

Université de Montréal

Méthodes d'évaluation en extraction d'information ouverte

par

Fabrice Lamarche

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de Maître ès sciences (M.Sc.) en
Informatique

30 Août 2023

© Fabrice Lamarche, 2023

Université de Montréal

Département d'informatique et de recherche opérationnelle, Faculté des arts et des sciences

Ce mémoire intitulé

Méthodes d'évaluation en extraction d'information ouverte

Présenté par

Fabrice Lamarche

A été évalué par un jury composé des personnes suivantes

Guy Lapalme

Président-rapporteur

Philippe Langlais

Directeur de recherche

Jian-Yun Nie

Membre du jury

Résumé

L'extraction d'information ouverte (OIE) est un domaine du traitement des langues naturelles qui a pour but de présenter les informations contenues dans un texte dans un format régulier permettant d'organiser, d'analyser et de réfléchir sur ces informations. De nombreux systèmes d'OIE existent, revendiquant des performances de plus en plus grandes. Afin d'établir leurs performances et de les comparer, il est nécessaire d'utiliser une référence. Celles-ci ont également évolué dans le temps et se veulent précises et objectives, permettant d'identifier les systèmes les plus performants. Dans ce mémoire, nous identifions certaines limitations des méthodes d'évaluation actuelles et nous proposons une nouvelle référence permettant d'y remédier. Cette nouvelle référence comprend deux composantes principales, soit une annotation manuelle de phrases candidates et une fonction permettant d'établir la concordance syntaxique entre différents faits extraits et annotés. De plus, nous proposons de nouvelles lignes directrices pour mieux encadrer et définir la tâche d'extraction d'information ouverte elle-même, ce qui permet de quantifier de manière plus précise les informations pertinentes extraites par les systèmes d'OIE. Nos expériences démontrent que notre référence suit de plus près ces lignes directrices que les références précédentes, qu'elle parvient à mieux juger de la concordance entre les faits extraits et les faits annotés et qu'elle est plus souple dans son approche que la référence qui est à ce jour l'état de l'art. Notre nouvelle référence permet de tirer des conclusions intéressantes sur les performances réelles des systèmes d'extraction d'information ouverte. Nous montrons notamment que les systèmes les plus récents ne sont pas forcément les meilleurs.

Mots-clés : extraction d'information ouverte, jeux de données, évaluation de tâche, traitement automatique des langues naturelles

Abstract

Open Information Extraction (OIE) is a field of natural language processing whose aim is to present the information contained in a text in a regular format that allows that information to be organized, analyzed and reflected upon. Numerous OIE systems exist, claiming ever-increasing levels of performance. In order to establish their performance and compare them, it is necessary to use a reference. These have also evolved over time, and are intended to be precise and objective, making it possible to identify the best-performing systems. In this thesis, we identify some of the limitations of current evaluation methods and propose a new benchmark to remedy them. This new benchmark comprises two main components : a manual annotation of candidate sentences and a function to establish syntactic concordance between different extracted and annotated facts. In addition, we propose new guidelines to frame and better define the open information extraction task itself, enabling us to better quantify and measure the amount of relevant information extracted by OIE systems. Our experiment shows that our benchmark follows these guidelines more closely than previous benchmarks, is better at judging the match between extracted and annotated facts, and is more flexible than the current state-of-the-art benchmarks. Our new benchmark allows us to draw some interesting conclusions about the actual performance of open information extraction systems. We show that the latest systems are not necessarily the best.

Keywords : open information extraction, datasets, task evaluation, natural language processing

Table des matières

Résumé	1
Abstract	2
Table des matières	5
Liste des Tables	7
Liste des Figures	8
Liste des Sigles	9
1 Introduction	10
1.1 Extraction d’information ouverte	10
1.2 Motivations	12
1.3 Contributions	14
1.4 Organisation du mémoire	14
2 Évaluation en OIE	16
2.1 Définition de la tâche	16
2.2 Évaluation générale	18
2.3 Métriques d’évaluation	19
3 Revue de littérature	21
3.1 Systèmes d’OIE	21
3.1.1 TextRunner	21
3.1.2 ReVerb	21
3.1.3 ClausIE	22
3.1.4 MinIE	24
3.1.5 RnnOIE	25
3.1.6 Seq2Seq	25
3.1.7 IMojIE	26
3.1.8 OpenIE6	27
3.1.9 M2OIE	27
3.1.10 CompactIE	28
3.2 Références d’OIE	28
3.2.1 OIE2016	29

3.2.2	WiRe57	29
3.2.3	CaRB	30
3.2.4	BenchIE	31
3.3	Limitations des références actuelles	33
3.3.1	Outil Power BI	33
3.3.2	Analyse de BenchIE	36
4	Approche proposée	42
4.1	Lignes directrices d’annotation	42
4.2	Lignes directrices de correspondance	45
4.3	Données	48
4.4	Annotations (Plateforme AnnIE)	49
4.5	Fonction de correspondance	49
5	Analyse	52
5.1	Annotations	52
5.2	Correspondance	53
5.3	Limitations de la fonction de correspondance	56
5.4	Comparaison avec les annotations manuelles	58
5.5	Comparaison des références	59
5.6	Évaluation sur tâche subséquente	61
	Conclusion	66
	Références	67
A	Lignes directrices d’annotation	73
A.1	Nombre d’argument	73
A.2	Information pertinente	73
A.3	Minimalité	74
A.4	Exhaustivité	75
A.5	Relation complètes	75
A.6	Coréférence	76
A.7	Inférence	76
A.8	Relations non verbales	78
A.9	Reformulation	79
A.10	Voix active ou passive	79
A.11	Argument unique	79
A.12	Spéculation et attribution	80

A.13	Correction	80
A.14	Type d'information spécifique	80
A.15	Langue	81
B	Lignes directrices de correspondance	82
B.1	Extractions égales	82
B.2	Spécificité de la relation	82
B.3	Erreurs de syntaxe	83
B.4	Mots équivalents	83
B.5	Niveau de détail	83
B.6	Relations complètes	84
C	Exemples d'annotation et de correspondance	85

Liste des tableaux

1	Exemples d’extractions sur une phrase tirée de la page Wikipedia de Chilly Gonzales (https://en.wikipedia.org/wiki/Chilly_Gonzales) par trois extracteurs.	11
2	Exemples d’extractions explicites et implicites sur une phrase du jeu de données original de la référence <i>OIE2016</i>	17
3	Exemple d’extraction acceptée dans <i>CaRB</i> adapté de la Table 1 de Gashteovski et al. (2022)	31
4	Exemples de faits manquants de <i>BenchIE</i>	37
5	Exemples de faits non pertinents de <i>BenchIE</i>	37
6	Exemples de faits erronés de <i>BenchIE</i>	38
7	Exemples de faits incomplets de <i>BenchIE</i>	38
8	Exemples d’erreurs de correspondance de <i>BenchIE</i>	39
9	Exemples de doubles annotations de <i>BenchIE</i>	39
10	Exemples de sens différents de <i>BenchIE</i>	40
11	Proportion des différents types d’erreurs sur le sous-ensemble de 50 phrases de <i>BenchIE</i>	40
12	Annotation : principe d’informativité	43
13	Annotation : principe de minimalité	44
14	Annotation : principe d’exhaustivité	44
15	Annotation : principe de complétude des relations	44
16	Annotation : principe de coréférence	45
17	Correspondance : principe de spécificité de la relation	46
18	Correspondance : erreurs de syntaxe	47
19	Correspondance : principe du choix de mot	47
20	Correspondance : niveau de détail	48
21	Exemple de correspondance selon les critères de la fonction de correspondance sur une phrase de WiRE57	51
22	Caractéristiques des annotations	52
23	Exemples des annotations de <i>BenchIE</i> et de <i>BenchIE^{FL}</i>	54
24	Performance des fonctions de correspondance	55
25	Comparaison de la sortie de la fonction avec les annotations manuelles. Les éléments entre parenthèse sont ceux obtenus sur le jeu de données d’entraînement, les autres sur le jeu de données de test.	57
26	Exemples de faux positifs par la fonction de correspondance	57
27	Exemples de faux négatifs par la fonction de correspondance	58

28	Corrélations entre les fonctions de correspondance et les annotations manuelles sur <i>BenchIE50^{FL}</i>	59
29	Résultats (F1) des systèmes sur différentes références en utilisant leur script d'évaluation respectifs.	61
30	Exemples de triplets question-passage-réponse de WebAssertions et WebAssertionsFL	63
31	Scores des systèmes sur <i>WebAssertions^{FL}</i>	64
32	Corrélation entre les scores des systèmes selon les références et leurs scores sur <i>WebAssertions^{FL}</i>	64

Table des figures

1	Expression régulière utilisée par ReVerb afin d'identifier les relations cohérentes.	22
2	Types de clauses utilisés par ClausIE selon la Table 1 de Del Corro et Gemulla (2013)	23
3	Organigramme décisionnel pour la détection des types de verbes et types de clauses utilisé par ClausIE repris de la Figure 2 de Del Corro et Gemulla (2013) .	24
4	Architecture du modèle RnnOIE repris de la Figure 2 de Stanovsky, Michael, Zettlemoyer, et Dagan (2018)	25
5	Architecture du modèle Seq2Seq repris de la Figure 1 de Cui, Wei, et Zhou (2018)	26
6	Architecture du modèle M2OIE (tiré de la figure 2 de Ro, Lee, et Kang (2020)) .	28
7	Exemple de <i>synsets</i> utilisés par <i>BenchIE</i> selon la Table 2 de Gashteovski et al. (2022)	32
8	1 ^e capture d'écran de notre plateforme Power BI portant sur <i>BenchIE</i>	34
9	2 ^e capture d'écran de notre plateforme Power BI portant sur <i>BenchIE</i>	35
10	Capture d'écran d' <i>AnnIE</i> démontrant les fonctionnalités de la plateforme	49
11	Score des systèmes selon les fonctions de correspondance sur <i>BenchIE50^{FL}</i> . . .	59
12	Résultats (F1) des systèmes sur différentes références en utilisant leur script d'évaluation respectifs.	60

Liste des Sigles

IE Extraction d'information.

OIE Extraction d'information ouverte.

TALN Traitement automatique des langues naturelles.

Chapitre 1 - Introduction

Dans cette section, nous décrivons de manière simplifiée le sujet de ce mémoire en introduisant l'extraction d'information et l'extraction d'information ouverte (nous utiliserons l'abréviation OIE par la suite) et en expliquant la tâche spécifique. Nous soulevons également quelques défis spécifiques à l'évaluation de l'OIE. Nous décrivons enfin les motivations derrière ces travaux ainsi que les contributions proposées.

1.1 Extraction d'information ouverte

Le traitement automatique des langues naturelles (TALN) est un domaine à l'intersection de la linguistique, de l'informatique et de l'intelligence artificielle, qui vise à développer des méthodes de compréhension, de traitement et de manipulation des langues. Un sous-domaine du TALN est l'extraction d'information (IE). Introduite informellement par (Cowie & Wilks, 1996), cette tâche vise à extraire de phrases les informations véhiculées par celles-ci dans un format donné. Généralement, dans le cadre de l'IE, on dispose d'une liste de *relations* à priori, et on cherche à trouver dans le texte toutes les *entités* qui satisfont ces relations. Ces informations extraites s'organisent dans la plupart des cas dans des tuples de la forme suivante :

(Argument 1 - Relation X - Argument 2 - Argument 3, ...)

La plupart des relations sont satisfaites par seulement deux arguments, mais certains systèmes permettent l'extraction de tuples à plus de deux arguments. Dans la suite de ce mémoire, nous référerons à ces tuples en utilisant les termes *tuple*, *extraction* ou *fait*, de manière interchangeable.

Le but premier de cette manipulation est d'organiser le texte, permettant ainsi d'autres opérations subséquentes (sur celui-ci). Cela permet également une forme de raisonnement sur le texte, chose généralement impossible sur du texte brut. Un des intérêts principaux pour cette tâche, ainsi que pour le TALN plus largement, provient de la quantité énorme de textes non structurés disponibles, principalement sur le web. L'analyse de ces documents peut être facilitée par des systèmes d'extraction d'information et par l'organisation de l'information que ceux-ci permettent.

(Yates et al., 2007) introduit un nouveau paradigme d'extraction d'information, l'extraction d'information ouverte (OIE), mettant de l'avant certaines des limitations de l'extraction d'information classique. La principale de ces limitations est que les systèmes d'IE nécessitent des données annotées manuellement ou des patrons créés à la main pour l'extraction et ce,

pour chaque relation d'intérêt, ce qui limite grandement la portée de ces systèmes. Les auteurs proposent un premier système d'OIE, *TextRunner*, dont nous explorerons le fonctionnement de manière plus détaillée à la section 3.1.1 et qui place les balises premières de ce qu'est l'OIE. La tâche ici est similaire à l'IE traditionnelle, à la différence près qu'aucune relation n'est spécifiée préalablement et que toutes celles étant exprimées dans le texte original doivent être extraites sous forme de triplets. Des exemples d'extractions retournées par différents systèmes d'OIE sur une phrase tirée de la page Wikipédia de Chilly Gonzales (https://en.wikipedia.org/wiki/Chilly_Gonzales) sont présentés à la Table 1. On y retrouve l'ensemble des faits extraits par les systèmes sur cette phrase. On observe la disparité dans la sortie des systèmes.

He is the younger brother of the prolific film composer Christophe Beck.

Reverb

(He - is the younger brother of - the prolific film composer Christopher Beck)

MinIE

(Christophe Beck - is - prolific film composer)

(He - is younger brother of - Christophe Beck)

(He - is - younger brother)

OpenIE6

(He - is - the younger brother of the prolific film composer Christophe Beck)

TABLE 1 – Exemples d'extractions sur une phrase tirée de la page Wikipedia de Chilly Gonzales (https://en.wikipedia.org/wiki/Chilly_Gonzales) par trois extracteurs.

On voit dès lors apparaître un point important de nos travaux qui est de différencier ce qui est exprimé de ce qui n'est pas exprimé dans une phrase donnée. Cette notion, de par sa nature subjective, n'est pas consensuelle, et est définie de façon différente par différents auteurs. (Wu & Weld, 2010) définissent par exemple la tâche en déclarant qu'un système d'OIE devrait

“produce one triple for every relation stated explicitly in the text, but is not required to infer implicit facts”

, alors que dans les faits, nombre de systèmes subséquents font des extractions implicites, à nos yeux non triviales. Un exemple de fait implicite est l'extraction (*Christopher Beck - is - prolific film composer*), faite par le système *MinIE*, présentée à la Table 1. Cette information n'est pas présente verbatim dans la phrase, mais peut facilement être déduite de celle-ci. Les notions d'inférence, de faits implicites et explicites sont explorées au chapitre 2. Nous tentons donc ici d'ajouter un cadre bien défini à la tâche permettant une rigueur et une justesse jusqu'alors manquantes.

Après cette première publication sur le sujet en 2007, différents systèmes d'OIE ont été développés. Nous présenterons en détail certains de ces systèmes à la section 3.1, principalement ceux utilisés dans nos travaux. Ces systèmes proposent généralement des performances accrues selon certaines métriques par rapport aux systèmes les précédant, justifiant leur adoption. Initialement, ces mesures étaient faites au niveau des extractions elles-mêmes, les extractions étant jugées pertinentes ou non, ce qui permettait d'établir un score de précision. Bien que certains papiers mesuraient également le rappel, cette notion cruciale n'était pas généralement utilisée.

(Stanovsky & Dagan, 2016) proposent une première référence complète, *OIE2016*, avec des annotations globales sur un ensemble de phrases, permettant de calculer à la fois la précision et le rappel à l'aide d'une fonction établissant la correspondance entre les extractions faites par les systèmes et les annotations manuelles. Cette fonction de correspondance permet de faire correspondre les faits extraits par les systèmes et ceux annotés manuellement de manière à regrouper en paires les faits qui véhiculent la même information. Cette opération est nécessaire pour mesurer la quantité d'information pertinente extraite par les systèmes ainsi que la quantité de l'ensemble théorique de l'information présente dans la référence capturée par les systèmes dans leurs extractions. Par la suite, différentes références ont été proposées : *Wire57* (Lechelle, Gotti, & Langlais, 2019), *CaRB* (Bhardwaj, Aggarwal, & Mausam, 2019) et *BenchIE* (Gashteovski et al., 2022), qui seront traitées à la section 3.2. Chacune de ces références tend à corriger certains problèmes et certains biais des références précédentes et tente de redéfinir le mieux possible la tâche de l'OIE.

1.2 Motivations

L'OIE n'est pratiquement jamais une fin en soi. En effet, les extractions des systèmes d'OIE, bien que structurant l'information comprise dans un texte, ne sont généralement pas utilisées directement. Ces extractions sont utilisées subséquemment par des systèmes traitant d'autres tâches comme le question-réponse (Fader, Zettlemoyer, & Etzioni, 2014) ou la compréhension de texte (Stanovsky, Dagan, & Mausam, 2015). Ces tâches secondaires prouvent la nécessité

d’identifier le système d’OIE le plus performant. Bien qu’en théorie il faudrait pour identifier le système le plus performant, tester tous les systèmes sur la tâche précise à accomplir, cela n’est pas réalisable en pratique, ce que nous aborderons dans la section 2.2.

Indépendamment du fait que nous voudrions potentiellement établir une référence pour l’OIE en utilisant des tâches subséquentes, la réalité est qu’un grand nombre, si ce n’est l’ensemble des méthodes d’OIE qui sont proposées s’évaluent sur des références générales, qui mesurent uniquement si les systèmes sont en mesure de répliquer ce que les annotateurs ont proposé en sortie. Par exemple, trois des systèmes d’OIE proposés en 2022, *MilIE* (Kotnis et al., 2022), *OIE@OIA* (Wang, Peng, Sun, & Li, 2022) et *CompactIE* (Fatahi Bayat, Bhutani, & Jagadish, 2022), utilisent respectivement à travers les trois publications, *OIE2016*, *WiRE57*, *CaRB* et *BenchIE*. Cette utilisation majoritaire motive l’existence de références générales. La référence état de l’art à ce jour en OIE est *BenchIE*. Cette référence innove par rapport aux références précédentes en introduisant le concept de *clusters* ou de *synset*, soit une liste de formulations pour un fait donné jugées sémantiquement équivalentes. Cependant, l’analyse de cette référence met en lumière certaines lacunes auxquelles nous avons tenté de remédier.

1.3 Contributions

Nous avons dans un premier temps développé un rapport PowerBI (environ 1500 lignes de codes, 2 pages de rapports et une vingtaine de tableaux interactifs) présenté en section 3.3.1. Cet outil d'analyse nous a permis d'identifier certains problèmes de BenchIE, à la fois en ce qui a trait à l'annotation des phrases utilisées ainsi qu'à la classification des faits contenus ou non dans la référence. Nous tentons de résoudre ces différents problèmes en proposant une nouvelle référence entière qui suit de plus près les nouvelles balises pour encadrer la tâche d'OIE que nous mettons sur pied. De plus, nous évaluons des systèmes d'OIE directement sur une tâche secondaire et montrons que les systèmes que notre référence désigne comme supérieurs performant également mieux sur ces tâches secondaires.

Nos contributions dans ce mémoire sont :

- Un rapport interactif PowerBI portant sur *BenchIE*
- Les lignes directrices du processus d'annotation des phrases de référence
- Les lignes directrices de la fonction de correspondance entre les faits extraits et les faits annotés
- La reannotation de corpus textuels de deux références selon nos lignes directrices d'annotation
- Une fonction de correspondance entre les faits extraits et les faits annotés selon nos lignes directrices de correspondance
- Une analyse des performances de notre fonction de correspondance et de la nouvelle référence dans son ensemble.

1.4 Organisation du mémoire

Le reste du mémoire est organisé comme suit :

Le chapitre 2 présente une définition plus formelle de l'évaluation de l'OIE, les défis que cela représente ainsi que certaines pistes sur les formes différentes que pourrait prendre cette évaluation.

Au chapitre 3, nous faisons un survol des différentes méthodes d'extraction ainsi que d'évaluation, et proposons une analyse en profondeur de la référence état de l'art à ce jour.

L'approche que nous proposons est décrite au chapitre 4. Nous y présentons nos lignes directrices d'annotation et de correspondance pour l'OIE, ainsi que nos annotations manuelles et décrivons la fonction de correspondance développée.

Le chapitre 5 comprend une analyse et une comparaison de la référence proposée avec les références précédemment utilisées en OIE. Cette comparaison est motivée et expliquée, ainsi que les métriques et les données utilisées pour ce faire. Une analyse de la corrélation entre les scores donnés par notre référence et ceux donnés par une évaluation sur une tâche subséquente est également présentée.

Chapitre 2 - Évaluation en OIE

Dans cette section, nous tentons de définir la tâche qu'est l'évaluation de l'OIE et motivons les choix que nous faisons en définissant un nouveau cadre pour cette évaluation. Pour ce faire, nous explorons le problème de l'inférence et de l'évaluation sur les tâches subséquentes à l'OIE.

2.1 Définition de la tâche

La grande majorité des textes au sujet de l'OIE s'accordent pour dire que les systèmes d'OIE devraient optimalement extraire une liste exhaustive de faits pour chaque phrase, signifiant qu'ils doivent capturer l'ensemble de l'information présente dans la phrase. Ceci étant à la fois hautement subjectif et déterminant pour une référence d'évaluation, nous tentons dans cette section de faire le tour de la question à savoir ce qui doit et ce qui ne doit pas se trouver dans la sortie d'un système d'OIE. Un des premiers papiers traitant directement de l'OIE est "Machine reading" (Etzioni & Cafarella., 2007), dans lequel on soutient :

"A key problem is that many of the beliefs of interest are only implied by the text in combination with a background theory. To recall Roger Schank's old example, if the text states that a person left a restaurant after a satisfactory meal, it is reasonable to infer that he is likely to have paid the bill and left a tip. Thus, inference is an integral part of text understanding."

L'article subséquent par les mêmes auteurs (Yates et al., 2007) introduit le premier système d'OIE, TextRunner. On voit dès lors apparaître une notion cruciale, l'inférence en OIE. On peut diviser les extractions, ou les faits exprimés dans une phrase en deux catégories, soit les faits explicites et les faits implicites. Les faits explicites sont directement impliqués par la phrase, sans qu'une quelconque opération logique soit nécessaire pour les déduire. Les faits implicites au contraire ne sont pas directement déclarés dans la phrase, mais peuvent être inférés par celle-ci. On fait également la distinction avec des faits qui sont potentiellement impliqués dans la phrase, mais pas nécessairement impliqués. Des exemples de ces différents types de faits sont présentés dans la Table 2. La phrase utilisée dans cette table est commune aux jeux de données de *OIE2016*, *CaRB* et *BenchIE*.

On peut observer que les faits explicites découlent directement de la phrase alors que les faits implicites nécessitent une réflexion. Le triplet (Paul Johanson - is - Monsanto's director of plant sciences) est implicite, car le verbe, *is*, n'est pas présent dans la phrase. On observe également que les faits implicites nécessaires sont obligatoirement vrais si la phrase est vraie alors que les faits implicites potentiels ne le sont pas nécessairement. C'est le cas des triplets (Monsanto's chemical spray - overcomes - these problems) et (Monsanto's chemical spray - is

However , Paul Johanson , Monsanto ’s director of plant sciences , said the company ’s chemical spray overcomes these problems and is “ gentle on the female organ . ”

Extraction explicites

(Paul Johanson - said - Monsanto’s chemical spray overcomes these problems)

(Paul Johanson - said - Monsanto’s chemical spray is gentle on the female organ)

Extractions implicites obligatoires

(Paul Johanson - is - Monsanto’s director of plant sciences)

Extractions implicites potentielles

(Monsanto’s chemical spray - overcomes - these problems)

(Monsanto’s chemical spray - is gentle on - the female organ)

TABLE 2 – Exemples d’extractions explicites et implicites sur une phrase du jeu de données original de la référence *OIE2016*

gentle on - the female organ), dont les relations sont présentes dans la phrase, mais dont la véracité n’est pas impliquée par celle de la phrase.

