Communautés Internet comme sources de préterminologie

Mohammad Daoud, Christian Boitet

Laboratoire LIG — Université Joseph Fourier — 385, rue de la Bibliothèque, 38041 Grenoble, France
{Mohammad.Daoud, Christian.Boitet }@imag.fr

Résumé

Cet article décrit deux expériences sur la construction de ressources terminologiques multilingues (preterminologies) préliminaires, mais grandes, grâce à des communautés Internet, et s'appuie sur ces expériences pour cibler des données terminologiques plus raffinées venant de communautés Internet et d'applications Web 2.0. La première expérience est une passerelle de contribution pour le site Web de la Route de la Soie numérique (DSR). Les visiteurs contribuent en effet à un référentiel lexical multilingue dédié, pendant qu'ils visitent et lisent les livres archivés, parce qu'ils sont intéressés par le domaine et ont tendance à être polygottes. Nous avons recueilli 1400 contributions lexicales en 4 mois. La seconde expérience est basée sur le JeuxDeMots arabe, où les joueurs en ligne contribuent à un réseau lexical arabe. L'expérience a entraîné une croissance régulière du nombre de joueurs et de contributions, ces dernières contenant des termes absents et des mots de dialectes oraux.

Mots-clés: terminologie, préterminologie, approches collaboratives, réseaux lexicaux, DSR, jeux sérieux.

Abstract

This paper describes two experiments on building preliminary but large multilingual terminological resources (preterminologies) through Internet communities, and draws on these experiments to target more refined terminological data from Internet communities and Web 2.0 applications. The first experiment is a contribution gateway for the Digital Silk Road (DSR) website. Visitors indeed contribute to a dedicated multilingual lexical repository while they visit and read the archived books, because they are interested in the domain and tend to be multilingual. We collected 1400 lexical contributions in 4 months. The second experiment is based on the Arabic JeuxDeMots, where online players contribute to an Arabic lexical network. The experiment resulted in a steady growth of number of players and contributions, the latter containing absent terms and spoken dialectic words.

Keywords: terminology, preterminology, collaborative approaches, lexical networks, DSR, serious games.

1 Introduction

Construire des ressources terminologiques multilingue pour un domaine est une tâche difficile et compliquée, car la terminologie représente la structure conceptuelle d'un domaine en utilisant un ensemble d'unités lexicales dans une langue particulière. Cette structure conceptuelle est plus dynamique et change plus vite que sa représentation symbolique (terminologie). En outre, toutes les communautés de langues différentes ne partagent pas le même intérêt dans un domaine donné. Par exemple, l'arabe est considéré comme une langue pauvrement dotée en ressources linguistiques (Yassin, 2003) (Diab, Habash, 2009), en particulier dans les domaines de la science et de la technologie, tandis que l'anglais et le français ont une terminologie beaucoup plus riche dans ces domaines. Classiquement, pour construire des ressources terminologiques, une banque de termes est construite par des terminologues: le domaine considéré est étudié et des documents sont consultés pour en extraire les termes pertinents. Cette approche dépend fortement de terminologues et de financement par de grandes organisations, elle est donc très coûteuse.

De nombreux chercheurs ont été à la recherche d'alternatives (Kageura, Umino, 1998) (Joubert, Lafourcade, 2008) (Nagata et al., 2001). Les approches contributives (Ahn, 2005) à la collecte de connaissances semblent prometteuses à

beaucoup d'entre eux, mais ne sont pas pratiques, car elles nécessitent des bénévoles motivés qui sont capables de participer à un processus de construction d'une base lexicale. Nous avons proposé d'abaisser les attentes (1) en nous appuyant sur des volontaires en ligne et (2) en collectant de la préterminologie (Daoud et al., 2009) plutôt que de la terminologie. Nous avons essayé de motiver les internautes à contribuer par des interactions directes et indirectes avec une communauté Internet. Nous avons expérimenté cette solution avec le site Web de la Route de la Soie Numérique (Digital Silk Road), et le JeuxDeMots arabe. La section suivante montre brièvement l'approche contributive proposée. La section 3 décrit les résultats expérimentaux de l'approche directe de collaboration sur DSR. Et la section 4 présente l'expérience faite avec JeuDeMots arabe.

