### 2.4.1 Rappels sur les mesures

#### 2.4.1.1 Exactitude

Tâche de **catégorisation** : On doit attribuer à chaque item une catégorie. Le nombre de catégories possibles n'est pas limité.

On définit la notion d'exactitude (accuracy/success rate) :

| | |
|---|---|
| $n$ | Nombre d'items (= nombre de prédictions) |
| $h$ | (*hits*) Nombre de valeurs correctement prédites |

$$\text{Accuracy} = \frac{h}{n}$$

#### 2.4.1.2 Précision/rappel

Pour une tâche de **repérage** : on me donne des items parmi lesquels je dois retrouver un type particuliers d'items.

On définit la paire de notions *précision* et *rappel*

| | |
|---|---|
| $N$ | Nombre total d'items traités |
| $n$ | Nombre d'éléments à trouver $n \leqslant N$ |
| $h$ | (*hits*) Nombre de catégorisations correctes $h \leqslant n$ |
| $f$ | (*false alarms*) Nombre de catégorisations fausses |
| $m$ | (*misses*) Nombre d'éléments non trouvés $m + h = n$ |
| $c$ | nombre total de catégorisations prononcées ($c = h + f$) |

La précision rend compte de la qualité de la prédiction quand elle est prononcée, mais ne tient pas compte des cas où on n'a pas fait de prédiction :

$$\text{Précision } p = \frac{h}{h+f} = \frac{h}{c}$$

Le rappel s'intéresse à la quantité d'éléments effectivement retrouvés, sans tenir compte du fait qu'on a pu en catégoriser beaucoup de façon erronée :

$$\text{Rappel } r = \frac{h}{m+h} = \frac{h}{n}$$

f-score : moyenne harmonique de la précision et du rappel :

$$f = 2 \times \frac{p.r}{p+r}$$

(moyenne harmonique *vs.* moyenne arithmétique : inverse de la moyenne arithmétique de l'inverse des termes)

### 2.4.2 Résolution d'anaphores

Ici on prend l'anaphore comme unité.

#### 2.4.2.1 Anaphores connues

Dans ce cas, on donne un nombre total d'anaphores $n$ à résoudre. Le système attribue à chaque anaphore un *antécédent* qui est soit correct soit incorrect.

$\Rightarrow$ dans ce cas c'est bien une mesure d'exactitude qu'il faut.

**Critiques**

- Manque de réalisme : la résolution anaphorique est aussi liée à la détection des anaphores.
- Non prise en compte des chaînes de coréférence.

**Variante** Avec prise en compte des chaînes de coréférence : on change la façon de décider qu'un antécédent est bon.

- Anaphore correcte : n'importe quel antécédent de la chaîne de coréférence du gold.
- Anaphore incorrecte : autre antécédent.

**Autre variante** Dans les campagnes d'évaluation, on a longtemps proposé une mesure encore légèrement différente : Pour chaque anaphore à résoudre, on donne l'une des notes suivantes :

**1** le pronom est correctement associé à une entité de la chaîne de coréférence. L'antécédent peut être pronominal lui-même, mais la chaîne doit contenir au moins une mention non pronominale (*ainsi on sait que le pronom en cours de résolution sera effectivement associé à une entité*).
**0.5** un pronom de la chaîne de coréférence du gold est choisi comme antécédent, mais ce pronom lui-même n'a pas été associé à une entité (ou n'est associé qu'à d'autres pronoms) (*pas de possibilité de relier le pronom à une entité*).
**0** aucune résolution dans la chaîne de coréférence.

#### 2.4.2.2 Anaphores à trouver

Dans ce cas, on a un double problème : trouver les anaphores, d'une part, et pour chaque anaphore, trouver l'antécédent.

Le résultat du calcul est une collection de paires $\langle$*antécédent*, *anaphore*$\rangle$. Il n'y a qu'un seul type de paire correcte (si on ne prend pas en compte les chaînes de coréférence), mais il y a deux types de paires incorrectes. Et il y a aussi des paires manquées.

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

On peut définir des notions de précision et de rappel, on parle de mesure **couples** parce qu'elle revient à prendre comme unité d'évaluation les couples.

La précision est donnée par $p = \frac{h}{h+i+f}$, et cette mesure agrège la performance dans la détection des anaphores (c'est-à-dire la capacité à limiter $f$) et la performance en résolution (c'est-à-dire la capacité à limiter $i$). Le rappel est donné par la formule $r = \frac{h}{h+i+m}$. Là encore, on agrège deux performances, celle qui concerne les anaphores, et celle qui concerne les antécédents).

