



# Traduction automatique

## Évaluation

**Enzo Doyen**

enzo.doyen@unistra.fr

2025

# Pourquoi évaluer ?

- ♦ comparaison des systèmes de TA avec les traductions humaines ;
- ♦ comparaison des systèmes de TA entre eux (systèmes complètement différents, ou versions différentes du même système) ;
- ♦ analyse (qualitative et quantitative) des erreurs produites par les systèmes ;
- ♦ en recherche scientifique, preuves de l'efficacité des systèmes par l'utilisation de mesures objectives.

# Types d'évaluation

## Évaluation humaine

- + Mesure d'évaluation « idéale »
- Coût humain important, chronophage et laborieux.
- Non réutilisable.
- Accord interannotateur souvent faible.

## Évaluation automatique

- + Très peu coûteux, facile, et réutilisable
- Résultats dépendants de la métrique.
- Pas toujours fiable.



# I. Évaluation humaine

# Évaluation humaine

Évaluation sur la base de questionnaires qui visent à mesurer :

- ♦ l'« **adequacy** » : la préservation du sens par rapport au texte original ;
- ♦ la « **fluency** » : la formulation dans la langue cible (en d'autres termes, la traduction est-elle « naturelle » ? Le texte aurait-il pu être produit par un ou une locutrice native ?)

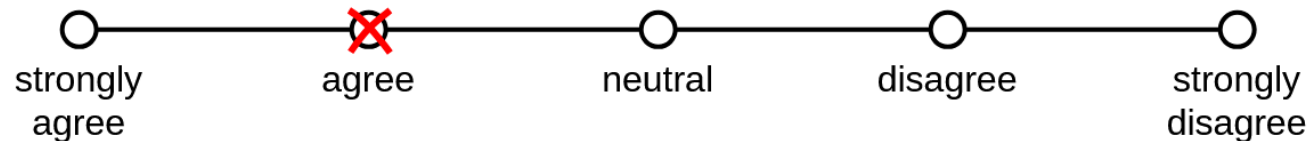
# Évaluation humaine

Évaluation sur la base de questionnaires qui visent à mesurer :

- ♦ l'« **adequacy** » : la préservation du sens par rapport au texte original ;
- ♦ la « **fluency** » : la formulation dans la langue cible (en d'autres termes, la traduction est-elle « naturelle » ? Le texte aurait-il pu être produit par un ou une locutrice native ?)

On utilise souvent une **échelle de Likert** pour évaluer ces deux aspects.

1. The website has a user friendly interface.



Source : [https://en.wikipedia.org/wiki/Likert\\_scale](https://en.wikipedia.org/wiki/Likert_scale)

# Évaluation humaine : exemple d'interface

## Evaluation Criteria:

**Adequacy:** How much of the meaning expressed in the source text is also expressed in the translation?

**Fluency:** How fluent/natural is the language in the translation?

Scale: 1 = Very Poor, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent

| Translation Pairs  | Adequacy  | Fluency   |
|--|---|---|
| <p>1. <b>ENGLISH</b></p> <p>The weather is beautiful today, and I plan to go for a walk in the park.</p> <p><b>FRENCH</b></p> <p>Le temps est magnifique aujourd'hui, et je prévois d'aller me promener dans le parc.</p>                | <div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div></div> | <div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div></div> |
| <p>2. <b>ENGLISH</b></p> <p>Machine learning algorithms are revolutionizing the way we process data.</p> <p><b>FRENCH</b></p> <p>Les algorithmes d'apprentissage automatique révolutionnent la façon dont nous traitons les données.</p> | <div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div></div> | <div><div>1</div><div>2</div><div>3</div><div>4</div><div>5</div></div> |

# Évaluation humaine

Évaluation sur la base de questionnaires qui visent à mesurer :

- ♦ l'« **adequacy** » : la préservation du sens par rapport au texte original ;
- ♦ la « **fluency** » : la formulation dans la langue cible (en d'autres termes, la traduction est-elle « naturelle » ? Le texte aurait-il pu être produit par un ou une locutrice native ?)

On utilise souvent une **échelle de Likert** pour évaluer ces deux aspects.

