Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Remplacement de mentions pour l'adaptation d'un corpus de reconnaissance d'entités nommées à un domaine cible

Arthur Amalvy[1] Vincent Labatut[1] Richard Dufour[2]
(1) Laboratoire Informatique d'Avignon (LIA), Avignon Université, France
(2) Laboratoire des Sciences du Numérique de Nantes (LS2N), Nantes Université, France
arthur.amalvy@univ-avignon.fr, vincent.labatut@univ-avignon.fr,
richard.dufour@univ-nantes.fr

RÉSUMÉ

La reconnaissance d'entités nommées est une tâche de traitement automatique du langage naturel bien étudiée et utile dans de nombreuses applications. Dernièrement, les modèles neuronaux permettent de la résoudre avec de très bonnes performances. Cependant, les jeux de données permettant l'entraînement et l'évaluation de ces modèles se concentrent sur un nombre restreint de domaines et types de documents (articles journalistiques, internet). Or, les performances d'un modèle entraîné sur un domaine ciblé sont en général moindres dans un autre : ceux moins couverts sont donc pénalisés. Pour tenter de remédier à ce problème, cet article propose d'utiliser une technique d'augmentation de données permettant d'adapter un corpus annoté en entités nommées d'un domaine source à un domaine cible où les types de noms rencontrés peuvent être différents. Nous l'appliquons dans le cadre de la littérature de fantasy, où nous montrons qu'elle peut apporter des gains de performance.

ABSTRACT

Mention replacement for adapting a named entity recognition dataset to a target domain

Named Entity Recognition is a well-studied natural language processing task, that is useful in a number of applications. Since recently, deep-learning models are able to solve this task with good performance. However, datasets used to train and evaluate those models cover a sparse number of domains (newswire, web). As performance for a model trained on a specific domain are generally lower on another one, this implies lower performance for less covered domains. In order to fix this issue, this article proposes to use a data augmentation technique that can be used to adapt a named entity recognition corpus from a source domain to a target domain where the encountered names can be different. We apply this technique to fantasy novels, and we show that it can yield performance gains in that context.

MOTS-CLÉS : apprentissage profond, augmentation de données, reconnaissance d'entités nommées.

KEYWORDS: deep learning, data augmentation, named entity recognition.

1 Introduction

La reconnaissance d'entités nommées (REN) a bénéficié récemment des avancées en apprentissage profond (Li *et al.*, 2020), et notamment de l'arrivée des modèles de langue neuronaux contextuels pré-entrainés, tels que BERT (Devlin *et al.*, 2019). Ceux-ci ont en effet permis d'améliorer les performances de nombreuses tâches de traitement automatique du langage naturel (TALN). Malgré

les performances élevées dont font preuve ces modèles pour la REN, celle-ci est encore loin d'être considérée comme résolue (Stanislawek *et al.*, 2019).

L'un des défauts de ces modèles provient de leurs données d'entraînement et d'évaluation. En effet, s'il existe de nombreux corpus annotés en entités nommées pour certains domaines (on peut citer le corpus journalistique CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), le corpus multi-domaines OntoNotes (Weischedel *et al.*, 2011) ou encore WikiGold (Balasuriya *et al.*, 2009), tiré de Wikipédia), d'autres sont plus délaissés et n'ont que peu ou pas de corpus dédiés : c'est le cas, par exemple, du domaine littéraire. Un modèle entraîné sur les données d'un domaine spécifique étant généralement moins performant lorsqu'il est appliqué sur celles d'un autre, ce délaissement se traduit par des performances moindres dans les domaines n'ayant pas de données annotées. Dans ce contexte, il est donc intéressant de disposer de techniques permettant d'adapter un corpus ou un modèle d'un domaine source, où l'on possède beaucoup de données, à un domaine cible plus pauvre en ressources.

