Défi TextMine'23 - Reconnaissance d'entités d'intérêts dans les signatures d'e-mails

Kévin Cousot (Emvista), Cédric Lopez (Emvista), Pascal Cuxac (INIST-CNRS), Vincent Lemaire (Orange Labs)

Vendredi 21 octobre 2022

1 Introduction

Le 21 octobre 2022, l'association Extraction et Gestion des Connaissances (EGC) a lancé le groupe de travail TextMine. Dans le cadre de ce groupe de travail, un objectif est de confronter l'état de l'art scientifique aux problèmes de text mining rencontrés par des industriels. Sous la forme de défis, le groupe de travail propose des jeux de données inédits et les partage avec la communauté scientifique. Le premier défi du groupe de travail TextMine a été lancé le 21 octobre en étroite collaboration avec la société Emvista, éditrice de logiciels fondés sur des technologies du Traitement Automatique du Langage Naturel, qui a fourni une partie des données. En particulier, la société s'intéresse à la structuration des informations véhiculées dans les e-mails.

Le défi proposé porte sur la reconnaissance d'entités d'intérêts dans les signatures d'e-mails dans le but de structurer l'information et de la stocker en base de données (par exemple un système de gestion de la relation client). Il s'agira de s'approcher de conditions réelles dans lesquelles les données disponibles à l'entraînement ne reflètent pas nécessairement la distribution des données auxquelles le système devra faire face en production. Le participant qui obtiendra le meilleur résultat se verra remettre un prix de 300 euros.

La suite du document s'organise comme suit. Nous commencerons par décrire tâche (section 2) et les données à utiliser (section 3). Les modalités d'évaluation et de participation sont ensuite expliquées section 4.

2 Tâche

La tâche se formule comme un problème de classification de tokens à 13 classes. On prend en entrée un texte brut dans lequel les entités d'intérêts sont identifiées. Il s'agit alors d'attribuer une étiquette à chacune des entités.

Donnée d'entrée :
Anna$_{?}$ Dupont$_{?}$
Directeur$_{?}$ Général$_{?}$
Téléphone : 01.55.52.12.96$_{?}$
www.anywaythewall.com$_{?}$

Résultat attendu :
Anna$_{\text{Human}}$ Dupont$_{\text{Human}}$
Directeur$_{\text{Function}}$ Général$_{\text{Function}}$
Téléphone : 01.55.52.12.96$_{\text{Phone_Number}}$
www.anywaythewall.com$_{\text{Url}}$

3 Les données

Face à l'absence de données annotées, le groupe de travail a produit trois jeux de données à l'aide de différentes stratégies. Toutes ces données sont partagées avec la communauté scientifique.

- Un Jeu de données authentique (JDA dans la suite) : un jeu de données composé de signatures authentiques pseudonymisées ;

Classe	Définition
Human	Noms et prénoms des personnes qui figurent dans la signature. Ex : Martin Dupond
Organization	L'organisation à laquelle l'auteur de la signature est rattachée.
Function	L'ensemble des fonctions assignées à la personne identifiée dans la signature. Ex : conseiller, assistant, professeur.
Project	Projet dans lequel la personne est impliquée. Ex : Comm & Partenariats, Master 2 Informatique, Direction de l'innovation.
Location	Bâtiments, bureaux, villes, numéros et noms des rues. Ex : 24 avenue Jean Jaurès, Montpellier
Reference_CEDEX	Courrier d'entreprise à distribution exceptionnelle. Ex : Cedex 05
Reference_CS	Course spéciale. Ex : CS 39521.
Reference_Code_Postal	Code postal. Ex : 34080.
Phone_Number	Numéro de télphone ou de fax. Ex : 01.23.45.67.89
Email	Adresses e-mails. Ex : martin.dupond@bboite.com
Url	URL, typiquement vers le site web de l'entreprise ou la personne. Ex : www.anywaythewall.com
Social_Network	Nom des réseaux sociaux. Ex : LinkedIn, Facebook, Twitter...
Reference_User	Identifiant d'une personne ou d'une organisation sur un réseau social. Ex : @Kalypso_immobilier

Table 1: Classes annotées dans les jeux de données.

