Décennie d'un article de journal par analyse statistique et lexicale

Pierre ALBERT Flora BADIN Maxime DELORME Nadège DEVOS
Sophie PAPAZOGLOU Jean SIMARD[1]
(1) CNRS–LIMSI, 91403 ORSAY
flora.badin@limsi.fr

Résumé. Dans le cadre du DÉfi de Fouille de Texte (DEFT) 2010, nous avons proposé une méthode permettant de dater des articles de journaux. L'étude du corpus a permis de relever les différentes caractéristiques des articles identifiables de façon automatique telles que la présence d'entités nommées et de variations orthographiques. Une approche mixte se basant à la fois sur ces propriétés linguistiques et sur les statistiques a été développée. Grâce à cette approche, les résultats ont permis d'obtenir une F-mesure allant jusqu'à 0.338.

Abstract. For the 2010 edition of DEFT, we proposed a method for dating newspaper articles. Studying the corpus allowed us to extract distinctive features of the articles. These features such as named entities and orthographic variations are automatically identificable. We developed a mixed approach based on the recognition of those linguistic properties and on statistics. Thanks to this method we have been able to achieve F-measures up to 0.338.

Mots-clés : Analyse statistique, entités nommées, fouille de texte, algorithme d'apprentissage.

Keywords: Statistical analysis, named entities, text-mining, learning algorithms.

1 Introduction

La fouille de texte consiste à extraire, en s'aidant de théories linguistiques, une information précise d'un document. Ainsi, la tâche proposée par le concours DEFT 2010 a pour objectif la datation d'articles de journaux. La connaissance de la date d'un document est très important en Traitement Automatique des Langues (TAL) puisqu'il permet de replacer le document dans son contexte historique afin de mieux l'interpréter.

L'utilisation de GALLICA, d'où sont extraits ces articles, est la seule ressource non-autorisée. Si un minimum de traitement logique est possible, notamment à travers les entités nommées, il est aussi intéressant de s'intéresser aux marqueurs caractéristiques d'une époque. Nonobstant l'efficacité de ces traitements, combiner ceux-ci avec un apprentissage statistique est indispensable pour éviter les biais et les absences d'information. Pour ce dernier, deux approches ont été utilisées, l'une portant sur la fréquence d'apparition des mots et l'autre sur leur enchainement. Par l'association de ces méthodes, nous avons estimé les décennies plausibles pour les extraits proposés. Nos résultats montrent la pertinence de cette approche, qui pourrait cependant bénéficier de développements supplémentaires afin d'être réellement efficace.

Dans une première section, nous revenons sur les caractéristiques du corpus d'entraînement qui nous a été fourni dans le cadre du concours. La seconde section expose les différentes solutions que nous avons retenues pour dater les articles. Une brève troisième section décrit la méthode de soumission de nos résultats. La dernière section permet de conclure et propose quelques évolutions possibles qui pourraient s'ajouter aux méthodes existantes ainsi que des perspectives qui pourraient être intéressantes.

2 Corpus

Durant ce DÉfi Fouille de Texte 2010, un corpus de plus de 6315 articles de journaux a été mis à notre disposition. Environ 60 % a été annoté et nous a servi de corpus d'entraînement. Le reste du corpus a été utilisé pour l'évaluation de notre modèle et a donné lieu à la soumission pour le concours DEFT 2010. Nous allons ici décrire plus en détails le corpus d'entraînement. Puis nous reviendrons sur l'ensemble des caractéristiques de ce corpus qui ont orienté notre réflexion vers une méthode d'analyse statistique et lexicale.

2.1 Le corpus d'entraînement

Le corpus d'entraînement est composé d'articles de journaux parus entre 1800 et 1944. Tous les articles sont issus de deux journaux : *La Croix* et le *Journal des Débats et des Décrets* devenu ensuite le *Journal de l'Empire* puis le *Journal des Débats Politiques et Littéraires*. Les sujets qui y sont traités sont très variés, allant de l'actualité internationale (concernant les guerres par exemple) à l'abolition de lois en passant par le simple fait divers (comme la météorologie). En dehors du journal *La Croix*, les articles sont des résumés de débats ayant eu lieu dans divers corps administratifs ou gouvernementaux. Ces articles ont été numérisés par reconnaissance optique de caractères (OCR pour *Optical Character Recognition*) à partir de la version papier des journaux puis découpés en 3594 articles. Ils ont ensuite été structurés au format XML pour les besoins de DEFT 2010. La structure XML donne pour chaque article les informations sur le nom du journal, la date de publication et la décennie. Toutes les années pouvant apparaître dans le corps

de texte des articles ont été masquées par des balises `<annee />`.

2.2 Caractéristiques du corpus

Une annotation manuelle d'une cinquantaine d'articles a permis de constater un ensemble de caractéristiques propres au corpus.

