Actes du neuvième défi fouille de texte, DEFT2013, Les Sables-d'Olonne, France, 21 juin 2013. Pages 3-16.
Proceedings of the Ninth DEFT Workshop, DEFT2013, Les Sables-d'Olonne, France, 21st June 2013. Pages 3-16.

DEFT2013 se met à table : présentation du défi et résultats

Cyril Grouin[1,2] Pierre Zweigenbaum[1] Patrick Paroubek[1]
(1) LIMSI–CNRS, Orsay (2) INSERM U872 Eq 20 & UPMC, Paris
prenom.nom@limsi.fr

RÉSUMÉ
L'édition 2013 du défi fouille de texte (DEFT) a porté sur l'analyse des recettes de cuisine sous quatre angles différents : le niveau de difficulté d'une recette, le type de plat préparé, le titre, et les ingrédients nécessaires à la réalisation de la recette. Dans cet article, nous présentons le déroulement du défi au travers des corpus produits et des tests humains réalisés, ainsi que les résultats obtenus par les équipes ayant participé. Les mesures d'évaluation utilisées sont la distance relative moyenne à la solution pour la difficulté du plat et le couple précision/rappel pour le type de plat, la MRR (moyenne de l'inverse du rang) pour l'association titre/recette et la MAP (moyenne des précisions moyennes) pour la liste des ingrédients d'une recette.

ABSTRACT
DEFT2013 spills the beans: challenge presentation and results

The DEFT 2013 challenge focused on the analysis of cooking recipes through four points of view: the level of difficulty of the recipe, the type of dish to be prepared, the title, and the ingredients needed to prepare the recipe. In this paper, we present the challenge: the corpora we produced, the human tests, as well as the results each participant achieved. The evaluation measures we used are the mean distance to the solution for the difficulty task, precision and recall for the type of dish, mean reciprocal rank (MRR) for the title/recipe association and mean average precision (MAP) for the list of ingredients.

MOTS-CLÉS : campagne d'évaluation, classification, mesures, recettes.

KEYWORDS: evaluation campaign, classification, metrics, recipes.

1 Présentation

1.1 Introduction

Le défi DEFT est un atelier annuel d'évaluation francophone en fouille de textes. La neuvième édition de ce défi a porté sur l'analyse des recettes de cuisine rédigées en langue française. Dans le cadre de cette nouvelle édition, nous avons orienté le défi vers un nouveau domaine d'application. La thématique retenue, les recettes de cuisine, a déjà fait par le passé l'objet d'une campagne d'évaluation *(Computer Cooking Contest*[1]*)*. Nous nous intéressons dans DEFT2013 à deux types de fonction d'analyse du langage, la classification de documents *(tâches 1 à 3)* et l'extraction d'information *(tâche 4)*, ceci dans un domaine de spécialité.

1. http://computercookingcontest.net

Pour cette nouvelle édition du défi, nous proposons quatre tâches d'analyse concernant les recettes de cuisine : identifier le niveau de difficulté de réalisation d'une recette, identifier le type de plat préparé, apparier une recette à son titre, identifier les ingrédients d'une recette.

1.2 Organisation

Une étape de réflexion relative au choix des thématiques et des tâches à proposer a été engagée en fin d'année 2012. Durant cette période, des tests de tâches ont été effectués auprès d'étudiants. Ces tests nous ont permis de confirmer les choix effectués, ou dans certains cas, de réorienter certaines tâches envisagées. Nous avons orienté cette nouvelle édition vers les recettes de cuisine du site Marmiton, [2] dont les catégories d'informations proposées correspondaient aux thématiques que nous souhaitions traiter dans ce défi et parce que les recettes sont librement accessibles.

Deux appels à participation ont été lancés sur la liste de diffusion de la communauté du traitement automatique des langues LN les 16 février et 5 mars. Les données d'apprentissage ont été mises à la disposition des participants à partir du 28 février, les données de test l'ont été entre le 25 avril et le 5 mai, dans une fenêtre de trois jours au libre choix de chaque participant. La référence et les résultats individuels *(avec classement, moyenne et médiane)* ont été communiqués aux participants le 6 mai.

Dix équipes se sont inscrites à cette édition du défi, parmi lesquelles deux sont issues du monde industriel. Les six équipes qui ont participé aux tests sont les suivantes :
- **Celi France :** Luca DINI, André BITTAR, Mathieu RUHLMANN ;
- **GREYC :** Gaël LEJEUNE, Charlotte LECLUZE, Romain BRIXTEL ;
- **LIA :** Xavier BOST, Ilaria BRUNETTI, Luis Adrián CABRERA-DIEGO, Jean-Valère COSSU, Andréa LINHARES, Mohamed MORCHID, Juan-Manuel TORRES-MORENO, Marc EL BÈZE, Richard DUFOUR ;
- **LIM&Bio :** Thierry HAMON, Natalia GRABAR, Amandine PÉRINET ;
- **Orange Labs :** Olivier COLLIN, Aleksandra GUERRAZ, Yannick HIOU, Nicolas VOISINE ;
- **Wikimeta Lab :** Eric CHARTON, Ludovic JEAN-LOUIS, Marie-Jean MEURS, Michel GAGNON.

2 Présentation

2.1 Tâches proposées

Nous avons proposé quatre tâches autour de la thématique des recettes de cuisine. La motivation des deux premières tâches concerne le jugement appliqué à une recette. Dans de nombreux cas, le type de plat, et plus encore, le niveau de difficulté de réalisation d'une recette peut se révéler variable d'un individu à un autre, avec pour conséquence une indexation biaisée des recettes présentes sur le site. L'indexation des recettes selon les ingrédients qu'elle contient est également une tâche complexe, qui fait appel à la distinction entre ingrédients obligatoires et ingrédients facultatifs. Enfin, l'appariement d'un titre avec la recette qui lui correspond permet de mettre en évidence les désaccords qui peuvent parfois exister dans ces associations. L'objectif global visé par les quatre tâches du défi vise ainsi à fournir des méthodes adaptées à l'indexation de sites web participatifs, dont le contenu, déposé par chaque utilisateur, est généralement classé et

2. http://www.marmiton.org/

évalué par l'utilisateur qui aura déposé le contenu en ligne, avec des écarts d'appréciation de classement assez importants.