**⚠ Problème :** composantes avec une part importante de subjectivité, difficiles à évaluer de manière objective.



# Évaluation humaine

D'autres manières d'évaluer :

- ♦ évaluation de la compréhension des textes traduits par des locutrices et locuteurs natifs (**Scarton et Specia, 2016**) ;
- ♦ tâche de *gap-filling* : on enlève des mots de la traduction de référence, et on demande à des locutrices et locuteurs natifs de remplir les blancs avec la traduction automatique comme indice (**Forcada et al., 2018**) ;

# Évaluation humaine : *gap-filling* (Forcada et al., 2018)

**Instructions:** Fill each one of the gaps in the "problem sentence" at the bottom with the most fitting **single word**, using only information from the *hint text* (if there is one).

**Hint text:** (you might need to scroll to find some highlighted text)

The Federal Republic of Germany after 1945 experienced a huge economic boom, which was the economic basis for a stable democracy.

**In the German Democratic Republic the socialist one-party dictatorship of the SED and the socialist planned economy have been introduced at the same time.**

Until 1989, the GDR had therefore great economic problems.

The consequences had a major impact on life in the GDR.

**Problem sentence:** At the same time in the German Democratic Republic , the socialist one-party dictatorship of the SED and

state-planned  were introduced .

✓ Submit

# Évaluation humaine

D'autres manières d'évaluer :

- ♦ évaluation de la compréhension des textes traduits par des locutrices et locuteurs natifs (**Scarton et Specia, 2016**) ;
- ♦ tâche de *gap-filling* : on enlève des mots de la traduction de référence, et on demande à des locutrices et locuteurs natifs de remplir les blancs avec la traduction automatique comme indice (**Forcada et al., 2018**) ;
- ♦ demander à des traductrices et traducteurs professionnels de reporter les erreurs de traduction (**Popović, 2020**).

# Évaluation humaine : accord interannotateur

Pour calculer l'accord interannotateur pour **2 personnes**, on utilise généralement le coefficient Kappa de Cohen (**Cohen, 1960**).

$$k = \frac{p_o - p_e}{1 - p_e}$$

Où  $p_o$  est la proportion d'accord observée entre les annotateurs et annotatrices, et  $p_e$  est la proportion d'accord attendue par hasard, calculée à partir des proportions de chaque catégorie.

$$p_o = \frac{\text{nombre d'accords observés}}{\text{nombre total de jugements}}$$

# Évaluation humaine : accord interannotateur

$p_e$  : proportion d'accord attendue par hasard, calculée à partir des proportions de chaque catégorie.

Par exemple, soit  $k$  le nombre de catégories, et  $p_{i1}$  et  $p_{i2}$  les proportions d'items assignés à la catégorie  $i$  par les deux annotateurs/annotatrices.

$$p_e = \sum_{i=1}^k p_{i1} \cdot p_{i2}$$

(Dans les cas de catégorisation non binaires, p. ex. échelle de Likert, on utilise généralement une version pondérée du coefficient Kappa, qui prend en compte la distance entre les catégories.)

# Évaluation humaine : coefficient Kappa, exemple

**Exemple de tâche** : soit 2 annotateurs/annotatrices, A et B, qui ont catégorisé 100 phrases en deux catégories : C (correct) et I (incorrect).

|              | Annot. B : C | Annot. B : I | Total |
|--------------|--------------|--------------|-------|
| Annot. A : C | 50           | 10           | 60    |
| Annot. A : I | 5            | 35           | 40    |
| Total        | 55           | 45           | 100   |

