Natural Language Processing (NLP)

1 — Introduction

Georges-André Silber

Session 2023/2024 — ES3A_MES-07

École des mines de Paris

2 milliards

C'est, en euros, la valorisation de Mistral AI, start-up française spécialisée dans l'intelligence artificielle, après sa nouvelle levée de fonds de 385 millions d'euros dimanche. L'entreprise d'une vingtaine d'employés suit *«une ambition claire : créer un champion européen à vocation mondiale dans l'intelligence artificielle»*, fanfaronne son patron, Arthur Mensch, 31 ans, cité dans le communiqué du groupe. Avec, en ligne de mire, son concurrent OpenAI, créateur de ChatGPT. **E.V.** **Lire notre décryptage sur *Libé.fr***

23 Les Echos Mardi 12 décembre 2023

HIGH-TECH&MEDIAS

Les opérateurs télécoms français s'emparent de l'IA générative

- Orange, SFR, Bouygues Telecom et Free n'ont pas attendu ChatGPT pour se mettre à l'IA.
- Tous espèrent maintenant faire un « saut quantique » avec l'IA générative.
- Les cas d'usages apparaissent dans le service client et sur les réseaux.

Quelques sources

- Chaire annuelle 2023—2024 de Benoît Sagot au Collège de France « *Apprendre les langues aux machines* »
- Cours du master MVA « *Algorithms for speech and language processing* »
- Cours de Stanford « *CS224N : Natural Language Processing with Deep Learning* »
- Livre « *Speech and Language Processing* » (Jurafsky/Martin)
- Livre/notes « *Natural Language Processing* » (Jacob Eisenstein)

Cours de Science des données (S. Mallat)

- Chaire « *Science des données* » du Collège de France
- Objectif : méthodes effectives d'extraction de connaissance des données
- Rencontre entre les maths app. et l'informatique
- Mathématiques : statistiques, probabilités, analyse harmonique, géométrie, groupes, ...
- Informatique : IA, langages, bases de données, calcul parallèle, ...
- Données : images, sons, textes, mesures physiques, ...
- 3 grands sous domaines : traitement du signal, modélisation (apprentissage non supervisé), prédiction (apprentissage supervisé).
- Très grandes dimension pour les deux derniers domaines.
- Importance des informations a priori (données + "modèle").

Natural Language Processing

- En français : Traitement Automatique des Langues (TAL)
- Lien avec l'IA : imiter et assister l'intelligence humaine
- Discipline pluri-disciplinaire : linguistes, informaticiens, mathématiciens
- « *Talistes, taliens, taleux* » (Lebarbé)
- Challenges principaux du NLP : analyse, génération, transformation de textes, interaction humain/machine
- Applications : linguistique, humanités, droit, santé, ...
- Années 90 : passage des règles à l'apprentissage automatique (ML)
- Les grands modèles de language (LLM) réalisent aujourd'hui la plupart des tâches du NLP de manière performante : état de l'art du domaine

Histoire sélective du NLP

- 1950, traduction automatique (contexte de guerre froide)
- 1950, *Computing Machinery and Intelligence* (A. Turing)
- 1954, expérience Georgetown-IBM, traduction du russe vers l'anglais
- 1966, ELIZA (Joseph Weizenbaum)
- 1968, SHRDLU (PhD de Terry Winograd au MIT)
- 1970–2000, « ontologies conceptuelles »
- 2018, BERT (Google)
- 2020, GPT-3 (OpenAI)
- 2023, ChatGPT (OpenAI)

Histoire sélective des réseaux neuronaux

- 1943, Notion de neurone artificiel (McCulloch & Pitts)
- 1957/1958, Apprentissage supervisé, Perceptron (Rosenblatt)
- 1962, Plusieurs couches en propagation avant (Rosenblatt)
- 1986, Rétropropagation du gradient (Rumelhart, Hinton, Williams)
- 1989, Réseaux convolutifs (Le Cun *et al.*)
- 1990, Réseaux récurrents (Elman)
- 1997, LSTM (Hochreiter)
- 2006, *Deep Learning*, $c \geq 3$ (Hinton, Bengio)
- 2017, Architecture *Transformer* (Vaswani *et al.*)

Mise en œuvre du NLP

- Correction orthographique
- Détection de contenus haineux
- Extraction de contenu structuré
- Moteurs de recherche
- *Data to text*
- Sous-titrage d'images
- Résumé de vidéos
- Traduction automatique
- Résumé de texte
- Simplification de texte
- Système de réponse aux questions
- Chatbots

Tâches et outils du NLP

- Analyse lexicale, lemmatisation, *tokenization*
- *POS Tagging* (étiquetage morpho-syntaxique)
- *Named Entity Recognition* (NER), reconnaissance d'entités nommées
- Analyse syntaxique symbolique ou probabiliste
- Modèles de langues
- Réseaux de neurones, *Large Language Models*
- Agents conversationnels

4e révolution de l'accès à l'information

Extrait de la leçon inaugurale de B. Sagot (11/2023) :

1. Écriture : stockage des informations de manière externe et pérenne. Outil d'accès à l'information
2. Imprimerie : externalisation et diffusion facilités
3. Web : numérisation massive, moteurs de recherche. Automatisation de l'identification des sources.
4. IA : restitution des informations et capacité externe de raisonnement