Efficacité combinée du flou et de l'exact des recettes de cuisine

Thierry Hamon[1] Amandine Périnet[1,2] Natalia Grabar[3]
(1) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, 74, rue Marcel Cachin, 93017 Bobigny
thierry.hamon@univ-paris13.fr
(2) Lingua et Machina c/o INRIA Rocquencourt BP 105 78153 Le Chesnay Cedex
amandine.perinet@lingua-et-machina.com
(3) CNRS UMR 8163 STL, Université Lille 1&3, 59653 Villeneuve d'Ascq
natalia.grabar@univ-lille3.fr

RÉSUMÉ

La cuisine et les recettes de cuisine occupent une place importante dans notre vie. L'alimentation a par ailleurs une influence directe sur notre bien-être, santé et bonne humeur. La compétition DEFT 2013 étant dédiée à l'analyse des recettes de cuisine, nous y avons participé pour étudier ce domaine de spécialité. Quatre tâches orientées catégorisation et extraction d'informations sont proposées. En fonction des tâches, nous utilisons des approches à base de règles et d'apprentissage. Nous avons aussi construit des ressources sémantiques pour traiter des informations exactes (par exemple, quantités exprimées en grammes ou litres, durées exprimées en minutes ou jours) et floues (par exemple, quantités exprimées en *chouilla* et *louche*, durées séquencées avec *après, ensuite, alors que*). En exploitant notre approches et nos ressources, et en naviguant entre l'exact et le flou, nous avons obtenu les résultats officiels suivants : 4e/6 sur la tâche 1, 3e/5 sur la tâche 2, 2e/3 sur la tâche 3, et 5e/5 sur la tâche 4. Plusieurs améliorations sont envisagées.

ABSTRACT

Combined efficiency of fuzziness and exactness in recipes.

An important place in our lifes is dedicated to *cuisine* and recipes. Besides, food has a direct impact on our well-being, health and mood. Because the DEFT 2013 challenge was dedicated to the analysis of recipes, we participated in this challenge to study this specialized area. Four tasks oriented on categorization and information extraction have been proposed. According to tasks, we use rule-based and machine learning approaches. We have also built semantic resources to process exact (*i.e.*, quantities expressed in grams or liters, durations expressed in minutes or days) and fuzzy (*i.e.*, quantities expressed in *chouilla* and *louche*, durations sequenced with *après, ensuite, alors que*) information. Using the approaches and resources, and sailing between the exactness and fuziness, we have obtained the following official results : 4e/6 on task 1, 3e/5 on task 2, 2e/3 on task 3, and 5e/5 on task 4. Several improvements are planned.

MOTS-CLÉS : Traitement Automatique des Langues, campagne d'évaluation, recettes de cuisine, annotation sémantique, création de ressources sémantiques, systèmes à base de règles, système d'apprentissage automatique.

KEYWORDS: Natural Language Processing, evaluation campaign, recipes, semantic annotation, semantic resource creation, rules-based system, machine learning system.

1 Introduction

La cuisine et les recettes de cuisine occupent une place importante dans notre vie. Elles font certainement aussi partie des facteurs principaux qui influencent notre bien-être et la bonne humeur au quotidien. En effet, selon un dicton russe, *Любовь и голод правят миром* (*L'amour et la faim gouvernent le monde*). D'autres dictons marquent aussi l'importance de la cuisine et de la nourriture dans notre vie (de Rudder, 2006; Lim, 2007) : *les carottes sont cuites*, *la fin des haricots*, *s'occuper de ses oignons* ou *avoir mangé son pain blanc*. Même nos hommes les plus illustres s'intéressent à la cuisine (Dumas, 1873).

Il n'est donc pas étonnant que les scientifiques aussi s'intéressent à l'alimentation et la nutrition. Des phénomènes connus au grand public comme ceux de *malbouffe* ou de *French paradox* sont étudiés. Si ces études expliquent et justifient les bienfaits du vin rouge et de ses substances, elles montrent aussi que la malbouffe (par exemple, la consommation de hamburgers, de hot-dogs, de frites, de chips ou de sodas) peut favoriser l'obésité, le diabète, les maladies cardiovasculaires, des dépressions et même certains cancers. Il ne faut donc pas sous-estimer la place que l'alimentation, et la bonne alimentation surtout, occupent dans notre vie. Pour mieux étudier les questions relatives à l'alimentation de la population, il existe même une sous-section 44.04 de CNU dédiée à la nutrition. L'effort principal des chercheurs consiste à attirer notre attention sur la qualité de notre alimentation et son influence sur notre santé. Depuis quelques années, l'étude NutriNet-Santé [1] semble avoir concentré les efforts et les initiatives autour de ces questions en France. Dans de telles études, une attention particulière peut être portée à certaines pathologies, comme le diabète, à la condition sociale ou à la provenance géographique de la population étudiée (Jones-Smith *et al.*, 2013; Andreeva *et al.*, 2013).

