LYSET08 – Analyse linguistique des modèles de langues

Diaporama – évaluation génération de texte, f-score

Andrea Briglia

andrea.briglia@sorbonne-nouvelle.fr

Crédits : M. Wauquier

Générer du texte

- Générer du texte n'est pas compliqué.
- *Hello world* est l'output traditionnel des programmes simples
  - Techniquement, ces programmes génèrent du texte, à savoir la séquence "Hello world"
  - Mais ça reste le niveau 0 de la génération de texte

Générer du texte

- Générer du texte n'est pas compliqué.
- *Hello world* est l'output traditionnel des programmes simples
  - Techniquement, ces programmes génèrent du texte, à savoir la séquence "Hello world"
  - Mais ça reste le niveau 0 de la génération de texte

Générer du texte

- Générer du texte n'est pas compliqué.
- *Hello world* est l'output traditionnel des programmes simples
  - Techniquement, ces programmes génèrent du texte, à savoir la séquence "Hello world"
  - Mais ça reste le niveau 0 de la génération de texte

Générer du texte

- Générer du texte n'est pas compliqué.
- *Hello world* est l'output traditionnel des programmes simples
  - Techniquement, ces programmes génèrent du texte, à savoir la séquence "Hello world"
  - Mais ça reste le niveau 0 de la génération de texte

Génération automatique aléatoire

- Générer automatiquement du contenu n’est pas difficile en soit
  - Pour rappel, Mozart proposait déjà un système pour générer aléatoirement de la musique
  - Cette approche est applicable à n’importe quel type de contenu, y compris textuel
    - Il suffit d’associer un mot à un nombre et de sélectionner aléatoirement *n* nombres
  - Testons

Génération automatique aléatoire

- Générer automatiquement du contenu n'est pas difficile en soit
  - Pour rappel, Mozart proposait déjà un système pour générer aléatoirement de la musique
  - Cette approche est applicable à n'importe quel type de contenu, y compris textuel
    - Il suffit d'associer un mot à un nombre et de sélectionner aléatoirement *n* nombres
  - Testons

Génération automatique aléatoire

- Générer automatiquement du contenu n'est pas difficile en soit
  - Pour rappel, Mozart proposait déjà un système pour générer aléatoirement de la musique
  - Cette approche est applicable à n'importe quel type de contenu, y compris textuel
    - Il suffit d'associer un mot à un nombre et de sélectionner aléatoirement *n* nombres
  - Testons

Génération automatique aléatoire

- Générer automatiquement du contenu n'est pas difficile en soit
  - Pour rappel, Mozart proposait déjà un système pour générer aléatoirement de la musique
  - Cette approche est applicable à n'importe quel type de contenu, y compris textuel
    - Il suffit d'associer un mot à un nombre et de sélectionner aléatoirement *n* nombres
  - Testons

Génération automatique aléatoire

- Générer automatiquement du contenu n'est pas difficile en soit
  - Pour rappel, Mozart proposait déjà un système pour générer aléatoirement de la musique
  - Cette approche est applicable à n'importe quel type de contenu, y compris textuel
    - Il suffit d'associer un mot à un nombre et de sélectionner aléatoirement *n* nombres
  - Testons

Génération semi-aléatoire

- La génération aléatoire est facile, peu coûteuse, mais terriblement problématique
- Le texte généré souffre de grosses lacunes linguistiques
  - Il est nécessaire d'insuffler quelques connaissances linguistiques
- Testons de nouveau

Génération semi-aléatoire

- La génération aléatoire est facile, peu coûteuse, mais terriblement problématique
- Le texte généré souffre de grosses lacunes linguistiques
  - Il est nécessaire d'insuffler quelques connaissances linguistiques
- Testons de nouveau

Génération semi-aléatoire

- La génération aléatoire est facile, peu coûteuse, mais terriblement problématique
- Le texte généré souffre de grosses lacunes linguistiques
  - Il est nécessaire d'insuffler quelques connaissances linguistiques
- Testons de nouveau

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Le bon et le mauvais texte généré

- Générer du texte n'est dans l'absolu pas difficile
  - On peut facilement avoir un niveau satisfaisant de résultats pour un coût limité
  - Mais la 'créativité' proposée par ces systèmes est relativement limitée
    - Le système est contraint par la taille du lexique, les règles définies, etc.
- Le véritable enjeu des systèmes de NLG est la qualité du texte généré
  - C'est d'autant plus important en fonction des applications visées
    - Pour un chatbot conversationnel, il faut que le texte soit logique et crédible
    - Pour un chatbot informationnel, il faut que l'information fournie soit claire et juste
    - etc

Nécessité d’une évaluation

- Comme pour toute application ou tâche, il est primordial d’évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d’un texte généré ?