# Évaluation humaine : coefficient Kappa, exemple

Accords observés :

|              | Annot. B : C | Annot. B : I | Total |
|--------------|--------------|--------------|-------|
| Annot. A : C | 50           | 10           | 60    |
| Annot. A : I | 5            | 35           | 40    |
| Total        | 55           | 45           | 100   |

$$p_o = \frac{50 + 35}{100} = 0.85$$

# Évaluation humaine : coefficient Kappa, exemple

Accords attendus par hasard pour chaque catégorie :

|              | Annot. B : C | Annot. B : I | Total |
|--------------|--------------|--------------|-------|
| Annot. A : C | 50           | 10           | 60    |
| Annot. A : I | 5            | 35           | 40    |
| Total        | 55           | 45           | 100   |

$$p_c = \frac{60}{100} \times \frac{55}{100} = 0.33$$

$$p_i = \frac{40}{100} \times \frac{45}{100} = 0.18$$



# Évaluation humaine : coefficient Kappa, exemple

Calcul du score final :

$$p_e = p_c + p_i = 0.33 + 0.18 = 0.51$$

$$k = \frac{0.85 - 0.51}{1 - 0.51} = \frac{0.34}{0.49} \approx 0.694$$

# Évaluation humaine : coefficient Kappa

| $k$         | Niveau d'accord |
|-------------|-----------------|
| < 0.20      | Très faible     |
| 0.21 – 0.40 | Faible          |
| 0.41 – 0.60 | Moyen           |
| 0.61 – 0.80 | Bon             |
| 0.81 – 1.00 | Parfait         |

Tableau 1. – **Interprétation du coefficient Kappa de Cohen (Landis et Koch, 1977)**



## II. Évaluation automatique

# Scepticisme initial sur l'évaluation automatique

(Bar-Hillel, 1960)

## A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation

One of the reasons why we do not as yet have any translation centers, not even in the planning stage, in which electronic computers, general or special purpose, are used to automate certain parts of the translation process, in spite of the fact that such centers would fulfill a vital function in saving a considerable amount of qualified human translator time per document translated, and thereby facilitate more, quicker and, after some time, cheaper translation, is the reluctance of many MT workers to recognize that the idea of inventing a method for fully automatic high quality translation (FAHQT) is just a dream which will not come true in the foreseeable future. By not realizing the practical futility of this aim, whatever its motivational importance for certain types of basic research, they have misled themselves and the agencies which sponsored their research into not being satisfied with a partly automated translation system whose principles are well understood today, and instead to wait for the real thing which was believed, and made to believe, to be just around the corner.

# Types de mesures d'évaluation automatique

- ♦ **Mesures basées sur la distance d'édition** : on compare la traduction automatique avec la traduction de référence, et on calcule le nombre d'opérations nécessaires pour transformer l'une en l'autre (insertion, suppression, substitution).
- ♦ **Mesures basées sur la précision et le rappel** : on compare les n-grammes de la traduction automatique avec ceux de la traduction de référence, et on calcule la précision et/ou le rappel.
- ♦ **Mesures basées sur la similarité sémantique** : on compare les traductions en utilisant des modèles de langue et des plongements lexicaux.

# Distance d'édition : WER (Word Error Rate)

Métrique basée sur la distance de Levenshtein (mesure de différence entre 2 chaînes de caractères), communément utilisée en reconnaissance vocale.

$S$  = Substitutions,  $D$  = Deletions,  $I$  = Insertions,  $N$  = nombre de mots dans la traduction de référence. Résultat généralement entre 0 et 1 (↓).

$$\text{WER} = \frac{S + D + I}{N}$$

Version pondérée proposée par **Hunt (1990)** :

$$\text{WWER} = \frac{S + 0.5D + 0.5I}{N}$$

# Distance d'édition : TER (Translation Edit Rate)

Approximation du nombre de modifications nécessaires pour transformer la traduction automatique en la traduction de référence.

Contrairement au WER, le TER compare les **séquences de mots** et prend en compte les réordonnements.

$$\text{TER} = \frac{S + D + I + \text{réordonnements (R)}}{N}$$

# Distance d'édition : WER et TER, exemple

**Phrase de référence** : "the quick brown fox jumps over the lazy dog"

**Phrase hypothèse** : "the brown quick fox jumps over lazy the dog"



# Distance d'édition : WER et TER, exemple

**Phrase de référence** : "the **quick brown** fox jumps over **the lazy** dog"

**Phrase hypothèse** : "the **brown quick** fox jumps over **lazy the** dog"

# Distance d'édition : WER et TER, exemple

## WER

**Phrase de référence** : "the **quick brown** fox jumps over **the lazy** dog"

**Phrase hypothèse** : "the **brown<sub>s</sub> quick<sub>s</sub>** fox jumps over **lazy<sub>s</sub> the<sub>s</sub>** dog"

