Corpus, ressources et linguistique outillée

Annotation et Qualité des Corpus

Dr Wiem Takrouni
Sorbonne Université
Master Langue et Informatique
UFR Sociologie et Informatique pour les Sciences Humaines

March 28, 2025

Overview

1. Introduction à l'Annotation des Corpus
2. Formats et Standards d'Annotation
3. Méthodes d'Annotation : Manuelle vs. Automatique
4. Outils d'Annotation
5. Cas Pratiques : Annotation en Action
6. Vers une Annotation à Grande Échelle

Objectif

- Objectif : Comprendre le rôle de l'annotation dans le traitement des corpus et les enjeux de qualité des données.

Pourquoi annoter un corpus ?

Définition d'un corpus annoté

Un corpus annoté est un ensemble de textes où des **informations linguistiques** ont été ajoutées.

Pourquoi l'annotation est essentielle ?

- **Améliorer l'apprentissage** des modèles d'IA
  (ex : un modèle de classification a besoin d'exemples annotés pour s'entraîner).
- **Faciliter l'analyse linguistique**
  (ex : repérer automatiquement les verbes dans un texte).
- **Aider les modèles de NLP** à mieux comprendre le texte.

Exemple

"Les annotations entraînent des modèles comme BERT ou GPT pour des tâches spécifiques."

Types d'annotations

Annotation Morphosyntaxique

Identification des **catégories grammaticales** des mots (*POS-tagging*).

Exemple

"chat" $\rightarrow$ **NOM**, "mange" $\rightarrow$ **VER**, "rouge" $\rightarrow$ **ADJ**

Modèles utilisés

SpaCy, **Stanza**, **TreeTagger**

Annotation Sémantique

Sémantique

Ajout d'informations sur le **sens des mots** et les **entités nommées**.

Exemple

"Barack Obama vit aux États-Unis."

Barack Obama → **PERSON**

États-Unis → **LOC (Lieu)**

Applications

Utilisé pour : la reconnaissance d'entités nommées (NER), la recherche d'information.

Modèles et outils

SpaCy, **Flair**, **Hugging Face (BERT NER)**

Annotation Pragmatique

Pragmatique

Analyse des **relations entre phrases** et expressions (ex : anaphores).

Exemple

"Marie a acheté un livre. Elle l'adore."

"Elle" $\rightarrow$ fait référence à *"Marie"*

"l'" $\rightarrow$ fait référence à *"un livre"*

Applications

Utilisé pour : la **résolution des coréférences** (relier des pronoms à leur référent).

Modèles et outils

CoreNLP, **NeuralCoref**

Introduction à l'Annotation des Corpus

Annotation = données d'entraînement pour l'IA

1. Les modèles d'IA (comme BERT) ont besoin de **milliers de phrases annotées** pour apprendre.

Exemple : Un modèle qui fait la reconnaissance des entités nommées (NER) a besoin de textes où les **entités (noms, lieux...)** sont déjà identifiées.

Amélioration des modèles NLP

1. Un bon corpus annoté **réduit les erreurs des modèles**.

Exemple : Un modèle de **traduction automatique** (ex. : Google Translate) s'améliore grâce aux corpus annotés.

Lien entre annotation et Machine Learning

1. En **apprentissage supervisé**, on a besoin d'un **corpus annoté** pour entraîner un modèle.

Exemple : Un **classificateur de sentiment** doit être entraîné sur un corpus avec des étiquettes "positif" ou "négatif".

Exemple concret - Texte brut vs. Texte annoté

1. **Texte brut** (sans annotation):
   Marie a acheté un livre à Paris.
   Elle l'adore !

Texte annoté (morphosyntaxique + sémantique):

Marie	NOM	B-PER
a	VER	O
acheté	VER	O
un	DET	O
livre	NOM	O
à	PRP	O
Paris	NOM	B-LOC
.	PUNCT	O
Elle	PRON	O
l'	PRON	O
adore	VERB	O
!	PUNCT	O

- "Marie" → annoté comme **PERSON**
- "Paris" → annoté comme **LOC** (Lieu)

Lien avec l'IA : Pourquoi c'est utile ?

