



# **Financial News Summarization**

Ahmed Rachid Hazourli

# Problématique



## Objectif

Solution pour automatiser la génération des résumés des News Financières.



## Apports professionnels

Aider les analystes à évaluer rapidement une situation financière avant de recommander toute action et aider les gérants dans la prise de décision.



## Méthodologie

Utiliser une méthode de l'état de l'art pour générer des bons résumés.

# Méthodes de génération de résumés



## Résumé abstractif

Paraphrase le contenu du document original tout en prenant en compte la cohésion et la concision du résumé en sortie.

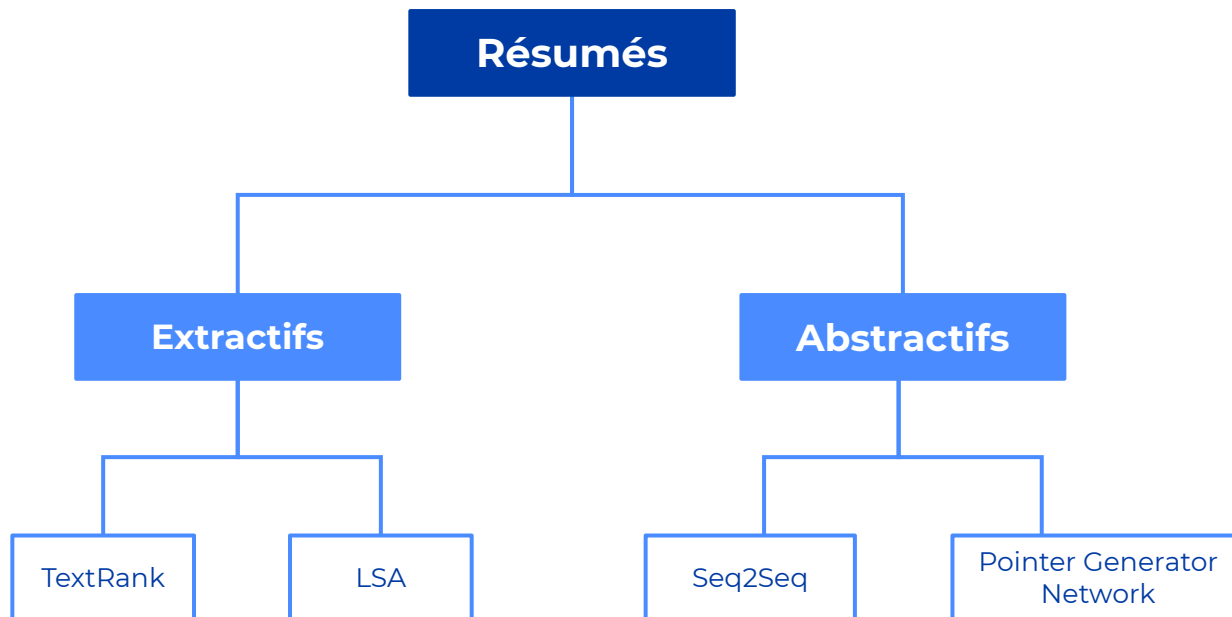
Capable de paraphraser le texte original, de générer de nouvelles phrases et d'inclure de nouveaux mots comme « companies » et « digest » ce qui rend le résumé plus concis.



## Résumé extractif

Choisir des parties du document d'origine, comme des phrases entières voire même un paragraphe


# Exemples de méthodes



# Avantages / Inconvénients



# Motivation

A decorative graphic on the left side of the slide consisting of two overlapping squares. The top square is a lighter blue, and the bottom square is a darker blue, extending from the bottom left towards the center.

J'ai choisi l'utilisation de la méthode abstractive pour la génération des résumés.

La sortie des modèles classiques de Deep Learning (Full Connected Network, ConvNet, RNN...) est de taille fixe ce qui n'est pas adapté à notre démarche.

La Seq2Seq qui est une méthode de l'état de l'art connue par la nature de sa sortie qui n'est pas contrainte par une taille fixe. J'ai donc choisi le modèle BERT qui utilise l'architecture Seq2Seq.

