Traitement Automatique des Langues, Biens Communs Informationnels et Industries de la Langue

Gaël de Chalendar
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Saclay, F-91191, France;
gael.de-chalendar@cea.fr

Résumé

Cet article étudie les interactions entre le domaine du Traitement Automatique des Langues et le concept de Biens Communs Informationnels. Il montre combien il est important que les sociétés des Industries de la Langue participent à leur développement.

1. Introduction

Comme dans tous les domaines, il y a aujourd'hui une grande diversité de logiciels de Traitement Automatique des Langues (TAL) publiés sous licence libre. Pour autant, certains, peut-être surtout dans le monde industriel, continuent de penser que s'ils sont gratuits, alors ces logiciels doivent être de moins bonne qualité. D'autres pensent que si leur code source est mis à disposition de tous, alors ils ne pourront pas assurer une activité commerciale rentable autour d'un tel logiciel. Nous voudrions dans cet article suggérer en quoi ces présuppositions sont fausses. Au contraire, nous verrons comment le logiciel libre permet d'améliorer rapidement la qualité et la fiabilité des logiciels de TAL. Nous évoquerons aussi divers modèles commerciaux (business models) permettant un essor économique autour de la liberté du code. Nous argumenterons sur le fait que les progrès scientifiques et techniques se trouvent accélérés par l'ouverture du code source, pour l'ensemble de la communauté et par conséquent pour les acteurs individuels.

2. Les Biens Communs Informationnels

Inutile d'en appeler à Dieu comme Bien commun ultime à l'image de St Thomas d'Aquin (Modde, 1949) pour justifier moralement l'existence et le développement des Biens Communs informationnels. Philippe Aigrain [1] définit dans son glossaire les Biens Communs comme "toute « chose » ou entité immatérielle à laquelle on a décidé de donner un statut de propriété commune, de la faire appartenir à tous, parce qu'elle n'appartient à personne". Les Biens Communs Informationnels (BCI) sont les "bien communs qui peuvent être créés, échangés et manipulés sous forme d'information, et dont les outils de création et le traitement sont souvent eux-mêmes informationnels (logiciels). Il peut s'agir de données, de connaissances, de créations dans tous les médias, d'idées, de logiciels". Le coût marginal de la duplication des BCI est quasiment nul. De plus, "leur valeur n'est pas réduite par le fait qu'un autre s'en serve, mais plus souvent augmentée par la faculté d'échange ou de communication accrue".

P. Aigrain décrit comment, dans son article de 1968 "The tragedy of the commons", l'économiste Garrett Hardin, avait porté un grand coup à l'idée de Communs en montrant "la fragilité des biens communs physiques en présence d'un usage croissant résultant de la démographie, de la quête du profit ou de l'éro- sion d'un système de valeurs communes". Seule alors une économie administrée ou la propriété individuelle permettrait d'éviter ces écueils.

Or Elinor Ostrom, prix Nobel d'économie 2009, et Charlotte Hess, dans leur ouvrage « Understanding knowledge as a commons » réduisent en poudre ce modèle [2]. Pour elles, le modèle de Hardin ne ressemble aucunement aux communs réels, tels qu'ils sont gérés collectivement depuis des millénaires. Pour Hardin, les communs sont uniquement des ressources disponibles, alors qu'en réalité ils sont avant tout des lieux de négociations (il n'y a pas de communs sans communauté), gérés par des individus qui communiquent, et parmi lesquels une partie au moins n'est pas guidée par un intérêt immédiat, mais par un sens collectif. De plus, et concernant les BCI, il existe une différence majeure entre ces Communs de la connaissance et les Communs naturels : les biens numériques ne sont plus soustractibles. L'usage par l'un ne remet nullement en cause l'usage par l'autre, car la reproduction d'un bien numérique a un coût marginal qui tend vers zéro.

Les logiciels sont un parfait exemple d'entités pouvant être considérées comme des Biens Communs Informationnels. Cela se réalise sous la forme des logiciels libres, diffusés selon une licence assurant le maintien des 4 libertés (utiliser, copier, étudier et modifier et redistribuer le logiciel).

1. Les éléments de définition et de réflexion autour des Biens Communs Informationnels évoqués dans cet article proviennent en grande partie du livre Cause Commune de Philippe Aigrain (Aigrain, 2005).

