Autour de la résolution automatique de la coréférence

définition de la tâche, modélisation cognitive, schémas Winograd

Pascal Amsili

avec le concours décisif de
Olga Seminck & Quentin Gliosca

Laboratoire de Linguistique Formelle,
Université Paris Diderot & CNRS

séminaire du CENTAL, octobre 2018

1. Résolution head-based *vs.* span-based
   - La tâche
   - Etat de l'art
   - Nouvelle proposition

2. Biais humains et statistiques en résolution de pronoms
   - Préférences
   - Méthode
   - Résultats

3. Des schémas Winograd français
   - Les schémas Winograd
   - Collection française
   - Robustesse statistique
   - Facilité pour les humains

- Pour une nouvelle définition de la tâche de résolution de coréférence
- nouvelle proposition,
- et premiers résultats

Résolution de co-référence

- Partitionnement des mentions de référents de discours en chaînes de coréférence

(1) Bill Clinton a prononcé un discours devant les sénateurs aujourd'hui. Le Président leur a présenté son nouveau projet de réforme du budget.

- {Bill Clinton, Le Président, son}, {les sénateurs, leur}

Deux sous-tâches

- identification des mentions
- partitionnement en chaînes de coréférence

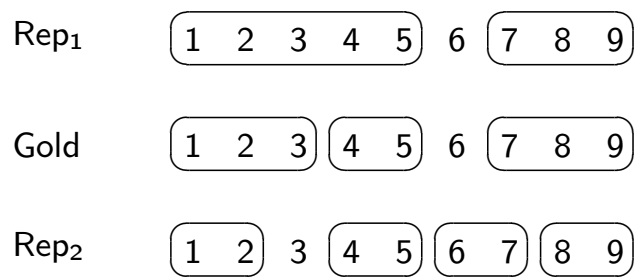
Evaluation de la tâche

- Corpus de référence (aussi pour l'apprentissage supervisé) : **Ontonotes**
  - Anglais, chinois et arabe
  - Des articles de journaux aux conversations téléphoniques
  - Singletons non annotés
  - Mentions : des syntagmes maximaux (sauf verbes)

(2) a. *One of [the two honorable guests] in [the studio] is [Professor Zhou Hanhua from the Institute of Law of the Chinese Academy of Social Sciences].*

b. *Hence, over here, [I] think [the municipal government, ah, including sub-district offices], [made] a lot of effort to take good care of resident households.*

- Métrique(s)
  - Détection des mentions : empans exacts : rappel & précision
  - Partitionnement : alignement de sous-ensembles



  - métriques spécifiques : MUC, B^3, $CEAF_e$, BLANC, LEA
  - usage : CoNLL

Modèle mention-mention classique

Structure générale des résolveurs par apprentissage depuis les années 2000

- Pré-traitement syntaxique pour identifier les mentions
- Pour chaque mention du document, évaluer les paires formées avec les mentions précédentes
- (plus une mention nulle pour les cas de première mention/singleton)
- Classer (*ranking*) les paires selon les descripteurs appris
- Choisir la mention co-référente la plus proche ou la meilleure
- Former les chaînes de coréférence par fermeture transitive

Descripteurs

Descripteurs unaires
Tête
Premier mot
Dernier mot
Mot précédent
Mot suivant
Longueur
Descripteurs binaires
Correspondance exacte
Correspondance des têtes
Distance en nombre de phrases
Distance en nombre de mentions

Etat de l'art - évaluation

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Table 1: Results on the test set on the English data from the CoNLL-2012 shared task. The final column (Avg. F1) is the main evaluation metric, computed by averaging the F1 of MUC, B^3, and $CEAF_{\phi_4}$. We improve state-of-the-art performance by 1.5 F1 for the single model and by 3.1 F1.

(Lee et al., 2017)

Etat de l'art - approche neuronale end-to-end

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Table 1: Results on the test set on the English data from the CoNLL-2012 shared task. The final column (Avg. F1) is the main evaluation metric, computed by averaging the F1 of MUC, B^3, and $CEAF_{\phi_4}$. We improve state-of-the-art performance by 1.5 F1 for the single model and by 3.1 F1.

(Lee et al., 2017)

Modèle de bout en bout (Lee et al., 2017)

- Tâches jointes : identification des mentions + coréférence

- chaque empan de phrase est une mention potentielle
- pas d'hypothèse sur les mentions, pas d'erreurs venant du pré-traitement

→ meilleure performance

→ phase d'analyse syntaxique inutile

Représentations des empans

- empans représentés par des vecteurs (*word embeddings*)
- chaque phrase du document est passée à un BLSTM
- un empan est représenté par les vecteurs BLSTM de ses bornes et sa longueur
- une paire est représentée par les représentations des deux empans,
- leur produit composante par composante et la distance entre les deux empans

Motivation

Analyse des erreurs de (Lee et al., 2017) :
Problèmes de frontières d'empans (hormis la longueur limitée à 10)

(3) I think in fact, we can sum up *many beneficial experiences* from this case and actually extend them to other areas. *(Gliosca, 2018)*

⇒ on essaie d'apprendre toute la syntaxe en ne considérant que les NP maximaux !

