# Chapitre 3) Qu’est ce qu’une relation saine et stable en NLP ?

L'extraction d'information (IE) en traitement automatique du langage (TAL) désigne un ensemble de techniques visant à analyser et à comprendre les relations entre les entités mentionnées dans un texte. L'objectif fondamental est de dévoiler les structures sous-jacentes pour, par la suite, permettre l'interrogation précise des liens entre ces entités dans de vastes bases de données, ainsi que la construction de graphes de connaissances ou d'ontologies (cf. chapitre 5).

Au fil du temps, cette discipline s'est affinée pour donner naissance à des tâches spécialisées, dont l'extraction de relations (Relationship Extraction, RE), l’extraction d’événements, pour lesquelles l’état de l’art est encore décevant. Contrairement à la reconnaissance d'entités nommées (NER), largement étudiée et appliquée, l'extraction de relations se concentre sur l'identification des liens sémantiques entre entités, constituant ainsi une sous-catégorie spécifique de l'information extraction. Cette approche, bien que plus ardue, offre un potentiel considérable pour enrichir les systèmes d'information et améliorer l'analyse sémantique des documents non structurés. Néanmoins le manque de corpus annotés et de consensus sur la définition d’une relation rend difficile la création de systèmes répondant à tous les cas d’usage. La démocratisation des LLMs promet une simplification des tâches d’extraction d’information depuis les textes structurés, mais interroge sur la pertinence d’employer des moyens aussi coûteux en ressource pour une tâche dont la définition est assez simple.

Comme dans la vraie vie, se mettre d’accord sur les relations est difficile en NLP. On étudiera d’abord les différentes définitions d’une relation proposées dans le temps avant de présenter une chronologie des méthodes employées pour la détection de relations ces 30 dernières années.

## 1) Qu’est ce qu’une relation ?

Pour identifier une relation dans un texte, il est essentiel de définir précisément ce qu'est une relation et ce qu'elle est censée décrire. Ce défi peut être décomposé en plusieurs questions fondamentales, chacune pouvant être abordée indépendamment :

- **Qu'est-ce qu'une relation ?**
  Dans le contexte du traitement automatique du langage naturel , une relation est un lien sémantique entre deux entités mentionnées dans un texte. Par exemple, dans la phrase "Frodon Saquet a trouvé l’anneau", la relation "trouver" relie les entités "Frodon" et "L’Anneau".
- **Quand peut-on affirmer qu'il y a une relation entre deux entités (portée) ?**
  La détection d'une relation dépend de la proximité contextuelle des entités et de la présence d'indices linguistiques indiquant un lien. Des techniques telles que l'analyse syntaxique et sémantique sont utilisées pour déterminer si une relation explicite ou implicite existe entre les entités.
- **Que signifie la relation (sémantisme de l'objet créé) ?**
  Comprendre le sens d'une relation implique d'interpréter le type de lien qui unit les entités. Cela nécessite une analyse approfondie du contexte et peut s'appuyer sur des ressources telles que des ontologies ou des bases de connaissances pour attribuer une signification précise à la relation identifiée.

- **Comment représenter une relation ?**
  Les relations peuvent être représentées sous forme de triplets (sujet, prédicat, objet), facilitant ainsi leur intégration dans des graphes de connaissances ou des bases de données relationnelles. Par exemple, le triplet ("Frodon", "porte", "l'Anneau").
- **Comment exploiter les relations qu'on repère ?**
  Les relations extraites peuvent être utilisées pour enrichir des graphes de connaissances, améliorer des systèmes de question-réponse, ou encore pour l'analyse sémantique de textes. Elles permettent de structurer l'information contenue dans des documents non structurés, facilitant ainsi son exploitation ultérieure.

La littérature actuelle n'apporte pas de réponses définitives à ces questions, ce qui complique la mise en place de systèmes pour les repérer. Des approches récentes, telles que l'extraction ouverte de relations, cherchent à identifier des relations sans les définir au préalable, mais elles posent également des défis en termes de précision et de pertinence des relations extraites

## a) Définition formelle

Une relation désigne techniquement avant tout un lien entre deux entités.
Ce qui implique déjà d'avoir défini précisément auparavant ce que l'on considère comme une entité possible (domaine de définition). Définition restrictive du TAL (Personne, Localisation, Organisation…, les sous-classes du NER) ou de façon plus large (incluant les dates, mais aussi les concepts abstraits) comme on a besoin de le faire dans la construction d'ontologies.
Le plus souvent, les relations sont notés comme un triplet type RDF (from, REL, to).

Traditionnellement en NLP, on a considéré trois variantes de l'extraction d'information, de la plus générique à la plus spécifique :
L'essentiel du débat porte sur la définition du contenu de cette relation.

**1) La simple cooccurrence ou connexion**

Une relation peut simplement indiquer l'existence d'un lien entre des entités, mais il est souvent essentiel de comprendre la nature de ce lien ou d'identifier le fragment de texte qui l'atteste. Par exemple, dans la phrase « Minas Tirith se situe dans le Gondor », on peut noter :

- (Minas Tirith, Gondor) : une relation existe entre les deux, les concepts sont liés.

Conserver toute la phrase est facile, mais beaucoup moins informatif. Cela oblige l'utilisateur à deviner la nature exacte du lien. Certaines relations sont plus importantes que d'autres selon le cas d'usage, et ne permettent aucun requêtage ni hiérarchisation de l'information ou du degré de connexion entre les entités.

De plus, si une phrase contient n entités, sans trier celles qui sont effectivement en relation pertinente, on se retrouve avec (n-1)! relations deux à deux, dont certaines peu pertinentes. Les entités présentes dans deux subordonnées indépendantes d'une phrase ont très peu de chances d'être sémantiquement reliées.

Un tri des relations est donc nécessaire ; autant dès lors tenter de les caractériser.

**2) les relations spécifiques**

Une relation spécifique peut être définie en précisant le lien exact entre les entités, souvent en mentionnant l'extrait de texte qui les relie syntaxiquement, généralement le pivot verbal. Par

exemple, dans la phrase « Minas Tirith se situe dans le Gondor », la relation « se_situe_dans » établit le lien entre « Minas Tirith » et « Gondor ».

Cependant, il arrive que le texte ne mentionne pas explicitement la relation, ce qui complique l'extraction de l'information. Prenons l'exemple de l'ellipse syntaxique dans la phrase « Minas Tirith, tout comme l'ancienne capitale Osgiliath, se situe dans le Gondor ». Ici, le lien entre « Minas Tirith » et « Osgiliath » est implicite, ce qui rend sa détection plus complexe pour un système automatisé, bien qu'un humain puisse facilement reconstituer le lien.

D'autres types de relations implicites incluent :

- **Lien d'apposition** : « Bilbo, le hobbit » implique que « Bilbo est un hobbit ».
- **Lien de fonction** : « Théoden, roi du Gondor, a lancé l'assaut » indique que « Théoden est le roi du Gondor ».
- **Lien de possession exprimé indirectement** : « La forteresse de Sauron » suggère que « Sauron possède une forteresse ».

Ces relations, bien que cruciales, ne sont pas toujours explicitement présentes dans le texte, ce qui rend l'extraction de relations plus complexe que le simple parsing de tokens connectant des entités.

Pour aborder ces défis, l'Open Information Extraction (OpenIE) a été développé. Cette approche vise à extraire des relations sans s'appuyer sur des schémas ou des catégories prédéfinies, permettant ainsi de découvrir des relations implicites. Elle est « open » car elle ne nécessite pas de connaissances préalables sur le domaine et est conçue pour être généralisée sur de vastes corpus.

Des recherches récentes ont exploré l'extraction de relations implicites dans des ensembles de données de compréhension de lecture, en développant des outils basés sur des analyses syntaxiques pour convertir ces ensembles en données exploitables pour l'OpenIE. Ces approches ont permis de créer des modèles neuronaux capables d'extraire des relations implicites avec une précision accrue, surpassant les méthodes précédentes.
En somme, bien que l'extraction de relations implicites présente des défis techniques considérables, des approches comme l'OpenIE offrent des solutions prometteuses pour surmonter ces obstacles et enrichir les bases de connaissances à partir de textes non structurés.

## 3) Les relations prédéfinies

L'extraction d'informations standard (Standard Information Extraction) en traitement du langage naturel (NLP) vise à identifier et à structurer des données spécifiques à partir de textes non structurés. Dans des domaines spécialisés, tels que le secteur médical, l'intérêt se porte souvent sur des relations très précises, comme les effets secondaires d'un médicament, son utilisation pour traiter une maladie particulière ou ses contre-indications avec d'autres substances. Dans ces contextes, de nombreuses relations lexicales générales ne sont pas pertinentes.

