Extraction automatique de réseaux de personnages : Fondation d'Asimov

Abdelhadi ESSABRI[1]*, Abdou NIANG*[2]

[1]Avignon Université
[2]Avignon Université
email@address

Abstract

Cet article présente une approche informatique pour l'extraction automatique de réseaux de personnages dans la série Fondation d'Isaac Asimov. En utilisant des techniques de Traitement Automatique du Langage Naturel (TALN) et des outils de modélisation de réseaux, notre méthode séquentielle explore les interactions entre les personnages au fil des chapitres. Les résultats sont discutés, et des suggestions d'amélioration sont proposées pour de futures investigations.

Termes clés: traitement du langage naturel, reconnaissance d'entités nommées, modélisation de réseaux, Isaac Asimov, série Fondation, analyse littéraire computationnelle

1. Introduction

Cet article présente une approche informatique novatrice pour l'extraction automatique de réseaux de personnages dans la série Fondation d'Isaac Asimov. En utilisant une séquence de modèles deTraitement Automatique du Langage Naturel (TALN) tels que Spacy, Camembert-NER et Flair. Notre méthode explore de manière séquentielle les interactions entre les personnages à travers les chapitres. L'analyse approfondie des personnages au sein de textes littéraires peut apporter une compréhension riche de la dynamique narrative et des relations interpersonnelles.

Nous mettons en œuvre une approche combinant la modélisation des entités nommées avec ces trois modèles successivement, et la construction de graphes avec NetworkX. L'objectif est d'extraire des informations détaillées sur les personnages, d'identifier d'éventuels alias, et de représenter graphiquement les relations entre eux pour chaque chapitre d'un livre. Notre méthodologie offre une perspective inédite sur la structure narrative, fournissant une base solide pour des analyses plus approfondies et des insights captivants dans le monde complexe des personnages littéraires.

Dans cette perspective, l'utilisation séquentielle de modèles de traitement du langage naturel (NLP) tels que Spacy, Camembert-NER et Flair offre une méthodologie puissante pour détecter et catégoriser les entités nommées, en mettant particulièrement l'accent sur les personnages. Nous explorons comment cette approche peut offrir une perspective inédite sur la saga Fondation, ouvrant ainsi la porte à de futures recherches explorant les relations complexes entre les personnages dans des œuvres littéraires emblématiques.

2. Méthodologie

Nous avons appliqué un pipeline séquentiel sur deux livres de la série Fondation, à savoir **Prelude à Fondation** qui compte 19 chapitres, et **Les Cavernes d'Acier** qui comprend 18 chapitres. Ce processus méthodique englobe plusieurs étapes clés visant à démystifier les interactions entre les personnages au fil des chapitres.

- **Identification des Personnages**: La première étape a consisté en l'identification des entités nommées (REN) en utilisant trois modèles distincts : **Spacy**, **Camembert NER** et **Flair**. Initialement, un modèle Spacy a été employé, mais les résultats obtenus ne se sont pas avérés satisfaisants. Par la suite, le modèle Camembert NER a été introduit, démontrant des performances nettement supérieures à son prédécesseur, suivi du modèle Flair. Les entités de type **Personne** ont été extraites de manière à obtenir une liste exhaustive des personnages, et des résolutions d'alias ont été mises en œuvre pour regrouper les différentes occurrences d'un même personnage.
- **Détection des Interactions**:Les interactions entre les personnages ont été définies en termes de co-occurrences dans le texte. Deux entités sont considérées en co-occurrence lorsqu'elles apparaissent à une distance de **25 tokens ou moins**. Cette approche pragmatique permet de saisir les moments où les personnages partagent l'espace narratif, créant ainsi une base pour une analyse plus approfondie de leurs relations.
- **Construction du Graphe**: Pour chaque chapitre, un graphe a été construit en utilisant les informations extraites par les trois modèles. Chaque nœud de ce graphe représente un personnage, et les liens entre ces nœuds reflètent les interactions détectées. Le poids de chaque lien est déterminé par le nombre de co-occurrences entre les deux personnages. Ainsi, les graphes obtenus illustrent visuellement les réseaux de personnages pour chaque chapitre, offrant une représentation tangible des dynamiques relationnelles au sein de la série Fondation.

2.1. Modèle Spacy

Le fonctionnement de la détection des entités nommées dans SpaCy repose sur l'utilisation de modèles statistiques pré-entraînés et d'algorithmes d'apprentissage automatique. Voici un aperçu du raisonnement technique derrière ce processus :

1. **Modèles statistiques** : SpaCy utilise des modèles statistiques basés sur des réseaux de neurones pour effectuer diverses tâches de traitement du langage naturel. Ces modèles sont entraînés sur de vastes ensembles de données annotées, où les entités nommées sont identifiées et catégorisées.
2. **Entraînement du modèle** : Pendant la phase d'entraînement, le modèle apprend à associer des caractéristiques spécifiques des mots et des contextes aux entités nommées qu'il doit détecter. Cela se fait en ajustant les poids des connexions entre les neurones du réseau de neurones en fonction des exemples d'entraînement fournis.