Réponses :

- faire une vraie tâche jointe parsing + coréférence → à explorer
- travailler avec les têtes des syntagmes

Interlude 1 : qu'est-ce qu'une mention ?

- mentions $\rightarrow$ référent de discours : objets abstraits
- plusieurs moyens de les identifier de façon univoque
- avant : syntagmes maximaux
- alternative : têtes
- équivalent à 99 % dans Ontonotes

Interlude 2 : applications avales (*downstream*)

Résumé automatique (Durrett et al., 2016) :

- sélection des phrases les plus importantes
- contraintes d'anaphoricité pour éviter les pronoms orphelins
- remplacement d'un pronom par son antécédent
- ou inclusion de la phrase de l'antécédent

Traduction automatique (Le Nagard and Koehn, 2010) :

(4) a. The window is open. It is black.
b. La fenêtre est ouverte. Il est noir.

- résolution des coréférences sur le document source
- pronoms factices porteurs du genre dans la langue cible

Syntagmes maximaux non nécessaires

Nouvelle définition

- mentions représentées par leurs têtes
- tout aussi pertinent pour les applications (voire plus)
- la syntaxe est laissée en dehors
- focalisation sur la résolution des coréférences
- évaluation basée sur les têtes

Principe :

- mention représentée uniquement par le vecteur BLSTM de sa tête
- plus de longueur des mentions
- plus de mécanisme d'attention

Pourquoi c'est possible ?

- le vecteur de la tête encode toute la phrase de la mention
- la similarité est calculée entre représentations apprises

Résultats

	Temps (s)	Max RAM (Mo)
Gliosca (2018)	**101**	**3181**
Lee et al. (2017)	240	6972

	Prec.	Rec.	F1
Gliosca (2018)	69.94	**69.80**	**69.87**
Lee et al. (2017)	71.12	65.62	68.25
Clark and Manning (2016)	**73.18**	63.23	67.83

Perspectives

- faire adopter la nouvelle définition de la tâche
- mécanisme d'attention à une fenêtre de mots
- apprentissage par renforcement (Clark and Manning, 2016)
- modèle mention-entité (Luo et al., 2004; Wiseman et al., 2016)

Perspectives

- faire adopter la nouvelle définition de la tâche
- mécanisme d'attention à une fenêtre de mots
- apprentissage par renforcement (Clark and Manning, 2016)
- modèle mention-entité (Luo et al., 2004; Wiseman et al., 2016)

Questions ?

Projet

- Biais humains étudiés en psycholinguistique :
  - Préférence pour le sujet
  - Parallélisme syntaxique
- Peut-on les découvrir en corpus ?

Préférences

- Préférence pour le sujet (en anglais)

(5) The postman met the streetsweeper before he$_?$ went home.
(Crawley et al., 1990; Hemforth et al., 2010)

- Explications : topicalité ? fréquence ? paramètre de langue ?

- Parallélisme syntaxique (Smyth, 1994)

(6) a. The postman met the streetsweeper before Lea met him$_?$.
b. The postman met the streetsweeper before he$_?$ met Lea.

- Problème de définition (même fonction ?, même structure ?)

Biais mis en évidence sur des items contrôlés

Méthode

- Construire par apprentissage un résolveur état-de-l'art
- de façon que le modèle appris soit interprétable (coefficients de régression)
- Confirmer (infirmer ?) le rôle des facteurs découvert par les psycholinguistes
- Evaluer numériquement leur interaction
- et le rôle d'autres facteurs éventuels

- algorithme de résolution mention-mention (aka pair-wise) :
- classification par régression logistique
- Ontonotes (Pradhan et al., 2011)
  - Seulement pronoms de 3^{e} personne
  - Annotation singletons
  - Annotation genre/nombre (automatiquement)
  - Exemples positifs et négatifs : (Soon et al., 2001)

Traits

	Estimate	Signif.
(Intercept)	-2.3533	***
match in gender	2.4206	***
match in number	0.2430	*
m_1 is a subject	1.5142	***
match in syntactic path	1.7318	***
m_1 is a proper noun	0.5007	***
m_1 is a possessive pronoun	1.9037	***
m_1 is a personal pronoun	0.7647	***
words between m_1 and m_2	-0.0114	***
m_1 & m_2 in the same sentence	0.3587	***
length of syntactic path m_1	-0.1361	***
m_1 is determined	-0.2825	*
m_1 is undetermined	-0.4422	**
m_1 has a demonstrative determiner	0.6045	*
m_1 is a common noun	-0.8967	***
m_1 spans m_2	-3.4372	***
length in words of m_1	-0.0201	*
m_1 is a geopolitical entity	-1.2885	***
m_1 is a date	-1.9416	***

(Seminck and Amsili, 2017, 2018)

Conclusion

- L'influence des deux biais étudiés est confirmée,
  et les coefficients permettent de quantifier leur interaction.
- Ces préférences humaines ont donc une base statistique
- Ça ne règle pas la question de l'origine/explication des préférences
  (priming *vs.* motivation cognitive)

Pespectives :

- Faire tourner le modèle sur des items expérimentaux ✓
- Utiliser le modèle pour prédire le temps de réaction ✓
- Trouver la "bonne" définition du parallélisme syntaxique
- Poursuivre la comparaison avec d'autres préférences

Conclusion

- L'influence des deux biais étudiés est confirmée,
  et les coefficients permettent de quantifier leur interaction.
- Ces préférences humaines ont donc une base statistique
- Ça ne règle pas la question de l'origine/explication des préférences
  (priming *vs.* motivation cognitive)

Pespectives :

- Faire tourner le modèle sur des items expérimentaux ✓
- Utiliser le modèle pour prédire le temps de réaction ✓
- Trouver la "bonne" définition du parallélisme syntaxique
- Poursuivre la comparaison avec d'autres préférences

Questions ?