Pour répondre à ces besoins spécifiques, il est plus pertinent de se concentrer sur des relations stéréotypées, également appelées « relations taxonomiques ». Ces relations sont souvent représentées par des noms composés, séparés par des underscores, tels que « is_located_in » (est situé dans) ou « treats_disease » (traite la maladie). Cette approche est couramment adoptée dans la construction d'ontologies et de graphes de connaissances.

Les ontologies fournissent un cadre formel pour définir les concepts, les relations et les attributs au sein d'un domaine spécifique, créant ainsi un vocabulaire commun pour la communication et la collaboration. Les graphes de connaissances, quant à eux, utilisent ces ontologies comme base pour

représenter les données sous forme de réseau d'entités interconnectées, facilitant ainsi l'intégration et l'analyse de données complexes et hétérogènes.

Cependant, cette approche présente des limitations. En s'appuyant uniquement sur des sources de données externes comme DBpedia ou WordNet, on peut uniquement identifier des relations déjà attestées ailleurs, sans la possibilité de découvrir de nouvelles relations spécifiques au corpus analysé. Cette méthode est donc limitée à la reconnaissance de relations préexistantes et ne permet pas d'explorer des relations inédites ou contextuellement spécifiques.

Table 1 ACE05 Entity Types and Subtypes

| Type | Subtypes |
|---|---|
| FAC (Facility) | Airport, Building-Grounds, Path, Plant, Subarea-Facility |
| GPE (Geo-Political Entity[3]) | Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province |
| LOC (Location) | Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body |
| ORG (Organization) | Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports |
| PER (Person) | Group, Indeterminate, Individual |
| VEH (Vehicle) | Air, Land, Subarea-Vehicle, Underspecified, Water |
| WEA (Weapon) | Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified |

Table 6 ACE05 Relation Types and Subtypes
(Relations marked with an * are symmetric relations.)

| Type | Subtype |
|---|---|
| ART (artifact) | User-Owner-Inventor-Manufacturer |
| GEN-AFF (Gen-affiliation) | Citizen-Resident-Religion-Ethnicity, Org-Location |
| METONYMY* | *none* |
| ORG-AFF (Org-affiliation) | Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership |
| PART-WHOLE (part-whole) | Artifact, Geographical, Subsidiary |
| PER-SOC* (person-social) | Business, Family, Lasting-Personal |
| PHYS* (physical) | Located, Near |

Les « méta relation » sont à définir pour chaque domaine, rendant impossible l’interopérabilité du système. Voir ici par exemple les relations définies pour ACE(2008) pour un cas business. Mais la plupart d’applications logicielles spécifiques ressemblent à ça…

L'extraction de relations en traitement du langage naturel (NLP) est souvent abordée comme une tâche de classification multiclasse, où chaque classe correspond à un type de relation spécifique, avec une classe supplémentaire pour l'absence de relation. Alternativement, elle peut être décomposée en plusieurs tâches de classification binaire, chacune déterminant la présence ou l'absence d'un type particulier de relation entre deux entités.

Une approche spécifique, appelée **Binary Relationship Extraction**, se concentre sur l'identification des participants à une relation précise. Cette méthode simplifie techniquement la tâche d'extraction de relations (ER) en se limitant à des relations bien définies et souvent binaires. Cependant, elle constitue une version plus restreinte de l'extraction de relations, car elle ne permet pas de découvrir de nouvelles relations ou de traiter des relations complexes impliquant plusieurs entités.

Cette approche nécessite une connaissance préalable du cas d'utilisation et/ou du corpus, car elle repose sur des relations prédéfinies. Par conséquent, elle est particulièrement adaptée aux domaines où les types de relations sont bien établis et limités, comme dans certaines applications biomédicales ou juridiques.

spécifiques dans des domaines spécialisés, elle présente des limitations en termes de découverte de nouvelles relations. L'adoption d'approches comme l'OpenIE permet de surmonter ces défis en offrant une flexibilité accrue et la possibilité d'explorer des relations non préalablement définies.sur lequel elle sera appliquée. Elle nécessite un tuning précis et n’est pas générique.

## b) la question de l’arité des relations

Jusqu'à présent, nous avons principalement abordé des relations reliant deux entités, représentées sous forme de triplets. Cependant, cette structure tripartite présente des limitations, notamment lorsqu'il s'agit de décrire des actions impliquant plus de deux participants, comme c'est souvent le cas avec les verbes intransitifs.
Prenons l'exemple de la phrase « Bilbo donne l'Anneau à Frodon ». Représenter cette action par un triplet unique, tel que (Bilbo, donne l'Anneau à, Frodon), introduit une relation spécifique par objet donné, ce qui manque de généralité. De plus, cette représentation ne capture pas la relation intrinsèque du don qui lie les trois participants.

Pour surmonter ces limitations, plusieurs approches peuvent être envisagées :

1. **Utilisation de tuples plus complexes** : On pourrait envisager des structures à quatre éléments, comme (Bilbo, donne, l'Anneau, Frodon). Cependant, cette approche peut devenir rapidement complexe et difficile à généraliser, surtout lorsqu'il s'agit de représenter des actions impliquant un nombre variable d'actants. Doit on inclure le complément de moyen ? De temps ? De lieux ? Si on tolère les quadruplets, alors pourquoi pas les relations d’arité n ? Donc on ne s’arrête plus...
2. **Enrichissement des triplets avec des cadres sémantiques** : Une autre solution consiste à intégrer des informations supplémentaires, telles que les rôles sémantiques des actants, en utilisant des cadres sémantiques (semantic frames) comme ceux proposés par Fillmore. Chaque relation aurait un patron à suivre qui nous indique de combien d’actants on a besoin et quel est le rôle de chacun des actants. Cela permettrait de préciser le rôle de chaque participant dans l'action, mais cela nécessite une analyse sémantique approfondie et peut être complexe à mettre en œuvre de manière automatisée. Il faudrait de plus gérer le cas des verbes qui ne se comportent pas tout le temps de la même façon (ex : il pleut verbe avalent vs il pleut des cordes, emploi transitif imagé, et la cascade pleut sur moi, usage aussi poétique.) et aussi gérer le cas d’expressions jamais rencontrées avant dans le corpus.

3. **Décomposition de la relation en sous-relations** : Une approche consiste à décomposer la relation complexe en plusieurs relations plus simples. Par exemple, (Bilbo, donne, l'Anneau) et (l'Anneau, est destiné à, Frodon). Cependant, cette décomposition dépend du sémantisme du verbe et peut ne pas toujours être triviale. De plus, elle peut entraîner une perte d'information sur la chaîne de relations entre les actants.

Il est également important de noter que certains verbes, dits **bitransitifs**, peuvent être accompagnés de deux compléments, comme dans « Bilbo donne l'Anneau à Frodon ». Dans ce cas, l'action implique trois participants : l'auteur de l'action (Bilbo), l'objet de l'action (l'Anneau) et le destinataire (Frodon). Représenter cette relation par un triplet classique est insuffisant, car il ne capture pas la complexité de l'interaction entre les trois entités.

En somme, bien que la structure tripartite soit utile pour représenter des relations simples, elle est insuffisante pour décrire des actions complexes impliquant plusieurs actants. Des approches plus sophistiquées, intégrant des informations sémantiques et des structures relationnelles plus complexes, sont nécessaires pour une représentation fidèle de ces interactions.

Cette question renvoie en fait à une question de syntaxe largement discutée depuis 50 ans.
La notion de **valence verbale**, introduite par Lucien Tesnière dans son ouvrage *Éléments de syntaxe structurale* (1959)[1], décrit la capacité d’un verbe à se combiner avec un certain nombre d’actants

1 Pour plus d’informations sur ce point : https://questionsdelangue.wordpress.com/2021/06/18/la-valence-verbale-une-introduction/

(ou arguments) nécessaires à la construction de son sens. Ce concept est central dans l'analyse syntaxique et sémantique, puisqu'il permet de modéliser la structure des phrases en fonction des propriétés intrinsèques des verbes.

Types de valence :

* Valence zéro (verbes avalents)
Ces verbes n'exigent aucun actant. Ils se limitent à des phénomènes naturels ou impersonnels. Exemples : *Il pleut, Il neige*.