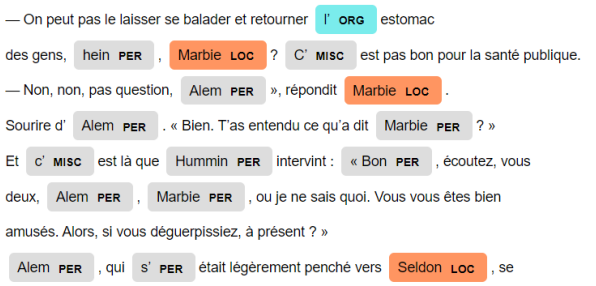

Figure 1: *example de détection des entités nommées avec Spacy*

3. **Représentation du texte** : Lorsqu'un texte est donné au modèle pour l'analyse, le modèle le transforme en une représentation vectorielle qui capture les informations importantes pour la tâche de détection des entités nommées. Cette représentation est basée sur la sémantique et la syntaxe du texte.
4. **Propagation avant du réseau de neurones** : La représentation vectorielle du texte est alors propagée à travers le réseau de neurones pour produire des prédictions sur la présence et le type d'entités nommées dans le texte.
5. **Détection des entités nommées** : Les sorties du réseau de neurones sont ensuite utilisées pour identifier les entités nommées dans le texte. Les entités sont généralement délimitées par des frontières de phrase, et le modèle attribue un label à chaque entité, indiquant son type (par exemple, personne, organisation, lieu).

Le modèle SpaCy rencontre quelques erreurs dans la détection des entités nommées. Quelques points notables incluent des confusions entre les catégories d'entités (par exemple, "l'" pour "ORG"), des erreurs d'attribution de types d'entités (comme **LOC** (lieu) pour des noms de personnes comme **Marbie** et **Seldon**), ainsi que des difficultés avec des expressions peu communes ou ambigües.

2.2. Modèle Camembert NER

Le processus de fonctionnement d'un modèle NER basé sur Camembert ou similaire pourrait être décrit de manière générale comme suit :

1. **Pré-entraînement** : Le modèle est pré-entraîné sur une grande quantité de données textuelles non annotées. Pendant cette phase, il apprend à représenter les relations sémantiques et syntaxiques entre les mots dans le langage naturel.
2. **Fine-tuning** (ajustement fin) : Le modèle est ensuite fine-tuné sur un ensemble de données annotées spécifiques à la tâche de détection des entités nommées en français. Cette étape permet au modèle de s'adapter à des entités spécifiques du français.
3. **Architecture Transformer** : Le modèle repose probablement sur une architecture Transformer, qui permet de prendre en compte les contextes bidirectionnels lors de l'analyse du texte. BERT et ses dérivés, comme Camembert, sont connus pour leur capacité à capturer des informations complexes dans le langage.
4. **Tokenization** : Le texte d'entrée est divisé en "tokens" (unités de base) adaptées au modèle, et ces tokens sont ensuite intégrés dans l'architecture du modèle.
5. **Calcul des scores de probabilité** : Le modèle attribue des scores de probabilité à chaque token pour chaque catégorie d'entité nommée possible (par exemple, personne, lieu, organisation). Ces scores sont calculés en fonction des informations apprises pendant l'entraînement.
6. **Sélection des entités nommées** : Les entités nommées sont finalement sélectionnées en fonction des scores de probabilité attribués à chaque token. Une fois que le modèle détecte des séquences de tokens avec des probabilités élevées pour une catégorie d'entité, il les identifie comme des entités nommées.

L'analyse approfondie des personnages au sein de textes littéraires peut apporter une compréhension riche de la dynamique narrative et des relations interpersonnelles. Dans cette perspective, l'utilisation de modèles de traitement du langage naturel (NLP) tels que Camembert-NER offre une méthodologie puissante pour détecter et catégoriser les entités nommées, en mettant particulièrement l'accent sur les personnages.

Notre script met en œuvre une approche novatrice en combinant la modélisation des entités nommées avec le modèle Camembert-NER et la construction de graphes avec NetworkX. L'objectif est d'extraire des informations détaillées sur les personnages, d'identifier d'éventuels alias et de représenter graphiquement les relations entre eux pour chaque chapitre d'un livre.

