Détermination des sens d'usage dans un réseau lexical construit à l'aide d'un jeu en ligne

Mathieu Lafourcade, Alain Joubert
{lafourcade, joubert}@lirmm.fr

LIRMM – Univ. Montpellier 2 - CNRS
Laboratoire d'Informatique, de Robotique
et de Microélectronique de Montpellier
161, rue Ada – 34392 Montpellier Cédex 5 – France

Abstract

Lexical information is indispensable for the tasks realized in NLP, but collecting lexical information is a difficult work. Indeed, when done manually, it requires the competence of experts and the duration can be prohibitive. When done automatically, the results can be biased by the corpus of texts. The approach we present here consists in having people take part in a collective project by offering them a playful application accessible on the web. From an already existing base of terms, the players themselves thus build the lexical network, by supplying associations which are validated only if they are suggested by a pair of users. Furthermore, these typed relations are weighted according to the number of pairs of users who provide them. Finally, we approach the question of the word usage determination for a term, by searching relations between this term and its neighbours in the network, before briefly presenting the realization and the first obtained results.

Keywords : Natural Language Processing, lexical network, typed and weighted relations, word usage, web-based game

Résumé

Les informations lexicales, indispensables pour les tâches réalisées en TALN, sont difficiles à collecter. En effet, effectuée manuellement, cette tâche nécessite la compétence d'experts et la durée nécessaire peut être prohibitive, alors que réalisée automatiquement, les résultats peuvent être biaisés par les corpus de textes retenus. L'approche présentée ici consiste à faire participer un grand nombre de personnes à un projet contributif en leur proposant une application ludique accessible sur le web. A partir d'une base de termes préexistante, ce sont ainsi les joueurs qui vont construire le réseau lexical, en fournissant des associations qui ne sont validées que si elles sont proposées par au moins une paire d'utilisateurs. De plus, ces relations typées sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées. Enfin, nous abordons la question de la détermination des différents sens d'usage d'un terme, en analysant les relations entre ce terme et ses voisins immédiats dans le réseau lexical, avant de présenter brièvement la réalisation et les premiers résultats obtenus.

Mots-clés : Traitement Automatique du Langage Naturel, réseau lexical, relations typées pondérées, sens d'usage d'un terme, jeu en ligne

1. Introduction

La connaissance de relations lexicales ou fonctionnelles entre termes est nécessaire pour l'exécution d'un très grand nombre de tâches en Traitement Automatique des Langues (TAL). Ces relations que l'on trouve généralement dans des thésaurus ou des ontologies peuvent être mises en évidence de façon manuelle ; par exemple, l'un des plus anciens thésaurus est le Roget, sa version actuelle étant (Kipfer 2001), ou le réseau lexical le plus célèbre est Wordnet (Miller 1990). De telles relations peuvent aussi être déterminées automatiquement à partir de

corpus de textes, par exemple (Robertson et Spark Jones 1976) ou (Lapata et Keller 2005), dans lesquels sont effectuées des études statistiques sur les distributions de mots. En outre, certaines applications de TAL requièrent des informations de différentes natures, comme la synonymie ou l'antonymie, mais également des relations d'hyperonymie/hyponymie, holonymie/méronymie, … L'établissement de telles relations, s'il est effectué manuellement par un ensemble d'experts, nécessite des ressources (en durée et en personnel) qui peuvent être prohibitives, alors que leur extraction automatique sur un corpus de textes est beaucoup trop dépendante des textes choisis.

La méthode développée ici s'appuie sur un système contributif, où ce sont les utilisateurs qui font évoluer la base, au travers d'une interface présentée sous forme d'un jeu. De plus, contrairement aux méthodes classiques qui permettent d'acquérir des informations lexicales généralement statiques, le prototype introduit ici réalise l'acquisition d'informations lexicales évolutives.

Dans cet article, nous présentons les principes d'un jeu (JeuxDeMots[1]) visant à construire la base de relations. L'objectif poursuivi ici concerne avant tout la fiabilité et la qualité des informations recueillies auprès des utilisateurs, l'un des éléments-clé étant qu'une relation ne peut être validée que si elle est proposée par au moins deux utilisateurs. Dans une deuxième partie, utilisant le réseau ainsi obtenu, nous abordons la problématique de la détermination de la polysémie d'usage.