Nécessité d'une évaluation

- Comme pour toute application ou tâche, il est primordial d'évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d'un texte généré ?

Nécessité d'une évaluation

- Comme pour toute application ou tâche, il est primordial d'évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d'un texte généré ?

Nécessité d'une évaluation

- Comme pour toute application ou tâche, il est primordial d'évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d'un texte généré ?

Nécessité d'une évaluation

- Comme pour toute application ou tâche, il est primordial d'évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d'un texte généré ?

Nécessité d’une évaluation

- Comme pour toute application ou tâche, il est primordial d’évaluer un système de NLG
  - Comprendre dans quelle mesure le système marche
  - Comprendre ce qui ne marche pas
  - Comprendre pourquoi ça ne marche pas
  - Comprendre comment améliorer le système
- Comment évaluer la qualité d’un texte généré ?

Illustration

- Pour chaque extrait de l’annexe 3, identifiez les forces et faiblesses des textes générés

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l’extrait 1

- Une efficacité redoutable
  - Transformation des amorces pour garantir la grammaticalité et l’idiomaticité de la phrase
    - *don’t* devient *simply never*
    - *I killed my boss* devient *I, too, killed my boss*
  - Mise en page
  - Gestion sémantique d’une locution figée métaphorique
    - En surface du moins, nous n’avons pas suffisamment de contexte pour s’assurer que n’est pas le fruit du hasard
- Avec quelques limites
  - L’adaptation des amorces n’est pas complètement aboutie
    - Emploi surprenant du pronom personnel à la 1e personne dans *you can kill my boss* et *go kill my boss*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 2

- Un extrait crédible
  - L'extrait présente un style littéraire plutôt proche de celui de J.K. Rowling
  - On retrouve la terminologie de la saga *Harry Potter* (*Death Eaters*, *Hagrid's hut*)
- Mais en surface uniquement
  - L'intégration de nouveaux personnages et/ou éléments non vus est relativement limitée
    - On retrouve beaucoup d'éléments basiques : *spider*, *Ron*, *Harry*, *Voldemort*, *Death Eaters*
    - Un nouveau personnage est proposé (Mr. Staircase, *a shabby-robed ghost*), mais ce n'est pas une nouvelle créature, et la création patronymique se limite à de l'antonomase inversée
  - Le texte contient de nombreux problèmes linguistiques, pragmatiques et/ou stylistiques
    - *Ron's Ron shirt was just as bad as Ron himself.*
    - *He saw Harry and immediately began to eat Hermione's family.*
    - *The password was "BEEF WOMEN," Hermione cried.*

Analyse de l'extrait 3

- Une performance globalement correcte
  - La syntaxe est bonne
    - Le texte contient des tournures anglophones idiomatiques (*So am I*, *Standing behind him was a giant elephant*
    - On observe une cohérence temporelle, avec du passé tout du long
  - Le contenu du texte est 'cohérent'
    - Positionnement et alternance des personnages
  - Mais quelques soucis demeurent, notamment pragmatiques
    - La relation entre *woman* et *elephant* de type *husband* est étrange
- L'outil offre par ailleurs la possibilité de moduler le texte
  - Longueur
  - Mots à intégrer

Analyse de l’extrait 3

- Une performance globalement correcte
  - La syntaxe est bonne
    - Le texte contient des tournures anglophones idiomatiques (*So am I*, *Standing behind him was a giant elephant*
    - On observe une cohérence temporelle, avec du passé tout du long
  - Le contenu du texte est ‘cohérent’
    - Positionnement et alternance des personnages
  - Mais quelques soucis demeurent, notamment pragmatiques
    - La relation entre *woman* et *elephant* de type *husband* est étrange
- L’outil offre par ailleurs la possibilité de moduler le texte
  - Longueur
  - Mots à intégrer

