La Déclaration Universelle des Droits de l'Homme : 329 langues pour la constitution automatique de corpus et de lexiques

Hubert Naets
Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M)
Commissariat à l'Énergie Atomique
Bat. 38-1 ; 18, rue du Panorama ; BP 6
92265 Fontenay aux Roses Cedex ; France
naetsh@zoe.cea.fr

Mots-clefs : langues-pi, corpus, world wide web, langues peu dotées

Keywords: pi-languages, corpus, world wide web, under-resourced languages

Résumé Dans cet article, nous présentons le prototype d'un système permettant la constitution automatique de corpus depuis le web à partir des 329 langues de la Déclaration Universelle des Droits de l'Homme. Ce système exploite au maximum la proximité existant entre les langues pour augmenter la précision des requêtes envoyées au moteur de recherche et éviter ainsi la récupération de documents écrits dans des langues de même famille. Cette démarche s'inscrit dans le cadre de la production de ressources linguistiques informatisées pour les langues minoritaires dont la survie dépend entre autres de ces données.

Abstract In this article, we present the prototype of a system allowing the automatic constitution of corpora from the web, with the 329 languages of the Universal Declaration of Human Rights as bootstrapping. This system exploits the proximity existing between the languages to increase the precision of the requests sent to the search engine and to avoid the crawling of documents written in a language of the same family.

1 Introduction

En 2000, dans un article du *Courrier* de l'Unesco[1], intitulé *6000 langues: un patrimoine en danger*, Ranka Bjeljac-Babic constatait que sur les quelque 6000 langues existant dans le monde, dix d'entre elles disparaissaient chaque année. Cinquante langues seraient donc mortes depuis la parution de son texte. Pour Bjeljac-Babic et d'autres, « une langue qui n'est pas employée sur Internet "n'existe plus" dans le monde moderne. Elle est hors circuit. Elle est exclue du "commerce" », dans un cadre où « la diversité des langues [est] perçue comme une entrave aux échanges et à la diffusion du savoir ».

Le système présenté dans cet article ne pourra certes pas entraver la disparition prévue de 50 à 90% des langues au cours de ce siècle mais a entre autres buts de fournir à certaines d'entre elles des ressources nécessaires à leur prise en compte dans des analyseurs morphosyntaxiques, des parseurs syntaxiques, des moteurs de recherche ou tout autre outil issu du monde du Traitement Automatique des Langues. L'objectif que nous essayons d'atteindre est l'automatisation de la production et de l'analyse des ressources, en limitant au strict minimum toute intervention humaine. Cela correspond à un besoin croissant des industries de la langue d'étendre rapidement leurs offres à de nouvelles langues, alors que ces entreprises ne disposent pas nécessairement, dès l'introduction de nouvelles langues dans leurs systèmes, des personnes compétentes pour ces langues. Les langues peu dotées en gagneront une reconnaissance implicite.

Le programme que nous nous sommes fixé pour correspondre à ces besoins comprend la collecte de corpus pour de nouvelles langues, l'analyse morphologique des formes de ces langues, ainsi que la constitution de dictionnaires multilingues et la mise à disposition de ces ressources sur le Web.

Dans cet article, nous nous centrons sur la constitution des corpus. Il convient de noter que le traitement ici proposé n'a pas encore été testé dans sa totalité.

2 Les systèmes existants de construction de corpus à partir du web pour les langues minoritaires

Kevin P. Scannell[2] (Scannell, 2003) a réalisé *An Crúbadán*, un crawler web spécialisé dans la constitution de corpus notamment pour des langues minoritaires. À partir de textes d'amorçage d'une centaine de mots, il combine plusieurs de ces mots pour générer des requêtes qui sont transmises à l'API de Google. Le moteur de recherche renvoie une liste de documents potentiellement écrits dans la langue cible. Ces documents sont récupérés depuis le web et sont traités à l'aide d'un ensemble de techniques statistiques afin de déterminer quels documents ou parties de ceux-ci sont écrits dans la langue recherchée. Le web crawler parcourt ensuite les liens présents dans les documents identifiés comme appartenant à la langue cible. Le nouveau corpus ainsi constitué sert à amorcer la génération d'un nouvel ensemble de requêtes. *An Crúbadán* est utilisé actuellement sur 136 langues.