* Valence un (verbes monovalents)
Ces verbes nécessitent un seul actant, généralement le sujet. Exemples : *Bilbo court, Frodon dort.*

* Valence deux (verbes bivalents)
Ces verbes requièrent deux actants, généralement un sujet et un objet direct ou un complément d'objet. Exemples : *Bilbo lit un livre, Frodon mange une pomme*.

* Valence trois (verbes trivalents)
Ces verbes impliquent trois actants, typiquement un sujet, un objet direct, et un objet indirect ou un complément circonstanciel. Exemples : *Bilbo donne l'Anneau à Frodon.*

Cependant, des constructions linguistiques plus complexes, telles que les expressions à verbe support et les périphrases verbales, illustrent les limites des triplets RDF pour décrire la réalité linguistique. Ces structures nécessitent des représentations plus nuancées pour capturer la richesse des relations entre les actants et les actions décrites.

Pire encore, la notion de valence s'applique également à d'autres classes de mots tels que les noms et les adjectifs. Des noms, notamment ceux issus d'un verbe (déverbaux) ont une valence similaire à un verbe auquel il son rattaché. Ex :

*Bilbo donne l'Anneau à Frodon => Le don de l'Anneau de Bilbo à Frodon*

Mais du coup par calque => *Le cadeau de l'Anneau de Bilbo à Frodon.*

Doit-on extraire de ce genre de groupes nominaux un triplet quand on fait de l'IE ? Je pense que oui.

Même chose pour des adjectifs : *apte à combattre les orques. Généreux de ses richesses…*

réduire la notion d'information à la structure basique du groupe verbale sujet/prédicat est limitant. C'est aussi une manière de raisonner influencée par le fonctionnement de nos langues indos européennes dans lequel le verbe est l'élément pivot.

## c) Les relations par l'usage

La définition d'une « bonne » relation dépend largement de son contexte d'utilisation. Dans les systèmes experts, où la précision et la spécificité sont cruciales, les relations sont souvent très détaillées et adaptées à des cas particuliers. À l'inverse, dans un web ouvert, où l'objectif est la généralité et l'interopérabilité, les relations tendent à être plus larges et flexibles.

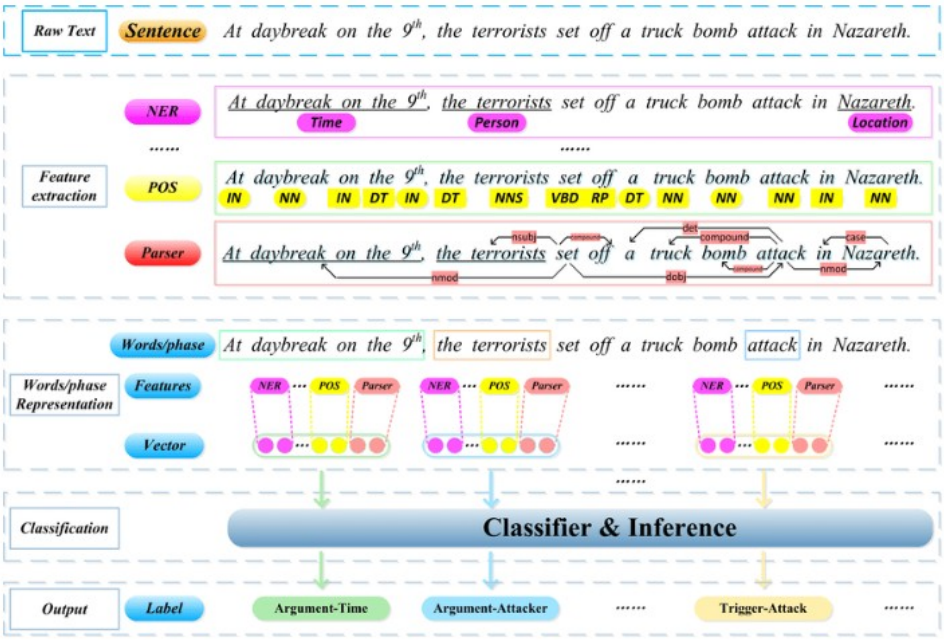
De plus, plusieurs types spécifiques de relations ont donné naissance à des sous-domaines distincts dans le champ de l'Information Retrieval (IR), chacun ayant ses propres exigences et méthodologies adaptées à des besoins particuliers :

*** L’extraction d’événements :**

L'extraction d'événements (event extraction) est une tâche bien explorée du NLP, consistant à identifier et à classer des actions ou des occurrences mentionnées dans un texte. Un événement est généralement défini comme une action se produisant à un moment et en un lieu donnés, impliquant des participants spécifiques. Contrairement à l'extraction de relations, qui se concentre sur les liens entre entités, l'extraction d'événements se focalise sur le verbe ou le prédicat central de la phrase, souvent associé à des marqueurs temporels. On peut voir un événement comme un duo action verbale particulière + marqueur temporel, ou comme une relation + marqueur temporel

Par exemple, dans la phrase «  Eowyn a tué le roi sorcier lors de la bataille des champs de Pélennor », l'événement est le combat, avec « Eowyn » comme acteur, « le roi sorcier » comme objet, et « lors de la bataille des champs de Pélennor » comme date.
L'extraction d'événements implique plusieurs étapes :

1. **Détection de l'événement** : Identifier la présence d'un événement dans le texte, souvent en repérant des déclencheurs tels que des verbes ou des noms déverbaux.
2. **Extraction des arguments** : Déterminer les entités associées à l'événement et leurs rôles respectifs (par exemple, acteur, objet, lieu, temps).
3. **Classification de l'événement** : Attribuer une catégorie à l'événement identifié, comme un lancement de produit, une acquisition, un incident, etc.




Cette tâche est complexe en raison de la variabilité linguistique et de la nécessité de comprendre le contexte pour interpréter correctement les événements. Elle est particulièrement utile dans des domaines tels que la politique ou la géopolitique, où la dimension temporelle des événements permet d'effectuer des analyses approfondies et de les représenter sur des frises chronologiques. Par exemple, suivre l'évolution des relations diplomatiques entre des pays en identifiant des événements clés tels que des accords, des conflits ou des visites officielles.

La détection d’évènements peut être prise en charge par un pipeline[2] de techniques NLP telles que la reconnaissance d'entités nommées, l'analyse syntaxique et sémantique, puis finalement un système de ML, ou bien par un système unique.

2 https://www.researchgate.net/publication/337638438_A_Survey_of_Event_Extraction_From_Text

*** L’extraction de relations causales.**

L’extraction de relations causales vise à identifier et structurer les relations de cause à effet exprimées dans des textes. Par exemple, dans la phrase *"L’industrialisation de la Compté a créé un cataclysme écologique"*, une relation causale entre *industrialisation* (cause) et *cataclysme écologique* (effet) est implicite. Cette tâche est essentielle pour comprendre les liens entre les événements et les entités dans des domaines variés.

Les applications de l’extraction de relations causales sont nombreuses :

Analyse de risques : Identifier les causes d'incidents dans des rapports industriels ou médicaux.

-Prédiction d’événements : Aider à anticiper des conséquences basées sur des causes identifiées dans des données économiques, climatiques, ou sociales.

-Aide à la décision : Enrichir les systèmes d’aide à la décision en offrant une compréhension des relations causales sous-jacentes à des phénomènes complexes.

-Recherche scientifique : Accélérer l’analyse des articles scientifiques en identifiant les relations causales dans des expériences ou des hypothèses.

La capacité à extraire ces relations de manière fiable est un défi clé pour le développement de systèmes experts.

L'extraction d'information (EI) joue un rôle essentiel dans le processus de conversion de données non structurées en données structurées, et permet la mise à jour et l'enrichissement des bases de connaissances. Cette tâche est particulièrement cruciale pour l'identification et la mise à jour d'informations concernant des entités telles que des personnes physiques ou morales (entreprises, organisations). Par exemple les ministres changent tout le temps…

Dans de nombreuses bases de données, il est courant que des informations essentielles soient manquantes (par exemple, la date de naissance) ou obsolètes (telles que l'adresse). Pour maintenir ces bases de données à jour, il est important de les alimenter en continu à partir de flux de données non structurées, c'est-à-dire du texte brut provenant de diverses sources.

L'extraction d'information, et plus spécifiquement l'extraction ouverte d'information (Open IE), devient alors cruciale pour la conception et la mise à jour de bases de connaissances ou de bases de données relationnelles (voire chapitre 4). Ces bases constituent un support précieux pour de nombreuses applications, notamment dans les domaines de la politique ou de la géopolitique, où la dimension temporelle des événements permet d'effectuer des analyses approfondies et de les représenter sur des frises chronologiques.

