Construire un lexique de sentiments par crowdsourcing et propagation

Mathieu Lafourcade, Nathalie Le Brun, Alain Joubert

Construire un lexique de sentiments par crowdsourcing et propagation

Mathieu Lafourcade[1], Nathalie Le Brun[2], Alain Joubert[1]
(1) Lirmm, Université Montpellier, France
(2) Imagin@t, 34400 Lunel, France
mathieu.lafourcade@lirmm.fr, imaginat@imaginat.name, alain.joubert@ lirmm.fr

RÉSUMÉ

Cet article présente une méthode de construction d'une ressource lexicale de sentiments/émotions. Son originalité est d'associer le crowdsourcing via un GWAP (Game With A Purpose) à un algorithme de propagation, les deux ayant pour support et source de données le réseau lexical JeuxDeMots. Nous décrivons le jeu permettant de collecter des informations de sentiments, ainsi que les principes et hypothèses qui sous-tendent le fonctionnement de l'algorithme qui les propage au sein du réseau. Enfin, nous donnons les résultats quantitatifs et expliquons les méthodes d'évaluation qualitative des données obtenues, à la fois par le jeu et par la propagation par l'algorithme. Ces méthodes incluent une comparaison avec Emolex, une autre ressource de sentiments/émotions.

ABSTRACT

Building a sentiment lexicon through crowdsourcing and spreading

This paper describes a method for building a sentiment lexicon. Its originality is to combine crowdsourcing via a Game With A Purpose (GWAP) with automated propagation of sentiments via a spreading algorithm, both using the lexical JeuxDeMots network as data source and substratum. We present the game designed to collect sentiment data, and the principles and assumptions underlying the action of the algorithm that propagates them within the network. Finally, we give quantitative results and explain the methods for qualitative evaluation of the data obtained for both the game and the spreading done by the algorithm, these methods including a comparison with Emolex, another resource sentiment/emotions.

MOTS-CLÉS : sentiments, crowdsourcing, GWAP, réseau lexical, propagation.
KEYWORDS: sentiments, crowdsourcing, GWAP, lexical network, spreading.

1 Introduction

La caractérisation automatique des sentiments présents dans les textes est devenue un enjeu majeur pour des applications telles que l'analyse de discours politiques, ou d'opinions relatives à la fourniture de services touristiques, culturels, ou de biens de grande consommation. La constitution d'une ressource lexicale de sentiments (associer à un terme un ensemble pondéré de sentiments ou d'émotions) est un préalable à ce type de recherche, que les approches pour sa construction soient statistiques supervisées, ou plus linguistiques (Brun, 2011).

Les valeurs de sentiments peuvent être exprimées à partir d'un ensemble fermé prédéterminé (Taboada *et al*., 2011) ou sur un ensemble ouvert, éventuellement moins précis mais potentiellement plus riche, pouvant rendre compte de la variété du vocabulaire des émotions (Whissell, 1989). Parmi les ressources existantes, Saif et Turney (2010, 2013) ont utilisé un ensemble de sept sentiments (confiance, peur, tristesse, colère, surprise, dégoût, et joie) pour leur ressource de polarité/sentiment pour l'anglais (EmoLex), obtenue par crowdsourcing à l'aide d'Amazon Mechanical Turk (ce qui peut poser problème, voir (Fort *et al*., 2014)). Chaque terme de leur lexique (environ 14000 termes) est lié à 0, 1, ou plusieurs des 7 sentiments proposés. La valeur "indifférence" sera affectée à un terme qui n'est relié à aucun de ces 7 sentiments/émotions. Une adaptation de cette ressource a été réalisée pour le français (Abdaoui, *et al*., 2014). Esuli et Sebastiani (2006) ont produit une ressource libre basée sur WordNet (Fellbaum, 1998) sous la forme d'un lexique associant à chaque *synset* trois valeurs de polarité (positif, négatif et neutre). Il ne s'agit pas à proprement parler d'un lexique de sentiments, même si dans la terminologie utilisée, on rencontre souvent une confusion entre polarité et sentiments.

La détection de sentiments et d'opinions (qui relèvent en fait souvent de la polarité) est l'objet de beaucoup d'efforts : ces données sont le plus souvent identifiées et extraites automatiquement depuis des corpus, voir (Kim et Hovy, 2006), (Mudinas *et al*., 2012), (Strapparava et Mihalcea, 2008) et (Kiritchenkos, *et al*., 2014). Nous proposons ici une autre démarche basée sur l'idée de propagation de valeurs de sentiment à travers une structure de réseau lexical (et non pas un corpus de textes). Une approche voisine a été proposée par Kamps *et al*. (2004) pour l'anglais avec WordNet via les relations de synonymie / antonymie ; elle se focalise sur la polarité (bon, mauvais), la potentialité (fort, faible) et l'activité (passif, actif).