# Application des modèles Seq2Seq

**01**

PoS Tagging

**02**

Named Entity Recognition

**03**

Question Answering

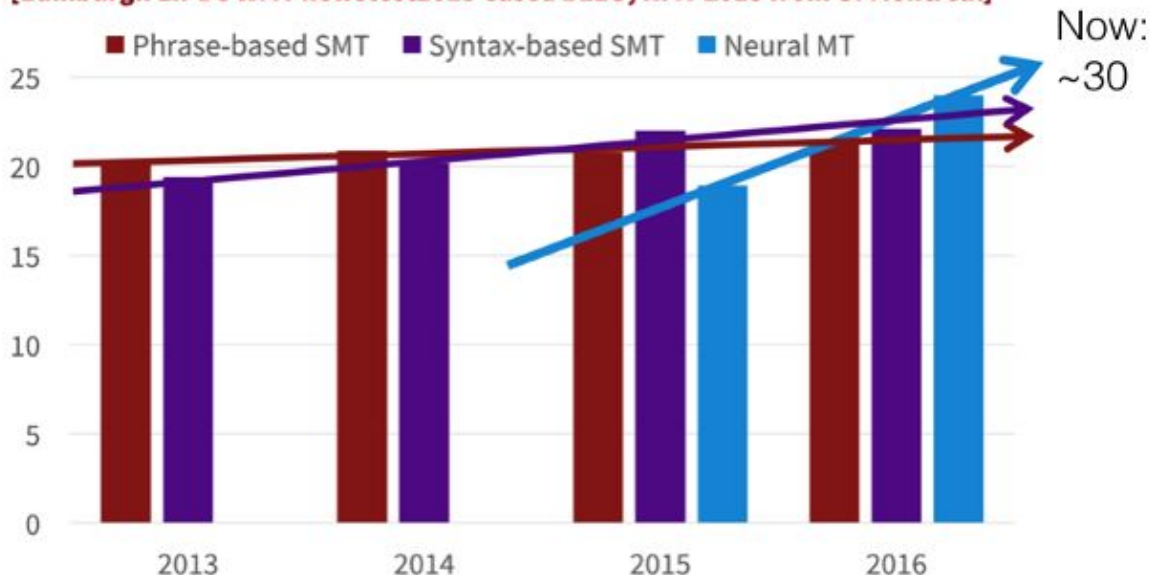
**04**

Text Summarization

# Seq2Seq pour la traduction des textes

## Progress in Machine Translation

[Edinburgh En-De WMT newstest2013 Cased BLEU; NMT 2015 from U. Montréal]



\* Next 3 slides are thanks to Prof Chris Manning (Stanford)

Progrès des modèles Seq2Seq de Deep Learning pour la Traduction des Textes depuis 2015



# Étapes de la réalisation du projet



# Data Exploration

- Notre dataset contient **2071 news financières** récupérées depuis Bloomberg sous formes de: document / résumé.

Exemples:

- Dad
- dadzadada

# Les étapes du prétraitement:

La première étape à chaque fois que l'on fait du NLP est de construire une pipeline de nettoyage de nos données.

Le plus gros travail du data scientist ne réside pas dans la création de modèle. Le nettoyage du dataset représente une part énorme du processus, nous avons appliqué ces étapes:

- La mise en minuscule des textes a été une étape importante dans notre démarche.
- **Suppression des stopwords:** les mots fréquents et courants sans valeur inhérente comme: the, and, while ...
- **Suppression des URLs:** comme <http://t.co/opjnrXlb...>
- **Suppression des ponctuations et caractères spéciaux:** comme ?, !, # ...
- **Suppression** des " 's "
- **Suppression des textes entre parenthèses ( )**
- **Tokenization:** séparation des mots

A decorative graphic on the left side of the slide consisting of two overlapping squares. The top square is a lighter blue, and the bottom square is a darker blue, creating a cross-like shape.

# Data Splitting & Model Training

L'idée est comme d'habitude de diviser notre jeu de données en un échantillon d'entraînement (80%) dans lequel nous allons apprendre les paramètres du modèle et un échantillon test (20%) dans lequel nous allons les tester.