La notion d’inférence ne fait pas du tout consensus en OIE, particulièrement pour ce qui est de son évaluation. La première référence complète d’OIE , *OIE2016*, ne fait pas mention directe de l’inférence, mais une analyse de ses annotations prouve que les faits implicites ne sont pas ajoutés aux annotations. Les auteurs de *WiRE57* soutiennent la chose suivante par rapport à l’inférence :

“Light inference, in the form of reformulation, is helpful to make use of the information, but full-fledged inference should be processed by a dedicated program, and is not part of the Open IE task.”

Ceci soulève une nuance intéressante en ce qui a trait aux faits implicites. Nous suivons une logique semblable dans nos propres recommandations d’annotation à la section 4.1. Nous soutenons que les faits implicites qui sont le résultat d’une inférence légère et qui sont nécessairement impliqués par la phrase doivent être inclus dans une référence, alors que les faits implicites qui nécessitent une inférence lourde, ou qui sont potentiellement impliqués par la phrase ne doivent pas être inclus. Les auteurs de *BenchIE*, expliquent pour leur part dans le papier présentant leur référence qu’ils se concentrent sur les extractions explicites uniquement, et que donc les extractions implicites qui contiennent des mots ne se trouvant pas initialement dans la phrase ne

sont pas annotées. Ceci illustre bien la dichotomie entre ces deux approches. Nous pensons qu'il est nécessaire d'inclure dans une référence certaines des extractions implicites et la prochaine section fera état des motifs qui nous poussent à cette décision.

2.2 Évaluation générale

Le fait que le résultat des systèmes d'OIE n'est généralement pas utilisé directement, mais est plutôt une étape intermédiaire dans d'autres processus pousse à penser que l'évaluation de l'OIE devrait passer par ces processus secondaires. On pourrait par exemple imaginer une évaluation basée sur les performances des systèmes d'OIE appliqués au question-réponse (QA), ou l'on disposerait d'un système de QA et d'une référence de QA prenant en entrée des triplets et retournant en résultat la performance du système sur la référence dépendamment des triplets. Il suffirait alors pour désigner le système d'OIE le plus performant de sélectionner celui donnant le meilleur score sur la référence de QA. Or, plusieurs problèmes rendent cette approche en pratique difficile.

Premièrement, de tels résultats ne seraient représentatifs que des performances des systèmes d'OIE sur la tâche précise sur laquelle ils auraient été testés. Un système d'OIE très performant pour le QA ne l'est pas nécessairement pour la compréhension de texte. De plus, bien que le nombre de tâches utilisant l'OIE soit à ce jour limité, il faudrait tout de même compter une référence par tâche et dès lors qu'une nouvelle tâche apparaîtrait il faudrait construire une nouvelle référence, ce qui n'est pas pratique et ne mène à aucune généralisation. Deuxièmement, le classement donné par la référence sur la tâche subséquente serait hautement biaisé par le système utilisé pour cette tâche. Autrement dit, si nous utilisons un système de QA donné, les résultats obtenus par cette procédure ne seraient valides que pour ce système précis, n'indiquant à priori pas les performances générales des systèmes d'OIE par rapport au QA.

Ces différentes problématiques motivent l'existence d'une référence d'OIE générale, qui évalue les systèmes sur l'OIE directement, même si cela n'est souvent pas un résultat final. De plus, cela nous indique que cette référence devrait être exhaustive, car c'est là une contrainte qui permet de s'assurer que l'ensemble des extractions que pourrait nécessiter une tâche subséquente sont présentes. Cela indique également la nécessité d'annoter les faits implicites, car il est facile d'imaginer une tâche qui ferait usage de ces extractions. C'est donc cette logique qui motive nos lignes directrices d'annotation qui sont présentées aux sections 4.1 et 4.2.

De plus, comme mentionné précédemment, l'utilité de références générales d'OIE est prouvée par leur utilisation prépondérante dans le développement de nouveaux modèles. Même si cette méthodologie n'est pas totalement satisfaisante, elle est néanmoins importante, et elle

participera fort probablement à une amélioration des systèmes d'OIE.

Ces références sont formées d'une annotation manuelle par des expert ou par *crowdsourcing* d'un corpus de référence, et d'une fonction de correspondance, permettant d'établir un score pour un ensemble d'annotations et d'extractions donné. Cette fonction vise à mesurer à quel point les extractions ressemblent aux annotations et contiennent l'ensemble des informations comprises dans celles-ci.

Bien qu'il ne soit pas pertinent de créer une référence pour l'OIE utilisant les résultats des systèmes sur des tâches subséquentes, une analyse de la corrélation entre les scores donnés par une référence d'OIE et ceux donnés par une référence sur une tâche subséquente peut nous donner une idée de la justesse de cette référence. Les résultats d'une expérience suivant ce principe sont présentés à la section 5.6.

2.3 Métriques d'évaluation

Lors de l'évaluation des systèmes d'OIE, mais aussi des fonctions de correspondance elles-mêmes, il est nécessaire d'utiliser certaines métriques. Ces métriques sont présentées dans la présente section.

Soit un ensemble d'objets O , dont un certain ensemble O_p est associé à une caractéristique positive. Soit également un système qui produit un autre ensemble d'objets, O_s , tentant d'identifier tous les objets possédant la caractéristique positive. Les métriques suivantes peuvent alors être définies :

$$Precision(P) = \frac{|O_p \cap O_s|}{|O_s|}$$

$$Rappel(R) = \frac{|O_p \cap O_s|}{|O_p|}$$

$$F1 = \frac{2*(P+R)}{(P*R)}$$

Dans le cadre de l'OIE, et plus particulièrement pour évaluer les performances des systèmes, pour calculer la précision, l'ensemble d'objets correspond aux différentes extractions retournées par un système donné, et la caractéristique positive est le fait qu'une extraction corresponde à

une des extractions annotées manuellement. Pour le calcul du rappel, l'ensemble d'objets est l'ensemble des faits annotés et la caractéristique est le fait qu'une annotation corresponde à une extraction faite par le système évalué. Pour avoir une haute précision, un système doit donc faire le plus petit nombre possible d'extractions qui ne correspondent pas à des faits annotés, alors que pour avoir un rappel élevé, le système doit faire le plus d'extractions possible qui correspondent à des faits annotés.

On cherche aussi à faire une évaluation des différentes fonctions de correspondance existantes. Pour ce faire on utilise également ces différentes métriques, par contre dans ce cas, l'ensemble des objets est l'ensemble des paires (extraction système, fait annoté) possibles pour chaque phrase et la caractéristique positive est la présence de correspondance entre une paire donnée. Les différentes fonctions de correspondance sont donc évaluées sur leur capacité à répliquer cette présence de correspondance entre les paires, à la fois en précision et en rappel.

Ces deux métriques permettent de calculer le F1, aussi appelé *F-score* ou *F-mesure*, qui est la moyenne harmonique entre la précision et le rappel et qui sera la principale métrique utilisée pour calculer à la fois la qualité des extracteurs ainsi que la qualité des fonctions de correspondance développées.

Une autre métrique que nous utiliserons est la corrélation entre deux statistiques. Soient deux variables aléatoires réelles, X et Y à variance finie, la corrélation entre ces deux variables, r , est définie comme suit :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Où $Cov(X, Y)$ est la covariance entre X et Y , et σ_X et σ_Y sont les écarts types de chacune des variables.

Cette métrique sera utilisée pour valider le classement obtenu en utilisant notre référence. Nous comparerons celui-ci à un classement obtenu en utilisant une correspondance annotée manuellement. Nous montrerons que le classement obtenu par notre référence corrèle plus avec le classement manuel que celui des autres références.

Chapitre 3 - Revue de littérature

Dans ce chapitre, nous faisons une revue de littérature sur le sujet de l'OIE, en débutant par la présentation de différents systèmes proposés pour la tâche, principalement ceux utilisés pour nos expérimentations. Ces systèmes ont été sélectionnés pour représenter une variété de techniques différentes, de systèmes plus récents et moins récents, et selon la disponibilité du code permettant de les exécuter. Par la suite, nous présentons les principales références d'évaluation de l'OIE et expliquons leur fonctionnement. Nous tentons ensuite de soulever les différentes problématiques liées à ces références, notamment à l'aide d'un outil d'analyse que nous avons développé.

3.1 Systèmes d'OIE

Cette section présente l'évolution des systèmes d'OIE en ordre chronologique de publication.

3.1.1 TextRunner

TextRunner (Yates et al., 2007), développé à l'Université de Washington, est le premier système d'OIE. Il répond aux premiers besoins présentés par l'OIE, soit celui d'être automatique, au sens où les relations ne sont pas spécifiées manuellement, celui d'être efficace, comme on souhaite faire des extractions sur de grandes collections de documents, et celui d'être performant sur des corpus hétérogènes comme le web. Le fonctionnement de TextRunner est le suivant : une passe est faite sur le texte, identifiant la catégorie grammaticale des mots ainsi que les groupes nominaux. Des extractions plausibles sont ensuite construites à l'aide de ces groupes nominaux et de différentes contraintes. Un classifieur supervisé à l'aide d'exemples annotés manuellement arrive ensuite à prédire avec haute précision si une extraction plausible est valide ou non. Un regroupement non supervisé des extractions est ensuite fait pour identifier les extractions désignant le même fait, et celles-ci sont regroupées en une seule extraction. Ce système a été évalué par les auteurs en l'exécutant sur un corpus de 9 millions de documents. De ces textes, TextRunner a extrait 7.8 millions de triplets, dont 400 ont été évalués manuellement et dont 80% ont été jugés corrects. Cela illustre bien les premières méthodes d'évaluation de l'OIE et les différences avec les méthodes plus récentes.

3.1.2 ReVerb

Les auteurs de *ReVerb*, (Fader, Soderland, & Etzioni, 2011) font le constat que les méthodes existantes présentent un problème d'incohérence et un problème de non-pertinence. L'incohérence fait référence à une relation qui est mal interprétée, donc formulée de la mauvaise façon ou reliée aux mauvaises entités. La non-pertinence fait référence à une extraction qui

véhicule une information qui n'est pas exprimée dans la phrase ou qui n'est pas utile dans le cadre de l'OIE, selon eux. Ils introduisent alors *ReVerb*, tentant de résoudre ces problèmes à l'aide de différentes contraintes lexicales. Pour limiter le nombre d'erreurs d'incohérence, une contrainte syntaxique est introduite : toutes les relations de plus d'un mot qui sont extraites doivent commencer avec un verbe, terminer avec une préposition et être contenues de manière continue dans la phrase originale, entre les deux entités. L'expression régulière présentée à la Figure 1, traduite de (Fader et al., 2011), permet d'identifier ces relations qui sont cohérentes.

$$\begin{aligned}
 & V | V P | V W^* P \\
 & V = \text{verbe particule? adv?} \\
 & W = (\text{nom} | \text{adj} | \text{adv} | \text{pron} | \text{det}) \\
 & P = (\text{prep} | \text{particule} | \text{infinitif})
 \end{aligned}$$

FIGURE 1 – Expression régulière utilisée par ReVerb afin d'identifier les relations cohérentes.

Pour résoudre le problème de non-pertinence, une contrainte lexicale est utilisée : toutes les relations qui sont extraites doivent avoir été rencontrées un certain nombre de fois dans un large corpus pour être considérées comme pertinentes. Cela évite d'extraire des relations trop spécifiques. On considère ici empiriquement qu'une relation est pertinente si elle a été rencontrée plus de 20 fois dans un corpus de 500 millions de phrases. L'algorithme de *ReVerb* est donc le suivant :

1. Pour tout verbe dans la phrase, trouver la séquence de mots s qui satisfait l'expression régulière de la Figure 1
2. Trouver les deux groupes nominaux x et y , à gauche et à droite de s respectivement, qui satisfont certaines contraintes, formant l'extraction finale (x, s, y)

Une fonction de confiance entraînée à l'aide d'une régression logistique sur une collection de phrases et utilisant quelques caractéristiques simples permet d'exprimer la certitude du système par rapport à chaque extraction. Les auteurs mesurent que leur expression régulière capture environ 85% des relations exprimées dans un corpus de 300 phrases. Leur papier présente une analyse comparant ReVerb à TextRunner et WOE (Wu & Weld, 2010) et des performances améliorées par rapport à ces deux systèmes.

3.1.3 ClausIE

ClausIE, (Del Corro & Gemulla, 2013) introduit le principe de *clauses*, signifiant une partie de phrase véhiculant une pièce d'information cohérente. Ces clauses sont formées obligatoirement d'un sujet (S), d'un verbe (V) et optionnellement de différents objets directs

(O), objets indirects (O_i), adverbes (A) et compléments (C). On peut classer les clauses selon le rôle grammatical de ces composants, et ce faisant, en anglais, on obtient uniquement 7 types de clauses qui sont présentés à la Figure 2, reprise de l’article original.

Pattern	Clause type	Example	Derived clauses
Basic patterns			
S_1 : SV_i	SV	AE died.	(AE, died)
S_2 : SV_eA	SVA	AE remained in Princeton.	(AE, remained, in Princeton)
S_3 : SV_eC	SVC	AE is smart.	(AE, is, smart)
S_4 : $SV_{mt}O$	SVO	AE has won the Nobel Prize.	(AE, has won, the Nobel Prize)
S_5 : $SV_{dt}O_iO$	SVOO	RSAS gave AE the Nobel Prize.	(RSAS, gave, AE, the Nobel Prize)
S_6 : $SV_{ct}OA$	SVOA	The doorman showed AE to his office.	(The doorman, showed, AE, to his office)
S_7 : $SV_{ct}OC$	SVOC	AE declared the meeting open.	(AE, declared, the meeting, open)

FIGURE 2 – Types de clauses utilisés par ClausIE selon la Table 1 de Del Corro et Gemulla (2013)

Le type de la clause est obtenu en enlevant tous les éléments qui ne sont pas nécessaires pour que l’extraction conserve sa cohérence. Par exemple, la clause “*AE remained in Princeton*”, qui contient un sujet, un verbe et un adverbe, est de type SVA, car la clause obtenue en enlevant une quelconque partie (S, V, A), n’est pas cohérente. Les sujets, verbes, objets et compléments sont toujours essentiels alors que les adverbes peuvent être optionnels ou non. L’algorithme de *ClausIE* est divisé comme suit :

1. Calcul de l’arbre de dépendance de la phrase en utilisant l’analyseur de dépendance non lexical de l’Université Stanford
2. Détermination de l’ensemble des clauses en utilisant l’arbre de dépendance
3. Détermination de l’ensemble des clauses dérivées cohérentes et de leur type en utilisant un arbre de décision, présenté à la Figure 3
4. Génération des extractions à l’aide des clauses

La conception de *ClausIE* permet une flexibilité au niveau de la génération des extractions, laissant à l’utilisateur le choix de l’inclusion ou non de certaines clauses basées sur leur type. ClausIE a pour avantage de ne pas nécessiter d’entraînement, donc de données annotées. Les performances de *ClausIE* au moment de la publication sont meilleures que les systèmes précédents, *OLLIE* (Mausam, Schmitz, Soderland, Bart, & Etzioni, 2012) et *ReVerb*, sur deux jeux de données dont celui fourni par ReVerb, un ensemble de 200 phrases tirées de Wikipédia, ainsi que le jeu de données NYC (Sandhaus, 2008). Notons ici que ces performances ne sont

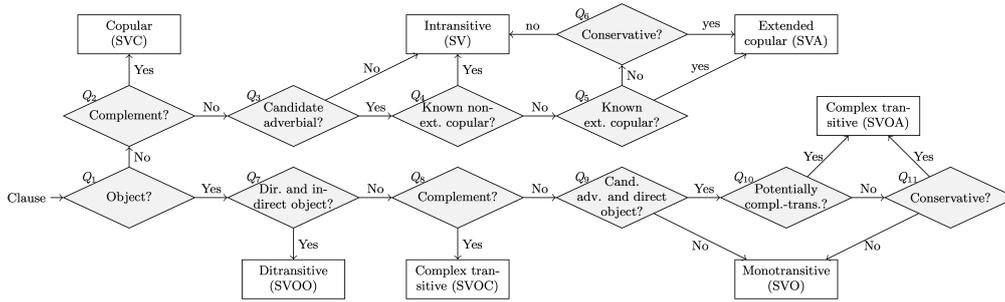


FIGURE 3 – Organigramme décisionnel pour la détection des types de verbes et types de clauses utilisé par *ClausIE* repris de la Figure 2 de Del Corro et Gemulla (2013)

mesurées qu’en précision et non pas en rappel, les auteurs de la publication soutenant qu’il est impossible d’obtenir une référence contenant l’ensemble des faits exprimés par une phrase.

3.1.4 MinIE

MinIE (Gashteovski, Gemulla, & del Corro, 2017), est un système construit par-dessus *ClausIE*. Son but premier est de produire des extractions qui soient plus compactes, ou minimales, que celles des systèmes précédents. Pour ce faire, les extractions produites par *ClausIE* sont bonifiées, principalement pour inclure plus d’extractions implicites, à l’aide des patrons identifiés par *FINET* (Del Corro, Abujabal, Gemulla, & Weikum, 2015), ainsi que d’autres patrons créés à la main. Ces différentes extractions identifiées, leurs relations sont ensuite modifiées pour les rendre plus riches sémantiquement, plus informatives, à l’aide de règles inspirées des patrons syntaxiques de *ReVerb*. Cette opération est très utile à nos yeux, transformant l’extraction (*Faust, made, a deal with the devil*) de *ClausIE*, dont la relation est peu informative, en l’extraction (*Faust, made a deal with, the devil*, beaucoup plus intéressante. En effet, les auteurs expliquent que la relation *made*, est hautement polysémique, possédant une cinquantaine de définitions différentes dans WordNet, au contraire de la relation *made a deal with*. Les extractions sont ensuite annotées pour identifier différents attributs comme leur polarité ou leur attribution à l’aide de règles créées manuellement. Finalement, les extractions sont minimisées, encore une fois selon des patrons, avec différents modes proposés (conservateur, agressif, etc.). Les auteurs évaluent leur nouveau système en mesurant sa précision, son nombre total d’extractions et la longueur moyenne de ces extractions sur un sous-ensemble de 10 000 phrases du New York Times Corpus. Ils comparent ces mesures à celles obtenues en utilisant d’autres systèmes (*ClausIE*, *OLLIE*, and *Stanford OIE*). Ils observent que *MinIE* est plus précis, produit des extractions plus compactes et produit plus d’extractions que *OLLIE*, autant que *Stanford OIE* et moins que *ClausIE*.

3.1.5 RnnOIE

RnnOIE, (Stanovsky et al., 2018) est un des premiers systèmes d'OIE basé sur de l'apprentissage supervisé. C'est le manque de données d'entraînement qui explique que les systèmes précédents n'aient pas exploité cette technique, mais qu'ils soient plutôt semi-supervisé ou basé sur des règles définies manuellement. Les auteurs ont donc présenté un jeu de données d'entraînement ainsi qu'une architecture pour leur système. Cette architecture est présentée à la figure 4 et suit l'architecture état de l'art à l'époque pour le SRL (Zhou & Xu, 2015). Cette implémentation est basée sur un bi-LSTM transducteur.

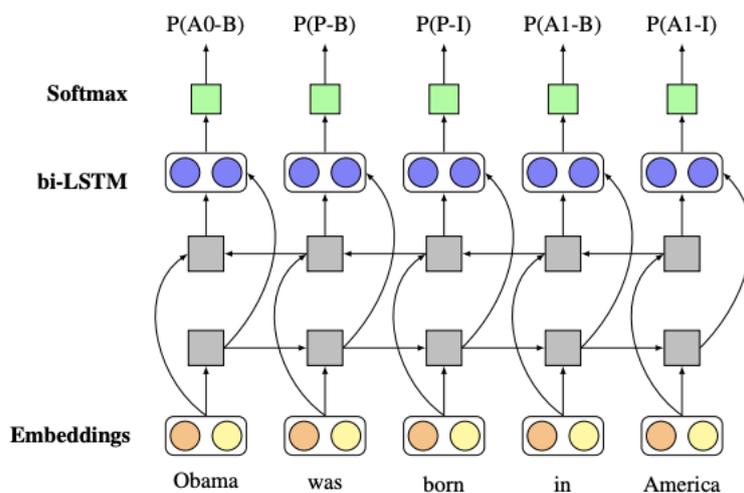


FIGURE 4 – Architecture du modèle RnnOIE repris de la Figure 2 de Stanovsky et al. (2018)

Les auteurs formulent l'OIE comme une tâche de *BIO-tagging*, ou d'étiquetage de séquence sur une séquence, la phrase, ou les arguments et la relation sont des séquences de mots continues. Le modèle est entraîné sur les données de développement de *OIE2016*. Des performances accrues sur la référence *OIE2016* proposée peu avant sont présentées dans l'article.

3.1.6 Seq2Seq

Seq2Seq, introduit par (Cui et al., 2018) est un autre des premiers systèmes d'OIE basés sur un modèle neuronal. À la différence de *RnnOIE*, il s'agit d'un modèle de séquence à séquence, prenant en entrée la phrase à analyser, la transformant en une représentation interne cachée de longueur variable et produisant en sortie la même phrase, mais découpée à l'aide de *tags* afin de produire l'extraction. L'encodeur et le décodeur du modèle séquence à séquence sont des LSTM à 3 couches. L'architecture de ce modèle utilise également les mécanismes d'attention

et de copie récemment proposés (Bahdanau, Cho, & Bengio, 2016) (Gu, Lu, Li, & Li, 2016), respectivement afin d'utiliser les informations de contexte local de la phrase et afin de limiter l'utilisation de *tokens unk* (ou *unknown*), mot inconnu). Le plan de cette architecture est présenté à la figure 5

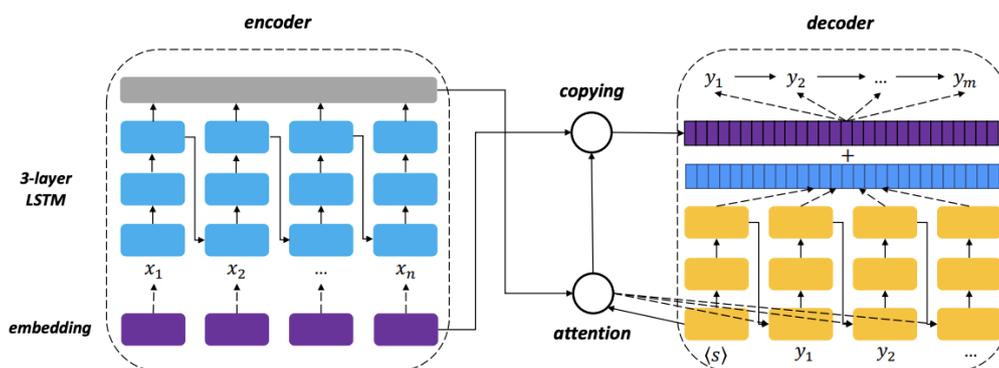


FIGURE 5 – Architecture du modèle Seq2Seq repris de la Figure 1 de Cui et al. (2018)

Le modèle est entraîné sur des données obtenues en utilisant le système d'OIE OpenIE4 (Christensen, Mausam, Soderland, & Etzioni, 2011) sur des phrases tirées de Wikipédia et en retirant les extractions dont la confiance est inférieure à 0.9. Les performances du modèle sont ensuite évaluées sur *OIE2016*, rapportant une aire sous la courbe (AUC) de 0.473, supérieure à celles de *OpenIE4*, *ClausIE*, *PropS* et *OLLIE*.