Tâche 1 — Niveau de difficulté de réalisation d'une recette. La première tâche a pour but d'évaluer la capacité d'un algorithme à inférer la difficulté d'une recette de cuisine en se basant sur toutes les informations qu'il est possible d'extraire à partir du texte de la recette et de son titre. Le niveau de difficulté de réalisation d'une recette est évalué sur une échelle à quatre valeurs : *très facile, facile, moyennement difficile, difficile*. La mesure d'évaluation principale retenue pour cette tâche est la distance moyenne à la solution, calculée en micro-moyenne sur une échelle à quatre valeurs. Nous avons aussi vérifié la corrélation des résultats ainsi obtenus avec ceux provenant d'une mesure de précision/rappel en micro-mesure.

Tâche 2 — Type de plat. La deuxième tâche propose de classer les recettes en fonction du type de plat préparé, selon une partition en trois classes : *entrée, plat principal, dessert*. Les mesures d'évaluation retenues sont la précision et le rappel calculées en micro-moyennes.

Tâche 3 — Appariement titre/recette. La troisième tâche demande au système de retrouver pour chaque texte de recette à traiter, son titre original dans une liste de titres de recettes.

Tâche 4 — Ingrédients d'une recette. La dernière tâche se démarque des précédentes car elle ne concerne pas la classification des recettes, mais l'extraction d'information. Il s'agit en effet dans cette tâche d'identifier la liste des ingrédients de la recette. Les ingrédients identifiés doivent être exprimés selon des libellés normalisés présents dans une liste globale fournie aux participants. Cette liste contient au moins tous les ingrédients à trouver dans la base des textes de recettes, mais peut aussi contenir des ingrédients qui ne sont présents dans aucune des recettes des corpus d'entraînement et de test.

2.2 Corpus

2.2.1 Présentation

Le corpus de recettes utilisées pour le défi provient du site de recettes participatif Marmiton. Sur ce site, chaque utilisateur dépose ses propres recettes. Ce type de dépôt implique que l'utilisateur renseigne lui-même certaines des caractéristiques liées à la recette, en particulier les informations de coût et de niveau de difficulté. Les participants ont été autorisés à utiliser des corpus de recettes complémentaire, à l'exclusion des recettes provenant de Marmiton.

2.2.2 Modalités de constitution

Collecte des données. Nous avons téléchargé les recettes du site pendant les deux premières semaines de février 2013, au moyen d'une chaîne effectuant les étapes suivantes :
- Collecte des 26 pages d'index des ingrédients utilisés dans les recettes du site ;
- Pour chaque ingrédient indexé dans les pages d'index précédentes, collecte des pages listant les recettes utilisant cet ingrédient ;
- Conversion des pages web téléchargées au format XML.

Au terme de cette phase de récupération, nous avons collecté 46 176 recettes que nous avons transformées en XML. La moitié de ces recettes a été aléatoirement conservée pour l'édition 2013 du défi, l'autre moitié pouvant servir de corpus pour l'année suivante.

Constitution de la référence. La référence des différentes tâches a été réalisée de manière automatique, sur la base des informations présentes dans les recettes.

– la référence des tâches 1 à 3 est directement inférée des informations présentes dans les recettes : *niveau de difficulté, type de plat, titre de la recette* ;
– la référence de la tâche 4 a été constituée lors de la collecte des données par la correspondance entre les pages d'index des ingrédients et les différentes recettes.

Concernant les trois premières tâches, la référence est issue des informations renseignées par l'utilisateur qui a posté la recette sur Marmiton. Ainsi, le niveau de difficulté de réalisation d'une recette est largement subjectif et dépend également du niveau d'expertise de chacun en cuisine ; d'autre part, l'auteur d'une recette aura tendance à minorer le niveau de difficulté de sa recette, de manière à inciter le plus grand nombre de visiteurs à tester la recette avec l'espoir d'obtenir des commentaires en retour. On observe ce comportement dans la répartition des recettes selon les quatre niveaux de difficultés (tableau 1 gauche). À l'inverse, la répartition en types de plat semble plus équilibrée et moins sujette à la subjectivité (tableau 1 droite). Nous sommes conscients de la présence d'erreurs dans les références mais nous avons considéré que ces erreurs étaient marginales et n'auraient pas un impact trop fort sur l'évaluation finale. En conséquence, aucune phase de nettoyage humain *(tâche forcément coûteuse)* n'a été réalisée sur les références.

Niveau	Nombre	Pourcentage	Type	Nombre	Pourcentage
Très facile	11 602	50,2 %	Entrée	5 407	23,4 %
Facile	9 584	41,5 %	Plat principal	10 742	46,5 %
Moyennement difficile	1 776	7,7 %	Dessert	6 945	30,1 %
Difficile	132	0,6 %			

TABLE 1 – Répartition des recettes par niveau de difficulté (gauche) et type de plat (droite)

En ce qui concerne la dernière tâche, chaque recette contient la liste des ingrédients fournie par l'utilisateur. Cependant, les ingrédients de cette liste ont une forme qui varie avec les recettes, et il serait très difficile d'arriver à prédire exactement le libellé utilisé par l'auteur. C'est pourquoi la référence a été constituée à partir de l'index des ingrédients des recettes, index qui constitue une liste de libellés normalisés d'ingrédients et qui associe à chacun l'ensemble des recettes qui l'utilisent. Si cette méthode a l'avantage de fournir une liste normalisée d'ingrédients pour chaque recette, elle présente néanmoins plusieurs défauts. D'une part, certains ingrédients d'une recette ont pu être oubliés lors de l'indexation, et seront considérés comme faux positifs s'ils sont (correctement) trouvés par un système. D'autre part, certains index peuvent être associés par erreur à une recette (c'est le cas du « maïs » mis comme index d'une recette dont un ingrédient contenait l'expression « mais compter un peu plus long pour la cuisson »), et ils compteront comme faux négatifs si un système ne les produit pas. À noter, la forme normalisée des ingrédients ne contient pas de caractères accentués (les diacritiques sont supprimés) ni d'espaces (ils sont remplacés par des tirets).

