Correction automatique de réponses textuelles : une approche basée sur des schémas conceptuels

Emel Dalgic*, Alexandre Lefort**, Iana Atanassova*,***

*Université de Franche-Comté, CRIT
F-25000 Besançon, France
**E-Cole, France
***Institut Universitaire de France (IUF)

Résumé. Dans cet article, nous proposons une méthode pour la correction automatique de définitions produites par les étudiants dans le domaine de la gestion et l'économie. Notre algorithme traite les réponses textuelles des étudiants afin d'y identifier les concepts et relations correctement restituées et ceux qui sont erronés ou manquants. Nous avons effectué une évaluation sur plusieurs corpus de définitions, et les résultats montrent un score de F1 d'autour de 0,91 pour l'identification des relations, et des scores entre 0,75 et 0,83 pour l'identification des groupes nominaux arguments des relations. Cette approche s'inscrit dans un projet plus vaste autour de la correction automatique des réponses textuelles et la production d'outils destinés aux enseignants et apprenants.

1 Introduction et problématique de recherche

Dans cet article nous proposons un algorithme de traitement de réponses textuelles à des questions d'examens à des fins de correction automatique. L'algorithme traite en particulier les définitions de notions en économie-gestion, formulées sous formes de texte libre. Ce travail s'inscrit dans un projet plus vaste d'outil destiné aux enseignants pour offrir une aide à l'évaluation de la justesse des réponses des étudiants et une évaluation automatique.

À partir d'un ensemble de définitions d'un concept $D = \{D_1, D_2, ...\}$ produites par l'enseignant comme des définitions correctes et une définition C produite par l'étudiant, notre objectif est de comparer le contenu sémantique de C de manière à identifier toutes les différences de sens qui existent entre C et la définition la plus proche de l'ensemble D. Pour être utilisés à des buts d'apprentissage et dans un système de correction semi-automatique, les résultats proposeraient les éléments nécessaires pour produire une description textuelle des éventuelles erreurs dans la définition C à destination de l'étudiant ou de l'enseignant en plus d'un score de similarité.

La correction des réponses textuelles est directement liée aux domaines de l'évaluation automatique (*Automated Scoring*, (Madnani et Cahill, 2018)) et de l'évaluation automatique des écrits (*Automatic Essay Scoring, AES*, (Ke et Ng, 2019)). La stratégie mise en place dans ces domaines est l'utilisation d'algorithmes de Machine Learning entraînés sur des corpus généraux de très grande taille afin de calculer le degré de similarité entre la réponse attendue,

supposée fournie, et la réponse donnée. Ces solutions ne sont toutefois pas exploitables pour la correction automatique de réponses textuelles à des questions ouvertes.

Les réponses sous forme de phrases écrites en langue naturelle (français) doivent être analysées afin d'identifier les concepts exprimés et les relations entre eux, et comparer ce contenu à un modèle correct fourni par le professeur. Les obstacles, non résolus par les approches actuelles en Traitement Automatique des Langues (TAL) (Bender et Koller, 2020), viennent du fait que l'algorithme doit pouvoir identifier de manière précise le contenu sémantique des réponses, afin de répondre à des exigences de précision (fiabilité des résultats) et de traçabilité. Ceci est nécessaire afin de proposer des indications aux usagers sur les éventuelles erreurs.

Les approches faisant appel à des connaissances linguistiques sont souvent préférées dans des contextes d'entreprise (Madnani et Cahill, 2018; Chiticariu et al., 2013) car contrairement aux approches par apprentissage (Otter et al., 2020), la qualité des résultats peut être contrôlée. De plus, dans un domaine restreint, il est en effet possible de créer automatiquement ou semi-automatiquement (c'est-à-dire de manière assistée) des structures qui décrivent la logique du texte sous forme de règles linguistiques. Ces structures étant explicites, l'explication des résultats de ces méthodes peut être proposée aux utilisateurs et des garanties de robustesse peuvent être données grâce à cette traçabilité. Une grande partie des applications développées actuellement en TAL fait appel à des approches hybrides (Gomez-Perez et al., 2020). Pour notre tâche qui consiste en la comparaison entre données textuelles, une représentation des définitions sous forme de graphes de relations, qui s'apparentent à des ontologies, est une approche pertinente (voir par ex. Krötzsch (2017); Ghanavati et Breaux (2015)). Dans Nakamura-Delloye (2011), les relations entre les entités nommées sont identifiées sur base d'analyse syntaxiques en dépendance afin d'enrichir une ontologie.