Plusieurs possibilités existent pour réaliser cette adaptation. Gururangan *et al.* (2020) montrent par exemple que continuer de pré-entraîner un modèle neuronal contextuel sur le domaine cible permet d'augmenter les performances dans celui-ci. Malheureusement, ce pré-entraînement est plus coûteux que l'affinage (*fine-tuning*) d'un modèle pré-entraîné. Une autre voie possible est l'adaptation du corpus original par augmentation. L'augmentation de données est une technique devenue standard en apprentissage profond pour le traitement de l'image (Shorten & Khoshgoftaar, 2019). Des opérations comme le recadrement (*cropping*), le retournement (*flipping*) ou l'ajout de bruit permettent de créer des exemples synthétiques supplémentaires, ce qui accroît le nombre de données d'entraînement. Comparativement, elle est moins explorée en TALN, bien que des recherches récentes commencent à s'y intéresser (Feng *et al.*, 2021). En effet, il ne paraît pas aussi facile de générer de nouveaux exemples à partir du texte, car cela requiert de garder la cohérence de la phrase tout en la modifiant. L'augmentation de données est encore moins explorée dans le cadre de la REN, car celle-ci se heurte à des problèmes spécifiques : certaines techniques classiques d'augmentation (comme le remplacement par un synonyme, l'échange de mots ou encore l'insertion aléatoire (Wei & Zou, 2019)) peuvent notamment engendrer des exemples synthétiques avec des labels incorrects (Dai & Adel, 2020).

Il existe tout de même quelques travaux utilisant cette technique pour adapter des corpus à un autre domaine. Ding *et al.* (2020) proposent d'entraîner un modèle de langage pour générer des exemples synthétiques dans le cas où peu d'exemples sont disponibles, mais supposent cependant l'existence d'un minimum d'exemples déjà annotés. Chen *et al.* (2021) décrivent une architecture neuronale capable d'apprendre à transformer une phrase d'un domaine source vers un domaine cible, et s'en servent pour générer des exemples synthétiques proches du domaine cible. Néanmoins, leur méthode nécessite d'entraîner cette architecture, ce qui demande aussi de disposer de quelques exemples annotés dans le domaine cible.

Nous présentons dans cet article une méthode d'augmentation des données, le remplacement de mentions (Dai & Adel, 2020), pour permettre d'adapter un corpus annoté en entités nommées d'un domaine à un autre. Cette méthode a l'avantage de la simplicité, et de ne pas demander l'annotation de données du domaine cible. Elle consiste à générer de nouveaux exemples en remplaçant, dans une phrase du corpus original, une entité nommée par une autre du même type. Nous décrivons notre méthode dans la Section 2. Pour montrer l'intérêt de cette technique, nous expérimentons en adaptant le corpus CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), qui couvre originellement le domaine journalistique, à celui de la littérature de fantasy. Nous discutons les résultats obtenus en Section 3.

2 Protocole Expérimental

2.1 Jeux de données

Nous utilisons deux jeux de données différents, provenant de deux sources distinctes, tous deux composés de textes en anglais. Nous nous concentrons sur les entités nommées désignant des personnes.

Jeu d'évaluation Dekker *et al.* (2019) ont procédé à une évaluation de plusieurs modèles de REN, en annotant un corpus littéraire pour les entités de type `PERSONNE`. Il se compose du premier chapitre de 40 romans. Cependant, d'après nos observations, ce jeu de données souffre de problèmes d'annotations (erreurs et incohérences), d'encodage et de tokenisation, que nous avons donc cherché à corriger. Nous avons rectifié les problèmes d'encodage et de tokenisation manuellement. Afin de pallier les erreurs et les incohérences, nous avons défini un ensemble de règles d'annotation que nous avons appliquées à tout le corpus grâce à un processus semi-automatique. Celui-ci s'articule en 3 étapes : 1) l'application d'heuristiques de correction supervisées par un humain, puis 2) une correction assistée à l'aide d'un modèle BERT (Devlin *et al.*, 2019), et finalement 3) une correction manuelle en dernier ressort. Nos résultats sont évalués sur le sous-ensemble de ce corpus formé par les 17 romans de fantasy qu'il contient.