2. Ce paragraphe, dans l'esprit de la licence Creative Commons, est réécrit à partir de (Le Crosnier, 2009), texte diffusé sous licence Creative commons v3 – attribution. À notre avis, cette licence permet et encourage même une telle réutilisation d'un texte. Suite au doute soulevé par un relecteur de la première version de notre article, nous avons demandé son avis à l'auteur initial qui a confirmé notre interprétation.

3. TAL et Biens Communs

Le Traitement Automatique des Langues a particulièrement besoin de logiciels libres. Les langues, prises toutes ensemble, sont le langage naturel, ce que l'on désignera ici par "le" langage. Le langage, donc, est, autant que l'eau, l'air ou les connaissances scientifiques un Bien Commun. Il est indispensable que les technologies qui sont au coeur d'applications fondamentales pour les libertés publiques (réseaux sociaux, ciblage commercial, surveillance par les autorités civiles et militaires) existent sous une forme open source, permettant au public d'en avoir connaissance. Cela ne signifie pas que tout code soit ouvert, mais que toute fonctionnalité ait la possibilité d'être implémentée sans être bloquée par des interdictions liées à une interprétation restrictive du droit d'auteur.

La mise à disposition sous licence libre de nombreux résultats de travaux permet aussi à notre avis d'accélérer le développement des connaissances scientifiques, même si nous n'en avons pas de preuve. Par exemple, la plupart des approches statistiques de ces dernières années (SVM, CRF, Deep Learning, etc.) ont connu des implémentations libres permettant à l'ensemble des chercheurs de se les approprier et de les faire évoluer. De plus, de plus en plus de chercheurs mettent aussi à disposition leurs données d'apprentissage et d'évaluation, ce qui permet d'une part de mettre en œuvre pour de vrai le principe de reproductibilité de la méthode scientifique mais aussi de poursuivre les travaux en question dans diverses directions.

3.1. Logiciels Libres

La nécessité de fournir à l'ensemble de la communauté des outils libres a été bien comprise par beaucoup de chercheurs en TAL et les outils sous licence libre sont très nombreux, avec des niveaux de performance et d'industrialisation très variés. Parmi les outils internationaux, certains des plus notables sont NLTK (Bird et al., 2009) [3], GATE (Cunningham et al., 2002) [4], ou FreeLing (Padró and Stanilovsky, 2012) [5]. En France, on trouve l'ensemble des outils de l'équipe INRIA Alpage [6] ou notre propre analyseur multilingue, LIMA, Libre Multilingual Analyzer [7] (de Chalendar, 2014). Nous verrons en section 4. les modèles économiques mis en œvre autour de ces logiciels.

3.2. Ressources linguistiques

La situation du côté des ressources linguistiques est contrastée. Oon trouve, au moins pour les langues bien dotées, de nombreuses ressources lexicales ou sémantiques. Par exemple en français les lexiques Lefff, Morphalou ou Glaff et les WordNets Wolf et WoNef. On trouve aussi un assez grand nombre de corpus librement téléchargeables (l'extrait du Penn Treebank de NLTK par exemple). Par contre, l'apparition de corpus annotés vraiment libres, donc autorisant la modification et la redistribution des textes annotés, est beaucoup plus récente. Le nombre de langues disposant de telles ressources n'est pas très élevé et les ressources sont encore assez limitées en terme de taille, de qualité et de variété. En se limitant à des corpus annotés en morphosyntaxe, aux États-Unis, les auteurs de l'American National Corpus ont décidé depuis 2006 ((Ide, 2008)) de n'offrir plus que des textes librement diffusables pour donner l'Open American National Corpus [8].

En France, le corpus Sequoia (M. and D., 2012) est libre et annoté avec les meilleurs standards de qualité mais reste de petite taille. Le Free French Treebank (Hernandez and Boudin, 2013) est beaucoup plus grand mais avec une qualité d'annotation encore inférieure puisque annoté automatiquement.

