Traitement automatique de la coréférence dans les textes économiques : un exemple d'application chez ReportLinker

Marilyne Latour, Paul Moncuquet

ReportLinker, 21 Quai Antoine Riboud, 69002, Lyon, France
pmo@reportlinker.com, mla@reportlinker.com
https://www.reportlinker.com/

TextMine @ EGC 2023, 17 Janvier 2023, Lyon

Introduction

Définition

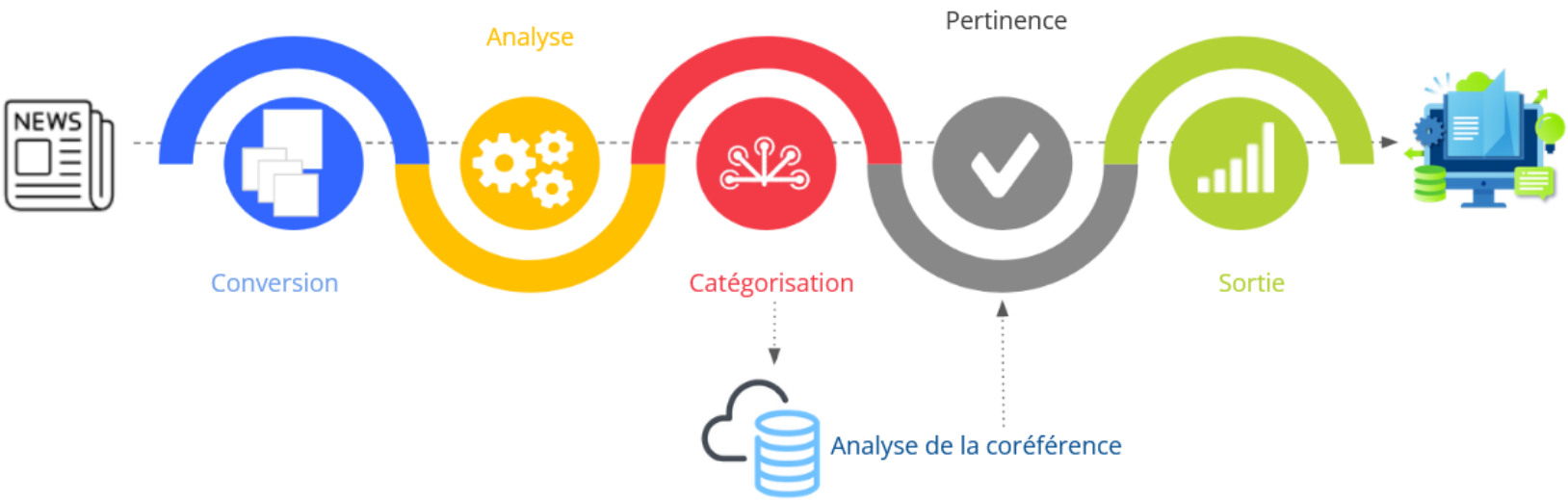
La coréférence est la relation entre plusieurs expressions référentielles qui désignent le même référent [Delaborde, 2021]. En français, elles sont nommées « chaîne de référence » par de nombreux linguistes francophones [Gobert and Fabre, 2017] ; [Landragin and Oberle, 2018].
Plusieurs travaux en linguistique (morphologie, syntaxe, rôles discoursifs et sémantique) portent sur les entités du discours et donnent des indications sur les critères qui peuvent aider les systèmes de traitement automatique à prédire si une expression référentielle sera reprise (coréférente) ou non (singleton).