2. Construction du réseau lexical

2.1. Principe du logiciel

Afin d'éviter les écueils d'un système où n'importe quel utilisateur pourrait écrire n'importe quoi, et donc pour assurer la qualité et la sécurité de la base, il a été décidé que les validations des relations proposées anonymement par un joueur seraient effectuées par d'autres joueurs, tout autant anonymement. Pratiquement, les validations sont faites par concordance des propositions entre paires de joueurs. Ce processus de validation rappelle celui utilisé par (von Ahn et Dabbish, 2004) pour l'indexation d'images ou plus récemment par (Lieberman et al., 2007) pour la collecte de ″connaissances de bon sens″. A notre connaissance, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

Une partie se déroule entre deux joueurs, en asynchrone, basée sur la concordance de leurs propositions. Lorsqu'un premier joueur que nous appellerons (A) débute une partie, une consigne concernant un type de compétence (synonymes, contraires, domaines …) est affichée, ainsi qu'un mot[2] M tiré aléatoirement dans une base de mots. Ce joueur (A) a alors un temps limité pour répondre en donnant des propositions correspondant, selon lui, à la consigne appliquée au mot M. Le nombre de propositions qu'il peut faire est limité pour éviter que des joueurs ne frappent n'importe quoi le plus rapidement possible : nous souhaitons que les joueurs réfléchissent ″un minimum″. Ce même mot, avec cette même consigne, est proposé par la suite à un autre joueur que nous appellerons (B) ; le processus est identique. Afin d'accroître l'aspect ludique, pour toute réponse commune dans les propositions de (A) et (B), ces deux joueurs gagnent un certain nombre de points. Le calcul de ce nombre de points est explicité en section 2.2.

---

[1] JeuxDeMots est accessible à l'adresse http://www.lirmm.fr/jeuxdemots. Depuis peu, il existe aussi une version anglaise, ainsi qu'une version thaï et une version japonaise (toutes deux en cours de développement), à l'adresse http://www.lirmm.fr/jeuxdemots/world-of-jeuxdemots.php

[2] Pour la suite de cet article, *mot* et *terme* sont considérés comme synonymes, bien qu'un terme puisse être constitué de plusieurs mots (exemple : *pomme de terre* ou *jeux olympiques*).

Pour le mot cible M, nous mémorisons les réponses communes aux joueurs (A) et (B). Nous ne mémorisons pas les réponses proposées uniquement par l'un des deux joueurs. Cela permet la construction d'un réseau lexical reliant les termes par des relations typées et pondérées, validées par paires de joueurs. Ces relations sont typées par la consigne imposée aux joueurs ; elles sont pondérées en fonction du nombre de paires de joueurs qui les ont proposées, comme explicité en section 2.2. La structure du réseau lexical que nous cherchons ainsi à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds, telles que rappelées par (Polguère, 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions lexicales, telles que présentées par (Mel'čuk et al., 1995). Initialement, les nœuds sont constitués des termes de notre base de départ, mais celle-ci peut s'accroître ; effectivement, si les deux joueurs (A) et (B) d'une même partie proposent un terme initialement inconnu, alors ce terme est ajouté à notre base. La figure 1 présente les relations acquises pour le terme *aile*.

relations ==>

aile ---r_associated:360--> voler
aile ---r_associated:360--> oiseau
aile ---r_associated:330--> avion
aile ---r_associated:240--> plume
aile ---r_associated:120--> poulet
aile ---r_associated:110--> vol
aile ---r_associated:100--> ange
aile ---r_associated:100--> cuisse
aile ---r_lieu:80--> oiseau
aile ---r_holo:80--> oiseau
aile ---r_holo:80--> avion
aile ---r_lieu:70--> avion
aile ---r_associated:70--> planer
aile ---r_lieu:60--> ange
aile ---r_holo:60--> aigle
aile ---r_associated:60--> deltaplane
aile ---r_syn:60--> aileron
aile ---r_has_part:60--> plume
aile ---r_syn:60--> bras
aile ---r_has_part:60--> os
aile ---r_associated:60--> pigeon
aile ---r_holo:50--> voiture
aile ---r_syn:50--> élytre
aile ---r_carac:50--> petite
aile ---r_lieu:50--> aigle
aile ---r_associated:50--> bras
aile ---r_holo:50--> ULM
aile ---r_associated:50--> moineau
aile ---r_carac:50--> grande
aile ---r_associated:50--> ailé
aile ---r_associated:50--> plumes
aile ---r_associated:50--> bâtiment
aile ---r_lieu:50--> ciel
aile ---r_lieu:50--> aéroport
aile ---r_associated:50--> planeur
aile ---r_holo:50--> pigeon

aile ---r_lieu:50--> pigeon
aile ---r_associated:50--> voiture
aile ---r_associated:50--> battre
aile ---r_has_part:50--> muscle
aile ---r_associated:50--> insecte
aile ---r_carac:50--> cassée
aile ---r_lieu:50--> vautour
aile ---r_syn:50--> voilure

relations <==

oiseau ---r_associated:170--> aile
oiseau ---r_has_part:160--> aile
plume ---r_associated:160--> aile
rapace ---r_has_part:140--> aile
papillon ---r_associated:130--> aile
poulet ---r_associated:110--> aile
avion ---r_has_part:100--> aile
insecte ---r_has_part:90--> aile
volaille ---r_has_part:90--> aile
nez ---r_has_part:80--> aile
cuisse ---r_associated:70--> aile
voler ---r_instr:70--> aile
plume ---r_holo:60--> aile
avion ---r_associated:60--> aile
Ailette ---r_associated:60--> aile
frégate ---r_associated:60--> aile
Icare ---r_associated:50--> aile
huîtrier-pie ---r_associated:50--> aile
toucan ---r_associated:50--> aile
voilure ---r_syn:50--> aile
voler ---r_associated:50--> aile
poule ---r_associated:50--> aile
battement ---r_associated:50--> aile
poule ---r_has_part:50--> aile
deltaplane ---r_has_part:50--> aile