Analyse de l’extrait 3

- Une performance globalement correcte
  - La syntaxe est bonne
    - Le texte contient des tournures anglophones idiomatiques (*So am I*, *Standing behind him was a giant elephant*
    - On observe une cohérence temporelle, avec du passé tout du long
  - Le contenu du texte est ‘cohérent’
    - Positionnement et alternance des personnages
  - Mais quelques soucis demeurent, notamment pragmatiques
    - La relation entre *woman* et *elephant* de type *husband* est étrange
- L’outil offre par ailleurs la possibilité de moduler le texte
  - Longueur
  - Mots à intégrer

Analyse de l’extrait 3

- Une performance globalement correcte
  - La syntaxe est bonne
    - Le texte contient des tournures anglophones idiomatiques (*So am I*, *Standing behind him was a giant elephant*
    - On observe une cohérence temporelle, avec du passé tout du long
  - Le contenu du texte est ‘cohérent’
    - Positionnement et alternance des personnages
  - Mais quelques soucis demeurent, notamment pragmatiques
    - La relation entre *woman* et *elephant* de type *husband* est étrange
- L’outil offre par ailleurs la possibilité de moduler le texte
  - Longueur
  - Mots à intégrer

Analyse de l’extrait 3

- Une performance globalement correcte
  - La syntaxe est bonne
    - Le texte contient des tournures anglophones idiomatiques (*So am I*, *Standing behind him was a giant elephant*
    - On observe une cohérence temporelle, avec du passé tout du long
  - Le contenu du texte est ‘cohérent’
    - Positionnement et alternance des personnages
  - Mais quelques soucis demeurent, notamment pragmatiques
    - La relation entre *woman* et *elephant* de type *husband* est étrange
- L’outil offre par ailleurs la possibilité de moduler le texte
  - Longueur
  - Mots à intégrer

Analyse de l’extrait 4

- Une performance bien plus nuancée
  - On observe une certaine cohérence syntaxique (1SG.M) ainsi que thématique (politique)
    - campagne électoral, chômage, mesures, élection présidentielle, tête de liste
  - On est sur un entre-deux sur le plan stylistique, entre texte littéraire et discours politique
    - *c’est l’espoir, la folie, la sérénité d’analyser la situation en quelques cieux et en quelques manières*
  - On note de nombreux problèmes de grammaticalité et/ou d’idiomaticité
    - *Le chômage à la tête!*
    - *l’espoir est généralement définitive*
    - *matinabondage*
    - *se lancer dans un grand cachet*
    - *il me semble certainement qu’une grande cause affiche autant d’insouciance*
    - *seront nécessaires et nécessaires*

Analyse de l’extrait 4

- Une performance bien plus nuancée
  - On observe une certaine cohérence syntaxique (1SG.M) ainsi que thématique (politique)
    - campagne électoral, chômage, mesures, élection présidentielle, tête de liste
  - On est sur un entre-deux sur le plan stylistique, entre texte littéraire et discours politique
    - *c’est l’espoir, la folie, la sérénité d’analyser la situation en quelques cieux et en quelques manières*
  - On note de nombreux problèmes de grammaticalité et/ou d’idiomaticité
    - *Le chômage à la tête!*
    - *l’espoir est généralement définitive*
    - *matinabondage*
    - *se lancer dans un grand cachet*
    - *il me semble certainement qu’une grande cause affiche autant d’insouciance*
    - *seront nécessaires et nécessaires*

Analyse de l’extrait 4

- Une performance bien plus nuancée
  - On observe une certaine cohérence syntaxique (1SG.M) ainsi que thématique (politique)
    - campagne électoral, chômage, mesures, élection présidentielle, tête de liste
  - On est sur un entre-deux sur le plan stylistique, entre texte littéraire et discours politique
    - *c’est l’espoir, la folie, la sérénité d’analyser la situation en quelques cieux et en quelques manières*
  - On note de nombreux problèmes de grammaticalité et/ou d’idiomaticité
    - *Le chômage à la tête!*
    - *l’espoir est généralement définitive*
    - *matinabondage*
    - *se lancer dans un grand cachet*
    - *il me semble certainement qu’une grande cause affiche autant d’insouciance*
    - *seront nécessaires et nécessaires*

