Détermination et pondération des raffinements d'un terme à partir de son arbre des usages nommés

Mathieu Lafourcade, Alain Joubert

LIRMM – Univ. Montpellier 2 - CNRS
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
161, rue Ada – 34392 Montpellier Cédex 5 – France
{lafourcade, joubert}@lirmm.fr

Résumé Grâce à la participation d'un grand nombre de personnes via des jeux accessibles sur le web, nous avons construit un réseau lexical évolutif de grande taille pour le Français. A partir de cette ressource, nous avons abordé la question de la détermination des sens d'usage d'un terme, puis après avoir introduit la notion de similarité entre ces différents usages, nous avons pu obtenir pour un terme son arbre des usages : la racine regroupe tous les usages du terme et une descente dans l'arbre correspond à un raffinement de ces usages. Le nommage des différents nœuds est effectué lors d'une descente en largeur. En simplifiant l'arbre des usages nommés, nous déterminons les différents sens d'un terme, sens que nous introduisons dans le réseau lexical en tant que nœuds de raffinement du terme considéré. Nous terminons par une évaluation empirique des résultats obtenus.

Abstract Thanks to the participation of a large number of persons via web-based games, a large-sized evolutionary lexical network is available for French. With this resource, we approached the question of the determination of the word usages of a term, and after introducing the notion of similarity between these various word usages, we were able to build for a term its word usage tree: the root groups together all possible usages of this term and a search in the tree corresponds to a refinement of these word usages. The labelling of the various nodes is made during a width-first search. From its labelled word usage tree, we obtain the different meanings of a term, which can be inserted in the lexical network as refinement nodes for this term. Lastly, we present an evaluation of the results we obtain.

Mots-clés : réseau lexical, arbre des usages nommés d'un terme, pondération des sens d'un terme
Keywords: lexical network, tree of labelled word usages for a term, weighting of the meanings

1. Introduction

Dans cet article, nous rappelons comment nous avons exploité un réseau obtenu grâce à deux jeux accessibles sur le Web pour déterminer les différents sens d'usage des termes qui le constituent et ainsi construire pour chaque terme son arbre des usages nommés. Dans une deuxième partie, considérant l'arbre des usages d'un terme, nous faisons l'hypothèse qu'il est possible d'en déterminer les différents sens : cela nous conduit à obtenir les raffinements du terme considéré, que nous implémentons sous forme de nœuds dans le réseau lexical. Ces nœuds de raffinement peuvent être reliés aux nœuds déjà existants dans notre réseau, participant ainsi à son accroissement. Nous abordons ensuite la notion de

pondération relative des sens d'un terme, avant de conclure en présentant les premiers éléments d'une évaluation empirique des résultas obtenus.

2. Construction de l'arbre des usages nommés d'un terme

2.1. Construction du réseau lexical

La construction progressive du réseau lexical selon un principe contributif grâce à deux jeux accessibles sur le web, JeuxDeMots[1] et PtiClic[1], a déjà précédemment été décrite par (Lafoucade et Joubert, 2009). Ces deux jeux reposent sur les propositions faites par les joueurs de termes destination à partir de termes origine et de types de relation, selon un principe qui rappelle celui utilisé par (Lieberman et al., 2007) pour la collecte de ″connaissances de bon sens″. La structure du réseau lexical obtenu s'appuie sur les notions de nœuds et de relations entre ces nœuds, selon un modèle proposé par (Collins et Quillian, 1969) et plus récemment décrit par (Polguère, 2006). Chaque nœud de notre réseau est constitué par un terme ou raffinement d'un terme et les relations[2] entre ces nœuds traduisent des fonctions lexicales, telles que présentées par (Mel'čuk et al., 1995) . Nous avons conservé pour la construction du réseau tous les types de relation proposés y compris la relation d'antinomie, car deux termes antinomiques appartiennent au même champ lexical.

2.2. Détermination des usages

Lorsqu'un terme est polysémique, les termes qui lui sont directement reliés dans le réseau lexical forment plusieurs groupes distincts, chacun de ces groupes constituant un sens d'usage de ce terme. La notion de sens d'usage, plus communément appelée usage, est beaucoup plus fine que celle de sens rencontrée dans les dictionnaires traditionnels, comme l'a montré (Véronis, 2001). Nous avons fait l'hypothèse que les usages d'un terme correspondent dans le réseau aux différentes cliques auxquelles appartient ce terme ; deux termes appartiennent à une même clique s'il existe au moins une relation de l'un vers l'autre et au moins une relation de l'autre vers l'un. Notre approche est analogue à celle développée par (Ploux et Victorri, 1998) à partir de dictionnaires de synonymes. Nous avons déjà introduit la notion de pertinence d'une clique dans (Lafourcade et Joubert, 2009). Rappelons que la pertinence d'une clique, notée Rel pour « Relevance » en Anglais, est la moyenne des poids des relations entre les termes de la clique, valeur qui en exprime la cohérence, que multiplie le logarithme du nombre de termes impliqués dans cette clique, afin de tenir compte de l'importance de sa couverture lexicale.