La reconnaissance optique de caractères est une technique encore imparfaite ne permettant pas de numériser un texte sans erreur. De plus, l'âge [1], la qualité du papier ou de l'encre, la présence de pliures, de taches ou même de traits de présentation (pour séparer les colonnes par exemple) sur les journaux de notre corpus ont augmenté de façon considérable le nombre d'erreurs commises par le moteur de reconnaissance optique de caractères. Celles-ci sont de natures diverses : erreur d'identification du caractère (*lecteyrs*), insertion d'espace dans un mot (*pro position*) ou omission d'espace entre deux mots (*laquestion*), trait ou pliure verticale sur le journal identifié en tant que caractère (*jour |**i** est demandé*) ou pliure horizontale perturbant la reconnaissance optique sur l'ensemble d'une ligne (*apeix:evoir*).

Ces erreurs ont mené à plusieurs pistes, certaines exploitées et expliquées dans la suite de ce papier, d'autres non-exploitées.

Par exemple, certaines erreurs se sont révélées être liées aux réformes de l'orthographe durant la période concernée par les articles (voir section 3.1).

Une autre piste aurait pu consister à essayer de trouver une corrélation entre le nombre, la fréquence ou le type d'erreurs dans un texte avec la période de l'article. Mais cela suppose de pouvoir identifier les erreurs ce qui est une tâche très difficile. Cette piste n'a donc pas été creusée.

Nous avons également constaté que les versions numériques des articles contiennent des doubles espaces à intervalles réguliers. Ces doubles espaces correspondent aux retours à la ligne en fin de colonnes. Un outil a été développé pour étudier la corrélation entre la largeur des colonnes, les noms des journaux et l'année de parution des articles. Malheureusement, ce travail n'a pas donné de résultat suffisamment pertinent pour être exploitable.

Les caractéristiques de ce corpus, combinées au faible nombre de mots par article (300 mots), ne favorisent pas l'utilisation de méthodes habituelles en Traitement Automatique des Langues (TAL). C'est pour ces raisons que nous avons délibérément évité toute étude sémantique des textes. Nous nous sommes plus concentrés sur des approches statistiques et lexicales.

3 Les méthodes

3.1 Les réformes de l'orthographe

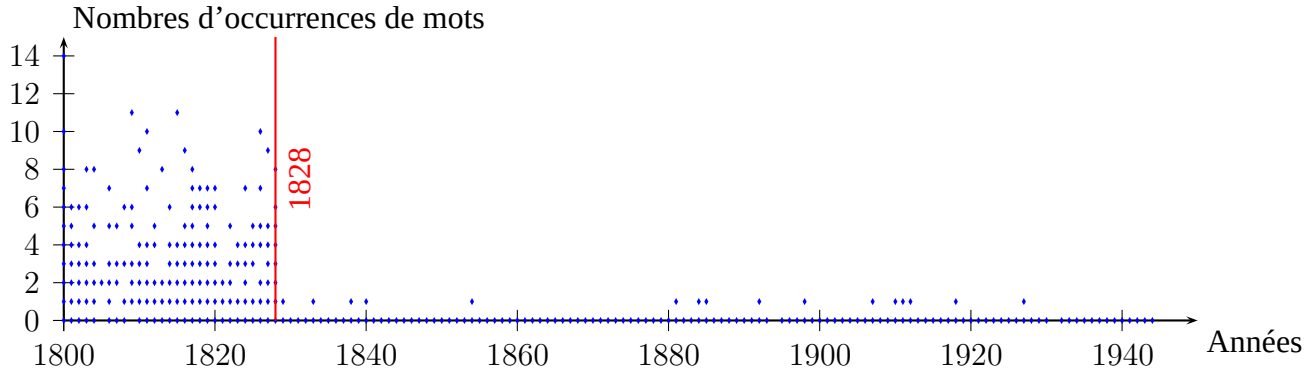
Concernant la période couverte par le corpus, deux réformes de l'orthographe ont eu lieu. Tout d'abord, celle de 1835 (voir [Académie française, 1835]) puis celle de 1878 (voir [Académie française, 1878]).

1. N'oublions pas que certains de ces articles ont plus de 200 ans.

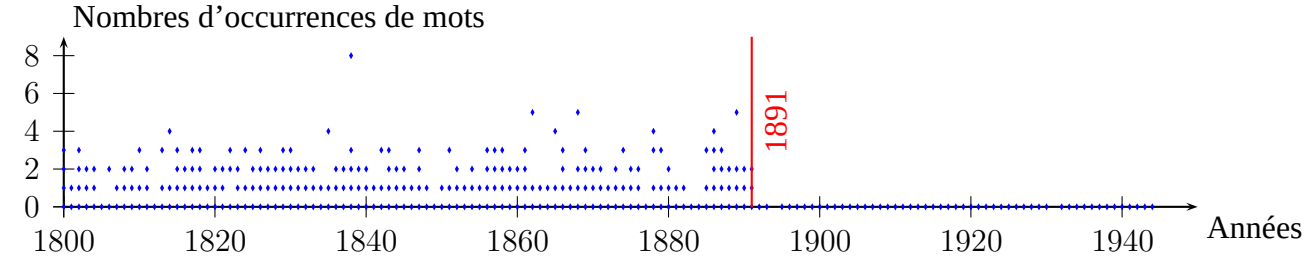
L'annotation manuelle de quelques articles nous a permis d'identifier deux éléments récurrents de la réforme de 1835 :
– La combinaison de lettres *oi* s'est vue transformée en *ai* dans une grande majorité des mots ;
– Le pluriel des mots en *nt* est passé de *ns* à *nts*.
Une étude a été réalisée sur le corpus d'entraînement afin de vérifier que ces deux règles fournissent une bonne indication concernant la date de l'article en question. Pour cela, un algorithme simple mais relativement efficace a été appliqué :

