Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français

Maroua Boudabous, Anna Pappa

▶ **To cite this version:**

Maroua Boudabous, Anna Pappa. Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français. Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL), Nov 2022, Marseille, France. pp.20-28. hal-03846827

HAL Id: hal-03846827
https://hal.science/hal-03846827v1

Submitted on 14 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Apprentissage actif pour l'extraction des aspects explicites : application à des avis non annotés en français

Maroua Bouadabous[1] Anna Pappa[1]
(1) LIASD, Université Paris 8, 2 rue de la liberté, 93526 Saint Denis, France
m.boudabous@ai.univ-paris8.fr, ap@up8.edu

RÉSUMÉ

Dans cet article, nous présentons un processus de bout en bout qui utilise l'apprentissage actif pour améliorer l'extraction des aspects explicites pour des langues à faibles ressources dans le cadre d'analyse des opinions et des sentiments. L'extraction des aspects explicites ou implicites, reste difficile en raison de la rareté des données labellisées et de la complexité du processus d'annotation manuelle. Nous avons conçu un processus en deux étapes : nous commençons avec une pré-labellisation via un CRF en utilisant l'apprentissage par transfert de connaissances. Ensuite, nous utilisons l'apprentissage actif pour améliorer la performance de labellisation, gérer les labels manquants et réduire les efforts d'annotation humaine. Nous avons utilisé deux corpus, composés des avis utilisateurs en langue française, sur des produits de beauté et des appareils électroniques pour l'évaluation des processus. Les résultats montrent que l'étape d'apprentissage actif améliore les performances de labellisation lorsque 30% des labels initiaux sont corrigés.

ABSTRACT

Active Learning for Explicit Aspect Term Extraction : a use case for unlabeled French reviews

In this paper, we present an end-to-end process using active learning to improve explicit aspect extraction for low-resource language in opinion and sentiment analysis. Handling Aspect Term Extraction (ATE) as a supervised sequence labeling task remains challenging due to labeled data scarcity and the complexity of the manual annotation process. We define a two-step process starting from defining a pseudo-labeler CRF using cross-domain learning. Then, we use active learning to deal with label uncertainty resulting from the first step and to reduce human annotation efforts. Two web-scraped datasets of French reviews on beauty products and electronic devices were used for process evaluation. Results show that the active learning step improves the learning model's performance when 30% of initial labels are corrected.

MOTS-CLÉS : Apprentissage actif, Extraction d'aspect explicite, Apprentissage inter domaines.

KEYWORDS: Active learning, Explicit Aspect Term Extraction, Cross-domain learning.

1 Introduction

Le World Wide Web est un gisement de données numériques couvrant une multitude de domaines d'intérêt (économie, politique, etc.) d'une richesse et d'une facilité d'accès sans précédent. En effet, l'explosion de l'usage d'internet engendre des volumes croissants de données langagières et multilingue disponibles en ligne. Ces données proviennent des sources différentes comme les réseaux sociaux, les plateformes de services ou les sites de commerces en ligne. Durant les dernières années,

les entreprises se sont intéressées à les exploiter afin d'y extraire des connaissances pertinentes pour l'amélioration de leurs produits et/ou services. Cet intérêt s'est également manifesté dans la communauté scientifique du traitement automatique des langues (TAL). L'apprentissage automatique et statistique est un des outils parmi les plus utilisés, pour analyser et traiter les données textuelles préalablement annotées. Cependant, les données étiquetées (ou labellisées) sont inégalement disponibles, en fonction de la langue. Nous désignons par langue à faible ressources, les langues ayant peu ou pas de données labellisées disponibles.

Dans cet article, nous nous intéressons à l'identification des aspects permettant la reconnaissance des termes utilisés pour désigner les attributs et caractéristiques des produits ou des services à partir des avis des utilisateurs. Cette tâche n'est pas triviale, elle est coûteuse, chronophage et suscite une attention considérable pour la création des modèles performants sur des données non labellisées. Nous distinguons deux types d'aspects : explicites et implicites. Dans ce qui suit, nous considérons l'identification des aspects explicites à partir de données textuelles en Français comme exemple de langue à faibles ressources. Ce contexte nous a motivé à proposer un processus d'identification des aspects adapté aux langues à faibles ressources. Notre processus de traitement se définit en deux étapes d'apprentissage : la première consiste à la pré-labellisation des données par transfert de connaissances et la deuxième à l'amélioration de cette labellisation par apprentissage actif.