Par ailleurs, l'EI peut également être utilisée pour le remplissage automatique de formulaires (template filling), permettant, par exemple, de pré-remplir automatiquement des documents administratifs ou comptables, automatisant ainsi des tâches humaines répétitives.

En d'autres termes, l'extraction d'information est aujourd'hui une tâche essentielle pour convertir la masse de données non structurées en données structurées, permettant ainsi de fusionner les connaissances issues de sources diverses et d'améliorer l'efficacité des systèmes d'information.

## 2) Défilé historique des méthodes d’extraction de relations

L'extraction d'information (EI) a considérablement évolué au fil des avancées en traitement du langage naturel. Selon le rapport "A Survey on Open Information Extraction from Rule-based Model to Large Language Model"[3] publié en octobre 2024, l'EI est passée de méthodes basées sur des règles à des approches neuronales, puis à l'intégration de grands modèles de langage (LLM). Cette évolution a permis d'améliorer la précision et la portée des systèmes d'EI, en les rendant capables de traiter des relations plus complexes et variées. Les LLM, en particulier, ont introduit une nouvelle ère d'extraction d'information générative, facilitant l'extraction de triplets relationnels à partir de textes non structurés. Cette progression reflète l'adaptation continue de l'EI aux technologies émergentes, renforçant son rôle dans des applications telles que le questionnement automatique, les moteurs de recherche et l'enrichissement des graphes de connaissances.

### a) Méthodes reposant sur les règles et les patterns

Utiliser des patrons lexico-syntaxiques est la technique la plus ancienne qui ait été utilisée pour repérer des relations dans des textes depuis les années 1990. On appelle aussi parfois cette méthode « Heast patterns » du nom de l’auteur d’un papier important de 199[4] ayant présenté cette approche.

Les patrons utilisés peuvent permettre de décrire des relations stéréotypées telle que l’hyperonymie ou la localisation, ou des patterns plus génériques (comme l’action d’un sujet sur l’objet).
On peut créer des patrons à partir de plusieurs éléments :
* des expressions lexicales (« X par exemple Y », « X, et pas Y »)… parfois conditionnées par la catégorie grammaticale des éléments.
* Sur l’identité du mot.
* Sur des suites de catégorie grammaticales
* Sur les chaînes de type d’entités (Pers Orga Lieu => Orga is_located_in Lieu)
* Sur la structure syntaxique de la phrase SVO => (S, V, O). evidemment ce n’est pas aussi simple que ça à cause de tous les problèmes linguistiques posés par la définition du groupe verbal…

| Type de Relation | Extrait | Patron Utilisé | Relation |
|---|---|---|---|
| **Hyperonymie** | "les terres de la Terre du Milieu, telles que le Gondor, le Rohan et le Mordor" | NPH such as {NP,}* {(or | and)} NP |
| **Localisation** | "où Frodo transporte l'Anneau Unique jusqu'à la Montagne du Destin" | Fonction syntaxique : locatif | *Montagne du Destin* (*lieu*) ← *où Frodo transporte* |
| **Sujet-Objet (Action)** | "Frodo transporte l'Anneau Unique jusqu'à la Montagne du Destin" | Relation Sujet-Verbe-Objet | *Frodo* (*sujet*) → *transporte* (*verbe*) → *Anneau Unique* (*objet*) |

3 Lecture fortement recommandée https://arxiv.org/abs/2208.08690
4 https://aclanthology.org/C92-2082.pdf

Si on considère par exemple la phrase suivante :
"Dans l'univers du Seigneur des Anneaux, les terres de la Terre du Milieu, telles que le Gondor, le Rohan et le Mordor, abritent des lieux emblématiques, y compris la Moria et Lothlórien, où Frodo transporte l'Anneau Unique jusqu'à la Montagne du Destin."

Les modèles basés sur des règles, tels que ReVerb (Fader et al., 2011), restent pertinents pour des tâches comme l'Open Information Extraction (OpenIE). ReVerb utilise Rohan ← hypernyme ← terres de la Terre du Milieu. un analyseur syntaxique basé sur des règles pour identifier des relations où les éléments sont des groupes nominaux, servant de and)} NP sujet ou d'objet à un prédicat. Après cette étape initiale, un modèle est entraîné sur un large corpus, mais l'analyse par motifs reste fondamentale.

Ces méthodes présentent plusieurs avantages, dont leur efficacité: ellessont peu coûteuses en termes de calcul, nécessitant une seule analyse syntaxique suivie de l'application de règles. Elles offrent une grande précision dans l'extraction des relations attendues.

En contrepartie, leur rappel est souvent limité: Elles ne peuvent extraire que les relations préalablement définies, ce qui peut être restrictif pour des documents variés. L'adaptation des règles à chaque langue et type de document est laborieuse, nécessitant des ajustements spécifiques pour des textes juridiques, des ordonnances ou des articles de presse. Ces méthodes gèrent de plus mal les structures syntaxiques complexes, comme les phrases avec subordonnées u structures imbriquées.

Des efforts continus sont déployés pour améliorer ces approches, et une perspective prometteuse est leur intégration en les combinant avec des techniques modernes d'apprentissage automatique de sorte à réduire le coût (du moins d’entraînement) de celles ci.

## b) Méthodes d’apprentissage supervisé

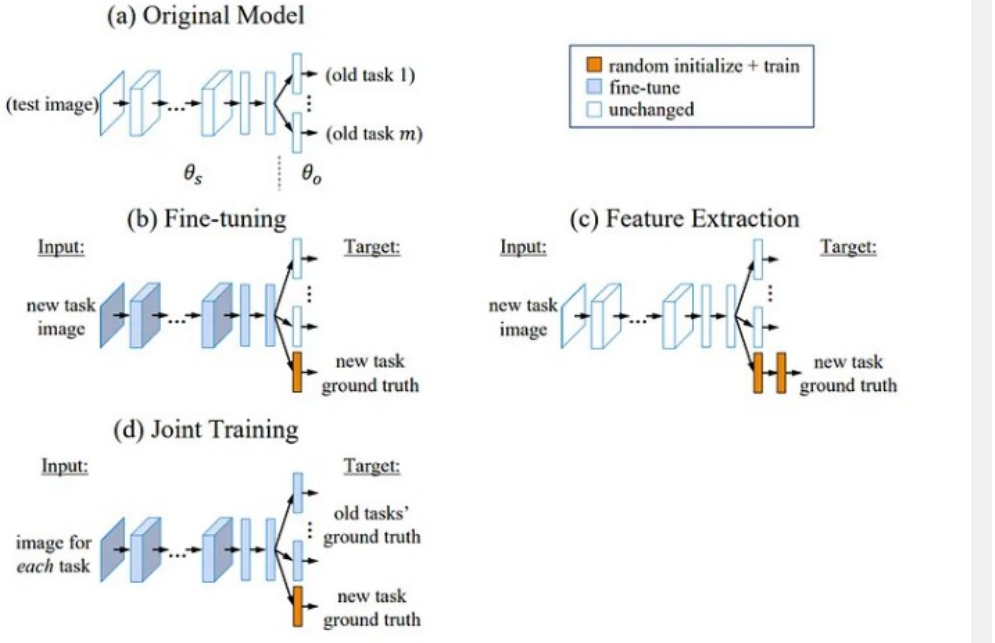
L'extraction de relations typées (ER est souvent abordée comme un problème de classification supervisée. Cette approche consiste à identifier les entités dans une phrase, puis à appliquer un modèle de classification pour déterminer la nature de la relation qui les unit. La tâche d’ER est traitée comme un problème de classification. Il faut au préalable repérer les entités et la phrase dans laquelle elles sont en relation puis appliquer un modèle de classification sur la phrase. L’architecture utilisée n’a aucune importance, le tout est que le modèle renvoie des résultats satisfaisants.

On peut construire un modèle :
* de classification simple : Une méthode consiste à représenter la phrase sous forme de vecteurs denses (embeddings) et à entraîner un classificateur pour attribuer une catégorie correspondant à la relation identifiée.
* Il est crucial de fournir au système des caractéristiques (features) à la fois sur les entités à relier et sur le segment de texte où la relation est recherchée. Cela peut être réalisé en concaténant les vecteurs des entités et du contexte, offrant ainsi une représentation riche pour le modèle.
Ces modèles permettent de spécifier explicitement les entités considérées en structurant l'entrée sous la forme : "phrase [SEP] entitéA [SEP] entitéB". Cette architecture est particulièrement adaptée aux contextes riches en entités nommées, évitant ainsi l'ambiguïté sur les entités concernées par la relation.

