Utilisation d'outils linguistiques pour trouver la date ou l'origine d'un fragment textuel

Laura Monceaux Annie Tartier
LINA, UMR 6241, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 03, France
laura.monceaux@univ-nantes.fr, annie.tartier@univ-nantes.fr

Résumé. Cet article décrit notre approche des deux tâches proposées, puis la mise en place de méthodes et stratégies pour affecter automatiquement une décennie ou une origine à un fragment textuel. Les méthodes appliquées sont basées sur le repérage d'éléments linguistiques tels que les entités nommées et les termes complexes. La stratégie a été choisie à partir de tests réalisés sur des extraits des corpus d'apprentissage. L'article se termine par une analyse des résultats obtenus et une ouverture vers des pistes plus prometteuses.

Abstract. This paper decribes our approch of the two proposed tasks, our methods and strategies to find automatically a decade or an origin for a piece of text. The methods that we applied are founded on marking linguistic elements such as named entities and complex terms. The choice of our strategy arises from tests made upon training corpora extracts. The paper ends with an analysis of the results we obtained, and with an somes propositions of improvement.)

Mots-clés : corpus d'apprentissage, corpus d'évaluation, extraction d'entités.

Keywords: training corpus, evaluation corpus, entities extraction.

1 Introduction

Cet article, plutôt technique, a pour objectif de présenter les méthodes que nous avons mises en œuvre pour répondre aux deux tâches de fouille de texte qui nous étaient proposées. Une première section présente notre approche de ces tâches et l'élaboration d'une méthode d'apprentissage. La section suivante est la description technique des phases d'apprentissage qui utilisent des sorties d'outils linguistiques. Elle est suivie d'une section qui explique le traitement des corpus d'évaluation . Les deux dernières sections présentent la stratégie adoptée pour générer les fichiers de sortie attendus. Nous terminons sur une tentative d'évaluation de nos résultats, au vu des éléments qui nous ont été renvoyés. Pour ne pas alourdir l'article, et parce que nous avons appliqué des méthodes similaires, nous avons pris le parti de ne pas séparer systématiquement la présentation des deux tâches. Cependant, lorsque c'est nécessaire, nous précisons ce qui est spécifique à l'une ou à l'autre.

2 Approche et élaboration des méthodes d'apprentissage

Les deux tâches proposées ont consisté à assigner une *classe* à chaque *portion* ou *article* d'un corpus d'évaluation. Pour chaque tâche un corpus d'apprentissage et un corpus d'évaluation, de structures analogues, nous ont été fournis.

2.1 Tâche n^o 1 : datation des portions

La première tâche est une tentative de datation automatique des *portions*. En effet les classes à affecter sont les quinze décennies qui constituent la période 1800-1944. Il faut noter que les *portions* des corpus sont des fragments textuels et non des articles à part entière. Ils peuvent commencer ou se terminer par une phrase incomplète et balayer plusieurs articles courts qui se suivent dans l'édition d'origine. Il faut noter aussi une forte dégradation du texte, due au procédé de numérisation, erreurs sur certains caractères et mots coupés à mauvais escient.

Les éléments d'un article de presse qui évoquent sa date de production sont sans doute d'abord les événements qui y sont mentionnés. C'est pourquoi, en plus du corpus d'apprentissage qui nous a été fourni, nous avons cherché ce qui se rapprochait le plus d'une base d'événements et nous nous sommes tournées vers les ressources historiques publiées par l'encyclopédie Wikipedia. Il existe en effet des pages qui listent les événements ayant eu lieu chaque année[1].

La question suivante concerne le repérage des événements. Les marqueurs les plus probables sont les *entités nommées*. Certains *termes complexes* comme, par exemple, "abolition de l'esclavage" peuvent aussi être de bons marqueurs parce qu'ils évoquent quelque chose de spécifique qui peut caractériser un événement. Nous avons donc décidé de caractériser les *portions* du corpus en extrayant ces deux catégories d'unités lexicales. Par crainte d'avoir une "couverture" trop faible des *portions* avec seulement les *entités nommées* et *termes complexes*, nous y avons ajouté les noms communs et les verbes présents dans les *portions*.

