Extraction de propriétés de produits

Patrick Marty, Tian Tian, Isabelle Tellier

Extraction de propriétés de produits

Patrick Marty* — **Tian Tian**** — **Isabelle Tellier****

* *LeGuide.com*
** *Lattice, UMR 8094, 1 rue Maurice Arnoux, 92 120 Montrouge*

RÉSUMÉ. *Le travail présenté dans cet article vise à extraire automatiquement certaines caractéristiques de produits à partir de descriptions textuelles fournies par un site marchand. La constitution d'un corpus de référence annoté révèle certains problèmes, provenant à la fois des textes et des particularités de la tâche. Pour l'aborder, nous avons testé deux approches : une méthode d'extraction fondée sur des dictionnaires et une méthode d'apprentissage automatique avec les CRF (Champs Aléatoires Conditionnels), pour lesquels nous avons essayé un grand nombre de modèles. Les résultats de nos expériences montrent les avantages et limites de ces deux méthodes.*

ABSTRACT. *In the work presented here, we try to automatically extract some product properties from descriptive texts provided by a merchant website. The constitution of an annotated reference corpus reveals some problems, not only due to the texts but also to the specificities of the task. To handle it, two distinct approaches have been tested : an extraction method based on dictionaries and a machine learning approach making use of CRFs (Conditional Random Fields), for which a large number of models have been tried. The results of our experiments outline the advantages and drawbacks of these two methods.*

MOTS-CLÉS : *descriptions de produits, extraction d'information, apprentissage automatique, CRF*

KEYWORDS: *product descriptions, information extraction, machine learning, CRFs*

CORIA 2014, pp. 121–136, Nancy, 18-21 mars 2014

1. Introduction

La pertinence d'un moteur de recherche repose beaucoup sur la qualité de son indexation. Dans le cas de sites marchands, les informations importantes à indexer, *i.e.* celles qui peuvent faire l'objet de requêtes d'internautes, figurent la plupart du temps dans des textes libres décrivant les offres de produits. L'objectif du travail présenté ici est d'extraire automatiquement de descriptions d'offres en texte libre des méta-données qui caractérisent le mieux leurs propriétés.

Notre source de données est LeGuide.com, un comparateur de prix. Ses clients sont des sites marchands, dont le contenu est indexé offre par offre. Chaque offre est caractérisée par plusieurs champs parmi lesquels : un identifiant, l'URL du site marchand, un titre et une description en textes libres, une image, un prix et des méta-données précisant certaines caractéristiques de l'offre ou du produit (notamment sa (ou ses) couleur(s), sa (ou ses) matière(s) et sa marque). Cependant, peu de marchands remplissent correctement les informations de leur catalogue dans les champs dédiés. La qualité des méta-données est en particulier très variable, alors que dans la plupart des cas les titres et/ou les descriptions des offres contiennent les informations pertinentes qui devraient y figurer. Pour améliorer la pertinence du moteur de recherche et l'expérience utilisateur, en lui proposant des possibilités de requêtage plus variées, il paraît donc naturel de chercher à extraire ces informations depuis les textes de description.

Pour étudier la faisabilité de cette tâche, abordée récemment dans (Ghani *et al.*, 2006 ; Putthividhya et Hu, 2011) nous nous sommes concentrés sur les offres disponibles en langue française sur une catégorie particulière de produits : les "chaussures femmes", qui est une de celles générant le plus de trafic. Nous précisons d'abord les difficultés identifiées lors de la constitution d'un corpus de référence annoté issu de cette catégorie de produits. Puis nous essayons d'explorer deux approches possibles pour aborder cette tâche : une à base de dictionnaires construits manuellement, une autre faisant appel à l'apprentissage automatique supervisé (en l'occurrence les CRF). Nous montrons que, suivant la nature de la propriété à extraire, l'une ou l'autre de ces méthodes est plus performante.

2. La tâche

2.1. *Spécificités et difficultés de l'extraction*

Les offres en français de la catégorie "chaussures femmes" indexées par LeGuide.com pour la France sont au nombre de 271125. Les méta-données que nous cherchons à extraire des descriptions en textes libres sont de trois types : la marque, la couleur et la matière. La marque est un champ à valeur unique, mais ce n'est pas nécessairement le cas des deux autres propriétés. Par exemple : " Derbies Type/Description : Chaussure à lacet Tbs pour femme. Dessus/Tige : cuir Intérieur : cuir Semelle : élastomère "

Dans cet exemple, il y a deux occurrences de "cuir" et une d'"élastomère" qui, toutes, correspondent au champ "matière". Ces trois valeurs sont toutes des cibles de l'extraction. Nous présentons dans ce qui suit les différentes difficultés de la tâche, telles qu'elles sont apparues lors de la constitution de corpus de référence (décrits dans la partie suivante).

2.1.1. *Problèmes de forme*

Les textes originaux de la base de données ne sont pas toujours bien formés. Certains sont mal segmentés, mal encodés ou incomplets.