$$\text{WER} = \frac{4}{9} \approx 0.444$$

## TER

**Phrase de référence** : "the **quick brown** fox jumps over **the lazy** dog"

**Phrase hypothèse** : "the **brown quick<sub>R</sub>** fox jumps over **lazy the<sub>R</sub>** dog"

$$\text{TER} = \frac{2}{9} \approx 0.222$$

# Précision/rappel : BLEU (Bilingual Evaluation Understudy)

- ♦ Métrique introduite par **Papineni et al. (2002)**, qui renvoie un score entre 0 et 1 (↑).
- ♦ Mesure la similarité entre la traduction automatique et la traduction de référence en utilisant les n-grammes.

# Précision/rappel : BLEU et n-grammes

**n-grammes** : séquences de n mots consécutifs dans un texte.

| 1-gramme  | 2-grammes  | 3-grammes   | 4-grammes                  |
|---|--|---|----------------------------|
| « the »,<br>« quick »,<br>« brown »,<br>« fox » | « the quick »,<br>« quick<br>brown »,<br>« brown fox » | « the quick<br>brown »,<br>« quick brown<br>fox » | « the quick<br>brown fox » |

# Précision/rappel : BLEU, n-grammes et précision

- On calcule la précision des n-grammes de la traduction automatique par rapport à ceux de la traduction de référence.

$$p_n = \frac{\# \text{ n-grammes candidats correspondants dans la référence}}{\# \text{ total de n-grammes dans phrase hypothèse}}$$

# Précision/rappel : BLEU, n-grammes et précision

- On calcule la précision des n-grammes de la traduction automatique par rapport à ceux de la traduction de référence.

$$p_n = \frac{\# \text{ n-grammes candidats correspondants dans la référence}}{\# \text{ total de n-grammes dans phrase hypothèse}}$$

**Problème :** si on prend uniquement une précision unigramme, la phrase hypothèse suivante pourrait avoir un score de 1, car le seul unigramme de l'hypothèse est aussi dans la référence :

**Phrase hypothèse :** "the the the the the the the"

**Phrase de référence :** "the quick brown fox jumps over the lazy dog"

# Précision/rappel : BLEU, n-grammes et précision

**Problème** : si on prend uniquement une précision unigramme, la phrase hypothèse suivante pourrait avoir un score de 1, car le seul unigramme de l'hypothèse est aussi dans la référence :

**Phrase hypothèse** : "**the the** the the the the the"

**Phrase de référence** : "**the** quick brown fox jumps over **the** lazy dog"

**Solution** : *clipped precision*. On limite le nombre de n-grammes comptés dans la phrase hypothèse à celui de la référence (ici, 2).

## 💡 Précision/rappel : BLEU, exemple global

**Target Sentence:**      The guard arrived late because it was raining  
                                 ↓        ↓        ↓        ↓        ↓  
**Predicted Sentence:** The guard arrived late because of the rain


Source : <https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics>

$$p_1 = \frac{5}{8}$$



## 💡 Précision/rappel : BLEU, exemple global

**Target Sentence:**      The guard arrived late because it was raining



**Predicted Sentence:** The guard arrived late because of the rain

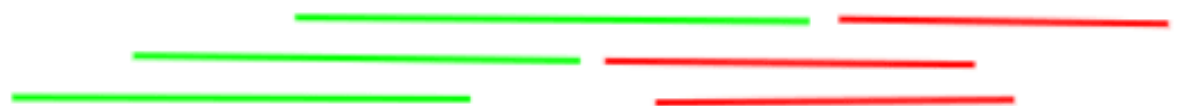
Source : <https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics>

$$p_2 = \frac{4}{7}$$

## 💡 Précision/rappel : BLEU, exemple global

**Target Sentence:** The guard arrived late because it was raining


**Predicted Sentence:** The guard arrived late because of the rain



Source : <https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics>

$$p_3 = \frac{3}{6}$$

# 💡 Précision/rappel : BLEU, exemple global



The diagram shows two sentences with horizontal lines above them representing n-grams. The Target Sentence is "The guard arrived late because it was raining" and the Predicted Sentence is "The guard arrived late because of the rain". Red lines above the Target Sentence represent 3-grams: "The guard arrived", "guard arrived late", and "arrived late because". A green line above the Predicted Sentence represents a 3-gram: "The guard arrived". The overlap between the red and green lines is the 3-gram "The guard arrived", which is highlighted in blue in the original image.