Nous apportons des modifications au niveau des deux étapes du processus pour s'adapter à la rareté des données labellisées. En effet, nous facilitons le transfert de connaissances, durant la phase de pré-labellisation, par la suppression des descripteurs dépendants du domaine d'application notamment les noms (e.g "restaurant" , "cuisine", "nourriture", "plats") et les verbes (e.g "déguster","savourer"). Nous agissons également sur la stratégie de sélection pour l'apprentissage actif, par la définition de l'indice de confiance. Ceci est la moyenne pondérée des probabilités de l'exactitude des termes identifiés comme "aspect" qui composent un avis utilisateur. Cette méthode de calcul permet de gérer le déséquilibre de fréquence entre termes "aspects " et "non-aspects".

La description de la méthode est organisée comme suit. La section 2 introduit la tâche d'identification des aspects dans la littérature. La section 3 décrit les différentes étapes du processus proposé pour la reconnaissance des aspects à partir de corpus non labellisés. Dans la section 4, nous présentons les données utilisées pour l'évaluation du processus , les expérimentations et les résultats. La section 5 conclut l'article et propose quelques perspectives.

2 État de l'art

L'identification des aspects présente une étape clé dans l'analyse des sentiments et des tendances. Or il s'agit d'une tâche complexe notamment pour les langues qui possèdent peu ou pas de données labellisées. L'utilisation d'apprentissage statistique supervisé a suscité de l'intérêt pour l'identification des aspects en définissant cette tâche sous forme d'un problème de labellisation séquentielle. Différentes architectures de réseaux profonds ont été utilisées pour résoudre le problème d'identification des aspects à savoir les réseaux récurrents de type LSTM (Liu *et al.*, 2015), le mécanisme d'attention (Li *et al.*, 2018), les réseaux de convolution CNN (Xu *et al.*, 2018) ainsi que les transformateurs (Santos *et al.*, 2021). Les performances de toutes ces architectures dépendent de la disponibilité des données d'apprentissage labellisées, ce qui rend leur utilisation problématique pour les langues à faibles ressources.

La rareté des données labellisées pour entraîner les modèles d'apprentissage supervisés a été considérée par (Li *et al.*, 2020) qui ont eu recours à l'augmentation des données. Ils génèrent des données supplémentaires à travers un modèle de génération masqué conditionnel appliqué sur les données d'apprentissage. Cette méthode est assez incontrôlable car elle risque de changer la sémantique des opinions générés. (Chen & Qian, 2020) ont utilisé des prototypes logiciels appris sur des données externes générées par des modèles de langage pré-entraînés pour pallier ce problème. Cette approche nécessite cependant des jeux de données volumineux pour entraîner les modèles de langages. Les auteurs soulignent que plus le volume des données est important mieux est la performance de leur modèle. Dans cet article, nous considérons la situation extrême où aucune donnée labellisée dans le domaine d'application n'est disponible. Le processus que nous proposons repose sur une étape de pré-labellisation qui permettra ensuite d'entraîner le modèle d'apprentissage via l'apprentissage actif pour améliorer l'identification des aspects par le biais de plusieurs itérations de correction des labels et d'enrichissement des données.

L'apprentissage actif (AA) représente une alternative qui permet de réduire l'effort de labellisation manuelle au profit de l'amélioration des performances.

Dans le domaine de traitement automatique des langages, AA a été considéré pour différentes tâches de labellisation séquentielle à savoir la reconnaissance d'entité nommée (NER) (Shelmanov *et al.*, 2019) et l'étiquetage morpho-syntaxique (PoS) (Ringger *et al.*, 2007). Dans ce contexte, l'AA a été dans un premier temps utilisé avec des modèles d'apprentissage automatique standards (Settles & Craven, 2008) (Settles, 2009). Ensuite, l'AA a été associé à des modèles d'apprentissage profonds (Schröder & Niekler, 2020) ; (Ren *et al.*, 2021), (Shen *et al.*, 2018). Cette association a révélé des résultats intéressants en termes de coût de calcul et de labellisation.