1. Un modèle d'IA utilise ces annotations pour apprendre à prédire les **étiquettes automatiquement**.

Ces annotations sont utilisées pour des tâches comme :

- **Traduction automatique**
- **Chatbots et assistants vocaux**
- **Recherche d'information avancée**

Formats et Standards d'Annotation

Pourquoi utiliser des formats standards ?

- **Interopérabilité** : Permet aux chercheurs et aux modèles IA de partager et comparer les corpus.
- **Automatisation** : Facilite l'utilisation par des outils comme SpaCy, NLTK, Stanza, Hugging Face.
- **Facilité d'analyse** : Les formats bien structurés permettent d'extraire et de traiter les données plus efficacement.

Exemples de standards les plus utilisés :

- **CoNLL** : Format tabulaire pour l'analyse morphosyntaxique et la reconnaissance d'entités.
- **XML-TEI** : Utilisé pour structurer des corpus complexes.
- **JSON, CSV** : Formats modernes utilisés pour le stockage et le deep learning.

Le format CoNLL (tabulaire, NLP)

Format en colonnes où chaque ligne = un mot annoté.

Très utilisé pour :

- Part-of-Speech tagging (POS)
- Reconnaissance d'entités nommées (NER)
- Dépendances syntaxiques

Exemple d'annotation :

- "Pierre" → NNP (nom propre), B-PER (Personne)
- "Paris" → NNP (nom propre), B-LOC (Lieu)
- "O" = Pas d'étiquette d'entité

```
Pierre  NNP   B-PER
mange   VB    O
une     DET   O
pomme   NN    O
à       PRP   O
Paris   NNP   B-LOC
.       PUNCT O
```

Formats et Standards d'Annotation

Pourquoi utiliser CoNLL ?

- Facile à lire et interpréter
- Compatible avec des outils populaires comme NLTK, SpaCy, Flair
- Utilisé dans les compétitions NLP (CoNLL-2003)

Formats et Standards d'Annotation

Le format XML-TEI (structuration avancée des corpus)

- Un standard XML pour annoter et structurer les corpus textuels.
- Utilisé en linguistique, édition numérique, archives historiques.

```xml
<text>
  <body>
    <p>
      <persName>Pierre</persName> mange <rs type="food">une pomme</rs> à
      <placeName>Paris</placeName>.
    </p>
  </body>
</text>
```

Explication :

- `<persName>` → Marque une personne
- `<placeName>` → Marque un lieu
- `<rs type="food">` → Marque une nourriture

Formats et Standards d'Annotation

Pourquoi utiliser XML-TEI ?

- **Hiérarchisation des données** : Permet de structurer des données complexes.
- **Interopérabilité** : Utilisé dans les projets de numérisation de textes historiques, favorise l'échange de données.
- **Compatible avec XSLT** : Permet des transformations de données via XSLT (Extensible Stylesheet Language Transformations).

Formats et Standards d'Annotation

JSON et CSV (Modernes, IA Deep Learning)

Pourquoi les utiliser ?

- Lisible par IA et compatible Python (pandas, TensorFlow, Hugging Face)
- Idéal pour bases de données, APIs, NLP, datasets ouverts

Exemple (NER - JSON) :

```
[
  {"text": "Pierre mange une pomme à Paris.",
   "entities": [
     {"start": 0, "end": 6, "label": "PER"},
     {"start": 24, "end": 29, "label": "LOC"}
   ]
  }
]
```

Explication :

- `"text"` : Phrase originale
- `"entities"` : Liste des entités avec positions (`start`, `end`)
- `"label"` : Type d'entité (PER = Personne, LOC = Lieu)

Formats et Standards d'Annotation

Pourquoi JSON ?