2 Projet d'expansion de ressources multilingues terminologiques

2.1 Objectif

La figure 1 montre une estimation simplifiée des ressources disponibles dans plusieurs langues. Nous essayons d'élargir la sphère terminologique, en trouvant la « terminologie cachée » à partir de la zone bleue (unités lexicales utilisées par les communautés comme des termes techniques).

Parce qu'il est difficile de produire une information terminologique complète à partir de la sphère lexicale, nous visons une ressource préliminaire appelée préterminologie (Daoud et al., 2009), et nous la représentons à l'aide d'une structure de graphe appelée MPG (Multilingual Preterminological Graph). Un nœud représente un préterme et un arc représente une relation non confirmée entre deux termes.

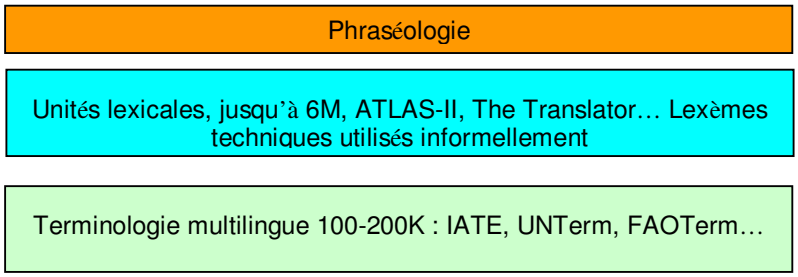

Figure 1: objectif principal (expansion de la sphère terminologique)

2.2 Ressources non classiques

Traditionnellement, une base de termes est construite par des terminologues, qui étudient le domaine, construisent sa taxonomie, consultent des documents pertinents pour extraire les termes (manuellement ou semi-automatiquement), puis traduisent les termes extraits. Le processus est difficile et influe sur la couverture de la base produite. Pour agrandir la sphère terminologique, nous proposons d'exploiter des ressources non conventionnelles, à partir de ce qui est disponible et ce qui peut être extrait à partir de ressources numériques, telles que des bases terminologiques, des corpus, des dictionnaires lisibles par machine (MRD), des systèmes de traduction automatique, des encyclopédies et des glossaires locaux.

A côté de cela, nous nous concentrons sur la communauté Internet comme une source de préterminologie multilingue. Comme les internautes interagissent avec des sites Web sur une base quotidienne, ces interactions peuvent conduire à des contributions. Nous classons ces interactions en:

- Explicites:
  - Interaction directe: PanImages (Etzioni et al. 2007)
  - Interaction indirecte: JeuxDeMots (www.lirmm.fr/jeuxdemots)
- Implicites:
  - Analyse de fichiers de log
  - Analyse du Web 2.0

L'initialisation du graphe préterminologique multilingue en utilisant des ressources numériques s'est révélée utile et essentielle. Toutefois, pour découvrir la terminologie cachée, l'intervention humaine est nécessaire. Comme les professionnels coûtent cher et ne sont pas disponibles, nous ciblons les contributions de bénévoles. Nous avons étudié

deux scénarios de contribution explicite (directe et indirecte), où les internautes sont activement impliqués dans le processus de contribution lexicale.

2.3 Communauté Internet du Web 2.0

La deuxième proposition d'intervention humaine dans la construction de banques de termes dépend des sites de partage et des réseaux sociaux. Cette participation n'est pas forcément explicite, ce qui signifie que les contributeurs peuvent ne pas être au courant du processus de contribution. En effet, leurs interactions sur Internet peuvent être analysées en arrière-plan pour en extraire des connaissances terminologiques multilingues.