3.1.7 IMojIE

(Kolluru, Aggarwal, Rathore, Mausam, & Chakrabarti, 2020) introduit *IMojIE*, un système d'OIE neuronal visant à améliorer certaines limitations des modèles précédents. Les auteurs observent le système *Seq2Seq* et soulèvent deux problèmes principaux. Tout d'abord, *Seq2Seq* génère un nombre fixe d'extractions, peu importe la phrase d'entrée. Ensuite, le modèle produit parfois des extractions très similaires pour une phrase donnée. Les auteurs proposent comme solution d'ajouter à l'entrée du modèle de génération les précédentes extractions générées de manière récursive jusqu'à ce qu'un *tag* spécifique de fin de génération soit retourné. Une autre modification à l'architecture de *Seq2Seq* est l'utilisation de *BERT* (Devlin, Chang, Lee, & Toutanova, 2019) comme encodeur à la place d'un LSTM. Le modèle est entraîné sur des données générées en utilisant trois systèmes précédents : *OpenIE4*, *RnnOIE* et *ClausIE* sur des phrases de Wikipédia, comme *Seq2Seq*, et en sélectionnant, à l'aide d'une fonction de filtrage,

les extractions de haute qualité. *IMojIE* est évalué sur la référence nouvellement proposée, *CaRB*, et obtient un score de F1 plus élevé que les autres systèmes utilisés pour la comparaison, *Stanford-IE*, *OpenIE4*, *OpenIE5*, *ClausIE*, *PropS*, *MinIE*, *OLLIE*, *Seq2Seq* et *RnnOIE*.

3.1.8 OpenIE6

OpenIE6 (Kolluru, Adlakha, Aggarwal, Mausam, & Chakrabarti, 2020) est un système d'OIE neuronal, et représente l'évolution de *OLLIE*, par la suite devenu *OpenIE4* puis *OpenIE5*. Ce système tente de concurrencer avec les systèmes de génération de séquence, comme l'état de l'art du moment, *IMojIE*, qui nécessite un grand temps d'exécution, et les systèmes d'étiquetage, comme *RnnOIE*, qui sont légèrement moins performants. Les auteurs proposent alors un nouveau schéma, le *Iterative Grid Labelling* (IGL) ou Étiquetage de Grille Itératif. Dans ce schéma, on représente une phrase et ces extractions comme une grille où les colonnes sont les mots et les lignes, les extractions, avec la case (i, j) le rôle du j -ème mot dans la i -ème extraction. Cette grille est ensuite remplie ligne par ligne par le modèle. Des plongements *BERT* sont utilisés par le modèle ainsi que le mécanisme d'attention. Certaines contraintes syntaxiques et grammaticales sont également ajoutées à la tâche de remplissage de la grille afin d'améliorer la performance et une fonction spécifique à la gestion des conjonctions est utilisée. Le modèle est entraîné sur les mêmes données qu'*IMojIE*, c'est-à-dire des extractions de différents systèmes d'OIE précédents. *OpenIE6* est évalué sur *CaRB* et obtient un F1 de 33.7, supérieur à *IMojIE* et aux autres systèmes testés. Notons également que les auteurs soutiennent que *OpenIE6* est dix fois plus rapide pour l'inférence que *IMojIE*.

3.1.9 M2OIE

Multi2OIE, ou *M2OIE*, (Ro et al., 2020) est un système d'OIE multilingue utilisant *BERT* et qui ne nécessite pas de données d'entraînement supplémentaires pour généraliser à une nouvelle langue, uniquement le modèle *BERT* correspondant. Le système extrait les faits en identifiant tout d'abord les relations, à l'aide d'un classifieur entraîné sur les plongements obtenus par *BERT*, puis en associant chaque relation extraite à des arguments à l'aide d'un transformeur avec des blocs d'attention multilatérale. L'architecture du modèle est présentée à la figure 6.

La différence entre *M2OIE* et *RnnOIE* réside principalement dans l'utilisation de blocs d'attention plutôt que d'une simple concaténation pour les informations concernant les relations extraites. Le modèle est entraîné sur les mêmes données qu'*IMojIE* et *OpenIE6*. Ce système présente des performances plus élevées que les autres systèmes testés sur *CaRB* et *OIE2016*, ainsi que la capacité d'extraire dans plusieurs langues.

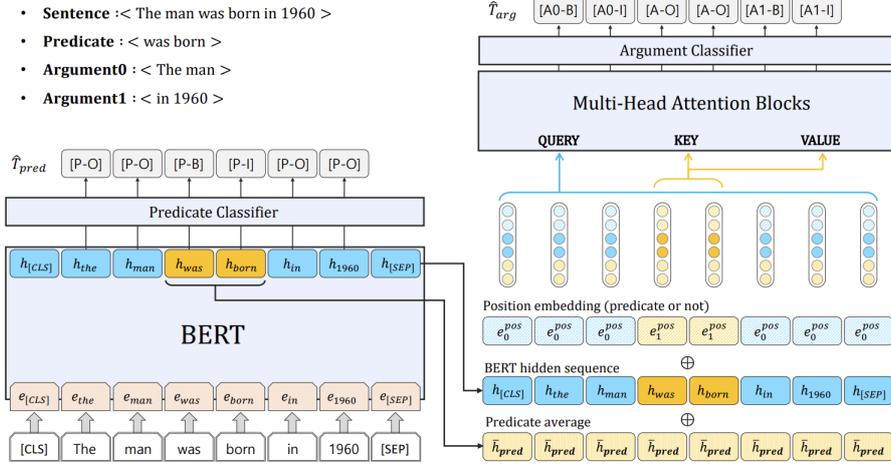


FIGURE 6 – Architecture du modèle M2OIE (tiré de la figure 2 de Ro et al. (2020))

3.1.10 CompactIE

(Fatahi Bayat et al., 2022) introduit *CompactIE*, en soulignant l’incapacité des systèmes d’OIE précédents à produire des tuples compacts, ou minimaux. Les auteurs expliquent que les différents systèmes jusqu’alors proposés produisent des extractions avec des arguments ou des relations inutilement longues ou spécifiques. L’approche privilégiée par *CompactIE* est de d’abord extraire les constituants (arguments et relation), à l’aide d’un premier modèle, puis de les lier pour former les extractions à l’aide d’un second modèle. Le modèle d’extraction est un modèle de remplissage de table, ou les types de constituants sont obtenus avec un modèle bi-affine avec attention, entraîné sur les plongements *BERT* des mots de la phrase d’origine. Cet entraînement est fait sous certaines contraintes : symétrie de la table, implication de la relation (une relation ne peut être présente si elle n’a pas d’arguments) et structure de triplets (les entités nécessaires pour compléter un triplet sont encouragées). La liaison des entités extraites est ensuite faite à l’aide d’un perceptron multicouche. Les données utilisées pour entraîner ces deux modèles sont obtenues en exécutant *IMojIE* sur les clauses extraites d’un ensemble de phrases provenant de la référence *OIE2016*. Les auteurs rapportent des performances accrues de leur système sur *WiRE57* et *CaRB*. Ils argumentent également que la moindre performance de leur système sur *BenchIE* par rapport à *MinIE* est due à la tendance de ce dernier système à produire plusieurs extractions différentes pour le même fait.

3.2 Références d’OIE

Cette section présente l’évolution des références d’OIE en ordre chronologique de publication.

3.2.1 OIE2016

(Stanovsky & Dagan, 2016) introduit la première référence complète d’OIE, *OIE2016*. Ici, complète signifie qu’une attention est portée à ce que la référence contienne, pour chacune des phrases du corpus, l’ensemble des faits devant être extraits par un système théoriquement parfait. Le besoin d’une telle référence vient du fait que les évaluations précédentes se font uniquement en précision, et pas en rappel. Les auteurs tentent tout d’abord d’identifier un consensus par rapport à ce que la sortie d’un système d’OIE idéal devrait contenir. Ils identifient trois critères principaux : l’assertivité, ou le fait que l’information soit exprimée dans la phrase, la minimalité, l’information contenue dans le triplet doit être la plus petite possible, et la complétude de la référence, mentionnée plus haut. Ces trois constats mènent les auteurs à considérer la tâche de QA-SRL (Question Answer Driven Semantic Role Labelling) (He, Lewis, & Zettlemoyer, 2015), qui est une tâche similaire au QA, mais où on présente à des annotateurs des relations et des phrases et ceux-ci doivent annoter certaines informations relatives à cette relation. Pour créer *OIE2016*, ses auteurs ont donc transformé deux références de QA-SRL présentées dans (He et al., 2015), en référence d’OIE. Pour ce faire, ils ont filtré certaines des réponses étant formées uniquement de pronoms, et certains autres critères, pour finalement obtenir une référence comprenant 3200 phrases, pour 10359 extractions.

Pour évaluer la correspondance entre les extractions des systèmes aux extractions annotées, la fonction suivante est proposée : une extraction système correspond à une extraction annotée si la tête grammaticale est la même pour la relation et pour chaque argument. Par exemple, l’annotation (*the **children** in class 5, should **watch**, less **television***) et l’extraction (*the **children**, should **watch**, more **television***), correspondrait dans ce cas car leur tête grammaticales sont respectivement *children*, *watch* et *television*. On voit directement que cette fonction peut faire correspondre des triplets qui n’expriment pas le même fait. Ceci en fait tout de même la première référence capable d’évaluer les systèmes à la fois en précision et en rappel. Les auteurs expliquent qu’ils ont validé leurs annotations en faisant une annotation manuelle de 100 phrases et en la comparant à leur annotation automatique des corpus de QA-SRL. Ils observent un F1 de 95.8 entre les deux annotations, mesurées à l’aide de leur fonction de correspondance.

3.2.2 WiRe57

WiRE57 (Lechelle et al., 2019) est une référence d’OIE composée d’annotations de qualité sur un relativement petit corpus de texte. Cette référence, qui fait suite à *OIE2016*, vise à corriger certaines des lacunes de cette dernière. Tout d’abord, il est exprimé que la complétude théorique d’*OIE2016* n’est pas avérée et que cette référence n’est pas exhaustive, ne permettant pas réellement de mesurer le rappel des systèmes. De plus, certains mots,

ajoutés aux annotations mais ne faisant pas partie du corpus original, ont été ajoutés dans le processus de transformation de la référence, et pour les auteurs de *WiRE57*, cela pose parfois problème, résultant en des triplets non pertinents. Finalement, les auteurs critiquent également la fonction de correspondance utilisée, expliquant qu’il est possible d’obtenir un score très élevé à l’aide d’un extracteur copiant uniquement les mots de la phrase avec un découpage relation-argument différent. Ils proposent plutôt une fonction de correspondance au niveau des mots eux-mêmes, pénalisant ainsi les extractions inutilement longues. La précision ainsi obtenue est le nombre de mots extraits qui sont présents dans la référence, le rappel est donc le nombre de mots annotés présents dans l’extraction.

Les auteurs proposent également des lignes directrices d’annotation, préconisant des annotations informatives, minimales, exhaustives et traitant l’inférence. Ils distribuent ces lignes directrices expliquées à l’aide d’exemples dans un document complet dont nous nous sommes inspirés pour nos propres lignes directrices. Ils annotent ensuite un corpus de 5 documents issus de différentes sources (Wikipedia, Reuters), pour un total de 57 phrases et 343 extractions. Leurs annotations présentent un fort accord interannotateurs, et leur référence permet de tirer des conclusions différentes de celles obtenues à l’aide d’*OIE2016*. Ils soulignent que le système le plus performant sur leur référence est *MinIE*, suivi par *ClausIE*, des conclusions proches de celles que nous formulons à l’aide de notre référence.

3.2.3 CaRB

CaRB est la première référence d’OIE résultant du *crowdsourcing*. Introduit par (Bhardwaj et al., 2019), il est le résultat d’annotations anonymes sur la plateforme Amazon Mechanical Turk (MTurk). Ces annotations sont faites sur le même ensemble de données que celui utilisé pour *OIE2016*. Les auteurs de la référence critiquent initialement *OIE2016* en utilisant les mêmes arguments que ceux soulevés par (Lechelle et al., 2019), soit la non-exhaustivité et le bruit présent dans la référence, ainsi que les problèmes quant à sa fonction de correspondance. Ils critiquent également *WiRE57* en raison de la petite taille de son corpus, expliquant que celle-ci ne peut représenter un ensemble de test exhaustif et objectif. Ils soutiennent cependant que la fonction de correspondance utilisée fait plusieurs améliorations par rapport à celles précédemment proposées et s’en inspirent pour leur propre référence.

La fonction développée est la suivante : Une grille est construite avec pour colonnes les extractions faites par le système évalué et pour lignes les annotations gold, le tout pour une phrase donnée. La précision et le rappel sont ensuite calculés pour chaque paire de la même façon que pour *WiRE57*, c’est-à-dire au niveau des mots. Le rappel global pour la phrase est ensuite calculé comme la moyenne des rappels maximaux sur chaque ligne. Cela permet de faire

correspondre les annotations aux extractions s’en rapprochant le plus. La précision est ensuite le score maximal sur les colonnes en ordre décroissant puis moyenné. Ce schéma de calcul permet de ne pas pénaliser les extracteurs qui produisent de longues extractions ou plusieurs informations sont contenues dans un seul tuple. Les auteurs valident le score obtenu par leur référence par rapport à celui obtenu avec *OIE2016* en prenant deux systèmes qui obtiennent un rang inverse selon les deux références et en demandant à des annotateurs anonymes de déterminer, en prenant des paires d’extractions issues des deux systèmes, lesquelles sont les meilleures ou si elles sont équivalentes. Cette méthodologie nous semble erronée, car la qualité des extractions semble être un critère subjectif et qui de plus n’est pas l’objectif de l’OIE. Cette mesure ne vérifie en aucun cas le rappel ou la précision des systèmes, ce qu’elle est censée valider. Malgré cela, *CaRB* devient une référence d’OIE largement utilisée dans la communauté, de nombreux auteurs de systèmes d’OIE rapportent leur performance sur ce jeu de données.

3.2.4 BenchIE

BenchIE (Gashteovski et al., 2022) est une nouvelle référence d’OIE qui tente de revenir vers une annotation manuelle extensive des phrases plutôt que de s’appuyer sur le *crowdsourcing*. Les auteurs soutiennent tout d’abord que les fonctions de correspondance basées sur un compte de mot précédemment utilisées par *WiRE57* et *CaRB* ne parviennent pas à s’assurer de la véracité des faits extraits par les systèmes. Un exemple d’une telle extraction est présenté à la table 3.

<i>Sen. Mitchell is confident he has sufficient votes to block such a measure with procedural actions</i>	
Annotation	Extraction candidate
Sen. Mitchell is confident he has sufficient votes ... with procedural actions	Sen. Mitchell, is confident he has, procedural actions

TABLE 3 – Exemple d’extraction acceptée dans *CaRB* adapté de la Table 1 de Gashteovski et al. (2022)

Dans cet exemple, l’extraction est fautive, n’est pas exprimée dans la phrase, mais obtiendrait dans le contexte de *CaRB*, une précision parfaite, car tous les mots de l’extraction se trouvent à la même position dans une annotation de la référence, et un rappel élevé, car tous les mots du premier argument, de la relation et deux des mots du deuxième argument de la référence se trouvent aux mêmes positions dans l’extraction système. Les auteurs proposent donc une approche centrée sur les faits, et non le compte des mots. Ils introduisent le concept de *synset* factuel, ou *cluster*. Ceux-ci sont utilisés lors de l’annotation d’un fait, afin de représenter plusieurs (ou toutes) les formulations possibles de ce fait à partir de la phrase dont il provient. Des exemples d’annotations utilisant les *synsets* sont présentés à la figure 7. Les annotations

de *BenchIE* utilisent les crochets ([]) pour dénoter des mots optionnels, ils peuvent être inclus dans les extractions des systèmes ou non, sans influence sur les résultats.

Input sentence: "Sen. Mitchell is confident he has sufficient votes to block such a measure with procedural actions."			
f_1	("Sen. Mitchell" \ "he";	"is";	"confident [he has sufficient ... actions]")
f_2	("Sen. Mitchell" \ "he"; ("Sen. Mitchell" \ "he";	"is confident he has"; "is confident he has";	"sufficient votes") "suff. votes to block [such] [a] measure")
f_3	("Sen. Mitchell" \ "he"; ("Sen. Mitchell" \ "he"; ("Sen. Mitchell" \ "he";	"is confident he has sufficient votes to block" "is confident he has ... to block [such]"; "is confident he has ... to block [such] [a]";	"[such] [a] measure") "[a] measure") "measure")
f_4	("Sen. Mitchell" \ "he"; ("Sen. Mitchell" \ "he";	"is confident he has ... [such] [a] measure with"; "is confident he has ... [such] [a] measure";	"procedural actions") "with procedural actions")

FIGURE 7 – Exemple de *synsets* utilisés par *BenchIE* selon la Table 2 de Gashteovski et al. (2022)

Pour remédier au problème de véracité, les auteurs proposent de s'appuyer sur une correspondance exacte entre les faits annotés et les faits extraits : seules les extractions donc chaque composante (arguments et relation) est exactement équivalente à celle d'une formulation d'un cluster seront considérés comme acceptées. Cette fonction de correspondance est très stricte, mais les auteurs la motive en expliquant avoir listé dans les clusters l'ensemble des formulations possibles d'un même fait, et ce pour chacun des faits de leur référence. Pour pouvoir comparer *BenchIE* aux références précédentes, les auteurs utilisent 300 des 1200 phrases de *CaRB* pour faire leurs annotations. Le processus d'annotation a été mené par deux annotateurs qui ont indépendamment annoté l'ensemble des faits à l'aide de la plateforme *AnnIE*, en s'assurant d'inclure les faits valides extraits par un ensemble prédéterminé de systèmes ainsi que ceux initialement annotés dans *CaRB*. Ils ont ensuite quantifié l'accord interannotateur (IAA), obtenant un score de 0.79 sur le rappel au niveau des clusters, c'est à dire la proportion moyenne des annotations des annotateurs qui sont couvertes par celles d'un autre annotateur. Les auteurs proposent également différentes versions de *BenchIE* dirigée vers les différentes tâches subséquentes à l'OIE, soit *BenchIE-E* pour les tâches reliées à la construction de bases de connaissances, dont les extractions qui contenaient un argument ne désignant pas une entité entière ont été retirées (c'est à dire les formulations dont un des arguments contient des mots appartenant réellement à la relation), *BenchIE-C*, une version à plat de la référence (dont tous les arguments et la relation ont été concaténés), utile aux tâches comme la simplification ou le résumé (Ponza, Del Corro, & Weikum, 2018), et autres. L'étude des résultats obtenus par différents systèmes sur leur référence permet aux auteurs de conclure que celle-ci mesure mieux les performances réelles des systèmes comparativement à *CaRB* qui surévalue légèrement la

précision et surévalue drastiquement le rappel des systèmes à cause de sa fonction de correspondance trop laxiste.

3.3 Limitations des références actuelles

Une observation préliminaire de la référence *BenchIE* nous a laissé croire qu'elle contenait certaines limitations, principalement dans son processus d'annotation et de correspondance. Pour vérifier cela, nous avons développé un outil d'analyse sous la forme d'un rapport interactif. Nous présentons cet outil et les résultats de l'analyse en découplant aux prochaines sections.

3.3.1 Outil Power BI

Power BI est une plateforme développée par Microsoft, principalement dédiée à l'intelligence d'affaires. Cette plateforme permet entre autres de créer des rapports interactifs de format .pbix et de les partager à l'aide du service Power BI Report Server. Nous avons utilisé cet outil pour développer un rapport permettant d'analyser les différentes facettes de la référence *BenchIE*. Ce rapport permet d'obtenir un grand nombre d'informations et de caractéristiques sur les différentes annotations manuelles et extractions des systèmes sur les 300 phrases de cette référence. Il permet également de trier ou de sélectionner un sous-ensemble de phrases selon ces différentes caractéristiques. La longueur des phrases, la longueur des annotations, la lisibilité ou encore la profondeur de l'arbre de dépendance de la phrase sont parmi ces caractéristiques pouvant être utilisées. Des visualisations sont également présentées permettant d'évaluer en un coup d'œil comment se comportent certaines caractéristiques sur des sous-ensembles de phrases. On peut également visualiser quels systèmes performant mieux selon certaines caractéristiques et voir quelles extractions sont associées à quelles annotations. Au moment du développement de cette application, nous n'utilisons que quatre systèmes d'OIE pour expérimenter, soit *ReVerb*, *IMojIE*, *OpenIE6* et *CompactIE*. Deux captures d'écran de cet outil sont présentées aux figures 8 et 9. Dans la figure 8, la partie du haut montre différents filtres permettant de sélectionner les phrases selon leur type d'erreur d'annotation, la partie à gauche montre un aperçu de différentes caractéristiques des phrases sélectionnées, le tableau central montre les phrases, les extractions des systèmes sur celles-ci et leur scores, et la partie du bas montre les distributions de certaines caractéristiques comme la longueur des phrases. La figure 9, montre pour sa part une vue plus détaillée du tableau central de la figure 8, où on observe les phrases de la référence, les annotations, ainsi que les extractions et les correspondances entre ces dernières. Cette page nous a été utile pour évaluer la performance de la fonction de correspondance de *BenchIE*, nous permettant d'identifier les correspondances valides et celles manquantes. Une licence est requise pour partager notre rapport en ligne, il est donc seulement disponible en version locale.