1. Filtrer tous les mots se terminant par *ois*, *oit* et *oient* ;
2. Dans cette liste de mots, retirer tous les mots se trouvant dans le dictionnaire contemporain de la langue française (comme *trois*, *aperçois*, *voient*) ;
3. Parmi les mots restants, remplacer la lettre *o* dans *ois*, *oit* et *oient* par la lettre *a* ;
4. Vérifier que le nouveau mot se trouve dans le dictionnaire contemporain de la langue française :
   – Si le mot se trouve dans le dictionnaire, il est comptabilisé ;
   – Si le mot ne se trouve pas dans le dictionnaire, il est enlevé de la liste.

Concernant les pluriels des mots se terminant par *nt*, le même algorithme a été utilisé. L'application de ces deux algorithmes sur le corpus d'entraînement a permis d'obtenir un couple (A, N) pour chaque article où A est l'année de l'article et N est le nombre d'occurrences constatées concernant la réforme dans ce même article. L'ensemble des couples répertoriés sont représentés sur la Figure 1 (un point sur le graphique pouvant représenter plusieurs articles). Ils permettent de mettre en évidence deux limites distinctes.



(a) La terminaison des mots en *oi* (réforme de 1835)



(b) Le pluriel des mots terminant par *nt* (réforme de 1878)

Figure 1 – Les réformes de l'orthographe.

La mesure de ces deux indices révèle avec certitude deux filtres sur les articles.

Le premier filtre concerne les mots dont la terminaison contient *oi*. Pour chaque article contenant plus de deux occurrences, la probabilité que la date de l'article soit supérieure à 1828 est nulle. Dans le cas où l'article contient une occurrence, la probabilité que la date de l'article soit supérieure à 1828 est $17/187$ soit 0.091 (proportion d'articles contenant une occurrence et dont la date est supérieure à 1828 par rapport au total des articles contenant une occurrence).

Le second filtre concerne les mots au pluriel et se terminant par *ns*. Pour chaque article contenant au moins une occurrence, la probabilité que la date de l'article soit supérieure à 1891 est nulle.

Dans le cas où aucun des deux indices n'est détecté dans l'article, aucun filtre n'est appliqué.

3.2 Les entités nommées

Concernant les entités nommées, les articles sont en général assez fournis. Nous trouvons aussi bien des entités nommées du type *M. Dupond*, *Napoléon III* que *France* ou *boulevard Haussmann*. Étant donnée la casse particulière des entités nommées, ce sont potentiellement des groupes de mots facilement identifiables dans un article.

Ce sont les noms propres désignant une personne qui ont été choisis comme axe de recherche. En effet, ils contiennent plusieurs caractéristiques (voir [Ehrmann, 2008]) :
– Il est possible d'associer une date de naissance à un nom de personne ce qui permettrait de filtrer les années possibles de l'article ;
– Le nom d'une personnalité est souvent accompagné de préfixes distinctifs tels que *M.*, *docteur* ou de suffixes tel qu'un chiffre romain (*Napoléon III*).
Voici une liste non-exhaustive des préfixes permettant d'identifier un nom : *M.*, *Madame*, *Mlle*, *Président*, *EVEQUE*, *Lord*… Concernant le nom propre, des préfixes ont également été identifiés tels que *Van den* comme dans *Van den Berghe* ou *de la* comme dans *de la Fontaine*. Pour les suffixes, seuls les chiffres romains ont été identifiés.

Si un ou plusieurs noms sont relevés dans un article, chaque nom fait l'objet d'une recherche sur Internet par le biais du site UNIVERSALIS afin de trouver une date de naissance. Bien évidemment, un résultat n'est pas systématiquement trouvé. Parfois, le résultat peut également être inutile. Par exemple, si le nom *Aristote* est trouvé dans un article, il est évident que sa date de naissance ne nous permettra pas d'en déduire une année pour un article datant de la période 1800–1944.

Si une date de naissance est trouvée, une faible probabilité sera donnée aux années antérieures à cette date. Dans notre cas, cette probabilité est 0.2. Ensuite, nous avons estimé que la probabilité qu'une personne soit citée dans la presse avant l'âge de ces 20 ans est relativement faible mais qu'elle augmentait linéairement avec l'âge. Nous avons donc une probabilité linéairement montante entre la date de naissance de la personne et la date de ces 20 ans. Après ces 20 ans, c'est une probabilité de 0.8 qui est donnée (voir Figure 2 page suivante).

3.3 Apprentissage par l'utilisation de *Conditional Random Fields*

Afin de pallier au manque de précision des outils lexicaux, une étude statistique par apprentissage, fondée en partie sur la méthode récente des *Conditional Random Fields* (CRF), est utilisée [Lafferty *et al.*, 2001].

