Résolution des coréférences

P. Amsili

fall 2020

Sorbonne Nouvelle

Plan

Sorbonne Nouvelle

Plan

Plan

Tâche

Tâche :

- Identification des mentions
  → classification binaire d'empans textuels
- Partitionnement en chaînes
  → clustering

Tâche

Tâche :

- Identification des mentions
  → classification binaire d'empans textuels
- Partitionnement en chaînes
  → clustering

Evaluation séparée :

- Utilisation des mentions gold → trop facile
- Statut des mentions "singleton"

Tâche

Tâche :

- Identification des mentions
  → classification binaire d'empans textuels
- Partitionnement en chaînes
  → clustering

Evaluation séparée :

- Utilisation des mentions gold → trop facile
- Statut des mentions "singleton"

→ Évaluation commune → importance des chaînes de coréférence

Sorbonne Nouvelle

Prédiction des mentions

Actuellement

bornes exactes : 1
bornes différentes : 0

MUC

prédiction correcte

$$\Updownarrow$$

mention prédite $\ni$ tête

et

mention prédite $\subset$ mention gold

SemEval 2010

bornes exactes : 1
cas précédent (bornes $\neq$) : 0,5
sinon : 0

Sorbonne Nouvelle

Anaphores connues

On donne un nombre total d'anaphores *n* à résoudre. Le système attribue à chaque anaphore un *antécédent* qui est soit correct soit incorrect.

Anaphores connues

On donne un nombre total d'anaphores *n* à résoudre. Le système attribue à chaque anaphore un *antécédent* qui est soit correct soit incorrect.

- Frontières de l'antécédent : *idem* bornes des mentions
- Variantes : prise en compte de la chaîne de coréférence, et aussi du fait que la chaîne contient autre chose que des pronoms.

Anaphores à trouver

Double problème : trouver les anaphores, d'une part, et pour chaque anaphore, trouver l'antécédent.

	antécédent	anaphore
h (hit)	correct	réelle
i (incorrect)	incorrect	réelle
f (false alarm)	-	pas anaphorique
m (miss)	-	non trouvée

Evaluation

mesure **couples**

précision : $p = \frac{h}{h+i+f}$

rappel : $r = \frac{h}{h+i+m}$

Si on appelle $\mathcal{K}$ (*key*, aussi appelé *gold*) l'ensemble des couples de référence, et $\mathcal{R}$ (réponse) la liste produite par le système qu'on évalue, les calculs précédents reviennent à comparer les cardinaux des ensembles en question. On définit $h = |\mathcal{R} \cap \mathcal{K}|$ (c'est le nombre de couples de la réponse qui se trouvent dans le gold). Alors la précision est donnée par $\frac{h}{|\mathcal{R}|}$, et le rappel est donné par $\frac{h}{|\mathcal{K}|}$.

Variantes sur les frontières et les chaînes

Sorbonne Nouvelle

Co-références : partition

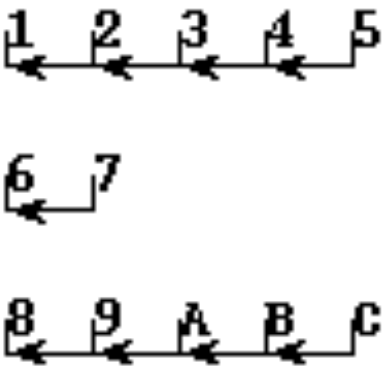

Figure 1: Truth

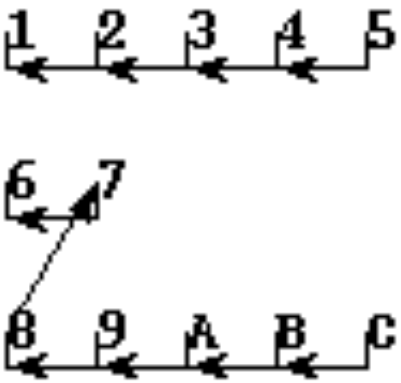

Figure 2: Response: Example 1

(Baldwin *et al.* , 1998)

MUC *link-based F-measure*

Comptage des liens dans $\mathcal{K}$ et dans $\mathcal{R}$
en considérant que la co-référence est une relation d'équivalence :
$\{(1,2), (1,3)\} = \{(1,2), (2,3)\} \rightarrow \{1, 2, 3\}$)
Principe : quel est le nombre **minimal** de liens à changer pour faire correspondre $\mathcal{R}$ et $\mathcal{K}$?
Le rappel est défini par $\frac{\text{nombre de liens trouvés (corrects)}}{\text{nombre de liens dans } \mathcal{K}}$, la précision par $\frac{\text{nombre de liens trouvés (corrects)}}{\text{nombre de liens dans } \mathcal{R}}$.