- Facile à traiter avec `json.loads()` en Python
- Utilisé pour l'entraînement des modèles IA (ex : datasets Hugging Face)

Comparaison des formats et cas d'usage

Format	Avantages	Inconvénients	Utilisation
CoNLL	Simple, lisible, compatible avec NLP	Pas adapté aux structures complexes	NER, POS-Tagging
XML-TEI	Très structuré, interopérable	Complexe, lourd	Édition numérique, corpus historiques
JSON	Idéal pour le Deep Learning, facile à manipuler	Moins lisible pour l'humain	Annotation pour l'IA (Hugging Face)
CSV	Facilement lisible, traitement rapide	Perte d'informations contextuelles	Stockage des annotations simples

Table: Tableau comparatif des formats

Quand utiliser chaque format ?

- CoNLL → NLP classique (POS, NER, dépendances syntaxiques)
- XML-TEI → Corpus complexes (projets de numérisation)
- JSON → Annotation pour le Machine Learning
- CSV → Annotation rapide et simple

Méthodes d'Annotation : Manuelle vs. Automatique

Pourquoi comparer les méthodes d'annotation ?

Les modèles NLP ont besoin de données annotées pour être efficaces.
Annotation = qualité du corpus = performance du modèle IA.

Trois méthodes principales :

Manuelle : Précise mais coûteuse et lente.
Automatique : Rapide mais risque d'erreurs.
Semi-Automatique : Un compromis entre les deux.

Méthodes d'Annotation : Manuelle vs. Automatique

L'Annotation Manuelle (Précise mais Lente)
Réalisée par des linguistes ou annotateurs humains. Chaque mot, phrase ou document est analysé et marqué manuellement.

Outils utilisés :

- Brat (outil d'annotation textuelle en ligne)
- WebAnno (annotation collaborative)
- Prodigy (interface avec machine learning assisté)

Avantages :

- Très précis
- Adapté aux corpus spécialisés (médical, juridique, etc.)

Inconvénients :

- Coûteux (nécessite des annotateurs formés)
- Lent (peut prendre des semaines/mois pour annoter un grand corpus)

Méthodes d'Annotation : Manuelle vs. Automatique

L'Annotation Automatique (Rapide mais Imparfaite)
Réalisée par des modèles d'IA ou de TAL. Utilise des algorithmes pour prédire les étiquettes des mots.

Outils et modèles utilisés :

- SpaCy (annotation morphosyntaxique et reconnaissance d'entités nommées)
- Stanza (Stanford NLP) (analyse syntaxique avancée)
- Flair (reconnaissance d'entités nommées basée sur deep learning)

Méthodes d'Annotation : Manuelle vs. Automatique

Exemple d'annotation automatique avec SpaCy

```
[
  {"text": "Pierre mange une pomme à Paris.",
   "entities": [
     {"start": 0, "end": 6, "label": "PER"},
     {"start": 24, "end": 29, "label": "LOC"}
   ]
  }
]
```

Sortie :

```
Emmanuel Macron → PER
France → LOC
```

Avantages :

- Très rapide (peut annoter des milliers de textes en quelques minutes)
- Facile à utiliser avec des modèles pré-entraînés

Inconvénients :

- Peut contenir des erreurs et des biais
- Moins fiable sur les textes spécialisés

Annotation Semi-Automatique (Compromis entre les deux)

Définition

L'IA propose des annotations, un annotateur humain les corrige et valide.
Permet de réduire les erreurs des modèles automatiques.

Exemples d'outils

- **Prodigy** : Permet d'ajuster les annotations en temps réel.
- **INCEpTION** : Annotation collaborative assistée par IA.
- **Hugging Face Model-assisted labeling** : Annotation à l'aide de NLP pré-entraînés.

Avantages :

- Plus rapide que l'annotation 100% manuelle
- Réduit le taux d'erreurs des modèles IA

Inconvénients :

- Nécessite quand même une validation humaine
- Pas toujours efficace si le modèle de base est trop imprécis

Quand utiliser quelle méthode ?

Annotation Manuelle

- **Utilisation :** Corpus spécialisés (médical, juridique, recherche)

Annotation Automatique

- **Utilisation :** Grands corpus de textes simples (Wikipedia, journaux)

Annotation Semi-Automatique

- **Utilisation :** Projets où il faut de la précision mais où l'automatisation est nécessaire

Outils d'Annotation

Brat

- Interface web pour annoter du texte.
- Utilisé pour la reconnaissance d'entités nommées et les relations syntaxiques.