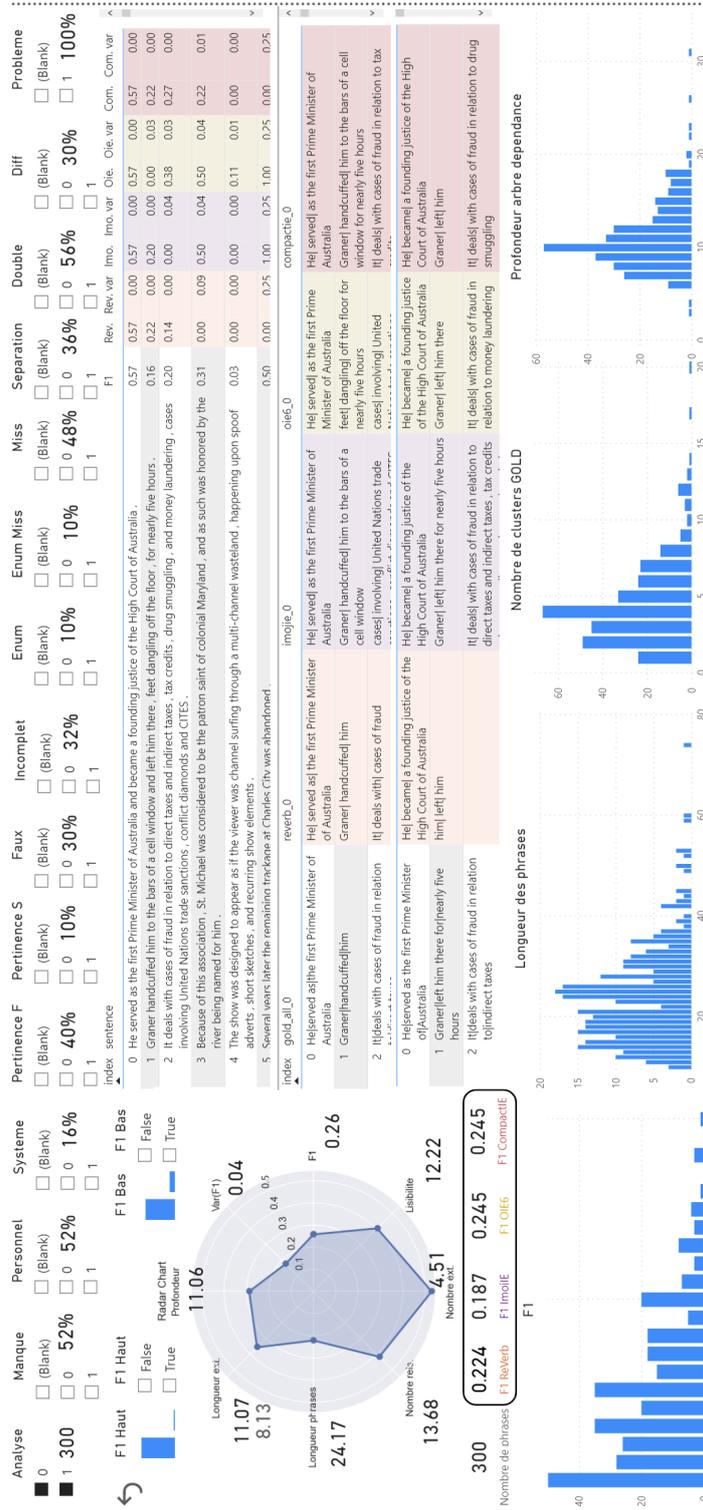


FIGURE 8 – 1^o capture d'écran de notre plateforme Power BI portant sur *BenchIE*

TP: Nombre de clusters couverts par au moins une extraction FN: Nombre de clusters non couverts FP: Nombre d'extractions n'ayant aucun match Fonction de match: Match exact par 'dot' ou 'lexical' (default: slot)

index	sentence	F1	Clusters	Rev.	RN	RP	RPN	RP	IN	IFN	TP	Olé	ON	OTF	ORF	Com.	CN	CTF	CFN	CFP			
0	He served as the first Prime Minister of Australia and became a founding justice of the High Court of Australia.	0.57	5	0.57	2	2	3	0	0.57	2	3	0	0.57	2	2	3	0	0.57	2	2	3	0	
1	Graver handcuffed him to the bars of a cell window and left him there, feet dangling off the floor, for nearly five hours.	0.16	7	0.22	2	1	6	1	0.20	3	1	6	2	0.00	3	0	7	3	0.22	2	1	6	1
2	It deals with cases of fraud in relation to direct taxes and indirect taxes, tax credits, drug smuggling, and money laundering, cases involving United Nations trade sanctions, conflict diamonds and CTES.	0.20	13	0.14	1	1	12	0	0.00	2	0	13	2	0.38	8	4	9	4	0.27	6	2	11	0
3	Because of this association, St. Michael was considered to be the patron saint of colonial Maryland, and as such was honored by the river being named for him.	0.31	5	0.00	0	0	5	0	0.50	3	2	3	1	0.50	3	2	3	1	0.22	4	1	4	3
4	The show was designed to appear as if the viewer was channel surfing through a multi-channel wasteland, happening upon spoof adverts, short sketches, and recurring show elements.	0.03	8	0.00	1	0	8	1	0.00	2	0	8	2	0.11	11	1	7	10	0.00	2	0	8	2
5	Several years later the remaining wreckage at Charles City was abandoned.	0.50	1	0.00	0	0	1	0	1.00	1	0	1	0	1.00	1	0	0	0.00	0	0	0	1	0
6	Dr. Jagan himself was personally involved in the organization of the strike, and helped to raise funds across the country to it.	0.50	5	0.57	2	2	3	0	0.57	2	2	3	0	0.57	3	2	3	0	0.29	2	1	4	1
7	In 1918, the municipality was renamed Jones in honor of American congressman William Jones, who authored the Philippine Autonomy Act of 1916 that provided for greater autonomy for the Philippines under American colonial rule.	0.26	7	0.22	2	1	6	1	0.40	3	2	5	1	0.40	3	2	5	1	0.00	2	0	7	2

index	gold_all_0	R	I	O	C	gold_all_1	R	I	O	C	gold_all_2	R	I	O
0	He served as the first Prime Minister of Australia	[0]	[0]	[0]	[0]	He served as the first Prime Minister of Australia	[1]	[0]	[1]	[1]	He became a founding justice	[1]	[1]	[1]
1	Graver handcuffed him	[0]				Graver handcuffed him	[0]				Graver handcuffed him to the bars of a cell window			
2	It deals with cases of fraud in relation to direct taxes					It deals with cases of fraud in relation to indirect taxes					It deals with cases of fraud in relation to tax credits			
3	St. Michael was considered to be the patron saint of colonial Maryland	[1]	[0]	[0]		St. Michael was considered to be the patron saint of colonial Maryland	[1]	[0]	[0]		St. Michael was honored by the river being named for him			
4	The viewer was happening upon spoof adverts					The viewer was happening upon short sketches					The viewer was happening upon recurring show elements			
5	The remaining wreckage at Charles City was abandoned several years later	[0]	[0]											
6	Dr. Jagan himself was personally involved in the organization of the strike	[0]	[1]	[0]	[0]	Dr. Jagan himself helped to raise funds to it	[1]	[0]	[1, 2]		Dr. Jagan himself helped to raise funds across the country			
7	the municipality was renamed Jones in 1918					the municipality was renamed Jones					the municipality was renamed Jones in honor of American congressman William Jones			
8	the laser cuts a portion of the prostate	[0]	[0]	[0]	[0]	a portion of the prostate is cut into smaller pieces					a portion of the prostate is flushed with irrigation fluid			
9	Komororan was the only Axis ship	[0]				Komororan was the only Axis ship to conduct attacks in Australian waters during 1941					Komororan was the only Axis ship to conduct attacks in Australian waters during 1941			

index	reverb_0	0	reverb_1	1	reverb_2	2	reverb_3
0	He served as the first Prime Minister of Australia	0.0	He became a founding justice of the High Court of Australia	2.0			
1	Graver handcuffed him	0.0	him left him	8.0			
2	It deals with cases of fraud						
3	the viewer was channel surfing						
4	Dr. Jagan was personally involved in the organization of the strike	0.0	Dr. Jagan helped to raise funds	1.0			
5	the municipality was renamed Jones	1.0	honored of American congressman William Jones authored the Philippine Autonomy Act of 1916 that				

index	image_0	0	image_1	1	image_2	2	image_3
0	He served as the first Prime Minister of Australia	0.0	He became a founding justice of the High Court of Australia	2.0			
1	Graver handcuffed him to the bars of a cell window		Graver left him there for nearly five hours				
2	cases involving United Nations trade sanctions, conflict diamonds and CTES		It deals with cases of fraud in relation to direct taxes and indirect taxes, tax credits, drug smuggling, and money laundering				3.0
3	the river being named for him	3.0	St. Michael to be the patron saint of colonial Maryland	0.0	St. Michael was considered to be the patron saint Because of this association		

FIGURE 9 – 2^e capture d'écran de notre plateforme Power BI portant sur *BenchIE*

Cet outil nous a permis d'échanger et de faire des observations plus facilement sur la référence, nous menant à identifier certaines lacunes de *BenchIE* dans les annotations et dans la correspondance. À la suite de ces observations, nous avons sélectionné un sous-ensemble de phrases de la référence et avons tenté de quantifier la gravité de ces lacunes en mesurant sur quelle proportion de phrases elles étaient présentes. La méthodologie et les résultats de cette expérience sont présentés à la section suivante.

3.3.2 Analyse de BenchIE

Les problèmes principaux que nous avons soulevés par rapport à BenchIE se trouvent dans l'annotation des phrases et dans la fonction de correspondance utilisée. Nous remarquons que certains faits extraits par les systèmes sont selon nous équivalents à ceux de la référence, mais qu'ils ne sont pas acceptés comme des faits valides. Nous observons également plusieurs faits que nous estimons être impliqués directement ou implicitement par la phrase qui ne sont pas inclus dans les annotations. Nous avons donc catégorisé plusieurs sortes d'erreurs potentielles que voici. Pour chacun des types d'erreur, deux exemples sont présentés, les phrases dont ils sont tirés sont en italique et les tuples manquant ou erronés sont entre parenthèses. Plusieurs formulations d'un même cluster sont séparées d'un saut de ligne, si des tuples proviennent de différents clusters, cette information est présentée.

1. Faits manquants : Certains faits manquent aux annotations (soit par observation ou par l'examen des extractions des systèmes). Le premier exemple de la Table 4, montre un fait qui est exprimé dans la phrase de manière implicite. Dans ce cas, la relation peut être exprimée de manière très simple par le verbe *is*, qui n'est pas présent dans la phrase, ou par une relation plus complexe qui définit mieux la relation entre les arguments. De plus, plus d'un système sur les sept systèmes utilisés dans nos expériences fait une extraction correspondante à celles omises dans les annotations de *BenchIE*. Dans le second exemple, il s'agit d'un fait implicite manquant. L'information par rapport à l'effet qu'a le rapprochement des particules, qui est le point principal véhiculé dans la phrase n'est présente dans aucun des faits annotés dans *BenchIE*.

<i>My Classical Way ” was released on 21 September 2010 on Marc ’s own label , Frazzy Frog Music</i>
(Marc’s own label - is - Frazzy Frog Music)
<i>For example , when two such hydrophobic particles come very close , the clathrate-like baskets surrounding them merge</i>
(when two such hydrophobic particles - come very close - [the] clathrate-like baskets surrounding them merge) ([the] clathrate-like baskets surrounding them - merge when - two such hydrophobic particles come very close)

TABLE 4 – Exemples de faits manquants de *BenchIE*

2. Non-pertinence : Certains faits dans les annotations ne sont pas pertinents, par manque de contexte ou autre. Dans le premier exemple de la table 5, l’annotation n’est pas pertinente car elle n’exprime pas une information présente dans la phrase, bien qu’elle soit formée de mots de celle-ci. Dans le deuxième exemple, il manque de contexte à l’annotation, soit l’attribution de l’information (*He accused Dow Jones of ...*).

<i>For example , when two such hydrophobic particles come very close , the clathrate-like baskets surrounding them merge .</i>
(two [such] hydrophobic particles - come - [very] close)
<i>He accused Dow Jones of “ using unfair means to obtain the stock at an unfair price . ”</i>
(Dow Jones - using - unfair means)

TABLE 5 – Exemples de faits non pertinents de *BenchIE*

3. Faits erronés : Certains faits dans les annotations sont erronés, faux ou pas nécessairement impliqués par la phrase. Le premier exemple de la table 6 est un cas d’information fausse. Le laser ne fait pas l’ablation du tissu ; la phrase exprime plutôt qu’au lieu de faire cela, il fait autre chose. Il est donc difficile de comprendre pourquoi cette annotation est faite. Dans le deuxième exemple, bien que l’information soit effectivement exprimée dans la phrase, il manque l’attribution de l’information (*He further stated that ...*). Le manque d’attribution est dans certains cas à la fois noté comme une erreur de non pertinence et un fait erroné comme il est parfois difficile de distinguer la différence entre les deux.

<i>Instead of ablating the tissue , the laser cuts a portion of the prostate , which is then cut into smaller pieces and flushed with irrigation fluid .</i>
([the] laser - ablating - the tissue)
<i>He further stated that Hauptmann looked different and that “ John ” was actually dead because he had been murdered by his confederates .</i>
(John - was - [actually] dead)

TABLE 6 – Exemples de faits erronés de *BenchIE*

4. Faits incomplets : Certains faits dans les annotations ne sont plus pertinents ou valides sans les mots optionnels. Dans le premier exemple de la Table 7, sans les mots optionnels, l’information est fausse. La prostate n’est pas coupée en pièce, uniquement une portion. Dans le deuxième exemple, sans les mots optionnels l’information n’est pas nécessairement fausse mais est différente de la version incluant les mots optionnels et n’est plus aussi pertinente.

<i>Instead of ablating the tissue , the laser cuts a portion of the prostate , which is then cut into smaller pieces and flushed with irrigation fluid .</i>
([a] [portion] [of] [the] prostate - is cut into - [smaller] pieces)
<i>They held the first Triangle workshop in 1982 for thirty sculptors and painters from the US , the UK and Canada at Pine Plains , New York .</i>
(They - held [the] [first] Triangle workshop in - 1982)

TABLE 7 – Exemples de faits incomplets de *BenchIE*

5. Erreurs de correspondance : un système fait une extraction équivalente aux annotations, mais qui n’a pas été évaluée comme correspondante. Dans les deux exemples de la table 8, le manque de correspondance peut s’expliquer soit par un manque de souplesse dans la fonction de correspondance ou par un manque de rigueur dans les annotations. Dans le premier exemple, c’est le déterminant *the* qui pose problème, il est présent dans l’extraction candidate mais n’a pas été inclut comme optionnel dans l’annotation. Une fonction de correspondance plus souple pourrait accepter l’extraction en ignorant ce genre de mot par exemple, ou le mot aurait pu être ajouté aux annotations de manière optionnelle. Dans le deuxième exemple, la même chose est vraie du point après *Cos*, présent dans l’annotation mais pas dans l’extraction.

Annotation	Extraction
<i>For example , when two such hydrophobic particles come very close , the clathrate-like baskets surrounding them merge .</i>	
(clathrate-like baskets - surrounding - them)	(the clathrate-like baskets - surrounding - them)
<i>QVC Network Inc. said it completed its acquisition of CVN Cos. for about \$ 423 million .</i>	
(it - completed - [its] acquisition [of CVN Cos.] for [about] \$ 423 million)	(it - completed - its acquisition of CVN Cos for about \$ 423 million)

TABLE 8 – Exemples d’erreurs de correspondance de *BenchIE*

6. Doubles annotations : Les annotations contiennent deux faits différents qui expriment la même information. Dans les deux exemples de la Table 9, les deux clusters présentent des formulations qui expriment la même information. Dans le premier exemple, sans les mots optionnels les deux extractions sont en tout point identiques. Dans le second cas le découpage entre la relation et le second argument est différent mais il s’agit tout de même de la même information.

<i>They held the first Triangle workshop in 1982 for thirty sculptors and painters from the US , the UK and Canada at Pine Plains , New York .</i>	
Cluster 10	
(They - held [the] [first] Triangle workshop at - [Pine Plains] New York)	
Cluster 12	
(They - held [the] [first] Triangle workshop at - New York)	
<i>The latter was lifted only as the b-side of “ Keep on Loving Me ”</i>	
Cluster 1	
([The] latter - [was] lifted - [only] as [the] b-side of Keep on Loving Me)	
Cluster 2	
([The] latter - [was] lifted [only] as [the] b-side of - Keep on Loving Me)	

TABLE 9 – Exemples de doubles annotations de *BenchIE*

7. Sens différents : Certains faits dans les annotations contiennent deux formulations qui n’ont pas la même signification. Dans les deux exemples de la Table 10, deux formulations différentes du même cluster ne signifient pas la même chose. Dans le premier cas l’annotation inclus soit une information d’endroit ou de temps, alors que dans le deuxième

cas on propose d’inclure un différent exemple de *hobbie* sans se soucier que chacun des exemples ne soient inclus dans un fait distinct.

<i>In 1717 Lady Mary Wortley Montagu observed the practice in Istanbul and attempted to popularize it in Britain , but encountered considerable resistance .</i>
Cluster 1 ([Lady] Mary Wortley Montagu - observed - [the] practice in Istanbul) ([Lady] Mary Wortley Montagu - observed - [the] practice In 1717)
<i>Hofmann was a below-average high school student , but he had many hobbies including magic , electronics , chemistry , and stamp and coin collecting .</i>
Cluster 2 (Hofmann - had - [many] hobbies) (Hofmann - had - [many] hobbies including coin collecting) (Hofmann - had - [many] hobbies including stamp collecting) (Hofmann - had - [many] hobbies including chemistry) (Hofmann - had - [many] hobbies including electronics) (Hofmann - had - [many] hobbies including magic)

TABLE 10 – Exemples de sens différents de *BenchIE*

Nous avons donc sélectionné au hasard 50 phrases de *BenchIE* et avons évalué pour chacune d’elles, en utilisant le document d’annotation fourni ainsi que les extractions des systèmes, la présence ou non des erreurs ci-dessus. La proportion de la présence de ces erreurs sur les 50 phrases évaluées est présentée à la Table 11.

Type d’erreur	Pourcentage d’occurrence	Compte
Fait manquant	52%	26
Non-pertinence	40%	20
Fait erroné	30%	15
Fait incomplet	32%	16
Énumération	10%	5
Erreur de correspondance	48%	24
Double annotation	56%	28
Sens différent	30%	15

TABLE 11 – Proportion des différents types d’erreurs sur le sous-ensemble de 50 phrases de *BenchIE*

On peut tout d’abord observer que plus de la moitié des phrases possèdent un ou plusieurs faits à nos yeux manquants. Dans la plupart des cas, il s’agit là de faits inférés (*BenchIE* n’inclut pas ce type de faits dans sa référence). Un exemple de ce type d’erreur est présenté à la Table 4. Malgré cela, certains faits ne sont pas des faits inférés et sont tout simplement manquants, comme le deuxième exemple de la Table 4. On observe également une haute proportion de 30% et 33% respectivement de faits erronés ou incomplets. Finalement on observe qu’environ la moitié des phrases présentent une erreur de correspondance. Des exemples de ce type d’erreur sont présentés à la Table 8. Cela nous indique que la correspondance exacte et le système de *clusters* utilisés dans *BenchIE* ne sont pas suffisamment fiables pour capturer une grande majorité des correspondances entre les faits extraits et les faits annotés. Les exemples d’erreur de correspondance à la table 8 montrent qu’il est soit nécessaire d’annoter plus de formulations d’un même fait (nous estimons qu’il est en pratique très compliqué, voir impossible de lister toutes les formulations valides d’un fait), soit d’utiliser une fonction de correspondance plus souple. Ces très hautes proportions d’erreurs motivent à nos yeux une réannotation des phrases de référence et la création de nouvelles lignes directrices d’annotation permettant de grandement réduire ces proportions d’erreurs. En effet, de telles erreurs introduisent le risque de biaiser les résultats en faveur de certains types de systèmes. Si par exemple une grande majorité des faits qui présentent une erreur de correspondance sont des faits qui contiennent une reformulation de la phrase d’origine, il est clair que les systèmes qui ne font pas de reformulation seront avantagés, comme ces faits même s’ils sont bien extraits par les systèmes qui sont en mesure de le faire ne sont pas acceptés comme valides. C’est donc cette analyse des lacunes de *BenchIE* qui motive notre nouvelle approche, proposée au chapitre 4. Autres faits notables, des 50 phrases évaluées, seulement deux ne présentaient aucune erreur. De plus, les phrases présentaient en moyenne 2.9 types d’erreurs différents. Par exemple, la phrase du premier exemple de la table 9 comprend à la fois un fait manquant, un fait incomplet, une erreur de correspondance, une double annotation et une erreur de sens différent.

Il est à noter que l’évaluation de ces erreurs a été menée avant la conclusion du développement des lignes directrices d’annotation et de correspondance. Il est donc possible que ce que nous ayons considéré comme une erreur de correspondance ou un fait manquant ait changé entre l’annotation de ces erreurs et la conception de notre propre référence. Cela étant dit, la proportion élevée de ces erreurs indique tout de même des lacunes dans l’annotation de la référence et dans la fonction de correspondance utilisée.

Chapitre 4 - Approche proposée

Dans ce chapitre, nous présentons les différentes contributions de ce mémoire. Nous expliquons les choix motivant nos nouvelles lignes directrices pour l'évaluation de l'OIE et démontrons les différentes composantes de notre référence, ainsi que les étapes ayant mené au produit final. Afin de développer ces ressources, nous avons annoté à la fois l'ensemble des faits exprimés selon nos lignes directrices d'annotation dans deux corpus différents, ainsi que la correspondance entre ces faits et un ensemble d'extractions par différents systèmes d'OIE selon nos lignes directrices de correspondance. C'est cette annotation des correspondances qui nous a permis d'entraîner et de tester différentes fonctions de correspondance. Ces quatre ressources représentent un effort important d'annotation et sont disponibles à des fins de validation et afin d'aider dans le développement de futures références.

4.1 Lignes directrices d'annotation

Au vu des différentes lacunes d'annotation soulevées par rapport à *BenchIE* au chapitre précédent, nous présentons ici de nouvelles lignes directrices tentant d'encadrer l'annotation de notre propre référence ainsi que de futures références dans le domaine de l'OIE. Ces lignes directrices se veulent complètes, permettant de laisser le moins de liberté possible dans l'interprétation de l'annotation. Nous présentons ici les grands principes de nos recommandations, le document Lignes directrices d'annotation en annexe A présente les différents principes et les illustre à l'aide d'exemples. Certains de ces exemples sont également présentés dans les tables suivantes. Les annotations en vert sont des exemples positifs (on souhaite que ces tuples soient extraits des phrases), alors que les annotations en rouge sont des exemples négatifs (nous ne souhaitons pas qu'ils soient inclus). Les exemples, ainsi que le guide d'annotation lui-même, sont fortement inspirés du guide d'annotation de (Lechelle et al., 2019). À noter également que nous ne listons pas l'ensemble des extractions à faire pour une phrase donnée, mais uniquement celles permettant d'illustrer le principe dont il est sujet.

Pour ces lignes directrices d'annotation et les lignes directrices de correspondance à la section suivante, les exemples sont tirés des jeux de données de *WiRE57* et *BenchIE*. Suivant la pratique des références précédentes, nous utilisons des mots optionnels, mis entre crochets dans les annotations, qui peuvent, mais ne doivent pas être inclus par les systèmes dans leurs extractions. Ceux-ci ne sont pas toujours inclus dans les Tables à des fins de lisibilité.

Nous décidons tout d’abord de présenter toutes les informations à l’aide de tuples contenant deux arguments et une relation ou un argument et une relation uniquement. Nous souhaitons éviter les extractions à plus de deux arguments, car nous émettons l’hypothèse que celles-ci peuvent pratiquement toujours être divisées en triplets présentant l’information de manière plus compacte et donc étant préférables aux longues extractions à plusieurs arguments. Par exemple, soit l’extraction composée suivante (He, was born, in Montreal, during winter). Elle peut être divisée en les deux extractions suivantes : (He, was born in, Montreal) et (He, was born during, winter). Ces deux dernières extractions ont pour avantage d’avoir des relations plus significatives à la place de *was born*, qui est plus général.

Le premier principe est le suivant : les tuples annotés doivent contenir une **information pertinente** qui est exprimée dans la phrase. Ils ne doivent pas contenir de généralité non informative. Cet aspect est subjectif, mais il s’agit plutôt d’un principe général. On souhaite que les tuples soient informatifs, donc dès qu’un tuple contient une information comprise dans le texte et que le tuple respecte tous les autres principes, on estime qu’il est valide. L’exemple négatif de la Table 12 est exclu des annotations car il n’est pas informatif. En effet, il présente une généralité qui ne nous offre pas d’information particulière sur les arguments.