Analyse de l’extrait 4

- Une performance bien plus nuancée
  - On observe une certaine cohérence syntaxique (1SG.M) ainsi que thématique (politique)
    - campagne électoral, chômage, mesures, élection présidentielle, tête de liste
  - On est sur un entre-deux sur le plan stylistique, entre texte littéraire et discours politique
    - *c’est l’espoir, la folie, la sérénité d’analyser la situation en quelques cieux et en quelques manières*
  - On note de nombreux problèmes de grammaticalité et/ou d’idiomaticité
    - *Le chômage à la tête!*
    - *l’espoir est généralement définitive*
    - *matinabondage*
    - *se lancer dans un grand cachet*
    - *il me semble certainement qu’une grande cause affiche autant d’insouciance*
    - *seront nécessaires et nécessaires*

Analyse de l'extrait 4

- Une performance bien plus nuancée
  - On observe une certaine cohérence syntaxique (1SG.M) ainsi que thématique (politique)
    - campagne électoral, chômage, mesures, élection présidentielle, tête de liste
  - On est sur un entre-deux sur le plan stylistique, entre texte littéraire et discours politique
    - *c'est l'espoir, la folie, la sérénité d'analyser la situation en quelques cieux et en quelques manières*
  - On note de nombreux problèmes de grammaticalité et/ou d'idiomaticité
    - *Le chômage à la tête!*
    - *l'espoir est généralement définitive*
    - *matinabondage*
    - *se lancer dans un grand cachet*
    - *il me semble certainement qu'une grande cause affiche autant d'insouciance*
    - *seront nécessaires et nécessaires*

Analyses documentaires

- Quelles conclusions tirer de ces quelques extraits ?
  - La nature des textes sur lesquels le système est entraîné a un impact très fort
    - Thématiquement
    - Linguistiquement
  - La grammaticalité pose parfois problème
  - La créativité des systèmes (dans son sens le plus artistique) est limitée

Analyses documentaires

- Quelles conclusions tirer de ces quelques extraits ?
  - La nature des textes sur lesquels le système est entraîné a un impact très fort
    - Thématiquement
    - Linguistiquement
  - La grammaticalité pose parfois problème
  - La créativité des systèmes (dans son sens le plus artistique) est limitée

Analyses documentaires

- Quelles conclusions tirer de ces quelques extraits ?
  - La nature des textes sur lesquels le système est entraîné a un impact très fort
    - Thématiquement
    - Linguistiquement
  - La grammaticalité pose parfois problème
  - La créativité des systèmes (dans son sens le plus artistique) est limitée

Analyses documentaires

- Quelles conclusions tirer de ces quelques extraits ?
  - La nature des textes sur lesquels le système est entraîné a un impact très fort
    - Thématiquement
    - Linguistiquement
  - La grammaticalité pose parfois problème
  - La créativité des systèmes (dans son sens le plus artistique) est limitée

Analyses documentaires

- Quelles conclusions tirer de ces quelques extraits ?
  - La nature des textes sur lesquels le système est entraîné a un impact très fort
    - Thématiquement
    - Linguistiquement
  - La grammaticalité pose parfois problème
  - La créativité des systèmes (dans son sens le plus artistique) est limitée

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

Enjeux

- Parmi les principaux enjeux pour la NLG, on retrouve :
  - Les contraintes syntaxiques
    - Grammaticalité
  - Contraintes sémantiques
    - Richesse lexicale
    - Gestion de la métaphore et de l'analogie
  - Contraintes stylistiques
    - Coréférences et anaphores, cohérence discursive
- Comment évaluer la qualité d'un texte généré ?

De l'évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs 'réponses' (c'est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d'évaluation peut se faire
  - Sur la base de l'objet que l'on évalue
  - Sur la base des types d'approches

De l'évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs 'réponses' (c'est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d'évaluation peut se faire
  - Sur la base de l'objet que l'on évalue
  - Sur la base des types d'approches

De l’évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs ‘réponses’ (c’est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d’évaluation peut se faire
  - Sur la base de l’objet que l’on évalue
  - Sur la base des types d’approches

De l'évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs 'réponses' (c'est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d'évaluation peut se faire
  - Sur la base de l'objet que l'on évalue
  - Sur la base des types d'approches

De l’évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs ‘réponses’ (c’est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d’évaluation peut se faire
  - Sur la base de l’objet que l’on évalue
  - Sur la base des types d’approches

De l’évaluation des textes générés

- Le NLG est une tâche dite *open-ended*
  - Plusieurs ‘réponses’ (c’est-à-dire textes) sont possibles/plausibles
- Cela rend son évaluation compliquée
- La typologie des méthodes d’évaluation peut se faire
  - Sur la base de l’objet que l’on évalue
  - Sur la base des types d’approches

