NLP ADAPTED TO THE CCUS PERIMETER

QUICK OVERVIEW OF USAGES, BENEFITS AND CHALLENGES

Jerome MASSOT, PhD., TOTAL EP RT USA
jerome.massot@total.com
Mary McLeod Bethune

The daughter of former slaves, Mary Jane McLeod Bethune became one of the most important black educators, civil and women’s rights leaders and government officials of the twentieth century. The college she founded set educational standards for today’s black colleges, and her role as an advisor to President Franklin Delano Roosevelt gave African Americans an advocate in government.

https://www.womenshistory.org/education-resources/biographies/mary-mcleod-bethune
A little bit of Jargon

The few words you need to know for this presentation ... and to shine in social events 😊

NLP = Natural (Neural) Language Processing
NLU = Natural (Neural) Language Understanding
NLG = Natural (Neural) Language Generation
LM = Language Model
KM = Knowledge Management
KG = Knowledge Graph
Q&A = Question-Answering

Summarization = producing abstract from paper content automatically
Ranked Corpus = bunch of documents sorted by relevance
Text Representation (Embeddings) = dense representation understood by AI
Contextualized Word Embeddings (ELMO) = not Sesame St. but word vectors
Transformer (BERT, T5, GPT-3, ...) = not the robot, but a model architecture
Named Entities = “smart” keywords linking words to concepts
DNN = Deep Neural Network

RNN, LSTM, Attention mechanism = Recurrent Neural Network with good memory and paying attention to the important information
Global Road Map

NLP/NLU TOOLKIT
UNDER THE HOOD

EXISTING AND FUTURE TOOLS

Q&A
Global Road Map

NLP/NLU TOOLKIT
UNDER THE HOOD

EXISTING AND FUTURE TOOLS

Q&A
ARTICLES TSUNAMI

Papers

Tsunami
Bibliographical Workflow

"Articles Lake"

Content Extraction and Text Processing

Text Understanding (NLP, NLU)

Topics Classification

Question Answering

Summarization (NLG)

Search Assistant (chatbot)
Content Extraction and Text Processing

- Content Extraction is very challenging
- Two strategies depending on the documents' nature:
  - **Native** files (pdf, word, …) : direct extraction
  - **Scanned** documents : Vision AI Segmentation
- Text processing (**Representations**) :
  - Cleaning
  - **Tokenization** (words, sentences, paragraphs)
  - N-grams extraction
  - Contextualized **Embeddings**
Text Representations \(\text{(from characters to numerical vectors)}\)

**Image content**

For image classification, the first feature is the organization of the pixels content. Advanced treatments are usually done.

**Semantic content (words)**

1 word but 2 representations

"character" or 
\[ ['c', 'h', 'a', 'r', 'a', 'c', 't', 'e', 'r'] \]

Words can be represented at different levels of **granularity**: n-grams (combination of words), unitary word, characters.

- \(N = 1\) : This is a sentence
- \(N = 2\) : This is a sentence
- \(N = 3\) : This is a sentence

Condensed as **Contextual Embedding Vectors**

**Document text content**

For text classification, the first feature is the words frequency of the content. **Advanced condensed representations** (vectors) are usually employed.
Today, AI does not “understand” text content. But it can mimic text understanding thanks to Rules and Probabilistic models:

- A Language Model is a probabilistic model able to predict a word knowing previous ones in the sequence (context).
- Its probabilistic foundations could come from:
  - Grammar, Semantic and Syntactical Rules
  - Documentary corpus used to compute conditional probabilities of occurrence for each word in a given vocabulary, knowing its context
  - Mathematic Engine is not very complex (Transformer), but global AI architecture (BERT) makes optimization tricky, especially when trained on the entire Web content.

\[
P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
\]
Domain Adaptation: BERT needs you!!

General Knowledge Domain

- New Language Model mastodonta every 6 months.
  - trained on foolishly big corpus,
  - Masking – Sentence Prediction Training

Beyond Language Model, Universal Knowledge Model.

Fine-Tuning (task specific training) is provided for:
- Text classification
- Knowledge Extraction
- Summarization

Specific Knowledge Domains

- BioBERT, FinBERT
  - Fine-Tune generic model with Medical or Financial corpora (technical vocabulary, language structure, …)

- No CCUSBERT
  - But such domain adaptation is not for free, and there is not equivalent for CCUS today …

What is needed?

Labelled data is dramatically missing in both quantity and quality -> very difficult Domain Adaptation and Transfer Learning for CCUS!!
How can you help with data?

- Academic Papers
- Technical Documents

± 50,000 pages
± 20,000 pages
(annnotated topics, entities)

Shared Open-Source Corpora

- Organize information into concepts
- Map data to knowledge
- Named Entities

CCUS Ontology(ies)

- Databases
- Taxonomies
- Gazetteers & Glossaries

Knowledge Base(s)

Saline aquifers
Carbonates etc.

“Amines”
“Solvents” etc.