Jeu d'entraînement Nous entraînons notre modèle avec une version modifiée du CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), qui est un jeu de données populaire composé d'articles journalistiques. Nos modifications consistent en l'inclusion systématique des titres de civilité (`Lord`, `Mr.`, `Lady`...) comme faisant partie des entités de type `PERSONNE`, pour rester cohérent avec les règles d'annotations établies lors de la correction du corpus de Dekker *et al.* (2019).

2.2 Modèle et Entraînement

Nous utilisons un modèle BERT-base (Devlin *et al.*, 2019) pré-entraîné, obtenu via la librairie `huggingface` (Wolf *et al.*, 2020), auquel nous apposons une couche neuronale de classification. Ce modèle est affiné (*fine-tuning*) sur notre corpus d'entraînement pendant 2 cycles d'apprentissage avec un taux d'apprentissage de $2 \cdot 10^{-5}$ (Devlin *et al.*, 2019).

2.3 Augmentation des données

Nous proposons l'utilisation d'une technique simple, le remplacement de mentions (Dai & Adel, 2020), pour adapter un corpus annoté en entités nommées d'un domaine source vers un domaine cible. Afin de générer un nouvel exemple, nous partons d'un exemple de départ et remplaçons une mention (un passage continu du texte annoté comme une entité) et ses mentions équivalentes (les autres passages identiques de l'exemple) par une autre entité de même type récupérée dans une liste prédéfinie. La longueur des passages remplacés pouvant être modifiée, les labels sont ajustés en fonction. La Figure 1 donne un cas de génération d'un exemple synthétique. Nous nous concentrons uniquement sur les entités de type `PERSONNE`, puisque notre corpus d'évaluation ne contient que

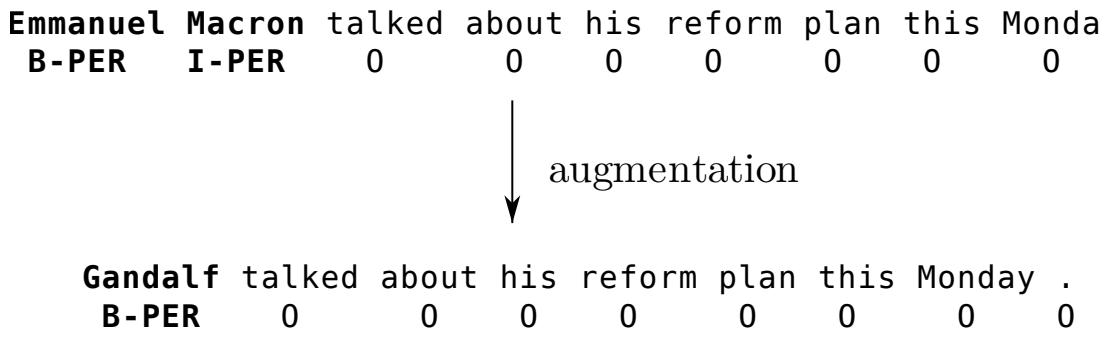

FIGURE 1 – Exemple de génération d'un nouvel exemple synthétique par remplacement de mention.

ces annotations. Nous espérons que cette augmentation permettra au modèle de reconnaître plus facilement les noms typiques du domaine cible.

Afin de mettre en évidence l'utilité de l'ajout de noms proches du domaine cible dans le corpus, nous comparons cinq configurations, détaillées ci-après :

Entraînement sans augmentation (`none`) Le modèle est entraîné sur notre version modifiée du CoNLL-2003 sans ajout de données synthétiques.

Entraînement avec augmentation interne (`conll`) Le modèle est entraîné sur notre version du CoNLL-2003 modifiée par l'application de notre technique d'augmentation. La liste d'augmentation est composée de toutes les entités de type PERSONNE du CoNLL-2003. Cette configuration nous permet de vérifier l'impact de notre technique d'augmentation lorsque aucune donnée d'un autre domaine n'est ajoutée au jeu de données initial.

Entraînement avec augmentation externe hors-domaine (`wgold`) Même configuration que la précédente, mais la liste d'augmentation est composée de toutes les entités de type PERSONNE du corpus annoté en entités nommées WikiGold (Balasuriya *et al.*, 2009), issu de Wikipédia. Cette configuration nous permet de contrôler l'apport de notre technique lorsque de nouvelles données sont ajoutées au corpus de départ, mais que celles-ci ne sont pas proches du domaine cible.