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité…)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité…)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité…)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité…)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité...)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité...)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité…)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité...)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité...)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Ce que l'on évalue

- Quand on parle de NLG, on parle à la fois d'un (sous)système et d'un résultat
- De fait, ces deux facettes mènent à deux types d'évaluation
  - Intrinsèque – évaluation de l'output du système selon des critères prédéfinis au regard de ce qui est attendu du système
    - En l'occurrence, on évalue la qualité du texte généré, dans son ensemble ou selon certains critères (cohérence, grammaticalité...)
    - Ces évaluations reposent généralement, d'une façon ou d'une autre, sur une évaluation humaine
    - Elles sont jugées plus faciles à mettre en place
  - Extrinsèque – évaluation de la performance du système ou de l'apport du module concerné dans une tâche.
    - En l'occurrence, on évalue si le texte généré de telle ou telle façon améliore les résultats du système dans une tâche (par exemple de recherche d'information ou de résumé automatique)
    - C'est l'approche la plus objective dans une approche TAL, mais elle implique l'évaluation d'un système entier
    - Ces évaluations sont donc davantage dépendantes des tâches elles-mêmes

Comment on l'évalue

- Indépendamment de ce que l'on évalue, plusieurs types de méthodes s'offrent à nous
  - Mesures '*human-centric*'
  - Mesures automatiques
  - Mesures automatiques par apprentissage (*machine-learned*)

Comment on l'évalue

- Indépendamment de ce que l'on évalue, plusieurs types de méthodes s'offrent à nous
  - Mesures '*human-centric*'
  - Mesures automatiques
  - Mesures automatiques par apprentissage (*machine-learned*)

Comment on l'évalue

- Indépendamment de ce que l'on évalue, plusieurs types de méthodes s'offrent à nous
  - Mesures '*human-centric*'
  - Mesures automatiques
  - Mesures automatiques par apprentissage (*machine-learned*)

Comment on l'évalue

- Indépendamment de ce que l'on évalue, plusieurs types de méthodes s'offrent à nous
  - Mesures '*human-centric*'
  - Mesures automatiques
  - Mesures automatiques par apprentissage (*machine-learned*)

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk (réf fin diaporama, mais aussi Prolific) et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine

- Cette évaluation constitue souvent la base de toute évaluation ultérieure
  - On parle de *gold standard*
- Cette approche est souvent préférée car elle est supposée être plus qualitative
  - On demande à des locuteurs, voire à des experts, de statuer sur certains éléments
  - C'est donc particulièrement coûteux et chronophage
    - La démocratisation des Amazon Mechanical Turk et autres systèmes de crowdsourcing rend cette solution plus accessible, mais non sans son lot de problème
  - Cette approche n'est pas sans défaut
    - Comment définir et objectiver la tâche ?
    - Comment uniformiser les évaluations ?
    - L'accord inter-annotateur permet de mesurer mais pas d'unifier l'annotation humaine
  - L'évaluation humaine couvre deux approches
    - Intrinsèque
    - Extrinsèque

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, overfitting
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Évaluation humaine intrinsèque

- On peut vouloir évaluer la qualité globale d'un texte ou un aspect spécifique
  - *fiuency*
    - grammaticalité, style, registre, ton, typicité, répétitivité
  - *adequacy* (notamment en traduction automatique)
  - *factuality* (notamment pour les résumés automatiques)
  - Ces mesures (et leur terminologie) sont assez variables, et les uniformiser n'est pas nécessairement souhaitable
    - Rigidité, *overfitting*, compromis (*trade-off*) entre précision et robustesse
- On peut procéder à une évaluation
  - Absolue - on évalue un texte seul
  - Relative - on compare deux ou plusieurs textes

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation (ex classification simple, *Likert scale*)
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accords inter-annotateurs

- Ces mesures visent à évaluer le degré d'accord des annotateurs
- Ces mesures peuvent varier en fonction du nombre d'annotateurs, de la nature de l'annotation
- L'interprétation de leur score (allant généralement de 0 à 1) n'est pas toujours évidente
- Parmi les plus utilisées, on trouve
  - L'accord brut
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux
  - Kappa de Cohen
    - C'est le pourcentage de cas où les annotateurs sont d'accord entre eux, au regard de l'accord attendu
  - F-score
    - Cette mesure permet d'évaluer l'accord lorsqu'il ne s'agit pas d'une tâche de catégorisation