Dans les domaines de l'informatique et du Traitement Automatique de Langues (TAL), les textes de recettes de cuisine ont aussi attiré l'attention des chercheurs. Le premier travail est certainement celui décrivant le système Epicure et son application aux recettes de cuisine (Dale, 1989). Ce système a pour objectif de générer une description linguistique des recettes de cuisine. Des représentations syntaxiques profonde et de surface sont proposées en utilisant la grammaire d'unification. Le fonctionnement du système s'appuie sur des connaissances du domaine, comme par exemple les objets (individuels, massifs, quantifiés ou non) encodés dans une ontologie. Les objets évoluent en fonction des états qu'ils subissent. Ce système de TAL prend aussi en compte la pronominalisation et les anaphores. L'objectif de ce système semble d'avoir implémenté une approche sémantique fine. Le passage à l'échelle et le traitement de grosses données n'étaient pas au centre d'intérêt à l'époque.

Un travail plus récent est fait dans le cadre du Computer Cooking Contest [2], qui a lieu tous les ans depuis 2008. Cette compétition propose de traiter les recettes de cuisine. En 2012, les objectifs étaient clairement orientés sur le Raisonnement à partir de cas. Par exemple, il s'agit de construire automatiquement un système à base de cas (1) d'une part pour en montrer la faisabilité à partir du texte libre et (2) d'autre part pour proposer de nouvelles recettes en se basant sur les recettes existantes et en fonction des spécifications données (Dufour-Lussier *et al.*, 2013). Dans ce travail cité, l'attention particulière est réservée aux actions (souvent des verbes) et à leurs arguments. En plus, plusieurs traitements de TAL sont appliqués : étiquetage morpho-syntaxique, résolution d'anaphores, analyse syntaxique. 15 recettes de cuisine sont traitées. Cette compétition

1. *https ://www.etude-nutrinet-sante.fr*
2. *http ://computercookingcontest.net*

a donné aussi lieu aux travaux interdisciplinaires sur la reconnaissance d'aliments dans notre frigo (Kamoda *et al.*, 2012), ou encore sur l'apprentissage et modélisation des gestes culinaires grâce à des gants spéciaux (Ota *et al.*, 2012).

La compétition de TAL DEFT 2013 propose aussi de travailler avec les recettes de cuisine. Devant cette perspective de découverte et de mise au point scientifique, culturelle et culinaire, nous avons décidé de participer au challenge. Les quatre tâches sympathiques de la compétition sont orientées sur la catégorisation des recettes (tâches 1 et 2) et l'extraction d'information (tâches 3 et 4). Ces tâches sont les suivantes :

1. Identifier à partir du titre et du texte de la recette son niveau de difficulté sur une échelle à 4 niveaux : très facile, facile, moyennement difficile, difficile.
2. Identifier à partir du titre et du texte de la recette le type de plat préparé : entrée, plat principal, dessert.
3. Apparier le texte d'une recette à son titre.
4. Extraire du titre et du texte d'une recette la liste de ses ingrédients.

Il va de soi que, pour aborder efficacement et positivement le traitement des recettes de cuisine, il faut être muni de bons ingrédients linguistiques (section 2) et de bons ustensiles de TAL (section 3). Nous présentons aussi les recettes proposées (section 4) et les résultats que nous avons pu obtenir en sortie (section 5). Comme le perfectionnement (au moins scientifique, culturel et culinaire) n'a pas de limites, nous mentionnons quelques perspectives que nous aimerions mettre en oeuvre dans l'avenir proche (section 6).

2 Ingrédients

Parmi les ingrédients utilisés, nous avons plusieurs ressources (résumées dans la table 1) :

– Une liste d'ingrédients de cuisine et d'aliments collectés dans des sources disponibles en ligne [3] [4] [5] [6], dans la partie française de l'UMLS (NLM, 2011) et à partir des ingrédients connus dans l'ensemble d'entraînement de la campagne. Une des difficultés a été de distinguer entre les ingrédients et les aliments. La solution que nous avons adoptée est de considérer que les ingrédients sont "bruts" alors que les aliments sont déjà préparés voire cuisinés. Logiquement, ce sont les ingrédients qui sont utilisés dans les recettes... Ce qui n'est bien sûr pas totalement vrai car de bons aliments comme les pâtes, les raviolis, les glaces, les sauces diverses et variées, sans parler des fromages, font aussi de très bons ingrédients irremplaçables dans notre cuisine. Devant ce flou inattendu, nous avons finalement décidé (1) d'introduire l'ambiguïté et de catégoriser les aliments concernés comme ingrédients aussi, ou bien (2) de simplifier la réalité et de les considérer comme de vrais ingrédients. Cette liste comporte 6 070 entrées catégorisées en viandes, légumes, fruits, produits de boulangerie, poissons, sucreries, etc.
– Une liste d'ustensiles collectés à partir de ressources disponibles en ligne [7] [8]. Cette liste comporte 222 entrées.