Cette approche est inspirée par le fait que les sites Web de partage produisent un nouveau type de corpus, et que les communautés implicites qu'ils créent façonnent dynamiquement des domaines de connaissance et créent des centaines de termes chaque semaine. Les approches automatiques actuelles ne sont pas suffisantes pour gérer cette grande quantité de connaissances produites quotidiennement, et les approches collaboratives ne peuvent pas y arriver non plus. Par exemple, chaque semaine, plus de 2 millions de microblogs (tweets) sont soumis à twitter.com. Ces tweets contiennent de nombreux termes techniques, et certains d'entre eux sont marqués par un dièse (hashtag). Ces termes peuvent être retrouvés et liés les uns aux autres dans un MPG en analysant en les contributions en ligne et les messages aux sites de partage. De cette façon, les utilisateurs contribuent implicitement. Plus l'instant, nous n'avons expérimenté nos approches à la contribution active que sur des sites Web traditionnels.

3 Passerelle contributive au site DSR (Route de la Soie numérique)

3.1 Expérimentation

Le projet Digital Silk Road (Ono et al., 2008) est une initiative lancée par le National Institute of Informatics (Tokyo) et l'Organisation des Nations Unies (ONU) en 2002, pour archiver les ressources culturelles historiques concernant la Route de la Soie, en les numérisant et les rendant accessibles et disponibles en ligne. Le projet comprend plusieurs sous-projets, chacun d'entre eux travaillant sur un type de données spécifique à numériser. L'un des plus importants sous-projets est l'archivage numérique des livres rares de la bibliothèque Toyo Bunko. Dans le cadre de ce projet, beaucoup de livres anciens rares disponibles à la bibliothèque Toyo Bunko ont été numérisés en utilisant la technologie OCR (Optical Character Recognition). La collection numérisée en ligne contient 113 livres dans 11 langues différentes (anglais, français, russe...), tous liés à la Route de la Soie historique, comme les deux volumes de l'Ancien Khotan par Marc Aurel Stein. Il est souhaitable de construire une base multilingue de termes pour les ressources de la DSR, car elle a une communauté d'utilisateurs multilingue. Les approches traditionnelles collaboratives dépendent de professionnels ou de non-professionnels dans le développement de ressources lexicales. Toutefois, l'activité de contribution elle-même est généralement une activité linguistique ou lexicale au cours de laquelle un contributeur apporte de la connaissance terminologique en le voulant, dans un souci de contribution. L'approche proposée pour la contribution active à la préterminologie dépend de la possibilité de lier l'activité de contribution à une activité très attrayante.

A titre expérimental, la base préterminologique MPG-DSR a été construite par des approches automatiques et a ensuite été étendue (1) par une passerelle contributive qui offre des activités attrayantes aux visiteurs de la DSR, et (2) en recevant des contributions par le biais de ces services, en particulier, par les activités Contribuer en Cherchant ((CWS, Contribute While Searching)), et Contribuer en Lisant ((CWR, Contribute While Reading)).

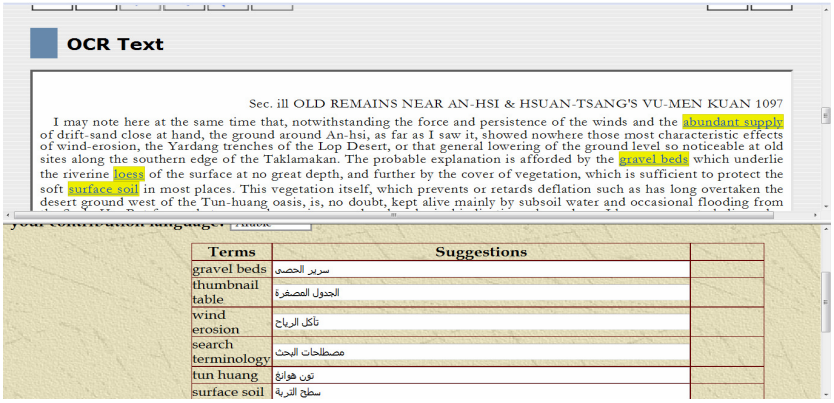

Figure 2: passerelle contributive pour la DSR

La figure 2 montre une capture d'écran de la passerelle contributive, où un utilisateur peut sélectionner un terme du livre, le traduire et offrir de nouvelles traductions directement à partir du même écran.