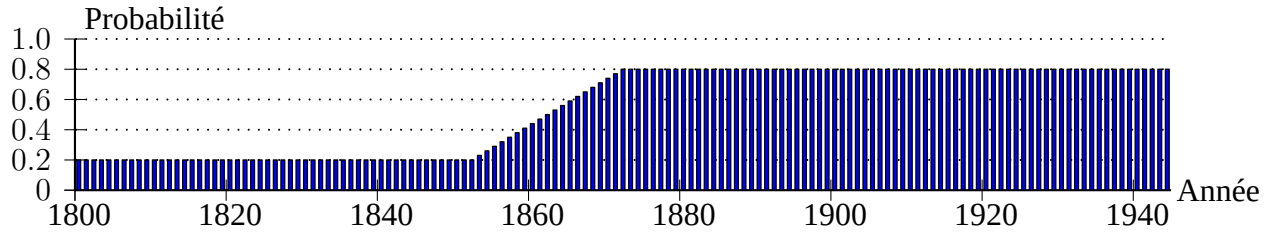


Figure 2 – Probabilité appliquée sur un article en fonction d'une date de naissance : Exemple avec Calamity JANE (1852–1903).

Dans ce cadre, un article est perçu comme un graphe linéaire de mots. Chaque mot est un nœud relié à une étiquette qui le caractérise (dans ce cas, sa date). L'apprentissage se fait alors en calculant les probabilités de transition du mot à chaque étiquette en fonction de son contexte. Le principe des CRF est voisin de celui des automates de MARKOV à états cachés. Dans ce dernier, la séquence de l'automate n'est pas connue et seul le résultat est visible. Le CRF en est une généralisation non-orientée, avec un nombre de contraintes illimitées, et dont les probabilités de transition varient en fonction de la séquence analysée [Wallach, 2004]. Les CRF sont particulièrement intéressants pour le traitement automatique des langues. Dans le cadre de l'étude de la variation diachronique, l'étiquetage se fait sur la période (la décennie dans notre cas). Parmi les différentes implémentations de l'algorithme, CRF++ est adapté au cadre des graphes linéaires.

La détermination de la date d'un mot n'étant pas liée aux mots qui l'entourent mais à leurs étiquettes, la règle d'entraînement est constituée des cinq unigrammes simples suivants :
– $U\%x[-2,0]$;
– $U\%x[-1,0]$;
– $U\%x[0,0]$;
– $U\%x[1,0]$;
– $U\%x[2,0]$;
Nous observons le contexte d'un mot sur une fenêtre totale de cinq mots et de leurs étiquettes.

3.3.1 Traitement du corpus

Le corpus a été découpé afin de prendre en compte les mots ainsi que la ponctuation qui a été ajoutée dans un second temps suite aux faibles résultats d'une première évaluation. L'apport de données sur ce corpus de taille relativement faible a eu une incidence non-négligeable. Les mots reconnus comme mal numérisés ont aussi été conservés afin de prendre en compte les erreurs récurrentes et potentiellement des schémas caractéristiques d'une période, même si ce dernier point semble avoir donné peu de résultats au regard de nos observations. L'apprentissage se faisant sur les mots et non sur leurs lettres, les erreurs ponctuelles n'ont aucune incidence. Il serait intéressant d'observer le gain apporté par une correction des erreurs de numérisation, comme nouvel apport d'informations justes.

L'étiquetage retenu est celui des décennies, limitant le nombre d'étiquettes possibles à quinze. Potentiellement plus fin, il a été favorisé par rapport à celui des années qui entraîne une multiplication des étiquettes (145 possibilités). En plus de diminuer les coûts d'apprentissage, cela permet de disposer virtuellement d'un plus grand corpus pour chaque étiquette. Le gain en précision qu'aurait apporté une première passe en année (regroupement de probabilités) est ici largement contrebalancé.

3.3.2 Expérimentations et résultats

L'entraînement étant particulièrement coûteux en ressources, seule une dizaine de configurations ont été testées (voir Source 1).

Source 1 – Résultats d'un apprentissage avec une fenêtre de cinq mots sur un article brut

```
resultats (total : 942):
difference  : nombre    pourcentage   cumul
ok  : 101   10.7218683651805    10.7218683651805
10  : 98    10.4033970276008    21.1252653927813
20  : 124   13.1634819532909    34.2887473460722
30  : 92    9.76645435244161    44.0552016985138
40  : 85    9.02335456475584    53.0785562632696
50  : 81    8.59872611464968    61.6772823779193
60  : 57    6.05095541401274    67.728237791932
70  : 61    6.4755838641189     74.2038216560509
80  : 53    5.62632696390658    79.8301486199575
90  : 54    5.73248407643312    85.5626326963906
100 : 46    4.88322717622081    90.4458598726115
110 : 43    4.56475583864119    95.0106157112526
120 : 20    2.12314225053079    97.1337579617834
130 : 16    1.69851380042463    98.8322717622081
140 : 11    1.16772823779193    100
```

Comparé à un système purement aléatoire, un écart significatif est relevé. L'algorithme se trouve bien au-dessus des résultats attendus pour les décennies proches. La moitié du corpus est identifié avec moins de quatre décennies d'écart (contre 37 % attendus). L'utilisation de la ponctuation a permis d'améliorer ces résultats de près de 20 %, avec cependant un temps d'entraînement quasiment doublé. Le résultat pour chaque article est retourné sous forme d'une densité de probabilité sur les quinze décennies.