# BERT

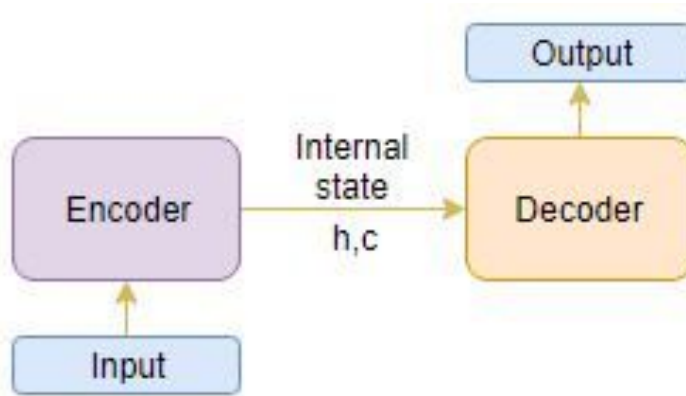
BERT est un modèle de représentation de textes écrit en langage naturel. La représentation faite par BERT à la particularité d'être contextuelle.

C'est-à-dire qu'un mot n'est pas représenté de façon statique comme dans un embedding classique mais en fonction du sens du mot dans le contexte du texte.

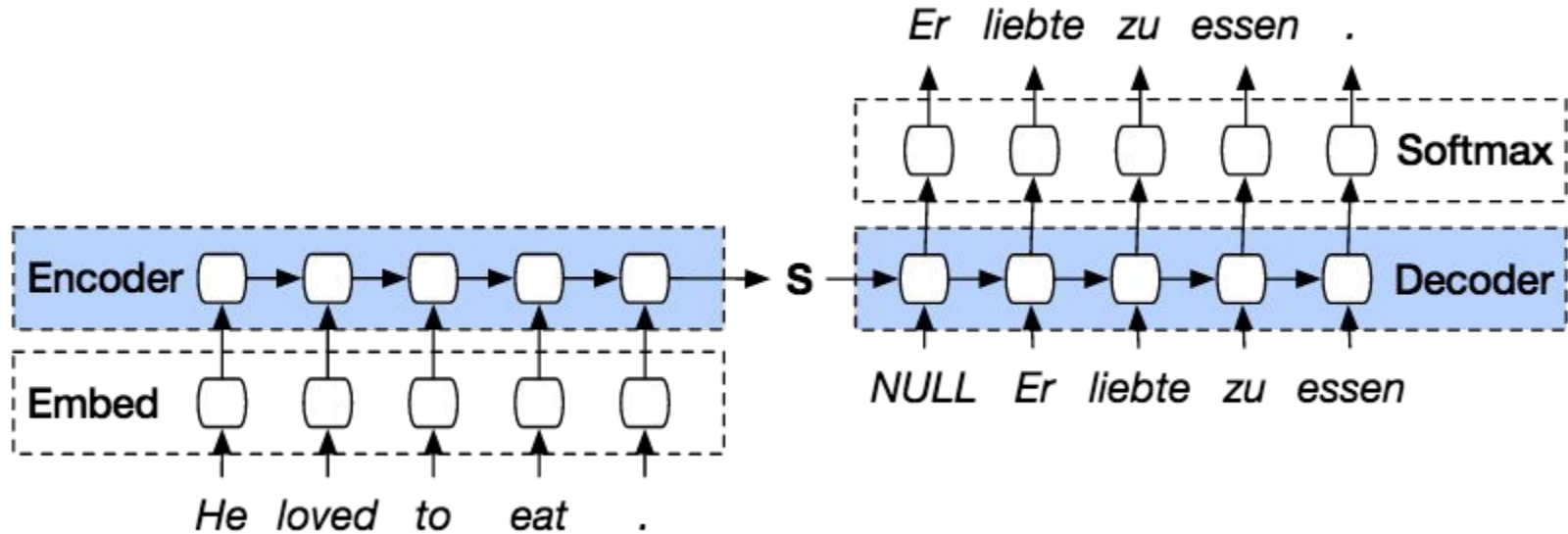
Par exemple, le mot "baguette" aura des représentations différentes dans "la baguette du magicien" et "la baguette du boulanger". En plus, le contexte de BERT est bi-directionnel, c'est-à-dire que la représentation d'un mot fait intervenir à la fois les mots qui le précèdent et les mots qui le suivent dans une phrase.

