Extraction automatique de réseaux de personnages par fusion de connaissances morphologiques

Résumé

Cette étude présente une méthodologie pour exploiter les technologies de traitement de langage naturel afin d'extraire automatiquement des réseaux de personnages de textes de la littérature de science-fiction. Notre étude se concentre sur l'identification des personnages, la résolution des problèmes d'alias et la détection de leurs interactions. Dans cette recherche, nous nous efforçons d'utiliser les technologies de traitement du langage naturel (NLP) en prenant comme jeux de données deux livres de la série de romans d'Isaac Asimov, "Fondation", avec pour principaux défis la reconnaissance d'entités nommées (NER). Nous avons mis en place un processus de traitement en plusieurs étapes, incluant le prétraitement du texte, la reconnaissance des entités, la résolution des alias et l'analyse de cooccurrence. Notre objectif est d'identifier et de classifier avec précision les noms de personnes et de lieux dans le texte, de regrouper les alias en une seule entité, et de détecter les interactions entre les personnages dans le texte. Les résultats de l'étude démontrent les performances de cette méthode pour cartographier les relations entre les personnages, ouvrant ainsi de nouvelles perspectives pour l'analyse littéraire.

Mots-clés : Traitement du langage naturel (TALN), Extraction automatique d'informations, Reconnaissance d'entités nommées (REN), Graphes de réseau de personnages

1. Introduction

La cartographie des relations entre personnages est particulièrement utile pour améliorer la compréhension en œuvres littéraires. Que ce soit dans la lecture de romans, de pièces de théâtre, de documents historiques ou d'autres textes, comprendre les relations entre les personnages peut aider les lecteurs à comprendre plus profondément le contenu et le contexte de l'histoire. En démêlant les relations entre les personnages, les lecteurs peuvent comprendre plus clairement le développement de l'histoire et les motivations des personnages. En particulier, face à des histoires complexes avec de nombreux personnages, la cartographie des relations entre les personnages peut aider les lecteurs à se souvenir de chaque personnage et de leurs caractéristiques. En analysant les relations entre les personnages, il est possible de révéler des significations plus profondes de l'histoire, telles que les thèmes, les conflits et les émotions. En particulier, lors de la lecture de textes riches en histoire, afin d'en comprendre les relations entre les personnages ainsi qu'appréhender une période où une société spécifique.

L'objectif de notre recherche est d'utiliser des systèmes de traitement du langage et des modèles de reconnaissance d'entités nommées pour construire automatiquement des réseaux de relations entre les personnages dans les romans de science-fiction, sans le support d'annotations humaines. Il est nécessaire d'identifier de façon exacte des entités nommées correspondant aux alias de personnages dans le texte, tout en évitant de détecter des entités non pertinentes pour notre tâche. La cartographie automatique des relations entre les personnages permettra de gagner beaucoup de temps et de réduire l'effort humain, en particulier lors du traitement de textes volumineux ou complexes.

Les applications sont nombreuses, ce type de système peuvent servir dans des environnements éducatifs, afin d'aider les élèves à mieux comprendre et analyser les œuvres littéraires. Pour les chercheurs en littérature, un tel outil peut servir d'outil d'assistance pour l'analyse d'œuvres littéraires complexes.

Dans notre étude, nous employons des techniques et outils spécifiques au traitement du langage naturel et à l'analyse de textes. Cela permet de faciliter la compréhension par le système du langage naturel et d'extraire des informations pertinentes et cruciales, telles que la catégorie grammaticale de chaque mot, les véritables entités nommées, et les alias des personnages. Par la suite, nous procédons à l'analyse de différents textes afin d'améliorer le traitement par le système pour des sujets spécifiques. Par exemple, notre corpus étant de la science-fiction, nous pouvons nous concentrer spécifiquement sur ce genre.

2. Méthode et Protocole Expérimental

2.1. Jeux de données

Nos jeux de données sont des fichiers textuels bruts et sans mise en forme résultant de l'extraction textuelle des documents PDF fournis. Ils contiennent des extraits d'une édition française de deux œuvres d'Isaac Asimov : les 18 chapitres du livre "Cavernes d'acier" ainsi que les 19 chapitres du livre "Prélude à Fondation.