Un réseau lexical, tel celui obtenu grâce au jeu en ligne (GWAP) JeuxDeMots (Lafourcade, 2007), comporte des termes associés par des relations lexico-sémantiques. Le projet JeuxDeMots (JDM) a permis non seulement la constitution d'un réseau lexical en perpétuelle extension et libre d'accès, mais également la validation/vérification des relations qui le constituent via un certain nombre de jeux et contre-jeux (Lafourcade *et al*., 2015b). Ainsi, une telle ressource est particulièrement propice à l'expérimentation de méthodes d'acquisition de données de sentiments/émotions. L'hypothèse que nous posons dans cet article est la suivante : pour construire une ressource de sentiments/émotions, il est intéressant de combiner l'approche par des jeux (où les données sont fournies par des locuteurs) avec l'utilisation d'un algorithme de propagation, afin d'affecter automatiquement à un grand nombre de termes du réseau des données de sentiments/émotions. Cette affectation se réalise par propagation des sentiments proposés par les joueurs/contributeurs vers les termes non renseignés. Elle repose sur l'idée implicite que les contributeurs fournissent des données de bonne qualité (Simperl, 2013).

Dans cet article, nous commencerons en section 2 par présenter les principes de l'acquisition de données de sentiments/émotions dans un cadre contributif ludique au sein d'un réseau lexical, à

l'aide du jeu *Emot*. Nous présenterons ensuite *Botemot*, un algorithme de diffusion des données de sentiments à travers le réseau. En section 3, nous indiquerons enfin les résultats obtenus via une analyse quantitative et qualitative. La méthodologie d'évaluation qualitative se fonde sur la comparaison entre les sentiments proposés par *Botemot* et ceux déjà présents dans le réseau lexical, et donc considérés comme valides. Nous conclurons sur l'intérêt de combiner diverses approches pour acquérir des données lexicales de qualité.

2 Crowdsourcing et réseau lexical

Le projet JeuxDeMots (Lafourcade, 2007) rassemble une collection de jeux et d'interfaces contributives qui concourent à la constitution d'un réseau lexico-sémantique de grande taille pour le français. Initié en 2007, à partir d'une base de 150 000 termes sans aucune relation entre eux, le réseau compte aujourd'hui environ 850 000 termes reliés par plus de 33 millions de relations lexicales et sémantiques. Notons que son extension est le produit combiné de l'activité des joueurs et de celle d'algorithmes, qui le parcourent en permanence en proposant de nouvelles relations, inférées à partir de celles existantes. Du fait de sa richesse, une telle structure est particulièrement appropriée pour tester de nouvelles combinaisons entre crowdsourcing et algorithmique (sous la forme de robots d'inférence) en vue d'établir de nouvelles relations. Ce projet a démontré que dans son contexte d'application, la création d'un réseau associatif généraliste, le crowdsourcing via des jeux était efficace, que ce soit qualitativement ou quantitativement (Lafourcade *et al*., 2015).

2.1 Structure de réseau lexical

Un réseau lexical est une structure de graphe où les mots représentent des objets lexicaux et les arcs des relations entre ces objets. Des structures comme Wordnet (Miller, 1995 et Fellbaum, 1998), BabelNet (Navigli, 2010) ou HowNet (Dong, 2007), qui sont bâties sur ce modèle, peuvent également être considérées comme des réseaux lexicaux. Dans l'ensemble, ces réseaux ont été construits manuellement, en faisant toutefois appel à des outils de vérification de cohérence. A notre connaissance, hormis JeuxDeMots, aucun projet de construction d'un réseau lexical de grande taille n'a impliqué une communauté de joueurs volontaires (Lafourcade *et al*., 2015b). Le réseau lexical JeuxDeMots définit environ une centaine de types de relations lexico-sémantiques binaires pouvant relier deux termes. Ces relations sont orientées et pondérées. Une relation affectée d'un poids négatif est considérée comme fausse. Par exemple :

voler : agent/-25 : autruche

Une telle relation négative peut être considérée comme inhibitrice (à l'image des réseaux de neurones) car elle ne laissera pas passer l'information à son voisinage lors de processus de propagation. De façon similaire, pour des systèmes d'inférence, une relation négative sera considérée comme logiquement fausse, et ce même si son terme plus général (le générique dans une ontologie) vérifie la relation (voler : agent/>0 : oiseau).

En janvier 2016, 824 434 termes étaient présents dans le réseau, et reliés par plus de 33 millions de relations lexicales et sémantiques (dont 256 608 relations inhibitrices). Plus de 12 000 termes polysémiques étaient raffinés en 37 146 sens et usages. Ce réseau est en construction permanente via des jeux, des activités de crowdsourcing et des processus d'inférence et de vérification de cohérence. La structure de graphe offre naturellement des stratégies d'inférence de nouvelles relations par diffusion. La diffusion consiste à propager des informations à travers le réseau à partir

de nœuds émetteurs et d'observer comment ces informations sont attribuées pour les termes voisins dans le réseau. Dans cet article, une hypothèse de travail est que les sentiments associés aux termes reliés au mot-cible vérifient globalement une forme de transitivité. Ceci nous permet donc d'inférer automatiquement les sentiments associés à un terme, pour peu qu'il soit suffisamment renseigné lui-même (en nombre de voisins) et que ses voisins le soient également (en termes de sentiments/émotions).

