# Présentation d’une seconde série d’articles

# Liste des articles

- The limits of automatic summarisation according to ROUGE
- Overview of the TAC 2008 Update Summarization Task
- Automatically Evaluating Content Selection in Summarization without human models
- Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE
- BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

The limits of automatic summarisation according to ROUGE

*https://aclanthology.org/E17-2007.pdf*

# The limits of automatic summarisation according to ROUGE

- (1) Perfect scores for extractive summarisation are theoretically computationally hard to achieve.
- (2) 100% perfect scores are impossible for higher quality datasets.
- (3) Relative perfect scores are highly diverse and unattainable by humans.
- (4) State-of-the-art automatic summarisation is unsupervised.

# The limits of automatic summarisation according to ROUGE

## 2. Preliminaries

- Rappel sur comment est calculé ROUGE-n(S)
- 3 datasets :

$$\text{ROUGE-}n(S) := \frac{\sum_{g \in S} |\{g | g \in S\} \cap \{g | g \in R\}|}{\sum_{g \in R} |\{g | g \in R\}|} \quad (1)$$

- DUC 2004

- judgment-summary pairs scraped from the European Court of Human Rights case-law website, HUDOC.3 The test set consists of 138 pairs.

- a comprehensive dump of English language Wikipedia articles.

# The limits of automatic summarisation according to ROUGE

## 3. ROUGE optimisation for extraction

- Preuve que la tâche d’optimisation des résumés extractifs est NP-hard.
- Qu'est-ce qui est compliqué ?
  - C'est de maximiser les scores et donc d'avoir un modèle que l'on peut vraiment améliorer. Ils font 2 preuves de cette complexité. Les deux expériences se basent sur ROUGE-N
- un résumé extractif qui maximise la métrique ROUGE-1 peut être réduit au problème de l'ensemble dominant pondéré maximum à k qui est un problème de théorie des graphes.

# The limits of automatic summarisation according to ROUGE

- le problème du max k-weighted dominating

  => l'objectif est de sélectionner un sous-ensemble de sommets de taille au plus k, maximisant la somme des poids et assurant une couverture majoritaire sur le graphe.

  => Pour les résumés :

  1. On imagine que chaque token du document est un sommet dans un graphe, les sommets représentent la similarité ou la connexion entre les tokens

  2. on pondère les sommets, chaque token à un poids =le nombre de fois qu'il apparaît dans le résumé de référence.

  3. Ensemble dominant de token : on choisi un sous-ensemble de token de sorte à ce que ce sous-ensemble maximise le chevauchement des tokens.

Re-evaluation Automatic Summarization with BLEU and 192 Shades of ROUGE

*https://aclanthology.org/D15-1013.pdf*

3 problématiques sont identifiées :

- Movement away from evaluation by correlation with human assessment
- Omission of important components of human assessment from evaluations
- Absence of methods of significance testing improvements over a baseline.

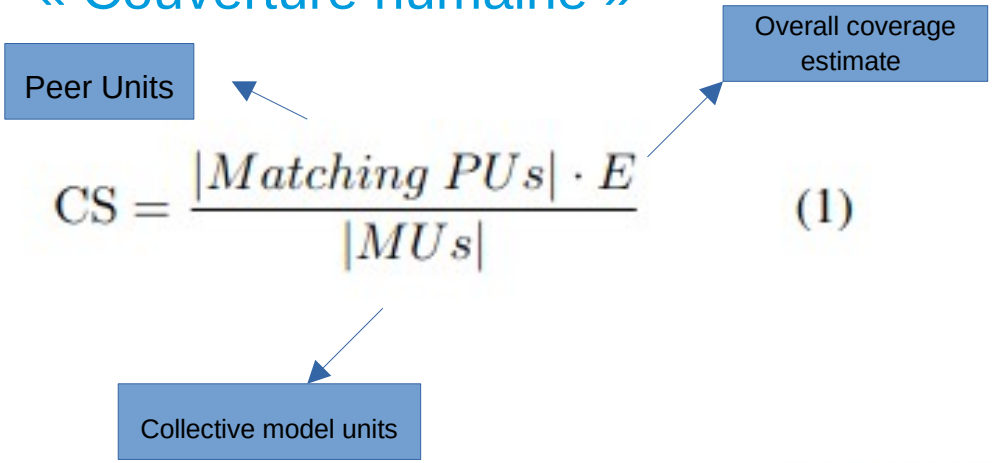
- ROUGE, adaptation du score BLEU pour l’évaluation des résumés
- Limites :

  - biais en faveur des métriques basées sur le rappel (évalue la capacité d’un système à identifier correctement toutes les instances pertinentes d’un ensemble de données).