- Jeu de données réaliste (JDR dans la suite) : un jeu de données composé de signatures construites manuellement par la société Isahit, plateforme de labellisation éthique des données pour l'IA[1] ; ce jeu de données contient des signatures réalistes, c'est-à-dire proches des signatures authentiques observées, mais non authentiques (elles n'ont jamais été utilisées dans des échanges d'e-mails) ;
- Jeu de données factice (JDF dans la suite) : un jeu de données composé de signatures créées automatiquement à partir d'une API de génération de fausses identités[2]

Les entités présentes dans les signatures sont annotées à l'aide d'une typologie de 13 classes, présentées dans le tableau 1.

3.1 Jeu de données authentique (JDA)

Face à l'absence de signatures authentiques parmi les jeux de données disponibles, la société Emvista a mis en place un formulaire Web permettant le "don" de signatures pendant 6 mois (Bendahman et al., 2022).

Pour préserver l'identité des contributeurs, un processus de pseudonymisation a été appliqué. Chaque entité d'intérêt a été remplacée par une autre entité de même classe afin d'assurer la cohérence du jeu de données. Certaines classes ont été traitées manuellement, d'autres de façon automatique. Notons également que les fonctions des personnes n'ont pas été remplacées, sauf dans le cas où la description permettait d'identifier la personne.

Un exemple de pseudonymisation est donné ci-dessous :

[1] https://fr.isahit.com/
[2] https://www.fakenamegenerator.com/

Classe	Quantité d'annotations
Human	1196
Organization	1537
Location	2680
Phone_Number	688
Function	1449
Email	344
Url	303
Social_Network	28
Reference_User	11
Reference_Code_Postal	349
Project	124
Reference_CEDEX	146
Reference_CS	33

Table 2: Nombre d'annotations dans JDA.

Signature:

Mary Margho
Directeur Général
Téléphone : +33.(0)1.52.62.32.65
6 Rue Jean-Paul Montagne
75001 Paris
www.saveprogramma.com

Signature pseudonymisée :

Anna Dupont
Directeur Général
Téléphone : 01.55.52.12.96
72, Rue Paul-Marie L'abbé
34090 MONTPELLIER
www.anywaythewall.com

Au total, le JDA contient 606 signatures. La répartition des classes utilisées est indiquée dans le tableau 2.

3.2 Jeu de données réaliste (JDR)

Le jeu de données réaliste a été produit par la société Isahit avec un suivi régulier par les organisateurs du défi. Le cahier des charges fourni initialement à Isahit contenait des contraintes émises à partir de l'observation des signatures authentiques (cf. section 3.1).

D'une part, il a été demandé que la taille des signatures générées soient comprise entre 2 et 132 tokens avec une moyenne d'environ 46 tokens sur la totalité du jeu de données produit. Il a également été demandé qu'entre 1 et 3 tokens par signature ne soit pas catégorisable dans l'une des classes prédéfinies (par exemple « Tel : » ou encore la coordination « et » entre deux noms de personnes). Enfin, des contraintes d'ordre statistiques ont été émises afin d'approcher les distributions observées dans le JDA.

Le jeu de données a fait l'objet d'une vérification de la qualité des annotations :

1. pour chaque signature, vérification automatique du respect des contraintes imposées sur l'utilisation des classes. Ce point a déclenché des aller-retours avec la société afin d'obtenir un respect total des contraintes énoncés dans le cahier des charges;

2. un échantillon de 50 signatures prises aléatoirement a été évalué manuellement : 100% des annotations observées sont correctes.

Au total, le JDR contient 473 signatures. La répartition des classes utilisées est indiquée dans le tableau 3.

Classe	Quantité d'annotations
Human	971
Organization	1023
Location	2533
Phone_Number	473
Function	567
Email	297
Url	227
Social_Network	18
Reference_User	9
Reference_Code_Postal	269
Project	98
Reference_CEDEX	150
Reference_CS	56

Table 3: Nombre d'annotations dans JDR.

Classe	Quantité d'annotations
Human	1023
Organization	943
Location	2150
Phone_Number	371
Function	0
Email	371
Url	0
Social_Network	0
Reference_User	0
Reference_Code_Postal	367
Project	0
Reference_CEDEX	0
Reference_CS	0

Table 4: Nombre d'annotations dans JDF.