Accord brut

- L'accord se calcule à l'aide d'une matrice de confusion

Accord =
(10+40) / 80 =

50/80 =
0,63 (63%)

JUGE 2 \ JUGE 1	oui	non	total
oui	10	20	30
non	10	40	50
total	20	60	80

Kappa de Cohen

- Cette mesure intègre la notion d'accord par chance, à l'aide de la formule :

$$F = \frac{A_{observ} - A_{attendu}}{1 - A_{attendu}}$$

Observé :

Ao = (10+40) / 80 = 0,63

JUGE 1

JUGE 2	oui	Non	total
oui	10	20	30
non	10	40	50
total	20	60	80

Attendu :

Aa = (7,5+37,5) / 80 = 0,56

JUGE 1

JUGE 2	oui	Non	total
oui	20x30/ 80 = 7,5	60x30/ 80= 22,5	30
non	12,5	37,5	50
total	20	60	80

K = (0,63 - 0,56)/(1- 0,56) = 0,29 (passable)

F-score

- Prenons une tâche où deux annotateurs doivent identifier des phrases agrammaticales
- L'accord se calculera selon la formule :

$$F = \frac{2xN_{commun}}{N_{annot1}+N_{annot2}}$$

Où N_{commun} est le nombre de phrases sélectionnées par les deux annotateurs, et N_{annot1} et N_{annot2} le nombre de phrases sélectionnées respectivement par l'annotateur 1 et l'annotateur 2 uniquement

- Si l'annotateur 1 a identifié les phrases **P1**, P3, **P5**, et l'annotateur 2 les phrases **P1**, P2, P4, **P5** et P6, on a donc la formule :

$$F = \frac{2x2}{3+5} = 0.5$$

- Le F-score peut aussi se décrire en termes de précision (combien d'éléments sélectionnés sont pertinents ?) et de rappel (combien d'éléments pertinents sont sélectionnés ?)

Évaluation humaine extrinsèque

- On peut évaluer la performance du système en termes de
  - Performance de l'utilisateur dans une tâche donnée
    - Dans quelle mesure l'utilisateur tire un bénéfice du système (procède-t-il à une meilleure prise de décision, fait-il moins d'erreurs, etc.)
  - Performance du système à réaliser sa tâche
    - Dans quelle mesure le système complète sa tâche (y a-t-il moins de post-editing à faire en aval, etc.)

Évaluation humaine extrinsèque

- On peut évaluer la performance du système en termes de
  - Performance de l'utilisateur dans une tâche donnée
    - Dans quelle mesure l'utilisateur tire un bénéfice du système (procède-t-il à une meilleure prise de décision, fait-il moins d'erreurs, etc.)
  - Performance du système à réaliser sa tâche
    - Dans quelle mesure le système complète sa tâche (y a-t-il moins de post-editing à faire en aval, etc.)

Évaluation humaine extrinsèque

- On peut évaluer la performance du système en termes de
  - Performance de l'utilisateur dans une tâche donnée
    - Dans quelle mesure l'utilisateur tire un bénéfice du système (procède-t-il à une meilleure prise de décision, fait-il moins d'erreurs, etc.)
  - Performance du système à réaliser sa tâche
    - Dans quelle mesure le système complète sa tâche (y a-t-il moins de post-editing à faire en aval, etc.)

Évaluation humaine extrinsèque

- On peut évaluer la performance du système en termes de
  - Performance de l'utilisateur dans une tâche donnée
    - Dans quelle mesure l'utilisateur tire un bénéfice du système (procède-t-il à une meilleure prise de décision, fait-il moins d'erreurs, etc.)
  - Performance du système à réaliser sa tâche
    - Dans quelle mesure le système complète sa tâche (y a-t-il moins de post-editing à faire en aval, etc.)