L'apprentissage actif est un processus itératif qui sélectionne un échantillon des données non labellisées selon une stratégie de sélection prédéfinie et invite un expert à les labelliser. Dans cet article, nous adaptons ce processus pour permettre à la fois d'enrichir des données et de corriger les pré-labels. En effet, l'expert est invité à vérifier l'exactitude de la labellisation des données sélectionnées à partir de l'ensemble de données pré-labellisées initial.

Définir la stratégie de sélection est incontournable pour tout processus d'apprentissage actif. La stratégie de sélection par incertitude est l'approche la plus répandue dans le cadre de l'apprentissage actif (Scheffer *et al.*, 2001) ; (Culotta & McCallum, 2005). Elle consiste à sélectionner les instances pour lesquelles le modèle d'apprentissage est le moins confiant. Notre approche redéfinit l'indice de confiance pour la labellisation séquentielle. Pour chaque instance, nous utilisons la moyenne pondérée des indices de confiance de chacun de ses termes "aspects" et "non aspects". La pondération est adoptée pour accorder plus d'importance aux aspects qui sont moins fréquents dans les avis des utilisateurs.

3 Description du processus d'identification des aspects

En effet, nous abordons le problème de la rareté des données labellisées pour l'identification des aspects explicites à partir des avis utilisateurs. Nous considérons le français comme un exemple de langues à faibles ressources. Tout d'abord, les données sont pré-labellisées via le transfert de connaissances par adaptation de domaine à l'aide d'un modèle de champs aléatoires conditionnels CRF. Ensuite, nous entraînons un modèle d'apprentissage profond de type BiLSTM-CNN-CRF pour

la reconnaissance des aspects sur les données précédemment pré-labellisées dans un cadre supervisé. Enfin nous appliquons l'apprentissage actif pour gérer les aspects manquants et incorrectement identifiés et améliorer les performances de pré-labellisation. La figure 1 illustre les différentes étapes du processus proposé.

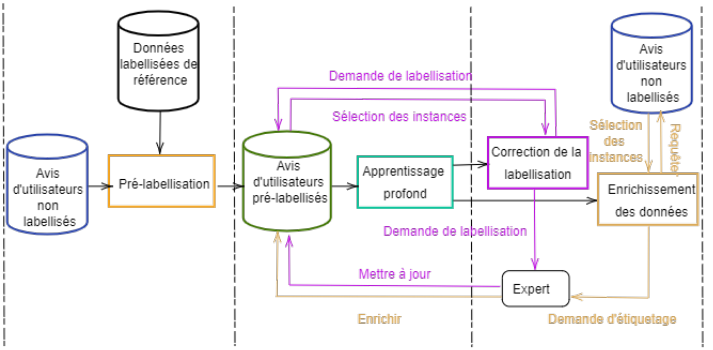

FIGURE 1 – Schéma global du processus proposé pour l'identification des aspects explicites .

Nous proposons la phase de pré-labellisation en tant que étape de pré-traitement pour l'annotation des données non labellisées en utilisant les champs aléatoires conditionnels (CRF). CRF est un modèle de prédiction probabiliste utilisé pour segmenter et labelliser les données séquentielles notamment les textes. Il prédit le label de chaque terme en considérant son voisinage.

Concrètement, les données en entrée pour le CRF sont des vecteurs de caractéristiques qui représentent différentes informations à propos du terme à savoir sa valeur littérale, le nombre de lettres, s'il s'agit d'un terme en majuscule ou minuscule, s'il s'agit d'un signe de ponctuation ou d'un chiffre. L'étiquette morpho-syntaxique (PoS) est également incluse. Nous considérons également les mêmes informations pour les K-mots voisins.

Pour renforcer l'indépendance de notre modèle CRF par rapport au domaine d'application, nous avons omis la caractéristique "valeur littérale" des mots dont l'étiquette morpho-syntaxique correspond aux catégories noms ("Noun") et verbes ("Verb") puisqu'il s'agit de catégories les plus dépendantes du domaine.

L'architecture BiLSTM-CNN-CRF se compose de deux couches de plongements de mots : la première couche utilise l'algorithme Glove, pré-entraîné sur le corpus wikipedia en Français, pour fournir les plongements des mots composant la séquence. La deuxième couche utilise un réseau de neurones convolutifs (CNN) pour extraire les informations morphologiques au niveau du caractère et les encoder dans des représentations neuronales. La sortie de ces couches passe par la suite dans une couche LSTM bidirectionnelle pour capturer des informations sur les mots passés et futurs de la séquence. Enfin, une couche CRF est ajoutée pour produire la séquence la plus probable des labels prédits (Voir la figure 2).