  - Omission des dimensions telles que l’ordre des unités de résumé.

## 3.1 : Combinaison des scores – human annotation

- Qualité linguistique
- « Couverture humaine »




$$\text{Human Assessment Score} = \frac{CS+MLQ}{2}$$

## 3.2 : ROUGE

- « Large number of distinct variants »
  - 8 choix de n-gram
  - Word-stemming
  - SW
  - P/R/F
- (8*2*2*3*2) → 192 variantes
- Avantage de ROUGE : score moyen ou médian de résumés individuels donc on peut effectuer des tests (system-level) scores/variantes.

## 3.3. Metric Evaluation by Pearson's r

- Permet de comparer les résultats des différentes variantes

| Metric | Stemming | RSW | Ave./Med | P/R/F | r | Metric | Stemming | RSW | Ave./Med | P/R/F | r | Metric | Stemming | RSW | Ave./Med | P/R/F | r |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | | | | | 0.797 • | R-2 | Y | N | M | F | 0.706 | R-L | Y | Y | A | F | 0.638 |
| R-2 | Y | Y | A | P | 0.786 • | R-3 | N | Y | M | P | 0.704 • | R-1 | N | N | A | F | 0.637 |
| R-3 | N | N | A | F | 0.785 • | R-1 | N | Y | A | P | 0.704 • | R-S4 | Y | N | M | F | 0.634 |
| R-2 | N | Y | A | P | 0.783 • | R-4 | N | N | M | R | 0.703 • | R-4 | Y | N | M | P | 0.634 |
| R-3 | N | Y | A | P | 0.781 • | R-L | N | Y | A | P | 0.700 • | R-1 | N | N | M | F | 0.634 |
| R-3 | Y | N | A | F | 0.779 • | R-W | Y | Y | A | P | 0.700 • | R-SU4 | N | Y | A | R | 0.633 |
| R-3 | N | N | A | R | 0.777 • | R-4 | N | Y | A | R | 0.700 • | R-L | Y | Y | M | P | 0.633 |
| R-4 | N | N | A | F | 0.771 • | R-1 | Y | N | M | P | 0.699 • | R-SU4 | Y | Y | M | R | 0.631 |
| R-3 | N | N | A | P | 0.771 • | R-S4 | N | Y | M | P | 0.698 | R-1 | Y | N | A | F | 0.630 |
| R-3 | Y | N | A | R | 0.770 • | R-1 | Y | Y | A | P | 0.698 • | R-1 | Y | Y | M | F | 0.629 |
| R-2 | N | Y | A | F | 0.769 • | R-3 | N | Y | M | F | 0.697 • | R-S4 | Y | Y | M | R | 0.626 |
| R-4 | N | N | A | R | 0.768 • | R-W | N | N | A | P | 0.696 • | R-S4 | N | N | A | R | 0.626 |
| R-2 | Y | Y | A | F | 0.768 • | R-W | Y | N | A | P | 0.695 • | R-SU4 | Y | N | M | F | 0.625 |
| R-3 | Y | N | A | P | 0.767 • | R-4 | N | N | M | F | 0.695 • | R-S4 | Y | N | A | R | 0.624 |
| R-3 | N | N | M | F | 0.766 • | R-S4 | N | Y | M | F | 0.693 | R-L | N | Y | M | F | 0.623 |
| R-3 | N | Y | A | F | 0.764 • | R-S4 | N | Y | A | F | 0.691 | R-SU4 | Y | Y | A | R | 0.622 |
| R-3 | Y | Y | A | P | 0.764 • | R-SU4 | N | Y | M | P | 0.690 | R-1 | Y | N | M | F | 0.617 |
| R-4 | Y | N | A | F | 0.763 • | R-1 | N | N | M | P | 0.690 • | R-1 | N | Y | M | R | 0.615 |
| R-4 | N | N | A | P | 0.762 • | R-2 | N | N | M | R | 0.689 | R-W | N | Y | A | R | 0.613 |
| R-4 | Y | N | A | R | 0.761 • | R-L | Y | Y | A | P | 0.688 • | R-S4 | N | N | M | R | 0.611 |
| R-3 | N | N | M | P | 0.760 • | R-3 | N | Y | M | R | 0.687 • | R-L | N | Y | M | R | 0.609 |
| R-4 | Y | Y | A | P | 0.759 • | R-S4 | N | N | M | P | 0.687 | R-1 | N | Y | A | R | 0.604 |
| R-2 | Y | N | A | P | 0.759 • | R-S4 | Y | N | A | F | 0.687 | R-L | N | Y | A | R | 0.601 |
| R-4 | N | Y | A | P | 0.758 • | R-S4 | N | N | A | F | 0.687 | R-W | N | N | M | F | 0.600 |
| R-2 | N | N | A | P | 0.757 • | R-4 | N | N | M | P | 0.687 • | R-L | N | N | M | F | 0.599 |
| R-3 | N | N | M | R | 0.753 • | R-L | N | N | A | P | 0.686 • | R-W | Y | Y | A | R | 0.598 |
| R-4 | Y | N | A | P | 0.752 • | R-SU4 | N | N | M | P | 0.686 | R-W | N | Y | M | R | 0.597 |
| R-3 | Y | Y | A | F | 0.748 • | R-L | Y | N | A | P | 0.683 • | R-1 | Y | Y | A | R | 0.595 |
| R-2 | N | N | A | F | 0.747 • | R-W | N | N | M | P | 0.682 • | R-1 | Y | Y | M | R | 0.591 |