Projet

- Créer une collection de schémas Winograd en français
- Vérifier empiriquement
  - qu'ils sont difficiles pour une machine
  - qu'ils sont faciles pour un humain

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?

R0 : Nicolas
R1 : son fils

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ? R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?
R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?
R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

(8) Nicolas n'a pas pu soulever son fils parce qu' il était trop lourd.
Qui était trop lourd ?

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?
R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

(8) Nicolas n'a pas pu soulever son fils parce qu' il était trop lourd.
Qui était trop lourd ?

- ... la bonne réponse change et elle est aussi évidente

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ?
R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

(8) Nicolas n'a pas pu soulever son fils parce qu' il était trop lourd.
Qui était trop lourd ?

- ... la bonne réponse change et elle est aussi évidente

$\Rightarrow$ indices linguistiques insuffisants

Définition (Levesque et al., 2012)

- une phrase avec un pronom ambigu...

(7) Nicolas n'a pas pu soulever son fils parce qu' il était trop faible.
Qui était trop faible ? R0 : Nicolas
R1 : son fils

- ... dont l'antécédent est évident pour un humain
- ... et dont il existe une variante obtenue en substituant un mot :

(8) Nicolas n'a pas pu soulever son fils parce qu' il était trop lourd.
Qui était trop lourd ?

- ... la bonne réponse change et elle est aussi évidente

$\Rightarrow$ indices linguistiques insuffisants

- Terminologie : 1 schéma = 2 items spe(cial) et alt(ernate)

Test d'intelligence artificielle

Alternative au test de Turing (simulation d'une conversation par une IA)

- besoin de raisonnement et de connaissances encyclopédiques
- résoud des problèmes liés au test de Turing :
  - *machine non humaine*
  - *détournement de la conversation* (Levesque et al., 2012)
- 2016 : premier Winograd Schema Challenge (Morgenstern et al., 2016)
  - 1er round : désambiguisation des pronoms (9)
  - 2e round : schémas Winograd

(9) Mrs. March gave the mother tea and gruel, while she dressed the little baby as tenderly as if it had been her own.

- 1er round : baseline (chance) : **42%**

→ meilleur système : Liu et al. (2016) : **58%**
(**66,7%** dans une version plus récente)

→ pas de 2e round

Motivation

Faire une collection pour le français

- Comparaison interlangues
- Benchmark pour les résolveurs en français

Deux expériences

- Vérifier systématiquement la *Google Proofness*
  - les schémas ne peuvent pas être résolus sans raisonnement et connaissances du monde
- Vérifier la facilité de résolution des humains

Problèmes d'adaptation (i)

- Des traits de genre et de nombre

(10) The drain is clogged with hair. It has to be ⟨cleaned/removed⟩.

- Traduction directe n'est pas possible, car *cheveux* est toujours au pluriel, alors que *siphon* est au singulier.
- Remplacé *cheveux* par *savon*.

(11) Il y a du savon dans le siphon de douche. Il faut le ⟨retirer/nettoyer⟩.

Problèmes d'adaptation (ii)

- Expression de but

(12) Mary tucked her daughter Anne into bed, so that **she** could ⟨sleep/work⟩.
Who is going to sleep ?

R0 : Mary
R1 : Anne

En français, les subordonnées finales sont de préférence à l'infinitif en cas de contrôle du sujet.

(13) a. Marie a couché sa fille$_i$ pour qu'elle$_i$ dorme.
b. *Marie$_i$ a couché sa fille pour qu'elle$_i$ dorme.
c. Marie a couché sa fille pour dormir.

⇒ Pas d'adaptation simple/directe de cet exemple en français.

Problèmes d'adaptation (iii)

- Difficultés lexicales

(14) Susan knows all about Ann's personal problems because **she** is ⟨nosy/indiscreet⟩.

'indiscreet' = *indiscrète* ? ?
Pas vraiment, puisque *indiscret* peut désigner :

– une personne qui révèle des choses qui devraient rester secrètes ; ou
– une personne qui essaie avec insistance de découvrir ce qui doit rester secret

Problèmes d'adaptation (iii)

- Difficultés lexicales

(14) Susan knows all about Ann's personal problems because **she** is ⟨nosy/indiscreet⟩.

'indiscreet' = *indiscrète* ? ?
Pas vraiment, puisque *indiscret* peut désigner :

– une personne qui révèle des choses qui devraient rester secrètes ; ou
– une personne qui essaie avec insistance de découvrir ce qui doit rester secret

→ a nosy person

Problèmes d'adaptation (iii)

- Difficultés lexicales

(14) Susan knows all about Ann's personal problems because **she** is ⟨nosy/indiscreet⟩.