La notion de similarité « fine » entre deux usages d'un même terme a récemment été introduite par (Lafourcade et Joubert, 2010) en se basant sur les principes énoncés par (Tversky, 1977) et leur application en TALN faite par (Fairon et Ho, 2004). Nous avons défini la similarité comme égale au rapport du poids des relations liant deux cliques sur le poids total des relations liant l'ensemble des termes de ces deux cliques : elle correspond à l'indice de Jaccard défini en statistiques.

2.3. Construction de l'arbre des usages nommés

L'arbre des usages d'un terme polysémique est une représentation structurée de ces différents usages. La racine de l'arbre regroupe tous les sens de ce terme et, plus on s'éloigne de la racine, plus on rencontre

[1] JeuxDeMots et PtiClic sont accessibles à l'adresse http://jeuxdemots.org et http://pticlic.org. Le réseau lexical résultant est disponible à l'adresse http://www.lirmm.fr/jeuxdemots/rezo.php et le « dictionnaire » qui en a été obtenu est disponible à l'adresse http://www.lirmm.fr/jeuxdemots/diko.php .

[2] Dans le réseau lexical obtenu, une relation peut être considérée comme un quadruplet : terme origine, terme destination, type et poids de la relation; entre deux mêmes termes, plusieurs relations de types différents peuvent donc exister.

des distinctions fines d'usages. La construction de l'arbre des usages d'un terme s'effectue par fusions successives des cliques selon une méthode de type « bottom-up ». Pour cela, nous fusionnons les cliques, deux à deux, en commençant par celles dont le coefficient de similarité est le plus élevé : nous constituons ainsi des quasi-cliques, regroupements de cliques, représentant donc des regroupements d'usages, proches lors des premières fusions, de moins en moins proches lors des fusions successives. L'algorithme de fusions s'arrête lorsque tous les coefficients de similarité sont nuls.

Le nommage des nœuds de l'arbre des usages d'un terme est effectué selon une méthode « top-down ». La racine est étiquetée par le terme lui-même. Chaque nœud de l'arbre est étiqueté par le terme issu de la clique (ou quasi-clique) qu'il représente dont la somme des poids des relations le liant au terme racine est la plus élevée, après élimination des termes déjà utilisés pour nommer les nœuds précédemment traités dans le nommage de l'arbre. Certains nœuds peuvent ne pas être étiquetables : dans ce cas, nous réalisons un élagage de l'arbre. La figure 1 représente l'arbre des usages nommés ainsi construit pour le terme *sapin*.

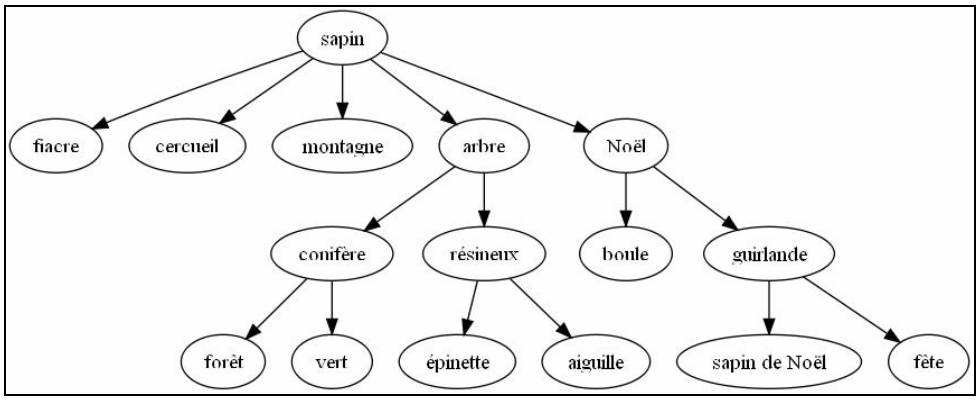

Figure 1 : Arbre des usages nommés pour le terme *sapin*.