Figure 1 : Ensemble des relations acquises pour le terme aile*. Sont présentées tout d'abord les relations dont le terme* aile *est origine, puis celles pour lesquelles le terme* aile *est destinataire. Pour chacune de ces relations, on a en outre son type ("idée associée", "a pour partie" …) ainsi que son poids. Le calcul de cette pondération est expliqué à la section 2.2.*

Il aurait pu être envisagé de mémoriser toutes les réponses, depuis le début du jeu, avec leurs fréquences. Notre base se serait accrue beaucoup plus rapidement, mais cela aurait été au détriment de sa qualité. L'intérêt de la solution retenue est de limiter de façon beaucoup plus drastique les réponses « fantaisistes » ou les erreurs dues à une mauvaise compréhension de la consigne, voire du mot M lui-même. L'émergence des solutions « originales » sera plus lente, mais elle se fera tout de même, après élimination des solutions les plus courantes, grâce au processus des termes « tabous ». Effectivement, lorsqu'une relation *mot M → terme proposé* a été faite par un grand nombre de couples de joueurs, elle devient banale ou taboue ; elle est affichée en même temps que le mot M, afin que les joueurs ne la proposent plus. Ainsi, les joueurs sont amenés à faire d'autres propositions, généralement plus originales. Ceci favorise l'émergence de relations plus rares, mais non l'émergence d'erreurs.

2.2. Emergence et pondération des relations entre termes

Côté jeu, il s'agit de définir le nombre de points gagnés par les joueurs (A) et (B), avec la même consigne sur un même mot M. Côté réseau lexical, il s'agit d'établir des relations entre termes, grâce aux propositions faites par (A) et (B). Pour cette partie, notons :

propositions de (A) : $x_1, x_2, \ldots x_i \ldots, x_n$

propositions de (B) : $y_1, y_2, \ldots y_j \ldots, y_m$

Pour tous les couples (i,j) tels que $x_i = y_j$, nous mémorisons la relation R : M → x_i.

L'un des avantages de notre méthode réside dans gestion de la pondération des relations entre termes. En effet, il est possible d'affecter un poids à la relation R : plus elle a été proposée de fois, plus son poids est important. Dans cette première version de notre prototype, nous avons envisagé un poids de 50 pour sa première occurrence, puis nous augmentons ce poids de 10 pour chaque occurrence suivante. Rappelons qu'une occurrence de R correspond à une proposition de R par le joueur (A) ainsi que le joueur (B) lors d'une même partie. A partir d'un certain nombre d'occurrences, une relation R est bien établie : elle devient alors banale, ou « taboue », et elle est indiquée en même temps que le mot M, afin que les joueurs ne la proposent plus. Ce processus permet de faire émerger plus facilement de nouvelles relations ; sa conséquence sur la base est donc une augmentation du taux de rappel[3].

Le nombre de points obtenus par (A) et (B) dépend du poids de la relation R. Ce nombre de points vaut actuellement : 10% (1000 - poids(R)). Plus la relation est récente, plus elle a de valeur : cela revient à « payer la primauté ». Cette fonction est décroissante : une relation rapporte de moins en moins de points. A partir d'un certain seuil, fixé actuellement à 300 pour la valeur du poids, la relation devient taboue ; elle est alors indiquée en même temps que la consigne et le mot M (c'est une solution donnée, exemple : *aile → oiseau* ou *aile → avion*). Jouer un mot tabou continue à rapporter des points, mais beaucoup moins : cette proposition n'est plus intéressante pour les joueurs qui donc sont invités à faire d'autres propositions. Avec les valeurs indiquées ci-dessus, une relation devient taboue quand elle a été proposée par 25 couples de joueurs.

Même lorsqu'une relation devient taboue, son poids n'est pas figé, mais il évolue beaucoup moins vite car cette relation est proposée moins souvent par les joueurs. Il est tout de même intéressant que le poids de la relation continue d'évoluer. En effet, au bout d'un certain temps, pour un même terme plusieurs relations peuvent être taboues. Si elles avaient le même poids,

[3] D'après (Salton 1968), le taux de rappel peut être défini par le rapport du nombre de relations pertinentes trouvées sur le nombre de relations pertinentes, la précision correspondant au rapport du nombre de relations pertinentes trouvées sur le nombre de relations proposées.

on ne saurait pas laquelle a atteint cet état en premier et donc on ignorerait celle qui est la plus « forte ».

Il a également été prévu un phénomène d'érosion des relations. Effectivement, une relation a pu être créée à la suite d'une erreur commune à deux joueurs, ou bien une relation a pu être conjoncturelle et être beaucoup moins forte, donc moins proposée, par la suite (exemple : *Paris → Jeux olympiques*). A chaque partie sur un mot M, le poids des relations existantes dans notre base à partir de ce mot M qui ne sont proposées par aucun des deux joueurs est très légèrement diminué (actuellement -1). Cela diminuera inexorablement le poids des relations accidentelles, mais nous espérons que cette érosion n'aura qu'un effet négligeable sur les relations fortes[4].