2 Création d'un corpus de définitions

Nous avons créé plusieurs corpus de définitions dans le domaine de l'économie-gestion, présentés dans la table 1.

Le corpus DEF-1 provient des anciennes épreuves de Management des Organisations et de Gestion et Finance du baccalauréat STMG en France [1]. Nous l'avons enrichi en utilisant ChatGPT [2].

Le corpus DEF-2 a été obtenu également à l'aide de ChatGPT [3] Tandis que les définitions du corpus DEF-1 débutent avec des formes de présentation comme *"Le stock correspond à [arg2]"* ou *"Un stock représente [arg2] "*, celles de DEF-2 commencent directement avec un groupe nominal correspondant à [arg2].

Le corpus DEF-3 contient les 21 premières définitions extraites du site web *"Financière Fonds Privés"* [4], dont les débuts ont été modifiées pour ressembler à celles de DEF-1. Des reformulations ont été obtenues par ChatGPT [5]. Dans ces trois corpus, toutes les définitions

1. Téléchargés à partir de `www.sujetdebac.fr`
2. ChatGPT version 3.5, consultée le 22/02/2023. Prompt utilisé : "Peux-tu reformuler la définition suivante de deux façons différentes : ...".
3. ChatGPT version 4, consulté le 21/03/2023. Prompts utilisés : "Donnez-moi un lexique de définitions de 50 termes de la comptabilité."; "Reformulez ces définitions de 3 manières différentes".
4. `https://www.financiere-fondsprives.com/vocabulaire/`
5. ChatGPT version 3.5, consulté le 22/06/2023. Prompt utilisé : "Peux-tu reformuler la définition suivante de cinq façons différentes : ... ".

obtenues par ChatGPT ont été vérifiées manuellement pour s'assurer qu'elles sont correctes. Nous avons écarté les reformulations qui introduisent de différences de sens.

Le corpus DEF-4 contient des définitions fournies par des étudiants comme réponses à des questions d'examens, réalisés à l'Université de la Réunion en 2022 et 2023 en Licence-1 d'Administration économique et sociale (AES).

TAB. 1 – *Description des corpus : nombre de définitions, de notions, et de relations qui ont été identifiées manuellement*

			Relations		
Corpus	Définitions	Notions	Équivalence	Possession	But
DEF-1	159	42	172	16	67
DEF-2	135	45	15	2	29
DEF-3	126	21	141	13	53
DEF-4	77	7	25	4	36

3 Méthode pour l'analyse et la correction des définitions

La construction d'un algorithme qui évalue et corrige les définitions nécessite une analyse du contenu textuel des réponses du point de vue des concepts qui y sont présents et les relations qui sont exprimées entre eux. En effet, une même définition peut être exprimée en utilisant une diversité de structures syntaxiques et/ou de reformulations basées sur des expressions synonymes. Pour cette raison, une analyse des occurrences des termes ne serait pas suffisante et nous avons besoin d'identifier les relations qui sont exprimées entre les différents concepts qui entrent en jeu. Les concepts et leurs relations s'organisent dans un réseau que nous appelons *schéma conceptuel*. Par exemple, la figure 1 montre le schéma conceptuel de la définition : *"Un stock est un ensemble de produits physiques possédés par une entreprise, en attendant d'être utilisés dans le processus de production ou d'être vendus sur un marché"*. Ce schéma propose quatre concepts qui sont reliés par les relations *"But"* et *"Possession"*. La restitution correcte de ces quatre concepts et des relations entre eux produirait une réponse correcte, quelle que soit la syntaxe de la phrase employées par un étudiant.