3. *http ://www.bioweight.com/glucides.html*
4. *http ://www.bioweight.com/proteines.html*
5. *http ://www.centre-clauderer.com/acides-bases/femme-2.htm*
6. *http ://les.calories.free.fr/*
7. *http ://popoblog.unblog.fr/liste-ustensiles-de-cuisine-mise-a-jour-le-130808/*
8. *fr.wikipedia.org/wiki/Ustensile_de_cuisine*

– Une liste de mots vides du français composée de 616 entrées et une liste d'exceptions composée de 37 entrées.
– Une ressource pour la détection de quantités des ingrédients, comme par exemple : *250 g de sucre, 3 oeufs, une bonne cuillère d'huile, 100 + 50 gr de beurre, 1/2 l de lait, deux graines de cardamone, un chouilla de sel, 2 louches de bouillon, beaucoup de menthe*. Nous avons distingué les quantités standards, déjà exprimées en grammes ou litres, et les quantités non standards (Delamasure, 2007), non exprimées en grammes ou en litres, mais qui font le charme de nos recettes de cuisine et en garantissent la réussite. Afin d'en assurer un traitement égalitaire et une utilisation possible dans un système de TAL, nous avons établi des heuristiques et avons converti toutes ces quantités en litres ou en kilogrammes. Nous nous sommes aidés pour ceci d'autres ressources et convertisseurs disponibles en ligne [9] [10]. Lorsque plusieurs équivalences étaient connues ou lorsque la question de conversion ne s'était pas encore posée, nous avons avancé nos propres propositions. Par exemple, les expressions comme *pincée, soupçon, chouilla, lichette, rasade, giclée, goutte, microgramme, 1 peu* signifient que la recette comporte 1 gramme de produit ; les expressions comme *louche, tube, fond, ravier, assiette creuse, poignée* signifient que la recette comporte 100 grammes de produit, etc. Lors de la normalisation, plusieurs opérations arithmétiques sont appliquées (somme, division, multiplication).
– Une ressource pour détecter des durées. Là aussi, nous distinguons la durée quantifiée et exprimée en minutes, heures, jours, nuits, etc., et la durée non quantifiée et séquencée par des expressions comme *puis, pendant que, alors que, ensuite*.
– Une liste d'actions (verbes et noms) de la ressource Verbaction (Hathout *et al.*, 2001).
– Une ressource distributionnelle (Harris, 1954) construite à partir des titres et du contenu des recettes (ensembles d'entraînement et de test). L'approche s'appuie sur deux méthodes d'acquisition de relations d'hyperonymie (Morin et Jacquemin, 2004; Bodenreider *et al.*, 2001) et une méthode l'acquisition de variantes terminologiques (Jacquemin, 1996). Pour délimiter les contextes de mots, nous utilisons des fenêtres graphiques de 10 mots à droite et 10 mots à gauche, en limitant les mots en relation aux mots appartenant à la même catégorie syntaxique (noms, adjectifs et termes identifiés par l'extracteur YaTeA (Aubin et Hamon, 2006)). En ce qui concerne la mesure de similarité, nous utilisons l'indice de Jaccard (Grefenstette et Tapanainen, 1994), qui normalise le nombre de contextes partagés par deux mots par le nombre total de contextes de ces mots. Nous avons également filtré les relations acquises avec la moyenne de trois seuils (nombre de contextes partagés, fréquence des contextes partagés, fréquence des mots pivots).
– Une ressource distributionnelle *FreDist* (Anguiano et Denis, 2011).
– Une ressource de synonymes de la langue générale fournis par le Petit Robert (Robert, 1990).

L'ingrédient principal reste bien sûr l'ensemble d'entraînement fourni par les organisateurs. Cet ensemble comporte 13 864 recettes. Chaque recette contient les informations suivantes : le titre, les ingrédients, les étapes ou le déroulement de la préparation, le coût, le niveau de difficulté et le type de plat. Mais seule le coût est disponible dans l'ensemble de test. Souvent, la boisson conseillée est aussi indiquée. L'ensemble d'entraînement, aidé par les ressources, a permis de créer le système de TAL et d'en faire les tests. Un autre ingrédient important sont les recettes du corpus de test, au nombre de 9 200 (2 300 par tâche). Ce corpus a été fourni pour une durée de trois jours : comme le feu de nos cuisinières ne chauffait pas assez fort et vite et pour ne pas risquer l'indigestion, le système de TAL était le seul moyen de test possible pour traiter les quatre tâches.

9. *http ://www.supertoinette.com/mesures-equivalences-culinaires.html*
10. *http ://webcafe.highbb.com/t1827-mesures-metriques-et-nord-americaine#13245*

Ressources	Exactes	Floues
Ressources lexicales	+ (ingrédients, ustensiles, synonymes)	+ (distributionnelles, aliments)
Quantités	+ exprimées en grammes et litres	+ exprimées autrement
Durées	+ quantifiées	+ non quantifiées
Actions	+	
Mots vides	+	
Recettes de cuisine	+ ensemble d'entraînement	+ ensemble de test

TABLE 1 – Ressources construites et utilisées afin de pouvoir traiter les informations exactes et floues contenues dans les recettes de cuisine.

3 Ustensiles de TAL

Nos ustensiles de TAL sont composés d'un système à base de règles et d'un système d'apprentissage automatique.

3.1 Système à base de règles

L'objectif du système à base de règles est d'assurer la reconnaissance de mots clés (ingrédients, ustensiles de cuisine...) et d'informations associées (quantités d'ingrédients, durées de préparation...). Ce système a quatre étapes principales. Trois étapes (reconnaissance de termes et d'informations associées, pondération de termes, et leur filtrage et sélection) sont adaptées du système utilisé lors de la compétition DEFT 2012 (Hamon, 2012). La quatrième étape, dédiée à l'appariement de titres, est adaptée du travail précédent (Hamon et Gagnayre, 2012).