2.1.1. Prétraitement des données

Tokenisation : Lors de la tokenisation, nous remplaçons d'abord toutes les ponctuations par des espaces, puis nous divisons le texte en fenêtre de 25 mots (ou tokens) chacun. Nous réalisons ce processus en créant des fenêtres segmentées qui se chevauchent. Chaque fenêtre contenant 25 mots, et chaque fenêtre se déplaçant d'un mot vers l'arrière par rapport à la fenêtre précédente. Ce processus se poursuit jusqu'à la fin du texte.

La lemmatisation et la racinisation : La lemmatisation est le processus de ramener différentes formes d'un mot à leur forme de base, tandis que la racinisation consiste à extraire la racine ou le radical de base d'un mot (ces formes n'ont pas nécessairement de signification). La lemmatisation et la racinisation sont toutes deux des méthodes importantes de normalisation des formes de mots, visant à réduire efficacement les différentes formes morphologiques d'un mot. Entre autres, l'outil Freeling est un outil qui peut fournir directement des résultats en appliquant une étape de lemmatisation. Cependant, nous avons choisi de ne pas procéder à une étape de lemmati-

sation et de racinisation dans notre projet, car pas adapter aux noms propres. Ce choix est dû au fait que, lors de nos essais, nous avons constaté que la lemmatisation et la racinisation pouvait à tort réduire certains alias de personnages à une forme de base correspondant à des verbes, entraînant ainsi la perte de nombreuses données importantes pouvant nous aider à résoudre notre tâche et induire en erreur notre algorithme de fusion.

Construction de la liste des mots vides : Après analyse des entités nommées PERSONNE et des étiquettes morphosyntaxiques (Part-Of-Speech - POS) des NP (nom propre) détectés par notre fusion de système 2.2.1, nous avons observé que des mots vides étaient détectés comme des personnes à cause de leurs casses présents en majuscule initiale. Nous avons donc essayé d'appliquer des listes de stopwords pour le français disponible dans les outils FreeLing et spaCy afin de retirer ses erreurs de détections comme personnage des mots tels que : "Space", "Merci", "Excusez-moi". Cependant, l'usage d'une liste de stopwords générique n'était pas adapté au domaine d'application et pouvait retirer des entités correctement détectées ainsi que de biaiser le système de détection des liens entre les personnages, à cause d'une distance plus faible entre les entités. Nous avions beaucoup plus d'entités PERSONNE dans une même fenêtre de 25 mots. Afin de résoudre ce problème, nous avons pris la décision de construire une liste de stopwords qui soit adapté à notre usage, sur la base des erreurs que nous observions suite à la fusion de nos systèmes. Après application de cette liste de stopwords plus adaptée, nous avons réussi à améliorer la métrique F1-mesure des nœuds et des arêtes représentant les liens, les personnages, passant de 0.36 à 0.58.

2.1.2. Statistiques

Afin d'améliorer le feedback manuel et l'analyse comparative des résultats, nous avons effectué quelques statistiques nécessaires sur diverses données :

	Les Cavernes d'acier	Prélude à Fondation
Chapitre		
Moy. # Mots	2957.17	4585.21
Moy. # Entité PERSONNE	62.4	82.9
Med. # Entité PERSONNE	62	87
Moy. % Entité PERSONNE	2.17	1.74
Moy. # Distance entités PERSONNE	43.47	52.49
Med. # Distance entités PERSONNE	33.22	44.42
Moy. % PROPN	4.83	4.58
Moy. % Uppercase	0.44	0.0
Moy. % Lowercase	72.33	67.79
Moy. % Capitalized	8.22	8.32
Moy. % Autre (incl. punctuation)	19.44	23.68
Moy. % Stopwords	62.44	60.26
Phrase		
Moy. # Mots	12.9	11.48
Paragraphe		
Moy. # Mots	31.39	27.95
Contexte (25 mots)		
Moy. # Entités PERSONNE	1.19	1.06
Med. # Entités PERSONNE	0.0	0.0

Table 1: *Tableau des statistiques des livres*

Les données du tableau ci-dessus sont calculées sur la base de la moyenne et de la médiane des données de chaque chapitre du livre correspondant.