'indiscreet' = *indiscrète* ? ?
Pas vraiment, puisque *indiscret* peut désigner :

- une personne qui révèle des choses qui devraient rester secrètes ; ou
- une personne qui essaie avec insistance de découvrir ce qui doit rester secret

→ a nosy person

Nous avons donc "traduit" (14) par ⟨bavarde⟩ *(talkative)*

(15) Sylvie est au courant de tous les problèmes personnels de Marie car **elle** est ⟨curieuse/bavarde⟩.

Produit

http://www.llf.cnrs.fr/winograd-fr

- 107 schémas en format xml,
- traduits/adaptés à partir du jeu original en anglais
- 2 stagiaires + 2 validateurs

```
<schema id="9" engn="46">
  <text>
    <txt1> Si l'escroc avait réussi à tromper Samuel, il aurait pu </txt1>
    <wordA>gagner</wordA>
    <wordB>perdre</wordB>
    <txt2> beaucoup d'argent. </txt2>
  </text>
  <question>
    <qn1>Qui aurait pu </qn1>
    <qwordA>gagner</qwordA>
    <qwordB>perdre</qwordB>
    <qn2> beaucoup d'argent ?</qn2>
  </question>
  <answer1>l'escroc</answer1>
  <answer2>Samuel</answer2>
</schema>
```

Google-proofness

- pas construction, les schémas ne peuvent pas être résolus sans l'intervention de raisonnement à propos de connaissances du monde

"... there should be no obvious statistical test over text corpora that will reliably disambiguate [the anaphor of a Winograd item] correctly."

(Levesque et al., 2012)

(16) Pendant la tempête, l'arbre est tombé et s'est écrasé sur le toit de ma maison. Maintenant je dois le ⟨déplacer/réparer⟩.

- Certains items de la collection anglaise ont été vérifiés de ce point de vue,
- mais nous voulions un test systématique applicable à toute la collection,
- d'où l'élaboration d'une mesure simple basée sur l'Information Mutuelle.

Information Mutuelle

- théorie de l'information (Shannon and Weaver, 1949)
- mesure la dépendance entre deux variables aléatoires
- peut mesurer la dépendance entre deux mots x et y (Church and Hanks, 1990)
- $MI(x, y)$ est positive si $P(x, y) > P(x) \times P(y)$

$$MI(x, y) = log_2 \left(\frac{P(x, y)}{P(x) \times P(y)} \right) \quad (1)$$

- comptages de fréquences (non-lissées) dans FrWaC (1.6 milliard de tokens du domaine `.fr`) (Baroni et al., 2009)
- fenêtre de voisinage : $-2 + 2$
- expressions multi-mots : tête lexicale
- lemmes plutôt que mots-formes (sauf cas particulier)

Illustration du calcul

(17) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/lourde⟩.

Illustration du calcul

(17) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/ ⟩.

Item Spe MI(sculpture, encombrer) = 4.23
MI(étagère, encombrer) = 10.01

Illustration du calcul

(17) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/ ⟩.

Item Spe

$\sout{MI(\text{sculpture}, \text{encombrer})} = 4.23$

$MI(\text{étagère}, \text{encombrer}) = 10.01$

Illustration du calcul

(17) La sculpture est tombée de l'étagère car elle était trop ⟨encombrée/ ⟩.

Item Spe

$\sout{MI(\text{sculpture}, \text{encombrer})} = 4.23$

$MI(\text{étagère}, \text{encombrer}) = 10.01$ ✓

Illustration du calcul

(17) La sculpture est tombée de l'étagère car [elle] était trop ⟨ /lourde⟩.

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23	
	MI(étagère, encombrer)	= 10.01	✓
Item Alt	*MI*(sculpture, lourd)	= 2.41	
	MI(étagère, lourd)	= 4.03	

Illustration du calcul

(17) La sculpture est tombée de l'étagère car [elle] était trop ⟨ /lourde⟩.

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23	
	MI(étagère, encombrer)	= 10.01	✓
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41	
	MI(étagère, lourd)	= 4.03	

Illustration du calcul

(17) La sculpture est tombée de l'étagère car [elle] était trop ⟨ /lourde⟩.

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23	
	MI(étagère, encombrer)	= 10.01	✓
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41	
	MI(étagère, lourd)	= 4.03	✗

Illustration du calcul

(17) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23	
	MI(étagère, encombrer)	= 10.01	✓
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41	
	MI(étagère, lourd)	= 4.03	✗

- Introduction d'un seuil de confiance (écart entre les deux valeurs)

Illustration du calcul

(17) La sculpture est tombée de l'étagère car [elle] était trop ⟨encombrée/lourde⟩.

