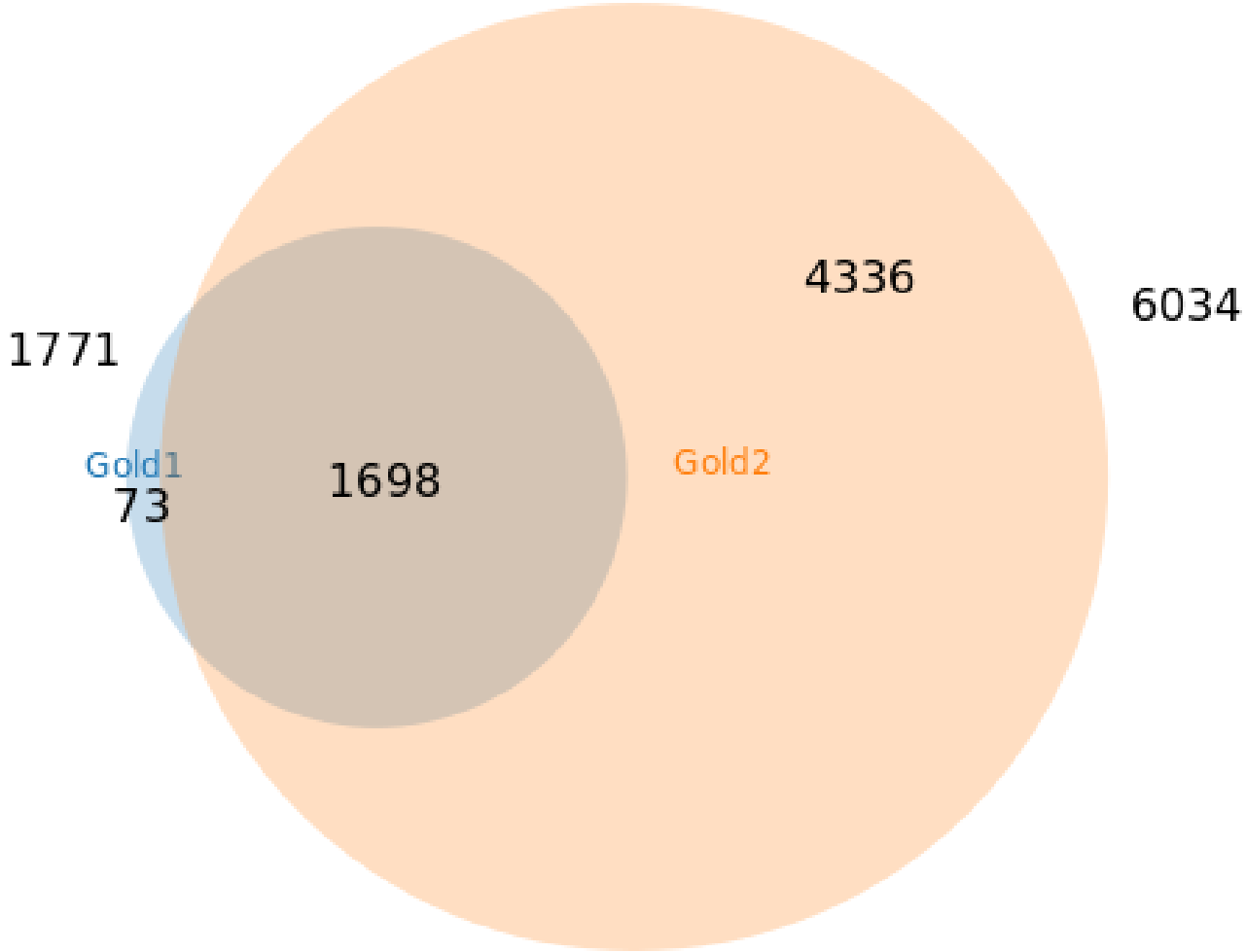
Comparaison des deux golds

Corrélation des scores

- Sur le premier gold, (N=1 771), $\rho = 0.42$.
- Sur le second gold, (N=13,089), $\rho = 0.61$ ($score_1$ de 0 pour les substituts absents du premier gold)



