Objectif du TP implémenter quelques-unes des mesures de précision et de rappel que l'on peut définir quand il s'agit de comparer un ensemble de liens de coréférence calculés avec un ensemble de liens qui sert de référence.

Principe les *mentions* (ou *entités*) sont représentées par des entiers, et l'ensemble des liens de co-référence est représenté par une liste de couples (i, j) [1].
On appellera G (gold) la liste de référence, et R (réponse) la liste produite par le système qu'il s'agit d'évaluer.

Manipulations

1. Implémenter une mesure de précision et de rappel qui prend comme élément les couples (i, j).
2. Ecrire une fonction qui crée, à partir d'une liste de couples, un ensemble de *chaînes* (classes d'équivalence). On ajoutera systématiquement sous forme de singletons les mentions qui ne sont pas dans une relation anaphorique [2].

▷ *Pour la suite, on supposera que G et R sont des suites (ou des ensembles) de classes d'équivalence.*

3. Implémenter la mesure de MUC qui est basée sur le comptage du nombre de liens dans les chaînes.
4. Implémenter la mesure B^3 qui est basée sur le comptage des entités dans les chaînes de référence.
5. [Bonus] Implémenter la mesure CEAF présentée dans [Luo, 2005], basée sur un appariement des classes de G et des classes de R.
6. Utiliser vos différentes métriques pour comparer les performances des réponses R_1, R_2, R_3, R_4, étant donnée la référence G.
   $G = \{(1,2),(2,3),(3,4),(4,5),(6,7),(8,9),(9,10),(10,11),(11,12)\}$
   $R_1 = \{\}$
   $R_2 = \{(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,9),(9,10),(10,11),(11,12)\}$
   $R_3 = \{(1,2),(2,3),(3,4),(4,5),(6,7),(7,8)(8,9),(9,10),(10,11),(11,12)\}$
   $R_4 = \{(1,2),(2,3),(3,4),(4,5),(6,7),(5,8),(8,9),(9,10),(10,11),(11,12)\}$
   Fournir les scores de rappel, précision et F-score retournées par les trois métriques. Quelle métrique vous semble la plus adéquate, et pour quelles raisons ?

On fournira un script auto-suffisant qui réalisera dans l'ordre les manipulations demandées.

1. On supposera, sans perte de généralité, que $i < j$ (par convention, l'ordre croissant des mentions représente l'ordre linéaire du texte).
2. Le plus simple étant de supposer qu'on a un inventaire fixe de mentions, par exemple les entiers entre 0 et 20.

I. mesure couples

Soit r le nombre d'éléments de R qui sont dans G. Alors la précision est donnée par $\frac{r}{|R|}$, et le rappel est donné par $\frac{r}{|G|}$.

II. mesure MUC

Le rappel est défini par $\frac{\text{nombre de liens trouvés (corrects)}}{\text{nombre de liens dans } G}$, la précision par $\frac{\text{nombre de liens trouvés (corrects)}}{\text{nombre de liens prédits}}$.

Méthode vue en cours Soit $\Delta^{R,G}$ le nombre total de liens "communs" entre R et G :

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

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

Autre méthode pour chaque classe G_i de G, on peut définir
– $p(G_i, R)$: la *partition* de G_i par intersection avec les classes de R
– $c(G_i)$: le nombre de liens *corrects* de G_i : en fait, $c(G_i) = |G_i| - 1$
– $m(G_i, R)$: le nombre de liens *manquants* entre G_i et R : $m(G_i, R) = |p(G_i, R)| - 1$
Le rappel pour une classe G_i est donné par $\frac{c(G_i) - m(G_i,R)}{c(G_i)}$ (ce qui peut se simplifier). Le rappel pour l'ensemble G est obtenu en faisant le rapport entre la somme des liens corrects et la somme des liens dans G.
La précision est obtenue en faisant le même calcul, mais en interchangeant les ensembles G et R.

II. mesure B^3

Soit i une entité (mention), soit $ch_X(i)$ la chaîne à laquelle cette entité appartient dans la liste X
La précision pour l'entité i est définie par $\frac{\text{nombre d'entités correctes dans } ch_R(i)}{\text{nombre d'entités dans } ch_R(i)}$
où le nombre d'entités correctes dans $ch_R(i)$ est le nombre d'entités de $ch_R(i)$ qui sont dans $ch_G(i)$.
Le rappel, sans surprise, est défini comme $\frac{\text{nombre d'entités correctes dans } ch_R(i)}{\text{nombre d'entités dans } ch_G(i)}$
Le calcul global fait intervenir un poids pour chaque entité :

$$\text{prec} = \sum_i w_i \times prec(i)$$

On peut envisager différents systèmes de pondération, mais ici on se contentera de $w_i = \frac{1}{N}$ avec N= nombre d'entités total.

Références

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

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.