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 $\frac{h}{|\mathcal{K}|}$.

On peut ajouter les variantes selon la prise en compte des chaînes de coréférence.

### 2.4.3 Résolution de coréférence

C'est plus délicat, puisqu'il faut envisager la **partition entière** (*i.e.* l'ensemble des chaînes) produites par le système et déterminer à quel point il y a correspondance avec la référence (*gold*). On peut illustrer le problème avec la figure 2.1, empruntée à Baldwin *et al.* (1998).

#### 2.4.3.1 Mesure « couples »

Cette mesure n'est pas utilisée pour la résolution de co-références.

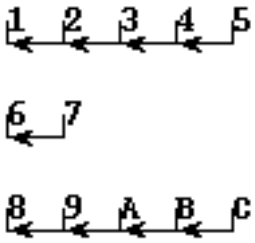


**Figure 1:** Truth

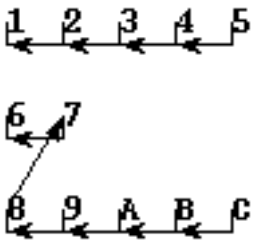


**Figure 2:** Response: Example 1

FIGURE 2.1 – Comparaison de chaînes de coréférences, from (Baldwin *et al.*, 1998)

#### 2.4.3.2 Métrique MUC

La métrique MUC est basée sur le comptage des liens (*link-based F-measure*) dans $\mathcal{K}$ et dans $\mathcal{R}$, mais en considérant que la co-référence est une relation d'équivalence : les couples {(1,2), (1,3)} et les couples {(1,2), (2,3)} dénotent la même classe d'équivalence (en l'occurrence {1, 2, 3}).

Principe de la mesure : on compare les classes de $\mathcal{R}$ et $\mathcal{K}$, et on cherche à calculer le nombre **minimal** de liens à changer pour faire correspondre $\mathcal{R}$ et $\mathcal{K}$. Par exemple, dans la réponse de la figure 2.1, il y a un lien en trop.

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

**Méthode de calcul 1** Soit $\Delta^{\mathcal{R},\mathcal{K}}$ le nombre total de liens "communs" entre $\mathcal{R}$ et $\mathcal{K}$ :

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

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

**Méthode de 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 $\mathcal{R}$.

Merci à Gliosca (2018) pour les paragraphes qui suivent.

#### 2.4.3.3 B³

**Définition** La métrique $B^3$ a été proposée par Baldwin *et al.* (1998) pour corriger les défauts de MUC. Elle permet en particulier de pénaliser plus fortement un lien erroné s'il connecte deux chaînes de coréférence de tailles importantes que deux chaînes plus petites.

Le rappel et la précision sont des moyennes de scores calculés respectivement pour chaque mention de référence et chaque mention prédite.

Le rappel $r$ est donné par :

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

Et la précision $p$ par :

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

**Critiques** 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.

#### 2.4.3.4 $\text{CEAF}_\text{e}$

**Définition** Introduite par Luo (2005), $\text{CEAF}_\text{e}$ aborde le problème de l'évaluation des systèmes de résolution des coréférences encore différemment puisqu'elle nécessite de déterminer au préalable un alignement entre les entités de référence et les entités prédites. Cet alignement est choisi au regard d'une mesure $\phi$ de similarité entre deux chaînes de coréférence $K_i$ et $R_j$ donnée par :

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

L'éventuelle différence entre le nombre d'entités de référence et le nombre d'entités prédites est gérée en considérant les alignements entre $m$ chaînes de coréférence de $\mathcal{K}$ et $m$ chaînes de $\mathcal{R}$ avec $m = min(|\mathcal{K}|, |\mathcal{R}|)$. 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 $\phi$ 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 $O(Mm^2log(m))$ par l'algorithme de Kuhn-Munkres de recherche du couplage de poids maximal avec $M = max(|\mathcal{K}|, |\mathcal{R}|)$.

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

et

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

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

**Critiques** Le défaut le plus évident de $\text{CEAF}_\text{e}$ 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.

De nouvelles métriques sont régulièrement proposées, mais les plus utilisées aujourd'hui sont MUC, $B^3$ et $\text{CEAF}_\text{e}$. 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. Voir cependant BLANC (*BiLateral Assesment of Noun Phrase Coreference* (Recasens et Hovy, 2011; Recasens *et al.*, 2013).