CorpusBuilder [(Ghani, Jones, Mladeni'c, 2001) et (Ghani, Jones, Mladeni'c, 2003)], qui a le même objectif de constitution de corpus de langues minoritaires que *An Crúbadán*, possède

[1] http://www.unesco.org/courier/2000_04/fr/doss01.htm
[2] http://borel.slu.edu/crubadan/apps.html

l'architecture suivante : une phase d'amorçage du système est assurée au moyen de deux petits ensembles de documents : des documents pertinents pour la langue recherchée et d'autres non. Une méthode de sélection de termes issus de ces deux groupes de documents est utilisée pour générer des requêtes composées de termes à inclure (provenant des documents pertinents) et de termes à exclure (issus des documents non pertinents). La requête ainsi produite est transmise à un moteur de recherche. Le document de plus haut rang renvoyé par ce moteur est téléchargé et passé à travers un filtre d'identification de langue. Selon la classification du filtre, le document est ensuite ajouté à l'ensemble des documents pertinents pour la langue ou à l'ensemble des documents non pertinents. La base de documents est alors mise à jour et le processus réitéré. R. Ghani, R. Jones et D. Mladenic insistent particulièrement sur les techniques de génération de requêtes en en testant six différentes (sélection uniforme des termes, sélection basée sur les fréquences, sélection probabiliste basée sur les fréquences, *rtfidf*, odds-ratio et odds-ratio probabiliste). *CorpusBuilder* a été utilisé pour traiter le slovène, le croate et le tchèque.

3 Vue d'ensemble du système

L'architecture de notre système (figure 1), qui est dans l'ensemble assez semblable à celles de K. P. Scannell et de R. Ghani, R. Jones et D. Mladenic, repose essentiellement sur une volonté de traiter un maximum de langues ou d'idiomes en parallèle et, partant, sur la nécessité de distinguer le plus tôt possible des idiomes très proches et ce, de la façon la plus automatisée possible.

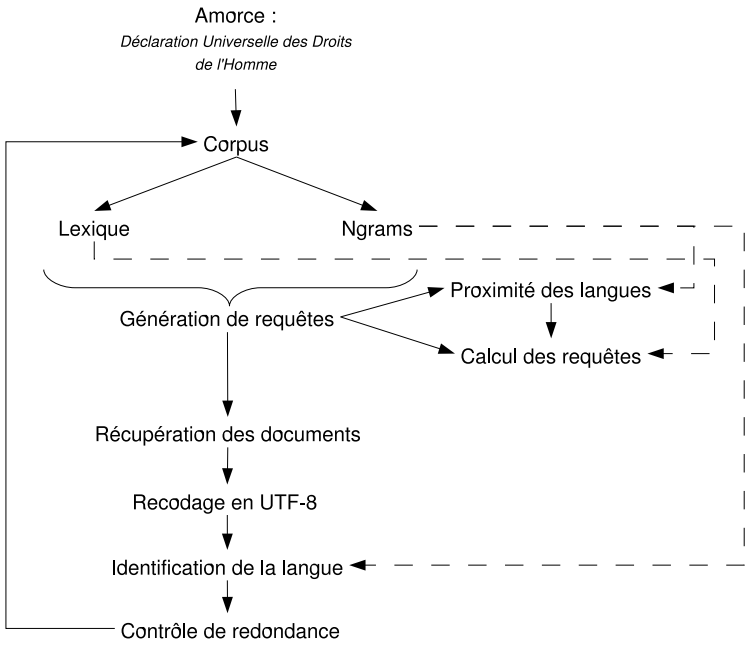

Figure 1: Schéma général du fonctionnement du système

4 La phase d'amorçage

À la suite de K. P. Scannell et de R. Ghani, R. Jones et D. Mladenic, nous partons d'un texte d'amorçage de relativement petite taille et présentant la caractéristique d'être actuellement traduit en 329 langues : la *Déclaration Universelle des Droits de l'Homme*[3] (DUDH). Parmi ces 329 versions, 275 sont au format texte, dans différents encodages de caractères, alors que les 56 documents restants sont représentés par une image (de type jpeg) ou par un fichier pdf, sans version texte. Les idiomes dans lesquels la DUDH est traduite sont très variés : des langues de grande ampleur (anglais, chinois, russe, espagnol,...) côtoient des langues régionales (breton, picard, galicien,...), des pidgins (du Nigeria par exemple) ou encore des idiomes presque éteints (l'arabela qui n'était plus utilisé que par 150 locuteurs en 1989 au Pérou, le yukaghir parlé par un nombre similaire de personnes en Russie, etc.). Nous avons remplacé un certain nombre d'images par des textes de la langue donnée, lorsque cela s'avérait possible.