3.2 Résultats et évaluation

La figure 3 montre le nombre total de visites de la passerelle jusqu'à la fin de chaque mois. D'avril 2010 à août 2010, la passerelle a reçu environ 580 visites, dont 180 (31%) ont conduit à une contribution.

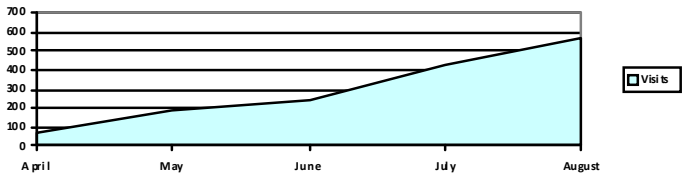

Figure 3: Nombre cumulé de visites à la passerelle contribution de la DSR-MPG

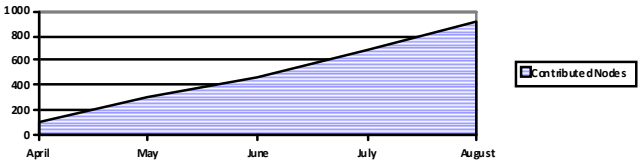

Figure 4: Nombre cumulé de nœuds contribués sur la période

Dans une session de contribution, plusieurs unités lexicales peuvent être contribuées. C'est pourquoi 920 unités lexicales ont été recueillies en 180 sessions de contribution (figure 4). Cela représente environ 5,1 contributions lexicales par session de contribution.

Au cours d'une période de 4 mois, CWS ((Contribuer en Cherchant)) a reçu 156 requêtes. La performance est similaire à la performance du système actuel, quand les termes de recherche sont en anglais. Toutefois, le CWS a reçu 68 requêtes dans d'autres langues que l'anglais et le japonais. 44 de ces requêtes ont été traduits par MPG-DSR avec succès, et des résultats ont été trouvés, alors que la recherche monolingue n'avait trouvé aucun résultat pour ces requêtes. Le taux d'accès est donc 44/68 = ~ 65%.

Les contributions en arabe ont été évaluées: 345 termes sur 390 ont été correctement traduits (88,4%). La plupart des termes étaient des termes multimots, et les 900 contributions ont ajouté 540 nouveaux nœuds au MPG.

Comme les 345 contributions provenaient d'environ 61 sessions contributives, 70 termes arabes de différentes séances (environ 1 terme de chaque session contributive) ont été sélectionnés et cherchés dans Alburaq et dans Google Dictionary. Les résultats ont été les suivants (tableau 1): Les résultats ont été les suivants (tableau 1):

	Exemples de prétermes contribués à partir de DSR-MPG	Dictionnaire Google	Alburaq
Nombre de prétermes	Nombre de prétermes	70	19
Pourcentage	100%	37,1%	27,1%

Tableau 1: résultats de la recherche d'un l'échantillon de prétermes contribués

4 JeuxDeMots arabe

4.1 Expérimentation

JeuxDeMots est un jeu en ligne qui vise à recueillir des unités lexicales, et à construire des fonctions lexicales entre elles grâce à un processus de divertissement ; les données collectées sont stockées sous forme d'un graphe. Les joueurs

créent un compte pour jouer à ce jeu. Après cela, au premier niveau du jeu, le jeu propose au joueur un terme et il a 1 minute pour entrer autant qu'il le peut d'unités lexicales associées au terme donné. Si sa réponse correspond à la réponse d'un joueur précédent sur même terme, les deux gagnent des points et leur classement monte.