* Avant de déterminer la nature exacte de la relation, il est possible d'entraîner un modèle pour prédire la présence ou non d'une relation entre deux entités (ex : Socher et al. 2013). Cette approche nécessite la création d'exemples négatifs à partir du corpus initial, une étape cruciale pour l'efficacité du modèle. Une fois la présence d'une relation confirmée, un second modèle plus complexe peut être appliqué pour classifier le type de relation, combinant ainsi rapidité et précision
On obtient une architecture à deux modèles qui combine un modèle plus rapide à un autre plus complexe utilisé seulement quand c'est nécessaire.

* Plutôt que de traiter l'extraction des entités (NER) et la classification des relations séparément, une approche intégrée permet de réduire les erreurs potentielles de propagation. Des modèles tels que CoType proposent une extraction conjointe en s'appuyant sur des bases de connaissances pour améliorer la précision.
On peut penser un entraînement conjoint (joint training) du modèle sur les deux tâches en même temps.
L'entraînement conjoint est un paradigme d'apprentissage développé initialement avec les modèles convolutifs mais qui peut s'étendre à toutes les architectures. On s'est rendu compte que on était obligé de fine tuner un modèle de fondation pour chaque tâche spécifique, par exemple à partir de BERT faire un modèle qui identifies les NER et un modèle qui identifie les POS. Or ces deux tâches sont en fin de compte en partie liées car le plus souvent les NER sont des noms. Les informations fournies par au moins une des deux tâches sont directement bénéficiaires à l'autre.
Mais on ne peut spécialiser un modèle qu'à une tâche à la fois. Si on se dit qu'entraîner un modèle au POS bénéficiera au modèle de NER, alors le second fine tuning viendra écraser l'essentiel du premier fine tuning. Il convient donc de trouver une façon d ' « apprendre sans oublier »[5]

L'idée est que on va scinder le réseau sur les dernières couches (notamment la feed forward layer qui permet d'obtenir la distribution finale des modèle sur le nombre de catégories, par exemple, en laissant quelques couches de paramètres spécifiques à chaque tâche, mais l'essentiel des couches du modèle est partagé entre les différentes tâches. L'espoir est que l'information utile à toutes les tâches est stockée dans les premières couches




Pour réaliser cet entraînement joint, on doit définir une fonction de loss pour chaque tâche et alterner l'entraînement sur chacune des tâches (quelques batches) avec une bonne alternance de

5 https://arxiv.org/abs/1606.09282

sorte à ce que les paramètres communs correspondent à toutes les tâches. On force ainsi la « spécialisation » à se faire dans les dernières couches du réseau.

Avantages des méthodes supervisées :
* meilleure précision, notamment sur les relations précises qu'on tente d'identifier
Limitations :
* nécessite un corpus annoté importante (plusieurs centaines d'exemples) pour avoir un bon modèle. Si le corpus est difficile à obtenir, on peut l'augmenter par *bootstrapping* (à partir de peu d'exemples, on recherche sur le Web des phrases qui contiennent exactement les mêmes entités que les seed examples, ce qui permet de gonfler à moindre frais la taille du corpus d'entraînement.) Malheureusement cela n'améliore pas sa diversité.
On peut à partir de ces exemples étendus définir des patterns (de structure syntaxique ou de chaînes d'entités) et s'en servir pour élargir la collecte, au risque toujours d'introduire du bruit dans les données.

On peut en 2024 utiliser des LLMs pour produire un corpus d'entraînement, avec les problèmes théoriques et pratiques que cela entraîne (voir littérature actuelle sur l'emploi de LLM pour la génération de datasets synthétiques).

Pour l'extraction ouverte de relations (Open ER) :
cette tâche peut également être abordée par des méthodes d'apprentissage supervisé. Cependant, cette approche est souvent coûteuse en termes de ressources, notamment en raison de la nécessité de disposer de vastes ensembles de données annotées manuellement. Par exemple, le jeu de données TACRED, introduit par Zhang et al. en 2017, comprend plus de 100 000 exemples annotés manuellement. Malgré sa taille, ce corpus se concentre principalement sur des relations impliquant des personnes et des organisations, ce qui limite son applicabilité à des domaines spécialisés, tels que l'analyse de rapports médicaux.
Les méthodes supervisées présentent l'inconvénient de produire des résultats avec une précision parfois insuffisante, surtout lorsqu'elles sont appliquées à des domaines spécifiques. Cette limitation les rend particulièrement inadaptées pour des extractions ciblées nécessitant une expertise approfondie. De plus, la dépendance à des ensembles de données annotées de grande envergure rend difficile leur adaptation à de nouveaux domaines ou à des types de relations non prévus initialement. Par conséquent, il est essentiel d'explorer des approches alternatives, telles que les méthodes semi-supervisées ou non supervisées, qui peuvent offrir une meilleure adaptabilité et nécessiter moins de ressources en annotation.

## c) Méthodes d'apprentissage non supervisées

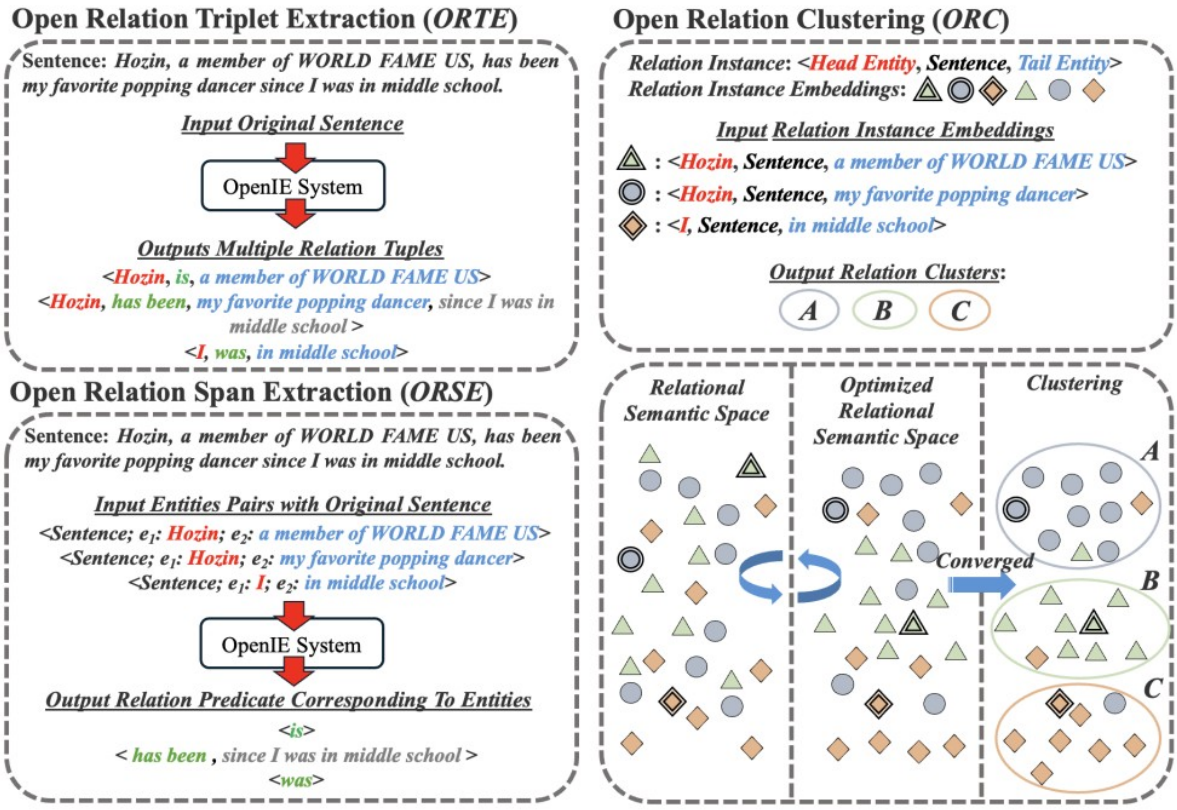
Si on peut penser, en utilisant la similarité des embeddings (cf TP 1) utiliser des méthodes non supervisées pour repérer des relations typées, c'est surtout pour la tâche dite d' Open Information Extraction (Open IE) que les méthodes non supervisées semblent les plus adaptées.
L'OPEN IE vise à extraire des relations à partir de textes sans recourir à des données annotées ou à une liste prédéfinie de relations. Les méthodes non supervisées sont particulièrement adaptées à cette tâche, notamment en exploitant la similarité des embeddings pour repérer des relations typées.