[1] Par exemple, la page de l'année 1827 se trouve à l'url `http://fr.wikipedia.org/wiki/1827`

2.2 Tâche n° 2 : affectation d'une origine à un article

Dans la seconde tâche il faut affecter à des *articles* deux informations dépendantes l'une de l'autre. Il s'agit d'une part du nom du journal dont est extrait l'article, d'autre part du pays d'origine de ce journal. Il est clair que si le nom du journal est connu avec certitude, il détermine celui du pays. Mais, une réflexion a priori peut laisser penser qu'il est plus facile de détecter les spécificités de la langue de chaque pays. Après avoir parcouru un certain nombre d'articles du corpus d'apprentissage qui nous a été fourni, nous n'avons pas repéré de spécificités notoires de la langue française de France ou de celle du Québec. Nous avons donc fait le choix de ne travailler que sur le nom du journal et d'en déduire automatiquement le nom du pays. Nous avons donc travaillé avec quatre classes *classes* qui sont les quatre noms de journaux. N'ayant pas non plus observé de différence notoire entre entre les langues de chaque pays pour les articles sportifs et pour ceux concernant les informations générales, nous n'avons pas utilisé la catégorie des articles qui nous était fournie dans les corpus.

Le corpus d'apprentissage proposé est constitué d'articles complets, récents, munis d'un titre. Pour classer un article "dans un journal", nous avons cherché à repérer dans quel journal du corpus d'apprentissage son lexique était le plus présent. Se contenter des mots simples aurait sans doute été peu discriminant, c'est pourquoi, en plus des noms communs et des verbes, nous avons considéré les *termes complexes* et les *entités nommées*. Nous n'avons pas distingué le titre du corps de l'article considérant que les éléments d'un titre se retrouvent en général dans le corps de l'article.

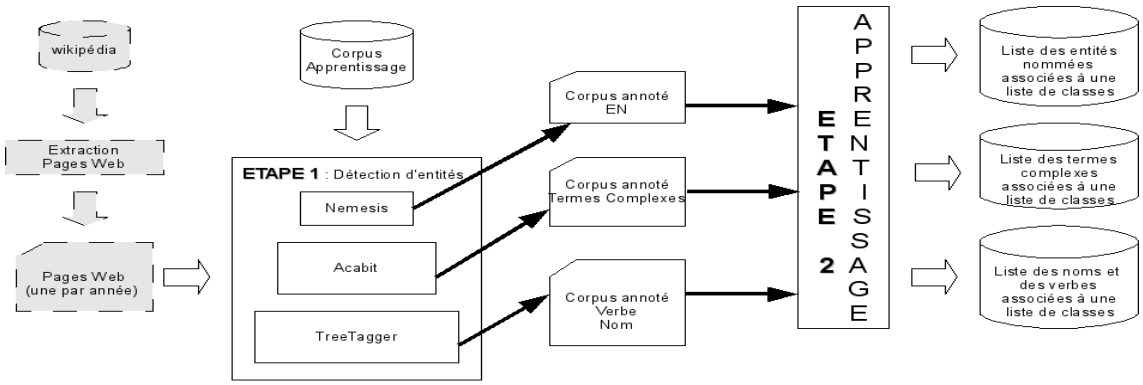

FIG. 1 – Apprentissage d'entités par classe

2.3 Plan de travail

D'un point de vue technique nous avons donc conduit un travail d'apprentissage similaire pour les deux tâches. Il se résume selon le plan ci-dessous et est détaillé dans la section suivante :

- Corpus d'apprentissage :
  - corpus fournis par DEFT2010,
  - pour la tâche n° 1 uniquement : corpus constitué par les pages d'événements de chaque année publiées dans *Wikipedia*.
- Entités extraites pour représenter une portion :
  - entités nommées,
  - termes complexes,
  - noms communs, et verbes.

2.4 Les outils

Pour réaliser les repérages d'entités dans les corpus nous avons utilisé les trois outils suivants dont les deux premiers ont été conçus au LINA :

– Les entités nommées ont été extaites avec NEMESIS, logiciel de reconnaissance, identification et catégorisation automatiques des entités nommées du français (Fourour, 2002).
– Les termes complexes ont été extraits avec ACABIT, programme d'acquisition de terminologie prenant en entrée un texte annoté linguistiquement et retournant une liste ordonnée de candidats termes (Daille, 2003). La vocation première de ce logiciel est de s'appliquer à des corpus de spécialité. Nous en avons un peu détourné l'usage en l'appliquant à un corpus journalistique, au risque, bien sûr, de dégrader les résultats.
– Les noms, verbes, adjectifs et adverbes ont été repérés par le logiciel TREETAGGER, lemmatiseur et outil d'annotation d'un texte en différentes parties du discours (Schmid, 1994).

Les documents résultant de l'apprentissage ont été générés avec des programmes, écrits spécialement pour cetta campagne, en Java, Perl ou Xslt.