JeuxDeMots arabe a été initialisé en utilisant un MPG basé sur l'arabe, en espérant cela élargirait le graphe et enrichirait son contenu. Ce paragraphe décrit les détails techniques de l'intégration de notre système, SEpT (Système pour l'Élicitation de Préterminologie), avec un jeu sérieux en ligne. Le JeuxDeMots arabe initial (JDMAR) (javalig.imag.fr/ jdmar/) a été localisé par la traduction de ses interfaces en novembre 2009; et le public a commencé à y jouer en janvier 2010.

4.2 Résultats

Le jeu a été "ensemencé" avec environ 750 termes arabes, choisis dans Alburaq.net. Il y a eu une croissance régulière du nombre des joueurs inscrits, à partir de janvier 2010. Notez que les visiteurs en ligne peuvent également jouer sans inscription à titre d'invités. Jusqu'à présent, nous avons environ 55 joueurs. La figure 5 montre la croissance du nombre de termes contribués au cours de la période de 8 mois partant du 1er janvier 2010. La figure 6 montre la croissance du nombre de relations ajoutées au graphe au cours de la même période de 8 mois.

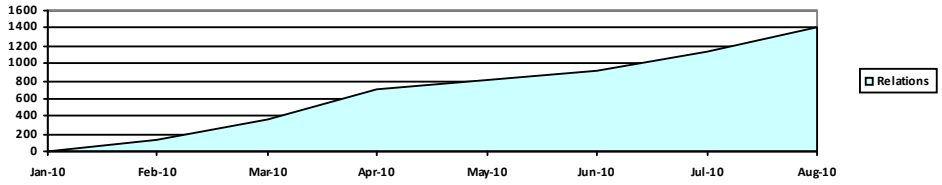

Figure 5: Nombre cumulé de termes contribués chaque mois

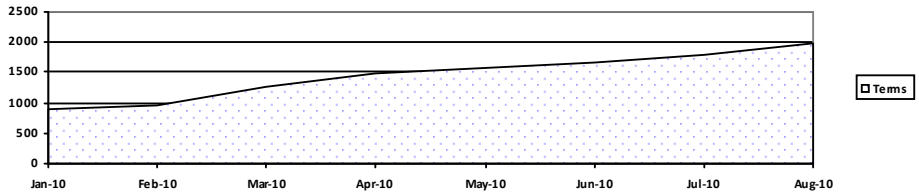

Figure 6: nombre cumulé de relations contribuées

JDMAR a prouvé le potentiel des contributions indirectes actives à la préterminologie d'une langue peu dotée telle que l'arabe, en maintenant un taux de croissance acceptable tout au long de l'expérience, par rapport à d'autres instances de JeuxDeMots, comme le JDM japonais.

Il a atteint son objectif de collecte de préterminologie à travers une activité non linguistique (non lexicale). Le deuxième objectif, secondaire, était d'attirer et de divertir des utilisateurs arabes. Le jeu a eu des joueurs fidèles (et des invités) de 18 pays différents. 20 joueurs inscrits ont répondu à une question d'enquête demandant s'ils jouent pour se divertir, ou seulement pour contribuer à une ressource lexicale : 3/20 ont répondu qu'ils jouent pour se divertir, 16/20 ont dit qu'ils jouent pour se divertir et pour contribuer, et 1/20 a dit qu'il ne joue que pour contribuer. Sur les 1400 relations contribuées, 200 relations ont été sélectionnées. Près de 175 sur 200 correspondaient à une relation fonctionnelle ou ontologique significative, et 25 étaient inexactes.

Pour tester la disponibilité du nouveau contenu dans la préterminologie produite, 200 prétermess de JDMAR-MPG ont été sélectionnés au hasard. Seulement 125 ont été trouvés dans Alburaq.net, et seulement 65 ont été trouvés dans le WordNet arabe. La raison en est que la contribution active à travers des jeux sérieux apporte de nouveaux prétermes et des termes dialectaux qui ne sont pas disponibles dans les ressources traditionnelles, comme par exemple les termes montrés dans le tableau 2.