3.1.1 Reconnaissance de termes (TermTagger) et d'informations associées

Les ressources décrites plus haut (listes d'ingrédients, d'ustensiles et d'actions) sont projetées sur les textes traités. Cette projection prend en compte les lemmes de mots du corpus. Nous utilisons le module Perl `Alvis::TermTagger`[11]. Parallèlement à la reconnaissance des termes, les entités nommées associées sont aussi recherchées. Par exemple, nous exploitons les ressources décrites plus haut pour reconnaître les quantités des ingrédients ou bien les durées nécessaires à la préparation d'un recette.

3.1.2 Pondération de termes

Pour identifier les termes les plus importants parmi ceux qui sont reconnus, nous les trions par ordre de pertinence avec plusieurs méthodes de pondération :

11. `http://search.cpan.org/~thhamon/Alvis-TermTagger/`

– La fréquence du terme dans le document (**tf**) ;
– Le nombre de documents du corpus où le terme apparaît (**df**) ;
– La position de la première occurrence du terme (**position**). Nous considérons ici que les termes situés au début du document ont un poids plus élevé que ceux situés à la fin ;
– Le cosinus de la position de la première occurrence du terme (**positionCos**). Nous faisons l'hypothèse que les termes qui apparaissent au début ou à la fin du texte sont les plus pertinents. Nous plaçons ainsi la première occurrence de chaque terme sur le cercle trigonométrique en considérant que le début du document est l'angle 0 et la fin l'angle 2π. Nous avons calculé le cosinus de l'angle formé par la position ;
– La fréquence de la forme canonique (lemmatisée) du terme (**canon**) ;
– Le fait que les termes sont quantifiés (**quant**). Ceci est essentiellement le cas des ingrédients.

3.1.3 Filtrage et sélection des termes

Les termes peuvent ensuite être filtrés ou regroupés suivant différents critères :

– Suppression des termes isolés étiquetés comme adjectifs (**filtrAdj**). Il s'agit alors de modifieurs de termes complexes, qui n'ont pas lieu d'apparaître tout seuls. Les adjectifs correspondant à des termes des ressources sont conservés ;
– Prise en compte de l'inclusion lexicale (**filtrInclLex**). Les termes en position tête d'un terme et ayant de rang plus élevé dans la liste triée sont supprimés. Par exemple, nous traitons de cette manière les termes comme {*crème anglaise*, *crème*}, {*pomme de terre*, *pomme*}, {*fève de tonka*, *fève*}. Les calculs des inclusions sont effectués au niveau syntaxique et au niveau des chaînes de caractères, lorsque la syntaxe ne détecte pas de relations ni d'inclusions ;
– Regroupement des termes en fonction de leur forme canonique (**filtrCan**). Il s'agit du regroupement des lemmes des composants et du filtrage par la forme fléchie la plus fréquente.

3.1.4 Appariement de titres

L'appariement de titres avec les recettes est effectué suivant les étapes suivantes :
– *Pré-traitement des titres et leur conversion en mots-clés.* À cette étape, les titres sont segmentés en mots, étiquetés morpho-syntaxiquement et lemmatisés (Schmid, 1994). En fonction des expériences, tous les mots ou seulement les mots "significatifs" (noms, adjectifs et verbes) sont retenus. Dans ce dernier cas, les mots grammaticaux et les mots de la liste d'exceptions ne sont pas considérés. Par exemple, le titre *Tarte au caramel et aux bananes* est converti en mots-clés suivants (*tarte, caramel, banane*).
– *Enrichissement de mots-clés avec les ressources linguistiques.* Les mots-clés sont étendus avec les ressources linguistiques : synonymes ou les mots associés selon les ressources. Chaque mot-clé, et les mots faisant partie de son extension, forment un cluster. Par exemple, le mot-clé *tarte* est enrichi avec *gâteau, pâtisserie, tartelette, gaufre, cake, galette, crêpe, beignet*, le mot-clé *banane* avec *pomme, fruit, abricot, poire*, et le mot-clé *caramel* avec *chocolat, beurre, vanille, croûton, pâte, sirop*. Les mots-clés sont considérés comme les labels sémantiques de leurs clusters : *tarte, banane, caramel*, respectivement. De cette manière, chaque titre est décrit par *n* clusters, en fonction du nombre de ses mots-clés.
– *Pré-traitement des recettes et leur sélection.* Les recettes sont pré-traitées de la même manière que les titres, et segmentées en phrases. Ensuite, les mots et termes des clusters sont appariés

avec les mots et les termes des recettes. C'est l'étape la plus importante : en plus de l'expansion et de l'appariement, une sélection des appariements entre les titres et les recettes est effectuée. Notons qu'à cette étape, il est possible d'effectuer l'appariement complet ou partiel (avec un pourcentage donné) entre les titres et les recettes. Par exemple, nous pouvons apparier le titre *Tarte au caramel et aux bananes* à la recette 90954 grâce à l'appariement direct et partiel (le mot-clé *tarte* n'apparaît pas dans le texte de la recette).