Entraînement avec augmentation externe intra-domaine (`fantasy`) Même configuration que la précédente, mais la liste d'augmentation est composée de noms de fantasy : il s'agit donc de noms proches du domaine cible. Ces noms ont été extraits à partir de la liste des personnages du jeu vidéo *The Elder Scrolls III : Morrowind*. Cette liste a été choisie car elle contient une bonne variété de noms typiques de la fantasy, sans que ceux-ci proviennent directement de notre corpus cible.

Entraînement avec augmentation parfaite (`dekker`) Même configuration que la précédente, mais la liste d'augmentation contient tous les noms de personnages de notre jeu d'évaluation. Cette configuration nous permet de tester la limite du potentiel de notre méthode d'augmentation.

Pour comprendre plus en détails l'influence de l'augmentation de données, nous expérimentons nos configurations en faisant varier la quantité d'exemples synthétiques générés. Non contrôlons cette quantité grâce à un *taux d'augmentation*, exprimé proportionnellement à la taille du corpus de départ. Nous considérons des augmentations de 5%, 10%, 50% et 100% de cette taille.

Le processus de génération des exemples est aléatoire : lors de nos expériences, la liste des exemples synthétiques est donc nouvellement générée à chacun des entraînements.

3 Résultats

Augmentation	Taux d'augmentation	Précision	Rappel	F1
`none`	0%	**93.11 (1.38)**	87.03 (4.20)	89.89 (2.11)
`conll`	5%	93.07 (1.66)	85.91 (4.21)	89.28 (2.22)
`conll`	10%	92.99 (1.33)	86.90 (4.31)	89.76 (2.10)
`conll`	50%	91.65 (2.17)**	89.50 (3.13)**	90.50 (1.05)*
`conll`	100%	90.82 (2.43)**	89.19 (3.70)**	89.91 (1.33)
`wgold`	5%	92.97 (1.75)	86.17 (7.19)	89.23 (3.89)
`wgold`	10%	92.69 (1.98)	88.61 (3.61)*	90.53 (1.29)
`wgold`	50%	91.82 (2.11)**	89.59 (3.02)**	90.63 (0.95)*
`wgold`	100%	90.98 (1.89)**	89.87 (3.19)**	90.37 (1.31)
`fantasy`	5%	91.26 (1.68)**	90.58 (3.05)**	**90.86 (1.13)****
`fantasy`	10%	89.80 (2.58)**	91.36 (2.63)**	90.51 (1.09)
`fantasy`	50%	88.42 (2.20)**	**92.35 (2.20)****	90.30 (0.83)
`fantasy`	100%	88.88 (2.93)**	91.57 (2.67)**	90.13 (0.95)
`dekker`	5%	92.28 (1.92)*	94.22 (2.09)**	93.21 (0.77)**
`dekker`	10%	91.17 (2.06)**	96.12 (1.15)**	93.56 (0.78)**
`dekker`	50%	89.34 (2.65)**	98.27 (0.51)**	93.56 (1.36)**
`dekker`	100%	89.38 (2.67)**	98.42 (0.52)**	93.65 (1.38)**

TABLE 1 – Moyenne (et écart type) sur 25 entraînements de la précision, du rappel et de la F-mesure des différentes configurations d'augmentation sur le sous-ensemble des 17 romans de fantasy de notre version corrigée du corpus de Dekker *et al.* (2019). Les étoiles dénotent les performances significativement différentes de la configuration `none` (test de Student apparié, * : $p < 0.1$, ** : $p < 0.05$).