Répartition des données. Le corpus de l'édition 2013 totalise 23 094 recettes que nous avons ensuite réparties entre corpus d'apprentissage *(13 864 recettes)* et corpus de test *(9 230 recettes)* selon une répartition respectivement fixée à 60 %/40 % comme utilisée dans les précédentes campagnes de DEFT. Cette répartition a été effectuée de telle sorte que la proportion de recettes en termes de niveau de difficulté et de type de plat soit équivalente dans les deux corpus.

Le corpus d'apprentissage est commun à l'ensemble des quatre tâches proposées cette année, dans la mesure où les données de référence sont incluses dans les méta-informations. En ce qui concerne le corpus de test, nous l'avons segmenté en quatre parties égales *(soit environ 2 300 documents par partie)*, chaque partie étant réservée pour une tâche, en garantissant également que chaque partie conserve la même proportion de recettes en niveaux de difficulté et type de plat *(même si cet impératif ne se révèle nécessaire que pour les tâches 1 et 2)*. Il en ressort que les documents des corpus de test de chacune des quatre tâches sont distincts pour chaque tâche.

Formatage des données. Nous donnons en figure 1 un exemple de recette issue du corpus d'apprentissage. Chaque document se compose de deux parties principales :
– des méta-données contenant : le titre *(référence de la tâche 3)*, le type de plat *(référence de la tâche 2)*, le niveau de difficulté *(référence de la tâche 1)*, le coût *(information complémentaire fournie dans l'apprentissage et le test)*, et la liste des ingrédients renseignés par l'utilisateur ;
– la préparation de la recette contenant les différentes étapes.

En dehors du document, la liste des ingrédients normalisés *(référence de la tâche 4)* a été fournie. La liste des ingrédients normalisés qui indexent cette recette sont les suivants : *ail, basilic, bouillon, citron, echalote, huile-d-olive, parmesan, pignon, poivre, tomate, truite*. Notons que les ingrédients normalisés, parce qu'ils servent à indexer le contenu du site Marmiton, étaient déjà tous désaccentués, avec espaces et apostrophes remplacés par des tirets.

2.3 Tests humains

Nous avons mis à contribution les étudiants de l'EBSI [3] et du M2 Pro d'Ingénierie Linguistique de l'INaLCO [4] en novembre 2012. Les étudiants ont travaillé sur les quatre tâches que nous envisagions à l'époque : (i) évaluer le niveau de difficulté *(très facile, facile, moyennement difficile, difficile)* d'une recette ; (ii) prédire le type de plat *(entrée, plat, dessert)* ; (iii) relier un titre à la recette correspondante ; et (iv) identifier l'ingrédient ajouté à la recette d'origine.

Niveau de difficulté. Les étudiants de l'INaLCO ont pris une demi-heure pour étudier le corpus et prédire le niveau de difficulté. Les résultats de ces tests sont rassemblés dans le tableau 2. La bonne réponse *(niveau de difficulté renseigné dans Marmiton)* est identifiée, au minimum par aucun annotateur et au maximum par 7 annotateurs, avec un taux moyen de réussite sur l'ensemble du corpus de 37,0 %. Aucune recette n'est donc évaluée de la même manière par l'ensemble des annotateurs. La majorité des erreurs effectuées (57,1 %) ne correspond toutefois

3. Ecole de Bibliothéconomie et des Sciences de l'Information, Université de Montréal. Cours assuré par Dominic Forest auprès d'un groupe de douze étudiants.

4. Institut National des Langues et Civilisations Orientales, Paris. Cours assuré par Cyril Grouin pour les parcours « *Ingénierie Multilingue* » et « *Traductiques et gestion de l'information* » du M2 Pro d'Ingénierie Linguistique auprès d'un groupe de dix étudiants : Abdennour Goumiri, Ardas Khalsa, Arthur Boyer, Asma Benahmed, Hamza Affane, Julie Arnal, Laure Chancerelle, Maria Goryainova, Selen Şahin et Thomas Moraine. Tous ces étudiants se sont acquittés de cette tâche d'évaluation humaine avec abnégation, à quelques heures seulement du dîner... Nous les en remercions vivement.

```
<?xml version="1.0" encoding="utf-8"?>
<recette id="54562">
  <titre>Cuisses de poulet au miel cuites au four</titre>
  <type>Plat principal</type>
  <niveau>Très facile</niveau>
  <cout>Bon marché</cout>
  <ingredients>
    <p>2 cuisses de poulet</p>
    <p>1 oignon</p>
    <p>10 cl d'huile d'olive (environ 1/2 verre à eau)</p>
    <p>3 cuillères à soupe de miel</p>
    <p>sel, poivre</p>
  </ingredients>
  <preparation>
  <![CDATA[
    Préchauffez le four à 200'C (thermostat 6–7).
    Peler les oignons et les émincer.
    Les disposer dans un plat à gratin (environ 40 x 20 cm).
    Disposer les cuisses de poulet sur les oignons.
    Dans un bol, mélangez l'huile et le miel.
    Badigeonner le poulet de ce mélange.
    Saler et poiver les cuisses.
    Mettre le plat dans le four pendant environ 50 minutes.
    Vos oignons sont caramélisés et de même pour vos cuisses de poulet.
    Bonne dégustation.
    Cette recette accompagnée par une bonne purée, c'est simple et c'est un vrai
        régal!Vous pouvez ajouter un deuxième oignon selon vos envies.
  ]]>
  </preparation>
</recette>
```

FIGURE 1 – Recette issue du corpus d'apprentissage avec la référence des différentes pistes en méta-données (sauf pour la tâche 4, dont les ingrédients normalisés pour cette recette étaient *cuisse-de-poulet, huile-d-olive, miel, oignon, poivre, sel*)

qu'à des écarts d'un seul niveau. Enfin, les annotateurs humains parviennent à identifier si la recette est plutôt d'un niveau facile ou d'un niveau difficile. La même expérience réalisée sur un second corpus de dix recettes auprès d'un groupe d'étudiants de l'EBSI permet l'obtention de résultats similaires avec un taux de réussite moyen de 34,0 %.

