Comment arpenter sans mètre ?

Les scores de résolution de chaînes de coréférences sont-ils des métriques ?


METHODE
POUR
LA MESURE DES SURFACES,
LA DIMENSION DES SOLIDES,
LEURS CENTRES DE PESANTEUR,
DE PERCUSSION ET D'OSCILLATION
Par l'Application du Calcul intégral.
Par M. CARRE' de l'Academie Royale des Sciences.
A PARIS,
Chez JEAN BOUDOT, Libraire de l'Academie Royale des Sciences, ruë S. Jacques au Soleil d'Or.
M. DCC.
AVEC PRIVILEGE DU ROI.

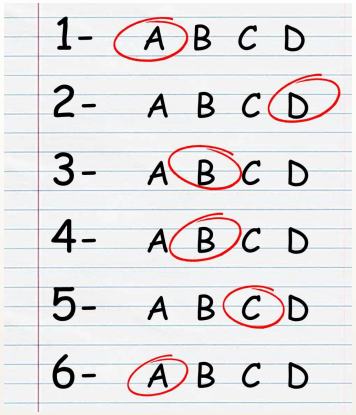
Adam LION-BOUTON, Jean-Yves ANTOINE
LIFAT, U. Tours, ICVL

Loïc GROBOL
LATTICE Montrouge

Sylvie BILLOT, Anaïs HALFTERMEYER
LIFO, U. Orléans, IVCL

Evaluation en TAL

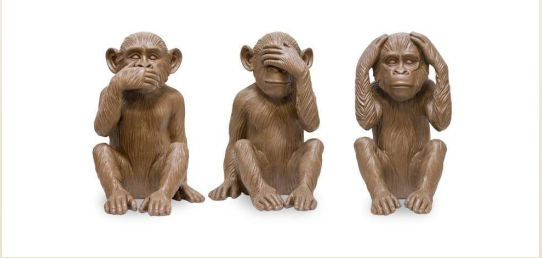
Ethique et déontologie scientifique – Nos méthodes d'évaluation orientent la conduite de nos recherches (modèle Poppérien de la réfutabilité par l'expérience)



Evaluation quantitative – Score de comparaison entre les sorties des systèmes et une référence idéale (GOLD standard)

Questions sur l'évaluation quantitative en TAL

- Qualité et représentativité des jeux de test
- Significativité statistique des résultats
- Interprétation en termes de qualité perçue
- Biais pouvant altérer les résultats



- **A quelle question répond l'évaluation ? Quel est le score le plus adapté ?**

Etude sur une tâche : résolution des chaînes de coréférence

Chaîne de coréférence

« Groupes de mots référant au même objet du discours »

Mention – Mot ou groupe de mots qui réfèrent à un élément du discours

Entité – Elément de l'univers du discours

Coréférence – Lorsque plusieurs mentions réfèrent à la même entité

J'suis bien chez moi mais j'comprends l'gitan quand il m'rappelle qu'il est libre

[Furax Barbarossa – Fin 2012]

J'suis bien chez moi mais j'comprends l'gitan quand il m'rappelle qu'il est libre

Document
Mention (pro)nominales

Furax

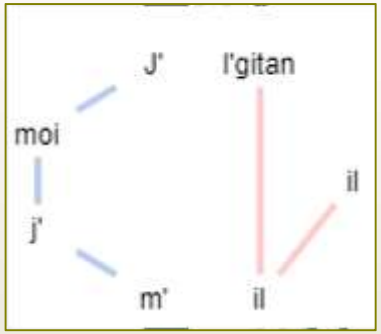
gitan

Discours
Entités

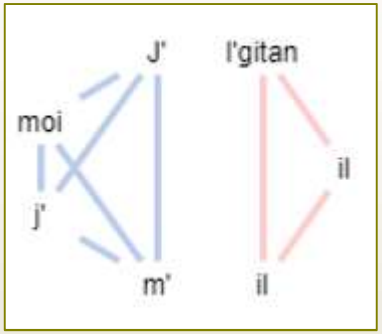
Chaîne de coréférence

Résolution des coréférences – Fournir une partition de l'ensemble des mentions détectées en termes de groupement par entités