Prodigy

- Outil interactif basé sur du machine learning assisté.
- Permet de valider/corriger les annotations proposées par un modèle NLP.

WebAnno

- Outil d'annotation collaborative, utilisé pour les projets multilingues.
- Supporte plusieurs types d'annotations : POS-tagging, NER, relations syntaxiques.

Outils d'Annotation Automatique

SpaCy

- Bibliothèque NLP rapide et efficace.
- Utilisée pour le POS-tagging, NER, dépendances syntaxiques.

Stanza (Stanford NLP)

- Développé par Stanford pour l'analyse syntaxique avancée.
- Fonctionne en multilingue (plus de 70 langues).

Flair

- Outil basé sur le deep learning (modèles neuronaux).
- Très performant sur la reconnaissance d'entités nommées.

Annotation d'un Corpus de Réseaux Sociaux

Problème :

- Langage informel : fautes, abréviations, emojis.
- Difficulté pour les modèles classiques à comprendre ce type de texte.

Solution :

- Utilisation de modèles spécialisés comme *Bertweet* (pré-entraîné sur Twitter).

Prétraitement des données :

- Suppression des emojis / stopwords.
- Normalisation des abréviations ("mtn" $\rightarrow$ "maintenant").

Annotation d'un Corpus Médical

Problème :

- Terminologie spécifique difficile à comprendre pour les modèles génériques.
- Risque d'erreurs critiques dans les annotations (ex : maladies, médicaments).

Solution :

- Utilisation d'outils spécialisés comme :
  - *CTakes* (annotation des textes médicaux avec des concepts UMLS).
  - *MedSpacy* (module NLP médical basé sur SpaCy).

Vers une Annotation à Grande Échelle

Utilisation du Deep Learning pour l'Annotation

Pourquoi utiliser l'IA ?

- Modèles avancés (*BERT*, *CamemBERT*, *XLM-R*) permettent d'annoter des millions de phrases rapidement.
- *Fine-tuning* : Ajuster un modèle sur un corpus spécifique pour améliorer la précision.

Exemple d'annotation avec un modèle pré-entraîné (BERT NER)

```python
from transformers import pipeline
annotateur = pipeline("ner", model="dbmdz/bert-large-cased-finetuned-conll03-english")
texte = "Google a été fondé par Larry Page et Sergey Brin à Stanford."
print(annotateur(texte))
```

```
[{'word': 'Google', 'entity': 'B-ORG'},
 {'word': 'Larry Page', 'entity': 'B-PER'},
 {'word': 'Stanford', 'entity': 'B-LOC'}]
```

Vers une Annotation à Grande Échelle

Annotation Automatique avec ChatGPT et GPT-like

Exemple : Génération d'annotations via ChatGPT

- On peut demander à ChatGPT d'annoter un texte directement.
- Exemple de prompt :

```
Annoter les entités nommées dans ce texte : "Marie est partie à Lyon en train."
```

```
Marie → PERSON
Lyon → LOCATION
```

Risques de cette approche ?

- Biais des modèles → Peut produire des annotations erronées.
- Pas toujours cohérent sur des textes complexes.

Défis et Perspectives de l'Annotation Automatisée

Les défis

- **Problèmes de qualité :** Certains modèles génèrent des erreurs importantes.
- **Dépendance aux corpus d'entraînement :** Un modèle entraîné sur du texte juridique peut être mauvais sur du texte médical.
- **Besoin de validation humaine :** Même avec des modèles avancés, une validation humaine est nécessaire.

Perspectives futures

- Combinaison Deep Learning + correction humaine pour un meilleur équilibre.
- Développement de modèles spécialisés plus précis (ex : BioBERT pour le médical).
- Annotation par apprentissage actif : l'IA propose des annotations et l'humain valide pour améliorer progressivement le modèle.