3.2 Système d'apprentissage automatique

Le système d'apprentissage est basé sur les fonctionnalités proposées par la plate-forme Weka (Witten et Frank, 2005). Plus particulièrement, nous utilisons et effectuons :
- le sur-échantillonnage avec l'algorithme *SMOTE* (Chawla *et al.*, 2002), particulièrement utile sur la tâche 1 (difficulté des recettes), où il existe un déséquilibre net dans les données ;
- la sélection d'attributs (avec *CFS combiné avec BestFirst* (Hall, 1998)), particulièrement utile pour rendre possibles les traitements étant donné le nombre total d'attributs ($n > 20\,000$) ;
- les algorithmes d'apprentissage testés sont : SVM, arbres de décision, régression logistique ;
- les attributs exploités sont assez variés, par exemple :
  - les lemmes ($n = 8\,145$) ;
  - les formes fléchies qui ne sont pas les lemmes ($n = 7\,830$) ;
  - les étiquettes morpho-syntaxiques ($n = 34$) ;
  - les étiquettes sémantiques provenant des ressources ($n = 61$), dont les types d'ingrédients, les ustensiles, les verbes d'action ;
  - le coût estimé d'une recette ;
  - la taille de la recette en nombre de caractères et de mots ;
  - le nombre d'étapes explicitement indiquées dans la recette ;
  - le nombre d'étapes implicites dans la recette ;
  - le nombre total d'étapes dans une recette ;
  - le nombre de lignes dans les ingrédients lorsque disponibles ;
  - la liste des actions ;
  - le nombre de documents du corpus où le terme apparaît (**df**) ;
  - le nombre d'occurrences des notions suivantes : sucre, sel, poivre, piment, huile, moutarde. Par exemple, lorsque *sucre, cassonade, sucre roux* sont reconnus, ils incrémentent la notion de sucre ;
  - les indications de temps exactes ou floues ;
  - les quantités exactes ou floues.

Les classes recherchées varient en fonction des tâches.

4 Recettes et tests réalisés

Pour aborder les tâches de la compétition, et comme *chaque tache a sa recette*, nous avons exploité le corpus d'entraînement pour mettre au point des stratégies différentes selon les tâches.

4.1 Tâche 1 : niveau de difficulté

Les textes des recettes sont d'abord traités avec le système à base de règles (section 3.1). Ensuite le système d'apprentissage (section 3.2) est appliqué. Les attributs exploités sont les suivants :
– les lemmes ;
– les étiquettes morpho-syntaxiques ;
– les étiquettes sémantiques provenant des ressources ;
– le coût estimé d'une recette ;
– la taille de la recette en nombre de caractères et de mots ;
– le nombre d'étapes explicitement indiquées dans la recette ;
– le nombre d'étapes implicites dans la recette ;
– le nombre total d'étapes dans une recette ;
– le nombre de lignes dans les ingrédients lorsque disponibles ;
– la liste des actions ;
– le nombre de documents du corpus où le terme apparaît (**df**) ;
– les indications de temps exactes ou floues ;
– les quantités exactes ou floues.
Il existe un déséquilibre dans les données du corpus d'entraînement, où les plat difficiles sont clairement sous-représentés :
– plats très faciles : 6 962,
– plats faciles : 5 752,
– plats moyennement difficiles : 1 068,
– plats difficiles : 80.
Nous effectuons donc un sur-échantillonage avec *SMOTE* : +1000% pour la classe 4, +200% pour la classe 3, combinaison des deux, tests d'autres valeurs par défaut. Nous pouvons ainsi constaté qu'il existe une amélioration nette lorsque le sur-échantillonage est effectué sur la classe 4 (la moins bien représentée).

Nous effectuons aussi une sélection d'attributs et pouvons par exemple constater que :
– très peu d'attributs apparaissent utiles ($n = 53$) : catégories morpho-syntaxiques, des lemmes, mais très peu de formes fléchies ce qui explique qu'elles ne sont pas utilisées,
– les catégories sémantiques ne sont pas saillantes,
– la taille et les sections des recettes sont utiles,
– les actions et le coût se révèlent importants aussi,
– les durées et le temps, exact ou flou, sont utiles,
– de même que la fréquence dans le corpus total de certains ingrédients.

4.2 Tâche 2 : type de plat

Les textes des recettes sont d'abord traités avec le système à base de règles (section 3.1). Ensuite le système d'apprentissage (section 3.2) est appliqué. Les attributs exploités sont les suivants :
– les lemmes ($n = 8\ 145$) ;
– les étiquettes morpho-syntaxiques ($n = 34$) ;
– les étiquettes sémantiques provenant des ressources ;
– le coût estimé d'une recette ;
– la taille de la recette en nombre de caractères et de mots ;
– le nombre d'étapes explicitement indiquées dans la recette ;

– le nombre d'étapes implicites dans la recette ;
– le nombre total d'étapes dans une recette ;
– le nombre de lignes dans les ingrédients lorsque disponibles ;
– la liste des actions ;
– la fréquence des ingrédients dans le corpus total (**df**) ;
– le nombre d'occurrences des notions suivantes : sucre, sel, poivre, piment, huile, moutarde.
– les indications de temps exactes ou floues ;
– les quantités exactes ou floues.