**Target Sentence:** The guard arrived late because it was raining

**Predicted Sentence:** The guard arrived late because of the rain

Source : <https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics>

$$p_4 = \frac{2}{5}$$

## Précision/rappel : BLEU, Brevity Penalty (BP)

Le score BLEU est multiplié par un facteur de pénalité de brièveté (*Brevity Penalty*, BP) pour pénaliser les phrases hypothèses plus courtes que la référence.

$c$  = longueur de la phrase hypothèse

$r$  = longueur de la phrase de référence

$$\text{BP} = \begin{cases} 1 & \text{si } c > r \\ e^{(1-r/c)} & \text{si } c \leq r \end{cases}$$

Valeur maximale limitée à 1 (logarithme).

## Précision/rappel : **BLEU**, formule finale

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right)$$

$N$  et  $w_n$  sont des paramètres modifiables ; ils correspondent respectivement au nombre de n-grammes pris en compte et aux poids associés (généralement,  $w_n = \frac{1}{N}$  ; p. ex. 0.25 avec 4-gram).

La formule BLEU classique a un  $N = 4$  ; les autres formes sont généralement appelées BLEU-5, BLEU-3, etc.

# Précision/rappel : METEOR

- ♦ Métrique introduite par **Banerjee et Lavie (2005)**, qui renvoie un score entre 0 et 1 (↑).
- ♦ Combine précision et rappel (moyenne harmonique).
- ♦ Possibilité d'intégrer synonymes et racines des mots dans le calcul.
- ♦ Meilleure corrélation avec scores humains que BLEU.

## Précision/rappel : METEOR

$$P = \frac{\# \text{ mots candidats correspondants dans la référence}}{\# \text{ total de mots dans phrase hypothèse}}$$

$$R = \frac{\# \text{ mots candidats correspondants dans la référence}}{\# \text{ total de mots dans phrase de référence}}$$

$$F = \frac{10PR}{R + 9P}$$

# Précision/rappel : METEOR

$$\text{Pénalité} = 0.5 \times \frac{\text{nombre de blocs}^1}{\# \text{ mots candidats correspondants dans la référence}}$$