3. Détermination des sens d'usage

3.1. Principe général

En première approximation, il est possible de considérer que si un terme T est polysémique, les termes qui lui sont reliés forment plusieurs groupes distincts, chacun de ces groupes constituant un sens d'usage de T. Nous faisons ici la distinction entre les notions de sens d'usage et de sens. La notion de sens d'usage (appelée plus communément usage) est beaucoup plus fine que celle de sens qui, comme l'a montré (Véronis 2001), est relativement pauvre lorsqu'on se réfère aux dictionnaires traditionnels ou à des ressources comme WordNet. L'usage est donc en TALN une notion plus importante que le sens. Pour citer l'exemple que nous prendrons à la section 3.3, *aile-plume-oiseau* et *aile-poulet-cuisse* constituent deux usages distincts du terme *aile*, alors qu'il s'agit manifestement du même sens de ce terme.

3.2. Détermination des cliques

Comment déterminer une clique ? C'est un ensemble de termes « fortement » reliés entre eux constituant un sous-graphe induit complet (ou clique) dans le réseau lexical. Les liens qui sont considérés ici sont uniquement des relations de type "*idée associée*" symétrique. Nous faisons abstraction des autres types de relation ("*tout de*", "*partie de*", …) en raison de leur caractère non symétrique. L'objectif est la construction d'un réseau lexical dans lequel chaque nœud est constitué par un usage d'un terme, et non plus un terme regroupant ses éventuels différents usages.

Dans le réseau lexical, un terme T est directement relié à n termes : $T_1,\ldots,T_i,\ldots,T_j,\ldots,T_n$. Les termes T_i et T_j appartiennent à deux usages différents de T si :

$$\text{poids}(T_i\text{-}T_j) < k * \min(\text{poids}(T\text{-}T_i), \text{poids}(T\text{-}T_j)) \quad \text{où k est un coefficient de seuil.}$$

Nous nous attachons ici aux différents usages du terme T ; le but est donc de regrouper ces n termes en un ou plusieurs groupes, chacun constituant un usage de T.

Les termes $T_{i1}, \ldots, T_{im}$ constituent le $i^{ème}$ usage de T si les (m+1)*m/2 relations entre ces (m+1) termes existent. Ainsi, un terme T_j n'appartiendra pas à ce $i^{ème}$ usage de T si au moins une des relations entre ce terme T_j et l'un des termes $T_{i1}, \ldots, T_{im}$ n'existe pas.

Il aurait été possible de considérer, non pas l'existence ou l'absence des relations, mais les poids relatifs de celles-ci. Dans ce cas, les termes $T_{i1}, \ldots, T_{im}$ auraient constitué le $i^{ème}$ usage

[4] Ce processus d'érosion est encore au stade expérimental : nous n'avons pas assez de recul pour voir ces effets dans la durée.

de T si le poids de chacune des m*(m-1)/2 relations entre deux termes T_{i1}, … ,T_{im} quelconques était supérieur ou égal à :

k * min (poids(T-T_{i1}), … ,poids(T-T_{im}))

où la valeur du coefficient k, probablement proche de 1, ne peut être donnée que par l'expérience.

3.3. Exemple du terme aile

Les figures 2 et 3 illustrent, sur un exemple simple, les résultats obtenus permettant de déterminer les différents usages d'un même terme, ainsi que la pertinence de chacun de ces usages (le principe de calcul de cette pondération est explicité à la section 3.4). Il est possible de remarquer sur cet exemple qu'un usage étant constitué d'une clique de termes, et non simplement d'une composante connexe, un même terme T_i relié à T peut appartenir à plusieurs cliques et donc un même terme T_i peut servir à définir plusieurs usages de T.

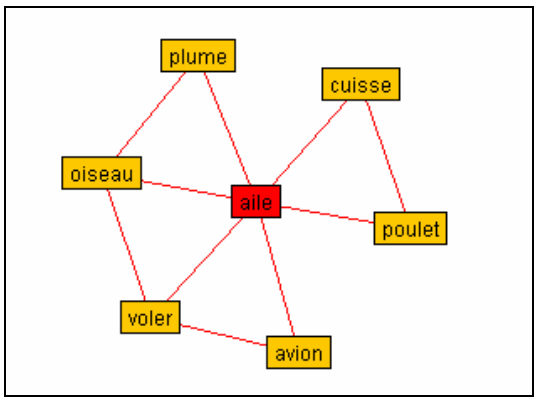

Figure 2 : Cette figure montre, autour du terme aile*, le réseau lexical en ne considérant que les relations de type "idées associées" symétrique. Ainsi reliés au terme* aile *se trouvent six termes qui correspondent selon les critères définis dans cet article à quatre usages.*