Semi-supervised labeling
Direct Named Entities Recognition
Compositional Semi-Automatic Data Labelling for NER thanks to BERT Fine-Tuning

“A regular expression (shortened as regex) is a sequence of characters that define a search pattern.”

\[(0?\[1-9]\|[12]\[0-9]\]|3\[01]\)\/[0-9]d{4}\]

01/01/2020 or 01-2020 = DATE

Named Entities Dictionary attack

Regexp

Deep Learning

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0.3745</td>
<td>0.3690</td>
</tr>
<tr>
<td>Value</td>
<td>0.6856</td>
<td>0.8265</td>
</tr>
<tr>
<td>Host</td>
<td>0.6520</td>
<td>0.7559</td>
</tr>
<tr>
<td>Acquisition</td>
<td>0.4609</td>
<td>0.5407</td>
</tr>
<tr>
<td>Well</td>
<td>0.7667</td>
<td>0.8425</td>
</tr>
<tr>
<td>Measurement</td>
<td>0.5763</td>
<td>0.6108</td>
</tr>
<tr>
<td>Quantifier</td>
<td>0.6103</td>
<td>0.6694</td>
</tr>
</tbody>
</table>

2.891 sentences ≈ 13,000 labels
Topic modeling: AI learns by itself some topics from the text content (clustering).

Latent Semantic Analysis (LSA) / Latent Dirichlet Allocation (LDA) have the same fundamental assumption: documents with the same topic use similar words.

Topic Classification: AI identifies in the text content one or several topics defined by humans (classification).

Rule-based methods, Machine Learning methods (Naive Bayes, Support Vector Machines), or Deep Learning methods map text features to predefined topics.

Today, BERT and Co. are the best tools!!
Question Answering Approaches

IBM Watson (Jeopardy)

Pipeline of conventional linguistically-based NLP techniques, such as parsing, part-of-speech tagging and coreference resolution. Very important step is the ranking of answer candidates and their curation with Open World Knowledge !!

Deep Learning Revolution

DNN revolutionized Q&A challenge !

- Smaller learning pipeline
- Significant amount of training.

The main idea is to find answer of the query in context by matching the representations of context and query thanks to a special attention mechanism.

Best model architectures are:

- Bi-LSTM
- BiDAF
- BERT and Co.
Successful Zero-Shot Learning Closed-Domain Extractive Q&A

- **Closed-Domain**: needs a context to find answers
- **Extractive**: answers are extracted as fractions of the context

**A** Named Entities are extracted from the question

**B** Ranking of contexts from filtered corpus thanks to entities

**C** Answer Processing using Reading Comprehension (Hugging Face)
Summarization

**Extractive Approaches**

Key Phrase Extraction techniques (TextRank, LexRank) are very popular methods to extract key sentences or rank sentences in a document.

Rules -> Ranking -> Extraction

Could be treated as a Classification task using specific features:
- Binary classification: key or not
- Multi-classification: from 1 to 5

Training -> Classification -> Extraction

**Generative Approaches (NLG)**

In generative approach, an AI model based on a Language Model generates a text summarizing the document to test.

Input Text -> Trained LM -> Summary

Commonly Sequence-to-Sequence (Seq2Seq)

Most efficient model architectures:
- Recurrent Neural Network (RNN)
- Long Short Time Memory (LSTM)
- Attention Mechanism
- Bi-Directional Stacked LSTM
- Transformers
- BERT fine-tuned for summarization
- Other Transformer architecture with similar fine-tuning

**Needed Datasets**

DOMAIN ADAPTATION & FINE-TUNING

unsupervised

supervised

supervised
Global Road Map

NLP/NLU TOOLKIT
UNDER THE HOOD

EXISTING AND FUTURE TOOLS

Q&A
Shared Corpora Solutions

Traditional repository manager:
GitHub, GitLab, ...

Traditional shared drive:
Google drive, Box, ...

New solution proposed by Google
https://journaliststudio.google.com/pinpoint
Text Analysis and Topics Modeling & Classification

StreamLit Web App (Python)  
Text Content Analysis
Bibliography Exploration with Graphs

Grobid
Native format

Beautifulsoup

Private Corpus

Connected papers
Integrated Bibliography Platform

Total’s gaia Explorer
Global Road Map

- NLP/NLU TOOLKIT
- UNDER THE HOOD
- EXISTING AND FUTURE TOOLS
- Q&A
Disclaimer and Copyright Reservation

THE TOTAL GROUP IS DEFINED AS TOTAL S.A. AND ITS AFFILIATES AND SHALL INCLUDE THE PARTY MAKING THE PRESENTATION

Disclaimer

This presentation may include forward-looking statements within the meaning of the Private Securities Litigation Reform Act of 1995 with respect to the financial condition, results of operations, business, strategy and plans of Total that are subject to risk factors and uncertainties caused by changes in, without limitation, technological development and innovation, supply sources, legal framework, market conditions, political or economic events. Total does not assume any obligation to update publicly any forward-looking statement, whether as a result of new information, future events or otherwise. Further information on factors which could affect the company’s financial results is provided in documents filed by the Group with the French Autorité des Marchés Financiers and the US Securities and Exchange Commission. Accordingly, no reliance may be placed on the accuracy or correctness of any such statements.

Copyright

All rights are reserved and all material in this presentation may not be reproduced without the express written permission of the Total Group.