Linguistique de corpus

Outils et méthodes de traitement de corpus

Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

mercredi 08 mars 2017 / L3 - Cours S2-5

Pratique Unitex
Normalisation

Suite des travaux pour normalisation du roman “Le tour du Monde en 80 jours”. Lorsque le document est importé dans Unitex, on s’apperçoit que les énoncés de dialogue ne sont pas segmentés commme des énoncés. L’objectif est donc d’ajouter un filtre de normalisation qui identifie les dialogues dans et ajoute un frontière de phrase {S} a la fin de chaque énoncé en créant un automate transducteur.

1. dans un premier temps sans ses soucier des balises d’énoncé existantes
2. en prenant compte les balises d’énoncé existantes, lorsqu’elles sont présente au milieu d’un énoncé

Rappels sur les automates
Définition

Un automate d'états finis déterministe est défini par :

1. un ensemble fini d'états
2. un ensemble fini de transitions (arcs reliants les états)
3. un alphabet d'entrée (liste finie de symboles d'entrée)
4. un état initial
5. un ou plusieurs états finaux.

L'automate est déterministe si dans un état donné, pour un symbole d'entrée donné, on ne peut emprunter qu'une seule transition.

Rappels sur les automates
Implémentation

L'implémentation d'un automate aura donc toujours :

- une bande de lecture unidimensionnelle, contenant dans chaque cellule un seul symbol pris dans l'alphabet d'entrée.
- une tête de lecture associée à une position unique sur la bande de lecture
- une mémoire contenant l'unique état interne courant de l'automate
- la fonction d'un automate d'états finis déterministe est de reconnaître si le texte stocké sur la bande de lecture fait parti du langage défini par la structure de l'automate.

Rappels sur les automates
Fonctionnement

- l'automate démarre la reconnaissance dans l'état initial avec la tête de lecture positionnée sur la première case de la bande et son état interne est l'unique état initial,
- pour une position de la tête de lecture, il ne peut y avoir qu'un transition d'état déterminée par le symbole contenu dans la cellule de la bande pointée par la tête de lecture, l'état interne change alors selon la transition et la tête de lecture se déplace vers la cellule suivante dans le sens de lecture de la bande.

Rappels sur les automates
Fonctionnement

- un texte est reconnu lorsque l'automate a pour état interne un des états finaux et que l'on a lu tout le texte de la bande de lecture (la tête de lecture se trouve à la fin, après la dernière cellule), on a reconnu le texte comme faisant partie du langage, l'automate produit donc la réponse OUI/VRAI/RECONNU.
- un texte n'est pas reconnu lorsque l'automate est BLOQUÉ (aucune transition n'est applicable) et que soit l'état interne n'est pas un des états finaux, soit que l'on n'est pas à la fin de la bande de lecture (il reste du texte à reconnaître), alors le texte n'est PAS reconnu comme faisant partie du langage, l'automate produit donc la réponse NON/FAUX/NON-RECONNU.

Rappels sur les transducteurs
Définition

- Un transducteur est un Automate d'états fini déterministe...avec en plus
- un alphabet de sortie
- des transition étendues, qui contiennent en plus de l'état courant, du symbole courant et de l'état de destination, un SYMBOLE DE SORTIE pris dans l'alphabet de sortie

Rappels sur les transducteurs
Implémentation

- par rapport à un automate d'états finis déterministe un transducteur aura donc en plus une bande de sortie unidimensionnelle, faite de cellules contenant chacune un caractère de l'alphabet de sortie
- et une tête d'écriture sur la bande de sortie

Rappels sur les automate
Fonctionnement

- par rapport à un automate d'états finis déterministe le fonctionnement est le même avec en plus d'une réponse OUI/NON une translitération du texte initial sur la bande de sortie.
- Notez que si la réponse de l'automate est NON, la translitération sur la bande de sortie peut être partielle, mais cela n'a pas d'importance car elle n'a pas d'intérêt.
- La translitération est obtenue par ajout successivement sur la bande de sortie du symbole de sortie associé à chaque transition qui est déclenchée lors de la reconnaissance.