Actes de l'atelier « Réseaux Lexicaux et Traitement des Langues Naturelles.

Michael Zock, Gemma Bel-Enguix et Reinhard Rapp
LIF, Aix Marseille Université, Marseille, France
michael.zock@lif.univ-marseille.fr, gemma.belenguix@gmail.com, reinhardrapp@gmx.de

Préface

1 Présentation du champ

La façon dont nous regardons les *unités lexicales*, leur organisation et utilisation a radicalement changé ces dernières décennies. Décrites dans des dictionnaires et considérées comme des annexes de la grammaire dans les années 80, nous les considérons aujourd'hui comme de la matière première en TAL. Si à l'époque on utilisait encore le terme 'dictionnaire', on parle aujourd'hui plutôt de 'ressource lexicale' pour souligner le fait que les données lexicales sont exploitables par la machine et qu'elles sont annotées et organisées différemment selon leurs finalités (lexiques, dictionnaires, thesaurus, ontologies ; ...). Il y a désormais un très grand nombre de ressources lexicales (WordNet et ses nombreux descendants, puis, FrameNet, VerbNet, PropBank ; ...), ressources que l'on a essayé de standardiser (http://en.wikipedia.org/wiki/UBY-LMF), de lier entre elles (http://verbs.colorado.edu/semlink/) ou de lier à des encyclopédies comme Wikipédia (BabelNet, http://en.wikipedia.org/wiki/BabelNet).

Au début de l'histoire des dictionnaires électroniques, on a essayé de construire les ressources (automatiquement) à partir des dictionnaires imprimés (Ide & Véronis, sites.univ-provence.fr/veronis/publis.html). Cependant, on a vite rencontré des problèmes à cause de la pauvreté de la source. Les informations contenues dans les dictionnaires papier étaient insuffisantes pour permettre ensuite une exploitation convenable par la machine (génération et analyse automatique de textes). Étant donné que le but principal était d'exploiter la ressource au moyen de la machine, et que l'on avait désormais accès à de vastes corpus, on s'est efforcé de construire des ressources contenant des informations plus riches, plus explicites et mieux structurées. Concernant ce dernier point, WordNet (WN) a joué un rôle capital. Paradoxalement WN a eu davantage de succès en TAL qu'il n'en a eu auprès des utilisateurs consultant la ressource (pour chercher des mots), ou auprès des psycholinguistes étudiant le lexique mental. Ceci dit, WN a eu un effet incontestable au niveau théorique. Il a profondément modifié notre manière de voir la structure des ressources lexicales. Dorénavant, on ne les considère plus comme des listes plates de mots, ou comme des listes structurées alphabétiquement (dictionnaire papier), mais plutôt comme des graphes (réseaux lexicaux) dont les noeuds sont les mots et liens les différents types de relations lexicales.

Parallèlement à l'évolution des ressources lexicales, on a pu observer une explosion de travaux portant sur les graphes (graphes complexes, phénomène 'petit monde', etc.). Ces derniers semblent se prêter à merveille à la modélisation de nombreux domaines (Barrat, 2008, Barabási, 2003) y compris la langue. En effet, il y a eu de nombreux travaux montrant leur pertinence pour capter le *sens* des mots et celui des phrases (Bieman, 2012; Mihalcea et Radev, 2011; Widdows, 2004; Sowa, 1991) ou pour modéliser divers aspects du « monde » lexical : *structures associatives* (Schvaneveldt, 1989, Nelson et al., 1998), *structure* du dictionnaire (Gaume et al. 2008), *densité lexicale*, *distance moyenne* entre les mots (Vitevitch, 2008), *accessibilité* (Ferrer i Cancho & Sole, 2001), *aspects dynamiques* des graphes (Dion, 2012), etc.

Les graphes sont essentiellement une forme de représentation mathématique et visuelle des relations entre des objets/entités. C'est une forme de langage. Les objets (noeuds) et les liens peuvent être de nature très différentes (pour ne pas dire, quelconque) et leur poids ou direction peuvent être variables (uni-/bi-directionnel). Par exemple, le noeuds peuvent être des *personnes* (réseaux sociaux), des *lieux* (stations, villes, pays), des *objets* (astres, galaxies) ou des unités de la langue. Dans ce dernier cas, les graphes permettent de représenter des informations de différentes nature à différents niveaux :

- le *sens des mots* (graphes définitions) ;
- le *sens de la phrase* (relations entre les mots formant une phrase : réseaux sémantique, graphes conceptuels) ;

- le *sens du texte* (la relation de phrases ou leur organisation pour former un texte,) ;
- l'*organisation* des *mots* dans le *lexique* mental au niveau micro- et macro-structurel, liant soit le sens à la forme, soit les mots entre eux (réseaux lexicaux, réseaux associatifs).

Dans tous ces cas, nous avons recours au même formalisme, seule la nature des liens et celle des objets liés (noeuds) sont différentes.