## 4. Metric Significance Testing

- Permet d’évaluer la significativité de la différence entre deux corrélations dépendantes.
- Est-ce que la différence entre deux corrélations est significatives ?

## 4. Metric Significance Testing

- Williams test

« It is formulated as follows
as a test of whether the population correlation between X1 and X3 equals the population correlation between X2 and X3:

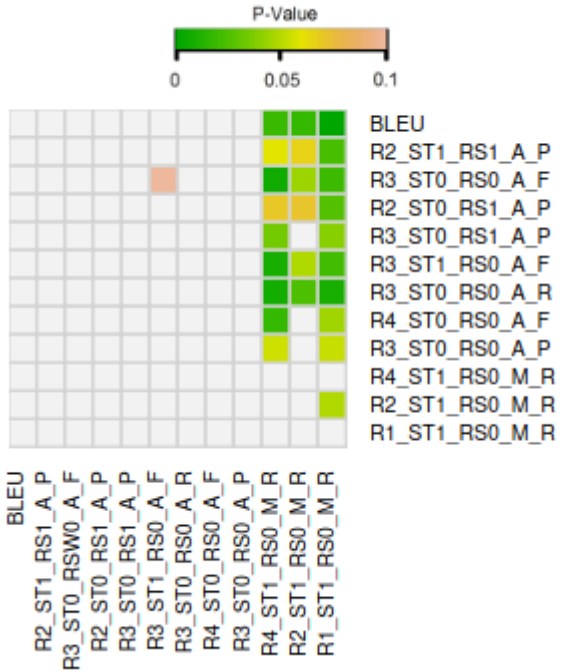
$$t(n-3) = \frac{(r_{13} - r_{23})\sqrt{(n-1)(1+r_{12})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{23}+r_{13})^2}{4}(1-r_{12})^3}},$$

where rij is the correlation between Xi and Xj , n is the size of the population, and:

$$K = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2r_{12}r_{13}r_{23}$$

## 4.1 Significance Test Results

- Assurer que les améliorations des métriques ne sont pas dues au hasard et sont statistiquement valides.
- le test a révélé une différence statistiquement significative dans la corrélation avec l'évaluation humaine entre ces métriques