3.4 Récurrence des termes

En s'inspirant de la méthode précédente, il est possible d'entraîner le système pour qu'il attribue un score différent aux articles en entrée. Le corpus d'entraînement est *déplié* de façon à ce que chaque chaîne de caractères de chaque article [2] se voit attribuer la décennie de l'article dont elle est extraite (voir Source 2).

Source 2 – Annotation automatique des chaînes de caractères du corpus d'entraînement

```
etant           1830
la              1830
representation  1830
```

2. Chaînes de caractères consécutifs séparés par au moins un espace de chaque côté.

```
la              1830
plus            1830
pure            1830
la              1830
plus            1830
noble           1830
de              1830
la              1830
pensee          1830
democratique    1830
,               1830
devrait         1830
au              1830
moins           1830
commencer       1830
```

La phase d'entraînement va attribuer à chaque chaîne de caractères trouvée dans le corpus un histogramme indiquant les occurrences de la chaîne pour chaque décennie. Le système va donc produire, à l'issue de l'entraînement, une table pour chaque chaîne de caractères rencontrée pendant l'entraînement. Chaque table contient quinze entiers (un par décennie) représentant le nombre d'occurrences rencontrées de cette chaîne pour chaque décennie. Une fois l'intégralité du corpus d'entraînement parcouru, les tables sont normalisées de façon à représenter une densité de probabilité.

Lors de la phase d'évaluation, un tableau de quinze scalaires initialisés à zéro est créé. La présence de chaque chaîne de caractères de l'article en entrée est vérifiée dans la table d'entraînement. Si la chaîne est introuvable, alors elle est ignorée. Sinon, les valeurs de la table de la chaîne sont additionnées à la table de l'article. Ainsi, une image approximative de la probabilité de répartition des chaînes de caractères de l'article est créée au fil des décennies.

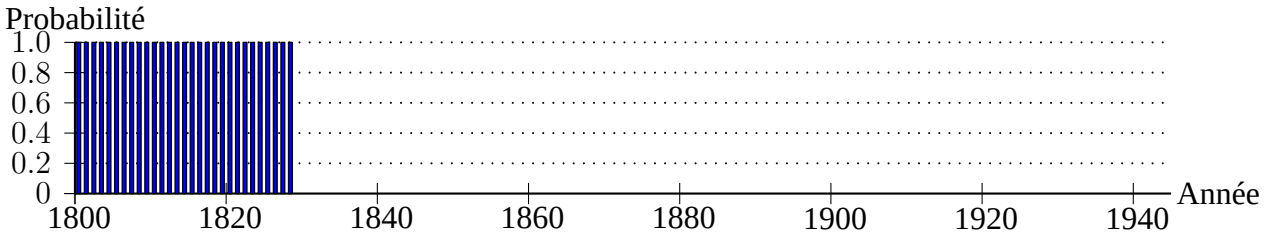
Une fois que l'intégralité d'un article est parcouru de cette manière, le tableau final est normalisé. Cette fois-ci, chaque valeur est normalisée en divisant chaque valeur du tableau par le maximum pour que chaque colonne se retrouve sur l'intervalle $[0; 1]$. Il ne s'agit plus ici d'une densité de probabilité. Ce tableau est ensuite transmis au processus de fusion qui se chargera de l'intégrer aux autres résultats.

3.5 Fusion des résultats

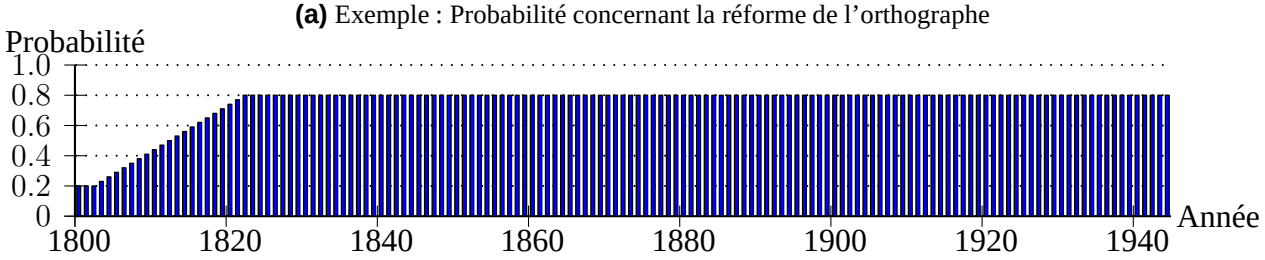
Tous les modules de traitement lexical présentés précédemment sont des filtres indépendants les uns des autres. Tous fournissent des résultats qui donnent pour un article, une probabilité sur chacune des années de 1800 à 1944. La fusion des résultats se fait par une simple multiplication des résultats de chaque filtre.