D'autre part, les annotateurs sont majoritairement d'accord entre eux. Si l'évaluation est réalisée non plus d'après le niveau de référence mais selon le niveau majoritairement attribué par les annotateurs, le taux moyen de réussite sur l'ensemble du corpus monte à 52,0 %. Cette observation témoigne du fait que le niveau de difficulté d'origine est mal choisi pour un grand nombre de recettes. Les concepteurs de sites participatifs qui évaluent la qualité d'un produit *(recettes, livres, films, etc.)* pourraient donc être intéressés par des outils d'évaluation automatique.

Type de plat. La tâche de prédiction du type de plat *(entrée, plat, dessert)* a été réalisée par les douze étudiants de l'EBSI. Sur cette tâche, ils ont fourni entre 6 et 9 bonnes réponses sur le corpus de dix recettes, avec un taux moyen de réussite de 77,5 %. À l'évidence, la tâche semble moins simple qu'il n'y paraît, a fortiori sur la distinction entre une entrée et un plat principal.

Recette	Référence	Annotateur									
		01	02	03	04	05	06	07	08	09	10
Couscous végétarien	1	4	3	**1**	3	3	3	2	3	4	3
Pain d'épice	1	2	2	**1**	**1**	**1**	**1**	2	3	2	**1**
Tartiflette courgettes	1	**1**	**1**	**1**	2	**1**	**1**	2	2	**1**	**1**
Garbure landaise	2	3	4	1	4	4	**2**	**2**	3	4	3
Rougets farcis	2	3	3	4	**2**	**2**	4	**2**	3	4	4
Lasagnes bolognaise	3	2	**3**	2	2	2	**3**	1	**3**	2	**3**
Sushi californien	3	**3**	4	**3**	**3**	**3**	4	**3**	2	4	**3**
Boulettes orientales	4	1	2	2	3	1	2	3	3	3	1
Lasagnes de légumes	4	3	2	**4**	3	3	**4**	2	**4**	3	3
Ravioles tartares	4	**4**	**4**	**4**	0	3	3	3	**4**	**4**	**4**

TABLE 2 – Prédictions des annotateurs humains en matière de niveau de difficulté d'une recette (0 : absence de réponse, 1 : très facile, 2 : facile, 3 : moyennement difficile, 4 : difficile)

Appariement titre/recette. La tâche consistant à relier chaque recette à son titre *(un corpus de dix recettes et de dix titres séparés)* a été réalisée par les étudiants de l'EBSI qui ont pratiquement tous réussi à effectuer les appariements corrects ; seul un(e) étudiant(e) a interverti deux titres. Le taux moyen de réussite est donc élevé sur cette tâche et se monte à 98,3 %. Le passage à l'échelle avec plusieurs milliers de titres et de recettes à apparier devrait se révéler plus complexe.

Ingrédient étranger. Sur la tâche d'identification de l'ingrédient étranger à une recette, [5] tous les étudiants de l'INaLCO ont correctement identifié l'ingrédient ajouté. Pour le défi, nous avons remplacé cette tâche par l'identification des ingrédients normalisés qui constituent la recette.

3 Évaluation et discussion

3.1 Modalités d'évaluation

Tâches 1 & 2 Les étiquettes de classe de la tâche 1 correspondent à des graduations sur une échelle de valeurs. Nous avons donc retenu comme mesure principale la distance relative moyenne à la solution, calculée en micro-moyenne comme mesure principale et en macro-moyenne comme mesure secondaire. L'échelle de valeurs comprend 4 valeurs réparties de manière symétrique sur une droite euclidienne (un espace à une dimension), de part et d'autre de l'origine. La distance entre les deux premières valeurs de classe de difficultés opposées est choisie comme double de celle séparant deux valeurs de difficulté de même signe (par ex. d(*très facile*, *facile*)= 1 mais d(*facile*, *moyennement difficile*)= 2), ceci afin de pénaliser plus un système qui ferait une erreur de genre de difficulté (facile/moyennement difficile) qu'un système faisant une erreur de qualification de difficulté (*très facile*/*facile*). Comme mesure secondaire, nous considérons le même calcul en macro-moyenne, afin de séparer des systèmes potentiellement ex-aequo en observant de manière plus fine leur performance sur les classes de recettes ayant des effectifs de

5. Pour chacune des cinq recettes constituant le corpus, nous avons ajouté un ingrédient provenant d'une recette du même type de plat. Sur la recette *« Crevettes au riz et à l'avocat en coque de tomate »* (entrée), nous avons ajouté des tomates cerises, tandis que sur la recette *« Cheesecake à la carotte »* (dessert) nous avons ajouté du sirop d'érable.

petite taille. À chacune des 4 valeurs possibles pour la donnée de référence r_i, correspond une valeur de distance maximale possible entre la réponse du système et cette donnée $dmax(h_i, r_i)$. Par exemple si la bonne réponse est *très facile*, la distance maximale possible est 4 si la réponse du système est *difficile*, mais si la bonne réponse est *facile*, la distance maximale possible sera 3, pour une réponse du système *difficile*. L'exactitude en distance relative à la solution moyenne (EDRM) se calcule en micro-moyenne comme indiqué dans l'équation 1.