$$\text{METEOR} = F \times (1 - \text{Pénalité})$$

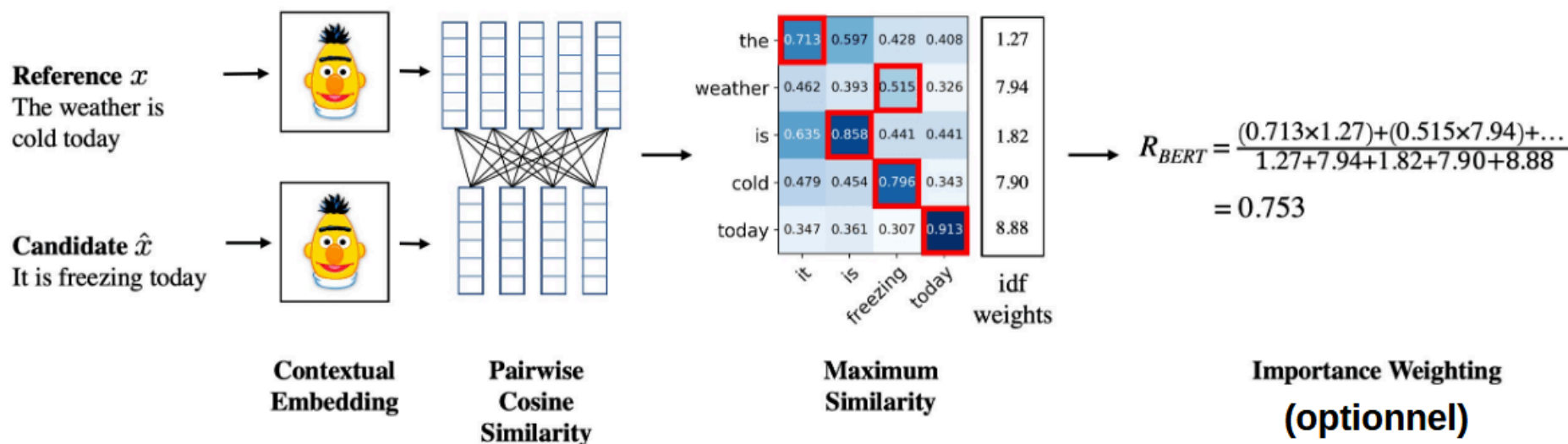
---

<sup>1</sup>Un « bloc » correspond à une séquence alignée entre la phrase hypothèse et la phrase de référence.



# Similarité sémantique : BERTScore

- ♦ Métrique introduite par **Zhang et al. (2020)**, qui utilise les plongements lexicaux de BERT pour évaluer la similarité sémantique entre les phrases.
- ♦ Repose sur la similarité cosinus entre la phrase hypothèse et la phrase de référence.



# BERTScoreVisualizer (Jaskowski et al., 2024)

Reference Text

Ensuring that the food supply is of a consistent and known quality.

Candidate Text

To be sure the food is of good quality.

Model

bert-base-uncased

CALCULATE  
AND  
VISUALIZE  
BERTSCORE

Reference Text Tokens

ensuring that the food supply is of a consistent and known quality .

Candidate Text Tokens

to be sure the food is of good quality .

Hover over a token to see more information.

Overall Scoring

Recall: 0.7263

Precision: 0.7646

F1: 0.745

# Similarité sémantique : COMET

- ♦ Métrique introduite par **Rei et al. (2020)**, basée sur XLM-RoBERTa (**Conneau et al., 2020**), un modèle de langue multilingue préentraîné.
- ♦ Plusieurs modes d'application :
  - *regression metric* : prédit un score, selon un entraînement sur des notes manuelles ;
  - *ranking* : se base sur la distance euclidienne entre les plongements lexicaux de la phrase hypothèse et ceux de la phrase de référence ;
  - *reference-free* : compare la phrase source avec la phrase cible (pour les cas où une traduction humaine de référence n'est pas disponible).

# Implémentation des métriques

- ♦ La plupart des métriques d'évaluation automatique sont implémentées dans des bibliothèques Python, comme sacrebleu, meteor, bertscore, comet, etc.
- ♦ **MATEO (Vanroy et al., 2023)** : interface Web pour évaluer les traductions automatiques, qui implémente plusieurs métriques d'évaluation automatique, dont BLEU, TER, BERTScore et COMET. <https://mateo.ivdnt.org/Evaluate>

# MATEO : mise en ligne des données

## EVALUATION - DATA

- Reference
- Source (when needed)
- Up to four MT systems
- First MT is baseline

### Input data

Add a reference file and one or more files with translations. One line per file. Cannot contain empty lines and must be in UTF8!

Reference file



Drag and drop file here

Limit 200MB per file

Browse files



newstest2021.en-de.ref.de.50.txt 9.3KB



Source file



Drag and drop file here

Limit 200MB per file

Browse files



newstest2021.en-de.src.en.50.txt 7.8KB



How many systems do you wish to compare? (max. 4)

3



System #1 (serves as baseline)