3.3 Jeu de données factice (JDF)

Le jeu de données factices (JDF) est composé de signatures créées automatiquement. La génération s'appuie sur des patrons définis manuellement, une API de création de fausses identités[3] et des heuristiques introduisant de la variabilité au sein des entités. L'API ne fournissant pas certaines classes, celles-ci sont absentes du JDF : Project, Url, Reference_User, Reference_CEDEX, Reference_CS et Function.

Au total, 500 signatures ont été générées.

3.4 Format

Les données sont distribuées en format JSON. Chaque signature possède trois champs :

- "identifier": l'identifiant, numérique, de la signature
- "text": le texte brut de la signature
- "annotations" : la liste des annotations d'entités d'intérêts portant sur la signature

Une annotations contient quatre champs :

[3] https://www.fakenamegenerator.com/

- "form" : la forme du token annoté
- "label" : la classe associée au token
- "begin" : l'index du début du token (inclus)
- "end" : l'index de fin du token (exclu)

Exemple :

```
[
  {
    "identifier" : 0,
    "text" : "Faustin Dupont",
    "annotations" : [
      {
        "form" : "Faustin",
        "label" : "Human",
        "begin" : 0,
        "end" : 7
      },
      {
        "form" : "Dupont",
        "label" : "Human",
        "begin" : 8,
        "end" : 14
      }
    ]
  }
]
```

4 Évaluation

Le défi consiste à obtenir la meilleure F-mesure[4] sur la tâche de reconnaissance d'entités d'intérêts dans le jeu de test. Tout système est le bienvenu : symbolique, connexionniste, à base de connaissances ou d'apprentissage etc.

Pour l'entraînement, les participants ont accès au JDF ainsi qu'au JDR. La répartition des classes est donnée tableau 5. Les participants ont la liberté d'utiliser ces jeux de données comme ils le souhaitent mais sans y apporter de modification. Les données sont téléchargeables sur le Github suivant : https://github.com/Emvista/Challenge_TextMine_2023.

Le jeu de données utilisé pour l'évaluation finale est le JDA. Celui-ci sera ajouté au dépôt github le 8 janvier 2023. Dès lors, les participants ont jusqu'à la clôture du défi, le 10 janvier 2023, pour soumettre leurs résultats par mail à l'une de ces deux adresses :

- kevin.cousot@emvista.com
- cedric.lopez@emvista.com

Les résultats d'évaluation leur seront communiqués après chaque soumission, mais le classement des participants ne sera dévoilé qu'à la remise du prix, le jour de l'atelier TextMine au sein de la conférence EGC, le 17 janvier 2023.

Le jeu de test respecte le même format que les jeux d'entraînement à l'exception du fait que le label y est absent. Ce champ est à remplir par les participants et doit être présent dans les résultats soumis à évaluation (voir listing 1 et 2).

[4] 2*(precision*rappel) / (précision + rappel)

Classe	Quantité d'annotations
Human	1994
Organization	1966
Location	4683
Phone_Number	844
Function	567
Email	668
Url	227
Social_Network	18
Reference_User	9
Reference_Code_Postal	636
Project	98
Reference_CEDEX	150
Reference_CS	56

Table 5: Nombre d'annotations dans les données d'entraînement JDR+JDF (11916 annotation).

Listing 1: Données de test

```
[
  {
    "identifier" : 0,
    "text" : "Faustin Dupont",
    "annotations" : [
      {
        "form" : "Faustin",

        "begin" : 0,
        "end" : 7
      },
      {
        "form" : "Dupont",

        "begin" : 8,
        "end" : 14
      }
    ]
  }
]
```

Listing 2: Soumission du participant

```
[
  {
    "identifier" : 0,
    "text" : "Faustin Dupont",
    "annotations" : [
      {
        "form" : "Faustin",
        "label" : "Human",
        "begin" : 0,
        "end" : 7
      },
      {
        "form" : "Dupont",
        "label" : "Human",
        "begin" : 8,
        "end" : 14
      }
    ]
  }
]
```

References

Nihed Bendahman, Kevin Cousot, and Cédric Lopez. Reconnaissance d'entités d'intérêt dans les signatures d'e-mails à partir d'un jeu de données authentique. *TextMine'22*, 2022.