2.2 Emot, un jeu de capture d'émotions

Le jeu Emot donne la possibilité au joueur d'associer une émotion ou un sentiment à un terme proposé. Le joueur peut choisir parmi une vingtaine d'émotions (amour, joie, tristesse, peur …), comme le montre la figure *1a* qui reproduit un écran typique. Le joueur a néanmoins la possibilité de saisir une émotion différente si aucune de celles proposées ne lui convient, ou s'il préfère préciser sa pensée. Il peut également passer, en particulier s'il ne connaît pas le mot. Emot est à ce titre un jeu semi-ouvert, selon la définition de Lafourcade *et al*. (2015a).

L'un des intérêts du jeu réside dans la comparaison de ses propres réponses avec celles déjà fournies par les autres joueurs (voir figure 1b). La possibilité de choisir entre termes faciles (les termes courants) et difficiles (des termes plus rarement utilisés), en offrant deux niveaux de jeu, génère un intérêt supplémentaire. Généralement, les joueurs commencent avec le niveau facile pour se familiariser avec le jeu, puis passent au niveau difficile, souvent jugé plus intéressant, donc plus motivant. La sélection par la mécanique du jeu d'un terme à proposer à un joueur se fait par tirage pseudo-aléatoire. Il peut s'agir, de façon équiprobable, 1) d'un terme disposant déjà d'au moins une relation de sentiment, ou 2) d'un terme voisin d'un terme ayant au moins une relation de sentiment.

FIGURES 1a et 1b : Exemple d'une partie d'Emot. Le joueur est invité à associer au terme "crise de larmes" une émotion ou un sentiment, soit en cliquant sur l'un des smileys proposés, soit en saisissant des mots dans la zone de texte. La réponse déclenche l'affichage du nombre de points gagnés (aspect jeu), ainsi que des réponses précédemment fournies par les autres joueurs (intérêt linguistique).

Les données ainsi fournies par les joueurs ont régulièrement et systématiquement été évaluées manuellement : il fallait s'assurer de leur validité avant d'en faire des valeurs de référence

susceptibles d'alimenter notre algorithme de diffusion. Sur 1500 contributions évalués manuellement par 4 locuteurs natifs du français, nous avons pu établir que 90% étaient parfaitement pertinentes, 9% discutables, et que seulement 1% des contributions étaient à rejeter, car inadéquates. Les contributions discutables relèvent quasi systématiquement de points de vue minoritaires mais néanmoins possibles (ex : *dégoût* associé à *saumon*). Les cas de rejet correspondent manifestement à des erreurs de sélection ou à du trollage (*timidité* associé à *pomme de terre*). L'ensemble des données de sentiments fournies par les joueurs via le jeu Emot est librement accessible ici : `http://www.jeuxdemots.org/emot.php?action=help`.

2.3 Algorithme d'inférence de relations

Botemot est un algorithme de propagation d'émotions dans le réseau lexical. Chaque terme se voit affecter les sentiments associés à ses voisins : pour chaque terme sélectionné (voir critères de sélection plus loin), Botemot va proposer tout ou partie des sentiments associés à ses voisins proches, c'est-à-dire aux termes qui lui sont liés par certaines relations. Les termes associés négativement sont ignorés, par contre un terme associé positivement mais doté de sentiments de poids négatif est normalement retenu, les poids négatifs étant alors soustraits de la somme des poids. L'exécution de cet algorithme se fait en boucle de façon continue (Never Ended Learning), conjointement à l'activité des joueurs.

Algorithme général

Dans le but d'associer des sentiments à chaque terme du réseau, nous appliquons la procédure suivante :

Algorithme **Botemot**

Entrées : un terme *T* dont on doit inférer les sentiments associés.
R le réseau lexical.
Sortie : *L* la liste pondérée des sentiments à inférer.

On initialise *L* à liste vide.

Le terme *T* est filtré
- en amont en fonction de ses polarités :
  si indifférence (polarités négative et positive < 25%),
  on arrête et on retourne la liste vide
- en fonction de son niveau de renseignement :
  si niveau trop faible,
  on arrête et on retourne la liste vide

Soit *E* l'ensemble pondéré des termes auquel *T* est relié dans *R*.

Pour chaque terme *t* de *E*, soit *subL* la liste pondérée des sentiments auquel il est relié dans *R* :
  $L = L + subL$

Filtrage aval de *L* en fonction du facteur de tolérance
retourner *L*

Ligne 1 : Nous restreignons notre expérience aux termes qui sont des noms (communs ou propres), des verbes, des adjectifs ou des adverbes. Le terme T est nécessairement voisin, dans le réseau lexical, d'au moins un terme pourvu d'un sentiment ;

Ligne 2 : dans nos expériences, *R* est le réseau lexical du projet JeuxDeMots ;

Ligne 10 : il s'agit de l'union pondérée des deux listes *L* et *subL*. Un sentiment apparaissant *n* fois avec une valeur moyenne de *p* aura un score de *n*p*. Un sentiment présent dans *subL* peut avoir un poids négatif, si ce sentiment était un sentiment rejeté (à poids négatif donc) d'un des termes de *E* ;

Ligne 7 : Les types de relations retenus pour inférer les sentiments sont les suivants:

- idées associées ;
- hyperonymes ;
- caractéristiques ;
- synonymes ;
- raffinements sémantiques (sens possibles pour un terme polysémique) ;
- conséquences.