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

- Celikyiklaz *et al.* (2021) proposent 5 grands groupes de mesures
  - les mesures de recouvrement des *n*-grammes (n-*gram overlap*)
    - Ces mesures évaluent le degré de correspondance entre des textes générés automatiquement et des textes de référence

    ex. F-score (ou F1), BLEU, ROUGE…
  - les mesures de distance
    - Ces mesures évalue la similarité entre deux unités (mots, phrases…), sur la base de la distance d'édition (le nombre de modifications à faire pour retomber sur la phrase de référence) ou à partir de la distance entre représentations vectorielles
  - les mesures de diversité
    - Ces mesures évaluent la diversité (ou la richesse) lexicale d'un texte

    ex. nombre de n-gram distincts, Type-Token ratio, SELF-BLEU
  - les mesures de recouvrement de contenu (*content overlap*)
    - Ces mesures évaluent la similarité d'un texte généré à un texte de référence en extrayant les informations sémantiques explicites
  - les mesures basées sur des critères grammaticaux
    - Ces mesures évaluent la similarité syntaxique d'un texte généré et d'un texte de référence.

Évaluation automatique

	Metric	Property	MT	IC	SR	SUM	DG	QG	RG
n-gram overlap	F-SCORE	precision and recall	✓	✓	✓	✓	✓	✓	✓
	BLEU	*n*-gram precision	✓	✓			✓	✓	✓
	METEOR	*n*-gram w/ synonym match	✓	✓			✓		
	CIDER	*tf-idf* weighted *n*-gram sim.		✓					
	NIST	*n*-gram precision	✓						
	GTM	*n*-gram metrics	✓						
	HLEPOR	unigrams harmonic mean	✓						
	RIBES	unigrams harmonic mean							
	MASI	attribute overlap							
	WER	% of insert,delete, replace			✓				
	TER	translation edit rate	✓						
	ROUGE	*n*-gram recall				✓	✓		
	DICE	attribute overlap							
distance-based	EDIT DIST.	cosine similarity	✓	✓	✓	✓	✓	✓	
	MEANT 2.0	vector based similarity	✓						
	YISI	weighted similarity	✓						
	WMD	EMD on words		✓		✓			
	SMD	EMD on sentences		✓	✓	✓			
	FRÉCHET	distributional similarity		✓					
content over-lap	PYRAMID	content selection			✓				
	SPICE	scene graph similarity		✓					
	SPIDER	scene graph similarity		✓					
diversity	TTR	richness of vocabulary			✓				
	SELF-BLEU	*n*-gram precision						✓	

Table 1: Untrained *automatic* and *retrieval-based* metrics based on string match, string distance, or context overlap. The acronyms for some of the NLP sub-research fields that each metric is commonly used to evaluate text generation are: **MT**: Machine Translation, **IC**: Image Captioning, **SR**: Speech Recognition, **SUM**: Summarization, **DG**: Document or Story Generation, Visual-Story Generation, **QG**: Question Generation, **RG**: Dialog Response Generation. EMD:Earth Mover's Distance; Sim.: Similarity.

Évaluation automatique

- Ces mesures permettent de mesurer l'efficacité d'un modèle à générer un texte
  - De façon rapide
  - De façon efficace
- Elles ne sont cependant pas adaptées à l'évaluation de systèmes *open-ended*
  - Le recouvrement entre un texte de référence et un texte généré est bien moins significatif lorsqu'il y a plusieurs outcomes possibles

Évaluation automatique

- Ces mesures permettent de mesurer l'efficacité d'un modèle à générer un texte
  - De façon rapide
  - De façon efficace
- Elles ne sont cependant pas adaptées à l'évaluation de systèmes *open-ended*
  - Le recouvrement entre un texte de référence et un texte généré est bien moins significatif lorsqu'il y a plusieurs outcomes possibles

Évaluation automatique

- Ces mesures permettent de mesurer l'efficacité d'un modèle à générer un texte
  - De façon rapide
  - De façon efficace
- Elles ne sont cependant pas adaptées à l'évaluation de systèmes *open-ended*
  - Le recouvrement entre un texte de référence et un texte généré est bien moins significatif lorsqu'il y a plusieurs outcomes possibles

Évaluation automatique

- Ces mesures permettent de mesurer l'efficacité d'un modèle à générer un texte
  - De façon rapide
  - De façon efficace
- Elles ne sont cependant pas adaptées à l'évaluation de systèmes *open-ended*
  - Le recouvrement entre un texte de référence et un texte généré est bien moins significatif lorsqu'il y a plusieurs outcomes possibles

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation automatique à l'aide de *machine learning*

- Celikyiklaz *et al.* (2021) évoquent plusieurs approches
  - Les approches à base d'embeddings
    - On crée des représentations vectorielles du sens d'un texte généré à un texte de référence pour en comparer la similarité sémantique
    - Une branche spécifique de ces mesures repose notamment sur l'utilisation de BERT
  - Les approches à base de régression
    - On traite l'évaluation d'un texte comme une tâche de régression, c'est-à-dire comme la prédiction d'une valeur mesurant la qualité d'un texte
  - Les approches basées sur les jugements humains
    - On entraîne des modèles à imiter l'évaluation humaine