$$\text{EDRM} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{d(h_i, r_i)}{dmax(h_i, r_i)}\right) \tag{1}$$

Cependant, la première tâche du défi peut aussi être considérée comme une tâche de classification automatique, comme l'est la tâche 2. À ce titre, les mesures habituelles de rappel, précision et F-mesure sont donc adaptées. Nous avons donc vérifié la corrélation des résultats obtenus avec l'EDRM sur la tâche 1 avec les mesures de précision et rappel, qui sont les mesures principales de la tâche 2. Ces mesures sont calculées à la fois en termes de micro-mesure (formules 2 et 3) et de macro-mesures (formules 4 et 5), le classement officiel étant établi sur la base des micro-mesures de manière à accorder un poids équivalent à chaque élément mesuré, même si cette mesure revient à privilégier les classes composées d'un grand nombre d'individus *(très facile, facile)* au détriment des classes faiblement représentées *(moyennement difficile, difficile)*.

$$\text{Micro-rappel} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux négatifs}(i)} \tag{2}$$

$$\text{Micro-précision} = \frac{\sum_{i=1}^{n} \text{vrais positifs}(i)}{\sum_{i=1}^{n} \text{vrais positifs}(i) + \sum_{i=1}^{n} \text{faux positifs}(i)} \tag{3}$$

$$\text{Macro-rappel} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i) + \text{faux négatifs}(i)}\right) \tag{4}$$

$$\text{Macro-précision} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\text{vrais positifs}(i)}{\text{vrais positifs}(i) + \text{faux positifs}(i)}\right) \tag{5}$$

Tâches 3 et 4. Les tâches 3 et 4 demandaient de renvoyer une liste ordonnée d'hypothèses, l'objectif étant de placer le plus haut possible dans cette liste la ou les bonnes hypothèses, et peuvent de ce fait être mises en parallèle avec le format et le mode d'évaluation des tâches de recherche d'information. Dans la tâche 3, une seule bonne réponse (le titre original de la recette) était possible pour chaque recette R_i. Une mesure basée sur le rang r_i de cette bonne réponse a donc été employée : l'inverse de ce rang, et donc pour l'ensemble des N recettes la moyenne de l'inverse du rang (MRR, voir formule 6). Dans la tâche 4, plusieurs bonnes réponses (les n_i ingrédients $\{I_i^1 \dots I_i^j \dots I_i^{n_i}\}$ de la recette R_i) étaient attendues pour chaque recette. La moyenne des précisions non interpolées $P(I_i^j)$ calculées à chaque position, dans la liste d'hypothèses, d'une des n_i réponses correctes I_i^j pour la recette R_i, est alors une mesure pertinente, et pour l'ensemble des recettes la moyenne de cette précision moyenne (MAP, voir formule 6).

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i} \qquad MAP = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}P(I_i^j) \tag{6}$$

On rappelle que si une bonne réponse n'est pas présente dans la liste fournie par un système, sa précision moyenne vaut zéro, comme si elle avait été classée à l'infini par le système. MRR et MAP ont été calculés à l'aide du programme `trec_eval` [6].

Pour la tâche 3, rappel, précision et F-mesure ont également été calculés en prenant comme classes les titres des recettes et en examinant le titre proposé au rang 1.

3.2 Résultats des participants

Tâche 1 — Niveau de difficulté. Nous indiquons dans le tableau 3 les résultats globaux obtenus par les participants sur la première tâche. Le tableau 4 donne les résultats par classe. Les meilleurs résultats et le vainqueur sont indiqués en gras. *rf* et *nb* sont présentés en section 3.3.

Équipe, soumission	Macro				**Micro**		Rang
	R	P	F	EDRM	P=R=F	**EDRM**	
02 — LIM&Bio, #1	0,399	0,332	0,363	0,609	0,580	0,826	
02 — LIM&Bio, #2	0,415	0,339	0,373	0,624	0,586	0,828	4
02 — LIM&Bio, #3	0,406	0,329	0,364	0,621	0,574	0,824	
04 — GREYC, #1	0,273	0,250	0,261	0,501	0,489	0,794	6
04 — GREYC, #2	0,265	0,248	0,256	0,501	0,465	0,787	
04 — GREYC, #3	0,289	0,281	0,285	0,543	0,360	0,624	
22 — LIA, #1	0,353	0,633	0,453	0,600	0,592	0,828	3
22 — LIA, #2	0,363	0,603	0,453	0,643	0,581	0,813	
22 — LIA, #3	0,364	0,613	0,457	0,638	0,588	0,820	
25 — Celi France, #1	0,304	0,342	0,322	0,558	0,511	0,799	5
26 — Orange Labs, #1	0,252	0,275	0,263	0,503	0,086	0,234	
26 — Orange Labs, #2	0,395	0,524	0,451	0,659	0,612	0,840	2
28 — Wikimeta Lab, #1	**0,419**	0,460	0,438	**0,690**	0,609	0,835	
28 — Wikimeta Lab, #2	0,375	**0,682**	**0,484**	0,612	**0,625**	**0,843**	1
comparaison : rf	0,301	0,337	0,318	0,511	0,543	0,810	
comparaison : nb	0,489	0,373	0,423	0,723	0,525	0,754	

TABLE 3 – Résultats globaux sur la tâche 1, *Niveau de difficulté*, évalués en termes de rappel, précision, F-mesure et EDRM en macro-moyenne, P=R=F et EDRM en micro-moyenne

Tâche 2 — Type de plat. Le tableau 5 donne les résultats globaux tandis que le tableau 6 fournit les résultats par classe. Les meilleurs résultats et le vainqueur sont indiqués en gras.

Tâche 3 — Appariement titre/recette. Le tableau 7 donne les résultats globaux. Les meilleurs résultats et le vainqueur sont indiqués en gras. Les équipes 02 et 26 ont renvoyé plusieurs fois le même titre au rang 1 pour des recettes différentes : ces titres (= classes) sont donc proposés plusieurs fois, dont toutes sauf au mieux une sont incorrectes, ce qui fait baisser leur précision.