Une approche consiste à combiner la détection et l'apprentissage de motifs (patterns) avec une évaluation mathématique de leur viabilité. Par exemple, le système ReVerb identifie automatiquement et extrait des relations binaires à partir de phrases en anglais, sans nécessiter de vocabulaire prédéfini. ReVerb est conçu pour l'extraction d'informations à l'échelle du Web, où les

relations cibles ne peuvent pas être spécifiées à l'avance et où la rapidité est essentielle. (voir le chapitre 20 de Jurafsky et Martin pour une bonne explication)
De plus, l'Open IE peut être envisagée comme un problème de prédiction de liens dans un graphe. Dans ce cadre, les entités sont représentées comme des nœuds, et l'objectif est de déterminer l'existence d'arêtes (relations) entre elles. Les embeddings des segments contenant les entités fournissent des informations contextuelles sur ces entités et leurs liens potentiels. Des recherches récentes ont exploré la possibilité d'inférer de nouveaux faits directement à partir de graphes de connaissances ouverts, sans aucune canonicalisation ni supervision à partir de connaissances organisées. Par exemple, l'étude "Can We Predict New Facts with Open Knowledge Graph Embeddings?" propose une tâche de prédiction de liens ouverts, c'est-à-dire prédire des faits en complétant des questions du type ("texte du sujet", "texte de la relation", ?).

Mais il serait intéressant également d'explorer d'autres pistes du monde de l'apprentissage non supervisé (clusterings).

Chaque point est ici l'embedding d'une relation et on essaie de grouper ensemble les relations similaires.
Le danger est que l'embedding contienne trop d'éléments bruités et ne parvienne pas à effectuer le rapprochements des relations, d'où le besoin de spécialiser des embeddings dans la représentation de relations[6]




**Avantages des méthodes non supervisées :**
absence de besoin de corpus annotés manuellement, ce qui leur confère une grande flexibilité et une capacité d'adaptation à divers domaines.
Limitations :

6 https://arxiv.org/abs/2208.08690

L'une des principales limites réside dans le fait que l'OpenIE peut être perçue comme un processus de simplification du texte, consistant à réduire une phrase à ses éléments essentiels pour identifier les relations entre les entités. Cette simplification peut entraîner une perte d'informations contextuelles cruciales pour une compréhension approfondie des relations.

De plus, sans l'utilisation d'embeddings, ces méthodes ne prennent pas en compte la sémantique des concepts, ce qui peut limiter leur capacité à saisir les nuances des relations. Les approches actuelles se concentrent principalement sur les relations verbales explicites, négligeant ainsi un grand nombre de relations implicites qui, bien que cruciales, restent difficiles à détecter sans une analyse sémantique approfondie.

En outre, les méthodes non supervisées peuvent générer des résultats avec une précision variable, en raison de l'absence de données annotées pour guider le processus d'extraction. Cette variabilité peut poser des défis, notamment dans des domaines spécialisés où la précision est essentielle.

## d) La manne des LLMs ?

Comme pour toutes les tâches complexes pour lesquelles même les modèles avancés de ML n'ont pas donné de résultat satisfaisant, l'arrivée des LLM a entraîné un nouveau regain d'intérêt pour la tâche, avec la nouvelle étiquette de « generative information extraction[7] ».
Grand espoir car les LLMs ont la capacité de comprendre la nature des relations qu'on veut identifier pour nourrir la base de données (contenu du *system prompt*) et le sens fin du texte que l'on demande d'analyser (contenu du *user prompt*). « double awareness » dont sont incapables tous les systèmes ou frameworks mentionnés jusqu'alors ou un de ces ingrédients (au moins) était figé.

Idée :
* pour le Binary Extraction :
Pour l'extraction binaire, il est envisageable de concevoir un *prompt* spécifique demandant de catégoriser la relation entre deux entités. Cette méthode remplace la modélisation traditionnelle par du *prompting*, en utilisant des techniques telles que le *Chain of Thought* (CoT) ou le *Few-Shot Learning*. Normalement devrait donner de bonnes performances.
**Avantages :**
Cette approche ne nécessite pas de phase d'entraînement ni de corpus annoté, hormis quelques exemples pour le *few-shot learning*.
Temps de développement à ne pas négliger (importance du prompt prouvée sur des tâches de classification simples)
Limitations :
L'utilisation des LLM est gourmande en ressources computationnelles, ce qui peut entraîner des coûts significatifs. Il est de plus essentiel de mettre en place un pré-filtrage pour ne soumettre aux modèles que les phrases pertinentes, afin d'optimiser les ressources et le temps de calcul.

* Pour l'Open IE :

Fonctionne aussi mais nécessite un plus lourd travail de prompt engineering pour définir ce que l'on entend par une bonne relation.
Avantage :
**:** L'utilisation de LLMs permet d'éviter le développement de modèles dédiés
Inconvénient :
-Risque d'hallucinations bien plus élevé que sur la détection de relations précises.

7 https://arxiv.org/pdf/2312.17617 survey semblant complet même si conceptuellement médiocre

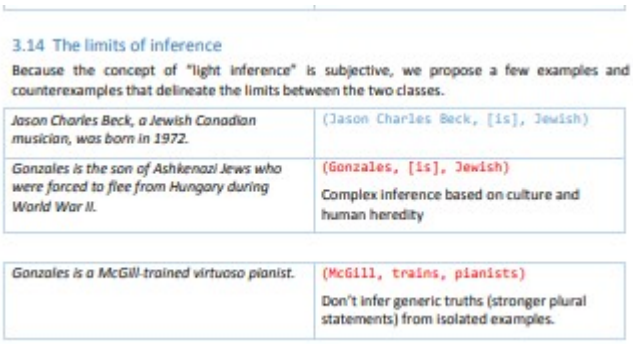
-Limite de rappel des LLMs qui ne donnent pas toutes les relations lorsque la phrase dépasse une certaine taille.
-Les meilleurs modèles comme GPT 4o repèrent en moyenne 1.5 fois plus de relations que des plus petits modèles 8B, sans pour autant être exhaustifs.

## e) L'évaluation de la tâche d'extraction d'information.

L'évaluation de la tâche d'IE / RE est complexe de par l'aspect fragmentaire d'abord de la recherche en RI, mais aussi du manque de travaux portant sur cette question spécifique de l'évaluation.
L'évaluation des approches supervisées est possible grâce à la création d'un test set, mais sa signifiance est limitée par le domaine d'usage souvent trop précis du domaine utilisé.
Pour les domaines non supervisés, en l'absence de corpus de référence, l'évaluation est très difficile et se fait souvent à la main en annotant le résultat du système sur un échantillon de texte aléatoire : définition des triplets sur le corpus puis évaluation avec des métriques de recouvrement classique (précision, rappel, F1 score). Les résultats obtenus sont fortement dépendants de la définition donnée de ce qui est un bon triplet. La capacité d'adaptation à un cas d'usage précis des systèmes génériques n'est pas traitée.

L'article de Léchelle et al. (2019), intitulé "WiRe57 : A Fine-Grained Benchmark for Open Information Extraction"[8] propose une réflexion intéressante sur la question.

La principales contribution de l'article est la création du corpus WiRe57, un corpus annoté de 57 phrases issues de cinq documents, en résolvant des défis tels que la coréférence, la granularité et l'inférence. Ce corpus vise à établir des critères précis pour déterminer ce qui doit être extrait par les systèmes Open IE, ce qui n'avait jamais été fait jusqu'alors Un travail de fond est proposé pour définir des lignes directrices d'annotation en définissant les informations à extraire, abordant des questions complexes liées à la coréférence et à la granularité des informations. Le corpus WiRe57 a été utilisé pour évaluer plusieurs systèmes d'extraction d'information ouverte. Parmi les sept systèmes comparés, MinIE a obtenu les meilleurs résultats, mais donnant des résultats mitigés : les rendus des systèmes d'OpenIE ne sont pas vraiment alignés aux attentes des utilisateurs et (du moins en 2019), la question de l'extraction d'information n'était pas du tout un problème résolu en NLP… Malgré la taille trop réduite du corpus pour être statistiquement significative, la réflexion de fond des auteurs sur la nature de la tâche est très précieuse dans un champs de recherche en général mal défini. Ils ont par exemple réfléchi à des questions comme l'usage du passif, les expressions prépositionnelles ou encore les inférences[9].