Évaluation

- Les méthodes peuvent être croisées
- Certaines sont développées pour des aspects en particulier (fluency, adequacy…)
- Les challenges (ou shared tasks) sont aussi des bons moyens de permettre des évaluations à grande échelle
- Les mesures et approches varient aussi en fonction des tâches ou applications

Évaluation

- Les méthodes peuvent être croisées
- Certaines sont développées pour des aspects en particulier (fluency, adequacy…)
- Les challenges (ou shared tasks) sont aussi des bons moyens de permettre des évaluations à grande échelle
- Les mesures et approches varient aussi en fonction des tâches ou applications

Évaluation

- Les méthodes peuvent être croisées
- Certaines sont développées pour des aspects en particulier (fluency, adequacy…)
- Les challenges (ou shared tasks) sont aussi des bons moyens de permettre des évaluations à grande échelle
- Les mesures et approches varient aussi en fonction des tâches ou applications

Évaluation

- Les méthodes peuvent être croisées
- Certaines sont développées pour des aspects en particulier (fluency, adequacy…)
- Les challenges (ou shared tasks) sont aussi des bons moyens de permettre des évaluations à grande échelle
- Les mesures et approches varient aussi en fonction des tâches ou applications

Oui mais ?

- L’évaluation vise à mesurer la qualité d’un texte
  - Pour l’améliorer et le rendre aussi *’human-like’* que possible
  - Pour identifier sa nature de texte généré
- L’évaluation seule ne peut suffire
  - La qualité des corpus est primordiale
    accuracy, timeliness, completeness, cleanness, unbiasedness

Oui mais ?

- L'évaluation vise à mesurer la qualité d'un texte
  - Pour l'améliorer et le rendre aussi '*human-like*' que possible
  - Pour identifier sa nature de texte généré
- L'évaluation seule ne peut suffire
  - La qualité des corpus est primordiale
    accuracy, timeliness, completeness, cleanness, unbiasedness

Oui mais ?

- L'évaluation vise à mesurer la qualité d'un texte
  - Pour l'améliorer et le rendre aussi '*human-like*' que possible
  - Pour identifier sa nature de texte généré
- L'évaluation seule ne peut suffire
  - La qualité des corpus est primordiale
    accuracy, timeliness, completeness, cleanness, unbiasedness

Oui mais ?

- L’évaluation vise à mesurer la qualité d’un texte
  - Pour l’améliorer et le rendre aussi ’*human-like*’ que possible
  - Pour identifier sa nature de texte généré
- L’évaluation seule ne peut suffire
  - La qualité des corpus est primordiale
    accuracy, timeliness, completeness, cleanness, unbiasedness

Oui mais ?

- L'évaluation vise à mesurer la qualité d'un texte
  - Pour l'améliorer et le rendre aussi '*human-like*' que possible
  - Pour identifier sa nature de texte généré
- L'évaluation seule ne peut suffire
  - La qualité des corpus est primordiale

    accuracy, timeliness, completeness, cleanness, unbiasedness

Oui mais ?

- L’évaluation vise à mesurer la qualité d’un texte
  - Pour l’améliorer et le rendre aussi ’*human-like*’ que possible
  - Pour identifier sa nature de texte généré
- L’évaluation seule ne peut suffire
  - La qualité des corpus est primordiale
    accuracy, *timeliness*, *completeness*, *cleanness*, *unbiasedness*

En pratique

- Dans les articles sur lesquels vous avez travaillé,
  - Quelles évaluations sont mises en place ?
    - En cas d'évaluation humaine, le nombre d'annotateurs est-il indiqué ? La tâche précisée ? L'accord inter-annotateur donné ?
  - Quelles mesures sont utilisées ?
    - En quoi cette mesure est pertinente ? Quels éléments sont donnés pour aider à leur interprétation ?

Pour en savoir plus sur Amazon Mechanical Turk :

Karen Fort, Gilles Adda, Kevin Bretonnel Cohen. Amazon Mechanical Turk: Gold Mine or Coal Mine ?. *Computational Linguistics*, 2011, 37 (2), pp.413-420.