6. `http://trec.nist.gov/trec_eval/` : pour la moyenne de l'inverse du rang,
`trec_eval -c -q -mrecip_rank reference.qrels hypotheses.txt`,
et pour la précision moyenne non interpolée,
`trec_eval -c -q -mmap reference.qrels hypotheses.txt`.

Équipe	Très facile			Facile			Moyennement diff.			Difficile		
	R	P	F	R	P	F	R	P	F	R	P	F
02 #1	0,769	0,625	0,689	0,478	0,528	0,502	0,000	0,000	0,000	0,350	0,175	0,233
02 #2	0,769	0,630	0,692	0,491	0,536	0,512	0,000	0,000	0,000	0,400	0,190	**0,258**
02 #3	0,797	0,606	0,689	0,429	0,535	0,476	0,000	0,000	0,000	0,400	0,174	0,242
04 #1	0,444	0,555	0,493	0,646	0,446	0,527	0,000	0,000	0,000	0,000	0,000	0,000
04 #2	0,292	0,557	0,384	0,768	0,433	0,554	0,000	0,000	0,000	0,000	0,000	0,000
04 #3	0,232	0,575	0,331	0,506	0,440	0,471	0,418	0,107	0,170	0,000	0,000	0,000
22 #1	0,762	0,624	0,686	0,499	0,555	0,525	0,101	0,352	0,156	0,050	1,000	0,095
22 #2	0,711	0,645	0,676	0,521	0,554	0,537	0,169	0,215	0,189	0,050	1,000	0,095
22 #3	0,719	0,644	0,679	0,529	0,553	0,541	0,159	0,254	0,195	0,050	1,000	0,095
25 #1	0,498	0,586	0,539	0,617	0,465	0,530	0,101	0,317	0,153	0,000	0,000	0,000
26 #1	0,002	0,667	0,004	0,007	0,350	0,014	1,000	0,083	0,153	0,000	0,000	0,000
26 #2	0,715	0,671	0,692	0,587	0,554	**0,570**	0,180	0,472	**0,261**	0,100	0,400	0,160
28 #1	0,759	0,662	0,707	0,525	0,570	0,547	0,190	0,343	0,245	0,200	0,267	0,229
28 #2	0,786	0,660	**0,717**	0,546	0,579	0,562	0,116	0,489	0,188	0,050	1,000	0,095
rf	0,648	0,591	0,618	0,533	0,491	0,511	0,021	0,267	0,039	0,000	0,000	0,000
nb	0,640	0,651	0,646	0,427	0,560	0,484	0,339	0,202	0,253	0,550	0,077	0,136

TABLE 4 – Résultats détaillés par classe sur la tâche 1, *Niveau de difficulté*

Équipe, soumission	Macro			Micro	Rang
	R	P	F		
02 — LIM&Bio #1	0,806	0,832	0,819	0,834	
02 — LIM&Bio #2	0,841	0,842	0,841	0,849	3
02 — LIM&Bio #3	0,806	0,832	0,819	0,834	
04 — GREYC #1	0,760	0,741	0,750	0,746	5
04 — GREYC #2	0,584	0,609	0,596	0,573	
04 — GREYC #3	0,381	0,418	0,398	0,331	
22 — LIA #1	0,873	0,868	0,871	0,876	
22 — LIA #2	0,879	0,880	0,879	0,886	
22 — LIA #3	**0,881**	**0,884**	**0,882**	**0,889**	1
26 — Orange Labs #1	0,767	0,857	0,810	0,821	
26 — Orange Labs #2	0,784	0,840	0,811	0,827	4
26 — Orange Labs #3	0,779	0,789	0,784	0,784	
28 — Wikimeta Lab, #1	0,843	0,850	0,847	0,856	2
comparaison : nb	0,862	0,852	0,857	0,859	

TABLE 5 – Résultats globaux sur la tâche 2, *Type de plat*, évalués en termes de rappel, précision et F-mesure en macro-moyenne, micro-moyenne

Pour ce qui est du rappel, il est à 1 pour un titre (= une classe) si ce titre est trouvé au moins une fois correctement au rang 1 : donc c'est la proportion des titres qui sont trouvés correctement pour au moins (= exactement) une recette. Cela explique pourquoi il est égal à la précision au rang 1 (P@1), proportion des recettes dont le titre correct a été trouvé au rang 1.

Équipe	Entrée			Plat principal			Dessert		
	R	P	F	R	P	F	R	P	F
02 #1	0,544	0,730	0,624	0,894	0,794	0,841	0,980	0,970	0,975
02 #2	0,687	0,702	0,694	0,851	0,844	0,848	0,983	0,980	*0,982*
02 #3	0,544	0,729	0,623	0,898	0,796	0,844	0,976	0,971	0,974
04 #1	0,730	0,568	0,639	0,677	0,828	0,745	0,873	0,825	0,849
04 #2	0,662	0,378	0,481	0,540	0,699	0,609	0,551	0,749	0,635
04 #3	0,792	0,265	0,397	0,208	0,528	0,298	0,142	0,461	0,217
22 #1	0,779	0,744	0,761	0,868	0,890	0,879	0,971	0,971	0,971
22 #2	0,756	0,777	0,766	0,890	0,882	0,886	0,989	0,982	**0,986**
22 #3	0,762	0,784	**0,773**	0,893	0,883	**0,888**	0,989	0,983	**0,986**
26 #1	0,345	0,882	0,496	0,965	0,755	0,847	0,991	0,934	0,962
26 #2	0,427	0,784	0,553	0,937	0,764	0,842	0,986	0,972	*0,979*
26 #3	0,514	0,623	0,563	0,836	0,774	0,804	0,988	0,970	*0,979*
28 #1	0,669	0,742	0,703	0,872	0,840	0,856	0,989	0,969	*0,979*
nb	0,776	0,696	0,734	0,825	0,883	0,853	0,986	0,976	0,981