He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.	(He, has written in, The Guardian)
	(He, has, written)

TABLE 12 – Annotation : principe d’informativité

Le second principe est celui de la **minimalité** des extractions : les tuples annotés doivent contenir une information minimale, qui ne peut être subdivisée en plus petite information. Cela permet de pénaliser les systèmes qui concatènent plusieurs informations dans une même extraction, sans spécificité accordée à la relation ou aux arguments. L’exemple de la Table 13 montre que le tuple négatif peut être séparé en trois tuples minimaux, plus pertinents.

Le troisième principe est l’**exhaustivité**. Certains arguments ou relations peuvent être affectés par des modifications, mais rester vrais sans celles-ci, il est alors nécessaire de lister toutes les formulations possibles qui respectent les autres principes. Dans l’exemple

The group was created in 2020 by three people	(The group, was, created)
	(The group, was created in, 2020)
	(The group, was created by, three people)
	(The group, was, created in 2020 by three people)

TABLE 13 – Annotation : principe de minimalité

de la Table 14, il est à la fois vrai qu'il a écrit des *opinion pieces*, des *newspaper opinion pieces* et des *magazine opinion pieces*, il faut donc lister ces trois faits distincts.

He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.	(He, has written, several newspaper opinion pieces)
	(He, has written, several magazine opinion pieces)
	(He, has written, several opinion pieces)

TABLE 14 – Annotation : principe d'exhaustivité

Le quatrième principe est la **complétude des relations**. Les relations sont les véhicules de l'information, les arguments ne doivent pas contenir d'information qui change le sens de celles-ci. Parfois, les relations peuvent être compliquées, mais nécessaires, alors qu'elles peuvent parfois être simplifiées. Il est obligatoire de les simplifier au maximum pour respecter le principe de minimalité sans pour autant qu'elles perdent leur sens original, celui exprimé dans la phrase. Dans l'exemple de la Table 15, le deuxième argument de l'exemple erroné contient le mot *over*, qui modifie le sens de la relation *is*.

Tokyo's population is over 13 Millions	(Tokyo's population, is over, 13 Millions)
	(Tokyo's population, is, over 13 Millions)

TABLE 15 – Annotation : principe de complétude des relations

Le cinquième principe porte sur la **coréférence**. Nous ne faisons pas de résolution de coréférence à l'extérieur des phrases. Même si une phrase donnée vient d'un document qui permet de régler une coréférence, nous sommes d'avis que l'OIE se veut une tâche s'exécutant sur des phrases isolées, nous résolvons donc uniquement les coréférences d'entités comprises dans les phrases prises isolément. Des tuples usant de pronoms dont

nous ne pouvons identifier l'élément de substitution peuvent sembler vides de sens, mais nous soutenons que la résolution de coréférence doit se faire à l'extérieur de l'OIE, étant une tâche en soi. Dans l'exemple de la Table 16, on ne fait pas de résolution de coréférence pour le pronom *He*, comme aucune information à ce sujet n'est disponible dans la phrase. Par contre, on remplace *them* par *tax reductions* dans l'annotation.

<p>He did not go as far as he could have in tax reductions ; indeed he combined them with increases in indirect taxes .</p>	<p>(He, combined tax reductions with, increases in indirect taxes)</p>
---	---

TABLE 16 – Annotation : principe de coréférence

L'inférence est le sixième principe. Nous soutenons que l'inférence est nécessaire, les faits directement impliqués par la phrase bien qu'ils ne soient pas exprimés verbatim sont des pièces d'information pertinentes. Une nuance est ici nécessaire en ce qui a trait aux faits implicites potentiels. Comme illustré précédemment, ceux-ci ne sont pas obligatoirement impliqués par la phrase et doivent donc être omis. Nous suivons également les recommandations de (Lechelle et al., 2019) par rapport à l'inférence légère et l'inférence lourde. Cette notion étant particulièrement propice à l'interprétation, nous tentons de l'encadrer à l'aide d'exemples présentés dans les lignes directrices en annexe A.

Ces six principes sont les principaux, d'autres règles et décisions prises sont présentées en détail dans le document complet, notamment des règles par rapport à la correction, l'attribution et la reformulation.

4.2 Lignes directrices de correspondance

Toute référence en OIE nécessite une fonction de correspondance permettant d'associer certaines extractions faites par les systèmes évalués aux extractions de référence. L'idée est d'associer des extractions dont le sens est le même, pour ainsi pouvoir déterminer quelle quantité de l'ensemble des informations présentes dans une phrase est capturée par les extractions des systèmes. Les références *OIE2016*, *WiRE57* et *CaRB* utilisent des fonctions de correspondance basées sur du *match* mot-à-mot alors que *BenchIE* fonctionne avec une correspondance exacte. Comme illustré précédemment, ces différentes approches composent avec certaines limitations, nous tentons donc d'établir ici des règles permettant de créer une correspondance plus performante. Comme pour

les annotations, nous présentons ici les grandes lignes de nos recommandations, qui sont présentées au complet dans le document de lignes directrices d’annotation en annexe A. Des exemples illustrant ces principes sont présentés aux Tables suivantes. Les annotations de référence y sont présentées en noir, les extractions que l’on souhaite faire correspondre, en vert. Les extractions qui ne correspondent pas sont en rouge. Encore une fois, toutes les annotations ne sont pas indiquées, seulement celles permettant d’illustrer le principe discuté.

Il est évident que deux extractions (exactement) équivalentes doivent correspondre. Cependant, pour des extractions qui ne sont pas en tout point identiques, on veut tout de même établir une correspondance entre les faits qui véhiculent la même pièce d’information. Certaines règles doivent être mises en place pour encadrer ce principe.

Premièrement, on veut accorder aux systèmes très peu de souplesse dans la **spécificité de la relation** extraite : c’est la relation qui véhicule l’information, il est donc important que celle-ci soit pratiquement aussi spécifique que celle de la référence. Cela étant dit, une formulation différente, mais autant spécifique devrait être acceptée. Dans l’exemple de la Table 17, la relation de l’extraction, *was thrown*, n’est pas aussi spécifique que celles de la référence.

The Finns party was thrown out of the government and the new “Blue Reform” group kept its cabinet seat.	([The] Finns party, was , thrown out of [the] government)
	([The] Finns party, was thrown out of , [the] government)
	(The Finns party, was thrown, out of the government)

TABLE 17 – Correspondance : principe de spécificité de la relation

Deuxièmement, certaines extractions faites par les systèmes peuvent présenter des **erreurs de syntaxe**, quand un mot est mal placé ou inutile. Si cette erreur change le sens de la relation ou d’un des arguments, cette extraction doit alors être refusée. Sinon, on l’accepte. Dans l’exemple de la table 18, le mot *also*, fait référence à la relation *is*, et ne change pas le sens de celle-ci alors que le mot *and* modifie le sens de l’extraction.

Troisièmement, certains **mots** peuvent être **équivalents** à ceux présents dans les annotations dans certains contextes même si nous avons choisi de ne pas les inclure dans la référence. Si ces mots sont utilisés dans les extractions système à la place de ceux utilisés dans la référence, on accepte tout de même l’extraction système. Certains choix

Known for his albums of classical piano compositions with a pop music sensibility, Solo Piano I and Solo Piano II, as well as his MC and electro albums, he is also a producer and songwriter.	(he, is [also], [a] songwriter)
	(he, is [also] [a], songwriter)
	(He, is, a songwriter also)
	(He, is also, a songwriter and)

TABLE 18 – Correspondance : erreurs de syntaxe

de mots peuvent parfois être erronés, mais on accepte quand même l’extraction si le sens demeure. Cela évite de pénaliser les systèmes qui ne reformulent pas les faits présentés par la phrase. Dans l’exemple de la Table 19, l’extraction est faite sans reformuler les mots de la phrase, en utilisant le déterminant *the* présent dans celle-ci, alors que nous avons annoté le fait avec l’article optionnel *a*, plus approprié. Cependant, comme ce choix de mot n’affecte pas le sens du fait, nous créditons l’extraction faite par le système.

He is the younger brother of the prolific film composer Christophe Beck.	(He, is, [a] younger brother)
	(He, is, the younger brother)

TABLE 19 – Correspondance : principe du choix de mot

Finalement, nous faisons le choix d’accepter une extraction système dont la relation correspond à celle d’une extraction de référence et qui présente dans un de ces arguments ce que l’on appelle un **niveau de détail** supérieur à celui de l’annotation. On entend par niveau de détail que l’argument contient des informations présentes dans deux faits distincts. Ce principe peut sembler arbitraire, mais en réalité il permet d’inclure dans la correspondance des extractions qui diffèrent dans la précision de l’un de leurs arguments, mais dont la relation et l’autre argument sont bien formulés. Ce choix est fait pour ne pas trop pénaliser les systèmes qui font des extractions longues. En effet, ces extractions comprennent tout de même des pièces d’information valides et comme les relations sont bien extraites, on veut accorder des points à ces systèmes. Dans l’exemple de la Table 20, l’extraction acceptée possède un seul niveau de détail supplémentaire, car elle combine les informations de deux faits différents alors que l’extraction refusée combine trois faits dans son deuxième argument, soient (*Gonzales, broadcasts, a web series*), (*Gonzales, broadcasts, Pop Music Masterclass*) et (*Gonzales, broadcasts Pop Music Masterclass on , WDR*).

Gonzales broadcasts a web series Pop Music Masterclass on WDR.	(Gonzales, broadcasts, [a] web series) (Gonzales, broadcasts, Pop Music Masterclass) (Gonzales, broadcasts Pop Music Masterclass on, WDR)
	(Gonzales, broadcasts, Pop Music Masterclass on WDR)
	(Gonzales, broadcasts, a web series Pop Music Masterclass on WDR)

TABLE 20 – Correspondance : niveau de détail

4.3 Données

Pour nos annotations, nous avons choisi deux jeux de données utilisés dans d’autres travaux. Premièrement nous avons décidé de reannoter les phrases utilisées à l’origine dans *WiRE57*. Ces phrases proviennent de cinq sources : trois sont des articles Wikipédia portant respectivement sur Chilly Gonzales, sur l’algorithme de maximisation de l’espérance et sur Tokyo, et les deux autres sont des articles de nouvelles de Reuters. Ces phrases ne feront pas partie de la référence à proprement parler. Elles ont été utilisées afin de développer la fonction de correspondance, et seront distribuées avec la référence, pouvant ainsi être utilisées pour l’entraînement de systèmes ou autres. Deuxièmement, nous avons reannoté l’ensemble des phrases de *BenchIE*. Ces phrases sont un sous-ensemble des phrases utilisées dans (Stanovsky & Dagan, 2016), qui proviennent d’une référence de question-réponse basée sur l’étiquetage sémantique des rôles (QA-SRL), proposé dans (He et al., 2015) et usant de phrases provenant de Wikipédia et PropBank (Palmer, Gildea, & Kingsbury, 2005).

Des 300 phrases de *BenchIE*, nous avons sélectionné un sous-ensemble de 50 phrases dont nous avons également annoté l’ensemble des correspondances entre les extractions des différents systèmes testés et les annotations. Nous avons également annoté les correspondances sur les phrases de *WiRE57*. Ce sont ces données qui nous ont permis de développer la fonction de correspondance. L’analyse de ces annotations est présentée au chapitre 5.

Résultent donc de ce processus quatre ressources, les annotations des phrases de *BenchIE* et *WiRE57*, que nous appellerons respectivement $BenchIE^{FL}$ et $WiRE57^{FL}$, ainsi que l’annotation des correspondances, à la fois sur *BenchIE50* et sur *WiRE57*.

4.4 Annotations (Plateforme AnnIE)

AnnIE (Friedrich et al., 2022) est une plateforme d’annotation développée par les auteurs de *BenchIE* qu’ils ont utilisée pour l’annotation de leur propre référence. Il s’agit d’une plateforme distribuée dans un package python permettant l’annotation manuelle dans un navigateur internet. Nous avons utilisé cette plateforme pour l’annotation de nos deux corpus de texte. *AnnIE* c’est avéré très utile, accélérant grandement le processus d’annotation. Elle permet de manipuler le texte et les annotations à l’aide du curseur plutôt que d’avoir à taper le texte. *AnnIE* propose également de pré-identifier des parties du texte qui peuvent être d’intérêt pour les annotations, comme les verbes et les groupes nominaux. Bien que nous ayons trouvé cette plateforme très utile et intuitive à utiliser, elle présentait certaines limitations dans la manipulation des *clusters* qui nous ont forcé par la suite à modifier le fichier d’annotations .txt directement.

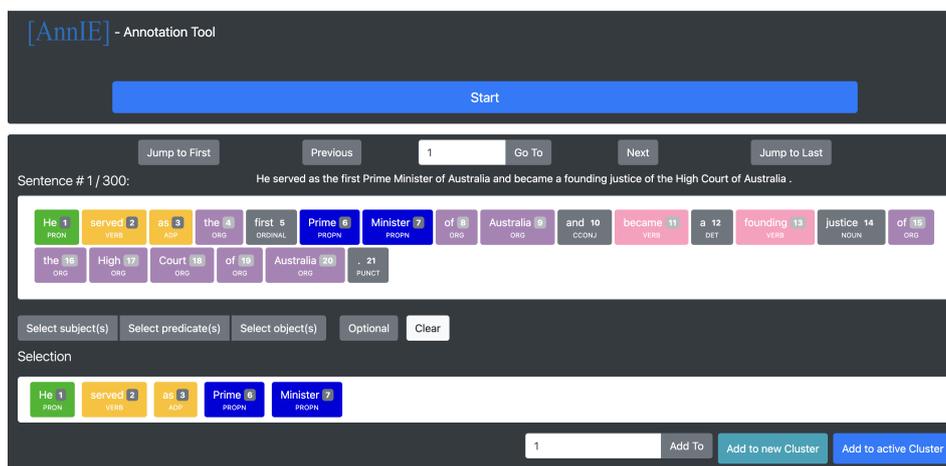


FIGURE 10 – Capture d’écran d’*AnnIE* démontrant les fonctionnalités de la plateforme

Nous avons également annoté la correspondance entre les extractions des systèmes sur deux ensembles de phrases et leurs annotations manuelles correspondantes, soit l’ensemble des phrases de *WiRE57*, et les 50 premières phrases de *BenchIE*. Ces annotations ont simplement été effectuées dans une feuille de calcul *Google Sheet*, en comparant simultanément les extractions des systèmes et les annotations manuelles.

4.5 Fonction de correspondance

Une fois les annotations effectuées, nous avons pu développer la fonction de correspondance entre les extractions système et les faits annotés. Nous avons commencé par implémenter les fonctions de correspondance originales de *WiRE* et *BenchIE*, puis avons analysé les erreurs faites par ces deux fonctions en les comparant aux annotations de

correspondance manuelles. Par la suite, nous avons développé différentes caractéristiques appartenant aux paires extractions - annotations qui tentent de capturer les différents principes de correspondance établis précédemment. Notre idée originale était d’entraîner un classifieur sur les données de *WiRE57^{FL}* qui ensuite serait utilisé sur les données de *BenchIE^{FL}* et formerait la référence finale. Cependant, au travers de nos expériences, qui seront expliquées au chapitre 5, la fonction de correspondance la plus performante sur le corpus final est plutôt une fonction statique dépendant d’un nombre limité de caractéristiques. Ces performances ont été mesurées à l’aide des métriques décrites à la section 2.3 : la précision, le rappel et le F1. Nous avons également mesuré à quel point le classement obtenu en utilisant ces différentes fonctions corrèle avec le classement obtenu en utilisant les annotations manuelles. Voici le fonctionnement de cette fonction :

Soit une extraction *ext*, faite par un système *sys*, sur une phrase *s*, et un ensemble d’annotations *a*, composé de différents clusters c_i , chacun composé de différentes formulations f_{i_j} . Pour chaque cluster c_i :

- Si *ext* est absolument identique à f_{i_j} pour toute formulation f_{i_j} , il y a correspondance. (Équivaut à la fonction de correspondance de *BenchIE*)
- Si *ext* est identique à l’une des formulations alternatives* de f_{i_j} pour toute formulation f_{i_j} , il y a correspondance.
- Si un des arguments de *ext* est identique à celui de f_{i_j} **et** que la relation d’*ext* est identique à celle de f_{i_j} **et** que l’autre argument de f_{i_j} est inclus dans l’argument de *ext* **et** que *ext* mise à plat (sans inclure de découpage entre la relation et les arguments) est identique à une formulation d’un autre cluster (f_{k_j}), il y a correspondance. Nous appelons ce critère celui du niveau de détail.
- Si aucun des trois critères précédents n’est rempli, il n’y a pas correspondance.

*Les formulations alternatives sont des formulations qui n’ont pas été annotées manuellement et qui sont formées en combinant les informations de différents clusters afin d’identifier les extractions qui respectent le principe d’un maximum d’un niveau de détail différent. La troisième condition sert également à identifier ce type d’extractions.

Des exemples pour chacun des critères sont présentés à la table 21. L’exemple de correspondance exacte ne nécessite pas d’explication, deux triplets identiques correspondent. L’exemple de correspondance par formulation alternative montre que si deux clusters dans les annotations ont la même relation et un argument équivalent, on crée une formulation alternative combinant les deux arguments différents, résultant en une formulation équivalente à celle de l’extraction, la correspondance est donc identifiée.

L'exemple de niveau de détail montre que l'extraction possède la même relation et un argument équivalent au cluster 1 (*Lugo, were*), et que, mise à plat, elle est équivalente au cluster 3, la correspondance avec le cluster 1 est donc identifiée.

<i>Lugo and Lozano were released in 1993 and continue to reside in Venezuela</i>		
Critère	Extraction	Annotation
Correspondance exacte	Lugo, resides in, Venezuela	Lugo, resides in, Venezuela
Formulation alternative	Lugo and Lozano, were, released	Cluster 1 : Lugo, were, released Cluster 2 : Lozano, were, released
Niveau de détail	Lugo, were, released in 1993	Cluster 1 : Lugo, were, released Cluster 3 : Lugo, were released in, 1993

TABLE 21 – Exemple de correspondance selon les critères de la fonction de correspondance sur une phrase de WiRE57

D'autres critères sont également utilisés pour s'assurer qu'une extraction ne correspond pas à différents faits ou que plusieurs extractions d'un même système ne correspondent pas à un même fait, ainsi que certaines modifications par rapport à la ponctuation des extractions système.

Cette fonction est la plus performante que nous avons obtenue. Cependant, elle présente toujours certaines limitations que nous explorerons au chapitre 5.

Chapitre 5 - Analyse

Ce chapitre comporte l'analyse des différentes contributions de ce mémoire. Nous explorons les annotations des différentes ressources, la conception et le résultat final de notre fonction de correspondance. Nous faisons également une comparaison des scores donnés par les différentes références existantes avec la nôtre et avec les scores obtenus par les systèmes en utilisant les annotations de correspondances manuelles. Nous terminons par l'analyse de l'évaluation sur une tâche subséquente.

5.1 Annotations

Dans cette section, nous comparons nos annotations manuelles, que nous appelons *BenchIE^{FL}*, à celles utilisées dans *BenchIE* et soulignons différentes caractéristiques de nos annotations qui résultent de nos lignes directrices proposées. La table 22 présente ces différentes statistiques.

Caractéristique des annotations	<i>BenchIE^{FL}</i>	<i>BenchIE</i>
Nombre de clusters total	1798	1354
Nombre moyen de clusters par phrase	6	4
Nombre moyen de formulations par clusters	3	6
Longueur moyenne des formulations (# mots)	11	13
Longueur médiane des formulations (# mots)	9	12
Longueur maximale des formulations (# mots)	37	37
Longueur moyenne des relations (# mots)	4	4
Longueur médiane des relations (# mots)	3	3
Longueur maximale des relations (# mots)	20	19

TABLE 22 – Caractéristiques des annotations

On y observe premièrement que nous avons annoté plus de faits au total que les annotateurs de *BenchIE*, ou un plus grand nombre de clusters moyen par phrase. Cela s'explique par le fait que nous avons inclus les informations implicites dans le texte et que nous avons suivi le principe de minimalité, signifiant que nous avons divisé les informations le plus possible, résultant en un plus grand nombre de clusters. Deuxièmement, on remarque que nous avons annoté beaucoup moins de formulations par cluster, à la fois parce que la spécificité de la relation est d'une grande importance dans nos lignes directrices, ce qui n'est pas le cas dans les annotations de *BenchIE*, et parce que nous

ne nous appuyons pas sur une correspondance exacte, il n'est alors pas nécessaire de lister toutes les formulations possibles d'un même fait. Cela implique que l'annotation d'une ressource suivant nos lignes directrices nécessite probablement moins de travail manuel d'annotation. Troisièmement, on remarque qu'en moyenne, ainsi qu'en médiane, nos annotations sont plus courtes. Cela suit encore une fois le principe de minimalité des extractions. On remarque par contre que la moyenne et la médiane des longueurs des relations sont plus ou moins équivalentes, de par notre désir de conserver la spécificité des relations. On peut donc émettre l'hypothèse que nos annotations sont plus concises, résumant mieux l'information tout en conservant la spécificité et l'informativité des relations, ce qui était un des objectifs principaux de notre réannotation. La table 23 montre un exemple des annotations de *BenchIE* et de *BenchIE^{FL}*. On y observe que nous avons listé des faits minimaux en moins de formulations et avec plus de spécificité par rapport à la relation en comparaison aux annotations originales.

5.2 Correspondance

Après avoir souligné certaines lacunes de la fonction de correspondance exacte de *BenchIE* au chapitre 3, nous cherchons à développer une fonction de correspondance plus performante. Après avoir annoté les correspondances entre les extractions des systèmes et nos annotations, suivant nos lignes directrices de correspondance, nous avons cherché à développer une fonction permettant d'identifier le mieux possible celle-ci. Pour ce faire, nous avons créé plus de cinquante caractéristiques en lien avec les paires (extraction, annotation), comme le nombre de mots les composant, le nombre de mots correspondant pour chaque partie (argument 1, relation, argument 2), la présence ou non de mots obligatoires manquant à l'extraction, ainsi que différentes métriques de similarité textuelle (BertScore (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020), BLEUScore (Papineni, Roukos, Ward, & Zhu, 2002), MeteorScore (Banerjee & Lavie, 2005)) entre les extractions et les annotations, etc. Nous avons également implémenté les fonctions de correspondance utilisées par *WiRE57* et *BenchIE*. Nous avons ensuite créé différentes fonctions de correspondance, certaines en entraînant un réseau de neurones, ou une forêt d'arbres décisionnels sur le corpus de *WiRE57^{FL}*, certaines étant des fonctions purement statiques. La performance de ces différentes fonctions a ensuite été évaluée sur le jeu de données de *BenchIE50^{FL}* et est présentée à la table 24.