3 Mise en œuvre de l'apprentissage

Dans cette section, et dans les suivantes, nous utilisons le terme générique *entité* pour désigner, selon le cas, une *entité nommée*, un *terme complexe*, un *nom commun* ou un *verbe*.

3.1 Construction d'une ressource externe pour la tâche n° 1

Nous avons téléchargé les 145 pages de Wikipédia correspondant chacune aux événements d'une année entre 1800 et 1944, frontières temporelles de l'étude. Puis nous avons fusionné ces pages par décennies, de manière à obtenir un fichier XML de structure identique à celle du corpus d'apprentissage fourni par DEFT2010. Ce « corpus d'événements » se compose de 145 portions, une par année, chaque portion renfermant des expressions relatives à des événements ayant eu lieu dans cette année.

Voici, à titre d'exemple, le début de la portion correspondant à l'année 1837 :

```
<portion id="1837">
<meta>
<journal>WIKIPEDIA</journal>
<date annee="2010" mois="04" jour="02"/>
</meta>
<periode>1830</periode>
<texte>
6 juin : Assassinat du président Diego Portales au Chili par des militaires mutins.
6 septembre : Révolte de la Sabinada à Bahia, au Brésil (fin le 16 mars 1838).
2 octobre : Le Racer's storm, un des ouragans les plus puissants et les plus dévastateurs
...
Septembre : Révolte des Canadiens français, rapidement matée par les forces régulières britanniques.
6 novembre : Affrontement à Montréal entre l'Association patriote « Les Fils de la Liberté »
et les membres du « Doric Club » d'allégeance loyaliste. Saccage de maisons de patriotes.
16 novembre : Arrestation de chefs patriotes. Louis-Joseph Papineau réussit à se rendre aux États-Unis.
19 novembre : Manifestations de Patriotes à Québec.
...
</texte>
</portion>
```

3.2 Extraction des entités des corpus d'apprentissage

Les différentes sortes d'entités (entités nommées, termes complexes, noms communs, verbes) ont été extraites des corpus d'apprentissage par des outils propres à chaque type et cités ci-dessus. Lors de cette extraction nous avons conservé le lien entre chaque entité et toutes les classes (decennies, journaux) dans lesquelles elle a été rencontrée. Pour avoir une image de la distribution des entités dans les corpus d'apprentissage, nous avons calculé le nombre de *portion/articles* différents ($nbport$ pour la tâche n° 1 et $nbart$ pour la tâche n° 2) dans lequel apparaît chaque entité, ainsi que son nombre d'occurrences ($frequence$) dans la classe. Il arrive en effet qu'une entité apparaisse plusieurs fois dans un même *portion/article* ($nbport \leq frequence$ et $nbart \leq frequence$).

À partir de ces extractions nous avons construit des images des corpus d'apprentissage (le corpus de DEFT et elui construit à partir de *Wikipedia* pour la tâche n° 1 et le corpus de DEFT pour la tâche n° 2). Ces images prennent la forme de la liste des entités qu'ils renferment associées aux classes dans lesquelles elles apparaissent. Voici l'exemple de l'entité nommée *Croix Rouge* extraite du corpus *Wikipedia* où elle apparaît dans 7 portions réparties dans 6 décennies, et du terme *action terrestre* extrait du corpus d'apprentissage DEFT sur les origines où elle apparaît dans 3 articles appartenant à 2 journaux :

```
<entite type="EN" ressource="wikipedia">
   <lemme>Croix-Rouge</lemme>
   <decennies>
     <dec nbport='1' frequence='1'>1940</dec>
     <dec nbport='1' frequence='1'>1910</dec>
     <dec nbport='1' frequence='1'>1870</dec>
     <dec nbport='2' frequence='2'>1860</dec>
     <dec nbport='1' frequence='2'>1850</dec>
     <dec nbport='1' frequence='1'>1820</dec>
   </decennies>
 </entite>
```

Une entité nommée et ses décennies extraite du corpus *Wikipedia*

```
<entite type='TC' >
  <lemme>action terrestre</lemme>
  <journaux>
    <journal nbart='1' frequence='1'>D</journal>
    <journal nbart='2' frequence='3'>M</journal>
  </journaux>
</entite>
```

Un terme complexe et ses journaux extrait du corpus DEFT sur les origines

4 Recherche de marqueurs dans les corpus d'évaluation

Afin de déterminer de manière précise la classe la plus probable associée à une portion ou à un article, selon la tâche (section 5), il faut au préalable repérer, dans ces derniers, les entités apprises dans les corpus d'apprentissage (étape 1 de la figure 2).