Contexte Applicatif

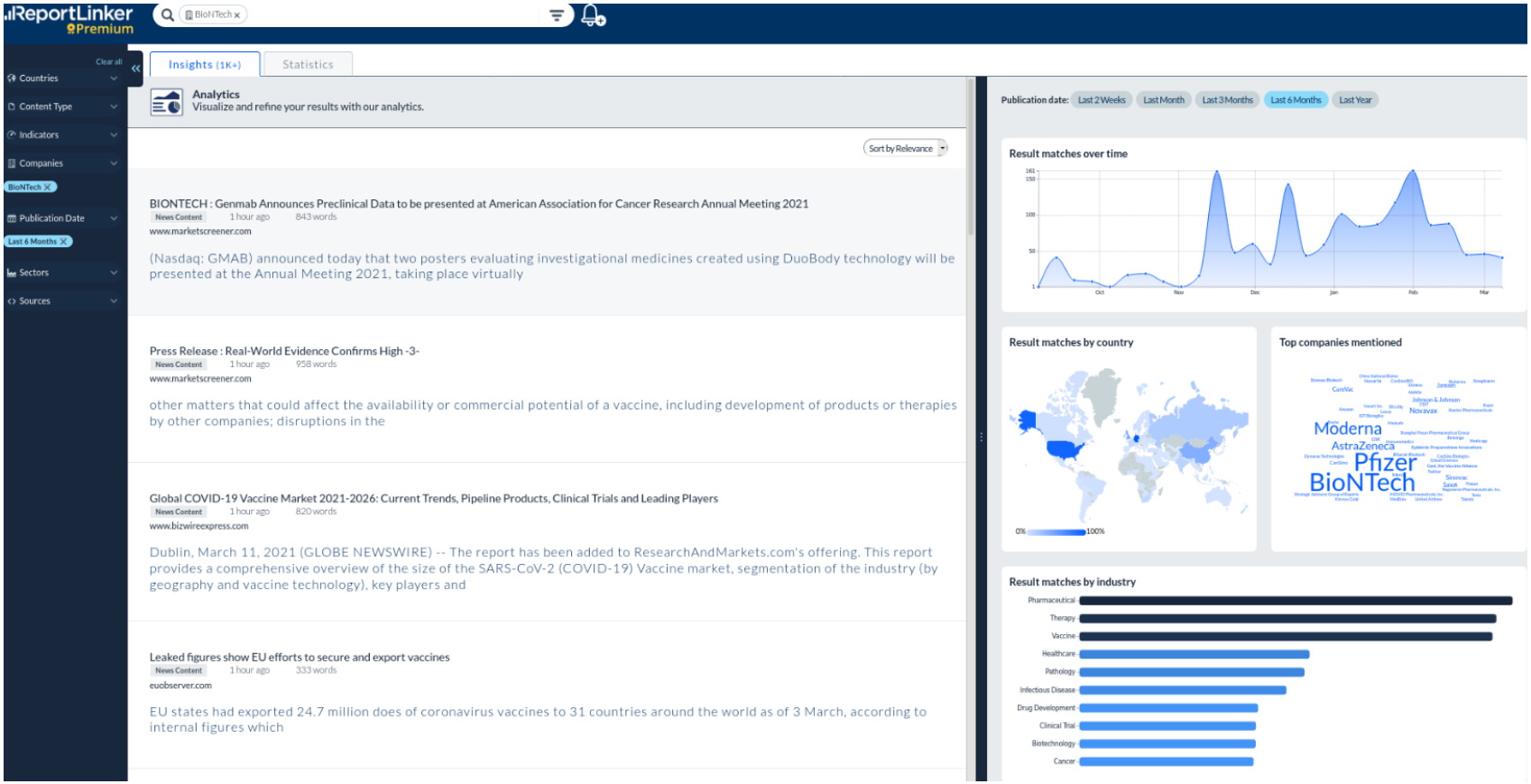
ReportLinker: Page d'accueil

- Agrégateur d'actualités généralistes
- 18000 journaux ou sites web
- 4 à 6 millions de dépêches d'actualités en langue anglaise chaque mois

ReportLinker: NLU



ReportLinker: Exemple de recherche



Objectif : ajouter du contexte aux expressions référentielles

Exemple 1 : Phrases cibles non compréhensibles

[Phrase_cible]« *The Swiss-based company is launching a slew of new products aimed especially at people working from home.* »

Exemple 2 : Phrases cibles compréhensibles

[Phrase_cible] « *Provident Healthcare Partners ("Provident"), a leading healthcare investment banking firm, announced that it has advised Texan Eye and Eye LASIK Austin in their partnership with Comprehensive EyeCare Partners ("CompEye").* »

Expérimentation

Corpus

	Nombre de phrases	Pourcentage
Phrases cibles compréhensibles	6 409	64%
Phrases cibles non compréhensibles	3 591	36%

Table: Répartition du nombre de phrases cibles compréhensibles et non compréhensibles parmi le corpus de 10 000 phrases

Tests effectués à partir des phrases non compréhensible

Nous avons testé quatre échantillons :

1. **Modèle 1** : un nom de société présent à la fois dans la phrase précèdente et dans le titre,
2. **Modèle 2** : un seul nom de société présent uniquement dans la phrase précèdente,
3. **Modèle 3** : un seul nom de société présent uniquement dans le titre,
4. **Modèle 4** : pas de mention de nom de société, ni dans la phrase précèdente, ni dans le titre

Exemple Modèle 1

- **[Phrase_cible]** « ***The Swiss-based company*** *is launching a slew of new products aimed especially at people working from home.* »
- **[Phrase_Precedente]** « *Working from home is something that* ***Logitech*** *is taking very seriously.* »
- **[Titre_Depeche]** « ***Logitech*** *Introduces All-In-One Dock to Declutter the Desktop and Make Joining Meetings Easy* »

Exemple Modèle 2

- **[Phrase_cible]** « *The tech company took a major step toward that goal earlier this year with its acquisition of Hyperconnect in a cash-and-stock transaction valued at $1.725 billion.* »
- **[Phrase_Precedente]** « *But **Match Group** is also looking to expand its offerings beyond dating.* »
- **[Titre_Depeche]** « *2 Stocks That Could Turn $200,000 Into $1,000,000 in 10 Years.* »