L'hypothèse de travail est la suivante : dès lors que deux termes sont liés par une de ces 6 relations, leurs sentiments associés sont globalement transmissibles. Par exemple, si un terme a une *conséquence* associée à un sentiment néfaste, il est probable que ce terme puisse lui-même être associé à ce sentiment. Ainsi, le schéma général suivant peut souvent être vérifié :

Si T : *conséquence* : C et C : *sentiments/émotions* : S
alors T : *sentiments/émotions* : S

Dont un exemple particulier pourrait être :

Si tumeur : *conséquence* : mort et mort : *sentiments/émotions* : peur
alors tumeur : *sentiments/émotions* : peur

Signalons, que dans le cas de deux termes reliés par la relation « idées associées », la propagation des sentiments peut s'avérer plus délicate de par le caractère général de ce type de relation.

L'union pondérée des listes est l'union ensembliste des éléments des listes avec somme des poids des éléments communs. L'algorithme *Botemot* est appliqué tour à tour à chacun des (850 000) termes du réseau lexical, et s'inscrit dans une boucle d'apprentissage permanent (Never Ended Learning). Pour un terme T donné, il peut retourner un ensemble vide, dans les cas suivants :

- le terme T n'a pas de polarité marquée (i.e >25%) ;
- le terme T n'a pas de termes liés pour les 6 types de relations considérés ;
- le terme T a des termes liés mais aucun d'eux n'a de sentiment associé.

La polarité, telle que définie dans le projet JeuxDeMots a été présentée dans (Lafourcade *et al*., 2015a) et nous l'exploitons comme filtre dans notre algorithme (cf. ci-après). La ressource JDM librement disponible contient les informations de polarité.

L'effet global de l'application itérée est une diffusion des sentiments selon la topologie du réseau. Cette diffusion se fait conjointement à l'activité des joueurs dont les contributions font office de valeurs de référence.

Filtrages

Le **filtrage par seuil** consiste à déterminer la partie la plus pertinente de la liste L de sentiments calculés afin d'en améliorer la précision (au détriment du *rappel*, cf. *évaluation qualitative*). Nous rappelons que les termes de L sont pondérés. Nous calculons la moyenne *mu* des pondérations de l'ensemble L. Soit *alpha* un facteur de tolérance, défini sur R+. Nous retenons les termes de L dont le poids est supérieur au seuil *mu* * *alpha*. Plus *alpha* est grand, plus le filtrage est strict. Par exemple, supposons l'ensemble suivant :

{peur:110, excitation 50, joie:30, angoisse:10}

La moyenne est de (110+50+30+10)/4 = 200/4=50. Avec alpha = 2, nous retenons les termes dont le poids est supérieur ou égal à 100, c'est-à-dire l'ensemble {peur:110}. Avec Alpha = 0.5, nous retenons les termes dont le poids est supérieur ou égal à 25, c'est-à-dire l'ensemble {peur:110, excitation 50, joie:30}.

En faisant baisser le seuil d'acceptation, un filtrage tolérant va augmenter la proportion de propositions à faible poids. Augmenter le rappel fera croître également le taux d'erreur, et baisser la précision.

Nous utilisons un deuxième type de **filtrage, par polarité,** où nous exploitons les polarités associées aux termes dans le réseau lexical JDM. En effet, nous pouvons, pour un terme à forte polarisation négative (et/ou positive) ne retenir que les sentiments polarisés dans le même sens. Dans nos expérimentations, nous avons limité l'action de l'algorithme aux termes affectés d'une polarité positive et/ou négative supérieure à 25%. (Notons que du fait de la polysémie ou du point de vue, de nombreux termes peuvent avoir une double polarité, positive et négative. La polarité neutre est ignorée, dans la mesure où elle ne donne généralement pas lieu à des sentiments autres que l'indifférence.)

En amont, nous effectuons un **filtrage par niveau de renseignement**, afin d'éviter les termes ayant trop peu de relations associées. En pratique, les termes qui ont un niveau de renseignement inférieur à 1000 (c'est-à-dire, ceux dont la somme des poids des relations extraites est inférieure à 1000) ne sont pas sélectionnés.

3 Evaluations

Nous présentons dans un premier temps une analyse quantitative suivie de plusieurs évaluations qualitatives.

Analyse quantitative

Concernant les sentiments, le réseau lexical JDM modifié par notre expérience contient :

- 112 643 termes associés à au moins une relation de sentiment, incluant l'indifférence ;

- 110 671 termes associés à au moins une relation de sentiment, l'indifférence exclue ;
- répartis selon 566 298 relations d'émotions/sentiments ;
- soit une moyenne d'environ 5 sentiments par terme.

Botemot a proposé un total de 972 467 sentiments pour 154 099 mots, dont environ 45 000 (soit 30%) n'ont aucun sentiment associé validé. Environ, 620 000 sont des propositions originales (non existantes dans le réseau) et 350 000 sont des propositions déjà présentes avant notre expérience.

L'hypothèse est la suivante : si une proposition de *Botemot* figure déjà parmi les sentiments/émotions validés, alors cette contribution est correcte. Ce faisant, par extrapolation nous pouvons avoir confiance dans les contributions de *Botemot* pour les termes n'ayant aucune relation de sentiments validée. Nous rappelons que bien évidemment *Botemot* ignore les sentiments déjà présents et validés du terme sur lequel il essaye de contribuer.