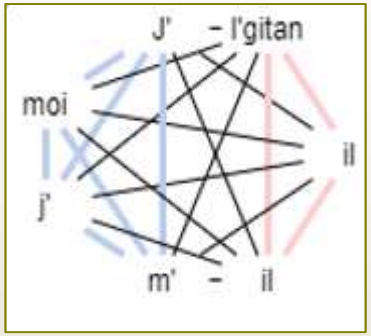
Plusieurs visions – Elément de l'univers du discours



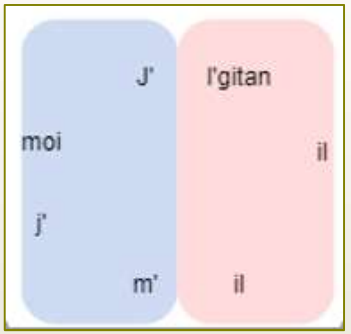
Séquence de liens (chaîne)



Ensemble des liens (relations)



Liens de coref et non coref



Ensembles de mentions

Link-centric model

Mention-entity model

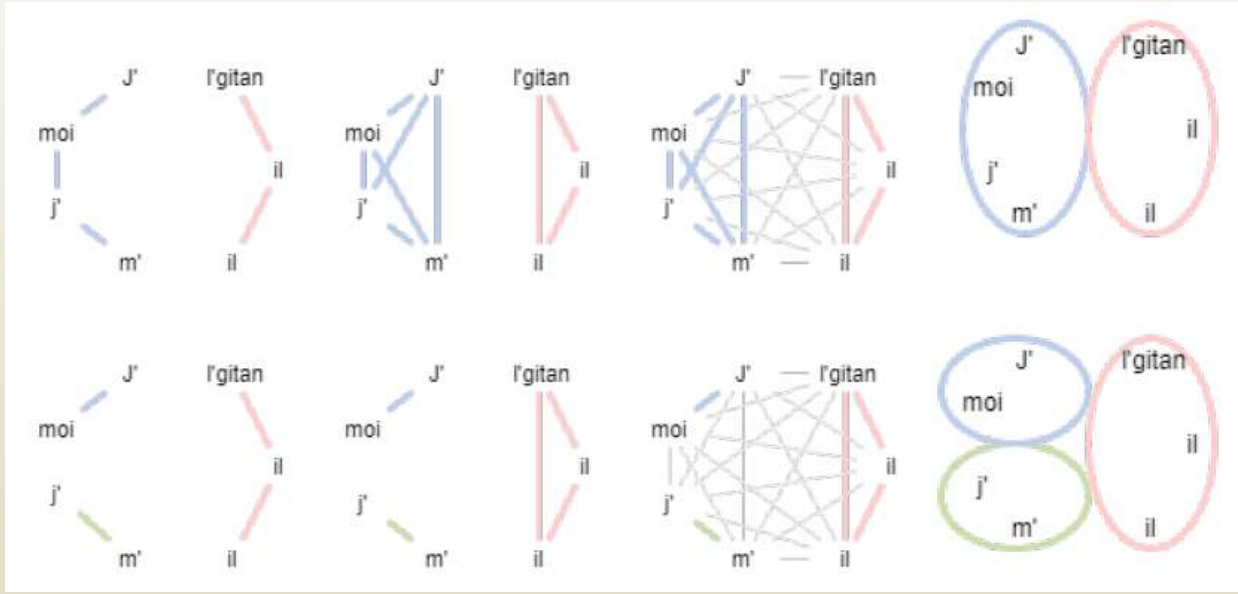
PhD Loïc Grobol
15 juillet 2020

Evaluation de la coréférence

Comparaison entre GOLD et système – Scores d'évaluation dépendent de la vision de la coréférence adoptée

GOLD J'$_1$ suis bien chez moi$_1$ mais j'$_1$ comprends l'gitan$_2$ quand il$_2$ m'$_1$ rappelle qu' il$_2$ est libre

SYS J'$_1$ suis bien chez moi$_1$ mais j'$_3$ comprends l'gitan$_2$ quand il$_2$ m'$_3$ rappelle qu' il$_2$ est libre