Clique 0: 'aile' 'oiseau' 'plume'	(P = 1970 / nl = 6 / moy = 328 / REL = 361)
Clique 1: 'aile' 'oiseau' 'voler'	(P = 1380 / nl = 6 / moy = 230 / REL = 253)
Clique 2: 'aile' 'voler' 'avion'	(P = 1030 / nl = 6 / moy = 172 / REL = 189)
Clique 3: 'aile' 'poulet' 'cuisse'	(P = 780 / nl = 6 / moy = 130 / REL = 143)

Figure 3 : Cette figure montre les quatre usages du terme aile *effectivement décelés :*

- aile-oiseau-plume => *aile : élément de l'anatomie d'un oiseau*
- aile-oiseau-voler => *aile : instrument du vol d'un oiseau*
- aile-voler-avion => *aile : partie d'un avion lui permettant de voler*
- aile-poulet-cuisse => *aspect gastronomique de l'aile*

Pour chacun de ces usages du terme aile *figure également une évaluation de sa pertinence (valeur* REL*). Le terme* aile *est polysémique, les quatre usages ci-dessus correspondent à deux sens distincts :* aile d'oiseau *et* aile d'avion*. On remarquera que notre réseau n'est pas encore complet, les sens* aile de moulin *ou* aile d'un bâtiment *par exemple, n'y figurent pas. Dans notre réseau, la relation* aile --> bâtiment *existe (voir fig.1), mais elle n'est pas symétrique, la relation* bâtiment --> aile *n'existant pas.*

3.4. Pertinence d'un usage

Evaluer la pertinence d'un usage consiste à obtenir une mesure de son importance à la fois en fréquence mais aussi en couverture lexicale. On émettra l'hypothèse que pour un terme donné

et en dehors de tout contexte spécifique, l'usage le plus pertinent est celui auquel on pense en premier en général. Ainsi donc, lors d'une analyse sémantique de texte, les usages peuvent être pondérés par défaut en fonction de leur pertinence a priori. Compte tenu du principe de la pondération des relations dans notre réseau lexical, le poids d'un usage est corrélé aux poids des relations entre les termes de la clique qui caractérise cet usage. Ainsi, pour une clique C de m termes et comprenant le terme T, le poids de l'usage correspondant sera égal à :

$$P(C) = \sum_{i,j} poids(T_i,T_j)$$

$$Rel(C) = Ln(m) * P(C) / [m*(m -1)]$$

Le terme *poids(T_i,T_j)* est le poids de la relation entre T_i etT_j. La pertinence est la moyenne des poids des relations existant entre les termes, valeur qui exprime la cohérence de la clique, que multiplie le logarithme du nombre de termes impliqué dans la clique.

4. Mise en œuvre et résultats

4.1. Réalisation

Le logiciel JeuxDeMots a été développé en PHP/MySQL; et certains programmes annexes ont été réalisés en langage JAVA et C++. L'interface, la comptabilisation de points, mais également des notions de niveau, honneur, captures de mots, procès entre joueurs … ainsi que l'affichage du classement des joueurs, ont été mis en œuvre afin d'accroître l'aspect attrayant du jeu. Le but recherché est d'inciter les joueurs à revenir régulièrement sur le site, et donc d'augmenter d'autant le nombre de relations acquises : c'est l'intérêt majeur de cette dimension jeu par rapport à un logiciel qui se contenterait de demander des relations à des utilisateurs qui, certes, auraient plus conscience de leur rôle d'«experts », mais qui, probablement, y consacreraient moins de temps.

4.2. Déroulement d'une partie

Chaque fois qu'un joueur se connecte au site et démarre une partie, une consigne est alors affichée pendant quelques secondes (par exemple : *"Donner des idées associées au terme suivant"*), avant que le terme sur lequel il doit appliquer cette consigne n'apparaisse à l'écran. Ce terme est tiré aléatoirement dans une base d'environ 150.000 termes. Il a alors une minute pour donner ses réponses. Si le joueur est (B), il est procédé à l'affichage immédiat du résultat de la partie : propositions qu'avait faites le joueur (A) et nombre de points gagnés. S'il est joueur (A), ces informations lui seront envoyées par mail après que (B) ait joué. Les parties proposées au joueur sont soit des parties en création où il est joueur (A), soit des parties à finir pour lesquelles il est joueur (B). Il y a donc en permanence un ensemble de parties à finir.

Si, à l'affichage d'un mot et de la consigne, un joueur estime n'avoir aucune idée, il a la possibilité de « passer » : la partie se termine alors prématurément. L'absence de réponse du joueur peut avoir deux causes principales : soit le terme n'est pas un terme courant (par exemple : *"gnomon"*), soit la consigne appliquée à ce terme n'a pas une grande signification (par exemple : *"contraires de pigeon ?"*). Le système mémorise alors le fait que ce terme est peu productif, en particulier par rapport à cette consigne ; ce terme appliqué à cette consigne sera moins souvent proposé.