En moyenne, *Botemot* parvient à sélectionner environ 19 termes pourvus de sentiments associés pour 1 terme qui ne suscite que l'indifférence, ce qui représente un bruit d'environ 5% (ratio 1/19). Ainsi, nous pouvons considérer l'algorithme comme relativement efficace dans la détection de termes auxquels l'association de sentiments (hors indifférence) est pertinente.

Evaluation qualitative

1) Par comptage et par poids

Pour évaluer les performances de *Botemot***,** nous calculons *rappel* et *précision*. Soient les définitions suivantes (nous rappelons ici qu'un sentiment *validé* est une contribution de sentiment issue du crowdsourcing, dont la pertinence a été vérifiée) :

rappel = nombre de propositions déjà présentes dans les sentiments validés / nombre total de sentiments validés ;
précision = nombre de propositions déjà présentes dans les sentiments validés / nombre total de propositions ;
F1-score = moyenne harmonique de la précision et du rappel : (2 * P * R) / (P + R) .

Une bonne proposition est un sentiment déjà validé (poids positif). Une mauvaise proposition est un sentiment invalidé (poids négatif). Une proposition nouvelle est un terme qui ne figure pas en tant que sentiment valide ou invalide pour le mot cible. Plus un terme est renseigné et est connecté à des termes eux-mêmes correctement renseignés, plus les sentiments inférés sont justes. Cette corrélation semble relativement logique.

En moyenne, nous obtenons une précision par comptage de 0.93, un rappel de 0.98 et un score F1 de 0.95 (pour environ 350 000 propositions de sentiments). Les chiffres de l'évaluation par comptage sont présentés dans le graphique 2. L'évaluation ci-dessus n'est basée que sur le comptage (c), c'est-à-dire la présence ou l'absence de mots, et ne tient aucun compte de la pondération des sentiments associés. Or, il semble pertinent d'estimer que pour l'algorithme, retrouver un sentiment déjà présent et fortement lié est plus performant que retrouver un sentiment plus faiblement lié. Nous avons donc également calculé précision (w) et rappel (w) et F1-score (w) en additionnant les poids des termes plutôt que leurs nombres, ceci en affectant aux propositions justes leur poids dans le réseau, et aux propositions nouvelles un poids de 25.

La *figure 2* présente les moyennes de précision, rappel et F1-score, par comptage et par poids, en fonction du niveau de renseignement des termes. La valeur en abscisse (le niveau de renseignement) est donc la somme des poids des relations déjà existantes pour ce terme pour les types de relations exploités par *Botemot*. On doit lire la courbe de gauche à droite, par exemple "la mesure F1 en comptage vaut en moyenne environ 0.95 pour les termes dont le niveau de renseignement est inférieur ou égal à 20000."

Nous observons que la précision en poids est systématiquement supérieure à celle en comptage de 3 à 4%. En effet, la prise en compte des poids permet une modulation beaucoup plus fine dans le calcul que le simple comptage du nombre de sentiments. Pour la même raison, la mesure F1 en poids est supérieure à celle en comptage. Les mesures de F1-score, qui se situent entre 0.94 et 0.98, valident notre hypothèse de départ : il est possible d'inférer des sentiments pertinents par propagation à partir de ceux déjà existants. Nous pouvons donc, par extrapolation, considérer comme valides les sentiments inférés de cette manière à un terme qui n'avait aucun sentiment validé, avec un niveau de confiance de plus de 94%.