Ainsi, dans l'exemple de la Figure 1, l'espace manque dans le token "tissuintérieur" qui est donc mal segmenté, alors que "tissu" devrait être repéré comme une matière. L'unité d'extraction étant le token, il ne sera pas possible de l'extraire correctement à partir du texte. D'autres textes présentent des problèmes d'apostrophes absentes, qui induisent le même genre de difficultés.

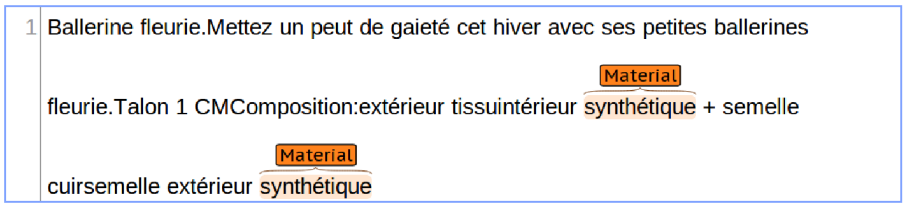

Figure 1 – Description avec mots collés

L'exemple de la Figure 2 est typique d'un problème de codage, dû à une mauvaise saisie de la part du marchand. Nous avons gardé ces textes comme des données dans notre base, mais sans en extraire les éventuelles valeurs qui y figurent.

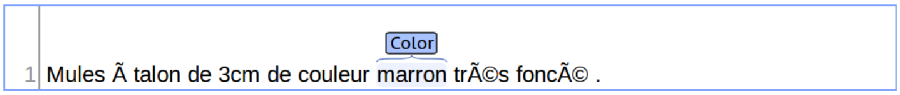

Figure 2 – Description avec problème de codage

Dans l'exemple de la Figure 3, un mot est écrit en abrégé. Un être humain pourrait compléter "synth" pour former "synthétique", qui est une matière. Mais nous avons fait le choix de ne chercher à extraire que les mots bien formés, cette valeur du champ matière sera donc ignorée.

2.1.2. *Mots composés*

Certaines valeurs de champs sont formées de plusieurs tokens constituant un mot composé. Le tableau 1 montre que, suivant les cas, la sémantique du mot composé

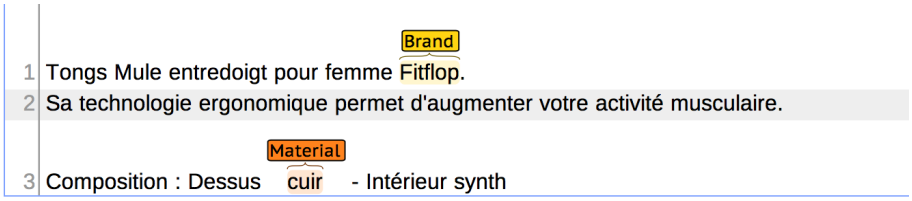

Figure 3 – Description avec mot incomplet

diffère fondamentalement -ou pas- de celle de sa tête : du "faux cuir" n'est pas du "cuir" mais "bleu ciel" est bien un type de "bleu". Un internaute ne s'offusquerait pas de se voir proposer un produit "bleu ciel" alors qu'il en a demandé un bleu, mais il pourrait légitimement se plaindre de voir apparaître un produit en "faux cuir" quand il en a demandé un en cuir.

à ne pas séparer	blanc cassé	faux cuir	velours côtelé
qui peuvent être séparés	bleu ciel	rose petunia	bleu marine
difficiles à décider	cuir vernis	daim stretch	mouton retourné

Tableau 1 – Exemples de mots composés

Pour les exemples de la dernière ligne du tableau, il est difficile de décider si "cuir vernis" est assimilable à "cuir" et "daim stretch" à "daim". Pour trancher cette question, nous avons essayé d'insérer des éléments (des adverbes par exemple) entre les mots composés. Ainsi "cuir très vernis"* et "mouton très retourné" ne sont pas attestés en français mais "daim très stretch" est acceptable. Il faudrait donc extraire "cuir vernis" et "mouton retourné" comme une seule valeur, tandis que "daim" tout court sera une matière.

2.1.3. *Problèmes syntaxiques*

Certaines descriptions d'offres ne sont pas syntaxiquement correctes. Dans l'exemple de la figure 4, la première phrase inclut le titre d'origine "Mules d'intérieur". D'autres textes se présentent comme de simples énumérations de propriétés. Nous ne pourrons donc pas procéder à des analyses syntaxiques poussées de nos textes, et devrons compter sur les contextes locaux pour identifier les valeurs des champs.