Pré-termes en arabe	Translitération	Traduction en anglais
فيســـــبوك	feesbuk	facebook

تويـــتر	tweeter	twitter
كاســيت شــريط	Kasit shareet	cassette
قطــايف	qatayef	une sorte de crêpe sucrée
توجيهـــي	tawjeehi	l'école secondaire
مفلــوز	mfalwez	quelqu'un qui a la grippe
بييـــاره	bayyara	plantation d'agrumes

Tableau 2: Exemple de prétermes

Conclusions

Traditionnellement, les terminologues sont chargés d'élaborer la terminologie. Leur travail sur un terme consiste à construire la correspondance entre la représentation symbolique, le concept et l'information terminologique, pour placer le terme dans la "sphère terminologique" (grâce à des informations telles que les définitions). Construire ces correspondances est une tâche épuisante qui influe sur le coût et la couverture. D'autre part, les approches collaboratives essayent de préparer le travail des terminologues par des bénévoles amateurs, ce qui est une tendance prometteuse dans l'acquisition de connaissances. Cependant, il peut être difficile d'amener les bénévoles à participer à une activité linguistique, même s'ils sont réellement familiers avec le domaine.

Les approches automatiques utilisent des ressources textuelles et lexicales pour développer la terminologie, mais elles sont limitées aux ressources disponibles et aux techniques classiques. Aussi, toutes les connaissances terminologiques ne sont pas disponibles en format textuel (terminologie latente). D'ailleurs, même lorsque les approches automatiques sont efficaces pour trouver des termes, il est difficile de construire les mêmes correspondances que celles établies par des terminologues.

Cet article décrit deux expériences préliminaires sur la création de ressources terminologiques multilingues préliminaires (préterminologie) à travers une approche alternative (la "société Internet"). Le JeuxDeMots arabe a atteint son but qui était de divertir des joueurs en ligne quelconques tout en recueillant une quantité importante de prétermes et de relations. Son contenu (prétermes, et correspondances monolingues) est unique et contient des mots du langage parlé qui ne sont pas disponibles dans les bases de données standard, et le taux de croissance a été stable et prometteur.

L'autre contribution technique active d'autres est directe, bien que le processus de contribution s'inscrive dans une activité non linguistique qui est une activité normale de la communauté en ligne elle-même. Les activités Contribuer en Lisant ((CWR)) et Contribuer en Cherchant ((CWS)) ont montré un bon potentiel pour attirer des contributions d'experts du domaine dans la moyenne. Les données contribuées ont un caractère distinctif parce qu'elles ne figuraient pas dans des ressources numériques, et qu'un faible pourcentage des termes contribués était couvert par les dictionnaires classiques.

References

AHN, L. V. (2005). Human Computation. Computer Science. Pittsburgh, Carnegie Mellon University PhD degree, 87 p.

DAOUD, M., C. BOITET, ET AL. (2009). Constructing multilingual preterminological graphs using various online community resources. SNLP2009, Thailand, 116-121.

DIAB, M. AND N. HABASH (2009). Arabic Dialect Processing. MEDAR09. April, 2009, Cairo, Egypt. Tutorial.

ETZIONI, O., K. REITER, ET AL. (2007). Lexical translation with application to image searching on the web. MT Summit XI, Copenhagen, Denmark, 175-182.

JOUBERT, A. AND M. LAFOURCADE (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. JADT'2008, Ecole normale supérieure Lettres et sciences humaines, Lyon, France, 657-666.

KAGEURA, K. AND B. UMINO (1998). Methods of automatic term recognition. Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, vol. 3 (2), 259-289.

NAGATA, M., T. SAITO, ET AL. (2001). Using the web as a bilingual dictionary. Proceedings of the workshop on Data-driven methods in machine translation, vol. 14, 1-8.

ONO, K., A. KITAMOTO, ET AL. (2008). Memory of the Silk Road — The Digital Silk Road Project. Proceedings of (VSMM08), Project Papers, Limassol, Cyprus, 437-444.

YASSIN, Y. A. (2003). Why Arabic Is the Most Difficult Language for Localization. Globalization Insider, Vol. XII, 5 p.