Nous explorerons comment cette méthodologie peut offrir une perspective inédite sur la structure narrative, fournissant une base solide pour des analyses plus approfondies et des insights captivants dans le monde complexe des personnages littéraires.

1. **Prétraitement du Texte :** Les chapitres des livres sont prétraités pour extraire le texte à partir de fichiers spécifiques.
2. **Modélisation des Entités Nomées (NER) :** Le modèle Camembert-NER est utilisé pour détecter les entités nommées, en particulier les personnages, marqués en tant que **PER**.
3. **Construction du Graphe avec NetworkX :** Pour chaque chapitre, un graphe est construit avec NetworkX, où les nœuds représentent les personnages et les arêtes indiquent une relation potentielle.
4. **Identification des Alias :** Les entités détectées sont comparées pour identifier d'éventuels alias. La distance entre les mots est utilisée pour déterminer la probabilité qu'ils soient des alias.
5. **Attribution des Alias aux Nœuds :** Les alias identifiés sont attribués aux nœuds du graphe, créant ainsi un regroupement de personnages et de leurs alias.
6. **Création et Visualisation du Graphe :** Les arêtes sont ajoutées en fonction de la distance entre les nœuds, et le graphe résultant est visualisé pour chaque chapitre.
7. **Export des Résultats :** Les données du graphe (au format GraphML) et d'autres informations sont stockées dans un DataFrame et exportées vers un fichier CSV pour la soumission dans Leaderboard de kaggle.

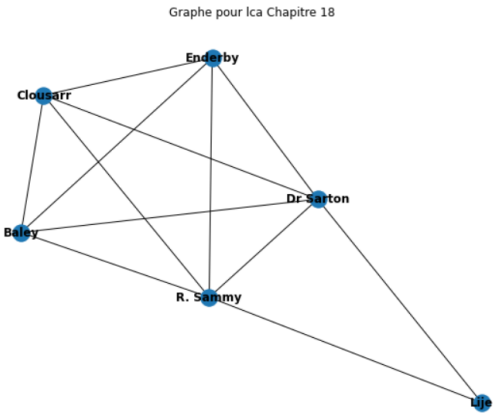

Figure 2: *Graphe des Interactions entre Personnages - Les Cavernes d'Acier, Chapitre 18*

3. Résultats et Discussion

Les graphes obtenus illustrent visuellement les réseaux de personnages pour chaque chapitre. Les alias ont été résolus et regroupés, fournissant des informations détaillées sur les relations entre les personnages principaux.

Dans le graphe généré pour le chapitre 18 de **Les Cavernes d'Acier** 2, nous observons plusieurs connexions entre les personnages principaux. Certains personnages, tels qu'Enderby, Clousarr, et Baley, sont fortement liés, partageant des interactions significatives. Ces liens suggèrent des relations étroites ou des échanges d'informations entre ces personnages. En particulier, le personnage R. Sammy est connecté à la fois à Enderby et à Clousarr, indiquant des relations complexes dans ce contexte narratif.

Le personnage Dr Sarton semble jouer un rôle central, étant relié à plusieurs autres personnages, dont Enderby, Clousarr, Baley, et Lije. Cette centralité suggère que le personnage du Dr Sarton pourrait être une figure clé dans le développement de l'intrigue du chapitre 18.

En examinant les arêtes du graphe, nous pouvons également noter des connexions indirectes, comme le lien entre R. Sammy et Lije à travers Dr Sarton. Ces connexions indirectes ajoutent une complexité supplémentaire aux relations entre les personnages.

4. Suggestions d'Amélioration

Bien que notre approche ait réussi à extraire des réseaux de personnages, des améliorations potentielles incluent l'exploration de méthodes plus avancées de résolution d'alias et l'ajustement des critères de co-occurrence pour une détection plus précise des interactions.

5. Conclusions

L'exploration automatisée des réseaux de personnages dans la saga Fondation offre une perspective unique sur les dynamiques narratives. Notre approche TALN séquentielle constitue un point de départ, et les résultats obtenus peuvent servir de base pour des analyses plus approfondies. Cette étude ouvre la porte à de futures recherches explorant les relations complexes entre les personnages dans des œuvres littéraires emblématiques.

6. References

[1] French NLP Model in spaCy, *spaCy Models*, https://huggingface.co/spacy/fr$_{c}ore_{n}ews_{s}m$

[2] French NER in Flair, *Flair Models*, https://huggingface.co/flair/ner-french

[3] CamemBERT NER, *Hugging Face Model Hub*, https://huggingface.co/Jean-Baptiste/camembert-ner

[4] Named Entity Recognition, *Wikipedia*, https://fr.wikipedia.org/wiki/Reconnaissance$'_{d}entit$