Nous avons testé le sur-échantillonnage, mais comme les données sont équilibrées pour cette tâche (3 246 entrées, 6 449 plats, 4 169 desserts), le sur-échantillonnage ne fournit pas de gain.

Nous effectuons aussi la sélection d'attributs et pouvons constater que :
– une réduction importante du nombre d'attributs est effectuée ($n = 72$),
– beaucoup de lemmes sont gardés, mais seulement 4 formes fléchies, qui ne seront donc pas utilisées pour cette tâche non plus,
– les catégories morpho-syntaxiques ne semblent pas être saillantes,
– les catégories sémantiques ne sont pas importantes,
– les quantités (exactes ou floues) de certains ingrédients (sel, poivre, moutarde) sont utiles,
– les indications sur les durées ne sont pas saillantes,
– certaines actions (comme *entrer*), dont la nominalisation est le type de plat *entrée*,
– le nombre de documents où apparaît l'ingrédient.

4.3 Tâche 3 : appariement du texte d'une recette à son titre

L'appariement des titres avec les textes des recettes est effectué avec l'approche décrite dans la section 3.1.4. Nous testons l'influence des ressources sémantiques (synonymes, distributionnelles) et constatons par exemple que les synonymes détériorent les résultats. Nous testons aussi l'utilisation de tous les mots ou des mots "significatifs" seulement et constatons que dans ce dernier cas, les résultats sont améliorés. Comme l'appariement à 100 % ne couvre pas la moitié de recettes, nous appliquons l'appariement partiel, ce qui permet d'augmenter la couverture. Afin de limiter l'appariement de plusieurs recettes à un même titre, lorsqu'un titre est affecté à une recette, celui-ci est retiré de l'ensemble des titres à assigner. Pour une recette donnée, les titres candidats sont ordonnés en fonction du rapport entre le nombre de mots appariés et le nombre maximum de mots appariés parmi tous les titres candidats. Parmi tous les titres candidats, le choix du titre est effectué de la manière suivante :
– le premier titre candidat, qui n'a pas déjà été assigné, est sélectionné,
– lorsqu'il existe aucun titre candidat non assigné, c'est le premier titre candidat qui est sélectionné.

Pour les recettes qui n'ont pas de titres avec des correspondances lexicales appariées ou bien lorsque cet appariement est inférieur à un seuil donné (par exemple, 75 % ou 50 %), aucun titre ne leur est assigné.

4.4 Tâche 4 : extraction de la liste des ingrédients

Pour l'extraction des ingrédients du texte des recettes, nous exploitons les étapes 3.1.1 à 3.1.3 du système basé sur les règles. Cela veut dire que les ressources sémantiques sont projetées sur les

Tâche/Run	Macro			Micro			MRR/MAP
	R	P	F	R	P	F	
T1/R1	0.399	0.332	0.363	0.580	0.580	0.580	
T1/R2	0.415	0.339	0.373	0.586	0.586	**0.586**	
T1/R3	0.406	0.329	0.364	0.574	0.574	0.574	
T2/R1	0.806	0.832	0.819	0.834	0.834	0.834	
T2/R2	0.841	0.842	0.841	0.849	0.849	**0.849**	
T2/R3	0.806	0.832	0.819	0.834	0.834	0.834	
T3/R1	0.036	0.019	0.025	0.036	0.036	0.036	0.0360
T3/R2	0.241	0.200	0.218	0.241	0.241	**0.241**	**0.2413**
T3/R3	0.133	0.119	0.125	0.133	0.133	0.133	0.1326
T4/R1							0.4115/0.4543
T4/R2							0.4170/0.4596
T4/R3							**0.4649**/0.4705

TABLE 2 – Performances officielles sur les quatre tâches sur le corpus de test

textes des recettes afin d'y reconnaître les ingrédients. Ensuite, les ingrédients sont pondérés avec les différentes méthodes. La solution qui s'avère être la plus efficace est la suivante : **position** * (**tf** des ingrédients quantifiés). Lorsque plusieurs ingrédients ont un poids égal, ils sont ordonnés en fonction de la fréquence **canon** dans la recette.

Les difficultés qui restent avec cette tâche sont :
– hésitation entre les formes courtes et étendues des ingrédients,
– hésitation entre les formes fléchies et lemmatisées des ingrédients.
Le calcul des inclusions lexicales syntaxiques et au niveau des chaînes de caractères a été effectué à cet effet. Cela a permis de faire des tests systématiquement, mais il reste difficile de reproduire la logique des données de référence et de la section *Ingrédients* des recettes. Le plus souvent, notre approche permet d'extraire les ingrédients mais leurs formes ou positions pondérées peuvent ne pas être correctes.

5 Résultats obtenus

Nous avons soumis trois runs pour chacune des quatre tâches de la compétition. Notre classement officiel est le suivant :
– tâche 1 : 4e/6
– tâche 2 : 3e/5
– tâche 3 : 2e/3
– tâche 4 : 5e/5
Les résultats détaillés pour ces tâches sont indiqués dans le tableau 2.