Imaginons un article dans lequel se trouveraient les mots *prouvoit* et *avoient* (respectivement *prouvait* et *avaient* en français contemporain) ainsi que la seule entité nommée *Victor Hugo*. Concernant la réforme de l'orthographe, l'apparition de deux occurrence nous permet de filtrer par rapport à l'année 1828 (voir Figure 3a page suivante). Pour les entités nommées, seule l'entité *Victor Hugo* a été identifiée et une recherche sur Internet nous permet de déterminer une date de naissance en 1802 (voir Figure 3b page ci-contre). En multipliant les deux résultats, nous obtenons la Figure 3c page suivante qui présente une

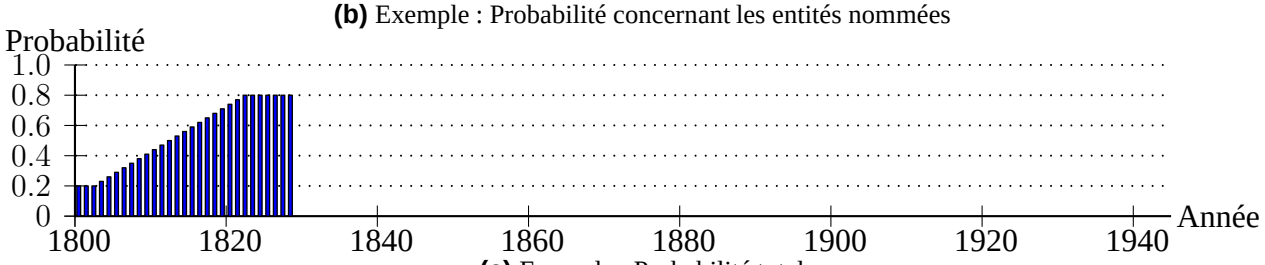
forte probabilité pour la décennie 1820.



(a) Exemple : Probabilité concernant la réforme de l'orthographe



(b) Exemple : Probabilité concernant les entités nommées



(c) Exemple : Probabilité total

Figure 3 – Exemple de fusion des résultats.

4 Soumission

La soumission finale concerne l'évaluation du corpus de test contenant 2721 textes pour lesquels aucune indication supplémentaire n'est donnée. De plus, comme pour le corpus d'entraînement, toutes les dates dans les corps de textes ont été masquées par des balises vides.

Avec notre algorithme, nous avons délibérément décidé d'appliquer certaines de nos méthodes année par année. Pourtant, le résultat final doit être une décennie ou, le cas échéant, fournir une probabilité pour chaque décennie. Le résultat doit se présenter sous la forme d'une densité de probabilité sur l'ensemble des décennies.

Nos résultats ne sont pas sous cette forme puisque une probabilité est fournie pour chaque année. De plus, la probabilité pour chaque année se trouve incluse dans l'intervalle $[0; 1]$ ce qui n'assure aucunement d'obtenir une densité de probabilité (*i.e.* la somme des probabilités n'est pas forcément égale à 1).

Pour commencer, nous avons effectué une moyenne de nos résultats par décennie.

$$p'_{\mathrm{d}} = \sum_{i=\mathrm{d}}^{\mathrm{d}+10} \frac{p_{\mathrm{d}+i}}{10} \tag{1}$$

d représentant la décennie concernée.

Puis, afin d'obtenir une vrai densité de probabilité, les probabilité de toutes les décennies ont été divisées par la somme totale des probabilités.

$$p''_{\mathrm{d}} = \frac{p_{\mathrm{d}}}{\sum_{i=1800}^{1940} p_{\mathrm{d}}} \tag{2}$$

Nous avons testé puis soumis trois différentes types de résultats puisque cela était permis dans les conditions du concours DEFT 2010. Tout d'abord, nous avons évalué la fiabilité de nos résultats avec un fichier contenant une probabilité pour l'ensemble des décennies. Ce fichier peut contenir des probabilités nulles comme dans le cas des filtres de la réforme de l'orthographe (voir section 3.1 page 3). Néanmoins, dans la plupart des cas, une probabilité non-nulle sera donnée pour chaque décennie. Ce type de résultat nous permet d'obtenir une précision de 0.297 et un rappel de 0.299. La F-mesure est de 0.298 (voir Figure 4 page suivante).

Puis nous avons tenté de voir si nous pouvions obtenir de meilleurs résultats en exploitant au mieux les sorties de notre algorithme. Par exemple, nous avons de nouveau évalué le corpus en ne conservant que la décennie ayant obtenu la plus grande probabilité. Dans le cas où plusieurs décennies sont concernées (plusieurs égalités), une probabilité égale est affectée à chacune. Cette seconde soumission donne de bien meilleurs résultats. Cette nouvelle stratégie nous a permis d'obtenir une précision de 0.336 et un rappel de 0.340 et donc une F-mesure de 0.338 (voir Figure 4 page ci-contre).