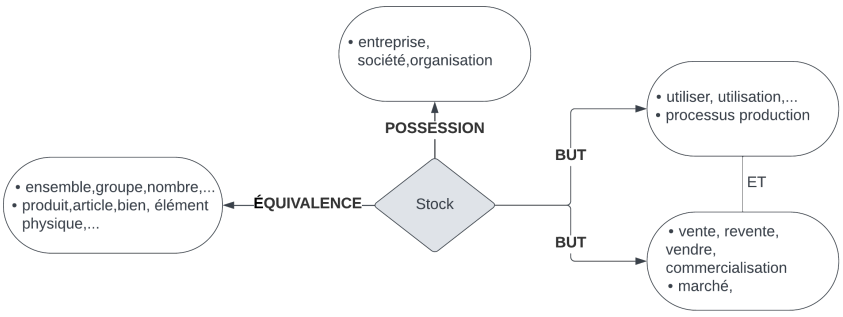

FIG. 1 – *Exemple d'un schéma conceptuel "stock"*

Le fonctionnement de l'algorithme que nous proposons est présenté sur la figure 2. Les principales étapes de traitement sont décrites comme suit :

1. L'ensemble des définitions correctes et la définition à corriger constituent l'entrée.
2. Les groupes nominaux (GN) et les relations présentes au sein des définitions sont identifiés par nos algorithmes de détection de relations et de GN (voir section 3.1). La liste des GN est enrichie par un dictionnaire de synonymes et d'antonymes.
3. Sorties obtenues : schémas conceptuels pour les définitions correctes (SC) et un schéma conceptuel de la définition de l'étudiant (SE).
4. Nous identifions le SC contenant le plus d'éléments similaires au SE.
5. Comparaison entre le SC identifié et le SE, en s'appuyant sur l'ensemble des relations, des GN et leurs synonymes dans les deux définitions.
6. Sorties obtenues : ensembles de relations correctes et manquantes, et entités correctes, incorrectes et manquantes. Ces éléments permettent de proposer une correction de la réponse de l'étudiant.

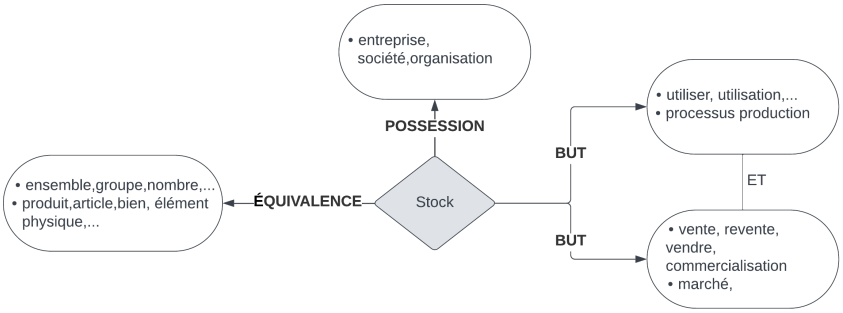

FIG. 2 – *Schéma général de l'algorithme correcteur*

3.1 Analyse des relations

Sur la base des corpus collectés, nous avons identifié manuellement les relations présentes dans les définitions, dont la grande majorité sont les 3 types de relations suivants : *équivalence*, *possession* et *but* (voir aussi le tableau 1).

Notre objectif est d'extraire ces relations sous forme de triplets (relation, arg1, arg2) pour les relations de possession et d'équivalence, et sous forme de doublet (relation, arg1) dans le cas de la relation de but (voir la figure 3, étape 1).

Nous nous appuyons sur les arbres syntaxiques des phrases, obtenus par SpaCy (Honnibal et Montani, 2017). Nous avons créé des règles de repérage pour chaque type de relation qui consistent en des ensembles de conditions sur les noeuds et les relations présents dans l'arbre syntaxique de la phrase. L'implémentation des règles est faite à l'aide des modules Dependency Matcher (DM) et Matcher [6] de SpaCy.