2.1.4. *Problèmes sémantiques et/ou pragmatiques*

Notre but est d'extraire la (ou les) couleur(s), la (ou les) matière(s) et la marque d'un produit dans les descriptions des offres, soit des chaussures pour notre corpus. Mais les descriptions peuvent aussi évoquer les caractéristiques d'autres objets liés au produit, comme les lacets ou les lanières : s'ils ne sont considérés que comme des

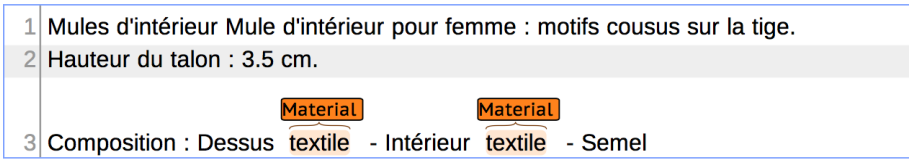

Figure 4 – Exemple de texte syntaxiquement incorrect

accessoires, ils ne devraient pas être extraits. Il est en effet peu probable qu'ils fassent l'objet d'une requête spécifique par le biais du moteur de recherche.

Dans l'exemple suivant, les "clous métal" sont des accessoires, donc "métal" ne doit pas être repéré en tant que matière des chaussures. La "boucle" est aussi un accessoires, et "argent" ne doit pas non plus être extrait, ni comme matière ni comme couleur :

> Escarpins à lanières cuir noir avec clous métal. Boucle argent avec inscription de la marque. Plateau de 2 cm et talon aiguille de 13 cm. Ces chaussures Guess sont livrées dans une boite Guess avec une poche...

Il est parfois difficile de définir et distinguer entre les parties principales et accessoires de chaussures. Une partie sera considérée comme principale si, en son absence, le produit ne peut plus assurer la fonction à laquelle il est destiné. Ainsi, les matières des parties intérieures, extérieures et des semelles sont extraites. Pour les talons, il y a déjà matière à discussion. Une paire de "chaussures avec les talons en bois" ne sont en effet pas des "chaussures en bois". Et si les talons sont retirés, les chaussures restent portables alors que les tongs ou les sandales perdent leur fonctionnalité si on les prive de leurs lanières ou de leurs "tiges". Ce raisonnement justifie d'extraire les parties soulignées dans les exemples suivants :

> Merrell Qesbour Thong Dark Earth Man. Tongs artisanales à semelle plate conçues pour les "missions" de type repos et balade. L-assise plantaire en daim est douce et sa semelle légère.Caractéristiques :TIGE/DOUBLURE-Tige cuir- Dessus d-assise plantaire en...

> Sandales Best Mountain Sandales légères à porter, Bride façon cuir à l'entre-doigt, Ensemble de larges lanières en tissus dessus,

Dans ce dernier exemple, "bride façon cuir" n'est pas du "cuir", il ne faut donc pas extraire "cuir" seul. Dans certains cas extrêmes, la description du produit mentionne une couleur non associée au produit lui-même, comme dans l'exemple qui suit :

Bottes Indiennes grises cloutées Coloris : gris anthracite / doublé fourré. Composition : 100% cuir Conseil + : Des bottes très recherchées cette saison ! Confortables, le cuir est doublé pour avoir bien chaud en tout saison. Le cuir est clouté sur le devant de la jambe d'un motif d'étoile en métal argent et bronze pour un effet so rock'. A porter avec un jean slim ou un leggings, un maxi gilet et un manteau de cachemire bleu ou noir pour un look working glam' ! ! ! ! !

2.2. *Etat de l'art*

La tâche que nous venons de décrire, même si elle a des spécificités (à part les marques, les valeurs à extraire ne sont pas des entités nommées), relève de l'extraction d'information. C'est donc naturellement les techniques usuelles de ce domaine que nous allons utiliser pour la traiter. Nous allons en tester deux : une à base de ressources et de dictionnaires (Poibeau, 2003 ; Ehrmann, 2008), une autre fondée sur l'apprentissage automatique supervisé, en particulier les modèles CRF (Lafferty *et al.*, 2001 ; McCallum et Li, 2003). C'est ce qu'ont fait par exemple (Raymond et Fayolle, 2010) dans deux corpus pour l'extraction d'entités nommées. Dans leur cas, les CRF sont plus performants que les SVM, eux-mêmes plus efficaces que des transducteurs à états finis.

La reconnaissance automatique d'attributs de produits a tout de même fait l'objet de travaux spécifiques : ainsi, dans (Putthividhya et Hu, 2011) les CRF, les SVM, MaxEnt et des HMM sont mis à contribution pour l'extraction de propriétés de produits comme leur marque, leur style (décontracté, sexy, urbain, etc), leur taille et leur(s) couleur(s) dans les titres de descriptions de vêtements et chaussures pour eBay. Dans ce cas aussi, ce sont les CRF qui se comportent le mieux.

3. Constitution de corpus annotés

3.1. *Les trois corpus : sélection, prétraitements et annotation*

Nous avons construit trois corpus différents, tous issus de la catégorie "chaussures femmes". Il ne comprennent aucune offre avec une description vide, mais certaines se répètent. Corpus1 est construit avec des offres ne provenant que d'un seul marchand, tandis que Corpus2 complète Corpus1 avec presque la même quantité d'offres d'autres marchands. Corpus3 est créé à part, avec des offres qui n'ont jamais été sélectionnées précédemment (mais qui pourraient contenir les offres d'un marchand présent dans Corpus2). Le tableau 2 montre les principales propriétés de ces trois corpus et la figure 5 les relations qu'ils entretiennent les uns envers les autres.