3.14 The limits of inference

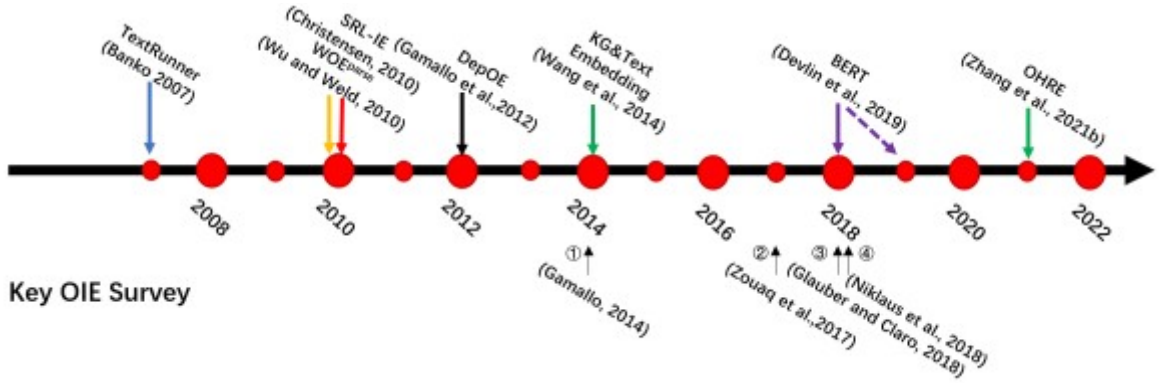
Because the concept of "light inference" is subjective, we propose a few examples and counterexamples that delineate the limits between the two classes.

| | |
|---|---|
| *Jason Charles Beck, a Jewish Canadian musician, was born in 1972.* | (Jason Charles Beck, [is], Jewish) |
| *Gonzales is the son of Ashkenazi Jews who were forced to flee from Hungary during World War II.* | (Gonzales, [is], Jewish)<br>Complex inference based on culture and human heredity |
| *Gonzales is a McGill-trained virtuoso pianist.* | (McGill, trains, pianists)<br>Don't infer generic truths (stronger plural statements) from isolated examples. |

Les travaux sur l'extraction d'information, et notamment sur l'OpenIE sont donc en perpétuelle poursuite du fait de l'importance capitale de cette brique pour le Big Data. Bien que

8 https://aclanthology.org/W19-4002.pdf
9 http://rali.iro.umontreal.ca/rali/sites/default/files/resources/wire57/WiRe57_Annotation_Guidelines.pdf

cette tâche ne soit pas la plus plébiscitée par la communauté NLP, elle fait l'objet de travaux intéressants qui à chaque fois tiennent compte des dernières avancées du domaine[10].




Un système d'Open IE qui serait capable de scaler (comprendre prendre en entrée des millions de pages de textes par jour) et d'en extraire de manière fine les informations essentielles et d'en rendre compte dans un format bien standardisé vaudrait de l'or et permettrait entre autre d'alimenter en continu des bases de données qui pourraient par exemple servir à mitiger les hallucinations des LLMs.

Définis ne serait-ce que sur des domaines de spécialité (médecine, juridique, monde RH), un tel système permettrait d'avoir une base de connaissance maintenue à jour dans des domaines où la connaissance évolue rapidement et où l'information est critique (informations sur les nouveaux médicaments, efficacité des protocoles de test, découverte de nouvelles contre indications…). Malgré ce que l'écosystème du NLP semble suggérer, l'OpenIE a un bel avenir…

# 3) Défis et débats

## a) La RI : tâche unitaire ou pipeline ?

L'extraction d'information est une tâche qui s'appuie sur plusieurs autres sous-tâches du traitement automatique du langage naturel (TAL). Parmi celles-ci, la reconnaissance des entités nommées (NER) joue un rôle essentiel en identifiant les éléments clés d'un texte, comme les personnes, les organisations ou les objets. Cependant, d'autres aspects du langage doivent être pris en compte pour obtenir un système performant, notamment la résolution des coréférences et le *linking* des entités.

Prenons un exemple concret : « Bilbo a trouvé l'Anneau de Gollum. Il l'avait perdu dans la rivière. »
Si l'on ne résout pas les coréférences, l'information contenue dans la seconde phrase reste ambiguë. Qui est « il » ? À quoi fait référence « le » ? Sans un système de résolution de coréférence, nous risquons de perdre une information cruciale : le fait que Gollum avait perdu l'Anneau. Une bonne résolution permet de rétablir explicitement les relations et de reconstruire des triplets d'information tels que (Gollum, a perdu, Anneau).

L'entité *linking* est une autre composante essentielle. Identifier une entité dans un texte ne suffit pas toujours : il faut aussi pouvoir la relier à une base de connaissances pour la désambiguïser. Dans le cas de notre exemple, un bon système ne se contentera pas d'identifier « Bilbo » comme un nom

10 https://arxiv.org/abs/2208.0869

propre, mais le reliera à *Bilbo Baggins*, personnage du *Seigneur des Anneaux*, et fera de même pour *l'Anneau unique* et *Gollum*. Ce processus est détaillé dans le chapitre 7 du manuel de référence.

En outre, il a été démontré qu'un système d'Open IE bénéficie également du *POS Tagging* (analyse morphosyntaxique). Connaître la nature grammaticale des mots (verbe, sujet, complément, etc.) permet d'améliorer l'extraction des relations, en identifiant par exemple plus précisément les verbes qui expriment une action et les noms qui désignent les acteurs et les objets de cette action.

Ainsi, pour construire un système efficace d'Open Information Extraction, plusieurs étapes sont nécessaires. Il faut d'abord un bon système de reconnaissance des entités nommé adapté aux types d'entités que l'on souhaite extraire. Par exemple, dans notre phrase d'exemple, nous devons pouvoir annoter correctement les entités :
**Bilbo** comme une personne, **l'Anneau** comme un objet, **Gollum** comme une personne et **la rivière** comme un lieu.

Ensuite, la résolution des coréférences est indispensable pour rétablir les relations implicites du texte. Il s'agit d'identifier que « il » renvoie à Gollum et que « le » fait référence à l'Anneau. Avec cette étape réalisée, notre phrase devient plus explicite :
**Bilbo a trouvé l'Anneau de Gollum. Gollum avait perdu l'Anneau dans la rivière.**
À partir de cette reformulation, un bon système d'Open IE peut alors extraire plusieurs relations pertinentes :

- (Bilbo, a trouvé, Anneau)
- (Gollum, a perdu, Anneau)

On pourrait même aller plus loin en proposant une reformulation plus informative et exploitable sous forme de triplets supplémentaires :

- (Bilbo, possède, Anneau)
- (Gollum, ne possède plus, Anneau)

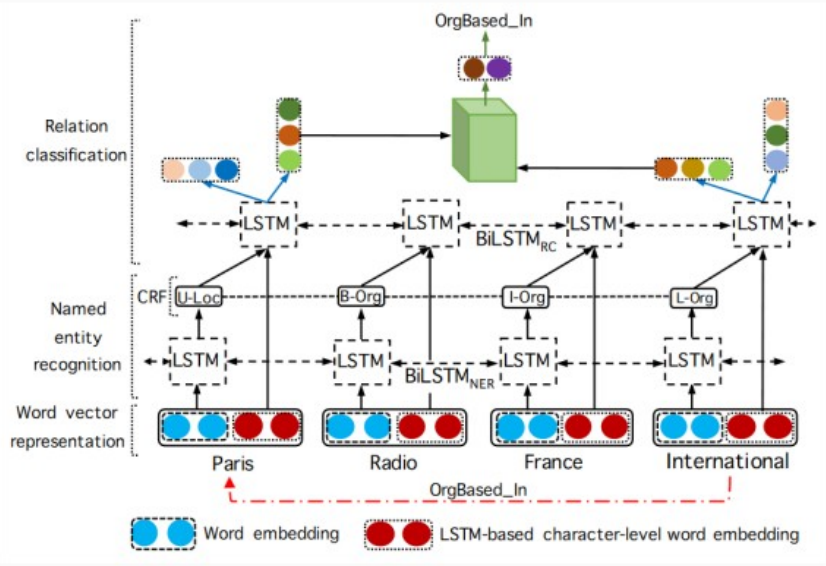
Ce dernier triplet, bien qu'implicite, permet d'enrichir l'information extraite et illustre comment un bon système d'extraction d'information peut non seulement identifier les relations présentes dans un texte, mais aussi inférer des informations supplémentaires à partir de la structure linguistique et du contexte.