4.1 Recherche des entités présentes dans le corpus d'évaluation

Pour chaque *portion/article* des corpus d'évaluation, nous avons repéré les différentes entités sur lesquelles nous avions décidé de travailler : les entités nommées (EN), les termes complexes (TC), les verbes (V) et les noms (N) avec les mêmes outils que ceux utilisés sur les corpus d'apprentissage. On obtient, pour chaque corpus d'évaluation, quatre listes (une par type d'entité) conservant la structure en *portion/article*, des entités qui en ont été extraites.

Voici par exemple quelques entités nommées d'une portion du corpus d'évaluation de la tâche n° 1 et quelques termes d'un article du corpus d'évaluation de la tâche n° 2 :

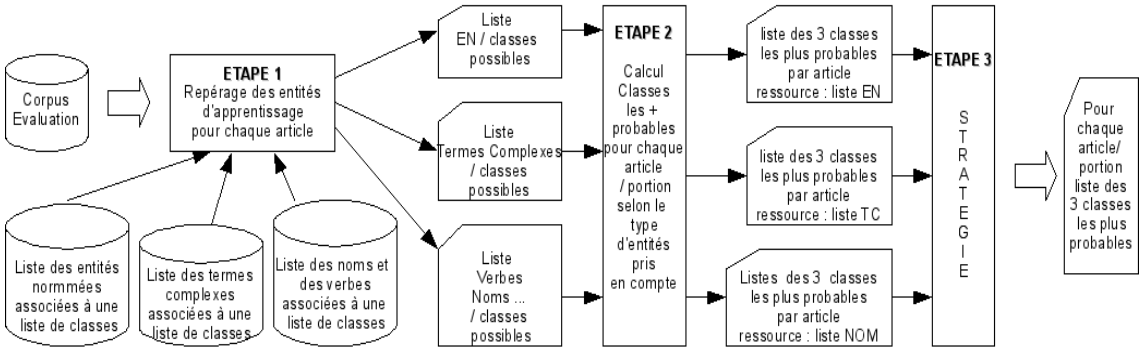

FIG. 2 – Détection de la classe d'une portion ou d'un article

```
<portion id="2143">
 ...
 <entite type="EN" nb="1">
  <lemme>Rodez</lemme>
 </entite>
 <entite type="EN" nb="1">
  <lemme>monts d' Aubrac</lemme>
 </entite>
 <entite type="EN" nb="1">
  <lemme>Louis-Philippe</lemme>
 </entite>
 ...
</portion>
```

Extrait du corpus d'évaluation de la tâche n° 1

```
<article id='1'>
   ...
    <entite type='TC' nb='1'>
      <lemme>jeu olympique de hiver</lemme>
    </entite>
    <entite type='TC' nb='1'>
      <lemme>médaille de or</lemme>
    </entite>
    <entite type='TC' nb='1'>
      <lemme>vache maigre</lemme>
    </entite>
    ...
</article>
```

Extrait du corpus d'évaluation de la tâche n° 2

4.2 Enrichissement des entités repérées dans les corpus d'évaluation par les données d'apprentissage

L'étape suivante consiste à enrichir les entités rencontrées dans les corpus d'évaluation par les différentes classes qui leur ont été associées lors de la phase d'apprentissage (étape 1 de la figure 2).

Voici les entités nommées enrichies de la portion de l'exemple précédent (tâche n° 1) :

```
<portion id="2143">
    ...
     <entite type="EN">
        <lemme>Rodez</lemme>
        <decennies>
           <dec nbport="1" ressource="appr">1890</dec>
           <dec nbport="1" ressource="appr">1850</dec>
           <dec nbport="1" ressource="appr">1910</dec>
           <dec nbport="1" ressource="appr">1930</dec>
           <dec nbport="1" ressource="wikipedia">1930</dec>
        </decennies>
     </entite>
     <entite type="EN">
        <lemme>monts d' Aubrac</lemme>
        <decennies/>
     </entite>
     <entite type="EN">
        <lemme>Louis-Philippe</lemme>
```

```
        <decennies>
           <dec nbport="1" ressource="appr">1870</dec>
           <dec nbport="2" ressource="appr">1830</dec>
           <dec nbport="1" ressource="appr">1840</dec>
           <dec nbport="1" ressource="wikipedia">1860</dec>
           <dec nbport="2" ressource="wikipedia">1840</dec>
           <dec nbport="2" ressource="wikipedia">1830</dec>
        </decennies>
     </entite>
    ...
  </portion>
```

Deux des trois entités nommées de la portion 2143 appartiennent aux listes d'entités nommées apprises avec le corpus d'apprentissage : "Rodez" et "Louis Philippe" : on enrichit ainsi ces deux entités par la liste des décennies probables en conservant pour chaque décennie, le nombre de portions (*nbport*)dans lesquels on a trouvé l'entité et l'information nous indiquant dans quel corpus d'apprentissage l'entité a été trouvée (*ressource*).