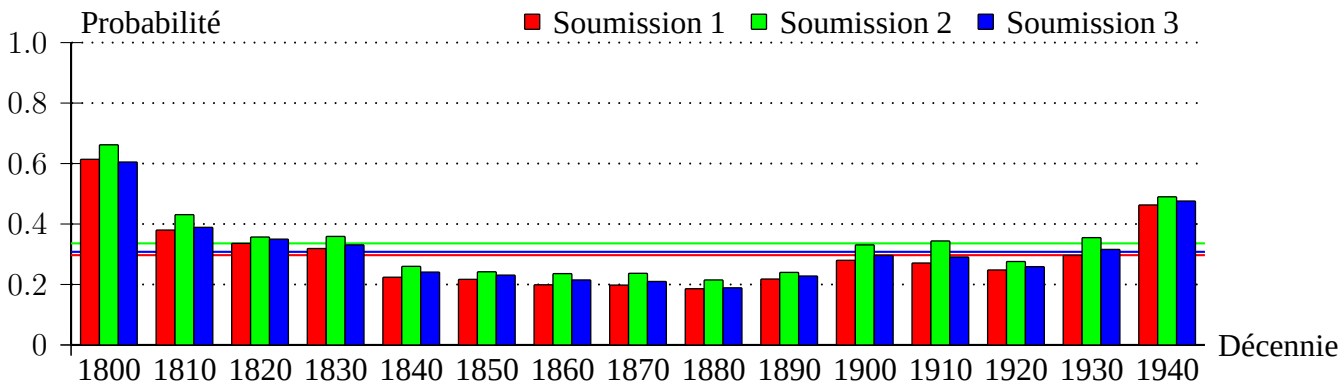
Cependant, la solution conservant uniquement la meilleure probabilité peut faire disparaître la décennie recherchée. En effet, nos méthodes étant en partie basées sur des statistiques, il est possible que la bonne décennie ne soit que la seconde voire la troisième meilleure probabilité. Nous avons donc décidé d'effectuer une troisième soumission en ne conservant que les trois meilleures probabilités. Dans le cas où plusieurs décennies possèdent la meilleure troisième probabilité, elles sont toutes conservées. La densité totale de probabilité est alors répartie sur l'ensemble des décennies concernées. Cette troisième et dernière proposition nous a permis d'obtenir une précision de 0.308 et un rappel de 0.313 et donc une F-mesure de 0.310 (voir Figure 4 page suivante).

La dernière solution est moins efficace que la seconde. Ceci est dû à la distribution des probabilités sur trois décennies (ou plus le cas échéant). Dans le cas où la seconde soumission supprimait la bonne décennie, cette troisième soumission va améliorer les résultats puisqu'elle aura plus de chance de conserver la bonne décennie. Dans le cas où la seconde soumission donnait déjà la bonne décennie, cette troisième soumission va avoir pour effet de répartir les probabilités sur deux autres décennies (ou plus) qui seront incorrectes ce qui diminue la fiabilité totale.

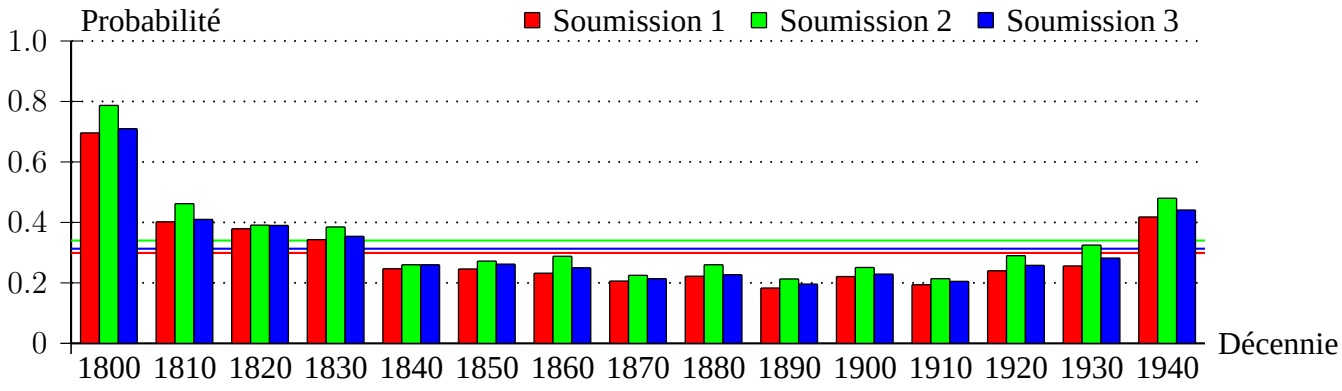
5 Conclusion

Dans cet article, nous avons présenté un système qui permet de dater des articles de journaux. Ce système se base sur différentes méthodes : méthodes basées sur le lexique, méthode statistique, apprentissage… Ces statistiques nous ont permis d'atteindre une précision de 33.6 % pour un rappel de 34 %. La F-mesure de nos résultats monte à 33.8 %.