La Table 1 indique les résultats de nos expérimentations. Nous calculons la précision, le rappel et la F-mesure selon la méthode du CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003) en utilisant la librairie python `seqeval` (Nakayama, 2018). L'augmentation externe intra-domaine (`fantasy`) avec un taux d'augmentation de 5 % donne une amélioration significative de la F-mesure, ce qui met en valeur l'utilité d'injecter des noms proches du corpus cible lors de l'entraînement. De manière générale, on observe une augmentation du rappel et une baisse de la précision pour toutes les configurations, effet qui semble augmenter avec l'accentuation du taux d'augmentation, jusqu'à un certain point. L'augmentation parfaite (`dekker`) démontre bien cet effet, avec un rappel très élevé mais une précision tout de même amoindrie. Elle démontre aussi une marge de progression de la performance importante, avec une augmentation du rappel de plus de 11 points par rapport à la configuration sans augmentation (`none`). Hors augmentation parfaite, l'augmentation externe

intra-domaine semble avoir le plus d'effets, avec l'accroissement du rappel le plus important mais également la baisse de précision la plus importante.

L'augmentation du rappel nous permet de remarquer que certaines classes de noms sont mieux reconnues par le modèle dans la configuration `fantasy` :

— Les noms qui sont aussi des noms communs (`"Bug", "Silent", "Weasel"`).
— Les noms de fantasy à apostrophes, tels que `"Rand al'Thor"` ou `"Bran al'Vere"`.

Ce constat parait indiquer que les gains de performances observés proviennent des noms spécifiques au domaine cible.

L'augmentation avec des données internes (`conll`) et l'augmentation hors-domaine (`wgold`) apparaissent avoir moins d'effet que l'augmentation externe intra-domaine (`fantasy`). Il est possible que, dans ces cas, l'information apportée par les exemples synthétiques ne soit pas pertinente par rapport au domaine cible, et pousse le modèle vers le sur-apprentissage. Pour appuyer cette hypothèse, la Table 2 présente les résultats de nos différentes configurations d'augmentation sur l'ensemble de test du CoNLL-2003. On observe que les résultats de nos stratégies sont proches du résultat original ou moins bons, ce qui est cohérent avec Dai & Adel (2020). Il semble donc que la technique d'augmentation testée n'apporte pas de bénéfice lorsque les informations ajoutées par les noms injectés sont redondantes (`conll`) ou non-pertinentes (`wgold`, `fantasy`, `dekker`).

Augmentation	Taux d'augmentation	Précision	Rappel	F1
`none`	0%	93.92 (1.03)	93.98 (1.18)	93.94 (0.48)
`conll`	5%	**94.13 (0.99)**	93.94 (0.93)	**94.02 (0.43)**
`conll`	10%	93.53 (1.64)	94.07 (1.22)	93.78 (0.50)
`conll`	50%	93.24 (1.65)*	**94.77 (0.63)****	93.99 (0.67)
`conll`	100%	92.72 (1.73)**	94.61 (0.69)**	93.64 (0.69)**
`wgold`	5%	93.89 (1.42)	93.87 (1.59)	93.86 (0.58)
`wgold`	10%	93.42 (1.93)	94.21 (1.45)	93.79 (0.88)
`wgold`	50%	93.14 (2.10)*	94.14 (1.13)	93.62 (0.90)*
`wgold`	100%	92.82 (1.78)**	94.70 (0.83)**	93.73 (0.67)*
`fantasy`	5%	93.53 (1.56)	94.37 (0.88)*	93.94 (0.55)
`fantasy`	10%	93.16 (1.67)**	94.26 (1.11)	93.69 (0.67)**
`fantasy`	50%	92.61 (1.64)**	94.66 (0.97)**	93.61 (0.64)**
`fantasy`	100%	92.63 (1.76)**	94.50 (0.92)**	93.54 (0.71)**
`dekker`	5%	93.14 (1.37)**	94.34 (0.82)	93.73 (0.57)
`dekker`	10%	93.16 (1.40)**	94.28 (1.08)	93.71 (0.51)**
`dekker`	50%	92.59 (1.40)**	94.50 (0.93)*	93.52 (0.65)**
`dekker`	100%	92.86 (1.29)**	94.42 (0.81)*	93.62 (0.45)**

TABLE 2 – Moyenne (et écart type) sur 25 entraînements de la précision, du rappel et de la F-mesure des différentes configurations d'augmentation sur le jeu de test du CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003). Seules les entités de type `PERSONNE` sont prises en compte. Les étoiles dénotent les performances significativement différentes de la configuration `none` (test de Student apparié, * : $p < 0.1$, ** : $p < 0.05$).