Selon les données du tableau, nous pouvons voir que la proportion moyenne des entités nommées de type PERSONNE dans les deux livres est respectivement de 2,17 % et 1,74 %, soit environ la moitié de celle des noms propres (4,83% et 4,58%). La proportion des noms propres est, quant à elle, la moitié de celle des mots commençant par une majuscule, indiquant donc qu'il ne faut pas se fier uniquement à la casse pour identifier les personnages et leurs alias. De plus, le nombre moyen de mots par chapitre dans ces deux livres diffère de plus de 1500, ce qui rend donc le nombre d'entités nommées plus cohérent, car la proportion des entités nommées y est à peu près la même. Même dans le livre ayant moins de mots, la proportion d'entités nommées dépasse légèrement celle de l'autre livre ayant plus de mots.

Dans ces deux livres, nous avons également constaté que la différence entre la moyenne et la médiane des distances entre deux entités nommées semblait être proportionnel avec le nombre de mots en moyenne par chapitres, indiquant peut-être que la prélude laisse plus de temps à l'histoire pour se construire en se focalisant sur moins de personnages. Un autre point intéressant qui renforce cette hypothèse est que, dans nos fenêtres segmentées de 25 mots, nous avons trouvé que la moyenne des occurrences des entités nommées par fenêtre segmentée par chapitre était respectivement de 1,19 et 1,06 pour les différents livres, tandis que la moyenne des médianes était de 0 pour les deux.

Bien que les données sur la distance entre les entités nommées (43,4 et 52,4) puissent nous amener à envisager une taille de fenêtre supérieure à 25, nos tests sur différentes longueurs de fenêtres segmentées, sans modifier d'autres paramètres, ont révélé que les prédictions les plus proches de la réalité étaient obtenues avec des fenêtres de 25 et 30 mots comme nous pouvons l'observer dans le Tableau 2.

Fenêtre de taille N							
15	**20**	**25**	**30**	**40**	**50**	**60**	**73**
0.51	0.55	0.58	0.58	0.46	0.52	0.55	0.45

Table 2: *Performances en fonction de la taille de la fenêtre de contexte.*

2.1.2.1. Distribution des entités PER dans les livres

Les graphiques présentés ci-dessus montrent la répartition des entités nommées PER au travers des fenêtres de contexte de 25 mots pour chaque chapitre de l'ensemble des livres du corpus.

Dans les graphiques, l'intensité de la couleur est utilisée pour indiquer la fréquence d'apparition des entités nommées PER dans la fenêtre. Plus la couleur est foncée, plus le nombre d'entités nommées apparaissant dans cette fenêtre segmentée est élevé. Blanc représente ici 0 entités et Noir 7.

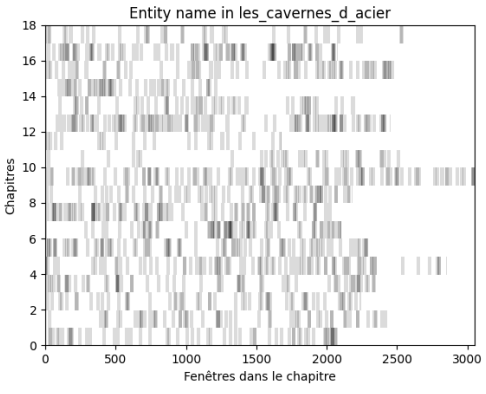

Figure 1: *Distribution des entités par chapitre pour le livre "Les Cavernes d'acier". Le blanc représente l'absence et le noir 7 entités dans la fenêtre.*

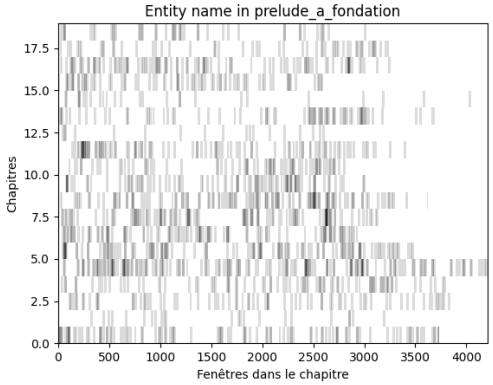

Figure 2: *Distribution des entités par chapitre pour le livre "Prélude à Fondation"*