## 5. Summarization System Evaluation

- Evaluation du système avec  la métrique ROUGE qui a eu la meilleure corrélation avec les évaluations humaines : ROUGE-2, precision, stemming et stop-words removed.
- Evaluation plus fidèle des résumés.

| System | ROUGE Best | ROUGE Original |
|---|---|---|
| DPP | 8.498 | 9.62 |
| ICSISumm | 8.317 | 9.78 |
| RegSum | 8.187 | 9.75 |
| Submodular | 8.047 | 9.35 |
| CLASSY11 | 7.717 | 9.20 |
| CLASSY04 | 7.690 | 8.96 |
| OCCAMS_V | 7.643 | 9.76 |
| GreedyKL | 6.918 | 8.53 |
| FreqSum | 6.838 | 8.11 |
| TsSum | 6.671 | 8.15 |
| Centroid | 6.660 | 7.97 |
| LexRank | 6.655 | 7.47 |

Table 3: Summarization systems originally included in Hong et al. (2014) evaluated with the best-performing ROUGE variant (Best): average ROUGE-2 precision with stemming and stop words removed; and evaluated with original suboptimal variant (median ROUGE-2 recall with stemming and without removal of stop-words)

## 6. Human Assessment Combination

- Un seul score d'évaluation humaine
- Comment attribuer une note qui représente bien la globalité des évaluations ?
- Moyennes arithmétiques, géométriques et harmoniques

| Metric | Stem. | RSW | Ave/Med | P/R/F | Mean | Geometric Mean | Harmonic Mean | Coverage Only | Ling. Qual. Only |
|---|---|---|---|---|---|---|---|---|---|
| BLEU | | | | | **0.797**• | **0.901**• | **0.936**• | **0.944**• | 0.642• |
| ROUGE-2 | Y | Y | A | P | 0.786• | 0.870• | 0.887• | 0.878 | 0.660• |
| ROUGE-3 | N | N | A | F | 0.785• | 0.869• | 0.893 | 0.894 | 0.650• |
| ROUGE-2 | N | Y | A | P | 0.783• | 0.868• | 0.885• | 0.876 | 0.658• |
| ROUGE-3 | N | Y | A | P | 0.781• | 0.836• | 0.840 | 0.826 | **0.682**• |
| ROUGE-3 | Y | N | A | F | 0.779• | 0.866• | 0.891 | 0.893 | 0.643• |
| ROUGE-3 | N | N | A | R | 0.777• | 0.871• | 0.901 | 0.907 | 0.632• |
| ROUGE-4 | N | N | A | F | 0.771• | 0.843• | 0.863 | 0.866 | 0.645• |
| ROUGE-3 | N | N | A | P | 0.771• | 0.837• | 0.849 | 0.843 | 0.658• |
| ROUGE-3 | Y | N | A | R | 0.770• | 0.867• | 0.899 | 0.905 | 0.624• |
| ROUGE-2 | N | Y | A | F | 0.769• | 0.877• | 0.909• | 0.910• | 0.619• |
| ROUGE-2 | Y | Y | A | F | 0.768• | 0.875• | 0.908• | 0.908• | 0.618• |
| ROUGE-3 | Y | N | A | P | 0.767• | 0.835• | 0.849 | 0.843 | 0.652• |
| ROUGE-3 | Y | Y | A | P | 0.764• | 0.825• | 0.832 | 0.821 | 0.660• |
| ROUGE-4 | N | N | A | P | 0.762• | 0.815• | 0.824 | 0.819 | 0.657• |
| ROUGE-4 | Y | Y | A | P | 0.759• | 0.794 | 0.790 | 0.774 | 0.678• |
| ROUGE-4 | N | Y | A | P | 0.758• | 0.793 | 0.789 | 0.772 | 0.678• |
| ROUGE-4 | Y | N | A | P | 0.752• | 0.809 | 0.819 | 0.815 | 0.646• |
| ROUGE-2 | N | N | A | F | 0.747• | 0.867• | 0.907• | 0.910• | 0.587• |
| ROUGE-2 | Y | N | A | F | 0.747• | 0.868• | 0.908• | 0.912• | 0.586• |
| ROUGE-2 | N | Y | A | R | 0.742• | 0.862• | 0.904• | 0.912• | 0.578• |
| ROUGE-2 | N | Y | M | F | 0.740• | 0.855• | 0.894• | 0.898• | 0.584• |
| ROUGE-2 | Y | Y | A | R | 0.737• | 0.858• | 0.900• | 0.908• | 0.575• |
| ROUGE-2 | N | Y | M | R | 0.722• | 0.848• | 0.895• | 0.905• | 0.553• |
| ROUGE-2 | N | N | M | R | 0.689 | 0.828 | 0.884• | 0.901• | 0.508 |

