Segmentation en phrases des textes de loi de la Cour Suprême américaine

Contexte

Dans un contexte de mondialisation, le droit américain a une forte incidence sur les systèmes juridiques des autres pays. Il est crucial que les juristes ayant l'anglais-américain en seconde langue puissent comprendre ces textes de loi. Le projet Lexhnology, financé par l'ANR, vise à montrer que l'explicitation de la structure argumentative des textes de loi aide les apprenants (étudiant en droit et langue) à monter plus rapidement en compétence sur la compréhension. Du point de vue du Traitement Automatique des Langues (TAL), cet objectif pose de nombreux problèmes et enjeux : ces textes de loi sont écrits dans une langue d'un domaine de spécialité, ils sont très longs et ne suivent aucune structure logique consensuelle.

Un problème particulier concerne le découpage en phrases. En effet, le '.' généralement utilisé pour marquer des fins de phrase a un usage ambigu dans les textes de loi. On le retrouve par exemple au sein de codes servant de référence à d'autres textes de loi. Le texte de la figure~1 illustre la complexité de la langue et l'usage du '.' dans des situations autres que la délimitation de phrases. Dans cet extrait, on compte 2 phrases et 3 codes faisant références à des textes de loi. Saurez-vous les retrouver ?

> *After the central criminal purposes of a conspiracy have been attained, a subsidiary conspiracy to conceal the crime may not be implied from circumstantial evidence showing merely that the conspiracy was kept a secret and that the conspirators took care to cover up their crime in order to escape detection and punishment. Krulewitch v. United States, 336 U. S. 440; Lutwak v. United States, 344 U. S. 604. Pp. 353 U. S. 399-402.*
>
> Figure 1: Extrait du texte de loi Grunewald v. United States, 353 U.S. 391 (1957)

Objectif

Le présent sujet vise à soutenir l'analyse automatique des textes de loi US en se penchant sur le problème de la segmentation des textes en phrases. Ce travail s'insère dans le cadre du projet Lexnology financé par l'ANR [2].

Missions

- Synthèse de l'état de l'art des outils et des approches
- Évaluation (en termes de précision, temps et énergie consommée) des outils existants (spacy, stanza, trankit, CoreNLP, deepsegment, nltk…) sur la tâche de segmentation en phrase. Un corpus de phrases sera mis à disposition.
- Proposition et évaluation d'adaptations de l'existant ou d'une solution tierce pour traiter les spécificités du domaine légal ; une piste pourra être de pré-traiter les textes en repérant les codes des citations (citeurl, legal-reference-extraction).
- Ecriture d'un rapport sous la forme d'un article scientifique

Pour aller plus loin, les étudiants pourront se pencher sur le problème de segmentation de l'en-tête qui ne suit pas un format textuel rédigé.

Références

- Sentence splitters benchmark, (Zavyalova et al., 2022)
- Unsupervised Multilingual Sentence Boundary Detection (Kiss & Strunk, 2006)
- Discourse-Based Sentence Splitting (Cripwell et al., 2021)
- Structural Text Segmentation of Legal Documents (Aumiller et al., 2021)