Item Spe	~~*MI*(sculpture, encombrer)~~	= 4.23	
	MI(étagère, encombrer)	= 10.01	✓
Item Alt	~~*MI*(sculpture, lourd)~~	= 2.41	
	MI(étagère, lourd)	= 4.03	✗

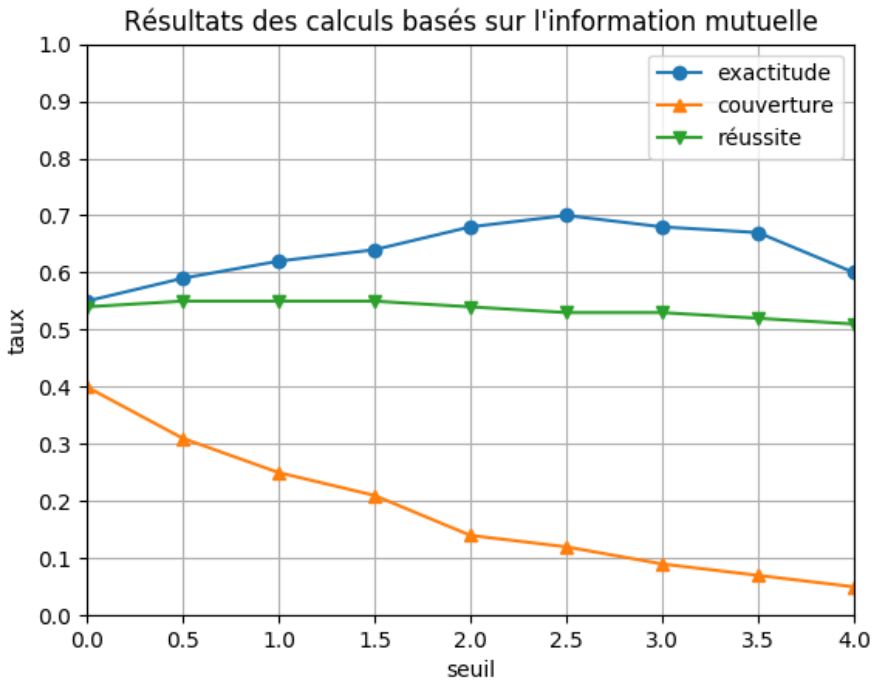
- Introduction d'un seuil de confiance (écart entre les deux valeurs)

- Au total, la méthode a pu être appliquée à 180 items (sur ≈ 220)
- parmi lesquels 131 ont pu recevoir un score d'information mutuelle (pas assez d'occurrences pour les items restants)

Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

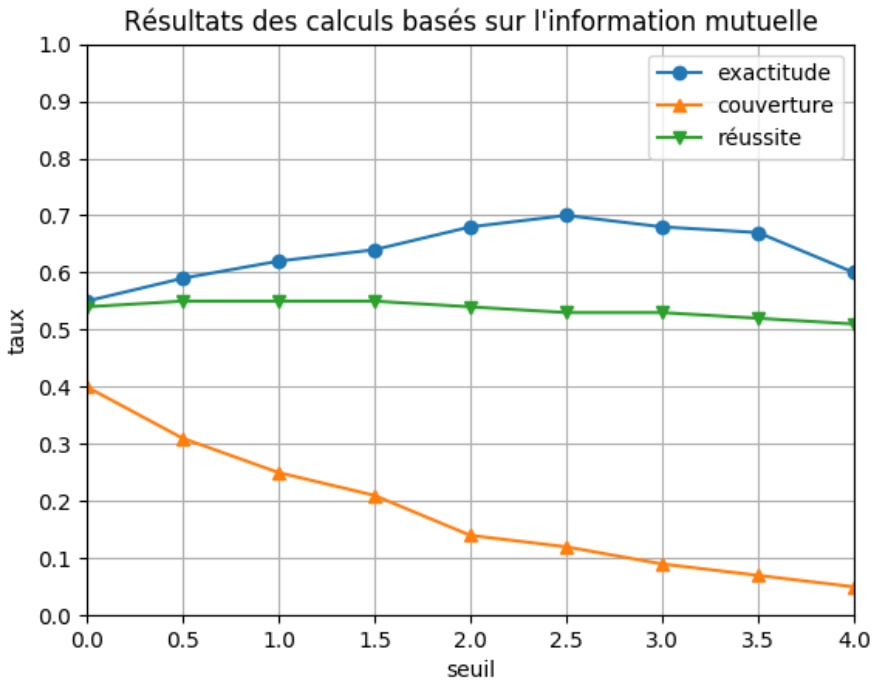
# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items



Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items

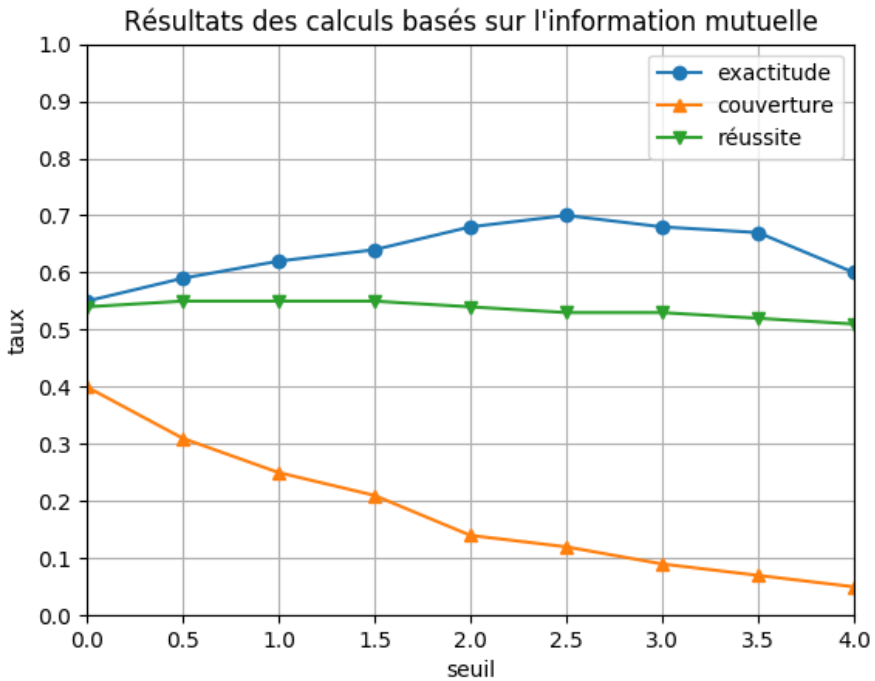


- répondre au hasard donne ≈ 50% de réussite

Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items

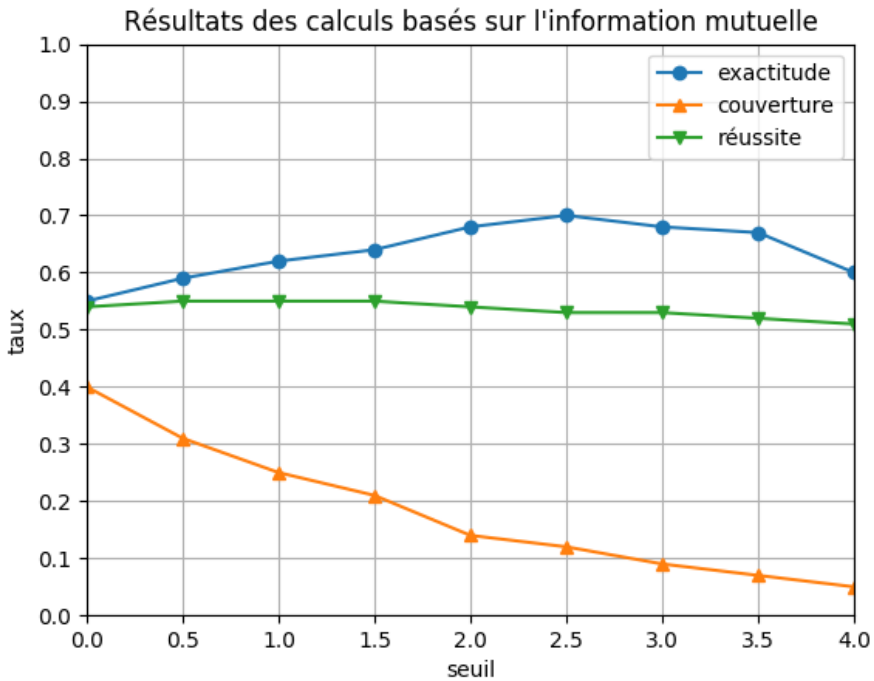


- répondre au hasard donne ≈ 50% de réussite
- le taux de réussite sans application de seuil reste bas (55%)

Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items

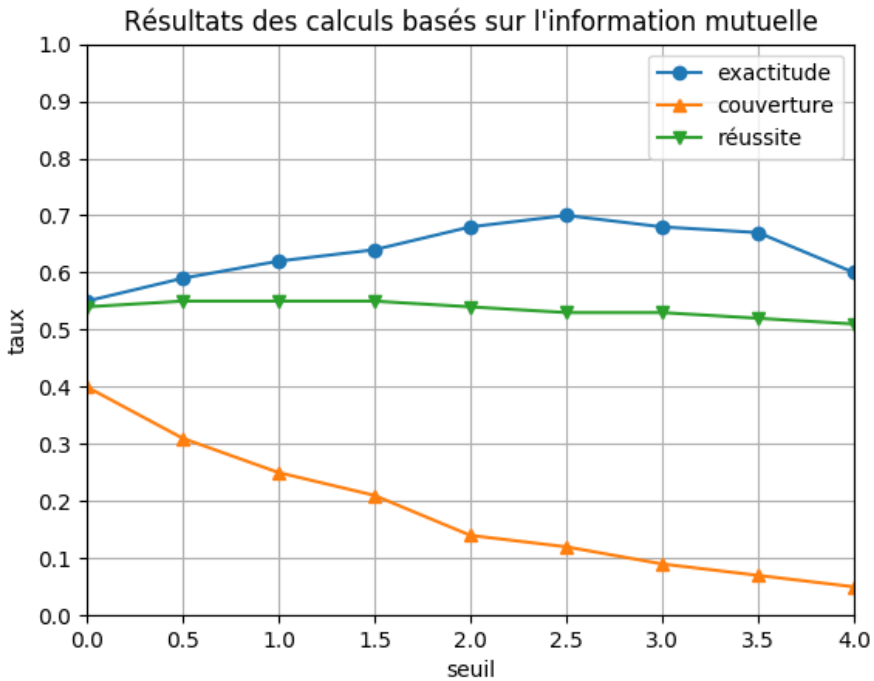


- répondre au hasard donne $\approx 50\%$ de réussite
- le taux de réussite sans application de seuil reste bas (55%)
- réussite dans 70% des cas avec Δ 2.5 **mais** $< 15\%$ des items.

Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items

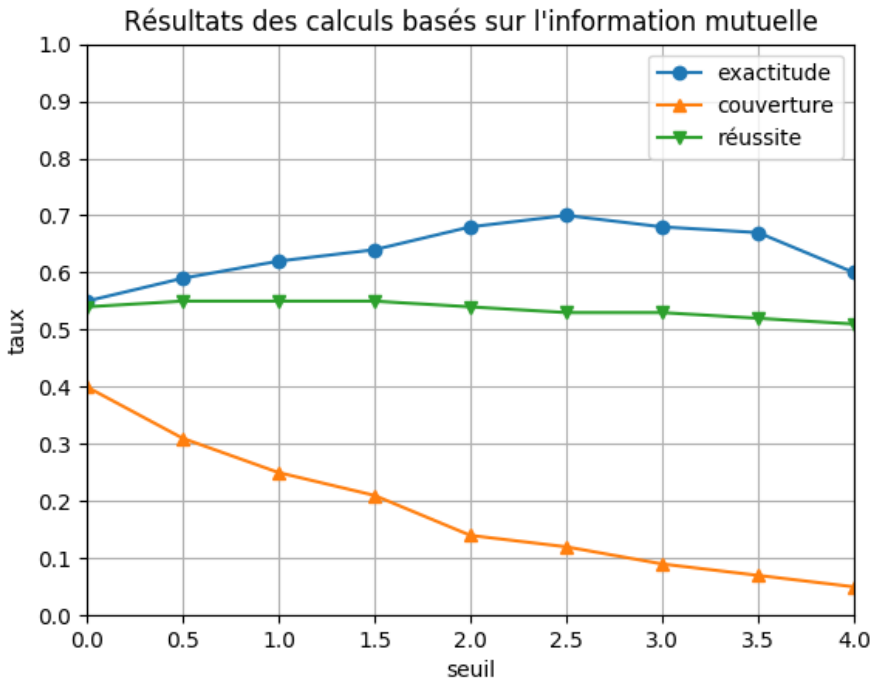


- répondre au hasard donne $\approx 50\%$ de réussite
- le taux de réussite sans application de seuil reste bas (55%)
- réussite dans 70% des cas avec Δ 2.5 **mais** $< 15\%$ des items.
- Dans son ensemble la collection est "Google-proof".

Résultats

Seuil	# Items	Exactitude	Couverture
None	131	0.55	0.40
Δ 0.5	95	0.59	0.31
Δ 1.0	73	0.62	0.25
Δ 1.5	59	0.64	0.21
Δ 2.0	38	0.68	0.14
Δ 2.5	30	0.70	0.12
Δ 3.0	25	0.68	0.09
Δ 3.5	18	0.67	0.07
Δ 4.0	15	0.60	0.05

# Items	nombre d'items auquels la méthode s'applique
Exactitude	taux de réussite sur les items ayant reçu une mesur
Coverage	pourcentage d'items ayant reçu une mesure
Réussite	réponse au hasard pour les autres items



- répondre au hasard donne ≈ 50% de réussite
- le taux de réussite sans application de seuil reste bas (55%)
- réussite dans 70% des cas avec Δ 2.5 **mais** < 15% des items.
- Dans son ensemble la collection est "Google-proof".
- Question ouverte : robustesse face à des méthodes plus sophistiquées

Facilité pour les humains

Taux de réussite pour les humains : 92% sur les schémas en anglais (Bender, 2015)

Expérience en ligne :

- 22 participants sur IbexFarm
- items randomisés
- réponses avec RT $> 1''$ et $< 60''$

Facilité pour les humains

Taux de réussite pour les humains : 92% sur les schémas en anglais (Bender, 2015)

Expérience en ligne :

- 22 participants sur IbexFarm
- items randomisés
- réponses avec RT $> 1''$ et $< 60''$

93,6 %

Facilité pour les humains

Taux de réussite pour les humains : 92% sur les schémas en anglais (Bender, 2015)

Expérience en ligne :

- 22 participants sur IbexFarm
- items randomisés
- réponses avec RT $> 1''$ et $< 60''$

93,6 %

OT

Type d'items difficiles

Ambigus :

(18) Pierre et Marc sont poursuivis pour diffamation. Pierre a écrit dans leur livre plusieurs faux témoignages que Marc a colportés. [Il] aurait dû être plus ⟨prudent/honnête⟩.

Réussite humaine : ⟨50%/18%⟩

Connecteur :

(19) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'[ils] venaient de plus loin.

Réussite humaine : ⟨70%/75%⟩

Complexes :

(20) Pierre jouait aux cartes avec Adam qui menait au score. Si la chance d'Adam n'avait pas tourné [il] aurait ⟨perdu/gagné⟩.