Overview of the TAC 2008 Update Summarization Task

https://tac.nist.gov/publications/2008/additional.papers/update_summ_overview08.proceedings.pdf

# Overview of the TAC 2008 Update Summarization Task

PLAN

1) Description de la tâche et des corpus

2) System Approaches

3) Résultats

# Overview of the TAC 2008 Update Summarization Task

Sources des données :

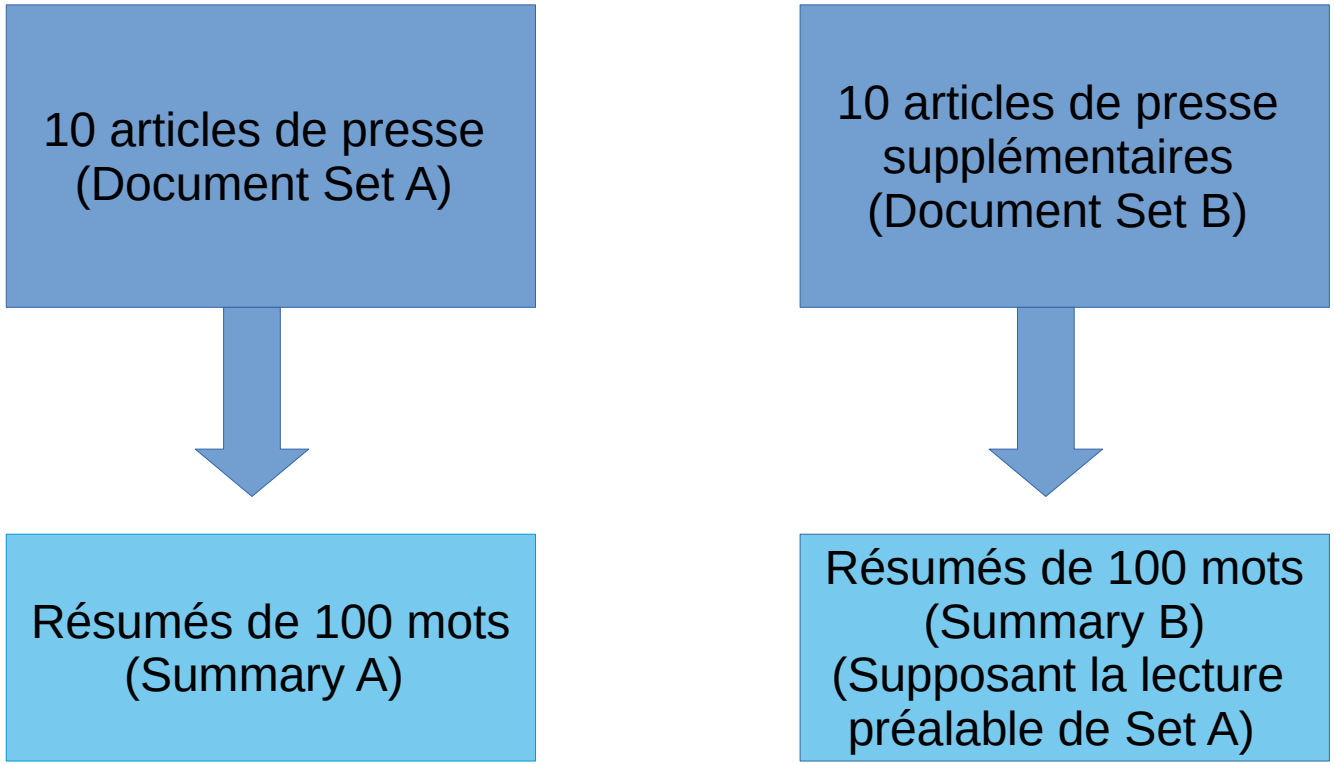
Collection AQUAINT-2

(qui contient environ 907 000 dépêches de presse)