De même pour la tâche 2, voici les termes complexes de l'article de l'exemple précédent, enrichis des informations sur les origines ramenées pa l'apprentissage :

```
<article id="1">
   ...
   <entite type="TC">
       <lemme>jeu olympique de hiver</lemme>
       <journaux>
          <journal nbart="6">D</journal>
          <journal nbart="1">M</journal>
          <journal nbart="2">P</journal>
       </journaux>
    </entite>
    <entite type="TC">
       <lemme>médaille de or</lemme>
       <journaux>
          <journal nbart="15">D</journal>
          <journal nbart="6">E</journal>
          <journal nbart="4">M</journal>
          <journal nbart="15">P</journal>
       </journaux>
    </entite>
        <entite type="TC">
       <lemme>vache maigre</lemme>
       <journaux>
          <journal nbart="1">M</journal>
       </journaux>
    </entite>
 </article>
```

Ici tous les termes complexes repérés dans l'article 1 ont été enrichis car ils sont tous présents dans le corpus d'apprentissage. Pour la tâche n° 2, on garde, pour chaque journal concerné par une entité, le nombre d'articles (*nbart*) dans lequel a été repérée cette entité pour ce journal.

4.3 Analyse du corpus d'évaluation

Suite à l'enrichissement de certaines entités du corpus d'évaluation, nous avons voulu étudier de manière plus précise ce corpus :

– en évaluant le nombre d'entités enrichies présentes dans ce corpus pour voir la couverture de nos listes d'apprentissage

– en mesurant l'ambiguïté des classes associées à chaque entité

4.3.1 Tâche 1

Dans le corpus d'évaluation nous avons reconnu un certain nombre d'occurrences d'entités de chaque type (Nbre d'occ), dont un certain nombre enrichies par nos listes d'apprentissage (Nbre d'occ enrichies) :

Type	Nbre d'occ.	Nbre d'occ. enrichies	Rapport
Entités Nommées	17760	9756	54,93 %
Termes Complexes	37311	9232	24,74 %
Noms	105727	102305	96,76 %
Verbes	60158	59578	99,04 %
	220956	180871	81,86 %

Comme nous l'avons constaté dans la section 2.1, les types "entités nommées" et "termes complexes" semblent les plus pertinents pour déterminer la décennie d'une portion, puisqu'ils permettent de faire référence à des événements ayant lieu lors de la décennie. Toutefois, on constate qu'une entité nommée sur deux du corpus d'évaluation n'est pas présente dans les corpus d'apprentissage et qu'il en est de même pour un terme complexe sur quatre. La tâche 1 semble donc difficile à résoudre, par un manque de connaissances initiales. De plus les entités enrichies semblent loin d'être spécifiques à une décennie particulière au regard du pourcentage d'entités n'étant associées qu'à une seule décennies :

Nbre de décennies associées : nb	$nb = 1$	$2 \leq nb \leq 4$	$5 \leq nb \leq 10$	$nb > 10$
Entités Nommées	15,44 %	15,71 %	14,84 %	54,01 %
Termes Complexes	42,40 %	32,06 %	14,61 %	10,93 %

On constate que pour les entités nommées la tâche est d'autant plus complexe que 54,01 % des entités sont associées à plus de dix décennies. En regardant de plus près, on constate qu'il s'agit essentiellement des noms de régions, de pays, de villes.

4.3.2 Tâche 2

On réalise les mêmes calculs pour la tâche 2 :

Type	Nbre d'occ.	Nbre d'occ. enrichies	Rapport
Entités Nommées	38048	28089	73,83 %
Termes Complexes	89118	37762	42,37 %
Noms	180865	172066	95,14 %
Verbes	104576	103915	99,37 %
	412607	341832	82,85 %

Pour la tâche 2, le nombre d'occurrences d'entités nommées et de termes complexes augmente considérablement et devrait donc permettre de résoudre plus facilement la tâche. Mais on constate que parmi les occurrences de ces entités, beaucoup sont présentes dans plus d'un journal et pas forcément du même pays.

Journaux associés à une entité	d'un seul journal	seulement français	seulement québécois
Entités Nommées	17,64 %	14,09 %	20,37 %
Termes Complexes	36,98 %	24,25 %	26,37 %

Les deux tâches ne semblent pas simples à résoudre au vue des connaissances acquises avec le corpus d'apprentissage.