TABLE 6 – Résultats détaillés par classe sur la tâche 2, *Type de plat*

Équipe, soumission	Macro			P@1	MRR	Rang
	R	P	F			
02 — LIM&Bio #1	0,036	0,019	0,025	0,036	0,036	
02 — LIM&Bio #2	0,241	0,200	0,218	0,241	0,241	2
02 — LIM&Bio #3	0,133	0,119	0,125	0,133	0,133	
04 — GREYC #1	0,127	0,127	0,127	0,127	0,127	3
26 — Orange Labs #1	0,309	0,249	0,276	0,308	0,430	
26 — Orange Labs #2	0,306	0,251	0,276	0,306	0,423	
26 — Orange Labs #3	**0,314**	**0,259**	**0,284**	**0,314**	**0,434**	1

TABLE 7 – Résultats globaux sur la tâche 3, *Appariement titre-recette*, évalués en termes de rappel, précision et F-mesure en macro-moyenne, précision au rang 1, et moyenne de l'inverse du rang (MRR)

Tâche 4 — Ingrédients d'une recette. Les résultats de la dernière tâche sont donnés dans le tableau 8. Notons que les systèmes 02 et 25 n'avaient pas normalisé les noms des ingrédients qu'ils produisaient : nous avons inclus cette normalisation (désaccentuation, remplacement des espaces par des tirets), ce qui n'a pas changé le classement. La meilleure MAP de l'équipe 25 serait sinon de 0,6559 (−0,01), et celle de l'équipe 02 serait de 0,4083 (−0,05). Les équipes 25 et 28 ont également laissé passer des ingrédients hors liste normalisée (respectivement 94 et 7 dans leur meilleur système), seules les équipes 04 et 22 se sont assurées que leur système produisait uniquement des ingrédients normalisés.

Équipe	02 LIM&Bio	04 GREYC	22 LIA	25 Celi France	28 Wikimeta Lab
#1	0,4115	0,4881	*0,6287*	**0,6662**	0,5675
#2	0,4170	0,5074	0,6218	—	*0,6428*
#3	0,4649	0,5556	0,6191	—	—
Rang	5	4	3	1	2

TABLE 8 – Résultats globaux sur la tâche 4, *Ingrédients d'une recette*, évalués en termes de moyenne de la précision non interpolée (MAP)

3.3 Discussion

Tâche 1 — Niveau de difficulté. Les classes de cette tâche étaient très déséquilibrées : la troisième classe (Moyennement difficile) avait un nombre d'instances un ordre de grandeur au-dessous des deux premières, et la quatrième (Difficile) était encore un ordre de grandeur plus bas. De ce fait, il était beaucoup plus difficile de détecter correctement les deux dernières classes, ce que reflètent les résultats des systèmes par classe : F-mesure généralement vers 0,6–0,7 pour Très facile, 0,5 pour Facile, 0,1–0,2 pour Moyennement difficile, et 0–0,2 pour Difficile. Une bonne performance sur les deux premières classes assurait un résultat honorable en micro-mesure, et rares sont les systèmes qui ont pu détecter un nombre non négligeable d'instances des deux classes difficiles (26-#2 et #28-1).

À titre de base de comparaison, nous avons mis au point une méthode simple fondée sur une représentation des recettes par les attributs suivants. Les trois champs textuels (titre, ingrédients, préparation) ont été segmentés en mots (en conservant les nombres et ponctuations), les mots des titres ont été représentés par des attributs binaires (présence / absence du mot) et les mots des deux autres champs par des attributs numériques (tf.idf, normalisé par la longueur du champ). Le nombre d'ingrédients (nombre de <p> dans la liste des ingrédients) a également été ajouté, ainsi que la longueur en mots et caractères de la liste d'ingrédients et de la préparation. Un classifieur bayésien naïf (*nb*, NaiveBayesMultinomial dans Weka) et un classifieur à forêt d'arbres (*rf*, RandomForest dans Weka) ont été entraînés sur ces vecteurs d'attributs : ces deux classifieurs ont un temps d'entraînement court à relativement court et sont assez robustes. La forêt d'arbres a obtenu une meilleure proportion d'instances bien classées (0,544) en validation croisée sur dix parties sur le corpus d'entraînement, mais sans savoir prédire la classe Difficile ; le classifieur bayésien, avec une correction plus faible (0,527), y trouvait en revanche près de la moitié des recettes difficiles. Les résultats sur le jeu de test (bas des tableaux 3 et 4, *rf* et *nb*) reflètent ces différences, qui induisent des différences importantes dans la macro F-mesure.

Tâche 2 — Type de plat. Quatre équipes sur cinq (Bost *et al.*, 2013; Hamon *et al.*, 2013; Collin *et al.*, 2013; Charton *et al.*, 2013) ont obtenu des résultats proches de la perfection pour la catégorisation du dessert (F=0,986, 0,982, 0,979, 0,979). Cela semble indiquer que cette catégorie était plus facile à trouver, ou mieux définie : les desserts contiennent en général du sucre, etc. Pour vérifier cela, nous avons entraîné pour cette tâche le classifieur bayésien naïf avec les mêmes attributs que ci-dessus (la forêt d'arbres donnait des résultats beaucoup moins bons sur le jeu d'entraînement). Ce classifieur obtient les résultats résumés au bas des tableaux 5 et 6. Dans la catégorie Dessert, il se placerait lui aussi dans le groupe des très bons résultats. Dans les deux autres catégories, il est plusieurs points derrière le meilleur système, ce qui montre l'intérêt

des traitements supplémentaires effectués par le LIA au-delà d'une simple segmentation en mots.