Corpus	Nombre d'offres	Nombre de marchands	Nombre d'offres répétées	Nombre de tokens	Taille du vocabulaire
Corpus1	530	1	42	15493	462
Corpus2	1000	16	212	32251	1833
Corpus3	200	32	5	6750	1328

Tableau 2 – Corpus construits

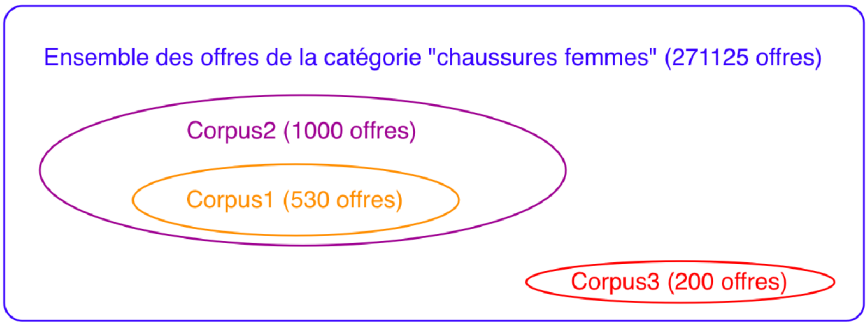

Figure 5 – Relations entre les corpus

Ces trois corpus ont été traités par le tokeniseur de NLTK (*Natural Language Tool-Kit* (Bird, 2006)) qui les réduit à des séquences de tokens (un token est un mot, une ponctuation, un chiffre ou un caractère spécial comme % ou &), mis en minuscule en conservant les accents. Les valeurs des trois champs à extraire ont été annotées à la main sur chacun desdits corpus. Le tableau 3 montre que les trois champs ne sont pas équitablement répartis dans les 1000 textes de Corpus2 (qui sera la principale source de nos expériences d'apprentissage), tandis que le tableau 4 donne le nombre d'entités de chaque type repérées dans chaque corpus, ainsi que la taille du vocabulaire concerné (nombre de tokens distincts). Le vocabulaire utilisé pour l'entité "marque" est le plus riche.

Corpus	marque	couleur	matière	aucune entité
Corpus2	552	165	827	93

Tableau 3 – Nombre d'offres contenant les entités

3.2. *Variabilité des trois corpus*

Les trois corpus sont destinés à étudier dans quelle mesure un système d'apprentissage automatique appris sur des données plus ou moins homogènes s'applique à

Corpus	**Nombre d'entités extraites**			**Taille du vocabulaire**		
	marque	couleur	matière	marque	couleur	matière
Corpus1	215	24	1121	47	15	13
Corpus2	590	757	2009	108	58	62
Corpus3	114	123	385	59	39	36

Tableau 4 – Nombre d'entités annotées et tailles des vocabulaires

d'autres textes, autrement dit s'il parvient à généraliser. Corpus1 ne provenant que d'un seul marchand, c'est le plus homogène stylistiquement.

Les nombres de textes minimum, maximum et moyen issus d'un même marchand dans Corpus2 et Corpus3 figurent dans le tableau 5.

Corpus	**Minimum**	**Maximum**	**Moyenne**	**Ecart type**	**Médianne**
Corpus2	1	530	62,5	128,36	12,5
Corpus3	1	20	6,25	5,93	4,0

Tableau 5 – Nombre d'occurrence de tokens pour Corpus2 et Corpus3

Chaque offre peut contenir de 0 à plusieurs entités annotées. Les nombres minimum, maximum et moyen d'entités par offre dans chacun des trois corpus sont assez similaires, comme le montre le tableau 6.

Corpus	**Minimum**	**Maximum**	**Moyenne**	**Ecart type**	**Médianne**
Corpus1	0	6	2,6	1,35	3
Corpus2	0	12	2,8	1,76	3
Corpus3	0	8	3,11	1,85	3

Tableau 6 – Nombre d'entités annotées par offre

4. Extraction par dictionnaires

4.1. *Constitution de dictionnaires*

La première méthode que nous avons testée sur notre tâche est l'application de dictionnaires (autant que de champs à extraire) construits manuellement. Pour le champ "marque", nous avons utilisé une base de marques (tous produits confondus) disponible chez LeGuide.com. Nous avons construit une liste de couleurs à partir d'informations disponibles sur Internet ainsi qu'une liste de matières. Le tableau 7 donne la taille de chaque liste.