5 Résolution des tâches

Pour décider de la stratégie la plus adaptée pour chaque tâche, nous avons partitionné chaque corpus d'apprentissage en deux : un corpus test et un corpus d'apprentissage partiel.

Ainsi pour la tâche 1, nous avons constitué un corpus de test de 350 portions et un corpus d'apprentissage partiel de 2371 portions à partir duquel nous avons réalisé un apprentissage comme il a été décrit dans la section 3.

Pour la tâche 2, nous avons constitué de même un corpus de test de 370 articles et un corpus d'apprentissage partiel de 3349 articles où comme pour la tâche 1, ce dernier a servi dans la phase d'apprentissage pour ce test.

L'objectif était d'observer les résultats sur ce test et de déterminer quelles étaient les types d'entités à utiliser pour répondre au mieux à chacune des tâches : les entités nommées ? fusionnées à un ou plusieurs autres types ? ou de fusionner d'autres entités ?

Tous les résultats de cette section porteront sur les corpus test que nous avons fabriqués selon la description ci-dessus.

5.1 Calcul des classes les plus probables pour chaque portion/article

La deuxième étape, pour résoudre les différentes tâches consiste, à partir des entités enrichies, à déterminer les classes les plus probables pour chaque portion / article (voir étape 2 figure 2).

Ainsi pour chaque type d'entité, il s'agit de retourner la liste des classes susceptibles d'être la classe recherchée. Pour se faire, on fusionne toutes les classes de même type retournées par les entités enrichies.

Ainsi pour l'article 1 de la tâche 2 concernant les termes complexes, on obtiendra la liste suivante :

```
<article id="1">
   <journal nbentites="3" nbart="6">M</journal>
   <journal nbentites="2" nbart="21">D</journal>
   <journal nbentites="2" nbart="17">P</journal>
   <journal nbentites="1" nbart="6">E</journal>
 </article>
```

Cela signifie que dans l'article 1, ont été repérés 3 termes complexes issus de 6 articles du journal *Le Monde* du corpus d'apprentissage, 2 termes complexes dans 21 articles du journal *Le Devoir*, 2 termes complexes dans 17 articles du journal *La Presse* et 1 terme complexe dans 6 articles du journal *L'est républicain*.

Une fois la tâche de fusion pour chaque type d'entité réalisée, il faut classer ces différentes propositions. Plusieurs tests ont été effectués pour trouver le meilleur tri possible :

1. en fonction du nombre d'entités,
2. en fonction du nombre de portions (*nbport*) ou d'articles (*nbart*) du corpus d'apprentissage où ont été apprises les entités,
3. en fonction du nombre d'entités puis du nombre de portions (*nbport*) ou d'articles (*nbart*),
4. en fonction d'un rapport entre le nombre d'entités et le nombre de portions ou d'articles ...

Au final, le tri 3 est celui qui amène le plus grand nombre de bons résultats pour la tâche 1, donc pour chaque article dans chaque fichier correspondant à un type d'entité, les classes seront triées selon le nombre d'entités puis le nombre de portions ou d'articles.

Maintenant il s'agit de déterminer quel types d'entités utiliser pour résoudre les différentes tâches.

5.2 Stratégie

C'est au niveau de la stratégie que l'on peut noter une différence entre les deux tâches auxquelles nous avons participé.

5.2.1 Recherche d'une stratégie pour la tâche 1

Pour la tâche 1, la stratégie doit tenir compte du nombre important de décennies associées à chaque entité (comme nous l'avons vu dans la section 4.3), quel que soit le type d'entité.

Prenons notre exemple, pour la portion 2143, nous aurons pour le type d'entités EN le résultat suivant :

```
 <portion id="2143">
     <decennies>
        <dec nbentites="1" nbport="4" ressource="appr;wikipedia">1830</dec>
        <dec nbentites="1" nbport="2" ressource="appr;wikipedia">1930</dec>
        <dec nbentites="1" nbport="3" ressource="appr">1840</dec>
<dec nbentites="1" nbport="1" ressource="appr">1890</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1850</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1910</dec>
        <dec nbentites="1" nbport="1" ressource="appr">1870</dec>
        <dec nbentites="1" nbport="1" ressource="wikipedia">1860</dec>
      </decennies>
   </portion>
```

Pour la portion 2143, 8 décennies sont proposées et la décennie 1830 semble la décennie la plus probable pour la publication de la portion. Nous rappelons, en effet, que les décennies proposées sont triées par nombre d'entités puis par nombre de portions.