Ces traductions de la DUDH, qui sont autant de corpus comparables, sont converties au format UTF-8 et servent à produire la liste des mots du texte, nécessaire pour la création des requêtes, ainsi que des ngrams de lettres permettant notamment d'identifier la langue des nouveaux documents. Les langues ne possédant pas de séparateurs sont découpées en bigrammes ou en trigrammes (en chinois par exemple) ou, lorsque cela est possible, en utilisant les changements de jeux de caractères (c'est le cas du japonais).

5 La génération des requêtes

La génération des requêtes constitue sans doute l'élément le plus important de la phase de constitution de corpus. Il s'agit en effet de produire des requêtes qui ramènent un maximum de documents dans la langue recherchée, à l'exclusion de toute autre langue. En d'autres termes, il s'agit d'atteindre les rappel et précision les plus importants possible. Il faut donc produire des requêtes composées de n mots spécifiques à une langue par rapport à des langues proches et éventuellement d'exclure m mots spécifiques aux langues proches de la langue recherchée.

> EXEMPLE : — Si l'on désire produire une requête permettant de découvrir des documents en asturien, on s'abstiendra de mettre dans la requêtes des formes telles que « que » ou « de » qui sont aussi très communes en catalan, en espagnol, en français, en galicien, en latin, en occitan auvergnat ou encore en occitan languedocien.

L'approche que nous proposons consiste donc à déterminer automatiquement, avant de générer la requête, quels sont les mots très fréquents et très spécifiques d'une langue par opposition aux langues proches. La fréquence importante de ces mots contribue à augmenter le nombre de documents découverts à l'aide d'un moteur de recherche ; quant à la spécificité, elle permet de s'assurer que le document (ou une partie de celui-ci) est bien écrit dans une langue donnée et non dans une autre.

Le problème peut être décomposé en deux sous-problèmes :

1. déterminer quelles sont les langues proches ;

[3] http://www.unhchr.ch/udhr/navigate/alpha.htm

2. déterminer les mots spécifiques à une langue par opposition à d'autres langues.

5.1 La proximité entre les langues

Le calcul de la proximité entre deux langues a pour objectif d'économiser un maximum de ressources et de temps : il est en effet peu efficace de déterminer le vocabulaire spécifique à l'asturien par rapport au japonais dans la mesure où l'intersection entre ces deux langues tend vers zéro, alors que l'asturien et l'espagnol ont une intersection beaucoup plus grande, ce qui implique une probabilité plus importante de confondre ces deux langues et donc de sélectionner erronément un document écrit dans une langue non recherchée.

Une approche naïve voudrait que l'on détermine la proximité de deux langues en comparant leurs lexiques. Plus le nombre de mots communs entre deux langues est important, plus les deux langues sont proches. C'est cependant sans tenir compte du fait que le corpus d'amorçage étant très restreint, il n'est absolument pas certain que tous les mots les plus fréquents d'une langue figurent dans le vocabulaire de ce corpus de départ. Il est donc nécessaire d'avoir un corpus d'une certaine taille avant que l'ordre des mots classés par fréquence ne se stabilise.

Une autre approche consiste à comparer les ngrams des deux langues plutôt que leur vocabulaire. Il s'agit de calculer les fréquences d'apparition de séquences de n lettres au sein du corpus. Ces séquences étant beaucoup plus fréquentes que des mots pris isolément, le modèle de langage se stabilise beaucoup plus rapidement et est donc utilisable avec de plus petits corpus.

C'est cette approche que nous avons sélectionnée pour la première version de notre collecteur de corpus. Nous sommes partis de l'algorithme de Cavnar et Trenkle (Cavnar, Trenkle, 1994), qui est utilisé également pour l'identification de la langue des documents (voir plus loin). Nous avons considéré que le corpus pour une langue donnée était le texte dont nous recherchions la langue. Une fois éliminée la meilleure langue possible pour le texte en question (cette meilleure langue est la langue du corpus), restent les n langues les plus proches, par ordre de proximité, où n peut être fixe ou variable.