Nom du dictionnaire	Taille du dictionnaire
marque	15391
couleur	596
matière	67

Tableau 7 – Tailles des vocabulaires pour chaque champ à extraire

4.2. *Résultats du reconnaisseur*

Pour chercher toutes les occurrences des entrées d'un dictionnaire dans un texte donné, nous utilisons l'algorithme de recherche de motif d'Aho-Corasick (Crochemore *et al.*, 2007), en raison de son efficatité (complexité linéaire en la taille du texte en entrée). Quelques pré-traitements supplémentaires ont été appliqués : les accents sont supprimés (contrainte d'implémentation) et un caractère espace est ajouté avant et après chaque token (afin d'indiquer le début et la fin du token). Le reconnaisseur est configuré pour extraire la séquence de tokens la plus longue correspondant à une valeur d'une liste (qui inclut des mots composés). Les résultats obtenus avec cette méthode pour chaque champ et chaque corpus sont fournis dans le tableau 8. Notons que, pour chaque offre, nous comparons les résultats du reconnaisseur avec l'intégralité des données annotées (un même champ peut donc prendre plusieurs valeurs) et que l'égalité stricte entre les valeurs des champs est requise. L'évaluation se fait en prenant en compte la localisation de l'entité, sa valeur et son type.

Champ	Corpus	Précision	Rappel	F-Mesure
marque	Corpus1	18,0%	66,5%	28,3%
	Corpus2	26%	74,1%	38,5%
	Corpus3	28,2%	81,6%	41,9%
couleur	Corpus1	100%	83,3%	90,9%
	Corpus2	82,9%	91,7%	87,1%
	Corpus3	89,1%	86,2%	87,6%
matière	Corpus1	98,9%	98,0%	98,4%
	Corpus2	95,6%	89,1%	92,2%
	Corpus3	90,7%	81,3%	85,7%

Tableau 8 – Evaluation de la méthode à base de dictionnaires par entité

4.3. *Analyse des erreurs*

Les entités "couleur" et "matière" apparaissent clairement comme les plus faciles à extraire. La précision et le rappel du reconnaisseur sur le champ "marque" sont en revanche plus faibles. Cela est notamment dû à la non-exhaustivité du dictionnaire de

marques qui, bien que volumineux, n'est pas spécialisé pour la catégorie chaussures. La faiblesse en précision s'explique aussi par le fait que cette liste de marques, une fois mise en minuscules et sans accents, contient beaucoup de mots de la langue courante, entrainant du bruit dans l'extraction. La liste de marques suivante en donne des exemples :

- elle : un magazine
- tous : une marque de bijoux
- plus : se trouve dans la base de marque
- boots : une marque de chaussures
- talon : une marque de chaussures
- look : se trouve dans la base de marque
- chic : se trouve dans la base de marque

Le même phénomène peut concerner aussi, mais à moindre échelle, les autres champs : la couleur "maïs", transformée en "mais" par le retrait des accents, a ainsi entrainé beaucoup d'extractions fautives.

Le deuxième phénomène source d'erreurs que nous avons repéré est celui de l'ambiguité de certains tokens, qui appartiennent à plusieurs listes. C'est le cas des exemples suivants :

- orange : marque d'un opérateur téléphonique ou couleur
- ciel : marque ou couleur
- chrome : couleur ou matière
- bronze : couleur ou matière
- mousse : couleur ou matière
- paille : couleur ou matière

Enfin, la non-prise en compte des contextes entraine des extractions abusives comme celle de "cuir" dans des expressions comme "faux cuir" ou "façon cuir". Un vrai système à base de règles devrait bien sûr prendre en compte ces phénomènes. Mais, plutôt que d'approfondir cette approche, nous avons préféré la mettre en concurrence avec une méthode d'apprentissage automatique.

5. Extraction par CRF

L'alternative aux méthodes à base de règles écrites manuellement (dont nos dictionnaires sont une version minimale) est bien entendu l'utilisation de techniques d'apprentissage automatique. La plus efficace à l'heure actuelle pour la tâche d'extraction d'entités est celle mettant en œuvre les CRF : nous en explorons dans cette section toutes les possibilités.

5.1. *Conditions de nos expériences*

Les CRF sont des modèles graphiques probabilistes non dirigés et discriminants pour la prédiction d'étiquettes pour des données structurées (Lafferty *et al.*, 2001 ; Tellier et Tommasi, 2011). Pour la tâche de détection d'entités dans des données textuelles, les types de CRF employés sont des CRF linéaires (Sha et Pereira, 2003). Ces derniers prédisent une séquence d'étiquettes à partir de la séquence textuelle segmentée (en tokens) en entrée.

Les prédictions d'un CRF sont basées sur la combinaison de fonctions caractéristiques qui modélisent des propriétés plus ou moins locales de la séquence de tokens observée. On distingue deux types de fonctions caractéristiques : les fonctions unigrammes qui prennent en entrée l'observation et l'étiquette du token courant, et les fonctions bigrammes qui prennent en entrée l'observation, l'étiquette du token courant et celle du token précédent. Un des intérêts des CRF linéaires est de pouvoir modéliser des dépendances longues car les fonctions caractéristiques peuvent accéder à toute la séquence d'observation. Cependant en pratique, les fonctions caractéristiques sont souvent utilisées pour modéliser des propriétés locales du texte. Ces fonctions sont générées à partir de patrons, qui définissent une conjonction de tests basiques sur les attributs de la séquence de tokens, et d'une fenêtre d'observation des données, qui contrôle la localité des fonctions caractéristiques.