D'après le graphique, il est apparent que dans 'Les Cavernes d'acier', la répartition des apparitions des personnages est assez uniforme, malgré l'existence de certains chapitres dans lesquels les personnages apparaissent rarement. De plus, il y a des chapitres où l'apparition des personnages ce fait principalement dans leur début. En revanche, dans 'Prélude à Fondation', les apparitions des personnages sont plus concentrées. Nous remarquons que les apparitions de personnages sont plus rares dans les premiers et derniers chapitres du livre, tandis qu'elles sont relativement plus fréquentes dans les chapitres centraux. Il est intéressant de noter que dans les chapitres les plus longs des deux livres, les apparitions des personnages sont également les plus nombreuses et la distribution est assez uniforme. Cette information est très importante, car ce type d'observation qui peuvent grandement aider à identifier les chapitres clés et à comprendre la trajectoire du développement de l'histoire, approfondissant ainsi notre compréhension et notre analyse du texte.

2.2. Identification des personnages

L'identification des personnages dans un texte est donc un défi majeur en traitement du langage naturel, nécessitant l'utilisation de modèles. Dans cette section, nous présentons plusieurs modèles, dont les jeux de données sont français, que nous avons évalués pour cette tâche.

Modèles	Type de modèles	Jeux de tests	Langue(s)
FreeLing [1]	HMM [2]	WOLF [3]	Multi-langues
NLTK [4]	HMM [2]	FTB [5]	Multi-langues
spaCy [6]	FastText [7]	WikiNER [8]	Multi-langues
Camembert [9]	BERT [10]	WikiNER [8]	Français

Table 3: *Liste des modèles évaluée.*

Nous avons donc évalué plusieurs modèles pour l'identification des entités nommées dans nos textes littéraires. NLTK, bien que polyvalent, avait des performances limitées pour le français. Camembert, bien qu'excellent pour le traitement du Français, il ne sait pas avérer aussi adaptés à notre tâche d'extraction de réseaux de personnages dû aux différents mots polysémiques qu'il contient. En revanche, spaCy et FreeLing ont montré des performances plus robustes au contexte très particulier que nous traitons, nous allons développer plus à ce propos dans la section suivante 2.2.1.

2.2.1. Fusion de spaCy et FreeLing

spaCy est souvent réputé pour sa précision dans l'identification des entités nommées, tandis que FreeLing peut offrir une couverture plus large en détectant des entités que spaCy pourrait manquer. En fusionnant les résultats des deux modèles, on cherche à tirer parti des forces individuelles de chaque modèle pour obtenir une détection plus complète et plus précise des personnages dans les textes analysés.

Le processus de fusion des résultats de spaCy et FreeLing découle de la volonté d'améliorer la précision et la couverture de l'identification des entités nommées. Dans notre démarche pour identifier les entités pertinentes dans le texte, nous avons pris la décision de créer un fichier qui correspond à l'intersection des étiquettes morphosyntaxiques (PER : personnes et MISC : étiquette indéterminée) détectées par spaCy avec les étiquettes morphosyntaxiques de type NP (Nom Propre) relevées par FreeLing.

La fusion des résultats de spaCy et FreeLing permet également de compenser les éventuelles lacunes ou erreurs spécifiques à un modèle, tout en augmentant le taux de précision et de rappel. Nous cherchons à couvrir à la fois les mentions explicites et les références indirectes aux personnages, pour obtenir une vue plus exhaustive du réseau de personnages dans le texte.