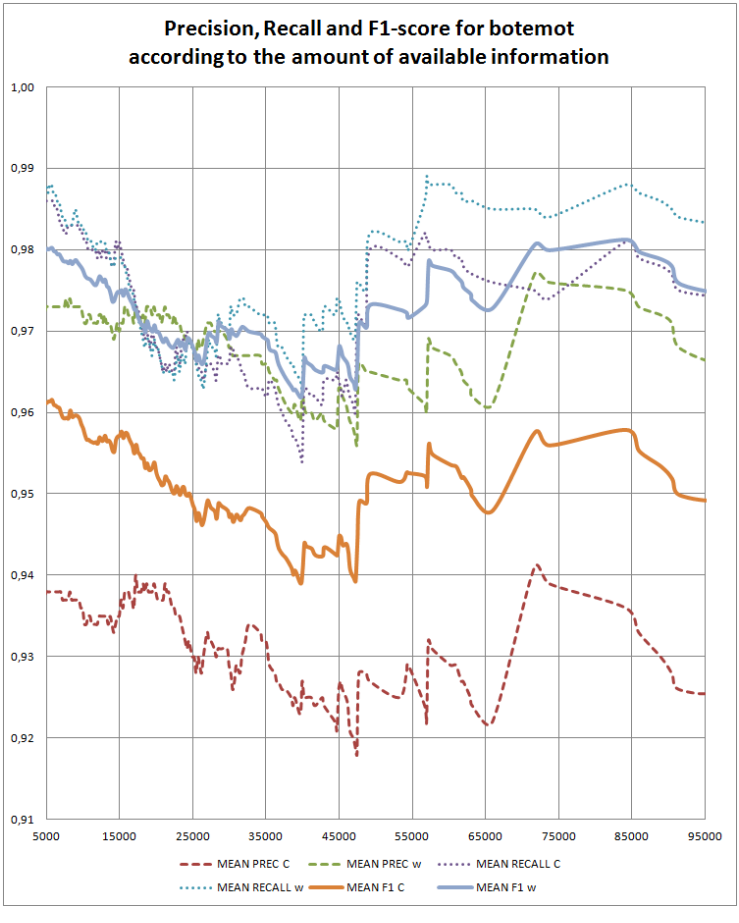

FIGURE 2 : Evolution de la précision, du rappel et du F1 score en fonction du niveau de renseignement des termes. Les courbes en comptage (marquées C) et en poids (marquées W) sont présentées. La tolérance a été fixée à 1 (seuil à la moyenne). La zone représentative ayant un nombre significatif de termes suffisamment renseignés se situe entre 5000 et 100000.

La *figure 3* présente le nombre de termes existant dans le réseau en fonction du niveau de renseignement (pour les relations exploitées par *Botemot*). Il est important de garder à l'esprit que les termes extrêmement bien renseignés (ceux dont le poids cumulés des voisins est supérieur à 100 000) sont relativement peu nombreux. Le but de *Botemot* étant de fournir des sentiments pertinents pour des termes qui justement n'en ont pas, nous devons plutôt focaliser notre attention sur la partie gauche de la courbe de la figure 2.

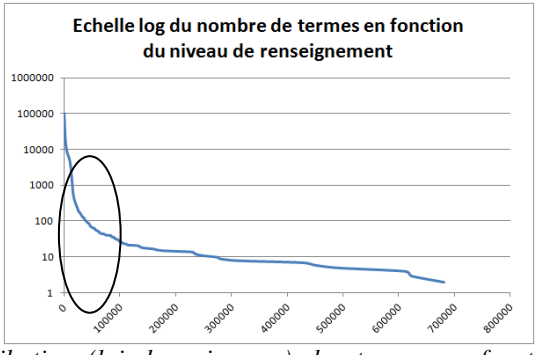

Figure 3 : Distribution (loi de puissance) des termes en fonction du niveau de renseignement. Les termes peu renseignés sont relativement nombreux (partie gauche de la courbe). Les termes entre 5 000 et 100 000 de niveau de renseignement sont ceux à considérer. Ceux inférieurs à 5 000 sont trop peu renseignés pour donner lieu à des inférences correctes. Ceux supérieurs à 100 000 sont trop peu nombreux pour donner lieu à des statistiques fiables.

2) *Par validation manuelle des nouveaux sentiments*

La validation manuelle des nouveaux sentiments proposés par *Botemot* pour un terme donné montre un taux d'acceptation allant de 93% à 97% (soit un taux de rejet entre 7% et 3%). Cette mesure, réalisée sur plus de 500 termes (au 10 octobre 2015), est globale par rapport au niveau de renseignement.

Ces deux approches d'évaluation semblent vérifier les points suivants :

- l'utilisation du poids pour discriminer sentiments importants et sentiments anecdotiques semble pertinente ;
- évaluer une méthode d'inférence par comparaison avec les données fournies par les contributeurs semble être une approche fiable.

Que faire des propositions nouvelles de *Botemot* ? On peut en tenir compte en modifiant le numérateur des formules de précision et rappel comme suit :

- approche optimiste, les supposer justes en leur donnant un score de ½. Cette approche augmente en moyenne le rappel de 0.5% et la précision de presque 2% (avec une tolérance de 1) ;
- approche pessimiste, les supposer fausses en leur donnant un score de -½. Cette approche diminue en moyenne le rappel de 0.5% et la précision d'environ 2% (avec une tolérance de 1) ;

Bien entendu, la méthode d'évaluation ne permet que d'apprécier globalement les performances de l'algorithme. Elle permet de mesurer les taux de rappel et précision que l'on obtiendrait si on validait en bloc les propositions de *Botemot* (ce que nous ne ferons pas).

Les données de sentiments fournis par les joueurs/contributeurs sont d'une très grande qualité tant au niveau de leur précision que dans leur diversité. C'est pour cela que nous les prenons comme référence pour l'algorithme de diffusion. Bien que très simple, l'algorithme de diffusion se révèle très précis. Le rappel varie en fonction du paramétrage. *Botemot* est peu créatif, il fait assez peu de nouvelles propositions (environ 10%). Cependant, on estime (manuellement) qu'environ 95% de nouvelles propositions sont justes et 5% sont acceptables (0,5% sont fausses). Pour les termes polysémiques, la précision et le rappel ne semblent pas inférieurs à ceux obtenus pour les termes monosémiques (mêmes valeurs à 2% près).

L'algorithme actuel ne propose pas de sentiments à activation négative, et en particulier ne repropose pas des sentiments qui ont été invalidés par les joueurs ou les contributeurs. Cela évite de "tourner en rond" en proposant et invalidant en boucle certains sentiments. Les sentiments invalidés ont un poids négatif et peuvent ainsi participer à l'inférence de sentiments pour d'autres termes avec un effet inhibiteur.