6. `https://spacy.io/api/dependencymatcher`; `https://spacy.io/api/matcher`

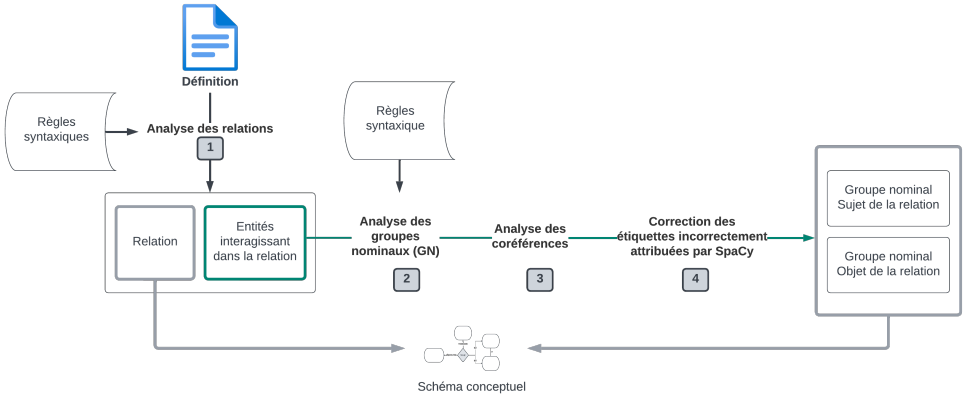

FIG. 3 – *Processus d'analyse des relations et des groupes nominaux*

La figure 4 présente une règle implémentée avec DM qui identifie la relation *but*. Le premier noeud est l'*ancre*, qui prend comme arguments deux éléments. RIGHT_ID est l'identifiant du nœud. RIGHT_ATTRS sont des conditions qui s'appliquent sur ce noeud. Elles peuvent concerner son orthographe, son lemme, son POS-tag, ou la relation de dépendance dans laquelle il entre. LEFT_ID est l'identifiant du nœud se trouvant à gauche de RIGHT_ID dans l'arbre de dépendance, et REL_OP décrit le lien entre LEFT_ID et RIGHT_ID. Par exemple, " > " annonce que LEFT_ID dépend immédiatement du nœud RIGHT_ID.

```
[{"RIGHT_ID": "ancre",
    "RIGHT_ATTRS": {'LEMMA': {"IN": ['vise','viser',  'servir','destiner']}}},
 {"LEFT_ID": "ancre",    "REL_OP": ">",   "RIGHT_ID": "but",
    "RIGHT_ATTRS": {"DEP": {"IN": ["xcomp","obl:arg"]}}    }]
```

FIG. 4 – *Exemple d'une règle syntaxique pour la relation but*

Au total 30 règles ont été implémentées, dont 14 pour la relation d'équivalence, 7 pour celle de possession et 9 pour celle de but. La version finale des règles a été obtenue par des tests successifs sur les corpus et des modifications itératives.

3.2 Analyse des Groupes Nominaux (GN)

L'extraction des GN utilise le module DM. La liste des GN considérés comme corrects pour une définition est obtenue par l'enrichissement à partir de dictionnaires de synonymes et d'antonymes que nous avons constitués spécifiquement pour cette tâche, à partir des sources suivantes : Le Larousse [7], le graphe de connaissances ConceptNet [8], ainsi que les définitions de nos corpus.

7. https://www.larousse.fr/dictionnaires/synonymes, consulté en mai 2023
8. https://conceptnet.io/, (Speer et al., 2018)

L'extraction des GN fait l'objet de traitements supplémentaires lorsque la phrase contient des coréférences (voir la figure 3, étapes 2 et 3).