5.2 Les mots spécifiques à une langue en vue de leur utilisation dans une requête

Comme nous l'avons dit précédemment, notre hypothèse est que, si l'on désire obtenir un maximum de documents pertinents pour une langue donnée, la partie positive de la requête doit être composée de termes les plus fréquents propres à la langue dont on recherche des documents ; quant à la partie négative, elle doit comporter les termes les plus fréquents dans les langues avec lesquelles la langue recherchée peut être confondue mais qui soient très peu fréquents dans la langue recherchée. Cette hypothèse correspond au score d'*odds-ratio*.

L'*odds-ratio* OR pour un mot m, étant donné le lexique d'une langue recherchée, $langue$, et le lexique confondu des n langues les proches de la langue recherchée, $languesproches$, se définit de la façon suivante :

$$OR = log_2(\frac{P(m|langue) * (1 - P(m|languesproches))}{P(m|languesproches) * (1 - P(m|langue))})$$

(Ghani, Jones, Mladeni'c, 2001) et (Ghani, Jones, Mladeni'c, 2003) comparent plusieurs méthodes de génération de requête et concluent que pour le slovène, le croate, le tchèque et le tagalog, la meilleure de ces méthodes (c'est-à-dire la méthode qui permet de récolter le plus de documents par requête) est celle des odds-ratios.

Exemple de requête produite pour l'espagnol :

```
+información +cinco +sentencias +ejecutado +partes -sans -au -peut -els -je -dels -pas
-mais -il
```

6 L'identification des langues

L'étape d'identification des langues présentes dans les documents récoltés est la partie la plus sensible de ce système. Il s'agit en effet de s'assurer que chaque document est bien écrit dans la langue recherchée et d'éliminer un maximum d'éléments écrits dans une autre langue. Pour ce faire, nous avons réimplémenté l'algorithme de catégorisation de textes basé sur des ngrams de Cavnar et Trenkle (Cavnar, Trenkle, 1994), en restant compatible avec l'implémentation *TextCat* de van Noord[4]. L'identificateur de langue a été entraîné à l'aide de chaque corpus de départ constitué pour chaque langue et est dynamiquement réentraîné à chaque série d'ajouts de textes au sein du corpus. Il est utilisé pour identifier la langue générale du texte et la langue de chaque phrase de celui-ci. En complément, un identificateur de langue basé sur des tokens a également été implémenté. Même si ce type de détecteur de langue fonctionne en général moins bien (Grefenstette, 1995), il s'avère très utile lorsque le nombre de mots dont il faut identifier la langue n'est pas suffisamment important pour permettre à l'identificateur ngrams de décider.

La reconnaissance de la langue porte sur le texte dans sa totalité mais également sur chaque phrase et/ou partie du document d'origine (Prager, 1999). Les parties dont la langue est douteuse ou qui ne sont pas écrites dans la bonne langue sont rejetées.

7 Contrôle de la redondance des documents

La dernière opération consiste à vérifier qu'un même document n'a pas été récupéré plusieurs fois par le système. Ceci s'avère important pour les langues très peu présentes sur le web, particulièrement dans les premiers moments de la phase d'amorçage où le lexique de ces langues est encore très pauvre. Un simple vérification de l'URL du document ne s'avère en effet pas suffisant dans la mesure où certains textes, comme par exemple la *Déclaration Universelle des Droits de l'Homme* dans une langue donnée, sont réutilisés à plusieurs endroits du web, au sein de mises en page parfois différentes. Le contrôle de la réutilisation des documents permet d'éliminer ces textes et ainsi d'éviter de biaiser la constitution du vocabulaire du texte qui, autrement, serait sur-représenté pour certains termes et sous-représenté pour les autres.

Pour la première version du système, nous avons choisi d'utiliser la technique de chevauchement de ngrams (*ngram overlap*) (Clough et al., 2002) : pour un texte source A et un texte potentiellement dérivé B, représentés par les ensembles de ngrams $E_n(A)$ et $E_n(B)$, et la proportion de ngrams présents à la fois dans A et dans B, la similarité entre les deux textes est la suivante :

[4] http://odur.let.rug.nl/ vannoord/TextCat/

$$SIM_n(A,B) = \frac{E_n(B) \cap E_n(A)}{E_n(B)}$$

Cette méthode permet de déterminer si deux textes partagent le même vocabulaire et si les mêmes séquences de mots sont présentes dans le même ordre au sein des deux textes. Elle permet ainsi d'écarter les textes provenant de la même source.