3. Sens d'un terme

3.1. Détermination des sens d'un terme

Nous avons fait l'hypothèse que les nœuds de profondeur égale à 1 dans l'arbre des usages, correspondant à des cliques (ou quasi-cliques)[3] dont la similarité est nulle, correspondent aux sens de ce terme tels qu'on les rencontre dans un dictionnaire ainsi que parfois à des champs lexicaux « forts ». Par expérience, cette hypothèse s'est trouvée largement vérifiée. Comme nous l'avons vu ci-dessus, dans l'arbre des usages, il est possible de nommer les nœuds : à la profondeur 1, le terme nommant un nœud constitue un raffinement du terme racine : il correspond à l'un de ses sens.

3.2. Utilisation des sens dans le réseau

Chaque raffinement, actuellement après validation par un expert, conduit à la création d'un nœud dans le réseau lexical. Ainsi, par exemple, les nœuds *voile (bateau)*[4], *voile (plaisance)* et *voile (coiffure)* ont été créés. Chaque nœud de raffinement est relié dans le réseau au terme racine par une relation typée *est_raffinement*, mais il est également possible de relier directement à ce nouveau nœud non seulement les termes (initialement reliés au terme racine) faisant partie des cliques constituant le sous arbre

[3] Pour simplifier, pour la suite de cette section, nous utiliserons le terme de clique pour désigner des cliques ou des quasi-cliques résultant de la fusion de cliques.
[4] Le nœud dénommé *voile (bateau)* représente le sens *bateau* du terme *voile*.

correspondant à ce raffinement, mais également les nœuds de raffinement de ces termes dont la clique comprend le terme racine initial. Il est ainsi possible d'obtenir des relations entre nœuds de raffinement.

Par exemple, les termes *voile* et *vaisseau* étaient déjà reliés dans le réseau lexical par une relation de type « association libre »; de plus, le terme *vaisseau* appartient à la clique du terme *voile* nommée *bateau* et le terme *voile* appartient à la clique du terme *vaisseau* nommée *navire* ; après processus de raffinement de ces deux termes *voile* et *vaisseau*, les nœuds *voile (bateau)* et *vaisseau (navire)* sont à présent reliés. La figure 2 illustre cet exemple en montrant un extrait des arbres des usages des termes *voile* et *vaisseau*, ainsi que la relation entre ces deux termes dans le réseau lexical. Les nouvelles relations créées sortant de *voile (bateau)* et *vaisseau (navire)* y sont également illustrées.

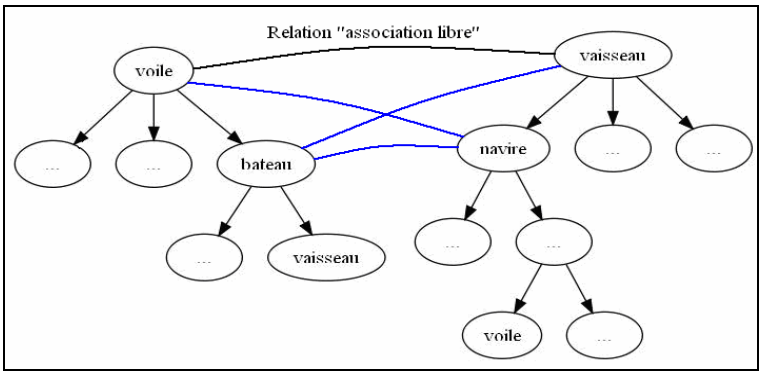

Figure 2 : Extraits des arbres des usages nommés des termes *voile* et *vaisseau*. Les liens représentés ici entre *voile* et *vaisseau*, ainsi qu'entre *voile (bateau)* et *vaisseau (navire)* sont des relations dans le réseau lexical. Cette dernière relation, automatiquement créée, relie deux nœuds de raffinement.

3.3. Réinjection des sens dans JDM

Afin d'améliorer la qualité de notre réseau en vue d'une utilisation en désambiguïsation, il a été décidé de réinjecter ces nœuds de raffinement dans JDM. Lorsqu'un joueur propose un terme polysémique dont les raffinements ont été validés, ces raffinements lui seront automatiquement proposés, et il aura la possibilité soit de conserver sa proposition (polysémique) initiale, soit d'indiquer parmi les raffinements proposés celui qui lui semble le plus approprié à son idée. Cela permet d'obtenir des relations entrantes sur le terme raffiné. Les termes raffinés sont également utilisés dans JDM en tant que termes origine : cela permet alors d'en obtenir des relations sortantes. Nous avons donc ainsi à la fois des relations entrantes et des relations sortantes de termes raffinés (après validation du raffinement par un expert).