Plusieurs questions se posent ainsi à l'issue du calcul :
– Quels types d'entités utiliser pour avoir les meilleurs résultats ?
– Comment combiner les résultats obtenus pour chaque type d'entité si nous utilisons plusieurs entités pour résoudre la tâche ?
– Doit on prendre en compte, pour chaque type d'entités, toutes les décennies qui lui sont associées ou seulement les 3 meilleures ?

A partir du corpus test que nous avons extrait du corpus d'apprentissage, plusieurs tests ont été menés en faisant varier plusieurs paramètres (comme le tri des décennies, le nombre de décennies pris en compte lors de la combinaison, etc.). Nous présentons ci-dessous l'intervalle des précisions[2] calculées sur les différents tests, pour les différents types d'entités et leur combinaison.

Types Entités	Précision Minimale	Précision Maximale
EN	30,57 %	32,86 %
NOM	37,14 %	40 %
TC	29,14 %	32,86 %
VERBE	28,57 %	31,43 %
EN-TC-NOM-VERBE	36,57 %	42,86 %

D'autres combinaisons ont été testées mais c'est la combinaison des 4 types d'entités qui a donné les meilleurs résultats.

Pour les 3 runs que nous avons soumis, nous avons donc pris en compte tous les types d'entités (EN, TC, NOM et VERBE).

La différence entre les 3 runs porte d'une part sur la méthode de combinaison des résultats :
– RUN 1 : combinaison pour chaque portion des 3 premières décennies retournées par chaque type d'entité,
– RUN 2 et 3 : combinaison pour chaque portion de TOUTES les décennies retournées par chaque type d'entité.
d'autre part sur la manière de trier le résultat de la combinaison :

– RUN 1 et 2 : tri en fonction du nombre d'entités du corpus d'apprentissage présentes dans le corpus d'évaluation, puis du nombre de portions dans lesquelles ces entités étaient présentes dans le corpus d'apprentissage, pour cette décennie,
– RUN 3 : tri en fonction du nombre d'entités du corpus d'apprentissage présentes dans le corpus d'évaluation, puis du nombre d'entités ayant retourné la décennie en première position, puis du nombre de portions.

Avec le corpus test, la combinaison de TOUTES les décennies retournées par chaque type d'entité améliore les résultats, mais pas de manière flagrante d'où les deux runs proposés. Le taux de confiance d'une décennie pour une portion est égal au nombre de portions dans lesquelles ses entités ont été apprises, par rapport à la somme des portions dans lesquelles les entités des trois décennies retournées ont été apprises.

5.2.2 Recherche d'une stratégie pour la tâche 2

Pour déterminer le journal dans lequel est paru l'article, nous avons fait varier également les différents types d'entités pour définir les 3 journaux les plus probables par article. Nous avons ainsi calculé le pourcentage de bonnes réponses (voir le tableau ci dessous), retournées en première position (Pos1), en deuxième position (Pos2) et en troisième (Pos3).

[2]Nombre de portions retournant la bonne décennie dans les trois premières proposées / Nombre de portions

Types Entité	Pos1	Pos2	Pos3
EN	58,65 %	26,76 %	8,65 %
NOM	43,24 %	26,22 %	17,84 %
TC	55,95 %	27,03 %	11,89 %
VERBE	32,16 %	24,59 %	21,08 %
EN-TC	59,73 %	26,49 %	10,54 %
EN-NOM	55,68 %	28,11 %	9,46 %
TC-NOM	52,97 %	26,22 %	13,78 %
EN-TC-NOM	58,92 %	27,84 %	9,46 %
EN-TC-VERBE	59,73 %	25,68 %	10,54 %
EN-TC-NOM-VERBE	60,81 %	24,05 %	10,54 %

L'utilisation des entités nommées apprises par le biais du corpus d'apprentissage partiel semble indispensable au vu du pourcentage de bonnes réponses en première position (58,65 %), comme les termes complexes (55,95 %).

Ainsi pour nos 3 runs, nous choisissons les 3 stratégies suivantes :

1. Entités Nommées + Termes Complexes
2. Entités Nommées + Termes Complexes + Noms
3. Entités Nommées + Termes Complexes + Noms + Verbes

Les deux premières stratégies nous permettent en effet d'obtenir les meilleurs résultats quant au nombre de bonnes réponses parmi les 3 retournées, et la troisième stratégie correspond au meilleur taux de bonnes réponses en première position.