# Architecture d'un modèle Seq2Seq

Comprenons cela du point de vue du résumé de texte. L'entrée est une longue séquence de mots et la sortie sera une version courte de la séquence d'entrée.

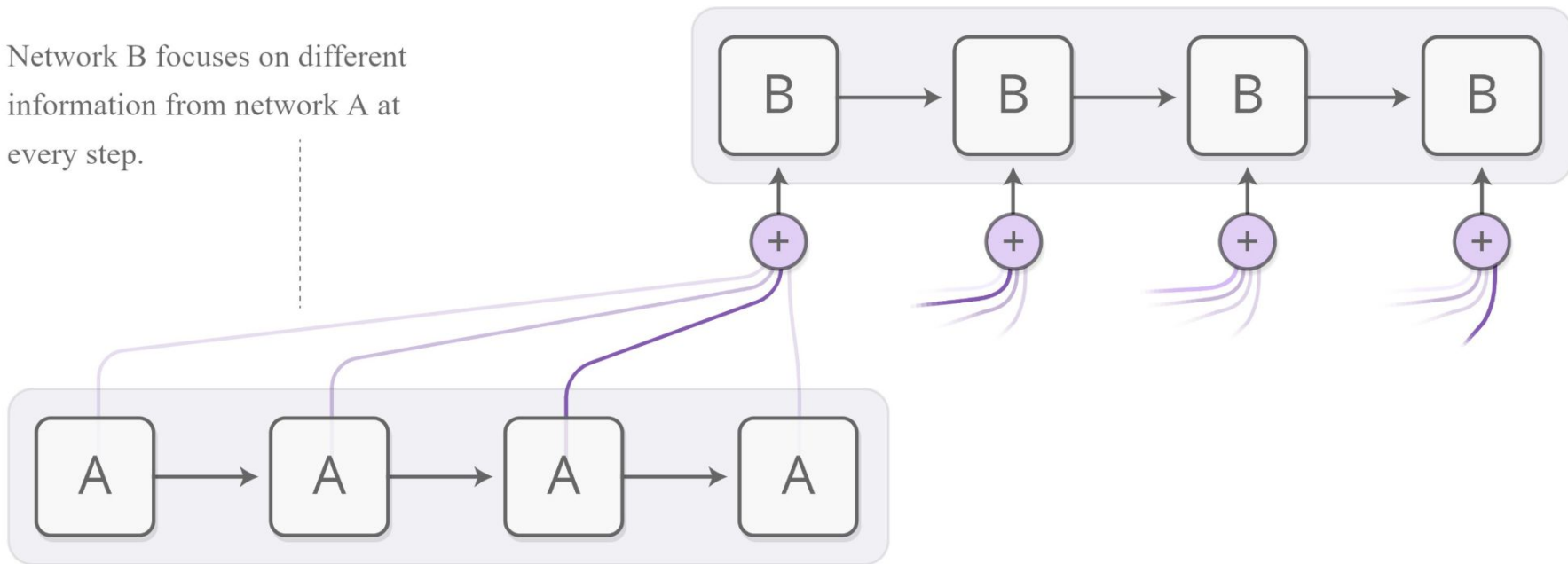


# Architecture d'un modèle Seq2Seq



# Mécanisme d'attention dans un modèle Seq2Seq

Network B focuses on different information from network A at every step.





# Pourquoi le BERT ?

Le principe d'utilisation de BERT est tout simple : Il est déjà pré-entraîné sur un dataset de documents en anglais nommé "XSUM" ( Extreme Summarization ) composé de 205 606 documents / résumés.

On le modifie pour une tâche précise puis on le re-entraîne avec nos propres données ( News financières ).

La modification dont il est question ici consiste en général à rajouter un réseau de neurones à la sortie de BERT. Cela s'appelle : le **fine-tuning**.

Nous avons utilisé le fine-tuning à l'aide du BERT pré-entraîné pour la tâche de génération de résumés.

# Evaluation des modèles

**ROUGE:** évalue les résumés en les comparant à des résumés modèles. C'est la mesure d'évaluation la plus adoptée par les travaux de l'état de l'art.