La métrique utilisée dans cette expérience est le F1 mesuré sur les correspondances annotées manuellement, comme mentionné à la section 2.3. La fonction *Match exact* correspond à la fonction de correspondance utilisée dans *BenchIE*. Elle performe moins bien

<i>Ms. Waleson is a free - lance writer based in New York .</i>	
<i>BenchIE</i>	<p>102->Cluster 1 :</p> <p>[Ms.] Waleson ->is ->[a] free [-] lance writer [based in New York] [Ms.] Waleson ->is ->[a] free-lance writer [based in New York] [Ms.] Waleson ->is ->[a] freelance writer [based in New York]</p> <p>102->Cluster 2 :</p> <p>[Ms.] Waleson ->[is] based in ->New York [Ms.] Waleson ->[is] based ->in New York [Ms.] Waleson ->is ->based in New York [a] free [-] lance writer ->is ->based in New York [a] free [-] lance writer ->[is] based ->in New York [a] free [-] lance writer ->[is] based in ->New York [Ms.] Waleson ->is [a] free [-] lance writer based in ->New York [Ms.] Waleson ->is [a] free [-] lance writer based ->in New York [Ms.] Waleson ->is [a] free [-] lance writer ->based in New York</p> <p>[Ms.] Waleson ->is [a] free-lance writer based in ->New York [Ms.] Waleson ->is [a] free-lance writer based ->in New York [Ms.] Waleson ->is [a] free-lance writer ->based in New York</p>
<i>BenchIE^{FL}</i>	<p>102->Cluster 1 :</p> <p>[Ms.] Waleson ->is ->[a] writer [Ms.] Waleson ->is [a] ->writer</p> <p>102->Cluster 2 :</p> <p>[Ms.] Waleson ->is [doing] ->free [-] lance [Ms.] Waleson ->is [a] ->free [-] lance [Ms.] Waleson ->is [a] ->free [-] lancer [Ms.] Waleson ->is [doing] ->free [-] lancing</p> <p>102->Cluster 3 :</p> <p>[Ms.] Waleson ->[is] based in ->New York</p>

TABLE 23 – Exemples des annotations de *BenchIE* et de *BenchIE^{FL}*.

Fonction de correspondance	F1 données d’entraînement (WiRE57)	F1 données d’évaluation (BenchIE50)
Match exact (BenchIE)	0.852	0.827
Match exact + Formulations alternatives	0.874	0.871
Match exact + Niveau de détail	0.860	0.870
Match Custom (Exact + Form. alt. + Niv. det.)	0.924	0.958
Match Custom + Ponctuation	0.930	0.965
MLPC Caractéristiques simples	0.968	0.860
MLPC Caractéristiques riches	0.982	0.855
LightGBM Caractéristiques simples	0.893	0.879
LightGBM Caractéristiques riches	0.902	0.854

TABLE 24 – Performance des fonctions de correspondance

que les autres fonctions évaluées à cause de sa rigidité inhérente. Elle possède par contre évidemment une précision parfaite. La fonction *Match exact + Formulations alternatives*, qui correspond aux deux premiers critères de la fonction décrite à la section 4.5, performe un peu mieux que cette dernière, elle parvient à capturer plus de correspondances, mais en identifie plusieurs qui sont fausses. La fonction *Match exact + Niveau de détail*, elle, correspond au premier et au troisième critère de la fonction finale décrite à la section 4.5. Cette version performe également mieux que la fonction de correspondance exacte mais à toujours un rappel trop bas et une précision moindre. La fonction *Match Custom* correspond à la fonction finale décrite à la section 4.5, composée des trois critères, sans les spécifications sur la ponctuation. Ces spécifications sont les suivantes : tout caractère de ponctuation apparaissant dans le package *string* de Python (c.à.d tout caractère dans `[! " # % & () * + , - . / : ; = < ? @ [\] ^ _ ‘ —]`) est retiré à la fois des extractions et des annotations, on calcule ensuite la fonction finale décrite à la section 4.5 sur les caractéristiques nécessaires évaluées sans cette ponctuation. C’est cette version de la fonction *Match custom*, *Match Custom + Ponctuation* qui est la plus performante sur l’ensemble de données d’évaluation, et qui est donc la fonction de correspondance utilisée dans notre référence. Les fonctions suivantes sont des modèles qui ont été entraînés sur les données d’entraînement composées des phrases de WiRE57. *MLPC* correspond au perceptron multicouche implémenté dans le package *scikit-learn* (Pedregosa et al., 2011). *LightGBM* (Ke et al., 2017) est une librairie de Gradient Boosting que nous avons utilisée pour entraîner des forêts d’arbres décisionnels. Ces deux types de modèles ont été entraînés, et leurs hyperparamètres ainsi que les caractéristiques utilisées, sélectionnés

à l'aide d'une recherche par grille. Pour ces deux types de modèles, il est évident que ceux-ci souffrent d'un problème de sur-apprentissage. Le F1 sur le jeu de données d'entraînement est très haut, meilleur que celui de notre fonction la plus performante, mais ne parvient pas à généraliser au jeu de données de test, présentant des performances bien moindres sur celui-ci. Cela s'explique potentiellement par le changement de domaine, les phrases de *WiRE57* portant sur des sujets différents et ayant potentiellement d'autres caractéristiques les différenciant de ceux de *BenchIE*. Ces mauvaises performances sur le jeu de données de test justifient que nous n'utilisons pas ces modèles dans notre référence finale, mais plutôt la fonction statique *Custom Match*. Nous avons également rapporté les performances des deux modèles entraînés sur deux ensembles de caractéristiques différents, simples et riches, les caractéristiques simples incluant les caractéristiques comme le compte de mots, le nombre de mots correspondant, etc., et les caractéristiques riches incluant les métriques de similarité, les distances d'édition entre les textes, etc., car ces ensembles de caractéristiques présentent le même motif pour le perceptron et la forêt d'arbres décisionnels. Dans les deux cas, les caractéristiques riches permettent d'accroître les performances sur le jeu de données d'entraînement, mais renforcent encore plus la non-généralisation aux données de test.

5.3 Limitations de la fonction de correspondance

Bien que la fonction de correspondance proposée performe mieux que toutes les autres fonctions testées, elle présente tout de même certaines limitations. Elle obtient une précision de 0.972 et un rappel de 0.891 sur le jeu de données d'entraînement, et une précision de 0.996 et un rappel de 0.936 sur le jeu de données de test. La comparaison de la sortie de la fonction aux annotations manuelles est présentée à la table 25.

On remarque que très peu de faux positifs sont retournés par la fonction. Cela s'explique par le fait que notre fonction de correspondance, malgré qu'elle soit plus souple que celle de *BenchIE*, reste tout de même conservatrice. On s'appuie sur des correspondances exactes entre les parties des extractions et une correspondance à plat entre les différents clusters d'une même phrase. Les seuls faux positifs s'expliquent par l'incapacité d'identifier quand un cluster correspond déjà à une autre extraction faite par le système, mais que l'extraction actuellement évaluée remplit les critères de correspondance de la fonction *match custom*. Des exemples de ce type d'erreur sont présentés à la table 26.

La principale source d'erreur est donc les faux négatifs, des paires (extraction - annotation) pour lesquelles la fonction de correspondance n'identifie pas la similarité, mais

	Correspondance Custom Match	Non-correspondance Custom Match
Correspondance annotée	Vrai positifs 279 (246)	Faux négatifs 19 (30)
Non-correspondance annotée	Faux positifs 1 (7)	Vrais négatifs 9101 (10690)

TABLE 25 – Comparaison de la sortie de la fonction avec les annotations manuelles. Les éléments entre parenthèse sont ceux obtenus sur le jeu de données d’entraînement, les autres sur le jeu de données de test.

Extraction	Annotation (cluster)
(The metropolitan government, administers , 39 municipalities in the western part of the prefecture)	([The] metropolitan government, [also] administers, municipalities)
(Arminius, is, leader of allies)	(Arminius, is [a], leader) (Arminius, is, [a] leader) (Arminius, is [the], leader) (Arminius, is, [the] leader)

TABLE 26 – Exemples de faux positifs par la fonction de correspondance

pour lesquelles nous avons annoté une correspondance. Des exemples de ce type d’erreur sont présentés à la table 27. Celles-ci ont différentes causes. Premièrement, comme le premier exemple, il peut s’agir d’un problème de choix de mots, le système ayant utilisé un mot ne se trouvant pas dans la phrase d’origine, comme *be*, à la place de ce que nous avons annoté, *was*. Pour résoudre ce problème, il serait possible d’introduire une liste de mots interchangeable, mais en pratique nous avons observé que cela introduit plus d’erreurs de correspondance que de corrections. Deuxièmement, des extractions qui possèdent un niveau de détail supplémentaire dans l’un de leurs arguments, comme le deuxième exemple, mais dont le niveau de détail n’a pas été annoté par manque de contexte ne peuvent pas être identifiées par la fonction, malgré qu’elles respectent tous les critères des lignes directrices de correspondance.

Extraction	Annotation (cluster)
(club policy, be promote talent into, senior team)	(club policy, was [to] promote talent into, [the] senior team)
(the band, released, Sahara “ in August of that year)	(([the] band, released, [‘] [‘] Sahara [’]) ([‘] [‘] Sahara [’], was released by, [the] band)

TABLE 27 – Exemples de faux négatifs par la fonction de correspondance

La fonction que nous avons développée remplit donc les objectifs de capturer plus de correspondances que les fonctions utilisées par d’autres références, mais elle fait toujours plusieurs erreurs qu’il est malheureusement plus difficile de corriger.

5.4 Comparaison avec les annotations manuelles

Nous comparons ici les performances des différents systèmes établies par notre référence à l’aide de la fonction de correspondance finale et de la correspondance exacte à celles obtenues à l’aide de nos annotations manuelles. Cette comparaison est encore une fois faite sur le jeu de données annotées *BenchIE50^{FL}*. La figure 11 présente ces différents scores, la métrique utilisée étant le F1. On peut remarquer que bien que les l’ordre dans lequel les systèmes sont ordonnés par les deux fonctions de correspondance soit le même, la fonction Custom Match capture beaucoup plus de correspondances, surtout pour les systèmes performants le moins bien (*IMojIE*, *M2OIE*).

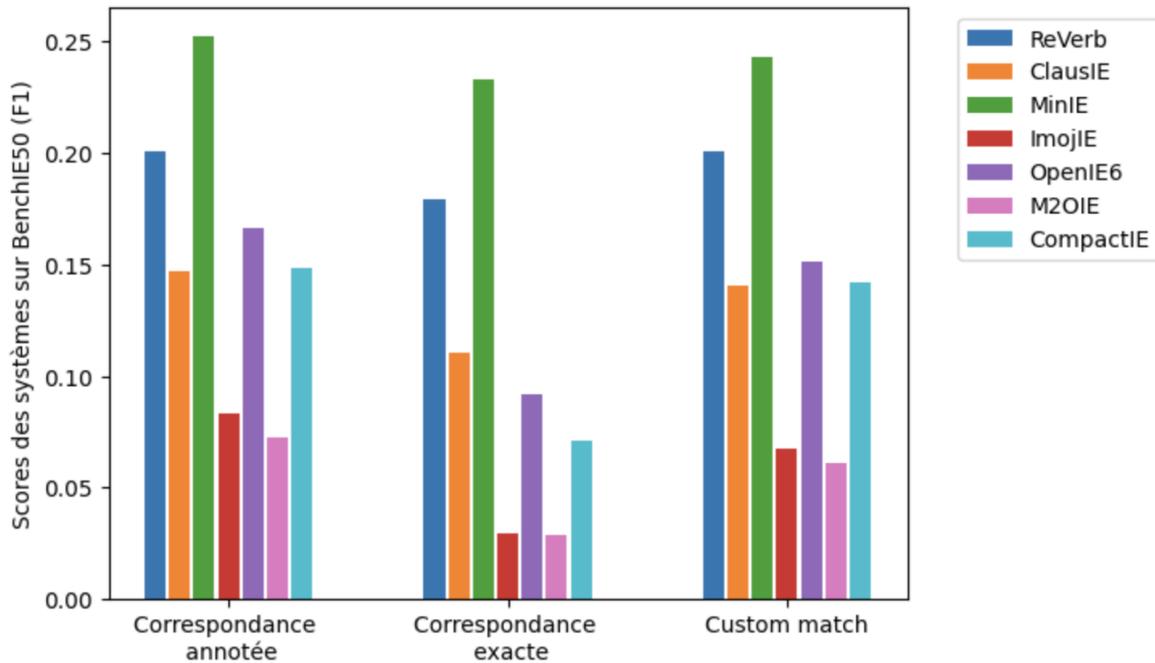


FIGURE 11 – Score des systèmes selon les fonctions de correspondance sur $BenchIE50^{FL}$

Cette observation est vérifiée par la mesure de la corrélation entre ces différents scores, présentée à la table 28. On voit que, bien que la corrélation obtenue par la fonction de correspondance exacte soit élevée, celle obtenue par Custom Match l’est encore plus. Nous présentons également la faible corrélation obtenue en utilisant la fonction de correspondance de *WiRE57* pour illustrer les critiques présentées précédemment par rapport aux fonctions basées sur des comptes de mots.

Fonction de correspondance	Corrélation entre la fonction de correspondance et l’annotation manuelle
WiRE57	0.219
BenchIE	0.961
Custom Match	0.997

TABLE 28 – Corrélations entre les fonctions de correspondance et les annotations manuelles sur $BenchIE50^{FL}$

5.5 Comparaison des références

Nous présentons les performances des différents systèmes sur les différentes références étudiées à la figure 12 et à la table 29. On y observe que chaque référence identifie très

différemment quels systèmes sont les plus performants. Comme démontré précédemment, c'est avec référence que les scores s'alignent le mieux avec ceux obtenus en annotant manuellement la correspondance entre les extractions système et les annotations, suivant nos lignes directrices de correspondance. On peut observer certaines similarités entre les résultats obtenus sur *BenchIE* et notre référence, qui désignent tous les deux *MinIE* comme étant le système le plus performant. Par contre, pour pratiquement tous les autres systèmes, il existe de grandes différences entre leurs classements finaux selon ces deux références.

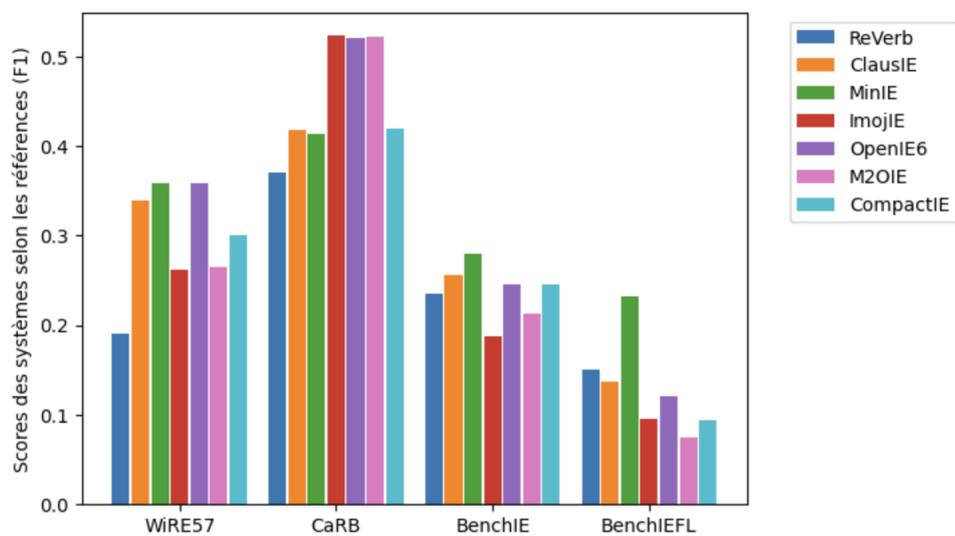


FIGURE 12 – Résultats (F1) des systèmes sur différentes références en utilisant leur script d'évaluation respectifs.

Système	Référence				
	WiRE57	CaRB	BenchIE	Annotations manuelles (BenchIE50)	$BenchIE^{FL}$
ReVerb	0.191	0.371	0.224	0.196	0.151
ClausIE	0.339	0.418	0.256	0.159	0.138
MinIE	0.359	0.413	0.279	0.261	0.232
IMojIE	0.262	0.523	0.187	0.083	0.096
OpenIE6	0.359	0.521	0.245	0.162	0.121
M2OIE	0.265	0.522	0.213	0.072	0.075
CompactIE	0.300	0.419	0.245	0.152	0.095

TABLE 29 – Résultats (F1) des systèmes sur différentes références en utilisant leur script d’évaluation respectifs.

Bien qu’il pourrait être tentant ici de tirer des conclusions sur les réelles avancées des systèmes d’OIE, au vu de la performance impressionnante d’un des premiers systèmes, *ReVerb*, en comparaison avec des systèmes beaucoup plus récents comme *M2OIE* et *CompactIE*, ce n’est ici pas l’objet de l’étude. En effet, en utilisant ces systèmes, nous n’avons aucunement cherché à optimiser leur sortie à l’aide des nombreux métaparamètres disponibles. Les résultats obtenus ne reflètent donc en rien la performance optimale de ces systèmes, uniquement leur performance dans leur état initial. Cela étant dit, nos comparaisons entre les références utilisant ces systèmes sont parfaitement valides, car on ne compare pas les performances optimales des systèmes, simplement leurs scores respectifs sur les différentes références.

5.6 Évaluation sur tâche subséquente

Une des tâches du traitement des langues qui se prête aux sorties des systèmes d’OIE est le question-réponse (QA). De manière générale, en QA, on dispose en entrée d’une question sous forme textuelle, et on cherche à produire en sortie une réponse à cette question, également sous forme de texte. Une formulation alternative de cette tâche est proposée dans (Yan et al., 2018), l’*Assertion Based Question-Answering* (ABQA). Dans cette tâche, on reçoit en entrée une question sous forme textuelle et un passage (une ou plusieurs phrases), lequel contient la réponse à la question, et on cherche à produire en sortie cette réponse dans le format (*Sujet, Relation, Argument 1, Argument 2, ...*). Dans leur article, les auteurs introduisent une référence d’ABQA, *WebAssertions*, composée d’environ 55 000 paires de passages-questions. On remarque tout de suite une grande

similarité dans le format de sortie avec l’OIE, les auteurs ont en effet utilisé le système d’OIE *ClausIE* pour extraire automatiquement les réponses des passages. Pour ce faire, ils ont extrait tous les tuples possibles et ont demandé à des annotateurs de noter ceux qui permettaient de répondre à la question associée au passage, pour finalement obtenir en sortie le triplet (question, passage, réponse).

Nous avons voulu utiliser cette tâche subséquente à l’OIE pour obtenir un classement des systèmes d’OIE nous permettant de valider en partie le classement obtenu par notre référence. Nous voulions initialement utiliser directement *WebAssertions* mais l’examen des réponses du jeu de données nous a indiqué que celui-ci comprenait beaucoup de réponses approximatives, bruitées ou non pertinentes. Des exemples des données de *WebAssertions* sont présentées à la table 30. Pour corriger ces problèmes, nous avons donc reannoté un sous-ensemble de 100 paires question-phrased avec la réponse à la question dans le format des clusters de notre référence, en utilisant nos lignes directrices d’annotation, produisant ainsi une cinquième ressource, *WebAssertions^{FL}*, elle aussi disponible. Des exemples de ces nouvelles annotations sont également présentés à la table 30.

Question	Réponse WebAssertion	Réponse WebAssertionFL
<i>Passage : the internal frame backpack is a recent innovation , invented in 1967 by greg lowe, who went on to found lowe alpine and lowepro, companies specializing in backpacks and other forms of carrying bags for various equipment .</i>		
when was the backpack invented ?	a recent innovation, invented, in 1967, by greg lowe	[the] backpack, [was]/[be]/[is] invented in, 1967 [the] internal frame backpack, [was]/[be]/[is] invented in, 1967
<i>Passage : the inner layer of a turtle 's shell is made up of about 60 bones that include portions of the backbone and the ribs , meaning the turtle cannot crawl out of its shell .</i>		
how many bones in a turtle shell ?	about 60 bones, include, portions of the backbone about 60 bones, include, the ribs	[a] turtle ['s] shell, [is] made up of [about], 60 bones [a] turtle ['s] shell, has [about], 60 bones ...

TABLE 30 – Exemples de triplets question-passage-réponse de WebAssertions et WebAssertionsFL

Pour évaluer les systèmes d'OIE, nous avons ensuite mesuré leur capacité à extraire la réponse dans l'ensemble de leurs extractions de sortie. Pour ce faire, nous avons utilisé la fonction de correspondance utilisée dans notre référence, et avons mesuré un point si au moins une des extractions système correspondait à une des annotations pour une question-passage donnée (certaines questions peuvent avoir plus d'une réponse dans un passage, ou une réponse qui nécessite plusieurs clusters). Cette métrique n'évalue en rien la capacité du système à identifier quelle réponse donner à la question, mais simplement si le système a pu extraire l'information nécessaire pour répondre à la question. Les scores obtenus par cette expérience sont présentés à la table 31.

Système	Score
ReVerb	0.23
ClausIE	0.18
MinIE	0.27
IMojIE	0.17
OpenIE6	0.17
M2OIE	0.16
CompactIE	0.17

TABLE 31 – Scores des systèmes sur $WebAssertions^{FL}$

Nous avons ensuite mesuré la corrélation entre les scores obtenus par les systèmes sur $WebAssertions^{FL}$, et ceux obtenus sur les différentes références étudiées et notre nouvelle référence. Ces corrélations sont présentées à la table 32. On y observe que la corrélation est la plus haute et très élevée entre les scores obtenus sur notre référence et ceux sur $WebAssertions^{FL}$. Cela nous indique que notre référence est un bon indicateur des performances des systèmes d’OIE sur l’ABQA, tâche découlant du QA. On ne peut pas ici généraliser en déclarant que notre référence peut déterminer les performances des systèmes d’OIE sur l’ensemble des tâches dans lesquelles ces systèmes pourraient être utilisés, mais il s’agit tout de même d’un indicateur positif de la pertinence de notre référence.

Référence	Corrélation
WiRE57	0.044
CaRB	-0.649
BenchIE	0.616
BenchIEFL	0.940

TABLE 32 – Corrélation entre les scores des systèmes selon les références et leurs scores sur $WebAssertions^{FL}$

Soulignons que les travaux évaluant l’utilisation des systèmes d’OIE pour des tâches particulières sont rares. *BenchIE* traite brièvement de ce sujet en introduisant plusieurs versions de la référence visant différents usages, mais ne mène pas une évaluation complète de l’impact de ces différentes versions. (Pei, Jindal, Chang, Zhai, & Li, 2023) propose une analyse des différentes références existantes et l’évaluation de différents systèmes d’OIE appliqués au *Complex QA*, une forme de QA, en utilisant le système QUEST (Lu et al., 2019). Leurs conclusions sont cependant différentes des nôtres, les auteurs comparent les systèmes d’OIE neuronnax sur *OIE2016*, *WiRE57* et *CaRB* et ils identifient M2OIE et

une variante de OpenIE6 comme étant les systèmes les plus performants.

Conclusion

Le développement de systèmes d'OIE performants est vraisemblablement freiné par le manque de références fiables (Liu, Gao, Dong, Huang, & Zhang, 2022). L'évaluation des systèmes d'OIE est une tâche complexe de par la nature de cette discipline. Le but premier de l'OIE est d'extraire les faits qui sont exprimés dans une phrase, or cette notion n'est pas clairement définie et ne fait pas consensus dans la communauté. Notre but est alors à la fois de recadrer l'OIE pour unifier les définitions nécessaires et de proposer une ressource d'évaluation la plus proche possible de ces nouvelles définitions.