Par exemple, sur une partie *spécifiques* de *fruit*, un joueur peut proposer *baie*. Lui seront alors indiqués les raffinements existant dans notre réseau : *baie (fenêtre)*, *baie (golfe)* et *baie (fruit)*. Il peut également être demandé à un joueur de proposer des *synonymes* de *baie (fenêtre)*, restreignant ainsi les propositions qu'il peut faire à ce raffinement de *baie*.

4. Pondération relative des sens d'un terme

Compte tenu de la pondération des relations dans notre réseau lexical, il est possible de définir un poids relatif pour chacun des sens d'un terme. Pour cela, nous considérons uniquement les cliques correspondant aux nœuds de profondeur 1 dans l'arbre des usages. Le $i^{ème}$ sens du terme T est représenté par la clique Ci. Cette clique Ci peut être une clique (clique effective et non une quasi-clique) initialement détectée dans notre réseau lexical : sa pertinence Rel(Ci) est alors celle définie dans (Lafourcade et Joubert, 2009). Plus fréquemment, Ci est une quasi-clique résultant de la fusion de cliques initialement détectées dans le réseau : dans ce cas, nous définirons sa pertinence Rel(Ci) comme

étant la somme des pertinences des cliques initiales dont elle est la fusion. Le poids relatif du $i^{ème}$ sens du terme T, représenté par la clique Ci, aura pour valeur le rapport entre la pertinence de la clique Ci et la somme des pertinences de l'ensemble des cliques représentant les différents sens du terme T. Ce poids relatif peut s'écrire :

$$PR(Ci) = Rel(Ci) / \sum_{j=1,n} Rel(Cj)$$

où Rel(C) correspond à la pertinence du sens représenté par la clique C et n est le nombre de sens que possède le terme T . La figure 3 montre l'arbre des usages nommés du terme *palais*, avec pour chacun de ses nœuds son poids relatif.

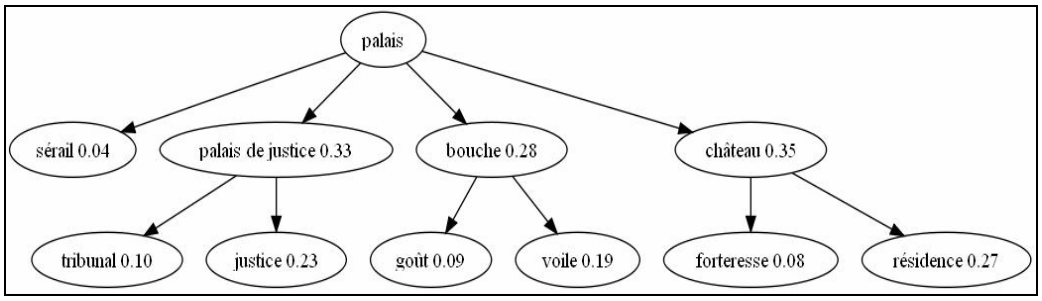

Figure 3 : Arbre des usages nommés du terme *palais* montrant le poids relatif de chaque clique.

5. Evaluation empirique des résultats

5.1. Principe de l'évaluation

Nous avons réalisé une évaluation sommaire sur près de 700 termes polysémiques dont les raffinements ont été validés par un expert. Les termes retenus pour cette évaluation sont tous des noms communs possédant au moins deux sens représentés par des cliques de pertinence supérieure à un certain seuil (50 pour cette évaluation).

L'évaluation des sens détectés dans notre réseau est quelque peu délicate : à notre connaissance, il n'existe pas une telle ressource, du moins pour la langue française. Nous avons donc fait appel à une trentaine de personnes, non expertes a priori, auxquelles il a été demandé quatre tâches différentes sur deux ensembles distincts d'une cinquantaine de termes chacun, un premier ensemble constitué de termes courants et un deuxième composé de termes plus rares. Sur certains ensembles de termes, les testeurs n'avaient pas le droit de consulter une quelconque ressource (tâches nommées Dict -), alors que pour d'autres ensembles ils avaient la possibilité de consulter un dictionnaire (tâches nommées Dict +) ; nous leur avons suggéré le dictionnaire en ligne Wiktionnaire, mais n'importe quel dictionnaire pouvait être utilisé. Par exemple, il était demandé aux testeurs :

Pour le terme palais*, nous proposons les sens suivants (par ordre de pertinence décroissante) :*

- *palais (château)*
- *palais (palais de justice)*
- *palais (bouche)*

Pensez-vous que certains sens soient absents de cette liste[5]*, et si oui combien ? Pensez-vous que parmi les sens proposés certains soient inappropriés, et si oui combien ?*

Les évaluateurs interrogés ont un niveau d'études d'approximativement Bac + 2 : nous pensons que cela correspond au profil moyen des joueurs de JDM. Les tableaux suivants synthétisent les résultats obtenus.

