Segmentation en phrases : ouvrez les guillemets sans perdre le fil

Sandrine Ollinger[1], Denis Maurel[2]

[1]ATILF, Université de Lorraine, CNRS – Sandrine.Ollinger@atilf.fr

[2]Université de Tours, Lifat – Denis.Maurel@univ-tours.fr

Abstract

This paper presents a graph cascade for sentence segmentation of XML documents. Our proposal offers sentences inside sentences for cases introduced by quotation marks and hyphens, and also pays particular attention to situations involving incises introduced by parentheses and lists introduced by colons. We present how the tool works and compare the results obtained with those available in 2019 on the same dataset, together with an evaluation of the system's performance on a test corpus.

Keywords: segmentation ; sentences ; inclusion ; Unitex ; graph cascades.

Résumé

Cet article présente une cascade de graphes pour la segmentation en phrases de documents XML. Notre proposition prévoit une inclusion de phrases pour les cas introduits par des guillemets et tirets et porte également une attention particulière aux situations d'incises introduites par des parenthèses et des listes introduites par des deux-points. Nous présentons le fonctionnement de l'outil et comparons les résultats obtenus à ceux disponibles en 2019 sur le même jeu de données, ainsi qu'une évaluation des performances du système sur un corpus test.

Mots clés : segmentation ; phrases ; inclusion ; Unitex ; cascades de graphes.