2.3. Détection des interactions entre personnages

La détection des interactions entre personnages repose sur une analyse des cooccurrences dans le texte. Cette phase se divise en deux parties :

2.3.1. Création de liste de personnages et gestion des alias

La création d'une liste de personnages dépends des entités trouvées issues de la fusion de spaCy et FreeLing 2.2.1. Les personnages sont comparés pour identifier leur correspondance. En effet, lorsqu'un personnage est présent dans la liste, on cherche si la chaîne de caractère représentant ce personnage est inclus dans un autre en utilisant une Regex. Si nous trouvons un lien, un alias est créé entre les deux personnages. Par exemple, pour un même chapitre, si nous avons dans la liste des personnages : ["Bentley, "Ben", "Ben Bailey", "Elijah Bailey"]. Nous associons leurs alias. Ainsi, nous obtenons la liste suivante : ["Ben Bailey": {"Bentley, "Ben"} et "Elijah Bailey": {}].

2.3.2. Cooccurrences entre les personnages

La dernière étape consiste à parcourir les tokens pour repérer la présence de noms de personnages à partir de la liste de personnages ou de leur alias. Après avoir identifié un personnage, la fonction explore les 25 tokens suivants pour détecter d'autres personnages. Si des noms de personnages sont identifiés dans cette fenêtre, la relation est considérée entre le premier personnage et les N personnages suivants dans la fenêtre. Nous incrémentons ensuite les arêtes dans le graphe.

2.4. Construction des graphes

Le processus de création des graphes débute par l'extraction des personnages à partir du texte 2.3. Une fois les entités extraites, la construction du graphe prend forme. Chaque personnage identifié constitue un nœud dans ce réseau.

Cette structure permet de capturer les relations entre les nœuds et de représenter l'intensité de ces relations via les

poids attribués aux arêtes. Elle offre ainsi une vue détaillée des connexions et des interactions entre les personnages du texte.

Chaque nœud, représentant un personnage, est défini par une balise node. Chaque balise node a un attribut id qui correspond à l'identifiant unique du personnage et les informations supplémentaires associées à ce nœud, comme les noms ou alias du personnage.

Les relations entre les personnages sont représentées par les balises edge (arête). Chaque edge a des attributs source et target, identifiant respectivement le nœud source et le nœud cible de l'arête. De plus, chaque edge les poids des arêtes, est accompagné par un nombre qui représente l'intensité de la relation entre les personnages.

3. Résultat

La soumission pour le leaderboard Kaggle exige un fichier CSV organisé : chaque ligne représente le graphe d'un chapitre du corpus pour l'évaluation. Les colonnes comportent un ID unique qui combine le code du livre et le numéro du chapitre. Ce format assure une évaluation uniforme des performances des modèles, permettant une identification précise des graphes pour l'extraction des personnages et de leurs alias.

Équipes	Scores
L'EKIP	0.70
Michel Marie Lamah	0.69
The explorer - Fusion (FreeLing + spaCy)	0.58
Squeezie le gotaga	0.58
ESSABRI ABDELHADI	0.45
Tiime	0.45
The explorer - NER spaCy	0.41
Maxime Renoux	0.36
The explorer - NER CamemBERT	0.28
none	0.25
The explorer - POS FreeLing	0.14
The explorer - POS NLTK	0.09

Table 4: *Classement des équipes en fonction des métriques F1*

En étant en troisième place ex æquo, notre équipe "The explorer" a obtenu un score de 0.58 avec le système de fusion entre le part-of-speech et la reconnaissance d'entités nommées. Ce score est calculé à partir de la F-mesure des nœuds et des arêtes pour chaque chapitre du corpus de test.

3.1. Métriques d'évaluation

Dans le contexte des personnages et de leurs alias par chapitre, la précision mesure la proportion de personnages correctement identifiés parmi ceux prédits par le modèle pour un chapitre donné. En utilisant la formule

$$Precision = \frac{TP}{TP + FP}$$

, où TP est le nombre de vrais positifs (personnages et leurs alias correctement prédits) et FP est le nombre de faux positifs (personnages et leurs alias prédits à tort), cette mesure évalue la justesse des prédictions de personnages et de leurs alias pour un chapitre.

Le rappel évalue la capacité du modèle à retrouver tous les personnages et leurs alias dans un chapitre donné. En utilisant la formule

$$Recall = \frac{TP}{TP + FN}$$

, où FN est le nombre de faux négatifs (personnages manqués), cette mesure évalue la capacité du modèle à identifier tous les personnages présents pour un chapitre.