4 Conclusion et Travaux Futurs

Nous avons proposé, dans cet article, l'utilisation d'une méthode simple d'augmentation des données permettant d'adapter un corpus de reconnaissance d'entités nommées d'un domaine source à un domaine cible. Les résultats de nos expériences montrent que cette méthode permet des gains de performance lors de l'adaptation du corpus journalistique du CoNLL-2003 au domaine de la littérature de fantasy pour la détection de noms de personnes. Sa simplicité a l'avantage de la rendre facilement adaptable à d'autres domaines, et combinable avec d'autres approches.

Ce travail reste néanmoins limité aux noms de personnes, et à un domaine spécifique. Les travaux futurs pourraient notamment s'intéresser à d'autres types d'entités (lieux, organisations...) et dans des contextes différents. La combinaison avec d'autres types d'augmentation (remplacement par un synonyme, échange de mots...) mérite probablement d'être explorée, car elle permettrait d'augmenter la variété des exemples générés, et donc potentiellement les performances.

Références

BALASURIYA D., RINGLAND N., NOTHMAN J., MURPHY T. & CURRAN J. R. (2009). Named entity recognition in Wikipedia. In *Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, p. 10–18.

CHEN S., AGUILAR G., NEVES L. & SOLORIO T. (2021). Data augmentation for cross-domain named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*, p. 5346–5356. DOI : 10.18653/v1/2021.emnlp-main.434.

DAI X. & ADEL H. (2020). An analysis of simple data augmentation for named entity recognition. In *International Conference on Computational Linguistics*, p. 3861–3867. DOI : 10.18653/v1/2020.coling-main.343.

DEKKER N., KUHN T. & VAN ERP M. (2019). Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, **5**, e189. DOI : 10.7717/peerj-cs.189.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, volume 1, p. 4171–4186. DOI : 10.18653/v1/N19-1423.

DING B., LIU L., BING L., KRUENGKRAI C., NGUYEN T. H., JOTY S., SI L. & MIAO C. (2020). DAGA : Data augmentation with a generation approach for low-resource tagging tasks. In *Conference on Empirical Methods in Natural Language Processing*, p. 6045–6057. DOI : 10.18653/v1/2020.emnlp-main.488.

FENG S. Y., GANGAL V., WEI J., CHANDAR S., VOSOUGHI S., MITAMURA T. & HOVY E. (2021). A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 968–988. DOI : 10.18653/v1/2021.findings-acl.84.

GURURANGAN S., MARASOVIĆ A., SWAYAMDIPTA S., LO K., BELTAGY I., DOWNEY D. & SMITH N. A. (2020). Don't stop pretraining : Adapt language models to domains and tasks. In *58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

LI J., SUN A., HAN J. & LI C. (2020). A survey on deep learning for named entity recognition. *arXiv*, **cs.CL**, 1812.09449. DOI : 10.1109/TKDE.2020.2981314.

NAKAYAMA H. (2018). seqeval : A python framework for sequence labeling evaluation.

SHORTEN C. & KHOSHGOFTAAR T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, **6**(1), 60. DOI : 10.1186/s40537-019-0197-0.

STANISLAWEK T., WRÓBLEWSKA A., WÓJCICKA A., ZIEMBICKI D. & BIECEK P. (2019). Named entity recognition - is there a glass ceiling ? In *23rd Conference on Computational Natural Language Learning*, p. 624–633. DOI : 10.18653/v1/K19-1058.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *7th Conference on Natural Language Learning*, p. 142–147.

WEI J. & ZOU K. (2019). EDA : Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, p. 6382–6388. DOI : 10.18653/v1/D19-1670.

WEISCHEDEL R., HOVY E., MARCUS M., PALMER M., BELVIN R., PRADHAN S., RAMSHAW L. & XUE N. (2011). OntoNotes : A large training corpus for enhanced processing.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., LE SCAO T., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45.