Une fois passée cette dernière étape, les documents ou parties de documents sélectionnés sont intégrés au corpus de la langue. Le lexique et les ngrams sont recalculés pour l'ensemble du corpus et le processus complet est répété.

8 Premiers résultats

Le système n'a pas encore pu être testé dans sa totalité mais les premiers résultats concernant chaque partie s'avèrent néanmoins extrêmement encourageants.

Ainsi, en divisant chacun de nos corpus d'amorçage en deux, la première partie servant à entraîner l'identificateur de langues ngrams et la seconde à le tester, nous avons obtenu une identification correcte des langues dans 97,8 % des cas. Les erreurs concernent des langues extrêmement proches et pour lesquelles la taille du demi-corpus d'amorçage utilisé ici n'était pas suffisante. Il est possible que nous devions enrichier manuellement le corpus d'amorçage pour certaines langues. Par ailleurs, nous n'avons pas encore pu tester les effets de l'entraînement dynamique de l'identificateur ngrams à partir des nouveaux documents récoltés.

La technique des odds-ratios produit également d'excellents résultats. Nos premiers essais indiquent néanmoins que les pages web de certaines langues peu dotées sont parfois polluées par des langues de diffusion plus importante. Nous n'avons pas encore pu évaluer cette proportion ni l'efficacité des deux identificateurs de langue dans ce cas. De la même façon que nous prenons en compte les langues les plus proches dans le calcul des odds-ratios, nous envisageons de déterminer pour chaque langue la liste des langues les plus "polluantes". Nous ne savons pas encore par contre si nous utiliserons cette liste pour filtrer les pages lors de la génération des requêtes ou si nous l'emploierons pour renforcer l'identification des langues. Il est probable qu'il faudra choisir cette deuxième solution pour les langues extrêmement peu dotées, sous peine de réduire drastiquement la taille de leur corpus. Par ailleurs, d'autres idiomes de la DUDH semblent absents du web.

Le codage des documents a également posé un certain nombre de problèmes : le codage indiqué dans l'en-tête du document HTML ne correspond pas toujours au codage réel de ce document, ce qui qui a pour effet de produire des conversions inappropriées en UTF-8. Un outil maison est en cours d'évaluation pour tenter de résoudre — du moins partiellement — cette difficulté.

9 Conclusion

Nous avons présenté un système permettant de récolter des corpus de textes pour 329 langues — dont de nombreuses langues faiblement dotées —, en utilisant comme amorce la "Déclaration Universelle des Droits de l'Homme". La proximité entre certaines langues pouvant engendrer

des confusions lors de la sélection de documents pertinents pour une langue donnée, l'accent a été mis sur l'exploitation de cette proximité afin d'augmenter au maximum la précision du système. Ce système, dont la réalisation est presque terminée, n'a pu être testé jusqu'à présent que morceau par morceau mais semble d'ores et déjà être très prometteur.

Remerciements

Nous tenons à remercier particulièrement M. Gregory Grefenstette (CEA) pour ses nombreux conseils au cours de l'élaboration de l'extracteur de corpus.

Références

Cavnar W., Trenkle J. (1994), N-Gram-Based Text Categorization, *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161-175.

Clough P., Gaizauskas R., Piao S., Wilks Y. (2002), METER: MEasuring TExt Reuse. Proceedings of *the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02)*, University of Pennsylvania, Philadelphia, USA, 152-159.

Ghani R., Jones R., Mladenic D. (2001), Building minority language corpora by learning to generate web search queries (Technical Report CMU-CALD-01-100).

Ghani R., Jones R., Mladenic D. (2003), Building minority language corpora by learning to generate web search queries, *Knowledge and Information Systems*.

Grefenstette G. (1995), Comparing Two language Identification Schemes, *JADT 1995: 3rd International conference on Statistical Analysis of Textual Data.*

Prager J. (1999), Linguini: Language Identification for Multilingual Documents, *Proceedings of the 32nd Hawaii International Conference on System Sciences.*

Scannell P. (2003), Automatic thesaurus generation for minority languages: an Irish example, *Proceedings of TALN 2003*, Batz-sur-Mer.