Sorbonne Nouvelle

Calcul 1

Soit $\Delta^{\mathcal{R},\mathcal{K}}$ le nombre total de liens "communs" entre $\mathcal{R}$ et $\mathcal{K}$:

$$\Delta^{\mathcal{R},\mathcal{K}} = \sum_{\gamma \in \mathcal{R},\, k \in \mathcal{K} : \gamma \cap k \neq \emptyset} (|\gamma \cap k| - 1)$$

$$R_{\text{muc}} = \frac{\Delta^{\mathcal{R},\mathcal{K}}}{\sum_{k \in G}(|k|-1)} \qquad P_{\text{muc}} = \frac{\Delta^{\mathcal{R},\mathcal{K}}}{\sum_{\gamma \in \mathcal{R}}(|\gamma|-1)}$$

Sorbonne Nouvelle

Calcul 2

pour chaque classe $\mathcal{K}_i$ de $\mathcal{K}$, on peut définir

- $p(\mathcal{K}_i, \mathcal{R})$: la *partition* de $\mathcal{K}_i$ par intersection avec les classes de $\mathcal{R}$
- $c(\mathcal{K}_i)$: le nombre de liens *corrects* de $\mathcal{K}_i$: en fait, $c(\mathcal{K}_i) = |\mathcal{K}_i| - 1$
- $m(\mathcal{K}_i, \mathcal{R})$: le nombre de liens *manquants* entre $\mathcal{K}_i$ et $\mathcal{R}$: $m(\mathcal{K}_i, \mathcal{R}) = |p(\mathcal{K}_i, \mathcal{R})| - 1$

Le rappel pour une classe $\mathcal{K}_i$ est donné par $\frac{c(\mathcal{K}_i) - m(\mathcal{K}_i, \mathcal{R})}{c(G_i)}$ (ce qui peut se simplifier). Le rappel pour l'ensemble $\mathcal{K}$ est obtenu en faisant le rapport entre la somme des liens corrects et la somme des liens dans $\mathcal{K}$.
La précision est obtenue en faisant le même calcul, mais en interchangeant les ensembles $\mathcal{K}$ et $\mathcal{R}$.

Sorbonne Nouvelle

B^3

Métrique proposée par Baldwin *et al.* (1998) pour corriger les défauts de MUC.
Pénalise plus les liens erronés qui regroupent des grosses chaînes.
Le rappel et la précision sont des moyennes de scores calculés respectivement pour chaque mention de référence et chaque mention prédite :

$$r = \frac{\sum_{K_i \in \mathcal{K}} \sum_{R_i \in \mathcal{R}} \frac{|K_i \cap R_i|^2}{|K_i|}}{\sum_{K_i \in \mathcal{K}} |K_i|} \qquad p = \frac{\sum_{R_i \in \mathcal{R}} \sum_{K_i \in \mathcal{K}} \frac{|K_i \cap R_i|^2}{|R_i|}}{\sum_{R_i \in \mathcal{R}} |R_i|}$$

Sorbonne Nouvelle

Critique

La métrique B^3 a pour défaut un comportement indésirable lorsqu'elle est utilisée pour évaluer une réponse fondée sur des mentions prédites, donc potentiellement inexactes. En effet, elle attribue systématiquement un rappel de 1 à une réponse dans laquelle toutes les mentions prédites sont ensemble, même s'il en manque, et une précision de 1 lorsque chaque mention prédite est dans sa propre chaîne de coréférence, même si toutes ne sont pas correctes.

$CEAF_e$

Basée sur un alignement préalable des entités.
Pour choisir l'alignement, on utilise une mesure ϕ de similarité entre deux chaînes de coréférence K_i et R_j donnée par :

$$\phi(K_i, R_j) = \frac{2|K_i \cap R_j|}{|K_i| + |R_j|}$$

(autres options possibles pour ϕ)

En cas de différence entre les nombres d'entités, on considère m entités ($m = min(|\mathcal{K}|, |\mathcal{R}|)$)

Sorbonne Nouvelle

Calcul de l'alignement

Appelons G_m l'ensemble de ces alignements et notons $\mathcal{D}_g$ le domaine de définition d'un alignement g. L'alignement optimal g^* pour la mesure de similarité ϕ est défini par

$$g^* = \underset{g \in G_m}{\text{argmax}} \sum_{K_i \in \mathcal{D}_g} \phi(K_i, g(K_i))$$

La complexité du calcul de cet alignement optimal par un algorithme naïf est factorielle en m, mais peut être ramenée en $O(Mm^2 log(m))$ par l'algorithme de Kuhn-Munkres de recherche du couplage de poids maximal avec $M = max(|\mathcal{K}|, |\mathcal{R}|)$.

