Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences

Benjamin Hervy, Matthieu Quantin, Pierre Teissier

Extraction et chaînage supervisés de connaissances d'un corpus d'entretiens en histoire des sciences

Benjamin Hervy**, Matthieu Quantin*,***, Pierre Teissier***

*École Centrale de Nantes, IRCCyN UMR CNRS 6597
prenom.nom@irccyn.ec-nantes.fr
**Université d'Angers, CERHIO UMR CNRS 6258
prenom.nom@univ-angers.fr
***Université de Nantes, Centre François Viète EA 1161
prenom.nom@univ-nantes.fr

1 Introduction

Les données des sciences de l'homme forment souvent des corpus de textes, qui sont hétérogènes par leurs forme et contenus ; spécifiques par leurs terminologie et signification. À partir d'un corpus d'entretiens en histoire des sciences, nous présentons une méthode supervisée générant un réseau de documents liés par leurs proximités de contenus. Il s'agit d'un graphe multiple flou, basé sur l'extraction de *n-grams* à taille variable.

1.1 Présentation du corpus

Le corpus est formé par la retranscription de 37 entretiens de chercheurs racontant leur carrière parfois depuis les années 1940. Le corpus global contient 293k mots et a été étudié manuellement (thèse, articles, livres) par Teissier (2007). Il entrecroise ainsi des questions techniques (recherche), des énoncés relationnels et affectifs (interpersonnels) et des positionnements identitaires (discipline, génération, genre, etc.).

1.2 Objectifs et hypothèses

L'approche proposée vise à mettre en évidence de nouveaux éléments de réflexion pour l'historien via la création d'un graphe de co-occurrences d'expressions entre documents.

Nous faisons l'hypothèse qu'il est préférable que l'automatisation génère des inférences bas-niveau s'articulant avec des inférences (manuelles) qualitatives haut-niveau de l'historien. L'objectif de la méthode proposée est de favoriser cette articulation pour saisir les structures et les dynamiques de communautés scientifiques.

L'hyper-spécialisation du corpus étudié participe de la richesse de l'analyse historique. La méthode numérique devra préserver cette richesse terminologique. Les connaissances extraites sont issues du corpus ou des choix de l'historien et non d'un modèle défini *a priori*. Dans cette optique, nous favorisons le *rappel* à la *précision* en produisant des indicateurs pour la supervision de l'historien.

2 Méthode d'analyse

2.1 Extraction de *multi-word expression* (MWE)

L'extraction se base sur l'algorithme ANA (Enguehard et Pantera, 1995) pour construire des MWE à partir des termes du corpus sans entraînement ni pré-traitement. Exemples de termes extraits : "Bronzes de vanadium", "Microscopie électronique à transmission à haute résolution".

2.2 Création des liens du graphe

Un arc pondéré est généré par chaque co-occurrence de MWE entre deux documents. Le calcul du poids d'un arc est basé sur :

1. le calcul du poids des MWE, adaptation d'**idf** (Salton et al., 1973) avec cosinus pour créer un effet de seuil : $poids = \left(\cos \frac{2-in\_docs}{nb\_docs}\right)^{factor_1}$ Ici, $factor_1 = 100$ empiriquement. $nb\_docs = 37$. in_docs : nombre de documents où la MWE apparaît.
2. le calcul de la proximité entre deux documents sur une MWE, adaptation de **tf** (Salton et al., 1973). Le minimum d'occurrences favorise l'équi-répartition des termes entre deux documents : $proximite = log_{10}((O^{t_i}_{p_1} + O^{t_i}_{p_2}) \times min(O^{t_i}_{p_1}, O^{t_i}_{p_2})^{factor_2})$ Ici, $factor_2 = 3$ et $O^{t_i}_{p_j}$: nombre d'occurrences du MWE t_i dans le document p_j.

La pondération de l'arc est obtenue par $poids \times proximite$ normalisée entre 0 et 1.

3 Conclusion et perspectives

Les résultats de la méthode numérique trouvent une répercussion dans l'état des connaissances de l'historien. Au delà d'illustrer des connaissances existantes, un "dialogue" heuristique s'instaure entre historien et analyse numérique. Premièrement, l'interprétation de relations "surprenantes" permet de braquer le regard sur un angle mort ou d'ouvrir une voie non explorée. Deuxièmement, le tracé de représentations numériques initié par des questionnements historiques met en évidence des réseaux de nœuds et des clusters inédits, qui peuvent renouveler les interprétations historiques ou, au contraire, s'avérer dénuées de sens.

Références

Enguehard, C. et L. Pantera (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics 2*(1).

Salton, G., C. S. Yang, et C. T. Yu (1973). Contribution to the Theory of Indexing. Technical report, Department of computer science, Cornell University, Ithaca, NY, USA.

Teissier, P. (2007). *L'émergence de la chimie du solide en France (1950-2000) : de la formation d'une communauté à sa dispersion.* Ph. D. thesis, Université Paris Ouest Nanterre.