Il y a donc deux grandes familles de chercheurs s'intéressant à des aspects complémentaires. Les uns s'intéressent à des données concrètes comme les *lemmes*, et les autres s'intéressent à la représentation de leur organisation (topologie) sous forme abstraite comme les *graphes*. C'est pour encourager l'échange d'idées entre ces deux mondes (les acteurs du monde TAL engagés dans la construction de ressources et les théoriciens des réseaux) que nous organisons cet atelier.

Se pose ensuite le problème de savoir comment se servir de ces graphes en TAL, ou comment se servir du TAL pour construire ce type de graphes. On pourrait également chercher à savoir comment l'un ou l'autre pourraient assister un être humain pour traiter la langue (accès lexical en production). Dans ce dernier cas, le TAL serait au service de l'être humain. On fait du TAL pour permettre du TIL (traitement interactif de la langue). Bien que très utile et tout à fait possible, cette dernière possibilité est rarement envisagée. Considérant cet aspect du traitement de la langue comme non pertinent on le laisse de côté, ce qui, vu son importance, est vraiment dommage. Peut-être cette rencontre est-elle une occasion d'y remédier.

2 Thèmes

Pour organiser cet atelier nous avions sollicité des soumissions portant sur l'ensemble des thèmes évoqués ci-dessus et plus particulièrement sur :

- l'*origine des données* permettant la construction des ressources : corpus, web, blogs, courriels, êtres humains (liste d'associations) ;
- la *méthode de construction* de la ressource: automatique, semi-automatique, collaborative (par des jeux) ;
- la *structuration des données* : alphabétique, thématique, liens sémantiques, liens associatifs ;
- la *caractérisation topologique du dictionnaire mental* (distribution, densité relative) et de son *évolution* ;
- *les facteurs affectant le poids des liens* ou des *noeuds* (aspects dynamiques des graphes) : fréquence, saillance, récence, changement de thème, etc. ;,
- l'*exploitation* ou *utilisation* de la ressource (ou d'une de ces transformations) : transformation du graphe en arbre pour assister l'accès lexical (navigation) ;
- l'*apprentissage automatique de liens* (repérage de relations sémantiques) ;
- la *visualisation* et *manipulation des graphes* (traduction en arbre, clustering, calcul de similarité sémantique) ;
- les *propriétés mathématiques* des réseaux lexicaux et l'*accessibilité des mots grâce à ces caractéristiques* (phénomène du 'petit monde') ;
- la *modélisation* des *variations linguistiques* et des *changements* de la langue (évolution du lexique).

3 Présentation des articles

Les articles retenus traitent les aspects suivants : désambigüisation lexicale, similarité structurelle entre des réseaux lexicaux, amélioration de navigation dans des ressources comme WordNet, facteurs socio-linguistiques affectant l'évolution d'une langue, accès lexical en mode production.

Gilles Sérasset (conférencier invité): *Réseaux Lexicaux, Traitement des Langues, et Données Liées Ouvertes*

S'appuyant sur les travaux réalisés dans le cadre des projets Papillon, LexALP et DBnary, l'auteur cherche à montrer en quoi, l'utilisation du format des donnés liées ouvertes, est logiquement l'étape suivante dans notre étude du lexique.

Laroussi Merhbene, Anis Zouaghi et Mounir Zrigui : *Approche basée sur les arbres sémantiques pour la désambiguïsation lexicale de la langue arabe en utilisant une procédure de vote*

Les auteurs proposent une approche semi-supervisée de désambigüisation lexicale des mots arabes. La partie supervisée a pour but de classifier les contextes des mots ambigus en tenant compte de leur sens. La partie non supervisée utilise la notion de vote pour classifier les mesures de collocations et pour choisir le sens convenable.

Bruno Gaume, Emmanuel Navarro, Yann Desalle et Benoît Gaillard : *Mesurer la similarité structurelle entre réseaux lexicaux*

L'objectif de ce travail est de comparer la structure topologique de différents réseaux lexicaux en utilisant la méthode des marches aléatoires. Au lieu de caractériser les paires de sommets selon un critère binaire de connectivité, les auteurs mesurent leur proximité structurelle par la probabilité relative d'atteindre un sommet à partir d'un autre. Comme cette méthode permet de rapprocher les sommets d'une même zone dense en arêtes, elle permet par la même occasion de comparer la structure topologique des réseaux lexicaux.

Guy Lapalme : *WordNet en XML-HTML*

L'auteur présente une version XML de WordNet permettant une consultation plus facile par l'être humain ou la machine que la version originale. Partant des fichiers XML ont peut générer des fichiers HTML ce qui permet d'explorer les synsets avec un simple navigateur internet. Un 'démonstrateur' en Java illustre la facilité d'accès à l'information en XML pour diverses applications de TAL.