4. Biens Communs Informationnels et Industries de la Langue

On peut lire que "les communs sont avant tout des lieux de négociations (il n'y a pas de communs sans communauté), gérés par des individus qui communiquent, et parmi lesquels une partie au moins n'est pas guidée par un intérêt immédiat, mais par un sens collectif." [9]. Et pourtant, dans une société organisée autour de l'entreprise individuelle, ou plutôt de l'entreprise à capitaux privés, qu'ils soient individuels ou eux-mêmes capitalistiques, il faut pouvoir justifier à ce niveau-là de l'utilisation de tels biens communs. Il faut démontrer leur efficacité économique. Même s'ils sont un idéal, ils doivent aussi être à l'origine de mécanismes efficaces pour la création de richesse au niveau de l'entreprise.

Étonnamment, malgré tous les modèles économiques réussis autour des logiciels libres (double licence, services, free/premium, etc.), on trouve assez peu d'outils de TAL dans ce cadre. Nous allons décrire les modèles adoptés par quelques outils de TAL, trois plateformes génériques (GATE, FreeLing et LIMA) et l'ensemble d'outils de l'INRIA Alpage. Nous remercions les personnes de ces divers projets qui ont bien voulu nous transmettre les éléments ci-dessous. Nous finirons avec quelques mots sur un certain nombre d'autres outils.

4.1. GATE, de l'université de Sheffield

GATE, déjà cité, est mis à disposition sous licence LGPL, permettant donc tous les usages, même commerciaux, sans contrepartie financière ni contribution au projet. Les auteurs proposent des services autour de sa mise en œuvre, comme des possibilités de customisation, des formations ou des développements à façon. Ils proposent aussi des programmes de sponsoring et des partenariats privilégiés permettant un financement réccurrent de leurs activités. Leurs clients sont entre autres (et en laissant de côté certains gros clients sous accord de confidentialité) le UK National Health Service, la UK Food and Environment Research Agency, Public Health England et de nombreuses PMEs et statups. Bien sur, il existe aussi de nombreux utilisateurs "anonymes" ou qui passent par l'intermédiaires d'autres fournisseurs de services.

3. http://www.nltk.org/
4. https://gate.ac.uk/
5. http://goo.gl/0EqmE1
6. http://goo.gl/XDodNo
7. https://github.com/aymara/lima/wiki
8. http://www.anc.org/
9. http://goo.gl/DgrnYQ

4.2. FreeLing, Université Polytechnique de Catalogne

FreeLing est pour le moment distribué sous double licence commerciale et GPL. Mais de nombreux utilisateurs outrepassent l'esprit de la GPL en fournissant des services commerciaux autour de FreeLing caché derrière un service Web (en mode SaaS, software as a service). On peut citer par exemple SentiLecto [10], un moteur d'analyse d'opinion ou Apicultur [11], des APIs Web basées sur freeling. Par conséquent, les auteurs ont décidé de fournir les prochaines versions sous licence Affero GPL (AGPL) qui impose les mêmes devoirs que la GPL (en particulier l'obligation de fournir le code source sous une licence compatible) même pour les outils mis à disposition à travers des applications Web (sites ou APIs).
Le modèle double licence pour l'intégration dans des logiciels propriétaires est mis en œuvre par exemple avec Ruby Reader, une application iPhone aidant des locuteurs japonais à comprendre des textes en anglais ou MediaTuners, un assistant automobile contrôlé par la voix (musique, appels, agenda, navigation, etc.).
Enfin, les auteurs sont en cours de mise en place en interne d'une offre SaaS de freeling et d'autres outils de TAL.

4.3. Outils de l'INRIA Alpage

La plupart des outils de l'équipe Alpage (INRIA, Université Paris-Diderot, Paris 7) sont distribués sous licences LGPL ou Cecill. À côté de son financement académique et de projets collaboratifs, l'équipe entretient des collaborations avec divers industriels qui financent ponctuellement le développement de nouvelles fonctionnalités ou ressources ou leur amélioration. On peut citer par exemple :

- Vera : traitement multilingue de parties libres de formulaires ;
- Kwaga : traitement de mails ;
- Viavoo : traitement d'avis de consommateurs ;
- Lingua et Machina : extraction de terminologie et acquisition de réseaux lexicaux dans des domaines spécialisés.