3.1.1. F-mesure des nœuds

Les nœuds dans ce contexte représentent les personnages extrait, et la F-mesure évalue l'extraction de ces personnages dans chaque chapitre. En effet, la F-mesure combine la précision et le rappel en une seule mesure, fournissant une évaluation globale de la performance du modèle par chapitre. En utilisant la formule

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

, cette métrique synthétise l'équilibre entre la précision et le rappel. Elle est particulièrement utile pour évaluer la performance dans les scénarios auxquels on souhaite atteindre un compromis entre la précision et le rappel.

3.1.2. F-mesure des arêtes

Après avoir évalué la F-mesure des nœuds 3.1.1, nous évaluons la cohérence des arêtes en utilisant le couplage biparti maximal entre les nœuds des graphes prédits et de référence. Cette procédure est étendue pour établir une correspondance entre les ensembles d'arêtes prédites et de référence, incorporant la similarité des noms associés aux nœuds à l'aide de l'index de Jaccard.

La précision des arêtes mesure la proportion d'arêtes prédites qui correspondent correctement aux arêtes de référence, tandis que le rappel des arêtes évalue si les arêtes de référence sont correctement associées aux arêtes prédites.

Il est à noter que l'évaluation ne prend pas en compte l'intensité des liens entre les personnages, se concentrant plutôt sur leur existence ou non.

4. Conclusion

Notre approche pour extraire automatiquement des réseaux de personnages à partir de textes littéraires a reposé sur l'utilisation combinée de modèles tels que spaCy et FreeLing. La fusion des résultats de ces modèles a permis d'améliorer la précision et la couverture de notre extraction, conduisant à des réseaux de personnages plus complets.

Pour améliorer notre système, nous pourrions envisager l'annotation manuelle d'entités nommées et l'entraînement d'un modèle de NER sur ces données annotées. Cela renforcerait la capacité du modèle à reconnaître des personnages et à s'adapter à d'autres œuvres littéraires.

De plus, une exploration des méthodes génératives, telles que l'utilisation de ChatGPT pour la détection des personnages ou des liens, pourrait offrir une perspective intéressante. Nous pourrions aussi envisager l'usage de ChatGPT, pour aider les humains à l'annotation de grandes quantités de données, dans le but d'apprendre un modèle FastText ou BERT. La totalité de notre code est disponible librement sur GitHub [1].

[1] Double blind

5. References

[1] Xavier Carreras and Isaac Chao and Lluís Padró and Muntsa Padró. (2004) Freeling: An open-source suite of language analyzers proceedings of the 4th international conference on language resources and evaluation (lrec'04). [Online]. Available: https://nlp.lsi.upc.edu/freeling/node/1

[2] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[3] Sagot Benoît et Fišer Darja. (2008) Building a free french wordnet from multilingual resources. In Ontolex 2008, Marrakech, Maroc. [Online]. Available: https://pauillac.inria.fr/~sagot/index.html#wolf

[4] S. Bird and E. Klein, "Acl workshop on effective tools and methodologies for teaching nlp and cl," in *NLTK: The Natural Language Toolkit.*, 2002, pp. 62–69.

[5] A. Abeillé, L. Clément, and L. Liégeois, "Un corpus arboré pour le français : le French treebank [a parsed corpus for French: the French treebank]," *Traitement Automatique des Langues*, vol. 60, no. 2, pp. 19–43, 2019. [Online]. Available: https://aclanthology.org/2019.tal-2.2

[6] Matt Honnibal and Ines Montani and Van Landeghem Sofie and Boyd Adriane and Henning Peters. (2023) spacy: Industrial-strength natural language processing in python. [Online]. Available: https://github.com/explosion/spaCy

[7] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[8] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, "Learning multilingual named entity recognition from wikipedia," *Artificial Intelligence*, vol. 194, pp. 151–175, 2013, artificial Intelligence, Wikipedia and Semi-Structured Resources. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370212000276

[9] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423