La distinction entre entrée et plat principal était plus difficile à établir : de fait, certains plats peuvent se servir indifféremment en entrée ou en plat principal, et les indications de quantité ne suffisent pas non plus à distinguer leur destination. Ceci étant posé, la bonne performance de cette méthode de base montre que cette tâche était nettement plus facile que la tâche 1, parce qu'elle avait une classe de moins, que ses classes étaient plus équilibrées, et que leur définition était moins subjective.

Tâche 3 — Appariement titre/recette. Les scores (MRR) pour la tâche 3 vont de 0,13 à 0,43. L'équipe 02 (Hamon *et al.*, 2013) a fourni zéro ou un titre par recette, l'équipe 04 (Lejeune *et al.*, 2013) exactement un titre par recette, et l'équipe 26 (Collin *et al.*, 2013) a produit 50 titres par recette. Cette dernière se détache des deux autres participants. Un score de 0,43 correspond à un placement du bon titre en moyenne entre les positions 2 et 3, ou encore, pour un système qui ne renverrait qu'un titre par recette, à trouver le bon titre une fois sur 2 à 3. En pratique, les systèmes des participants ont placé le bon titre en première position (P@1, précision au rang 1) dans respectivement 31,4 % des cas (éq. 26), 24,1 % des cas (02), et 12,7 % des cas (04). On vérifie ainsi que seul un système qui classe plusieurs titres pour chaque recette, comme le 26, peut obtenir un MRR supérieur à sa précision au rang 1.

Tâche 4 — Ingrédients d'une recette. Les scores (MAP) pour la tâche 4 s'étagent de 0,36 à 0,66. Le groupe des trois équipes de tête a bien réussi la tâche, avec une meilleure MAP entre 0,63 et 0,66. Ce score correspond par exemple au fait de classer en premier, dans chaque recette, les deux tiers des ingrédients, ce qui est très bon étant donné les défauts de la référence signalés plus haut.

Le tableau 9 montre le nombre de recettes pour lesquelles tous les ingrédients ont été trouvés et classés en premier (précision moyenne = 1) : les équipes 28 (Charton *et al.*, 2013) et 22 (Bost *et al.*, 2013) font beaucoup plus souvent le grand chelem que l'équipe 25 (Dini *et al.*, 2013), qui obtient pourtant la meilleure MAP (et met donc sans doute plus régulièrement les bons ingrédients vers le haut, même si ce n'est pas tout en haut). On reconnaît donc ici une stratégie

Équipe	02 LIM&Bio	04 GREYC	22 LIA	25 Celi France	28 Wikimeta Lab
#1	9	19	**173**	84	53
#2	11	68	147	—	**176**
#3	24	44	141	—	—

TABLE 9 – Tâche 4 : nombre de résultats parfaits (tous les ingrédients corrects en tête de liste)

différente entre l'équipe 25 et les équipes 22 et 28. L'équipe 25 est la seule à avoir appliqué le précepte de la MAP : classer tous les ingrédients est toujours mieux que de n'en classer qu'une partie. Elle a ainsi systématiquement classé une longue liste de quelque 850 ingrédients pour chaque recette. Si par exemple elle n'avait classé que 20 ingrédients par recette, sa MAP aurait été de 0,6513 (−0,015) — on constate que cela n'aurait néanmoins pas modifié le classement.

4 Conclusion

L'édition DEFT2013 a porté sur l'analyse des recettes de cuisine au travers de quatre tâches. La référence a été constituée en reprenant directement les informations présentes sur Marmiton, le site utilisé pour constituer les corpus. Du fait de cette référence « naturelle », les tâches présentaient des difficultés importantes : définitions subjectives et classes très déséquilibrées pour la difficulté de la recette, similarité entre entrée et plat principal dans les types de plats, créativité des titres de recettes, référence imparfaite et difficultés de normalisation dans la détection des ingrédients.

Au regard de ces difficultés, il apparaît que les participants ont réussi fort honorablement à prédire les valeurs renseignées par les humains qui déposent une recette sur le site. On peut même supposer que les méthodes développées pour estimer la difficulté d'une recette ou détecter ses ingrédients pourraient être utiles pour indexer automatiquement les recettes du site et lui apporter une plus grande cohérence.

Remerciements. Nous remercions Flora Badin (INIST–CNRS) et Dominic Forest (EBSI–Université de Montréal) pour les réflexions qu'ils ont menées sur les tâches à proposer aux participants du défi cette année. Nous remercions également les étudiants français (INaLCO) et canadiens (EBSI) pour les tests humains qu'ils ont réalisés et l'aide qu'ils nous ont ainsi apportée. Merci à tous les participants pour les efforts déployés et la richesse des méthodes utilisées. Nous les félicitons également pour la manière dont ils ont filé la métaphore culinaire dans leurs articles ! Enfin, nous remercions les organisateurs de TALN/Recital 2013 pour leur aide logistique dans la préparation de l'atelier de clôture de DEFT.

Références

Bost, X., Brunetti, I., Cabrera-Diego, L. A., Cossu, J.-V., Linhares, A., Morchid, M., Torres-Moreno, J.-M., El Bèze, M. et Dufour, R. (2013). Système du LIA à DEFT'13. *In Actes de DEFT*.

Charton, E., Jean-Louis, L., Meurs, M.-J. et Gagnon, M. (2013). Trois recettes d'apprentissage automatique pour un système d'extraction d'information et de classification de recettes de cuisine. *In Actes de DEFT*.

Collin, O., Guerraz, A., Hiou, Y. et Voisine, N. (2013). Participation de Orange Labs à DEFT 2013. *In Actes de DEFT*.

Dini, L., Bittar, A. et Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine. *In Actes de DEFT*.

Hamon, T., Périnet, A. et Grabar, N. (2013). Efficacité combinée du flou et de l'exact des recettes de cuisine. *In Actes de DEFT*.

Lejeune, G., Lecluze, C. et Brixtel, R. (2013). DEFT2013, une cuisine de caractères. *In Actes de DEFT*.