Recouvrement des deux golds



Différences

Substituts du premier gold rejetés (score moyen = 0)

73 cas sur 1 771 – tous proposés par un seul juge
Pas de schéma qui ressort, généralement une mauvaise interprétation du contexte

Nouvelles propositions

4 336 nouveaux substituts
Variantes morphologiques : *écartement* en plus de *écart* (pour *espace*), *directorat* en plus de *directeur* (pour *direction*), etc.
Mots rares ou plus recherchés : *sybillin* (*vaseux*), *malappris* et *malséant* (*incorrect*), *fangeux* et *estuarien* (*vaseux*), *amphigourique* (*obscur*), *férule* (*direction*), etc.
Proximité sémantique plus lâche : *progression* (pour *montée*)

Ré-évaluation des runs

Préalable

Dans le second gold, bien souvent plus de 10 substituts possibles pour chaque phrase

Les scores OOT et BEST baissent donc drastiquement pour certains items

Solution : normaliser les mesures

Diviser OOT et BEST par le meilleur score atteignable pour cet item

Nouveau classement (BEST)

System	BEST (1)	Rank	BEST (2)	Rank	Rank dif.
Prox_JDM_Syn	0.29	1	0.48	1	0
Prox_AxePara_JDM_Syn	0.20	3	0.37	2	-1
CEA_LIST-word_cos_sent	0.23	2	0.33	3	+1
Alpage_WoDiS	0.17	4	0.29	4	0
CEA_LIST-fredist_cos_sent	0.12	7	0.25	5	-2
CEA_LIST-isc_cos_w2	0.12	8	0.22	6	-2
Prox_LM	0.15	5	0.18	7	+2
CEA_LIST-isc_cos_sent	0.11	9	0.18	8	-1
Baseline	0.13	6	0.17	9	+3
CEA_LIST-isc_l2_sent	0.03	10	0.09	10	0

Nouveau classement (OOT)

System	OOT (1)	Rank	OOT (2)	Rank	Rank dif.
Prox_JDM_Syn	0.41	1	0.38	1	0
Prox_AxePara_JDM_Syn	0.37	2	0.35	2	0
CEA_LIST-isc_cos_sent	0.29	4	0.33	3	-1
CEA_LIST-isc_cos_w2	0.29	5	0.33	4	-1
Baseline	0.33	3	0.28	5	+2
CEA_LIST-fredist_cos_sent	0.24	6	0.23	6	0
CEA_LIST-isc_l2_sent	0.23	8	0.23	7	-1
Prox_LM	0.23	9	0.22	8	-1
CEA_LIST-word_cos_sent	0.24	7	0.19	9	+2
Alpage_WoDiS	0.22	10	0.19	10	0

Au-delà d'une simple tâche

Analyser a posteriori

- Permet de mieux comprendre la tâche
  - du côté des annotateurs
  - du côté des systèmes
  - pour identifier les différences entre les deux
- Permet de corriger les données
- Permet d'adapter la méthodologie pour une tâche future

L'analyse des difficultés en TAL

- Une problématique très répandue en RI depuis quelques années
- Plusieurs objectifs :
  - ne traiter que les cas difficiles (cf. TREC Hard Track)
  - identifier les approches complémentaires
  - faire de la fusion de systèmes ou des systèmes adaptatifs

Tout ça pour ça ?

La substitution lexicale :

Un tâche de TAL avec des labos qui participent et à la fin c'est Toulouse qui gagne.
(même si on change les règles du jeu)

Gold *a priori* ou *a posteriori* ?

Tout dépend :

- Si on cherche juste à évaluer, le gold *a priori* est moins coûteux
- Si on cherche à avoir un jeu de données fiable, réutilisable et qui permet d'y voir plus clair, mieux vaut faire de l'*a posteriori*.