Toute partie créée avec un joueur (A) génère deux parties à finir avec des joueurs (B). En effet, si tel n'était pas le cas, il suffirait que le joueur (B) passe le mot sans faire de proposition pour initier un sentiment de frustration chez le joueur (A), ce qui le démotiverait à revenir participer. Il est donc proposé à tout joueur qui se connecte un plus grand nombre de

parties à finir que de parties en création ; cela est plus motivant pour le joueur qui donc voit plus souvent le résultat immédiat de ses propositions.

Afin de permettre à tout joueur de se comparer aux autres membres de la communauté, il est possible d'afficher un tableau récapitulatif des joueurs enregistrés, avec leurs performances, par ordre de classement selon les points d'honneur, ainsi que les meilleurs scores obtenus sur une partie.

4.3. Résultats

Cette première version de JeuxDeMots est relativement récente : son lancement a eu lieu en juillet 2007. En environ six mois, plus de 500 joueurs se sont enregistrés et la plupart d'entre eux se connectent plusieurs fois par semaine. 50.000 parties ont été jouées : elles ont fait émerger près de 70.000 relations, dont 32.000 de type "idées associées". Actuellement, plus de 600 relations sont taboues, soit près de 1% du nombre total de relations. Il y a une émergence rapide des relations et on constate que les plus fortes sont statistiquement créées en premier. L'évolution de la base de termes est nécessairement plus lente : elle compte à ce jour environ 154.000 termes ; les joueurs y ont déjà ajouté plus de 2.000 nouveaux termes, principalement conjoncturels ou liés à l'actualité. La figure 4 présente un exemple partiel du réseau lexical obtenu.

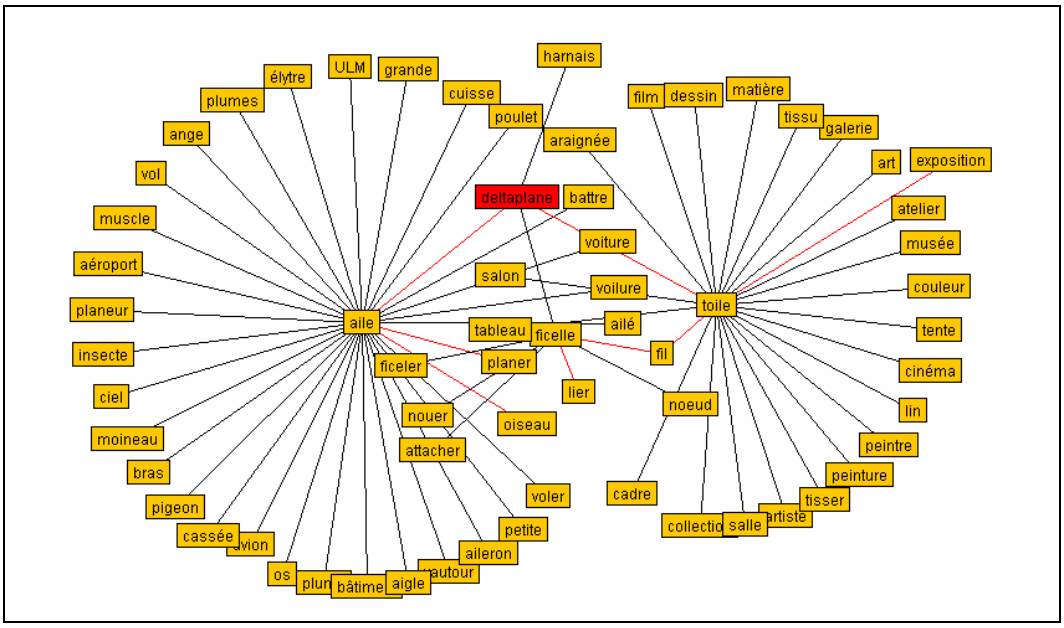

Figure 4 : Exemple partiel du réseau lexical avec une distance de 2, c'est-à-dire que l'on présente un terme central (ici : deltaplane*), tous les termes directement reliés à ce terme central, ainsi que tous les termes qui leur sont reliés. On constate que l'une des relations principales est* deltaplane → aile*, d'où l'affichage des relations au départ de* aile*. Les nœuds correspondant à des termes très généraux (ici :* aile *ou* toile*) sont des « hubs » dans le réseau, destinataires d'un grand nombre de relations. Le graphe présenté ici ne tient pas compte de la direction des relations.*

Nous avons comparé les résultats obtenus à ce jour grâce à JeuxDeMots (JDM) avec les données d'Euro Wordnet Français (EWF). JDM possède environ 6 fois plus de termes qu'EWF qui en comporte environ 23.000. Par contre, EWF est légèrement plus riche au niveau des relations : il en compte actuellement un peu plus de 100.000 ; mais le nombre de relations dans JDM est en progression d'environ 10.000 relations supplémentaires par mois. JDM devrait donc prochainement atteindre le même nombre de relations qu'EWF.

En ce qui concerne la détermination des usages d'un terme, la figure 3, ainsi que les annexes ci-après, montrent des exemples de cliques détectées dans le réseau lexical, avec leurs poids respectifs (poids total de la clique et moyenne des poids de ses relations). Nous indiquons également le nombre de relations "*idée associée*" dont ce terme est origine, avant de lister parmi ces relations celles qui sont symétriques, avec leur poids.

