Annotation automatique des adverbiaux temporels à l'aide d'Unitex

Travail à déposer sur Cours en Ligne
au plus tard le 30/01/2023 à minuit

Ce travail fait suite à l'exercice 9 du TD3 consacré à l'analyse automatique des adverbiaux temporels en français à l'aide de l'outil UNITEX. Dans un fichier .pdf, nommé **M1 TAL_*NOM1_NOM2***, vous fournirez (sur 8 pages maximum) :

A.

1. La biographie du journal Le Monde consacrée à de Villepin (fichier texteVillepin.txt) annotée manuellement (vous rendrez lisibles, à l'aide d'un code couleurs explicité, les choix de catégorisation retenus)
2. Une copie écran des principaux graphes définis sous Unitex (*cf*. Annexe I) ainsi qu'une copie écran de cette biographie annotée automatiquement à l'aide de vos graphes
3. Les mesures de rappel et de précision (*cf*. Annexe II) qui permettent de comparer les annotations de 1. et 2.

B.

4. Le texte journalistique consacré au mal-logement issu du journal pour enfants Le Ptit Libé (fichier texteL'immeuble d'Inès et d'Adam.txt) annoté manuellement (vous conserverez évidemment vos choix de catégorisation et de codes couleurs décidés en A.)
5. Une copie écran de ce texte annoté automatiquement à l'aide des mêmes graphes définis et utilisés en A.
6. Les mesures de rappel et de précision qui permettent de comparer les annotations de 4. et 5.

C.

7. Une présentation de la différence de votre approche de catégorisation avec celle retenue dans HeidelTime. Pour montrer cette différence, vous soumettrez à l'outil HeidelTime (accessible ici : http://heideltime.ifi.uni-heidelberg.de/heideltime/) le texte de A.

8. Une synthèse du travail présenté (incluant si besoin la présentation des difficultés sous Unitex rencontrées)

ANNEXES

Annexe I. Pistes pour la création des dictionnaires sous Unitex

Il s'agit ici de pistes. Vous pouvez tout à fait définir des dictionnaires différents selon les choix de catégorisation que vous aurez privilégiés au vu d'un premier travail manuel d'analyse du texte. Il est en tout cas clair que l'on va avoir besoin de plusieurs dictionnaires pour faire fonctionner les futurs automates. Tout d'abord, reprenons la liste des mots susceptibles d'apparaitre dans une date, comme dans les exemples [1] et [2].

[1] Ce mercredi 2 octobre 1872

[2] Le 29 septembre

Pour former une date, nous avons besoin de mots tels que le nom des jours de la semaine et le nom des mois de l'année. Ainsi, dans le dictionnaire jour.dic, nous trouverons les mots lundi, mardi, etc., et dans le fichier mois.dic, les mots janvier, février, etc. (cf. Figure 1). Dans les futurs transducteurs concernant les dates, nous trouverons donc ces dictionnaires appelés sous la forme <JOUR> et <MOIS>.

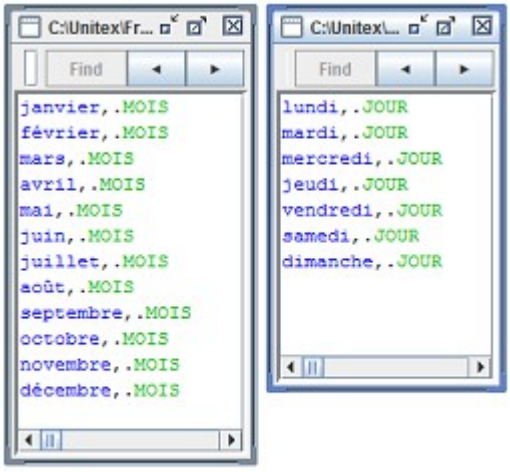

Figure 1 : Dictionnaires des noms de mois et de jours

On peut ensuite s'intéresser à créer les dictionnaires N2, N3, N4, N6 et N7 (cf. Figure 2) afin de récupérer certains noms de temps tels que « époque », « matinée » ou « hiver » dans les différents automates.

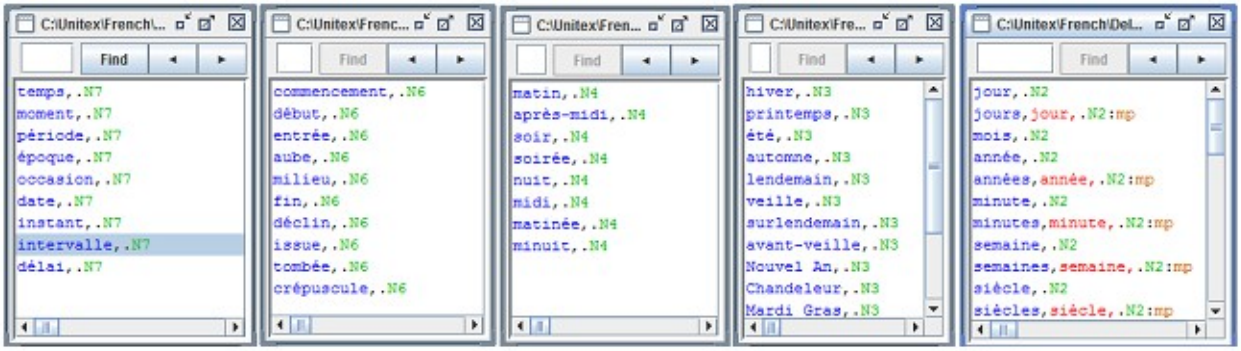

Figure 2 : Dictionnaires des noms de temps

On peut remarquer également dans le tableau des relevés que les adverbes apparaissaient souvent seuls et sans virgule. On pourra donc créer un dictionnaire pour eux, appelé ici Advt.dic (cf. Figure 3).

Il est cependant des adverbes qui peuvent être employés dans plusieurs domaines, et notamment dans le spatial autant que dans le temporel. Par exemple, nous avons fait le choix d'entrer dans la liste ci-après l'adverbe « là-dessus », en pensant au sens de « après cela, sur ces paroles ». Or cet adverbe peut également s'utiliser dans le sens spatial de « sur ». Nous supposons que cet emploi spatial est moins fréquent dans la littérature, car plus familier.

Figure 3 : Dictionnaire des adverbes de temps

Annexe II. Méthodes d'évaluation

Les **mesures d'évaluation** les plus fréquemment utilisées pour les systèmes de TAL sont le rappel et la précision. Ces mesures sont issues du domaine de la recherche d'information et se

sont largement répandues dans l'évaluation des systèmes de TAL (extraction d'information, analyse syntaxique automatique,…). Ces deux notions sont souvent utilisées, car elles reflètent le point de vue de l'utilisateur : si la précision est faible, l'utilisateur sera insatisfait, car il devra perdre du temps à lire des informations qui ne l'intéressent pas. Si le rappel est faible, l'utilisateur n'aura pas accès à une information qu'il souhaitait avoir.

De manière générale, le rappel mesure le silence du système, c'est-à-dire les informations pertinentes qui n'ont pas été trouvées, tandis que la précision en mesure le bruit, c'est-à-dire les informations non pertinentes trouvées. Plus le taux de rappel est haut, moins il y a de silence, plus la précision est élevée, moins il y a de bruit. Les formules ci-après indiquent comment calculer ces taux.

Rappel = nombre d'informations bien repérées / nombre d'informations à repérer $0 \leq \text{Rappel} \leq 1$

Précision = nombre d'informations bien repérées / nombre d'informations repérées $0 \leq \text{Précision} \leq 1$