Drag and drop file here

Limit 200MB per file

Browse files



nllb.50.txt 8.2KB



System #2 file



Drag and drop file here

Limit 200MB per file

Browse files



opus.50.txt 8.8KB



System #3 file



Drag and drop file here

Limit 200MB per file

Browse files



gpt-3.5-turbo.50.txt 9.2KB

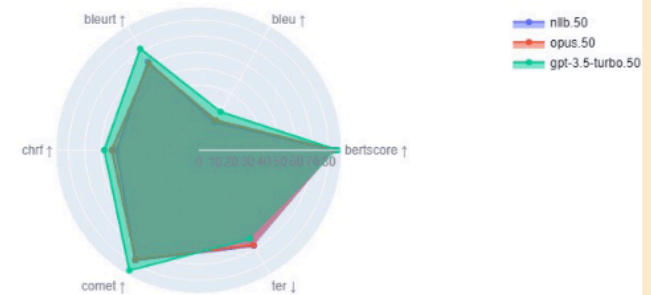


# MATEO : résultats et visualisation automatique

## EVALUATION - RESULTS

### Corpus level

- Bar plot
- Radar plot





## III. Mise en pratique

# Mise en pratique

- ♦ Notebook disponible sur Moodle pour vous exercer à implémenter les métriques d'évaluation automatique à l'aide de code Python et des bibliothèques correspondantes.
- ♦ Vous pouvez éventuellement aussi tester MATEO : <https://mateo.ivdnt.org/Evaluate>



# Bibliographie

- Banerjee, S., et Lavie, A. (juin 2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, et C. Voss (éds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.
- Bar-Hillel, Y. (janvier 1960). The Present Status of Automatic Translation of Languages. In F. L. Alt (éd.), *Advances in Computers: Vol. 1. Advances in Computers* (p. 91-163). Elsevier. [10.1016/S0065-2458\(08\)60607-5](#)
- Cohen, J. (avril 1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46. [10.1177/001316446002000104](#)
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., et Stoyanov, V. (avril 2020). *Unsupervised Cross-lingual Representation Learning at Scale* (Numéro arXiv:1911.02116). arXiv. [10.48550/arXiv.1911.02116](#)
- Forcada, M. L., Scarton, C., Specia, L., Haddow, B., et Birch, A. (octobre 2018). Exploring Gap Filling as a Cheaper Alternative to Reading Comprehension Questionnaires When Evaluating Machine Translation for Gisting. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névél, M. Neves, M. Post, L. Specia, M. Turchi, et K. Verspoor (éds.), *Proceedings of the Third Conference on Machine Translation: Research Papers: Proceedings of the Third Conference on Machine Translation: Research Papers*. [10.18653/v1/W18-6320](#)
- Hunt, M. J. (août 1990). Figures of Merit for Assessing Connected-Word Recognisers. *Speech Communication*, 9(4), 329-336. [10.1016/0167-6393\(90\)90008-W](#)
- Jaskowski, S., Chava, S., et Shah, A. (septembre 2024). *BERTScoreVisualizer: A Web Tool for Understanding Simplified Text Evaluation with BERTScore* (Numéro arXiv:2409.17160). arXiv. [10.48550/arXiv.2409.17160](#)
- Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press. [10.1017/CBO9780511815829](#)

- Landis, J. R., et Koch, G. G. (mars 1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174.
- Papineni, K., Roukos, S., Ward, T., et Zhu, W.-J. (juillet 2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, et D. Lin (éds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. **10.3115/1073083.1073135**
- Popović, M. (décembre 2020). Informative Manual Evaluation of Machine Translation Output. In D. Scott, N. Bel, et C. Zong (éds.), *Proceedings of the 28th International Conference on Computational Linguistics: Proceedings of the 28th International Conference on Computational Linguistics*. **10.18653/v1/2020.coling-main.444**
- Rei, R., Stewart, C., Farinha, A. C., et Lavie, A. (novembre 2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, et Y. Liu (éds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. **10.18653/v1/2020.emnlp-main.213**
- Scarton, C., et Specia, L. (mai 2016). A Reading Comprehension Corpus for Machine Translation Evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, et S. Piperidis (éds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16): Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.
- Vanroy, B., Tezcan, A., et Macken, L. (juin 2023). MATEO: MACHine Translation Evaluation Online. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 499–500.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., et Artzi, Y. (avril 2020). BERTScore: Evaluating Text Generation with BERT. *Eighth International Conference on Learning Representations*.

# Remerciements

- ♦ Pablo Ruiz Fabo pour le contenu de certaines diapositives.