Nous avons effectué des mesures sur les 1000 termes les plus fréquemment proposés par les joueurs. Nous avons divisé les termes en quatre quartiles en fonction de leur fréquence d'utilisation dans le jeu par les joueurs. Nous avons obtenu les résultats reproduits sur le tableau de la figure 5 (extraits à mi février 2008) :

quartile	NB termes	NB moy cliques	<100	100-199	200-299	>300
Q1	66	8.5	2.4	3.8	1.7	0.6
Q2	169	4.9	1.9	1.9	0.88	0.26
Q3	299	3.4	1.4	1.2	0.6	0.2
Q4	466	1.8	1	0.6	0.2	<0.1

Figure 5 : Les données de ce tableau correspondent aux 1000 termes les plus fréquemment rencontrés dans JDM. Leurs fréquences d'utilisation ont été divisées en quatre quartiles sur la fréquence, Q1 correspondant aux termes dont la fréquence d'utilisation est la plus élevée. Pour chaque quartile, nous avons indiqué le nombre de termes, le nombre moyen de cliques (donc d'usages) de chaque terme, ainsi que la répartition des pertinences de ces cliques.

On observe donc qu'en moyenne plus un terme est utilisé fréquemment, plus il a d'usages différents. Ce n'est certes pas une découverte lexicale, mais que cela soit observé dans les données issues de JeuxDeMots est un indice positif sur la validité linguistique de celles-ci. Ce résultat s'explique aussi par le taux plus faible de parties jouées sur les termes rares que pour les termes fréquents. Les cliques dont la pertinence est très faible (< 100) relèvent de sens peu découvert ou d'hapax. Seules les cliques raisonnablement formées (> 200) peuvent être considérées comme correspondant à des usages pertinents.

5. Conclusion

Le prototype JeuxDeMots est un jeu en ligne sur le web dont l'objectif est la construction d'un réseau lexical. L'émergence de relations typées et pondérées entre termes s'effectue grâce au concours d'un grand nombre d'utilisateurs dont l'activité a pour effet de bord la construction de ce réseau. Ces utilisateurs ne sont certes pas des linguistes, mais nous pensons que leur nombre permettra d'obtenir un réseau évolutif de bonne qualité, avec une couverture satisfaisante de l'ensemble des connaissances générales. Notre but n'est pas la constitution d'une base d'experts, mais d'une base de connaissances ″moyennes″, représentant une culture générale commune.

De plus, au vu des résultats actuels, bien que récents et nécessairement partiels, nous pensons arriver à séparer les différents sens d'usage parmi ceux représentés pour chaque terme du réseau. Ce dernier travail n'en est qu'à ses débuts : il serait possible de considérer en plus de la relation "*idée associée*" d'autres relations symétriques dans la détermination des cliques[5], ou pourquoi ne pas aussi considérer les quasi-cliques (sous-graphe induit presque complet)

[5] Nous avons récemment implémenté une version de notre logiciel dans laquelle nous considérons les relations de tout type pour déterminer les cliques, mais nous n'avons pu encore faire d'évaluation à ce sujet.

dans la détermination des usages d'un terme. Pour un même terme, deux cliques qui ont une forte proportion de termes en commun correspondent-elles réellement à deux usages distincts de ce terme ? Cette question reste actuellement ouverte : il est possible que, malgré l'évolution de notre réseau, des cliques actuellement séparées ne fusionneront pas, alors qu'on peut manifestement les considérer comme un même usage. Par exemple, pour le terme *ciel*, parmi les cliques existantes on trouve *ciel-nuage-gris* et *ciel-nuage-soleil* qui ne fusionneront probablement pas, car la relation *soleil-gris* a peu de chances d'émerger. Ce biais est-il induit par notre méthodologie ou est-il dû à la représentation en réseau lexical ?

Références

vonAhn L. et Dabbish L. (2004) Labelling Images with a Computer Game. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 319-326.

Kipfer B.A. (2001). *Roget's International Thesaurus*, sixth edition, Harper Resource (First Edition : 1852)

Lapata M. et Keller F. (2005) Web-based Models for Natural Language Processing. In *ACM Transactions on Speech and Language Processing*, vol.2, n°1, pp. 1-30.

Lieberman H., Smith D.A. and Teeters A. (2007) Common Consensus: a web-based game for collecting commonsense goals, *International Conference on Intelligent User Interfaces (IUI'07)*, Hawaii, USA

Mel'čuk I.A., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot AUPELF-UREF

Miller G.A., Beckwith R., Fellbaum C., Gross D. and Miller K.J. (1990) Introduction to WordNet: an on-line lexical database. In: *International Journal of Lexicography* 3 (4), pp. 235-244.

Polguère A. (2006) Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability* (COLING/ACL 2006), Sydney, pp. 50-59.

Robertson S. et Spark Jones K. (1976) Relevance weighting of search terms, *Journal of the American Society for Information Science*, n° 27, pp. 129-146.