L'apprentissage d'un CRF linéaire se fait de manière supervisée à partir d'un ensemble de couples (séquence de tokens , séquence d'étiquettes) et consiste à déterminer l'ensemble des poids des fonctions caractéristiques, généralement par maximum de vraisemblance et descente de gradient. Le calcul de la séquence d'étiquettes la plus probable pour une séquence de tokens se fait à l'aide de l'algorithme de Viterbi.

Pour nos expériences avec les CRF, nous avons utilisé le logiciel Wapiti (Lavergne *et al.*, 2010). Le codage classique d'étiquettes BIO pour "Begin/In/Out" (Sarawagi, 2008) est associé aux noms des différents champs à trouver (marque, couleur et matière) pour annoter les textes, ce qui fait un total de 7 étiquettes distinctes (B et I associés à chacun des trois champs, plus O). Nous avons en outre essayé d'exploiter différents attributs, soit internes (propriétés des textes lisibles dans les données) soit issus de ressources linguistiques externes (listes, étiqueteur morpho-syntaxique). La liste de ces attributs est la suivante :

– la valeur du token en minuscule et/ou sans accent ;

– la présence de majuscule(s) en début, au milieu ou partout dans le token (attributs booléens) ;

– le type de token (uniquement des lettres, présence de chiffres, de ponctuation ou de caractères spéciaux) ;

– la longueur en nombre de caractères du token ;

– la position dans le texte du token : chaque texte est découpé en 6 blocs en fonction du nombre total de tokens, la position d'un token est un nombre entre 0 et 5 correspondant à l'index du bloc où il se trouve ;

– la présence du token (en tant que B ou I) dans un de nos dictionnaires d'entités (celui de marque, de matière ou de couleur) ;

– la racine du token (obtenue avec NLTK (Bird, 2006)) ;

– l'étiquette morpho-syntaxique prédite par SEM, un analyseur morpho-syntaxique à base de CRF (Tellier *et al.*, 2012).

Cette liste d'attributs, combinée au choix de la fenêtre d'observation sur les données, donne une combinatoire très vaste de patrons possibles pour définir les fonctions caractéristiques du CRF. Nous avons exploré cet espace de façon progressive et systématique. Les Corpus 1 à 3 seront par ailleurs mis à contribution pour évaluer la capacité de généralisation des modèles sur des données plus ou moins homogènes. Nos expériences visent bien sûr à obtenir des résultats si possible au moins aussi bons que ceux atteints par les dictionnaires, mais elles poursuivent également différents autres objectifs :

– trouver les meilleurs attributs et les meilleurs patrons pour l'apprentissage d'un CRF sur cette tâche ;

– déterminer s'il vaut mieux apprendre un modèle distinct pour chaque champ ou si un unique modèle est capable d'extraire efficacement tous les champs en même temps ;

– déterminer s'il vaut mieux utiliser un corpus d'apprentissage très spécialisé (constitué par exemple d'un seul marchand, comme Corpus1) ou plus varié.

5.2. *Résultats des expériences*

5.2.1. *Baseline*

A titre de baseline, nous avons défini un patron bigramme "minimal" ne faisant appel qu'à la valeur du token courant (sans exploiter d'autres attributs). Un seul modèle est appris pour reconnaître simultanément les trois types de champs. Ses résultats sur Corpus2 en validation croisée à 5 plis sont donnés dans le tableau 9.

Corpus	**Champ**	**Précision**	**Rappel**	**F-mesure**
Corpus2	marque	86,18%	80,84%	83,41%
	couleur	76,27%	63,71%	69,22%
	matière	85,13%	85,02%	85,07%

Tableau 9 – Résultat de la baseline sur Corpus2

Nous constatons d'ores et déjà que les marques sont nettement mieux traitées par cette approche, mais les performances sur les autres champs sont en revanche inférieures à celles obtenues par les dictionnaires. Les différences de performance entre champs peuvent s'expliquer par leurs taux de présence dans les corpus : d'après le ta-

bleau 3, "couleur" est en effet moins présent que "marque", lui-même moins fréquent que "matière", dans les données de Corpus2.