Selon les organisateurs de la compétition :
– Sur la première tâche (niveau de difficulté), six équipes ont participé. Les résultats globaux obtenus en termes de micro-mesure varient de 0,625 à 0,489 sur les meilleures soumissions de chaque équipe (micro mesure moyenne = 0,569, médiane = 0,589).
– Sur la deuxième tâche (type de plat), cinq équipes ont participé. Les résultats globaux

obtenus varient entre 0,889 et 0,746 (micro-mesure) sur les meilleures soumissions de chacun (moyenne de 0,833, médiane de 0,849).
– Sur la troisième tâche (appariement titre/recette), trois équipes ont participé. Les micro-mesures varient de 0,314 à 0,127 (moyenne de 0,227, médiane de 0,241). Le MRR (Mean Reciprocal Rank) calculé varie de 0,434 à 0,196 sur ces mêmes soumissions.
– Sur la quatrième tâche (normalisation des ingrédients), cinq équipes ont participé. Les résultats, évalués en termes de MAP (Mean Average Precision) varient, sur les meilleures soumissions de chaque équipe, entre 0,6662 et 0,4649 (moyenne de 0,5916, médiane de 0,6287).

Par rapport à d'autres participants, nous sommes au-dessus des moyennes sauf pour la tâche 4. Sur la tâche 4, les résultats produits comportent un défaut de format. Les résultats obtenus après la correction sont : 0.4543, 0.4596, 0.4705 pour les trois runs respectivement. Ces nouveaux résultats sont indiqués après / dans le tableau.

5.1 Tâche 1 : niveau de difficulté

Les runs évalués ont été obtenus avec l'algorithme SVM, le sur-échantillonnage avec *SMOTE*, et la sélection d'attributs. La différence entre les runs est la suivante :
– run 1 : tous les attributs sont utilisés,
– run 2 : les actions ne sont pas utilisées,
– run 3 : les lemmes ne sont pas utilisés, mais les actions sont utilisées.

Au sein de ces tests, c'est la configuration SVM, sur-échantillonnage, sélection d'attributs, et sans les actions qui montre de meilleurs résultats.

5.2 Tâche 2 : type de plat

Les runs évalués ont été obtenus avec l'algorithme SVM. La différence entre les runs est la suivante :
– run 1 : sélection d'attributs, utilisation des actions,
– run 2 : pas de sélection d'attributs, utilisation des actions,
– run 3 : sélection d'attributs, sans l'utilisation des actions.

C'est la configuration SVM, sans sélection d'attributs, et avec utilisation des actions qui donne les meilleurs résultats.

5.3 Tâche 3 : appariement du texte d'une recette à son titre

Les paramètres des runs de la tâche 3 sont :
– Run 1 : catégories majeures (noms, verbes, adjectifs) ; prise en compte de la liste d'exceptions, appariement partiel à 75 % ;
– Run 2 : tous les mots du titre, prise en compte de la liste d'exceptions, appariement partiel à 50 % ;
– Run 3 : catégories majeures (noms, verbes, adjectifs) ; prise en compte de la liste d'exceptions ; utilisation de la ressource distributionnelle ; appariement complet.

C'est le run 2 qui a montré de meilleurs résultats sur les données de test.

5.4 Tâche 4 : extraction de la liste des ingrédients

Pour cette tâche, nous utilisons le module Perl TermTagger. Les paramètres communs des runs soumis sont :
– utilisation de la liste des ingrédients constituée à partir de ressources en ligne,
– assignation du poids comme spécifié dans la section 4.4,
– les termes les plus grands calculés avec les inclusions syntaxiques.
La différence entre les runs est la suivante :
– run 1 : ajout de la liste des ingrédients du corpus d'entraînement,
– run 2 : paramétrage par défaut,
– run 3 : ajout des termes les plus grands calculés avec les inclusions au niveau des chaînes de caractères, et utilisation de FLEMM (Namer, 2000).
Le meilleur run est obtenu avec l'ajout des termes les plus grands calculés avec les inclusions au niveau des chaînes de caractères, et avec l'utilisation de FLEMM. Comme nous l'avons noté plus haut, souvent notre approche permet d'extraire les ingrédients mais leurs formes ou positions pondérées peuvent ne pas être correctes. Il reste en effet difficile de reproduire la logique des données de référence.

6 Conclusions et perspectives de perfectionnement

Nous avons présenté les expériences et tests effectués dans le cadre de la compétition DEFT 2013 dédiée au traitement des recettes de cuisine. Il s'agit d'un domaine intéressant où les informations exactes et floues co-existent et garantissent certainement la réussite de nos recettes de cuisine. Face aux systèmes de TAL, ces informations doivent être traitées de manière spéciale. Nous avons ainsi effectué une normalisation du flou vers de l'exact, en nous inspirant de la norme pifométrique dédiée (Delamasure, 2007) et des instructions précieuses trouvées en ligne. L'approche développée est basée sur un système à base de règles et un système d'apprentissage, de même que différentes ressources. Nous avons obtenus les résultats officiels suivants : 4e/6 sur la tâche 1, 3e/5 sur la tâche 2, 2e/3 sur la tâche 3, et 5e/5 sur la tâche 4.