[5] Le sens *palais (sérail)* n'était pas proposé dans cette liste, sa pertinence égale à 17 dans notre réseau est inférieure au seuil de 50.

Mots courants	**Dict -**	**Dict +**
Sens absents	0.45	1.52
Sens inappropriés	0.66	0.37

Mots peu courants	**Dict -**	**Dict +**
Sens absents	0.25	1.67
Sens inappropriés	0.76	0.28

Tableaux 1 et 2 : Résultats de l'évaluation sans (Dict -) et avec (Dict +) utilisation d'un dictionnaire.

5.2. Interprétation des résultats

Sans l'aide de dictionnaire (Dict -), il y a systématiquement moins d'un sens d'écart en plus ou en moins. Dans les évaluations avec l'aide de dictionnaire (Dict +), les sens absents sont plus nombreux, alors que les sens inappropriés sont moins nombreux. Il semble que globalement il manque plus de sens dans notre réseau lexical que nous n'introduisons de sens erronés. Les sens absents dans notre réseau sont soit des sens très peu connus, soit des sens très spécifiques (exemples : *apparence, prétexte* pour le terme *voile* ou *absence de contraste* en parlant du *voile d'une photographie*). Les sens jugés inappropriés supplémentaires dans notre réseau diminuent lorsque les évaluateurs utilisent un dictionnaire : ils correspondent selon leurs dires à des sens particuliers qui leur étaient inconnus (exemple : *sapin(fiacre)*). Les sens supplémentaires résiduels (tâche Dict +), inférieurs à 0,4, correspondent le plus souvent à des champs lexicaux forts (exemple : *Noël* comme sens du terme *sapin*).

6. Conclusion

A partir de l'arbre des usages d'un terme, construit grâce au réseau lexical obtenu par JDM, nous avons pu en déterminer les sens. Ceci nous a permis de créer des nœuds de raffinement et ainsi d'établir de nouvelles relations entre ces nœuds dans le réseau. L'évaluation réalisée a montré que les résultats obtenus sont dans l'ensemble corrects si l'on tient compte du fait que le réseau lexical construit correspond à des connaissances relatives à une culture générale commune ; il ne s'agissait pas ici de rivaliser avec les dictionnaires traditionnels qui indiquent également des sens peu usités auxquels les joueurs de JDM ne pensent pas spontanément. Notre réseau lexical constitue une sorte de dictionnaire électronique généraliste, utilisable pour l'enseignement, l'aide à la rédaction …

En outre, nous avons remarqué que les dictionnaires traditionnels proposent généralement une hiérarchisation des sens. Par exemple, pour le terme *bouteille*, notre réseau propose les sens *bouteille en verre*, *bouteille de gaz* et *bouteille de plongée* ; dans un dictionnaire traditionnel, il est généralement indiqué le sens de *récipient* qui se subdivise en trois sous sens (qui sont effectivement ceux que nous proposons). Ce n'est pas (pas encore) le cas avec la méthode proposée ici. A suivre…

Références

COLLINS A., QUILLIAN M.R. (1969) Retrieval time from semantic memory, *Journal of verbal learning and verbal behaviour*, 8 (2), pp. 240-248

FAIRON C., HO N.D. (2004) Quantité d'information échangée : une nouvelle mesure de la similarité des mots, *Journées internationales d'Analyses statistiques des Données Textuelles (JADT'04)*, Louvain-la-Neuve (Belgique)

LAFOURCADE M., JOUBERT A. (2009) Similitude entre les sens d'usage d'un terme dans un réseau lexical, *Traitement Automatique des Langues*, vol.50/1, pp. 177-200

LAFOURCADE M., JOUBERT A. (2010) Construction de l'arbre des usages nommés d'un terme dans un réseau lexical évolutif, *Journées internationales d'Analyses statistiques des Données Textuelles (JADT'10)*, Rome (à paraître)

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

MEL'CUK I.A., CLAS A., POLGUERE A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot AUPELF-UREF

PLOUX S., VICTORRI B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement Automatique des Langues*, vol.39/1, pp. 161-182

POLGUERE 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.

TVERSKY A. (1977) Features of similarity, *Psychological Review*, 84, pp. 327-352

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