Par exemple, dans la définition : *"Le Bilan est un tableau résumant l'ensemble des actifs et passifs d'une entreprise, [...]. Il représente le patrimoine de l'entité."*, plusieurs relations d'équivalence sont présentes, dont la dernière indique que "bilan" désigne "patrimoine de l'entité". L'analyse des coréférences, avec le module *coreferee* de SpaCy [9], nous permet d'identifier les arguments corrects de la relation d'équivalence : (arg1 : '['Bilan']', arg2 : '['patrimoine', 'entité']').

Nous avons identifié deux cas spécifiques, où SpaCy attribue des POS-tag et des étiquettes de dépendances incorrects : le déterminant 'tout' est toujours identifié comme adjectif ; un participe passé employé comme adjectif est toujours identifié comme verbe. Dans ces deux cas, nous modifions ces étiquettes pour obtenir l'analyse correcte de la phrase, avant d'appliquer les règles de DM.

4 Résultats et discussion

Nous avons évalué l'identification des relations et des GN avec des mesures de précision et rappel sur tous les corpus, qui ont été annotés manuellement au préalable. Le tableau 1 présente le nombre de relations annotées dans chaque corpus. Dans certains cas, où les groupes nominaux soient identifiés par l'algorithme mais restitués de manière incomplète, nous considérons que le résultat est à la fois un faux négatif et un faux positif.

Le tableau 2 présente les résultats de l'évaluation de l'identification des relations et des groupes nominaux interagissant dans les relations. Pour les relations, nous remarquons que les scores de précision sont plus élevés que ceux du rappel, la majorité des erreurs correspondant à des faux négatifs. Les résultats pour le F1 sont autour des 91 %. La relation d'équivalence est celle qui est la mieux identifiée avec F1 de 92,1 %. Le corpus DEF-4 obtient des résultats moins élevés pour les relations de possession et de but. Ceci est dû au fait que les phrases de ce corpus, produites par les étudiants, ne sont pas toujours grammaticalement correctes. Pour les GN, les valeurs des F1 sont plus faibles que celles des relations, parce que la détection des relations a un impact sur celle des GN. De plus, un certain nombre des groupes nominaux identifiés ne sont que partiellement corrects et dans ce cas ils sont considérés comme faux. Les scores de F1 sont entre 75,1 % et 83,5 %, les GN de la relation de but étant les mieux identifiés, sauf pour DEF-4.

L'analyse des erreurs de notre algorithme indique que les résultats de l'identification des relations et des GN peuvent être améliorés par la prise en compte de nouvelles structures syntaxiques.

Dans d'autres cas, le rappel peut être amélioré par l'ajout de nouvelles règles syntaxiques. Etant donné que notre travail se focalise sur les définitions d'un domaine restreint, qui est l'économie-gestion, il est possible de créer des ensembles de règles qui couvrent la totalité des formulations de manière satisfaisante. Les résultats assez élevés de nos évaluations vont dans ce sens. Certaines erreurs sont dues à l'analyse de SpaCy, qui dans certains cas ne reconnaît pas correctement les GN. Il pourrait donc être pertinent de rajouter des traitements supplémentaires qui corrigent ce type d'erreurs.

9. Voir `https://spacy.io/universe/project/coreferee`.

TAB. 2 – *Résultats de l'identification des relations*

		P	R	F1	P	R	F1	P	R	F1
		Équivalence			Possession			But		
Relations	DEF-1	100	90,0	94,7	100	100	100	96,8	91,0	93,8
	DEF-2	100	72,2	82,8	100	100	100	100	96,5	98,2
	DEF-3	99,1	83,0	90,3	86,7	100	92,9	95,7	84,9	90,0
	DEF-4	100	80,0	88.9	100	60,0	75,0	100	66,7	80,0
	Total	99,7	85,6	92,1	100	97,1	91,0	97,5	85,4	91,0
Gr. nominaux	DEF-1	93,4	78,5	85,4	80,0	66,7	72,7	89,0	87,5	88,3
	DEF-2	83,3	29,4	43,5	100	100	100	100	93,1	96,5
	DEF-3	88,5	51,0	64,6	87,5	77,8	82,3	92,5	77,0	84,1
	DEF-4	94,1	59,2	72,7	100	60,0	75,0	88,2	44,1	58,8
	Total	91,6	63,7	75,1	90,0	74,0	81,2	92,0	76,6	83,5