→ Ensemble A

→ Ensemble B

# Overview of the TAC 2008 Update Summarization Task



# Overview of the TAC 2008 Update Summarization Task

**Summarization task :**

- La même pour l'homme et la machine : créer depuis les documents, 2 résumés brefs, fluide et bien organisé.
- A et B (les résumés) doivent répondrent au besoin d'information exprimé dans un « topic »
- 4 résumés humains par document,

  et parmi les 4, forcément 1 écrit

  par le créateur du « topic ».
- Équipe : 3 « prioritized run »

**num:** D0842G
**title:** Natural Gas Pipeline
**narr:** Follow the progress of pipelines being built to move natural gas from Asia to Europe. Include any problems encountered and implications resulting from the pipeline construction.

Rappel de la structure d'un topic.

# Overview of the TAC 2008 Update Summarization Task

2) System Approaches

- Extractive summaries (Limite de 100 mots)
- Pour les mises à jour : on évite les redondances

# Overview of the TAC 2008 Update Summarization Task

1) Pour améliorer la requête, élargissement de la requête initiale
 - ajout de WordNet (base de données lexicale)
 - Wikipedia (encyclopédie en ligne)
 - co-occurence de mots

2) Analyse linguistique approfondie
 - Lemmatisation
 - POS
 - NER
 - Analyse syntaxique

3) Sélection des phrases
 - Classement : (tf-idf)
 - Clustering : clusters de similarités

Automatically Evaluating Content Selection in Summarization without human models
*https://aclanthology.org/D09-1032.pdf*

- But : créer un modèle complètement automatique.
- On compare le résumé généré directement à l’input plutôt qu’a une résumé humain.
- **Principe que la distribution des mots dans l’input est similaire à celle du résumé automatique.**

- Données et Évaluation

  - Ensemble de Données : Utilisation de l'ensemble de test de la conférence TAC 2008, qui comprend 48 ensembles de documents.
  - Types de Résumés : Résumés focalisés sur une requête (query) et résumés de mise à jour (update summary).
  - Plusieurs métriques d'évaluation :

**Métriques d’évaluation (TAC 2008)**

- Pyramid evaluation
- Responsiveness evaluation
- ROUGE evaluation
- Linguistic quality evaluation

- 1. Similitude distributionnelle :
  - **Divergence de Kullback-Leibler (KL)** : Mesure la divergence entre deux distributions de probabilité.
  - **Divergence de Jensen-Shannon (JS) :** Une version symétrique de KL, toujours définie et souvent plus performante.
  - **Similitude Cosinus :** Utilisée pour comparer les représentations vectorielles tf-idf des mots des documents sources et des résumés.

- 1. Similitude distributionnelle :
  - **Divergence de Kullback-Leibler (KL)** : Mesure la divergence entre deux distributions de probabilité.

$$D(P||Q) = \sum_{w} p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)}$$

$$p(w) = \frac{C + \delta}{N + \delta * B}$$

- 1. Similitude distributionnelle :
  - **Divergence de Jensen-Shannon (JS) :** Une version symétrique de KL, toujours définie et souvent plus performante.

$$J(P||Q) = \frac{1}{2}[D(P||A) + D(Q||A)],$$

$$\text{where } A = \frac{P+Q}{2}$$

- 1. Similitude distributionnelle :
  - **Similitude Cosinus :** Utilisée pour comparer les représentations vectorielles tf-idf des mots des documents sources et des résumés.

$$cos\theta = \frac{v_{inp}.v_{summ}}{||v_{inp}||||v_{summ}||} \quad (4)$$

  - **2** variantes :

1. Vectors contain all words from input and summary
2. Vectors contain only topic signatures from the input and all words of the summary

- 2. Summary likehood :
  - Probabilité que les mots d'un résumé apparaissent dans le document source, évaluée par des modèles unigramme et multinomial.