5.2.2. *Recherche des meilleurs patrons*

Nous cherchons maintenant à améliorer ces résultats en exploitant des propriétés plus fines dans les patrons définissant les fonctions caractéristiques des CRF. Les choix à faire sont la nature de l'attribut testé (parmi les 10 listés précédemment) et la taille de la fenêtre (nombre impair compris entre 1 et 33) prises en compte sur les données d'observation. Pour trouver la meilleure fenêtre adaptée à chaque attribut, nous avons tout d'abord testé indépendamment les 170 patrons bigrammes possibles, utilisés seuls, en validation croisée à 5 plis avec Corpus2, pour chacun des trois champs à extraire. Cette première série d'expériences sert à sélectionner, pour chaque attribut, le patron qui mène à la meilleure précision, et par ailleurs tous ceux qui permettent d'atteindre un meilleur rappel que celui-ci. Quand plus de deux patrons différents pour un même attribut étaient conservés, nous avons procédé à une sélection supplémentaire consistant à ne garder que les meilleurs en F-mesure. Une fois K patrons portant chacun sur une seule propriété choisis, nous devons chercher la meilleure façon de les utiliser conjointement pour produire les fonctions caractéristiques du CRF. Nous avons le choix de garder entre 1 et K d'entre eux conjointement, ce qui fait une combinatoire de $C_K^2 + C_K^3 + ... + C_K^K$ ensemble de patrons possibles. Pour le champ "marque" nous avons ainsi testé 120 combinaisons de patrons possibles, 502 pour les couleurs et 2036 pour les matières.

Le tableau 10 montre les résultats des deux meilleures combinaisons de patrons trouvées (la meilleure en précision et la meilleure en F-mesure) pour chaque champ.

Champ	**Modèle**	**Précision**	**Rappel**	**F-mesure**
marque (120 modèles testés)	baseline	86,18%	80,84%	83,41
	meilleur en précision	**88,9%**	81,6%	85,06%
	meilleur en F-mesure	88,83%	83,43%	**86,03%**
couleur (502 modèles testés)	baseline	76,27%	63,71%	69,22%
	meilleur en précision	**86,71%**	71,02%	77,91%
	meilleur en F-mesure	85,59%	74,43%	**79,47%**
matière (2036 modèles testés)	baseline	85,13%	85,02%	85,07%
	meilleur en précision	**86,51%**	84,67%	85,57%
	meilleur en F-mesure	85,51%	86,31%	**85,9%**

Tableau 10 – Meilleures combinaisons de patrons

Nous constatons que les meilleurs modèles pour "couleur" augmentent significativement les scores de la baseline, tandis que l'extraction des champs "marque" et "matière" progresse plus difficilement de 1% à 3%. Les meilleurs patrons sélectionnés pour les trois champs contiennent des propriétés différentes, mais leur point commun est une taille de fenêtre presque toujours comprise entre 3 et 7. Nous avons aussi procédé à des expériences similaires pour les patrons unigrammes (que nous ne détaillons

pas ici) : ils donnent des résultats souvent très proches tout en prenant moins de temps de calcul.

5.2.3. *Autres expériences*

Comme nous l'avons vu en section 3.1, Corpus1 contient des offres en provenance d'un seul et unique marchand : est-ce avantageux de créer un modèle par marchand ou une certaine diversité est-elle préférable ? Pour trancher cette question, nous avons fait des expériences où seul Corpus1 était utilisé, en validation croisée. Les résultats obtenus étaient moins bons que ceux de la baseline sur Corpus1, il ne semble donc pas avantageux de sur-spécialiser les modèles CRF marchand par marchand.

Nous avons également testé les meilleures combinaisons de patrons précédemment affinées sur Corpus2 sur Corpus3, pour vérifier que nos combinatoires ne sont pas trop spécifiques de l'ensemble d'apprentissage, et nous avons comparé systématiquement sur ce corpus les modèles spécialisés sur un seul champ et ceux cherchant à extraire tous les champs en même temps. Les meilleurs résultats obtenus pour la F-mesure figurent dans le tableau 11.

	Précision	**Rappel**	**F-mesure**
meilleurs résultats sur marque	91,43%	26,45%	41,03%
meilleurs résultats sur couleur	84,21%	65,04%	73,39%
meilleurs résultats sur matière	86,32%	62,60%	72,57%

Tableau 11 – Meilleurs résultats obtenus sur Corpus3

Ces meilleurs résultats ont tous été obtenus à l'aide de modèles (pas les mêmes sur chaque ligne) qui extraient simultanément tous les champs. Si leurs précisions sont très proches de celles obtenues en validation croisée sur Corpus2 (et meilleures que celle de la baseline), les scores de rappel sont en revanche dégradés. Cela correspond sans doute à un effet de sur-apprentissage dû à la procédure de choix des patrons : les modèles n'ont pas réussi à bien généraliser sur un corpus nouveau.

6. Conclusion

Dans cet article, nous nous sommes confrontés à une tâche d'extraction de caractéristiques de produits dans des textes de sites marchands. Cette tâche, potentiellement très utile, relève de l'extraction d'information, mais elle présente des caractéristiques spécifiques : la multiplicité des données à extraire est variable d'un texte à un autre, et toutes ne sont pas des entités nommées. Les textes sont écrits avec des styles très variés, pouvant aller de la simple énumération de propriétés à des descriptions riches et élaborées, évoquant même d'autres produits dont les propriétés ne devraient pas être prises en compte dans l'extraction.