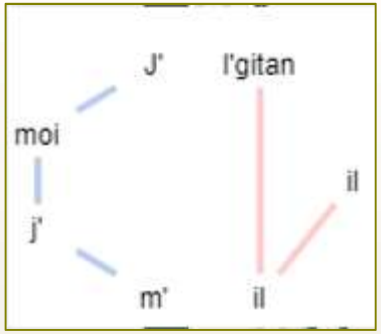


Différence

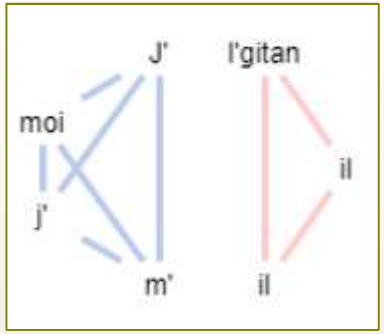
- 1 suppression de lien
- 3 suppressions de liens
- 2 déplacements de mentions

Evaluation de la coréférence

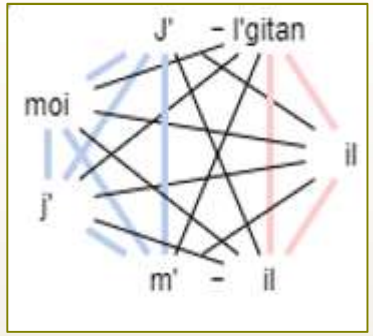
Multiplicité de scores, relevant de logiques différentes



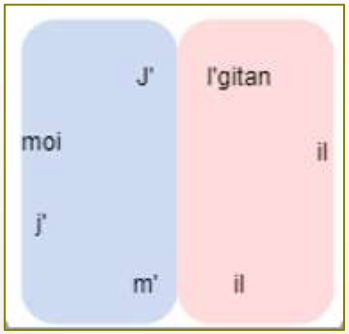
Séquence de liens (chaîne)



Ensemble des liens (relations)



Liens de coref et non coref



Ensembles de mentions

MUC
[Villain et al. 1998]

LEA
[Moosavi & Strube 2016]

BLANC
[Recasens & Hovy 2011]

$CEAF_m$
[Luo2005]

$CEAF_e$
[Luo2005]

B^3
[Bagga & Baldwin 1998]

MELA
[Denis & Baldridge 2009]

CoNLL
[Pradham & al. 2012]

Evaluation de la coréférence

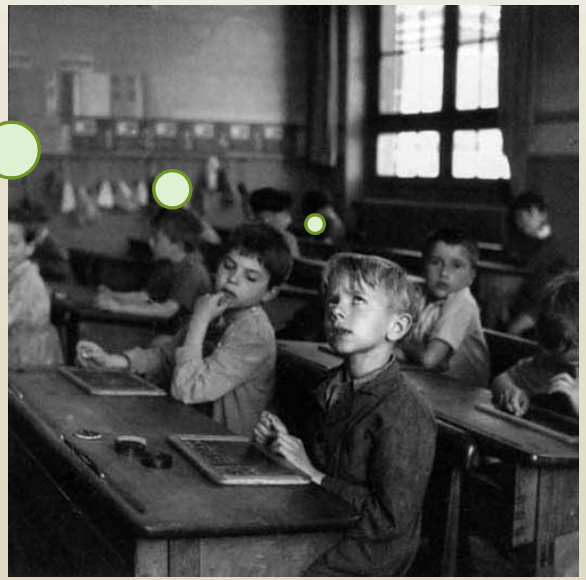
On évalue, mais on ne sait pas trop avec quoi…

- Biais pouvant affecter chaque score
- Lien entre scores objectifs et interprétation subjective
- Propriétés formelles de chaque score

Question

Les scores de résolution de chaînes de coréférences sont-ils des **métriques de similarité** ?

GOLD → SYS



Métrique de similarité normalisée

Idée – Métrique de **similarité** : pendant d'une métrique de **distance**

Définition formelle [Chen et al. 2009] – Fonction $s : X \times X \rightarrow \Re$ tel que :

$s(a,b) = s(b,a)$	(symétrie)	(1)
$s(a,a) \geq 0$		(2)
$s(a,b) \leq s(a,a)$		(3)
$s(a,b) + s(b,c) \leq s(b,b) + s(a,c)$	(inégalité triangulaire)	(4)
$s(a,a) = s(b,b) = s(a,b)$ si et seulement si $a = b$	(identité des indiscernables)	(5)

Métrique de similarité **normalisée** si de plus : $s(a,b) \leq 1$ (6)

***(1-s)* est de plus une métrique de distance normalisée** si la métrique *s* de similarité **normalisée** vérifie de plus :