Nous adaptons le processus de l'apprentissage actif, pour permettre à la fois d'enrichir des données et de corriger les pré-labels. En effet, l'expert est invité à vérifier l'exactitude de la labellisation des données sélectionnées à partir de l'ensemble initial de données pré-labellisées.

Nous adaptons la stratégie de sélection pour permettre la classification multi-label et pallier au

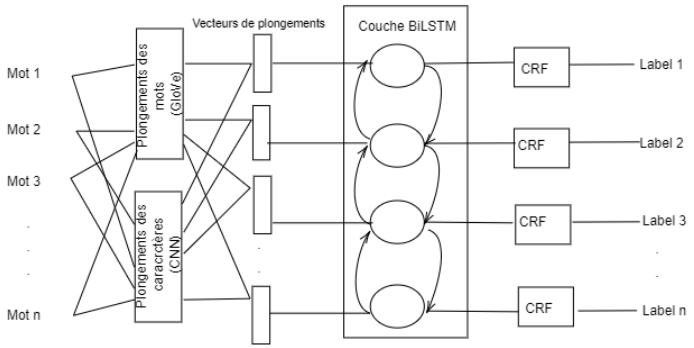

FIGURE 2 – Architecture du modèle d'apprentissage BiLSTM-CNN-CRF

déséquilibre de représentativité des classes «aspect» et «non-aspect». En effet, nous sélectionnons des instances dont la prédiction est la moins confiante. Pour s'adapter au contexte de la labellisation séquentielle, nous associons à chaque instance la moyenne des scores d'incertitude des termes qui la composent. Par ailleurs, les termes qui ne représentent pas des aspects sont supprimés du calcul. Ceci permet d'accélérer la convergence de l'apprentissage actif.

	Paramètre	Valeur
Modèle CRF	Coefficient de Régularisation L1	0.5
	Coefficient de Régularisation L2	0.1
	Nombre maximal d'itérations	50
	Algorithme d'optimisation	lbfgs
Modèle BiLSTM-CNN-CRF	Initialisation des plongements	U(-0.5,0.5)
	Nombre des couches convolutions	3
	Taille des couches convolutions	30
	Dimension de la couche BiLSTM	200
	Taux d'abandon (Dropout)	0.5
	Momentum	0.9
	Taux d'apprentissage	0.01
	Taille du batch	32

TABLE 1 – Définition des paramètres d'apprentissage pour le modèle CRF et le modèle BiLSTM-CNN-CRF

4 Évaluation

Pour évaluer les différentes étapes du processus proposé, nous nous sommes appuyés essentiellement sur deux types de jeux de données : des jeux de données labellisés des avis sur des restaurants (en français), mis à disposition dans la compétition SemEval (SemEval-2016 Tâche 5 «Aspect-Based Sentiment Analysis»), et des corpus non labellisés extraits à partir du web dont deux sous-ensembles ont été manuellement labellisés pour les utiliser comme jeu de données de référence.

Nous évaluons les deux étapes du processus en termes de F1-score qui combine les mesures de précision et de rappel et s'adapte mieux à notre tâche d'identification des aspects où les termes qui ne désignent pas des aspects sont plus présents que ceux les désignant. Pour la pré-labellisation, nous avons réglé les paramètres du CRF en l'entraînant sur les seules données disponibles en français pour l'identification des aspects explicites (Jeu de données sur les restaurants de SemEval 2016). Les valeurs des paramètres retenues sont décrites dans la Table 1.

La figure 3 montre la comparaison entre l'approche utilisant tous les descripteurs de CRF (Version 1, en bleu) à celle favorisant l'adaptation du domaine en omettant le descripteur "valeur littérale" aux termes ayant comme étiquette morpho-syntaxique "Noun" et "Verb" (version 2, en orange). Les résultats confirment que la deuxième approche favorise le transfert des connaissances. Nous remarquons une amélioration de $13,3\%$, $17,5\%$, et $12,5\%$ en termes de F1-score respectivement sur le corpus des musées, celui des appareils technologiques et le corpus des produits de beauté.