Exemple Modèle 3

- **[Phrase_cible]** « ***The firm*** *had revenue of C$308.26 million during the quarter, compared to the consensus estimate of C$305.00 million.* »
- **[Phrase_Precedente]** « *The company reported C$0.51 earnings per share for the quarter.* »
- **[Titre_Depeche]** « ***High Liner Foods (TSE:HLF)*** *Shares Cross Above 200 Day Moving Average of $0.00* »

Exemple Modèle 4

- **[Phrase_cible]** « *The **company** is seeing strong same-store sales growth, up 51% in the first quarter of 2021 over Q1 2020.* »
- **[Phrase_Precedente]** « *It sells all of the equipment and supplies for growing, including climate controls, lighting, soils, fertilizer, and additives.* »
- **[Titre_Depeche]** « *Looking for a U.S. Cannabis Stock? This Might Be the One* »

Résultats

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Fusions et acquisitions	29%	6%	58%	7%
Données de vente	23%	2%	69%	6%
Données financières	65%	1%	33%	1%
Lancements de produits	25%	10%	47%	18%
Descriptions d'activités	40%	7%	44%	9%

Table: Répartition du nombre phrases « cibles » non compréhensible par type de news parmi le corpus de 10 000 phrases

Résultats

	Modèle 1	Modèle 2	Modèle 3	Modèle
Fusions et acquisitions	29*(6.4)*%	6*(14)*%	58*(26.3)*%	7*(45.6)*
Données de vente	23*(12)*%	2*(9.5)*%	69*(41.4)*%	6*(37.2)*
Données financières	65*(25.6)*%	1*(36)*%	33*(21.5)*%	1*(16.9)*
Lancements de produits	25*(8.4)*%	10*(3.9)*%	47*(27.2)*%	18*(60.7*
Descriptions d'activités	40*(14)*%	7*(14)*%	44*(26.3)*%	9*(45.6)*

Table: Répartition du nombre phrases « cibles » dans l'application RLK

Résultats

	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Précision	98%	99%	96%	–

Table: Précision de la coréférence en fonction des 4 modèles

Précision

$$Précision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Résultats

Source d'érreurs

- **[modèle 3]** Lorsque plusieurs noms de sociétés sont cités dans le [Titre_Depeche], cela apporte de la confusion lors de l'interprétation des expressions référentielles.
- **[modèle 2]** Même cas de figure avec la [Phrase_Precedente] mais qui est plus rare. Il semblerait donc que la distance « inter-maillonnaire » ait un impact sur la fiabilité de la coréférence.

Résultats

Exemple d'erreurs:

- **[Phrase_cible]** « *The company recorded slow year-on-year growth for hair products, and the sales under hair saloon business were also down by a double-digit percentage.* »
- **[Phrase_Precedente]** « *For instance,* ***Henkel*** *and* ***Unilever*** *witnessed a decline in shampoo consumption in the first quarter of 2020 when the coronavirus has just started to transmit worldwide.* »
- **[Titre_Depeche]** « *Shampoo Market to Hit USD 39.58 Billion by 2028; Custom Made Hair Cleanser to Augment Market Growth, Says Fortune Business Insights™* »

Deux sociétés sont mentionnées dans la [Phrase_Precedente]: « Henkel » et « Unilever ».

Conclusion et ouvertures

Résumé

Notre objectif était double :

1. Constituer un corpus spécifique aux dépêches d'actualités en anglais dans le domaine de l'économie et des finances afin de pouvoir annoter les coréférences des noms de société.
2. Fournir des recommandations pour le traitement de la coréférence en corpus qui pourront servir de base d'entraînement à des outils automatiques ou à des analyses linguistiques.

Discussion

1. Les résultats obtenus sont déjà concluants. Ces bons chiffres peuvent éventuellement s'expliquer par le choix du corpus : des dépêches d'actualité sur la finance avec des typologies de classes spécifiques aux sociétés.
2. Cette approche est assez facilement généralisable et pourrait être un apport pour l'aide à la résolution de coréférence.

Ouvertures

1. Travailler sur des marqueurs linguistiques pour décider s'il y a continuité ou discontinuité dans la chaîne de référence.
2. Etendre notre étude non seulement à la phrase précèdente et au titre pour mesurer la coréférence en fonction de la proximité textuelle des maillons coréférents.

Références

Delaborde, M. (2021).
Analyse en corpus de chaînes de coréférence: la coréférence non-stricte à l'épreuve de la linguistique outillée.
Thèse de doctorat, Université de Sorbonne Nouvelle - Paris 3.

Gobert, E. and Fabre, B. (2017).
Détection de coréférences de bout en bout en français.
In *Actes JEP TALN RECICL 2017*.

Landragin, F. and Oberle, B. (2018).
Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage.
In *Actes PFIA 2018*.

Fin

Merci

Des questions ?

Pour nous contacter : pmo ou mla@reportlinker.com