Salton G. (1968) *Automatic Information Organization and Retrieval*, Mac Graw Hill, NY.

Véronis J. (2001) Sense tagging: does it make sense? *Corpus linguistics' 2001 Conference*, Lancaster, U.K.

Annexes

Ci-après, les différents usages actuellement décelés dans notre réseau lexical pour le terme *ciel*. Il parait peu probable que les cliques *1:ciel-nuage-gris* et *2:ciel-nuage-soleil* puissent fusionner, alors que les cliques *5:ciel-soleil-astronomie* et *6:ciel-soleil-étoile* fusionneront probablement prochainement.

ciel ---r_associated#0:280--> nuage
ciel ---r_associated#0:270--> bleu
ciel ---r_associated#0:200--> soleil
ciel ---r_associated#0:130--> avion
ciel ---r_associated#0:130--> oiseau
ciel ---r_associated#0:110--> étoile
ciel ---r_associated#0:60--> gris
ciel ---r_associated#0:50--> astronomie

0: 'ciel' 'bleu'
(P = 780 / nl = 2 / moy = 390 / REL =270)

1: 'ciel' 'nuage' 'gris'
(P = 680 / nl = 6 / moy = 113 / REL =125)

2: 'ciel' 'nuage' 'soleil'
(P = 860 / nl = 6 / moy = 143 / REL =157)

3: 'ciel' 'avion'
(P = 200 / nl = 2 / moy = 100 / REL =69)

4: 'ciel' 'oiseau'
(P = 180 / nl = 2 / moy = 90 / REL =62)

5: 'ciel' 'soleil' 'astronomie'
(P = 510 / nl = 6 / moy = 85 / REL =93)

6: 'ciel' 'soleil' 'étoile'
(P = 860 / nl = 6 / moy = 143 / REL =157)

Ci-dessous, sont donnés quelques exemples montrant les différents usages décelés.

verre ---r_associated#0:100--> boire
verre ---r_associated#0:70--> vin
verre ---r_associated#0:70--> vitre
verre ---r_associated#0:60--> boisson
verre ---r_associated#0:50--> fenêtre

0: 'verre' 'boire' 'vin' 'boisson'
(P = 950 / nl = 12 / moy = 79 / REL =110)

1: 'verre' 'vitre' 'fenêtre'
(P = 880 / nl = 6 / moy = 147 / REL =161)

barreau ---r_associated#0:140--> avocat
barreau ---r_associated#0:120--> prison
barreau ---r_associated#0:80--> chaise
barreau ---r_associated#0:60--> prisonnier

0: 'barreau' 'avocat'
(P = 200 / nl = 2 / moy = 100 / REL = 69)

1: 'barreau' 'prison' 'prisonnier'
(P = 820 / nl = 6 / moy = 137 / REL = 150)

2: 'barreau' 'chaise'
(P = 140 / nl = 2 / moy = 70 / REL = 49)

cuisse ---r_associated#0:140--> jambe
cuisse ---r_associated#0:130--> poulet
cuisse ---r_associated#0:100--> aile

0: 'cuisse' 'poulet' 'aile'
(P = 780 / nl = 6 / moy = 130 / REL = 143)

1: 'cuisse' 'jambe'
(P= 190 / nl = 2 / moy = 95 / REL = 66)

plume ---r_associated#0:750--> oiseau
plume ---r_associated#0:210--> léger
plume ---r_associated#0:160--> aile
plume ---r_associated#0:150--> stylo
plume ---r_associated#0:150--> écrire
plume ---r_associated#0:140--> encre
plume ---r_associated#0:100--> oreiller
plume ---r_associated#0:60--> poids
plume ---r_associated#0:50--> chatouiller

0: 'plume' 'oiseau' 'aile'
(P = 1970 / nl = 6 / moy = 328 / REL =361)

1: 'plume' 'léger' 'poids'
(P = 620 / nl = 6 / moy = 103 / REL =114)

2: 'plume' 'stylo' 'écrire' 'encre'
(P = 1320 / nl = 12 / moy = 110 / REL =152)

3: 'plume' 'oreiller'
(P = 170 / nl = 2 / moy = 85 / REL =59)

4: 'plume' 'chatouiller'
(P = 100 / nl = 2 / moy = 50 / REL =35)

parfum ---r_associated#0:430--> odeur
parfum ---r_associated#0:210--> senteur
parfum ---r_associated#0:100--> odorat
parfum ---r_associated#0:90--> fleur
parfum ---r_associated#0:90--> nez
parfum ---r_associated#0:90--> sentir
parfum ---r_associated#0:60--> essence

0: 'parfum' 'odeur' 'senteur' 'odorat' 'nez' 'sentir'
(P = 3980 / nl = 30 / moy = 133 / REL = 238)

1: 'parfum' 'fleur'
(P = 140 / nl = 2 / moy = 70 / REL = 49)

2: 'parfum' 'essence'
(P = 130 / nl = 2 / moy = 65 / REL = 45)