Comme conclusion de cette évaluation, *Botemot* semble fonctionner efficacement pour la tâche visée, à savoir proposer des valeurs de sentiments probables pour des termes n'en disposant pas encore dans le lexique, mais ayant des voisins renseignés. Il semble naturel de penser que si les données de départ sont de mauvaise qualité, alors les sentiments proposés le seront également. L'approche par crowdsourcing avec un grand nombre de contributeurs pour chaque terme renseigné a priori (plusieurs dizaines) garantit que les données de départ sont très majoritairement correctes.

3) Evaluation par comparaison avec le lexique de Saif Mohammad

Nous avons entrepris une comparaison des propositions de *Botemot* avec le lexique Emolex de Saif & Turney (2010) dans sa version traduite par Abdaoui *et al*. (2014). Nous rappelons que ce lexique associe à environ 14 000 termes un nombre variable de sentiments, choisis dans un ensemble prédéterminé de 7 sentiments (confiance, peur, tristesse, colère, surprise, dégoût, et joie). Pour chaque terme, ce nombre varie donc entre 0 (indifférence) et 7, chaque sentiment étant activé (valeur 1) ou non (valeur 0). Les 14 000 termes d'Emolex (dans sa version traduite) étant présents dans le réseau JeuxDeMots nous les avons soumis à l'action de *Botemot*, afin de procéder à une comparaison automatique entre les sentiments présents dans Emolex et ceux attribués par notre algorithme. Nous obtenons les résultats suivants :

- Pour 99,9% des mots de Emolex, au moins 1 des sentiments associés dans Emolex fait également partie des propositions de notre algorithme ;
- Pour 69% des mots sans aucun sentiment activé dans Emolex, *Botemot* propose au moins "indifférence" ;
- inclusion totale : dans 92% des cas, tous les sentiments associés dans Emolex font partie des propositions de *Botemot* ;
- inclusion stricte : pour 5% des mots, les sentiments figurant dans Emolex coïncident exactement avec les propositions de *Botemot*.

Une vingtaine de termes du lexique de Saif et Turney n'existent pas dans le réseau JDM. Il s'agit de mots malformés (sans doute en raison d'une traduction automatique approximative).

Pour l'inclusion stricte, il est normal de n'avoir qu'environ 5% de termes la vérifiant car *Botemot* propose une grande quantité de sentiments dont certains sont synonymes (*peur*, *crainte*, *angoisse*, *inquiétude*, …). Ce pourcentage ne peut globalement que diminuer avec le temps.

Une évaluation semi-manuelle de la pertinence des termes proposés par *Botemot* ne figurant pas dans le panel d'Emolex a été conduite sur un échantillon de 1500 termes associés à 8000 sentiments. Elle montre que 70% des sentiments fournis par notre algorithme sont des synonymes de l'un des 7 sentiments de la ressource Emolex. Pour les 30% restant, on ne relève que 0,5% d'erreur (soit 43 sentiments inadéquats.)

Que conclure ? Si Emolex est une ressource relativement précise et correcte, on peut éventuellement lui reprocher d'être basée sur 7 sentiments canoniques (ce qui était un choix de conception). La grande variété et subtilité des sentiments que peuvent éprouver les personnes et, plus encore, la richesse du vocabulaire mis en jeu lors de leur évocation (voir Ekman, 1992) mais aussi (Tausczik et Pennebaker, 2010) pour les aspects psycholinguistiques des émotions) fait penser qu'un vocabulaire émotionnel ouvert serait beaucoup plus opérationnel dans un lexique de sentiments. Par exemple *crainte*, *peur*, *angoisse*, *terreur* sont autant de variantes ayant des traits propres. Il en est de même pour *envie*, *jalousie*, *concupiscence*, *désir*, *avidité*, *appétence*.

4 Conclusion

Utiliser des GWAP (Game With a Purpose) est une approche qui s'avère tout à fait performante concernant le crowdsourcing lexical, mais on peut en augmenter l'efficacité (en couverture) en y adjoignant des mécanismes d'inférence par propagation. Notons que ce qui concerne ici les sentiments peut être appliqué à tout autre type d'information.

L'inférence prend ici la forme d'un algorithme de propagation qui "contamine" un terme avec les informations glanées chez ses voisins. Ce mode de diffusion permet d'accélérer la constitution du lexique, à partir d'un noyau construit par les utilisateurs/joueurs. L'ensemble s'inscrit dans une approche d'apprentissage permanent. Le noyau ne constitue pas une donnée figée, fournie comme paramètre de départ à l'algorithme, mais il évolue et s'accroît en permanence sous l'action conjuguée des joueurs et de l'algorithme. Ainsi, on peut légitimement espérer que les sentiments inférés vont gagner en pertinence et en précision au fil du temps.

Les informations nouvelles, la richesse et la diversité viennent des utilisateurs : *Botemot* ne peut pas deviner un sentiment qui ne serait associé à aucun des termes liés au terme-cible. On remarque également que grâce à la redondance du réseau lexical, la polysémie des termes ne fausse pas l'inférence des sentiments. Mais les sens suscitant les sentiments les plus marqués, en s'imposant, ont tendance à contaminer le sens général. Les points de vue sont divers et potentiellement opposés - par exemple *voiture*, *impôts* peuvent générer des sentiments contradictoires, ce qui fait la richesse de la ressource obtenue. Les sentiments associés à de tels termes sont donc éminemment subjectifs, et fortement dépendant du contexte. Les raffinements sémantiques permette parfois de répondre partiellement, par exemple *table (gastronomie)* et *table (index)* n'évoquent pas les mêmes sentiments.