Il est évident qu'il aurait été aussi très intéressant d'évaluer la tâche 2 avec un seul type d'entité, notamment pour les entités nommées et les termes complexes, puisque les résultats sont similaires. Le tri des résultats des combinaisons de la tâche 2 est identique à celui des runs 1 et 2 de la tâche 1. Le taux de confiance en un journal pour un article est égal au nombre d'articles dans lesquels ses entités ont été apprises, par rapport à la somme des articles dans lesquels les entités des trois journaux retournés ont été apprises.

Comme nous l'avons dit au début de l'article, nous avons réalisé notre apprentissage sur le nom de journal. Pour déterminer le pays où a été publié l'article, nous nous basons sur les résultats obtenus pour la détermination du journal (3 propositions maximum). Le taux de confiance dans le pays est déterminé en fonction de la somme des taux de confiance des journaux qui lui sont associés (*Le Devoir* et *La Presse* pour le *Québec* et L'*Est Républicain* et *Le Monde* pour la France).

6 Évaluation des résultats et conclusion

6.1 Tâche 1

Nos résultats :

	Macro Rappel	Macro Précision	Macro F-mesure
Run 1	5,1 %	5 %	5 %
Run 2	5,3 %	5,2 %	5,3 %
Run 3	5,3 %	5,2 %	5,3 %

À la lecture de quelques portions du corpus d'apprentissage, nous n'avons pas relevé de marques linguistiques spécifiques à une période particulière, encore moins à une décennie. C'est pourquoi la seule méthode qui nous a semblé possible était de repérer des éléments de lexique dans le corpus d'apprentissage.

Au vu de la moyenne des résultats obtenus par l'ensemble des participants, le corpus d'apprentissage semble ne pas renfermer suffisamment d'informations pour dater les portions du corpus d'évaluation. Il aurait d'autre part été intéressant de pouvoir travailler sur des données non dégradées par l'OCR car nous n'avons pas eu le temps de mesurer l'impact de ces erreurs.

En ce qui concerne notre travail nous sommes conscientes qu'il nous aurait fallu mieux cibler chaque catégorie d'entité :
– ne conserver comme entités nommées que celles qui sont des vrais marqueurs d'événements c'est à dire les noms propres de personnes, de lieux (typiques comme, par exemple, *jardin public* mais non géographiques), d'événements (comme par exemple *Exposition Universelle* ou *Jeux Olympiques*)
– conduire un travail de réflexion approfondi sur la nature des termes complexes à conserver,
– ne conserver que certaines catégories de noms et de verbes, ce qui aurait nécessité d'autres ressources externes.

6.2 Tâche 2

	Rappel	Précision	F-mesure
* Run 1 - Pays	72,1 %	72,5 %	72,3 %
Run 1 - Journal	41,9 %	43 %	42,5 %
Run 2 - Pays	69,2 %	69,5 %	69,4 %
Run 2 - Journal	39,6 %	41,3 %	40,4 %
Run 3 - Pays	68,5 %	68,8 %	68,7 %
Run 3 - Journal	39,3 %	41,4 %	40,3 %

De manière analogue nous n'avons pas trouvé dans le corpus d'apprentissage de marques linguistiques spécifiques à la langue française du Québec ou de la France. Ceci nous a donc conduites vers des méthodes analogues pour les deux tâches.

Toutefois le corpus d'apprentissage a constitué une ressource mieux adaptée à la tâche demandée.

Comme nous l'avons constaté sur notre corpus test, ce sont les entités nommées et les termes complexes qui permettent d'obtenir les meilleurs résultats. Comme pour la tâche précédente il aurait fallu mieux cibler chaque catégorie d'entités.

Nous avons ramené la tâche à la reconnaissance des journaux, à cause de la dépendance pays, journal. Il serait sans doute intéressant de tester une stratégie en deux temps :
– apprentissage puis reconnaissance du pays indépendamment du journal,
– apprentissage puis reconnaissance du journal sur des corpus réduits aux pays.
Nous avons été très intéressées par ce travail d'investigation. Nous pensons qu'il y a encore beaucoup de travail a effectuer sur chacun de ces thèmes, tant dans la construction des ressources d'apprentissage que dans la mise en œuvre de nos méthodes.

Références

DAILLE B. (2003). Information Extraction in the Web Era. In M. PAZIENZA, Ed., *Terminology Mining*, p. 29–44. Springer.

FOUROUR N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In J.-M. PIERREL, Ed., *Actes de TALN 2002 (Traitement automatique des langues naturelles)*, p. 265–274, Nancy : ATALA ATILF.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing(NeMLaP-1)*, p. 44–49.