Réussite humaine : ⟨50%/67%⟩

Conclusion

- La collection dans son ensemble est Google Proof

Conclusion

- La collection dans son ensemble est Google Proof
- Il y a peu de schémas que les humains réussissent mal

Conclusion

- La collection dans son ensemble est Google Proof
- Il y a peu de schémas que les humains réussissent mal
- On peut envisager de “nettoyer” la collection

Conclusion

- La collection dans son ensemble est Google Proof
- Il y a peu de schémas que les humains réussissent mal
- On peut envisager de “nettoyer” la collection
- Perspective : comparaison interlangue, mais correspondance non directe

Conclusion

- La collection dans son ensemble est Google Proof
- Il y a peu de schémas que les humains réussissent mal
- On peut envisager de “nettoyer” la collection
- Perspective : comparaison interlangue, mais correspondance non directe
- Prochaine étape : plus de schémas, mais construits directement

Conclusion

- La collection dans son ensemble est Google Proof
- Il y a peu de schémas que les humains réussissent mal
- On peut envisager de “nettoyer” la collection
- Perspective : comparaison interlangue, mais correspondance non directe
- Prochaine étape : plus de schémas, mais construits directement
- Winograd Schema Challenge pour le français ?

Remerciements

Olga Seminck – Quentin Gliosca

Remerciements

Olga Seminck – Quentin Gliosca

Nous remercions nos stagiaires pour les items parce qu'on ne peut pas les ⟨oublier/résoudre⟩.

Sarah Ghumundee
Biljana Knežević
Nicolas Bénichou
Hugo Taquet
Dara Nguyen
Ryan Hunt

Yann Castellvi
Lara Perinetti
Léonard Fromond
Pierre-Louis Lugiery
Diana Khabarova

Labex « Empirical Foundations of Linguistics » (ANR-10- LABX-0083)
École Doctorale Frontières du Vivant — Programme Bettencourt

Références I

Bailey, D., Harrison, A., Lierler, Y., Lifschitz, V., and Michael, J. (2015). The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3) :209–226.

Bender, D. (2015). Establishing a human baseline for the winograd schema challenge. In *MAICS*, pages 39–45.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*.

Clark, K. and Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Crawley, R. A., Stevenson, R. J., and Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, 19(4) :245–264.

Davis, E., Morgenstern, L., and Ortiz, C. (2015). A collection of winograd schemas. Web page collecting 144 Winograd pairs, with comments and references.

Durrett, G., Berg-Kirkpatrick, T., and Klein, D. (2016). Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. In *Proceedings of the 54th Annual Meeting of the ACL*, volume 1, pages 1998–2008.

Emami, A., Trischler, A., Suleman, K., and Cheung, J. C. K. (2018). A generalized knowledge hunting framework for the winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 25–31. Association for Computational Linguistics.

Références II

Gliosca, Q. (2018). *Revisiter la résolution des coréférences*. Mémoire de m2 recherche, Université Paris Diderot.

Hemforth, B., Konieczny, L., Scheepers, C., Colonna, S., and Pynte, J. (2010). Language specific preferences in anaphor resolution : Exposure or gricean maxims ? In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 2218–2223, Portland, USA.

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation*, pages 252–261.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of EMNLP 2017*, pages 188–197.

Levesque, H. J., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Knowledge Representation and Reasoning Conference*, North America.

Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., and Hu, Y. (2016). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *arXiv preprint arXiv :1611.04146*.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A Mention-Synchronous Coreference Resolution Algorithm Based On the Bell Tree. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 135–142.

Morgenstern, L., Davis, E., and Ortiz Jr., C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1) :50–54.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). Conll-2011 shared task : Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, pages 1–27. Association for Computational Linguistics.

Schüller, P. (2014). Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Références III

Seminck, O. and Amsili, P. (2017). A computational model of human preferences for pronoun resolution. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 53–63. Association for Computational Linguistics.

Seminck, O. and Amsili, P. (2018). A gold anaphora annotation layer on an eye movement corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki (Japan). ELRA.

Shannon, C. E. and Weaver, W. (1949). The mathematical theory of information.

Sharma, A., Vo, N. H., Aditya, S., and Baral, C. (2015). Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence. AAAI*.

Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23(3) :197–229.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4) :521–544.

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP : A Balanced Corpus of Gendered Ambiguous Pronouns. *ArXiv e-prints*.

Wiseman, S., M. Rush, A., and M. Shieber, S. (2016). Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the NAACL*, pages 994–1004.

Exemples de schémas

(21) J'ai sorti le portable de mon sac pour qu'[il] soit ⟨plus accessible/moins lourd⟩. *(101)*

(22) Le frère jumeau de Joël arrive toujours à le battre au tennis, même s'[il] a suivi deux ans de cours en ⟨moins/plus⟩. *(99)*

(23) Sandrine a appris que le fils d'Anne avait eu un accident ⟨donc/car⟩ [elle] l'a prévenue. *(98)*

(24) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'[ils] venaient de plus loin. *(93)*

(25) Fred est le seul homme encore vivant à se rappeler de mon arrière grand-père. [C'] ⟨est/était⟩ un homme remarquable. *(25)*