L'algorithme *Botemot*, qui demeure simple dans son principe, produit des propositions très pertinentes, à en juger par nos différentes évaluations (guère plus de 2% de propositions invalidées). La méthode d'évaluation générale des données inférées par calcul de la précision et du rappel peut, quant à elle, être facilement et efficacement transposée à tout domaine où une ressource est en construction permanente.

Références

ABDAOUI A., AZÉ J., BRINGAY S. & PONCELET P. (2014). *FEEL: French Extended Emotional Lexicon*. 2014. ISLRN: 041-639-484-224-2

BRUN C. (2011). *Detecting opinions using Deep Syntactic Analysis*. Proceedings of Recent Advances in Natural Language Processing, (RANLP 2011), Hissar, Bulgaria, pp. 392-398.

DONG, Z. D., DONG, Q., & HAO, C. L. (2007). *Theoretical findings of HowNet*. Journal of Chinese Information Processing, 21(4), 3-9.

EKMAN P. (1992). *An argument for basic emotions*. volume 6, p. 169–200.

ESULI A. AND SEBASTIANI F. (2006). *SentiWordNet: a publicly available lexical resource for opinion mining*. Proceedings of LREC-06, Gênes, Italie, 6 p.

FELLBAUM C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

FORT K., ADDA G., SAGOT B., MARIANI J. & COUILLAULT A. (2014) *Crowdsourcing for Language Resource Development : Criticisms about Amazon Mechanical Turk Overpowering Use*. Lecture Notes in Artificial Intelligence. Springer, pp. 303-314, 2014, 978-3-319-08957-7.

KAMPS J., MARX M., MOKKEN R. J., DE RIJKE M. (2004) *Using WordNet to Measure Semantic Orientations of Adjectives*. In Proceedings of LREC-04, 4th international conference on language resources and evaluation, Lisbon, PT, volume 4, 2004.

KIM S.-M. & HOVY E. (2004). *Determining the sentiment of opinions*. In Proceedings of the International Conference on Computational Linguistics (COLING), p. 1367–1373.

KIM S.-M. & HOVY E. (2006). *Extracting opinions, opinion holders, and topics expressed in online news media text*. In Proceedings of the Workshop on Sentiment and Subjectivity in Text, SST '06, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

KIRITCHENKO S., XIAODAN Z. & SAIF M. (2014). *Sentiment Analysis of Short Informal Texts*. Journal of Artificial Intelligence Research, volume 50, pages 723-762, August 2014.

LAFOURCADE M., (2007). *Making people play for Lexical Acquisition*. In Proc. SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya, Thailande, 13-15 December 2007, 8 p.

LAFOURCADE M., LE BRUN N., & JOUBERT A. (2015a) *Collecting and Evaluating Lexical Polarity with a Game with a Purpose*. In proc International Conference on Recent Advances in Natural Language Processing (RANLP 2015), Hissar, Bulgaria, September 5-11, 2015, 9 p.

LAFOURCADE M., LE BRUN N., & JOUBERT A. (2015b) *Games with a Purpose (GWAPS)* ISBN: 978-1-84821-803-1 July 2015, Wiley-ISTE, 158 p.

MILLER G. A. (1995). *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.

MUDINAS A., ZHANG D. & LEVENE M. (2012). *Combining lexicon and learning based approaches for concept-level sentiment analysis*. In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining , WISDOM '12, p. 5 :1–5 :8, New York, NY, USA : ACM.

NAVIGLI R. & PONZETTO, S. P. (2010) *BabelNet: Building a Very Large Multilingual Semantic Network*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 216-225.

SAIF. M. & TURNEY P. D. (2010). *Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon*. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California, p. 26–34.

SAIF M. & TURNEY P. (2013). *Crowdsourcing a Word-Emotion Association Lexicon*. Computational Intelligence, 29 (3), pp. 436-465.

SIMPERL E. & TURNEY P. (2013). *Knowledge engineering via human computation*. In Handbook of Human Computation. Springer New York, pp. 131-151.

STRAPPARAVA C. & MIHALCEA R. (2008). *Learning to identify emotions in text*. In Proceedings of the 2008 ACM Symposium on Applied Computing , SAC '08, p. 1556–1560, New York, NY, USA: ACM.

TABOADA M., BROOKE J., TOFILOSKI M., VOLL K. & STEDE M. (2011). *Lexicon-based methods for sentiment analysis*. Computational Linguistics, Volume 37 (2), pp. 267-307.

TAUSCZIK Y. R. & PENNEBAKER J. W. (2010). T*he psychological meaning of words : LIWC and computerized text analysis methods*. Journal of Language and Social Psychology. volume 29, p. 24–54.

TURNEY P. D. & LITTMAN M. L. (2003) *Measuring praise and criticism: Inference of semantic orientation from association*. ACM Trans. Inf. Syst. 21, 4 (October 2003), 315-346.

WHISSELL C. (1989). *The dictionary of affect in language*. Academic Press.