Sorbonne Nouvelle

Mesure

Une fois l'alignement g^* déterminé, le rappel et la précision sont respectivement donnés par :

$$r = \frac{\sum_{K_i \in \mathcal{D}_g^*} \phi(K_i, g^*(K_i))}{\sum_{K_i \in \mathcal{K}} \phi(K_i, K_i)} \qquad p = \frac{\sum_{K_i \in \mathcal{D}_g^*} \phi(K_i, g^*(K_i))}{\sum_{R_i \in \mathcal{R}} \phi(R_i, R_i)}$$

Critiques

Le défaut le plus évident de $CEAF_e$ est d'ignorer complétement les chaînes de coréférence prédites qui ne participent pas à l'alignement optimal, alors qu'elles peuvent être partiellement correctes. Un système qui prédit une multitude de petites entités homogènes sera par exemple fortement pénalisé.

BLANC

BiLateral Assesment of Noun Phrase Coreference (Recasens & Hovy, 2011; Recasens *et al.* , 2013)
Réponse aux problèmes des mesures précédentes sur les singletons.
On se focalise sur les liens entre mentions : liens de coréférence (soit C_k l'ensemble des liens de coréférence selon K) et liens de non-coréférence (soit N_K l'ensemble des liens de non-coréférence selon K).

mesure : moyenne (arithmétique) des f-scores pour la détection de liens de (non-)coréférence

Mesure

Table 2.1: Definition of the BLANC metric

	Coreference	Non-coreference	Average
Precision	$P_C = \frac{\lvert C_R \cap C_K \rvert}{\lvert C_R \rvert}$	$P_N = \frac{\lvert N_R \cap N_K \rvert}{\lvert N_R \rvert}$	$\text{BLANC}_P = \frac{P_C + P_N}{2}$
Recall	$R_C = \frac{\lvert C_K \cap C_R \rvert}{\lvert C_K \rvert}$	$R_N = \frac{\lvert N_K \cap N_R \rvert}{\lvert N_K \rvert}$	$\text{BLANC}_R = \frac{R_C + R_N}{2}$
F_1	$F_C = \frac{2 \cdot P_C \cdot R_C}{P_C + R_C}$	$F_N = \frac{2 \cdot P_N \cdot R_N}{P_N + R_N}$	$\boxed{\text{BLANC} = \frac{F_C + F_N}{2}}$

(Grobol, 2020)

Sorbonne Nouvelle

Evaluation : synthèse

De nouvelles métriques sont régulièrement proposées,
mais les plus utilisées aujourd'hui sont MUC, B^3 et $CEAF_e$.
En particulier, la moyenne de ces trois métriques a été utilisée pour départager les systèmes lors des campagnes d'évaluation CoNLL 2011 et 2012.
Elle est depuis largement utilisée sous le nom de score CoNLL car elle permet d'attribuer un score unique aux systèmes de résolution des coréférences.
Seuls les travaux qui tentent de modifier le statut des singletons utilisent (en plus) la métrique BLANC.

Plan

Sorbonne Nouvelle

References

Baldwin, Breck, Morton, Tom, Bagga, Amit, Baldridge, Jason, Chandraseker, Raman, Dimitriadis, Alexis, Snyder, Kieran, & Wolska, Magdalena. 1998. Description of the University of Pennsylvania CAMP system as used for coreference. *In : Proceedings of the 7th Message Understanding Conference (MUC-7)*.

Ferreira Cruz, André, Rocha, Gil, & Lopes Cardoso, Henrique. 2020. Coreference Resolution : Toward End-to-End and Cross-Lingual Systems. *Information*, 11(2), 74.

Grobol, Loïc. 2020. *Coreference Resolution for spoken French*. Ph.D. thesis, Sorbonne Nouvelle, ED 622.

Levesque, Hector J, Davis, Ernest, & Morgenstern, Leora. 2012 (May). The Winograd schema challenge. *In : Knowledge Representation and Reasoning Conference*.

Recasens, Marta, & Hovy, Eduard. 2011. BLANC : Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, **17**(04), 485–510.

Recasens, Marta, de Marneffe, Marie-Catherine, & Potts, Christopher. 2013. The Life and Death of Discourse Entities : Identifying Singleton Mentions. *Pages 627–633 of : HLT-NAACL*.

Sorbonne Nouvelle