**Unigram summary probability:**

$$(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}...(p_{inp}w_r)^{n_r} \quad (5)$$

where $p_{inp}w_i$ is the probability in the input of word $w_i$, $n_i$ is the number of times $w_i$ appears in the summary, and $w_1...w_r$ are all words in the summary vocabulary.

**Multinomial summary probability:**

$$\frac{N!}{n_1!n_2!...n_r!}(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}...(p_{inp}w_r)^{n_r} \quad (6)$$

where $N = n_1 + n_2 + ... + n_r$ is the total number of words in the summary.

- 3. Utilisation des « topic words »
  - Nombre de mots-clés du document source présents dans le résumé.
  - Pourcentage de mots-clés du document source apparaissant dans le résumé.

- Résultats
  - Corrélations :
    - Les corrélations entre les évaluations automatiques et manuelles montrent des résultats prometteurs.
    - La divergence de Jensen-Shannon (JS) obtient des corrélations élevées avec les évaluations manuelles : 0,88 avec le score de pyramide et 0,73 avec l'évaluation de la réactivité
  - Implications :
    - Les méthodes automatiques peuvent remplacer les évaluations basées sur des modèles humains lors du développement de systèmes de synthèse textuelle.
    - Elles permettent d'évaluer la performance des systèmes sur des collections de tests standard sans nécessiter de modèles humains.

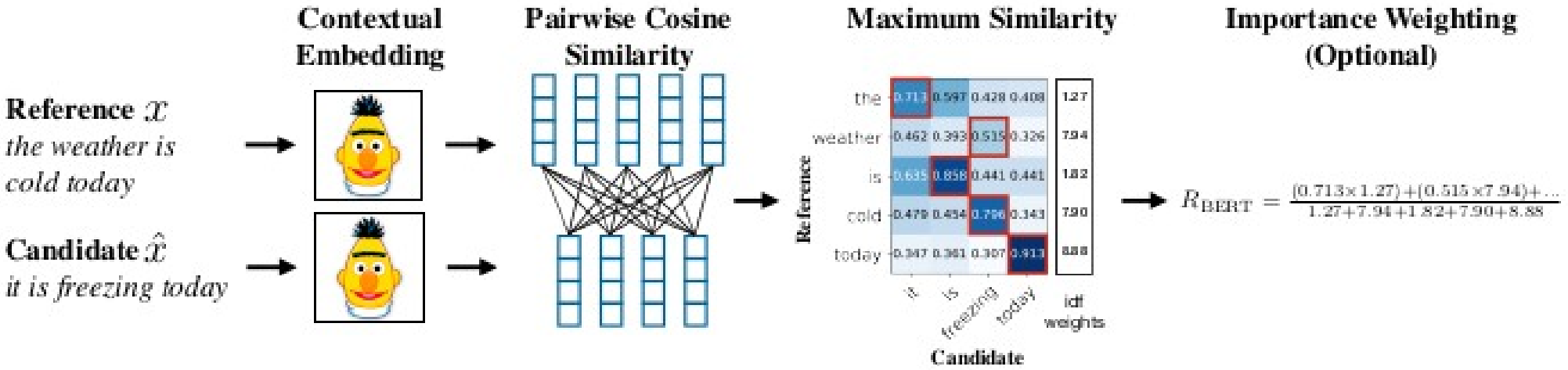
BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

https://arxiv.org/pdf/1904.09675

# BERTSCORE: EVALUATING TEXT GENERATION WITH BERT

- Computing semantic similarity vs token level syntactical similarity
- Calcul la similarité de 2 phrases comme une somme de similarités cosinus entre les embeddings des tokens.

# BERTSCORE: EVALUATING TEXT GENERATION WITH BERT



**Importance Weighting :** Les mots rares peuvent être plus important que des mots qui apparaissent souvent donc : inverse

$$\text{idf}(w) = -\log \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}[w \in x^{(i)}]$$

**Baseline Rescaling :**

$$\hat{R}_{\text{BERT}} = \frac{R_{\text{BERT}} - b}{1 - b} \quad .$$