$s(a,a) = 1$	(identité à 1)	(7)
$s(a,b) \geq 0$	(positivité)	(8)

Certaines propriétés formelles « naturelles » peuvent alors être observées :

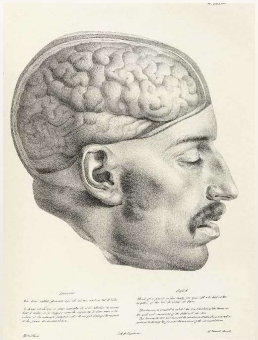
$s(a,a) = 1$	(équivalent à la propriété 7)	(9)
$s(a,b) \leq 1$	(équivalent à la propriété 6)	(10)
$s(a,b) + s(b,c) \leq 1 + s(a,c)$		(11)
$s(a,b) = 1 \Leftrightarrow a = b$		(12)

Métrique de similarité normalisée

Quel intérêt ?

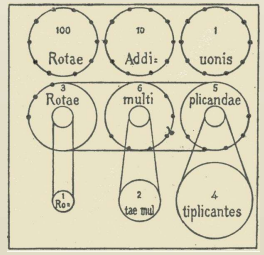
Cognitivement

- **Plausibilité cognitive** – Les jugements cognitifs de distance sont appréhendés dans des espaces métriques [Shepard 1962].
- **Intelligibilité** – L'inégalité triangulaire favorise l'intelligibilité des comparaisons d'estimations de similarité [Tversky & Gati 1982]



Pratiquement

- **Normalisation entre 0 et 1** – L'application d'une fonction convexe positive strictement croissante préserve le statut de métrique de similarité [Chen & al. 2009]
- **Métriques combinées (CoNLL)** – La somme et le produit préservent le statut de métrique de similarité [Chen & al. 2009]



Méthodologie

Recherche de falsification expérimentale et non formelle

Principes

- Jeux de tests basés sur les propriétés formelles
- Permet l'étude de fonctions de score non définies formellement (estimations statistiques)
- A force de preuve en cas de falsification d'une propriété

Protocole de test

- **Idée** : la résolution des coréférence revient à former des partitions
- **Jeux de tests systématiques** (toutes les partitions possibles) jusque *N=5* ou *N*=6 éléments suivant les propriétés: 40 000 à 140 000 partitions

$$s(a,b) \leq s(a,a)$$

Propriété 3

→ *N = 2*

$$s(\{\{1,2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1,2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1,2\}\},\{\{1,2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1,2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$
$$s(\{\{1\},\{2\}\},\{\{1\},\{2\}\}) \overset{?}{\leq} s(\{\{1\},\{2\}\},\{\{1\},\{2\}\})$$

Résultats

Situations de falsification

	1	2	3	4	5	6	7	8
MUC				X			X	
B^3				X				
$CEAF_m$								
$CEAF_e$				X				
CoNLL				X			X	
BLANC	X			X				
LEA				X				

Symétrie – BLANC ne respecte pas la symétrie: cas limites des entités uniques

$$BLANC(\{\{1,2,3\}\},\{\{1,2\},\{3\}\}) = 0{,}5 \neq 0{,}25 = BLANC(\{\{1,2\},\{3\}\},\{\{1,2,3\}\})$$

Identité à 1 – *s(a,a) = 1* n'est pas respecté par MUC (et donc CoNLL) si *a* est composé uniquement de singletons (MUC ne gère pas les singletons)

Inégalité triangulaire – Non respectée par tous sauf $CEAF_m$: cas de falsification

$a = \{\{1,2,3\},\{4,5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2\},\{3\},\{4\},\{5\}\}$ MUC,B^3,LEA

$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4\},\{5\}\}, c = \{\{1,2,3,5\},\{4\}\}$ $CEAF_e$, BLANC

$a = \{\{1,2,3,4\},\{5\}\}, b = \{\{1,2,3\},\{4,5\}\}, c = \{\{1\},\{2\},\{3\},\{4,5\}\}$ CoNLL

Conclusion

CEAFm est potentiellement la seule métrique de similarité normalisée

Poursuite des travaux sur d'autres aspects méthodologiques :

- Influence de biais potentiels (nombre d'entités, % de singleton…)
- Corrélation entre jugement humain et évaluation objective