Nous avons abordé cette tâche avec les deux familles d'outils traditionnels utilisés en extraction d'information : les règles écrites manuellement (réduites ici à des

dictionnaires) et l'apprentissage automatique avec des CRF. Les dictionnaires apparaissent particulièrement performants pour l'extraction de propriétés dont les valeurs peuvent être facilement listées (les couleurs et les matières), même si le traitement de tokens ambigus requiert certainement la prise en compte de contextes plus riches que ceux que nous avons utilisés, ainsi que des prétraitements pour éliminer les confusions possibles entre les champs (*i.e.* pas d'intersection entre les différents dictionnaires) ou avec des mots courants. Pour la reconnaissance des marques, en revanche, qui sont des entités nommées, la solution de l'apprentissage automatique semble la plus prometteuse. Les marques sont en effet les propriétés les plus évolutives et les plus spécifiques du type de produit décrit. Notre dictionnaire de marques qui, malgré sa taille, n'est ni spécialisé ni exhaustif, a favorisé le rappel mais s'est avéré très mauvais en précision. C'est au contraire sur ce champ que nos modèles CRF ont fourni leurs meilleurs résultats. Notons d'ailleurs que les dictionnaires utilisés comme règles étaient intégrés aux attributs des CRF appris, c'est pourtant seulement dans le cas des marques que les modèles en question ont dépassé la capacité d'extraction de ces dictionnaires. Les CRF n'ont pas réussi à tirer parti de cet attribut en conjonction avec les autres, comme on aurait pu l'espérer.

Cette première exploration suggère à l'avenir une combinaison de méthodes : pour certains champs "faciles à circonscrire" et relativement indépendants des produits décrits, les listes et les règles peuvent s'avérer suffisantes. Mais l'apprentissage automatique devient nécessaire pour ceux qui présentent une plus grande variabilité ou une plus grande spécificité. Il reste évidemment à valider ces intuitions sur des données nouvelles.

Ces travaux mettent aussi en évidence une difficulté d'utilisation des CRF : quelle stratégie employer pour sélectionner les caractéristiques produisant le meilleur modèle ? Peu de travaux abordent ce point, l'approche couramment utilisée étant de déléguer cette sélection à l'algorithme d'apprentissage des CRF, qui va assigner un poids nul aux fonctions caractéristiques non pertinentes (McCallum, 2003). Or nos expériences ont montré que cette approche n'est pas suffisante et qu'une exploration plus systématique des combinaisons possibles des fonctions caractéristiques permet d'obtenir de meilleurs résultats. Encore une fois, cette constatation empirique se doit d'être validée sur d'autres données et analysée sur le plan théorique.

7. Bibliographie

Bird S., « NLTK : the natural language toolkit », *COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, ACL, Stroudsburg, PA, USA, p. 69-72, 2006.

Crochemore M., Hancart C., Lecroq T., *Algorithms on strings*, Cambridge University Press, 2007.

Ehrmann M., Les Entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguisation, PhD thesis, Université Paris 7 – Denis Diderot, Juin, 2008.

Ghani R., Probst K., Liu Y., Krema M., Fano A., « Text mining for product attribute extraction », *SIGKDD Explor. Newsl.*, vol. 8, n° 1, p. 41-48, June, 2006.

Lafferty J., McCallum A., Pereira F., « Conditional Random Fields : Probabilistic models for segmenting and labeling sequence data », *Proceedings of ICML 2001*, p. 282-289, 2001.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *ACL*, ACL, p. 504-513, July, 2010.

McCallum A., « Efficiently inducing features of Conditional Random Fields », *Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03)*, 2003.

McCallum A., Li W., « Early results for named entity recognition with Conditional Random Fields », *CoNLL'2003*, 2003.

Poibeau T., *Extraction automatique d'information*, Hermés, Paris, 2003.

Putthividhya D. P., Hu J., « Bootstrapped named entity recognition for product attribute extraction », *EMNLP '11*, ACL, Stroudsburg, PA, USA, p. 1557-1567, 2011.

Raymond C., Fayolle J., « Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement », *TALN'10*, ATALA, Montréal, Québec, Canada, July, 2010.

Sarawagi S., « Information extraction », *Foundations and trends in databases*, vol. 1, n° 3, p. 261-377, 2008.

Sha F., Pereira F., « Shallow parsing with Conditional Random Fields », *HLT-NAACL 2003*, p. 213 - 220, 2003.

Tellier I., Dupont Y., Courmet A., « Un segmenteur-étiqueteur et un chunker pour le français », *TALN 2012, session démo*, 2012.

Tellier I., Tommasi M., « Champs Markoviens Conditionnels pour l'extraction d'information », *in* Eric Gaussier, François Yvon (eds), *Modèles probabilistes pour l'accès à l'information textuelle*, Hermès, 2011.