Nos résultats mettent en évidence la difficulté de traiter les données réelles, c'est-à-dire les réponses produites par les étudiants lors d'examens, du fait des erreurs qu'elles comportent. En effet, notre algorithme étant dépendant de la qualité de l'analyse morphosyntaxique par SpaCy, les scores obtenus pour le corpus DEF-4 sont moins élevés que ceux des corpus de définitions correctes. L'intégration de corrections orthographiques et grammaticales en amont des traitements pourrait améliorer ces résultats.

5 Conclusion

Nous avons proposé une approche innovante pour la création de schémas conceptuels à partir de phrases définitoires dans le but de proposer une correction de réponses textuelles en gestion et économie. Notre approche, base sur des analyses syntaxiques et des règles linguistiques, permet de produire des sorties qui ont pour objectif d'expliquer les erreurs faites par les apprenants et d'indiquer les éléments incorrects ou manquants dans une définition. Nous avons constitué et analysé plusieurs corpus de définitions, produites par des professeurs et par des étudiants. Les résultats des évaluations montrent que notre algorithme produit des résultats satisfaisants, avec F1 score d'autour de 0,94 pour l'extraction des relations et une moyenne de F1 de 0,81 pour l'identification des GN.

Ce travail fait partie d'un projet d'envergure autour des outils et interfaces à destination des apprenants et professeurs dans des domaines de spécialité. Nos futurs travaux s'orientent vers la généralisation de cette approche pour prendre en compte d'autres domaines disciplinaires et l'amélioration progressive des règles d'identification par l'analyse de données réelles.

Références

Bender, E. M. et A. Koller (2020). Climbing towards nlu : On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198.

Chiticariu, L., Y. Li, et F. Reiss (2013). Rule-based information extraction is dead ! long live rule-based information extraction systems ! In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 827–832.

Ghanavati, S. et T. D. Breaux (2015). Comparing and analyzing definitions in multi-jurisdictions. In *2015 IEEE Eighth International Workshop on Requirements Engineering and Law (RELAW)*, pp. 47–56. IEEE.

Gomez-Perez, J. M., R. Denaux, et A. Garcia-Silva (2020). Hybrid natural language processing : An introduction. In *A Practical Guide to Hybrid Natural Language Processing*, pp. 3–6. Springer.

Honnibal, M. et I. Montani (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ke, Z. et V. Ng (2019). Automated essay scoring : A survey of the state of the art. In *IJCAI*, Volume 19, pp. 6300–6308.

Krötzsch, M. (2017). Ontologies for knowledge graphs ? In A. Artale, B. Glimm, et R. Kontchakov (Eds.), *Proceedings of the 30th International Workshop on Description Logics (DL 2017)*, Volume 1879 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Madnani, N. et A. Cahill (2018). Automated scoring : Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 1099–1109. Association for Computational Linguistics.
French

Nakamura-Delloye, Y. (2011). Extraction non-supervisée de relations basée sur la dualité de la représentation (unsupervised relation extraction based on the dual representation). In M. Lafourcade et V. Prince (Eds.), *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, Montpellier, France, pp. 194–199. ATALA.

Otter, D. W., J. R. Medina, et J. K. Kalita (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems 32*(2), 604–624.

Speer, R., J. Chin, et C. Havasi (2018). Conceptnet 5.5 : An open multilingual graph of general knowledge.

Summary

In this article, we propose a method for the automatic correction of definitions produced by students in the field of management and economics. Our algorithm processes the students' textual responses to identify which concepts and relations are correctly restituted and which are erroneous or missing. We carried out an evaluation on several corpora of definitions, and the results show an F1 score of around 0.94 for the identification of relations, and between 0.75 and 0.88 for the identification of nominal groups which are elements of the relations. This approach is part of a wider project involving the automatic correction of textual responses and the production of tools for teachers and learners.