En travaillant sur les recettes de cuisine, nous nous sommes rendus compte que ces recettes sont bien plus que de simples instructions pour bien remplir nos petits ventres. Les recettes de cuisines deviennent en effet une vitrine d'expression. Par exemple, nous pouvons y trouver l'appel à la consommation de la viande des alligators du Canada, la fierté pour les recettes familiales proposées, les conseils de vie et de bonne nourriture, sans parler des émotions, de l'humour et des opinions.

Il reste plusieurs perspectives à notre travail. Nous avons plusieurs perspectives scientifiques :
– tester mieux l'influence de différents attributs dans les tâches 1 et 2 ;
– faire d'autres tests sur l'ordonnancement des ingrédients grâce à l'utilisation des CRF dans la tâche 4 ;
– effectuer d'autres tests avec les ressources pour l'expansion de titres et l'ordonnancement de titres dans la tâche 3 ;
– prendre en compte d'autres facteurs du flou comme les erreurs d'orthographe ;
– prendre en compte les anaphores ;
– exploiter les relations sémantiques entre les ingrédients pour réduire le nombre d'attributs.

Parmi les perspectives culinaires, nous aimerions tester autant de recettes que possible, moyennant la possibilité de trouver les ingrédients nécessaires (viandes d'alligator, jésus...).

Références

ANDREEVA, V., MARTIN, C., ISSANCHOU, S., HERCBERG, S., KESSE-GUYOT, E. et MÉJEAN, C. (2013). Sociodemographic profiles regarding bitter food consumption. cross-sectional evidence from a general french population. *Appetite*, 67(3):53–60.

ANGUIANO, E. et DENIS, P. (2011). FreDist : Automatic construction of distributional thesauri for French. *In TALN*, pages 119–124.

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 380–387. Springer.

BODENREIDER, O., BURGUN, A. et RINDFLESCH, T. (2001). Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. *In* URI INIST CNRS, éditeur : *Terminologie et Intelligence artificielle (TIA)*, pages 11–21, Nancy.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. et KEGELMEYER, W. P. (2002). Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

DALE, R. (1989). Cooking up referring expressions. *In Annual meeting on Association for Computational Linguistics*, pages 68–75.

DE RUDDER, O. (2006). *Aux petits oignons ! : Cuisine et nourriture dans les expressions de la langue française*. Larousse.

DELAMASURE, M. (2007). Norme française NF UNM 00-003. système d'unités pifométriques. Rapport technique, NA.

DUFOUR-LUSSIER, V., LE BER, F., LIEBER, J. et NAUER, E. (2013). Automatic case acquisition from texts for process-oriented case-based reasoning. *Information Systems*. Comming soon.

DUMAS, A. (1873). *Le grand dictionnaire de cuisine*. Editions Pierre Grobel.

GREFENSTETTE, G. et TAPANAINEN, P. (1994). What is a word, what is a sentence ? Problems of tokenization. *In COMPLEX (Computational lexicography and text research)*, pages 79–87.

HALL, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. Thèse de doctorat, University of Waikato, Hamilton, New Zealand.

HAMON, T. (2012). Acquisition terminologique pour identifier les mots clés d'articles scientifiques. *In Actes de l'atelier DEFT 2012*, pages 25–31, Grenoble, France.

HAMON, T. et GAGNAYRE, R. (2012). Linking expert and lay knowledge with distributional and synonymy resources : application to the mining of diabetes fora. *In Proceedings of The Fourth Swedish Language Technology Conference (SLTC 2012)*, pages 35–36, Lund, Sweden.

HARRIS, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

HATHOUT, N., NAMER, F. et DAL, G. (2001). An experimental constructional database : the MorTAL project. *In* BOUCHER, P., éditeur : *Morphology book*. Cascadilla Press, Cambridge, MA.

JACQUEMIN, C. (1996). A symbolic and surgical acquisition of terms through variation. *In* WERMTER, S., RILOFF, E. et SCHELER, G., éditeurs : *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

JONES-SMITH, J., KARTER, A., WARTON, E., KELLY, M., KERSTEN, E., MOFFET, H., ADLER, N., SCHILLINGER, D. et LARAIA, B. (2013). Obesity and the food environment : Income and ethnicity differences among people with diabetes : The diabetes study of Northern California (DISTANCE). *Diabetes Care*, 36(1):1200–1208.

KAMODA, R., UEDA, M., FUNATOMI, T., IIYAMA, M. et MINOH, M. (2012). Grocery re-identification using load balance feature on the shelf for monitoring grocery inventory. *In Proceedings of the Cooking with Computers workshop (CwC)*, pages 8–18.

LIM, J. S. (2007). Description structurale des proverbes coréens autour des « noms de cuisine ». *In Proceedings of the 26th conference on Lexis and Grammar*, Bonifacio.

MORIN, E. et JACQUEMIN, C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, 38(4):363–396.

NAMER, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, 41(2):523–547.

NLM (2011). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

OTA, S., SOGA, M., YAMAMOTO, N. et TAKI, H. (2012). Design and development of a learning support environment for apple peeling using data gloves. *In Proceedings of the Cooking with Computers workshop (CwC)*, pages 7–12.

ROBERT, L. (1990). *Le petit Robert*. Le Robert, Paris.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

WITTEN, I. et FRANK, E. (2005). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.