Nous proposons donc une nouvelle référence pour l'extraction d'information ouverte, *BenchIE^{FL}*, composée d'une réannotation d'un corpus précédemment utilisé et d'une nouvelle fonction de correspondance. Cette proposition est motivée par la mise en évidence de certaines lacunes des références préexistantes. Le développement de cette référence nous permet également de proposer de nouvelles lignes directrices pour le domaine de l'OIE, à la fois pour l'annotation des références et des correspondances. Ces travaux nous ont amenés à créer un outil d'analyse pour la référence *BenchIE*, ainsi nous proposons également deux autres corpus reannotés, *WiRE57^{FL}*, une autre référence d'OIE utilisée comme données d'entraînement, et *WebAssertions^{FL}*, une référence d'ABQA, également une réannotation d'un jeu de données existant. Cette dernière nous a permis de tester la robustesse de notre référence en comparant ses scores à ceux d'une tâche utilisant la sortie des systèmes d'OIE.

Nous publions ces différentes ressources, ainsi que le code permettant de les utiliser pour l'évaluation, avec les lignes directrices qui, nous l'espérons, pourront mieux définir l'OIE dans son ensemble et permettront de cerner plus précisément ce vers quoi nous estimons que les systèmes devraient tendre dans leurs sorties. Cela étant dit, notre référence est elle-même limitée, certains cas d'exception dans l'annotation mériteraient un examen plus en profondeur ou une solution permettant de contourner le problème d'ambiguïté dans la phrase d'origine. Il existe aussi certaines limites à la portée de notre fonction de correspondance qui ne parvient pas à capturer toutes les extractions que nous voudrions juger comme valides. Finalement, le problème du contexte et de la coréférence en est un que nous avons contourné dans notre référence en ne considérant que les références internes aux phrases, sans tenir compte du contexte global, mais cette approche n'est pas nécessairement celle à suivre pour l'OIE en général.

Références

- Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate*.
- Banerjee, S., & Lavie, A. (2005, juin). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72). Ann Arbor, Michigan : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/W05-0909>
- Bhardwaj, S., Aggarwal, S., & Mausam, M. (2019, novembre). CaRB : A crowdsourced benchmark for open IE. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 6262–6267). Hong Kong, China : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D19-1651> doi: 10.18653/v1/D19-1651
- Christensen, J., Mausam, Soderland, S., & Etzioni, O. (2011). An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on knowledge capture* (pp. 113–120).
- Cowie, J., & Wilks, Y. (1996). Information extraction.
- Cui, L., Wei, F., & Zhou, M. (2018, juillet). Neural open information extraction. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2 : Short papers)* (pp. 407–413). Melbourne, Australia : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P18-2065> doi: 10.18653/v1/P18-2065
- Del Corro, L., Abujabal, A., Gemulla, R., & Weikum, G. (2015, septembre). FI-NET : Context-aware fine-grained named entity typing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 868–878). Lisbon, Portugal : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D15-1103> doi: 10.18653/v1/D15-1103
- Del Corro, L., & Gemulla, R. (2013). Clausie : Clause-based open information extraction. In *Proceedings of the 22nd international conference on world wide web* (p. 355–366). New York, NY, USA : Association for Computing Machinery. Consulté sur <https://doi.org/10.1145/2488388.2488420> doi: 10.1145/2488388.2488420
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, juin). 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)* (pp.

- 4171–4186). Minneapolis, Minnesota : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/N19-1423> doi: 10.18653/v1/N19-1423
- Etzioni, O., & Cafarella., M. B. M. J. (2007). *Machine reading*. AAAI Spring Symposium.
- Fader, A., Soderland, S., & Etzioni, O. (2011, juillet). Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1535–1545). Edinburgh, Scotland, UK. : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D11-1142>
- Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (p. 1156–1165). New York, NY, USA : Association for Computing Machinery. Consulté sur <https://doi.org/10.1145/2623330.2623677> doi: 10.1145/2623330.2623677
- Fatahi Bayat, F., Bhutani, N., & Jagadish, H. (2022, juillet). CompactIE : Compact facts in open information extraction. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics : Human language technologies* (pp. 900–910). Seattle, United States : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2022.naacl-main.65> doi: 10.18653/v1/2022.naacl-main.65
- Friedrich, N., Gashteovski, K., Yu, M., Kotnis, B., Lawrence, C., Niepert, M., & Glavaš, G. (2022, mai). AnnIE : An annotation platform for constructing complete open information extraction benchmark. In *Proceedings of the 60th annual meeting of the association for computational linguistics : System demonstrations* (pp. 44–60). Dublin, Ireland : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2022.acl-demo.5> doi: 10.18653/v1/2022.acl-demo.5
- Gashteovski, K., Gemulla, R., & del Corro, L. (2017, septembre). MinIE : Minimizing facts in open information extraction. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2630–2640). Copenhagen, Denmark : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D17-1278> doi: 10.18653/v1/D17-1278
- Gashteovski, K., Yu, M., Kotnis, B., Lawrence, C., Niepert, M., & Glavaš, G. (2022, mai). BenchIE : A framework for multi-faceted fact-based open information extraction evaluation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 4472–4490). Dublin, Ireland : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2022.acl-long.307> doi: 10.18653/v1/2022.acl-long.307

- Gu, J., Lu, Z., Li, H., & Li, V. O. (2016, août). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 1631–1640). Berlin, Germany : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P16-1154> doi: 10.18653/v1/P16-1154
- He, L., Lewis, M., & Zettlemoyer, L. (2015, septembre). Question-answer driven semantic role labeling : Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 643–653). Lisbon, Portugal : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D15-1076> doi: 10.18653/v1/D15-1076
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 3146–3154.
- Kolluru, K., Adlakha, V., Aggarwal, S., Mausam, & Chakrabarti, S. (2020, novembre). OpenIE6 : Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 3748–3761). Online : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2020.emnlp-main.306> doi: 10.18653/v1/2020.emnlp-main.306
- Kolluru, K., Aggarwal, S., Rathore, V., Mausam, & Chakrabarti, S. (2020, juillet). IMO-JIE : Iterative memory-based joint open information extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5871–5886). Online : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2020.acl-main.521> doi: 10.18653/v1/2020.acl-main.521
- Kotnis, B., Gashteovski, K., Rubio, D., Shaker, A., Rodriguez-Tembras, V., Takamoto, M., ... Lawrence, C. (2022, mai). MILIE : Modular & iterative multilingual open information extraction. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 6939–6950). Dublin, Ireland : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2022.acl-long.478> doi: 10.18653/v1/2022.acl-long.478
- Lechelle, W., Gotti, F., & Langlais, P. (2019, août). WiRe57 : A fine-grained benchmark for open information extraction. In *Proceedings of the 13th linguistic annotation workshop* (pp. 6–15). Florence, Italy : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/W19-4002> doi: 10.18653/v1/W19-4002
- Liu, P., Gao, W., Dong, W., Huang, S., & Zhang, Y. (2022). *Open information extraction from 2007 to 2022 – a survey*.

- Lu, X., Pramanik, S., Saha Roy, R., Abujabal, A., Wang, Y., & Weikum, G. (2019). Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval* (p. 105–114). New York, NY, USA : Association for Computing Machinery. Consulté sur <https://doi.org/10.1145/3331184.3331252> doi: 10.1145/3331184.3331252
- Mausam, Schmitz, M., Soderland, S., Bart, R., & Etzioni, O. (2012, juillet). Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 523–534). Jeju Island, Korea : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D12-1048>
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank : An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–106. Consulté sur <https://aclanthology.org/J05-1004> doi: 10.1162/0891201053630264
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, juillet). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P02-1040> doi: 10.3115/1073083.1073135
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pei, K., Jindal, I., Chang, K. C.-C., Zhai, C., & Li, Y. (2023, juillet). When to use what : An in-depth comparative empirical analysis of OpenIE systems for downstream applications. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 929–949). Toronto, Canada : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2023.acl-long.53> doi: 10.18653/v1/2023.acl-long.53
- Ponza, M., Del Corro, L., & Weikum, G. (2018, octobre-novembre). Facts that matter. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1043–1048). Brussels, Belgium : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D18-1129> doi: 10.18653/v1/D18-1129
- Ro, Y., Lee, Y., & Kang, P. (2020, novembre). Multi²OIE : Multilingual open information extraction based on multi-head attention with BERT. In *Findings of the association for computational linguistics : Emnlp 2020* (pp. 1107–1117). Online : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/>

2020.findings-emnlp.99 doi: 10.18653/v1/2020.findings-emnlp.99

- Sandhaus, E. (2008). *The new york times annotated corpus*. Philadelphia : Linguistic Data Consortium.
- Stanovsky, G., & Dagan, I. (2016, novembre). Creating a large benchmark for open information extraction. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2300–2305). Austin, Texas : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/D16-1252> doi: 10.18653/v1/D16-1252
- Stanovsky, G., Dagan, I., & Mausam. (2015, juillet). Open IE as an intermediate structure for semantic tasks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2 : Short papers)* (pp. 303–308). Beijing, China : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P15-2050> doi: 10.3115/v1/P15-2050
- Stanovsky, G., Michael, J., Zettlemoyer, L., & Dagan, I. (2018, juin). Supervised open information extraction. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics : Human language technologies, volume 1 (long papers)* (pp. 885–895). New Orleans, Louisiana : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/N18-1081> doi: 10.18653/v1/N18-1081
- Wang, X., Peng, M., Sun, M., & Li, P. (2022, mai). OIE@OIA : an adaptable and efficient open information extraction framework. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1 : Long papers)* (pp. 6213–6226). Dublin, Ireland : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/2022.acl-long.430> doi: 10.18653/v1/2022.acl-long.430
- Wu, F., & Weld, D. S. (2010, juillet). Open information extraction using Wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 118–127). Uppsala, Sweden : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P10-1013>
- Yan, Z., Tang, D., Duan, N., Liu, S., Wang, W., Jiang, D., ... Li, Z. (2018). *Assertion-based qa with question-aware open information extraction*.
- Yates, A., Banko, M., Broadhead, M., Cafarella, M., Etzioni, O., & Soderland, S. (2007, avril). TextRunner : Open information extraction on the web. In *Proceedings of human language technologies : The annual conference of the north American chapter of the association for computational linguistics (NAACL-HLT)* (pp. 25–26). Rochester, New York, USA : Association for Computational Linguistics. Consulté

sur <https://aclanthology.org/N07-4013>

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *Bertscore : Evaluating text generation with bert*.

Zhou, J., & Xu, W. (2015, juillet). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1 : Long papers)* (pp. 1127–1137). Beijing, China : Association for Computational Linguistics. Consulté sur <https://aclanthology.org/P15-1109> doi: 10.3115/v1/P15-1109

Annexe A - Lignes directrices d'annotation

Ce document contient les lignes directrices d'annotation pour la tâche d'extraction d'information ouverte et a été utilisé dans le processus d'annotation de la référence *BenchIE^{FL}*. Les différents principes, en ordre d'importance, dictent les faits qui doivent et ne doivent pas être annotés. Les extractions en noir représentent les annotations manuelles, les exemples en vert représentent une correspondance alors que les exemples en rouge ne doivent pas correspondre. Dans ce document nous faisons référence aux faits annotés en utilisant de manière interchangeable les mots *tuple*, *fait*, et *triplet*. Les informations sont présentées selon le format suivant : les différentes formulations d'un même cluster (d'un même fait) sont dans un paragraphe et un saut de ligne les sépare.

<i>Phrase</i>	Cluster 1 Formulation 1 Cluster 1 Formulation 2 Cluster 2 Formulation 1 etc.
---------------	---

A.1 Nombre d'argument

Tous les tuples doivent contenir entre 1 et 2 arguments. Les extractions possédant plus de deux arguments peuvent être divisés en extractions plus compactes.

<i>Kyle left for school on Monday</i>	(Kyle, left for, school) (Kyle, left for school on, Monday) (Kyle, left on, Monday) (Kyle, left for, school, on Monday)
---------------------------------------	--

A.2 Information pertinente

Les tuples annotés doivent contenir une information pertinente qui est exprimée dans la phrase. Les tuples doivent être informatifs et pertinents. Ils ne doivent pas contenir de généralité ou de mots vides de sens, qui ne véhiculent aucune information.

<i>He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.</i>	(He, has written, opinion pieces) (He, has, written)
---	---

A.3 Minimalité

Les tuples annotés doivent contenir une information minimale, qui ne peut être subdivisée en plus petite information. Aucun tuple ne doit contenir d'information sur deux entités différentes si celle-ci est vraie pour les deux, et aucun tuple ne doit contenir plus d'une information sur une entité si celles-ci peuvent être divisées. Dans certains cas spécifiques comme le troisième exemple, certaines relations comme *combines* se prêtent mal à la minimalité, il est alors nécessaire de créer versions de l'annotation, une ou les arguments sont séparés dans deux clusters différents (*elements of a city* et *elements of a prefecture*) et une version où les deux arguments sont regroupés. Cette deuxième version doit être ajoutée aux deux clusters précédents. Ainsi, deux systèmes qui séparent ou qui regroupent les arguments ne seront distingués que par une précision légèrement inférieure pour le système qui sépare les extractions.

<i>Bob is a canadian musician</i>	(Bob, is, canadian) (Bob, is, a musician) (Bob, is, a canadian musician)
<i>The group was created in 2020 by three people</i>	(The group, was, created) (The group, was created in, 2020) (The group, was created by, three people)
<i>Tokyo is often referred to as a city, but is officially known and governed as a "metropolitan prefecture", which differs from and combines elements of a city and a prefecture, a characteristic unique to Tokyo.</i>	(Tokyo, combines, elements of a prefecture) (Tokyo, combines, elements of a city and a prefecture) (Tokyo, combines, elements of a city) (Tokyo, combines, elements of a city and a prefecture)

Il est parfois nécessaire de séparer les informations, si et seulement si elles sont également vraies lorsque séparées. Dans le deuxième exemple, le chien n'est pas noir ni brun, il est noir ET brun, alors que dans le premier exemple, Il possède des ancêtres corniques et Il possède des ancêtres gallois.

<i>He has Cornish as well as Welsh ancestry.</i>	(He, has, Cornish ancestry) (He, has, Welsh ancestry)
<i>This dog is black and brown.</i>	(The dog, is, black and brown)

A.4 Exhaustivité

Toutes les informations minimales présentées dans la phrase doivent être incluses dans un tuple. Certains arguments ou relations peuvent être affectés par des modifications, mais rester vrais sans celles-ci, il est alors nécessaire de lister toutes les formulations possibles qui respectent les autres principes. Dans l'exemple, il est à la fois vrai qu'il a écrit des *opinion pieces*, des *newspaper opinion pieces* et des *magazine opinion pieces*, il faut donc lister ces trois faits distincts.

<i>He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.</i>	(He, has written, several newspaper opinion pieces)
	(He, has written, several magazine opinion pieces)
	(He, has written, several opinion pieces)

A.5 Relation complètes

Les relations sont les véhicules de l'information, les arguments ne doivent pas contenir d'information qui changent le sens de celles-ci. Les relations peuvent être compliquées, mais nécessaires, alors qu'elles peuvent parfois être simplifiées. Il est obligatoire de les simplifier au maximum pour respecter le principe de minimalité sans pour autant qu'elles perdent leur sens original, celui exprimé dans la phrase. Dans l'exemple, le deuxième argument de l'annotation erronée contient le mot *over*, qui modifie le sens de la relation *is*, alors que dans l'exemple positif, *13 million* est uniquement l'objet de la relation.

<i>Tokyo's population is over 13 Millions</i>	(Tokyo's population, is over, 13 Millions)
	(Tokyo's population, is, over 13 Millions)

Parfois, les relations peuvent être compliquées, mais nécessaires, alors qu'elles peuvent parfois être simplifiées, en gardant la partie optionnelle seulement si elle est impliquée par le manque d'autres tuples, comme dans le premier exemple où la partie *from Hungary* dans le deuxième cluster est optionnelle, car l'endroit d'origine de leur fuite est présent dans le premier cluster. Dans le second exemple *in Paris* n'est pas optionnel, car sans cette information, la relation ne tient plus.

<i>His parents are Ashkenazi Jews who had to flee from Hungary during World War II.</i>	(His parents, had to flee from, Hungary) (His parents, had to flee from Hungary during, World War II) (His parents, had to flee during, World War II)
<i>Chilly Gonzales (born Jason Charles Beck; 20 March 1972) is a Grammy-winning Canadian musician who resided in Paris, France for several years, and now lives in Cologne, Germany.</i>	(Chilly Gonzales, resided in Paris for, [several] years)

A.6 Coréférence

Aucune résolution de coréférence n'est faite à l'extérieur des phrases. Même si une phrase donnée vient d'un document qui permet de régler une coréférence, comme l'OIE se veut une tâche s'exécutant sur des phrases isolées, nous résolvons uniquement les coréférences d'entités comprises dans les phrases prises isolément. Des tuples usant de pronoms dont nous ne pouvons identifier l'élément de substitution peuvent sembler vides de sens, mais la résolution de coréférence doit se faire à l'extérieur de l'OIE, étant une tâche en soi. Dans l'exemple, on ne fait pas de résolution de coréférence pour le pronom *He*, comme aucune information à ce sujet n'est disponible dans la phrase. Par contre, on remplace *them* par *tax reductions* dans l'annotation.

<i>He did not go as far as he could have in tax reductions ; indeed he combined them with increases in indirect taxes .</i>	(He, combined them with, increases in indirect taxes) (He, combined tax reductions with, increases in indirect taxes)
---	--

A.7 Inférence

L'inférence est nécessaire, les faits directement impliqués par la phrase bien qu'ils ne soient pas exprimés verbatim sont des pièces d'information pertinente. Une nuance est ici nécessaire en ce qui a trait aux faits implicites potentiels. Ceux-ci ne sont pas obligatoirement impliqués par la phrase et doivent donc être omis. Dans le premier exemple, il est nécessairement impliqué par la phrase que Paul Johanson est le directeur des sciences de Monsanto. Par contre, le fait que leur vaporisateur soit doux sur l'organe féminin n'est pas nécessairement vrai, il est par contre vrai que cette information est dite par Paul Johanson.

<p><i>However , Paul Johanson , Monsanto 's director of plant sciences , said the company 's chemical spray overcomes these problems and is " gentle on the female organ .</i></p>	<p>(Paul Johanson, is, Monsanto's director of science)</p> <p>(Paul Johanson, says, the company's chemical spray is gentle on the female organ)</p> <p>(the company's chemical spray, is gentle on, the female organ)</p>
--	---

Il est ensuite nécessaire de faire une distinction entre l'inférence légère et l'inférence lourde. L'inférence légère est une forme d'inférence qui ne nécessite pas de réflexion logique par rapport à la phrase pour déduire le fait, celui-ci est simplement vrai tant que la phrase est elle aussi vraie. Dans le premier et le quatrième exemple, la relation est implicite, mais le fait annoté est évidemment vrai. L'inférence lourde nécessite pour sa part une partie de réflexion ou de combinaison d'opérations logiques pour impliquer le fait. Un cas d'inférence lourde est celui qui nécessite des **connaissances externes**, comme le deuxième exemple ou il est nécessaire de connaître la culture humaine et le principe d'hérédité pour faire l'inférence. Un autre exemple d'inférence lourde est la **généralisation**, comme dans le troisième exemple ou on implique un fait plus fort, une généralisation de ce qui est exprimé dans la phrase à l'aide d'un seul exemple. Un dernier exemple d'inférence lourde est celui des **limites inférieures ou supérieures**. Comme dans le cinquième exemple, on ne souhaite pas généraliser des limites inférieures ou supérieures sur des entités qui ne sont pas directement exprimées dans la phrase. On inclut donc à la référence les faits qui peuvent être inférés à l'aide d'inférence légère, mais pas ceux résultant d'une inférence lourde.

<p><i>Jason Charles Beck, a Jewish Canadian musician, was born in 1972.</i></p>	<p>(Jason Charles Beck, is, Jewish)</p>
<p><i>Gonzales is the son of Ashkenazi Jews who were forced to flee from Hungary during World War II.</i></p>	<p>(Gonzales, is, Jewish)</p>
<p><i>Gonzales is a McGill-trained virtuoso pianist.</i></p>	<p>(McGill, trains, pianists)</p>
<p><i>Nomura is amongst the more bullish.</i></p>	<p>(Nomura, is, bullish)</p>
<p><i>The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.</i></p>	<p>(Arthur Dempster, wrote, a classic paper)</p>
<p><i>The prefecture is part of the world's most populous metropolitan area with upwards of 37.8 million people and the world's largest urban agglomeration economy.</i></p>	<p>(the world, has more than, 37.8 million people)</p> <p>(the prefecture, has upwards of, 37.8 million people)</p>

Autre nuance, il faut inclure les faits même s'ils peuvent être inférés par une combinaison d'opérations logiques sur d'autres faits. En effet, il est généralement nécessaire

de connaître la langue et le fonctionnement du monde pour faire ce genre d'inférence, ce qu'on ne peut attendre d'un système d'OIE. Dans le premier exemple, le troisième cluster peut être deviné par une combinaison des deux premiers, mais on l'inclut tout de même, car cette opération de combinaison n'est pas attendue d'un système d'OIE. Dans le deuxième exemple, le fait d'être connu pour un album implique que c'est son propre album, on inclut néanmoins ce deuxième fait à la référence.

<p><i>The school was founded in 1851 as the German-English Academy by a group of Milwaukee German Americans that included educationist Peter Engelmann and hardware wholesaler William Frankfurth</i></p>	<p>(The school, was founded by, a group of Milwaukee German Americans)</p> <p>(a group of Milwaukee German Americans, included, Peter Engelmann)</p> <p>(The school, was founded by, Peter Engelmann)</p>
<p><i>Known for his albums of classical piano compositions with a pop music sensibility, Solo Piano I and Solo Piano II, as well as his MC and electro albums, he is also a producer and songwriter.</i></p>	<p>(He, is known for, Solo Piano 1)</p> <p>(Solo Piano 1, is, his album)</p>

Si la phrase implique que deux choses différentes peuvent être vraies, mais qu'il est impossible de savoir de laquelle il s'agit en lisant uniquement la phrase, on n'ajoute pas cette information à la référence. Dans l'exemple, *The Guardian* est soit un *newspaper* ou un *magazine*, impossible de savoir lequel selon la phrase. Le sujet, *He*, a également écrit au moins une pièce dans chacune des publications, mais impossible de savoir s'il s'agit de *newspaper* ou de *magazine*, ni combien d'entre elles il a écrit, on omet alors ces informations.

<p><i>He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.</i></p>	<p>(He, has written [several] magazine opinion pieces in, The Guardian)</p>
--	---

Certaines relations sont implicites et peuvent donc être formulées de plusieurs façons, à la liberté de l'annotateur, il est alors important de proposer différentes formulations en tâchant d'utiliser des relations fréquemment utilisées dans le langage courant et par les systèmes d'OIE.

A.8 Relations non verbales

Certaines relations pertinentes ne sont pas directement écrites dans le texte, même si leur information est bien présentée. Ces faits doivent être annotés et les mots qui composent la

relation doivent être ajoutés, en tentant d'utiliser une relation des plus simples, courante et qui décrit la relation de la manière la plus générale possible. Si plusieurs relations sont possibles et aussi probables, elles peuvent être listées dans le même cluster si leur sens est le même.

<i>1923 Kantō earthquake</i>	(Kantō earthquake, [happened in], 1923)
<i>The United Nations (UN)</i>	(UN, [stands for], The United Nations)

A.9 Reformulation

Si une relation ou un argument est exprimé de manière complexe dans le texte, on rajoute dans le même cluster une reformulation plus simple du même fait, et ce même si la relation dans les deux formulations n'est pas la même et que le niveau de détail peut être différent. Cela est un compromis entre le but de l'OIE de récolter toutes les informations factuelles exprimées dans le texte et l'importance de formuler ces faits dans un langage simple, qui est à la fois pertinent, mais pas nécessairement le but premier de l'OIE.

