Traitement Automatique du Langage Naturel (TALN) - 1

Résumé

Le document porte sur le **Traitement Automatique du Langage Naturel (TALN)** ou **Natural Language Processing (NLP)**, une branche de l'intelligence artificielle visant à permettre aux machines de comprendre et manipuler le langage humain. Il commence par une introduction aux approches utilisées en TALN : **symbolique** (basée sur des règles linguistiques), **statistique** (fondée sur des modèles probabilistes), et **connexionniste** (reposant sur des réseaux neuronaux et le deep learning).

L'historique du TALN est retracé, depuis les premières tentatives de traduction automatique des années 1950, en passant par l'émergence des modèles statistiques dans les années 1980, jusqu'à l'utilisation des réseaux neuronaux dans les années 1990 et 2000. L'évolution récente est marquée par l'apparition des **modèles de langage à grande échelle** comme **GPT, BERT et LLaMa**.

Le document décrit ensuite les **prétraitements linguistiques**, indispensables pour transformer le texte brut en données exploitables. Ces étapes incluent la **tokenisation** (segmentation en unités linguistiques), le **codage des caractères** (ASCII, Unicode, UTF-8) et l'**embedding lexical** (Word2Vec, GloVe, FastText), qui permet de représenter les mots sous forme de vecteurs numériques.

Les architectures neuronales spécifiques au traitement du langage sont également abordées : les **Réseaux de Neurones Récurrents (RNN)** et leurs variantes comme **LSTM** pour la gestion de longues dépendances, les modèles **Encodeur-Décodeur** pour la traduction, et les **Transformers** qui ont révolutionné le domaine avec des mécanismes d'attention, ouvrant la voie aux modèles modernes comme **GPT et BERT**.

Enfin, plusieurs applications du TALN sont mises en avant, telles que la **reconnaissance vocale, les chatbots, la traduction automatique, la fouille de texte et la génération automatique de contenus.**

Terme	Définition
TALN (Traitement Automatique du Langage Naturel)	Discipline de l'IA permettant aux machines de comprendre et traiter le langage humain.
NLP (Natural Language Processing)	Equivalent anglais du TALN.
Tokenisation	Processus de découpage du texte en unités linguistiques (mots, phrases, sous-mots).
Word Embedding	Technique de représentation des mots sous forme de vecteurs pour capturer leurs relations sémantiques.
Word2Vec	Modèle de plongement lexical qui représente les mots en fonction de leur contexte.
GloVe	Modèle d'embedding basé sur les cooccurrences de mots dans un corpus.
FastText	Algorithme de Facebook qui prend en compte la morphologie des mots en plus de leur contexte.
RNN (Recurrent Neural Network)	Réseau de neurones conçu pour traiter les données séquentielles comme le texte.
LSTM (Long Short-Term Memory)	Variante des RNN capable de capturer des dépendances à long terme.
Transformers	Architecture neuronale qui repose sur l'attention et a remplacé les RNN/LSTM dans de nombreuses applications.
BERT (Bidirectional Encoder Representations from Transformers)	Modèle de Google qui comprend le contexte des mots dans les deux directions.
GPT (Generative Pre-trained Transformer)	Modèle développé par OpenAI pour la génération de texte.
Seq2Seq (Sequence to Sequence)	Architecture de réseaux neuronaux utilisée pour la traduction et la génération de texte.
Byte-Pair Encoding (BPE)	Algorithme de compression utilisé pour la tokenisation en sous-mots.
Mécanisme d'attention	Technique permettant aux modèles de pondérer différemment l'importance des mots dans une séquence.