Récapitulatif :
1) Un système de NER adapté aux entités / concepts que l'on veut découvrir pour tagger les entités
[PERS] Bilbo[/PERS] a trouvé [OBJ]l'Anneau[OBJ] de [PERS]Gollum[/PERS]. Il l'avait perdu [LOC]dans la rivière [/LOC]
2) Un système de résolution de coréférence
[PERS] Bilbo[/PERS] a trouvé [OBJ]l'Anneau[OBJ] de [PERS]Gollum[/PERS]. Il=[PERS]Gollum[/PERS] l=[OBJ]l'Anneau[OBJ]'avait perdu [LOC]dans la rivière [/LOC]
3) Enfin le système d'Open IE

[PERS] Bilbo[/PERS] a trouvé [OBJ]l'Anneau[OBJ] de [PERS]Gollum[/PERS]. Il=[PERS]Gollum[/PERS] l=[OBJ]l'Anneau[OBJ]'avait perdu [LOC]dans la rivière [/LOC]
=> (Bilbo, a trouvé, Anneau)
=>(Gollum, a perdu, Anneau)
Voire => (Bilbo, a , Anneau), (Gollum, n'a plus, Anneau)…

Ce n'est donc pas un modèle de NLP dont on a besoin, mais de tout un logiciel si on veut être capable de capturer toutes les informations…

Le problème fondamental des architectures logicielles dès lors qu'elles se complexifient est qu'elles sont sensibles au phénomène de propagation de l'erreur, à savoir que l'erreur (que ce soit de « précision » ou de « rappel » sur chacun des composant est directement reportée voire amplifiée sur le résultat final. C'est donc une pile de brique à constamment changer, tuner, pour améliorer parfois de façon négligeable les performances du système entier, on a un manque fondamental de contrôle sur les pipeline complexe.




Une alternative serait de définir un modèle unique pour relier tout en entraînant un unique modèle à réaliser toutes les tâches en même temps.

Voire par exemple l'architecture de ce LSTM
C'est possible par le paradigme dit de joint training déjà présentée. Aujourd'hui, la création de l'architecture dite Mixture of Experts semble aussi prometteuse pour ce genre de tentatives.

## b) Le problème de l'open world

Une fois les triplets extraits, plusieurs défis se posent quant à leur exploitation et leur structuration. L'un des premiers enjeux est la résolution des coréférences afin d'éviter les ambiguïtés et de garantir une représentation cohérente des relations. Par exemple, si deux phrases mentionnent « Sauron » et « leSeigneur des Ténèbres », il est essentiel de reconnaître qu'il s'agit de la même entité pour éviter la fragmentation des informations.

Un autre défi majeur est la gestion des contradictions. Si un triplet extrait indique (X, est le président de, Y) et qu'un autre affirme (Z, est le président de, Y), il devient crucial de savoir si X et Z sont la même personne, si l'information est périmée ou s'il y a une véritable contradiction. Ce problème est particulièrement aigu dans des contextes dynamiques comme l'actualité.

Mais l'un des obstacles les plus fondamentaux reste la diversité et l'hétérogénéité des triplets extraits. L'absence de standardisation s'avère un frein majeur à l'intégration efficace des informations, car des relations exprimées différemment peuvent en réalité véhiculer le même sens. Par exemple, les deux triplets suivants sont équivalents mais formellement distincts :

- (Sauron, a dominé, le Mordor)
- (Napoléon, était le maître de, le Mordor)

Deux principales approches peuvent être envisagées pour normaliser ces relations :

1. **Normalisation par lemmatisation du verbe**
   Une solution simple consiste à réduire le verbe à son lemme pour unifier certaines relations. Cela permettrait de faire correspondre des variations morphologiques d'un même verbe, comme « dirige », « dirigeait » et « dirigera », qui seraient toutes ramenées à « diriger ». Cependant, cette méthode rencontre des limites face aux verbes polysémiques (ex. « tenir » peut signifier « posséder » ou « organiser ») et aux expressions à verbe support (« avoir lieu », « prendre une décision »), qui nécessitent une analyse plus fine.
2. **Clustering des relations pour subsumer des catégories plus générales**
   Une autre approche, plus ambitieuse, consisterait à regrouper automatiquement les relations similaires en exploitant des techniques non supervisées. En appliquant des méthodes de clustering sur les relations extraites, il serait possible d'identifier des schémas récurrents et de généraliser certaines relations spécifiques sous des catégories plus larges. Par exemple, les triplets précédents correspondaient tous à l'idée de pouvoir.

   Cette approche permettrait de rendre la base de connaissances plus exploitable en réduisant la fragmentation des relations. Cependant, un tel regroupement pourrait entraîner une perte d'information, notamment sur les nuances de temps et d'aspect des verbes. Par exemple, fusionner « a dirigé » et « dirige » sous une même relation ne permettrait plus de savoir si l'action est toujours en cours ou appartient au passé.

Ainsi, toute tentative de structuration des triplets implique un compromis entre **standardisation et conservation des nuances**. Trouver un équilibre optimal entre ces deux aspects serait une piste de recherche particulièrement intéressante à explorer.

**<u>Différents formats de stockage</u>**

Le triplet a toujours été l'élément initial de l'IE : « *In seminal work (Banko et al., 2007), the task setting was to extract entities relation triples (entity1, relation, entity2*) »

Certains ont abandonné le format contraignant du triplet Task : entities + text → relation span

Par exemple, le cadre QuORE, proposé par Yang et al. en 2022, adopte une approche différente en se concentrant sur le clustering des instances de relations sans extraire explicitement les spans de relations ou attribuer des étiquettes prédéfinies. Dans cette approche, chaque instance est représentée par une paire d'entités (tête et queue) et la phrase correspondante. Les relations sont ensuite regroupées en clusters basés sur des similarités, sans nécessiter de définition explicite des spans de relations ou de leurs étiquettes

« Entities + Text→ Clustering without Explicit Relation Span or Label Open relation clustering (ORC), also known as open relation extraction, clusters relation instances (h, t, s), where h and t denote head entity and tail entity respectively, and s denotes the sentence corresponding to two entities. » cf lecture recommandée. C'est ce qu'on va essayer de faire dans la seconde partie du TP1.

Cette diversificati

on des approches, bien que prometteuse, introduit une complexité supplémentaire en termes de standardisation des sorties de l'Open IE. L'absence de format uniforme complique l'intégration et l'utilisation ultérieure des informations extraites. Pour pallier ce problème, il est essentiel de mettre en place des processus de normalisation des relations. Cela peut inclure la lemmatisation des verbes pour réduire les variations morphologiques ou l'application de techniques de clustering non supervisées pour regrouper des relations similaires sous des catégories plus générales. Cependant, ces méthodes peuvent entraîner une perte d'informations, notamment en ce qui concerne les nuances temporelles des verbes.

## c) La RI : rêve de savant fou ou techniquement accessible ?

La Recherche d'Information (RI) est-elle un rêve de savant fou ou une réalité techniquement accessible ? Cette question soulève des enjeux cruciaux en matière de scalabilité et d'éthique.

**Scalabilité des systèmes de RI**

La scalabilité désigne la capacité d'un système à gérer efficacement une augmentation du volume de données ou du nombre d'utilisateurs sans compromettre ses performances. Dans le contexte de la RI, cela implique la capacité à traiter de vastes ensembles de données hétérogènes en temps réel ou quasi-réel. Des architectures distribuées et des algorithmes optimisés sont essentiels pour atteindre cette scalabilité mais ne suffisent pas forcément à gérer le flux d'information.

**Enjeux éthiques liés à l'analyse de données hétérogènes**

L'utilisation de systèmes capables d'ingérer et d'analyser des données hétérogènes soulève des préoccupations éthiques majeures. Prenons l'exemple d'un système conçu pour surveiller les communications des citoyens :

Exemple de cas d'usage de la RI système de détection de la critique politique : ne pas critiquer le chef de l'état. Suffit de passer les communications des citoyens à la moulinette NLP Repérer les messages parlant du chef de l'état. Passer un système d'analyse de sentiments. En quelques secondes, on repère un dissident.

Bien que la RI soit techniquement réalisable grâce aux avancées en matière de scalabilité, son déploiement soulève des questions éthiques complexes. Il est impératif d'équilibrer les capacités technologiques avec des considérations éthiques pour garantir le respect des droits individuels et collectifs.