<i>Sam managed to convince John</i>	(Sam, managed to convince, John) (Sam, convinced, John)
<i>Gonzales broadcasts a web series Pop Music Masterclass on WDR, the documentary Classical Connections on BBC Radio 1, The History of Music on Arte, and Music's Cool with Chilly Gonzales on Apple Music's Beats1 radio show.</i>	(Gonzales, broadcasts Pop Music Masterclass on, WDR) (Pop Music Masterclass, is broadcasted on, WDR)

A.10 Voix active ou passive

Les clusters permettent de regrouper dans le même fait les formulations actives et passives d'une extraction. Même si une des deux formulations n'est pas présente textuellement, on doit l'ajouter si celle-ci est plus simple que la formulation originale de la phrase.

<i>The apple was eaten by Kyle</i>	(The apple, was eaten by , Kyle) (Kyle, ate , the apple)
------------------------------------	---

A.11 Argument unique

Certains faits n'ont pas de deuxième argument, on fait alors le choix de le laisser vide (on annote XXX pour le deuxième argument à des fins pratiques).

<i>From 1970 to 1985 , Gideon Rodan taught at the University of Connecticut School of Dental Medicine until he switched over to Merck .</i>	(Gideon Rodan, taught, XXX)
---	-----------------------------

A.12 Spéculation et attribution

Certaines informations dans le texte sont spéculatives ou attribuées à une entité, cette caractéristique doit alors être incluse dans la relation, de la manière dont c'est formulé dans la phrase. Cela permet de conserver cette information sans avoir à introduire une structure particulière. De plus, cette information doit être incluse dans la relation, car elle n'est en rien reliée aux arguments.

<i>An Apple Valley, California man plans on launching himself in a homemade rocket to prepare for obtaining evidence that the Earth is flat, according to The Washington Post.</i>	(The earth, is according to an Apple Valley man ,flat) (The earth, is, flat)
--	--

A.13 Correction

Parfois, certains tuples peuvent être formés de mots de la phrase originale, mais contenir des erreurs grammaticales. Il faut alors inclure le tuple formé des mots originaux ainsi que le tuple corrigé dans le même cluster. Cela permet de ne pas pénaliser les systèmes qui font la correction ni ceux qui ne la font pas.

<i>He has written several newspaper and magazine opinion pieces in The Guardian, Vice, Billboard, and others.</i>	(He, has written in, newspaper) (He, has written in, newspapers)
---	---

A.14 Type d'information spécifique

Les dates sont des informations spéciales, on choisit donc de les traiter un peu différemment en n'incluant qu'une seule extraction, la plus spécifique possible. On inclut la journée, le mois, l'année, l'heure, etc., si disponible. Pour les informations reliées à un endroit, si plusieurs spécifications existent (Continent, Pays, Ville, etc.) dans la phrase, on inclut une extraction par niveau de détail.

<i>He was born on september 26, 2017</i>	(He, was born on, september 26 2017) (He, was born in, september)
<i>He represented the riding of Nickel Belt in the Sudbury , Ontario area .</i>	(He, represented the riding in, the Sudbury area) (He, represented the riding in, Ontario)

A.15 Langue

Si certaines parties d'une phrase ou certaines informations sont exprimées dans des langues autres que celle de la référence d'origine, on omet d'inclure ces informations.

Annexe B - Lignes directrices de correspondance

Ce document contient les lignes directrices de correspondance pour la tâche d'extraction d'information ouverte et a été utilisée dans le développement de la fonction de correspondance de la référence *BenchIE^{FL}*. Les différents principes dictent les paires d'extractions faites par les systèmes et d'annotations qui doivent et ne doivent pas correspondre. Les exemples en vert sont des exemples qui correspondent à une annotation de la référence alors que les exemples en rouge ne correspondent pas. Dans ce document nous faisons référence aux faits annotés (Gold) en utilisant de manière interchangeable les mots *tuple*, *fait*, et *triplet*.

Nous cherchons ici à répondre à la question suivante : quand peut-on affirmer que deux extractions sont équivalentes d'un point de vue informatif ?

Les informations sont présentées selon le format suivant : les différentes formulations d'un même cluster (d'un même fait) sont dans un paragraphe et un saut de ligne les sépare.

<i>Phrase</i>	Cluster 1 Formulation 1 Gold Cluster 1 Formulation 2 Gold Cluster 2 Formulation 1 Gold etc. Extraction 1 Extraction 2 etc.
---------------	--

B.1 Extractions égales

Deux extractions égales en tout point doivent correspondre.

B.2 Spécificité de la relation

On accorde aux extractions très peu de souplesse dans la spécificité de la relation : c'est la relation qui est véhicule de l'information, il est alors important que celle-ci soit pratiquement aussi spécifique que celle de la référence. Cela étant dit, une formulation différente, mais aussi spécifique devrait être acceptée. Dans l'exemple, *was thrown* n'est pas une relation pertinente dans le contexte de cette extraction comme l'auraient été *was* ou *was thrown out of*.

<i>The Finns party was thrown out of the government and the new "Blue Reform" group kept its cabinet seat.</i>	(The Finns party, was , thrown out of the government)
	(The Finns party, was thrown out of , the government)
	(The Finns party, was thrown , out of the government)

B.3 Erreurs de syntaxe

Certaines extractions faites par les systèmes peuvent présenter des erreurs de syntaxe, quand un mot est mal placé ou inutile. Si cette erreur change le sens de la relation ou d'un des arguments, cette extraction doit alors être refusée. Sinon, on l'accepte. Dans l'exemple, le mot *also* fait référence à la relation *is*, et ne change pas le sens de celle-ci alors que le mot *and* modifie le sens de l'extraction.

<i>Known for his albums of classical piano compositions with a pop music sensibility, Solo Piano I and Solo Piano II, as well as his MC and electro albums, he is also a producer and songwriter.</i>	(He, is [also], a songwriter)
	(He, is also, a songwriter and)
	(He, is, a songwriter also)

B.4 Mots équivalents

Certains mots peuvent être équivalents à ceux présents dans les annotations dans certains contextes même si nous avons choisi de ne pas les inclure dans la référence. Si ces mots sont utilisés dans les extractions système à la place de ceux utilisés dans la référence, on accepte tout de même l'extraction système. Certains choix de mots peuvent être erronés, mais on accepte quand même l'extraction si le sens demeure. Dans le premier exemple, le déterminant *a* est utilisé à la place de *the* dans l'extraction, car c'est celui qui se trouve dans la phrase d'origine, cependant *the* est tout aussi voir plus approprié, on accepte donc l'extraction.

<i>He is the younger brother of the prolific film composer Christophe Beck.</i>	(He, is, a younger brother)
	(He, is , the younger brother)
<i>His parents are Ashkenazi Jews who had to flee from Hungary during World War II.</i>	(He , has, parents)
	(His , has, parents)

B.5 Niveau de détail

Une extraction système dont la relation correspond à celle d'une extraction de référence et qui présente dans un de ces arguments ce que l'on appelle un niveau de détail supérieur à celui de l'annotation doit être acceptée. On entend par niveau de détail que l'argument contient des informations présentes dans deux faits distincts. Ce principe peut sembler

arbitraire, mais en réalité il permet d'inclure dans la correspondance des extractions qui diffèrent dans la précision de l'un de leurs arguments, mais dont la relation et l'autre argument sont bien formulés. Ce choix est fait pour ne pas trop pénaliser les systèmes qui font des extractions longues. En effet, ces extractions comprennent tout de même des pièces d'information valides et comme les relations sont bien extraites, elles doivent être acceptées. Dans l'exemple, l'extraction acceptée possède un seul niveau de détail supplémentaire, car elle combine les informations de deux faits différents alors que l'extraction refusée combine trois faits dans son deuxième argument.

<p><i>Gonzales broadcasts a web series Pop Music Masterclass on WDR, the documentary Classical Connections on BBC Radio 1, The History of Music on Arte, and Music's Cool with Chilly Gonzales on Apple Music's Beats1 radio show.</i></p>	<p>(Gonzales, broadcasts, The History of Music) (Gonzales, broadcasts The History of Music on, Arte) (Gonzales, broadcasts, The History of Music on Arte) (Gonzales, broadcasts, a web series Pop Music Masterclass on WDR)</p>
--	--

B.6 Relations complètes

Les relations sont les véhicules de l'information, les arguments ne doivent pas contenir d'information qui changent le sens de la relation, sinon, elles doivent être refusées. Dans l'exemple, le deuxième argument de l'extraction contient le mot *over*, qui modifie le sens de la relation *is*, alors que dans l'annotation, *13 million* est uniquement l'objet de la relation.

<p><i>Tokyo's population is over 13 Millions</i></p>	<p>(Tokyo's population, is over, 13 Millions) (Tokyo's population, is, over 13 Millions)</p>
--	---

Annexe C - Exemples d'annotation et de correspondance

Cette annexe présente des exemples d'annotations de *BenchIE* et de *BenchIE^{FL}* sur des phrases de la référence, ainsi que les extractions des différents systèmes utilisés sur ces mêmes phrase, et la correspondance entre les extractions et les deux ensembles d'annotations selon la fonction de correspondance propre à la référence. Les colonnes de gauche indiquent les correspondance, le numéro faisant référence au cluster à laquelle l'extraction correspond. Les correspondance en bleu sont celles avec *BenchIE*, celle en vert avec *BenchIE^{FL}*. Ces exemples permettent de démontrer à la fois les différences dans les annotations et les nuances dans la fonction de correspondance utilisée.

He served as the first Prime Minister of Australia and became a founding justice of the High Court of Australia .		
ReVerb	Annotations BenchIE	Annotations BenchIE ^{FL}
1 2 He - served as - the first Prime Minister of Australia	1-> Cluster 1:	1-> Cluster 1:
3 4 He - became - a founding justice of the High Court of Australia	He --> served as --> [the] [first] Prime Minister [of Australia]	He --> served as --> [the] Prime Minister
	He --> served --> as [the] [first] Prime Minister [of Australia]	He --> served as [a] --> Prime Minister
	1-> Cluster 2:	He --> served as --> [a] Prime Minister
1 He - served - as the first Prime Minister of Australia	He --> served as [the] [first] Prime Minister --> Australia	1-> Cluster 2:
3 He - became - a founding justice of the High Court of Australia	He --> served as [the] [first] Prime Minister --> of Australia	He --> served as [the] first Prime Minister of --> Australia
4 He - became - a founding justice	1-> Cluster 3:	He --> served as --> [the] first Prime Minister of Australia
	He --> became --> [a] [founding] justice	1-> Cluster 3:
	He --> became --> [a] [founding] justice of [the] High Court [of Australia]	Australia --> has [had] --> [a] Prime Minister
2 2 He - served as first Prime Minister of - Australia	1-> Cluster 4:	Australia --> has [had] [a] --> Prime Minister
3 4 He - became - founding justice of High Court of Australia	He --> became [a] [founding] justice of --> [the] High Court [of Australia]	Australia --> had --> [a] Prime Minister
	He --> became [a] [founding] justice --> of [the] High Court [of Australia]	Australia --> had [a] --> Prime Minister
	1-> Cluster 5:	1-> Cluster 4:
1 He - served - as the first Prime Minister of Australia	He --> became [a] [founding] justice of [the] High Court of --> Australia	He --> became --> [a] founding justice
3 4 He - became - a founding justice of the High Court of Australia	He --> became [a] [founding] justice of [the] High Court --> of Australia	He --> became [a] --> founding justice
		1-> Cluster 5:
		He --> became [a] founding justice of --> [the] High Court of Australia
1 He - served - as the first Prime Minister of Australia		1-> Cluster 6:
3 4 He - became - a founding justice of the High Court of Australia		Australia --> has [had] --> [a] High Court
		Australia --> has [had] [a] --> High Court
		Australia --> had --> [a] High Court
		Australia --> had [a] --> High Court
		1-> Cluster 7:
1 He - served - as the first Prime Minister of Australia		[the] High Court of Australia --> has [had] --> [a] founding justice
3 4 He - became - a founding justice of the High Court of Australia		[the] High Court of Australia --> has [had] [a] --> founding justice
		[the] High Court of Australia --> had --> [a] founding justice
		[the] High Court of Australia --> had [a] --> founding justice

	In 1879 , he was re-elected U.S. Senator and was tipped as a Presidential candidate , but died suddenly after giving a speech in Chicago .				
	ReVerb		Anotations BenchIE		Annotations BenchIEFL
4 3	he - was re-elected - U.S. Senator		13--> Cluster 1:		13--> Cluster 1:
1 6	he - was tipped as - a Presidential candidate		he --> was tipped as --> [a] Presidential candidate		he --> was --> re-elected
			he --> was --> tipped as [a] Presidential candidate		he --> was --> re [-] elected
	ClausIE		he --> was tipped --> as [a] Presidential candidate		13--> Cluster 2:
	he - was re-elected - U.S. Senator In 1879		he --> was tipped In 1879 --> as [a] Presidential candidate		he --> [was] re-elected In --> 1879
4 3	he - was re-elected - U.S. Senator		he --> was tipped In 1879 as --> [a] Presidential candidate		he --> [was] re-elected in --> 1879
	he - died - In 1879		U.S. Senator --> was tipped as --> [a] Presidential candidate		he --> [was] re [-] elected In --> 1879
8	he - died - suddenly		U.S. Senator --> was --> tipped as [a] Presidential candidate		he --> [was] re [-] elected in --> 1879
	he - died - after giving a speech in Chicago		U.S. Senator --> was tipped --> as [a] Presidential candidate		he --> [was] re-elected [as] [a] U.S. Senator in --> 1879
	he - was tipped - as a Presidential candidate In 1879		U.S. Senator --> was tipped In 1879 --> as [a] Presidential candidate		he --> [was] re-elected [as] [a] U.S. Senator In --> 1879
1	he - was tipped - as a Presidential candidate		U.S. Senator --> was tipped In 1879 as --> [a] Presidential candidate		13--> Cluster 3:
			...		he --> [was] re-elected as [a] --> U.S. Senator
	MinIE		13--> Cluster 2:		he --> [was] re [-] elected as [a] --> U.S. Senator
3 2	he - was re-elected U.S. Senator In - 1879		he --> died [suddenly] after --> giving [a] speech [in Chicago]		he --> [was] re-elected as --> [a] U.S. Senator
	he - was - re-elected was tipped as Presidential candidate In 1879		he --> died [suddenly] after giving --> [a] speech [in Chicago]		he --> [was] re [-] elected as --> [a] U.S. Senator
	he - was - re-elected died suddenly after giving speech in Chicago In 1879		he --> died [suddenly] after --> giving [a] speech [in Chicago]		he --> [was] re-elected as [an] --> U.S. Senator
4 1	he - was - re-elected U.S. Senator		U.S. Senator --> died [suddenly] after --> giving [a] speech [in Chicago]		he --> [was] re [-] elected as --> [an] U.S. Senator
	he - was re-elected was tipped as - Presidential candidate		U.S. Senator --> died [suddenly] after giving --> [a] speech [in Chicago]		he --> [was] re-elected as --> [an] U.S. Senator
	he - was - re-elected died suddenly after giving speech in Chicago		...		he --> [was] re [-] elected as --> [an] U.S. Senator
			13--> Cluster 3:		he --> [was] re-elected --> U.S. Senator
	ImoIE		he --> was re-elected [U.S. Senator] In --> 1879		13--> Cluster 4:
3	he - was re-elected - U.S. Senator In 1879		he --> was re-elected [U.S. Senator] --> In 1879		he --> was [a] --> U.S. Senator
1	he - was tipped - as a Presidential candidate		he --> was re-elected [U. S. Senator] In --> 1879		he --> was [an] --> U.S. Senator
	he - died suddenly - after giving a speech in Chicago		U.S. Senator --> was re-elected In --> 1879		he --> was --> [a] U.S. Senator
			U.S. Senator --> was re-elected --> In 1879		he --> was --> [an] U.S. Senator
			...		13--> Cluster 5:
3	he - was re-elected - U.S. Senator In 1879		13--> Cluster 4:		he --> was U.S. Senator In --> 1879
	he - after giving - a speech in Chicago		he --> was re-elected --> U.S. Senator		he --> was U.S. Senator in --> 1879
	he - died suddenly - after giving a speech in Chicago In 1879		he --> was re [-] elected --> U.S. Senator		13--> Cluster 6:
	he - was tipped - as a Presidential candidate In 1879		he --> was --> re-elected U.S. Senator		he --> was tipped as --> [a] Presidential candidate
			he --> was --> re [-] elected U.S. Senator		he --> was tipped as [a] --> Presidential candidate
			he --> was re-elected In 1879 --> U.S. Senator		13--> Cluster 7:
			he --> was re [-] elected In 1879 --> U.S. Senator		he --> was tipped as [a] Presidential candidate In --> 1879
1	he - was tipped - as a Presidential candidate		...		he --> was tipped as [a] Presidential candidate in --> 1879
	he - died - suddenly a speech in		13--> Cluster 5:		13--> Cluster 8:
			he --> was tipped as [a] Presidential candidate In --> 1879		he --> died --> suddenly
			he --> was tipped as [a] Presidential candidate --> In 1879		13--> Cluster 9:
4 3	he - was re-elected - U.S. Senator		U.S. Senator --> was tipped as [a] Presidential candidate In --> 1879		he --> died In --> 1879
1	he - was tipped - as a Presidential candidate		U.S. Senator --> was tipped as [a] Presidential candidate --> In 1879		he --> died in --> 1879
	he - died suddenly - after giving a speech in Chicago		...		13--> Cluster 10:
			13--> Cluster 6:		he --> died after --> giving [a] speech
			he --> giving --> [a] speech		13--> Cluster 11:
			he --> giving --> [a] speech in Chicago		he --> [was] giving [a] speech in --> Chicago
			U.S. Senator --> giving --> [a] speech		he --> gave [a] speech in --> Chicago
			U.S. Senator --> giving --> [a] speech in Chicago		
			...		
			13--> Cluster 7:		
			he --> giving [a] speech in --> Chicago		
			he --> giving [a] speech --> in Chicago		

Vaccinations against other viral diseases followed , including the successful rabies vaccination by Louis Pasteur in 1886 .			
ReVerb		Annotations BenchIE	
		23--> Cluster 1:	
ClausIE		Vaccinations [against other viral diseases] --> including --> [the] [successful] rabies vaccination	
Vaccinations against other viral diseases		Vaccinations [against other viral diseases] --> including --> [the] [successful] rabies vaccination by Louis Pasteur	
followed		Vaccinations [against other viral diseases] --> including --> [the] [successful] rabies vaccination in 1886	
including the successful rabies vaccination by Louis Pasteur in 1886		Vaccinations [against viral diseases] --> including --> [the] [successful] rabies vaccination	
		Vaccinations [against viral diseases] --> including --> [the] [successful] rabies vaccination by Louis Pasteur	
MinIE		Vaccinations [against viral diseases] --> including --> [the] [successful] rabies vaccination in 1886	
Vaccinations against other viral diseases		23--> Cluster 2:	
followed		Vaccinations [against other viral diseases] --> including [the] [successful] rabies vaccination [in 1886] by --> Louis Pasteur	
including the successful rabies vaccination by Louis Pasteur in 1886		Vaccinations [against other viral diseases] --> including [the] [successful] rabies vaccination [in 1886] --> by Louis Pasteur	
		Vaccinations [against viral diseases] --> including [the] [successful] rabies vaccination [in 1886] by --> Louis Pasteur	
ImojIE		Vaccinations [against viral diseases] --> including [the] [successful] rabies vaccination [in 1886] --> by Louis Pasteur	
Vaccinations against other viral diseases		23--> Cluster 3:	
followed		Vaccinations [against other viral diseases] --> including [the] [successful] rabies vaccination [by Louis Pasteur] in --> 1886	
including the successful rabies vaccination by Louis Pasteur in 1886		Vaccinations [against other viral diseases] --> including [the] [successful] rabies vaccination [by Louis Pasteur] --> in 1886	
		Vaccinations [against viral diseases] --> including [the] [successful] rabies vaccination [by Louis Pasteur] in --> 1886	
OpenIE6		Vaccinations [against viral diseases] --> including [the] [successful] rabies vaccination [by Louis Pasteur] --> in 1886	
M2OIE		Annotations BenchIEFL	
Vaccinations against other viral diseases		23--> Cluster 1:	
followed		Vaccinations against other viral diseases --> followed --> XXX	
including the successful rabies		23--> Cluster 2:	
		Vaccinations against other viral diseases --> included --> [the] [successful]rabies vaccination	
CompactIE		Vaccinations against other viral diseases --> include --> [the] [successful]rabies vaccination	
Vaccinations against other viral diseases		Vaccinations against other viral diseases --> [are] including --> [the] [successful] rabies vaccination	
followed		23--> Cluster 3:	
including the successful rabies vaccination by Louis Pasteur in 1886		[the] [successful] rabies vaccination --> is by --> Louis Pasteur	
		[the] [successful] rabies vaccination --> was [done] by --> Louis Pasteur	
		Louis Pasteur --> did --> [the] [successful] rabies vaccination	
		23--> Cluster 4:	
		[the] [successful] rabies vaccination --> was [done] in --> 1886	
		[the] [successful] rabies vaccination --> happened in --> 1886	
		23--> Cluster 5:	
		[the] rabies vaccination --> was --> successful	
		[the] rabies vaccination --> is --> successful	
		23--> Cluster 6:	
		[the] [successful] rabies vaccination --> followed --> XXX	
		23--> Cluster 7:	
		rabies --> is [a] --> viral diseases	
		rabies --> is --> [a] viral diseases	
		rabies --> was [a] --> viral diseases	
		rabies --> was --> [a] viral diseases	

Arminius , leader of the Cherusci and allies , now had a free hand .			
ReVerb		Annotations BenchIE	Annotations BenchIEFL
4 Arminius - now had - a free hand		27--> Cluster 1:	27--> Cluster 1:
		Arminius --> had [now] --> [a] free hand	Arminius --> is [a] --> leader
ClausIE		leader of [the] Cherusci --> had [now] --> [a] free hand	Arminius --> is --> [a] leader
Arminius - is - leader of the Cherusci		leader of [the] allies --> had [now] --> [a] free hand	Arminius --> is [the] --> leader
Arminius - is - leader of allies			Arminius --> is --> [the] leader
Arminius - had - a free hand now			27--> Cluster 2:
1 4 Arminius - had - a free hand			Arminius --> is [the] leader of --> [the] Cherusci
			27--> Cluster 3:
MinIE			Arminius --> is [the] leader of --> [the] [Cherusci] allies
2 Arminius - is leader of - Cherusci			Arminius --> is [the] leader of --> [the] allies of Cherusci
3 Arminius - is leader of - allies			27--> Cluster 4:
1 4 Arminius - had - free hand			Arminius --> [now] had --> [a] free hand
			Arminius --> [now] had [a] --> free hand
ImojIE			
Arminius - had - a free hand now			
2 Arminius - is leader of - the Cherusci and allies			
OpenIE6			
Arminius - had - a free hand now			
2 Arminius - is leader of - the Cherusci			
3 Arminius - is leader of - the allies			
M2OIE			
Arminius , leader of the Cherusci and allies - had - a free hand			

