COÛT D'ALIGNEMENT ET DISTANCE DE LEVENSHTEÏN

Pour parler de coût d'alignement et par la suite de distance entre les mots, on se ramène à donner un coût aux opérations.

SUB On considérera que la substitution présente un coût qui ne dépend que des lettres que l'on échange, et non de la position concernée.
De plus on considère assez naturellement ce coût nul si on substitue a à a, et symétrique.
On le note $sub(a,b)$

DEL On considère le coût de cette opération constant, on le note del
INS ins

Rq Dans certains modèles il peut être intéressant de considérer qu'en début ou en fin, les suppressions et les insertions sont moins coûteuses, notamment lorsque les deux mots que l'on cherche à aligner sont de longueurs très différentes. Cf alignement ac traîches.

ex
x = ABRACADABRA
y = BRAS

Comme $|x| > |y|$ on "fait une promo" sur la suppression en début ou en fin de mot.

ainsi faire ABRA_CADABRA ou ABRACADABRAA_
_BRAS_____ _____BRAS

seront moins coûteuses que ABRACADABRA_
_B_____RAS.

Déf Le coût d'un alignement est la somme des coûts des opérations d'édition auxquelles il correspond, on le note $C(\tilde{x},\tilde{y})$.

Pté $$C(\tilde{x},\tilde{y}) = \sum_{i=1}^{|\tilde{x}|} c(\tilde{x}_i,\tilde{y}_i) \quad \text{où} \quad c(a,b) = \begin{cases} del & \text{si } b = \_ \\ ins & \text{si } a = \_ \\ sub(a,b) & \text{sinon} \end{cases}$$

ex
$del = ins = 1$
$sub(a,b) = \begin{cases} 1 & \text{si } a \neq b \\ 0 & \text{si } a = b \end{cases}$

$\tilde{x}$ = SALLE_
$\tilde{y}$ = SE_L_S

$C(\tilde{x},\tilde{y}) = 0+1+1+0+1+1 = 4$

Déf On introduit la "distance" de Levenshtein comme étant pour deux mots x et y.

$\text{Lev}(x,y) = \min \{ C(\tilde{x},\tilde{y}) \mid (\tilde{x},\tilde{y}) \text{ est un alignement de } (x,y) \}$

C'est le coût minimal d'un alignement entre x et y.
C'est aussi le coût minimal d'une succession d'opérations transformant x en y.

ex pour x = SALLE, y = SELS

$C\begin{pmatrix} \hat{x} = \text{SA L L E} \\ \tilde{y} = \text{SE \_ L S} \end{pmatrix} = 0+1+1+0+1 = 3 < 4.$

On ne peut pas faire moins donc Lev (SALLE, SELS) = 3

Pté Si del = ins > 0 et si sub est une distance sur Σ alors Lev est effectivement une distance sur Σ^*

Preuve
- sub en tant que distance est symétrique.
Puisque del = ins on est assuré que c est symétrique.
Par suite C et Lev sont symétriques.

- Si $\text{Lev}(x,y) = 0$, il existe $(\tilde{x},\tilde{y})$ un alignement de (x,y) tel que $C(\tilde{x},\tilde{y}) = 0$, alors $\forall i \in [\![1..|\tilde{x}|]\!]$ $c(\tilde{x}_i,\tilde{y}_i) = 0$
Comme del et ins sont > 0, cela implique $\tilde{x}_i \neq -$ et $\tilde{y}_i \neq -$.
Donc nécessairement $x = \hat{x}$ et $y = \hat{y}$.
De plus comme sub est une distance $\text{sub}(\tilde{x}_i,\tilde{y}_i) = 0$ ssi $\tilde{x}_i = \tilde{y}_i$.
D'où $\hat{x} = \hat{y}$. Finalement $x = y$.
D'où le caractère défini de Lev.

- Si on considère x, y, z trois mots, en enchaînant les opérations (dans) justifiant $\text{Lev}(x,y)$, et celles justifiant $\text{Lev}(y,z)$ on passe de x à z pour un coût $\text{Lev}(x,y) + \text{Lev}(y,z)$. Or $\text{Lev}(x,z)$ est minorée (définition).
D'où l'inégalité triangulaire.

Ainsi LEV est bien une distance.