A l'issue de la phase de pré-labellisation, nous constatons un taux de labels incorrects qui s'élève à 40% et un taux de labels "aspects" manquants de l'ordre de 59.5% sur le corpus (sans aucune annotation préalable) des produits de beauté (pour rappel le transfert s'effectue à partir des labels du corpus restaurants).

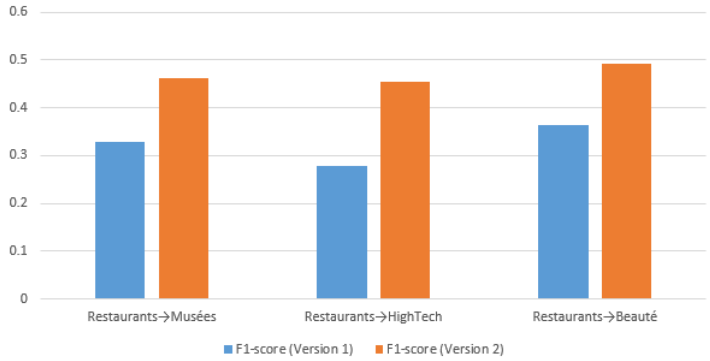

FIGURE 3 – Évaluation de la pré-labellisation par adaptation de domaines en prenant en compte tous les descripteurs (Version 1) et en ajustant les descripteurs (Version 2)

Pour la deuxième étape, nous utilisons le modèle BiLSTM-CNN-CRF comme modèle d'apprentissage profond pour l'identification des aspects explicites par apprentissage actif. Nous initialisons les plongements de mots en utilisant les plongements pré-entraînés de type Glove avec 300 dimensions (GloVe.840B.300d) fournit par (Pennington *et al.*, 2014). Nous définissons également les différents paramètres d'apprentissages comme décrit dans la table 1.

Nous réalisons les expérimentations pour l'apprentissage actif sur deux corpus pré-labellisés composés chacun de 5000 avis utilisateurs qui portent respectivement sur les appareils électroniques et les produits de beauté. Le modèle d'apprentissage BiLSTM-CNN-CRF est initialement entraînés sur 20 épochs, puis nous itérons le processus 10 fois.

La figure 4 illustre les résultats d'évaluation, en termes de F1-score, de différentes expérimentations décrites ci-dessus sur les jeux de données des avis concernant les produits de beauté et les produits technologiques. Nous comparons la stratégie de sélection par incertitude que nous avons adaptée à la stratégie de sélection aléatoire. Nous désignons par TCL le taux de correction de labels et par TED

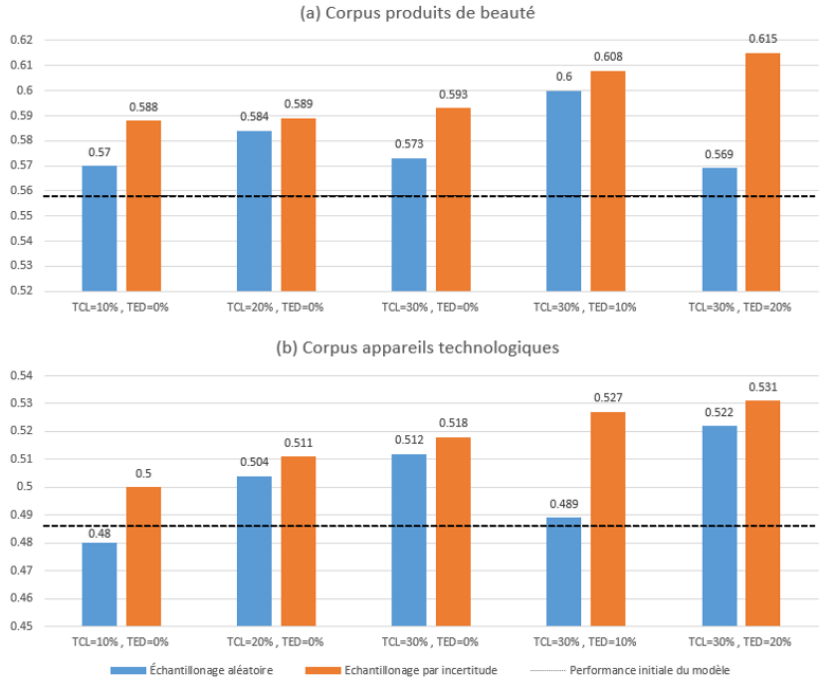

FIGURE 4 – Évaluation de l'apprentissage actif en terme de F1-score sous différentes configurations sur les corpus : (a) produits de beauté et (b) appareils technologiques.