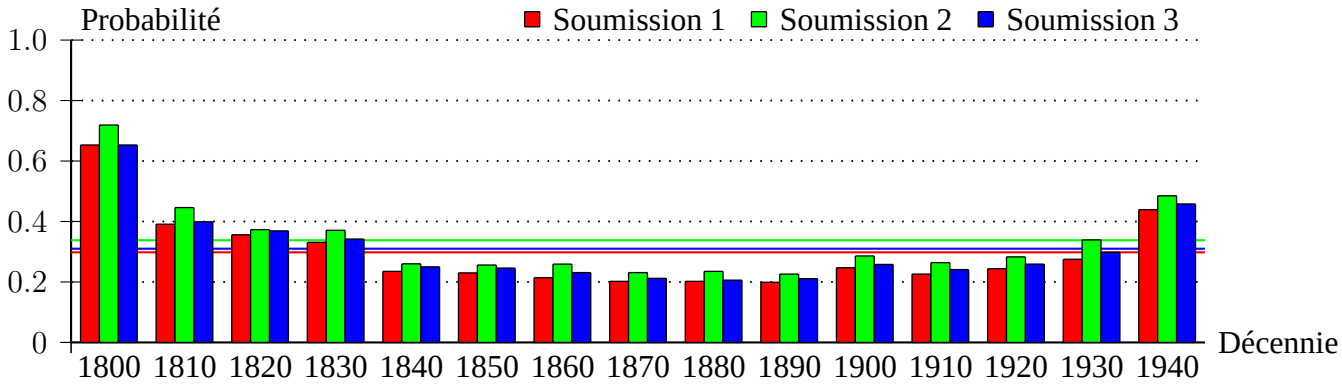
En ajoutant des modules complémentaires, nous pensons que notre système pourrait être plus efficace.



(a) Précision



(b) Rappel



(c) F-mesure

Figure 4 – Résultats des trois soumissions (précision, rappel et F-mesure).

Parmi ces modules complémentaires, certains n'ont pas pu être intégrés pour des raisons techniques alors que d'autres ont seulement été envisagés.

Afin de renforcer l'amplitude des pistes lexicales, il est logique de s'intéresser à l'étymologie de l'ensemble des mots des articles. Cette période étant très riche en inventions et technologies, le lexique associé à celles-ci a une plus grande fréquence d'apparition, principalement lors des guerres. La ressource de référence utilisée dans ce cas est le Trésor de la Langue Française Informatisé (TLFI) consulté par l'intermédiaire du site du Centre National de Ressources Textuelles et Lexicales (CNRTL). Les mots mal identifiés, présentant des caractères non-alphabétiques, sont filtrés, de même que les mots outils. Ce filtrage permet de diminuer le nombre de requêtes et ainsi diminue le temps de traitement du corpus d'environ 35 %.

Aucun traitement n'étant effectué d'un point de vue sémantique, l'ensemble des définitions sont prises en compte et la date répertoriée la plus ancienne de l'utilisation du terme est conservée. Pour un article, la date du mot le plus récent détermine ensuite avec une très grande probabilité la limite basse. Cette piste a été abandonnée en raison de problèmes techniques. Les requêtes au serveur nécessitant une charge importante, le site du CNRTL bloque notre adresse IP. Un accès local au dictionnaire ainsi qu'un cache des termes déjà traités permettrait de diminuer de façon importante le temps de recherche.

Pour améliorer le module de recherche des entités nommées, il faudrait utiliser une plus grande variété de préfixes tels que *Mgr* pour *Monseigneur* ou *S. M.* pour *Sa Majesté*. Les préfixes désignant la fonction des personnes (*Colonel* ou *Baron* par exemple) pourraient également être utilisées pour mieux filtrer les recherches sur Internet et ainsi permettre la distinction entre plusieurs personnes ayant le même nom.

Enfin, sur le modèle de la méthode des entités nommées, nous envisagerions de créer un module utilisant les dates des inventions (*macadam* ou *dynamomètre* par exemple) pour établir une limite basse. Ce module se baserait sur la supposition qu'un nom désignant une invention n'est pas utilisé tant que l'invention n'a pas été créée. Il est tout de même nécessaire de faire attention car ce module pourrait bien amener quelques erreurs comme par exemple avec le mot *voiture* qui, à l'époque, ne désignait pas une automobile mais simplement un engin tracté par des chevaux.

Toutes ces améliorations pourraient permettre d'obtenir, pour cette tâche, une meilleure F-mesure.

Références

ACADÉMIE FRANÇAISE (1835). *Dictionnaire de l'Académie française*. P. Dupont, 6e edition.

ACADÉMIE FRANÇAISE (1878). *Dictionnaire de l'Académie française*. Firmin–Didot, 7e edition.

EHRMANN M. (2008). *Les entités nommées de la linguistique au* TAL *: statut théorique et méthodes de désambiguïsation*. PhD thesis, Université PARIS VII.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional Random Fields : probabilistic models for segmenting and labeling sequence data. In [Society, 2001], p. 282–289.

I. M. L. SOCIETY, Ed. (2001). *International Conference on Machine Learning*, Princeton.

WALLACH H. M. (2004). *Conditional Random Fields : an introduction*. Rapport interne, Université de Pennsylvanie.