Gemma Bel-Enguix : *Linguistic Convergence in Societies with Asymmetrically Distributed Reputation*

L'auteur essaie de modéliser l'évolution d'une langue, par exemple, l'évolution du sens de mots, en jouant sur plusieurs paramètres socio-linguistiques. Ce type de recherche permet de simuler l'importance des structures sociales sur l'évolution d'une langue ou le changement d'une structure linguistique particulière.

Michael Zock et Didier Schwab : *Stocker des Mots ne Garantit nullement leur Accès*.

Les auteurs tentent de montrer (a) que mémoriser une forme lexicale ne garantit nullement son accès et (b) comment construire une aide navigationnelle permettant à un auteur (locuteur, rédacteur) de trouver le mot bloqué sur le bout de sa langue (ou de sa plume), car, si les dictionnaires sont relativement bien faits pour les récepteurs (lecteurs), ils ne sont pas toujours à la hauteur des attentes des producteurs (problèmes d'entrée, problèmes de navigation).

4 Conclusion

Vu le dynamisme du domaine où de 'nouvelles' théories comme les *méthodes vectorielles* (Widdows, 2004, Vitevitch, 2008), la *sémantique distributionnelle* (Sahlgren, 2008), et la *mémoire distributionelle* (Baroni et Lenci, 2010) etc., ont vu le jour et se sont généralisées, et vu le vivier du monde francophone travaillant sur les ressources lexicales nous étions très surpris du faible nombre de soumissions. Il n'est pas facile de savoir ce qui a pu causer ce 'silence', car il contraste énormément avec le succès d'un autre événement, consacré à des problèmes très voisins : CogALex (http://pageperso.lif.univ-mrs.fr/~michael.zock/CogALex-IV/cogalex-webpage/index.html). Il est vrai qu'étant lié à une conférence majeure de notre discipline, Coling, cet atelier attire naturellement un bien plus grand nombre de collègues, d'autant plus qu'il contient une tâche partagée consacrée à un des grands défis de la lexicographie informatique, la navigation dans une ressource lexicale afin de trouver le mot que l'on a sur le bout de la langue, mot qui est stocké dans la ressource, sans que l'on puisse nécessairement le localiser pour autant.

Références

BARABÁSI, A.-L. (2003). Linked: How Everything is Connected to Everything Else and What It Means for Business, Science, and Everyday Life. Plume

BARONI, M. et A. LENCI. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics* **36** (4): 673-721.

BARRAT, A. et al. (2008). Dynamical Processes on Complex Networks, Oxford University Press

BIEMANN, C. (2012). Structure Discovery in Natural Language . Theory and Applications of Natural Language Processing. Springer Berlin / Heidelberg.

DION, D. (2012). Dynamiques d'évolution de graphes de cooccurrences lexicales. Thèse de doctorat, Bordeaux.

FERRER i CANCHO, R., et SOLE, R. V. (2001). The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences, 268, 2261–2265.

GAUME, B., DUVIGNAU, K., PREVOT, L. et DESALLE, Y. (2008). Toward a cognitive organization for electronic dictionaries, the case for semantic proxemy. Cogalex-1, Coling, Manchester

MIHALCEA, R. et RADEV, D. (2011) Graph-based natural language processing and information retrieval. Cambridge University Press, Cambridge,

NELSON, D., McEVOY, C. & SCHREIBER, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/

SAHLGREN, M. (2008). The Distributional Hypothesis. *Rivista di Linguistica* 20 (1): 33–53.

SCHVANEVELDT, R. editor. (1989). Pathfinder Associative Networks: studies in knowledge organization. Norwood. N.J.

SOWA, J. (1991) Principles of Semantic Networks: Explorations in the Representation of Knowledge, edited by J. F. Sowa, Morgan Kaufmann Publishers, San Mateo, CA

VITEVITCH, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? Journal of Speech, Language, and Hearing Research, 51, 408–422.

WIDDOWS, D. (2004). Geometry and Meaning. Stanford, CA: CSLI. (http://www.puttypeg.net/book/)

5 Membres du Comité de Programme

Cristea, Dan	(University A.I.Cuza, Iasi, Romania)
Ferrer i Cancho, Ramon	(LARCA, université polytechnique de Catalogne, Barcelone, Espagne)
Ferret, Olivier	(CEA LIST, Gif sur Yvette, France)
Francopoulo, Gil	(Tagmatica, Paris, France)
Grefenstette, Gregory	(INRIA, Saclay, France)
Lenci, Alessandro	(Université de Pise, Italie)
L'Homme, Marie-Claude	(Université de Montréal, Canada)
Ploux, Sabine	(L2C2, Institut des Sciences Cognitives, Lyon, France)
Prévot, Laurent	(LPL, Université Aix Marseille, Aix en Provence)
Schwab, Didier	(LIG-GETALP, Grenoble, France)
Sérasset, Gilles	(LIG, Grenoble, France)