le taux d'enrichissement des données. Nous limitons les valeurs de TCL et le TED à 30% et 20% respectivement, pour garantir un gain de 50% en termes d'effort et de temps nécessaires pour une labellisation totale du jeu de données d'apprentissage. Nous avons également favorisé la correction des labels afin de réduire l'impact des erreurs de pré-labellisation sur la performance du modèle. La ligne en pointillé représente la performance du modèle BiLSTM-CNN-CRF avant l'application de l'apprentissage actif.

Globalement, les résultats soulignent l'intérêt de la mise en place d'un processus d'apprentissage actif puisque nous observons une amélioration autour de 5% et de 6% en terme de F1-score respectivement sur le corpus des avis utilisateurs sur les appareils technologiques et celui sur les produits de beauté par rapport aux résultats obtenus par le modèle BiLSTM-CNN-CRF sans application de l'apprentissage actif (sans enrichir les données ni les corriger par l'expert).

5 Conclusion

Dans cet article, nous avons proposé un processus de bout en bout qui s'appuie sur l'apprentissage actif pour améliorer la labellisation des aspects explicites, pour les langues à faibles ressources. Pour effectuer les tests, nous avons utilisé un corpus composé d'avis utilisateurs (en français) sur les produits de beauté et un autre sur les appareils électroniques, les deux corpus créés à partir de

Web. Notre approche consiste en un processus en deux étapes, commençant par une pré-labellisation pour remédier à la rareté des données labellisées via le transfert de connaissances par adaptation de domaines. Ensuite, le processus d'apprentissage actif est utilisé pour corriger les erreurs de pré-labellisation et réduire ainsi les coûts d'annotation manuelle. Les expériences montrent que l'apprentissage actif améliore considérablement les performances du modèle d'apprentissage lorsque 30% des labels initiaux sont corrigés. Dans les travaux futurs nous adapterons le processus proposé pour permettre l'identification et labellisation des termes d'aspect sur des ensembles de données composés de textes exprimés en différentes langues. De plus, nous ajouterons une autre phase de traitement pour catégoriser les aspects identifiés.

Références

CHEN Z. & QIAN T. (2020). Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2107–2117.

CULOTTA A. & MCCALLUM A. (2005). Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, p. 746–751.

LI K., CHEN C., QUAN X., LING Q. & SONG Y. (2020). Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online : Association for Computational Linguistics.

LI X., BING L., LI P., LAM W. & YANG Z. (2018). Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

LIU P., JOTY S. & MENG H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, p. 1433–1443.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.

REN P., XIAO Y., CHANG X., HUANG P.-Y., LI Z., GUPTA B. B., CHEN X. & WANG X. (2021). A survey of deep active learning. *ACM Computing Surveys (CSUR)*, **54**(9), 1–40.

RINGGER E., MCCLANAHAN P., HAERTEL R., BUSBY G., CARMEN M., CARROLL J., SEPPI K. & LONSDALE D. (2007). Active learning for part-of-speech tagging : Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, p. 101–108.

SANTOS B. N. D., MARCACINI R. M. & REZENDE S. O. (2021). Multi-domain aspect extraction using bidirectional encoder representations from transformers. *IEEE Access*, **9**, 91604–91613.

SCHEFFER T., DECOMAIN C. & WROBEL S. (2001). Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, p. 309–318 : Springer.

SCHRÖDER C. & NIEKLER A. (2020). A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv :2008.07267*.

SETTLES B. (2009). Active learning literature survey.

SETTLES B. & CRAVEN M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, p. 1070–1079.

SHELMANOV A., LIVENTSEV V., KIREEV D., KHROMOV N., PANCHENKO A., FEDULOVA I. & DYLOV D. V. (2019). Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 482–489.

SHEN Y., YUN H., LIPTON Z. C., KRONROD Y. & ANANDKUMAR A. (2018). Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

XU H., LIU B., SHU L. & YU P. S. (2018). Double embeddings and CNN-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 592–598.