4.4. LIMA, CEA LIST

Nous avons, au CEA LIST, placé en janvier 2014 notre analyseur linguistique multilingue LIMA sous double licence AGPL et propriétaire, avec comme objectif la poursuite de l'amélioration de LIMA et sa plus large diffusion. Le choix de la licence AGPL a été dicté par l'objectif d'éviter des utilisations commerciales allant à l'encontre de l'esprit de la licence GPL, comme c'était le cas pour FreeLing (section 4.2.).
Nous finalisons en ce moment l'adaptation des ressources linguistiques libres choisies pour la version libre de LIMA. LIMA libre est disponible en français et en anglais. Nous avons pour le moment maintenu uniquement propriétaires les modules et ressources pour l'allemand, l'arabe, le chinois et l'espagnol.
La version commerciale de LIMA est utilisée par la société ANT'inno [12] au cœur de l'outil de veille ANT'box. Ses clients sont, entre autres, Eurocopter, Dipta, Inéris, Sodiaal, Cyrus pour les industriels. L'IRSN, le gouvernement Mauritanien et des Ministères au Maroc et au Cameroun pour les organismes gouvernementaux.
ANT'inno offrira à l'avenir, en collaboration avec le CEA LIST, un support et des services autour de la version libre de LIMA.

4.5. Autres exemples de business models du libre dans le TAL

Voici rapidement quelques autres exemples de mise en œuvre de business models autour de logiciels libres de TAL.
Le groupe Natural Language Processing du l'université de Stanford fournit certains de ses outils (PoS tagger, Parser, NER) sous licence GPL et indique qu'une licence commerciale est disponible.
La société aliasi développe sous double licence commerciale/AGPL l'outil LingPipe. Elle a pour clients Thomson, Endeca, NLM, DARPA, MITRE, etc.
RapidMiner était un logiciel libre sous AGPL mais il a changé de licence pour redevenir propriétaire (Business source). Toutefois, la version libre a été "forkée" au moins une fois [13].
Gensim est supporté par sa communauté et son auteur principal Radim Řehůřek propose un support commercial et des développements spécifiques par l'intermédiaire de sa société de consultant [14].
Enfin, dernier exemple, le projet Carrot2 (Open Source Search Results Clustering Engine) est intégré par la société Carrot Search dans son produit commercial Lingo3G de clustering de documents.
Plus généralement, à côté d'outils spécifiquement de TAL et exploités comme tels, un grand nombre d'éditeurs d'autres types d'outils libres utilisant des fonctionnalités de TAL intègrent d'autres outils libres et reversent à la communauté les améliorations qu'ils apportent (XWiki, Nuxeo, WebLab,...)

5. Conclusion

De par sa place centrale dans les activités humaines, le langage est à notre avis un Bien Commun Informationnel aussi important à préserver que les Biens Communs que sont l'eau ou l'air. Par conséquent, il est fondamental de disposer de logiciels de TAL et de ressources linguistiques libres pour toutes les langues. De façon à ne pas faire porter les efforts uniquement sur les développeurs individuels ou les chercheurs du domaine, il faut que des industriels s'engagent dans leur développement, leur exploitation et leur diffusion. Les modèles économiques le permettant sont nombreux.

6. Bibliographie

Aigrain, P. (2005). *Cause commune : l'information entre bien commun et propriete*. Transversales. Editions Fayard.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. Book about NLTK.

10. http://tecnolecto.com/sentilecto
11. http://store.apicultur.com/
12. http://www.antinno.fr
13. http://goo.gl/hyimAe
14. http://radimrehurek.com/gensim/

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE : A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL, 2002.*

de Chalendar, G. (2014). The LIMA Multilingual Analyzer Made Free : FLOSS Resources Adaptation and Correction. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2932–2937.

Hernandez, N. and Boudin, F. (2013). Construction d'un large corpus écrit libre annoté morpho-syntaxiquement en français. In *Actes de la conférence TALN-RECITAL 2013*, Sables d'Olonne, France, June.

Ide, N. (2008). The American National Corpus : Then, now and tomorrow. Keynote paper presented at the HCS-Net Workshop on Designing the Australian National Corpus, 4-5 December, UNSW, Sydney, Australia.

Le Crosnier, H. (2009). Une bonne nouvelle pour la théorie des Biens Communs, October.

M., C. and D., S. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN'2012*.

Modde, A. (1949). Le Bien Commun dans la philosophie de saint Thomas. *Revue Philosophique de Louvain*, 47(14) :221–247.

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0 : Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.