Elle est déduite à partir du recouvrement entre les N-grammes des deux textes. Elle utilise trois métriques pour quantifier la comparaison.

- Précision: traduit à quel point les données sélectionnées sont pertinentes.
- Rappel: reflète à quel degré le résumé évoque des données pertinentes qu'il est censé inclure.
- F1-score: la moyenne harmonique de la précision et du rappel, elle montre une forte corrélation avec les jugements humains.

# Evaluation des modèles

$$Rappel(Res) = \frac{\sum_{r \in Ref} \sum_{Ngrammes \in Res} card_{match(r, Res)}(Ngrammes)}{\sum_{r \in Ref} \sum_{Ngrammes \in r} card(Ngrammes)}$$

$$Precision(Res) = \frac{\sum_{r \in Ref} \sum_{Ngrammes \in Res} card_{match(r, Res)}(Ngrammes)}{\sum_{r \in Res} \sum_{Ngrammes \in r} card(Ngrammes)}$$

$$Fmeasure = \frac{(1 + \beta)^2 Rappel \cdot Precision}{Rappel + \beta^2 Precision}$$

# Evaluation du modèle

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-S
<b>Fine-tuned BERT</b>	23.60	7.01	18.19	21.51

# Exemples de générations de résumés

Une évaluation humaine est aussi importante pour juger la pertinence d'un modèle de génération de résumés. Voici des exemples d'articles résumés:

- **Texte original:** HSBC (HSBA.L) has no plans to launch a cryptocurrency trading desk or offer the digital coins as an investment to customers, because they are too volatile and lack transparency, its Chief Executive Noel Quinn told Reuters. Europe's largest bank's stance on cryptocurrencies comes as the world's biggest and best-known, Bitcoin, has tumbled nearly 50% from the year's high, after China cracked down on mining the currency and prominent advocate Elon Musk tempered his support. read more It marks it out against rivals such as Goldman Sachs (GS.N), which Reuters in March reported had restarted its cryptocurrency trading desk, and UBS (UBSG.S) which other media said was exploring ways to offer the currencies as an investment product. read more "Given the volatility we are not into Bitcoin as an asset class, if our clients want to be there then of course they are, but we are not promoting it as an asset class within our wealth management business," Quinn said.
- **Résumé:** Europe's largest bank has no plans to launch a trading desk. CEO says they are too volatile and lack transparency

- **Texte original:** European stocks hit record highs on Monday as another run of gains in automakers more than offset early declines in commodity-linked shares sparked by downbeat China export data. The European automobiles and parts index (.SXAP) rose 0.9% to reach its highest since March 2015, extending a 5.3% rally from last week. The continent-wide STOXX 600 index (.STOXX) added 0.2%, with global investors now eyeing a European Central Bank meeting later this week. Euro zone banks (.SX7P) were broadly higher as government yields were steady near one-month lows ahead of the ECB meeting on Thursday when policymakers are expected to stick to their dovish policy stance. "We expect the ECB to maintain its current pace of asset purchases even as the economic restart gains traction," strategists at BlackRock wrote in a client note. "We wouldn't view a decision to slow purchases as a hawkish policy signal, as the ECB is focused on keeping financing conditions easy. This, and a Federal Reserve that we see keeping policy easy, provides a positive backdrop for risk assets including European equities." Miners (.SXPP) fell 1.6% as copper prices dipped after a slower-than-forecast growth in Chinese exports sparked concerns about weakening demand for the red metal. read more Oil and gas stocks (.SXEP) declined 0.3% as crude prices pulled back ahead of talks this week between Iran and world powers over a nuclear deal that, if clinched, is expected to boost crude supplies. Global stocks have been pinned near life highs as major economies reopen from coronavirus lockdowns, but concerns that the economic recovery may not be as fast as thought, and signs of quickening inflation, have slowed the pace of gains.
- **Résumé:** Automakers extend rally from last week, while commodity-linked shares fall.

# Limitations

Même si le modèle est performant mais il a montré plusieurs lacunes concernant la tâche de génération de résumé de News:

- Parfois incapable de reproduire des informations précises.
- Résumés générés répétitifs et contiennent des informations redondantes.



**Conclusion**