Construction de corpus multilingues : état de l'art.

Manuela Yapomo

Construction de corpus multilingues : état de l'art

Manuela Yapomo

LiLPa, EA 1339 – ICube, UMR 7357
Université de Strasbourg

1. Introduction

Corpus multilingues

- **Corpus parallèles** : corpus constitués de textes sources et leurs traductions [McEnery et Xiao, 2007]
- **Corpus comparables** : collections de textes similaires rassemblés sur la base d'un ensemble de critères [Skadiņa et al., 2010]

Objectif

Former un corpus multilingue structuré à partir de sous-corpus homogènes (clusters) ou d'alignements obtenus à partir d'une large collection de textes

Problèmes

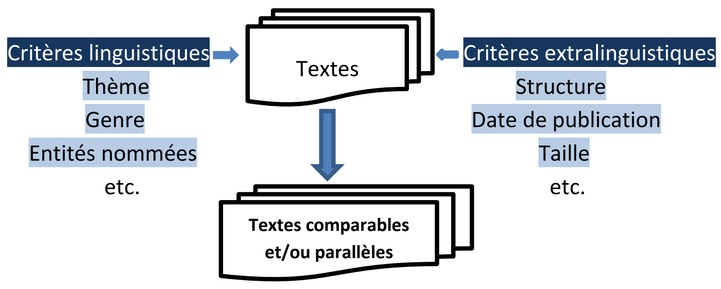
- Quand y a-t-il comparabilité ?
- Comment détecter des différences fines de comparabilité ?
- Comment définir la similarité dans un contexte translingue ?
- Comment mettre à jour un corpus multilingue structuré ?

2. La comparabilité en corpus multilingues

Échelles établies pour le jugement humain

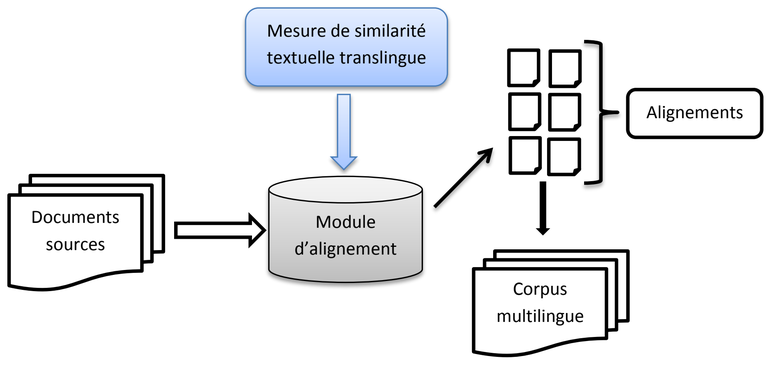
	Bekavac et al. (2004)	Skadiņa et al. (2010b)	Braschler & Schäuble (1998)	Pouliquen et al. (2004)
Critères linguistiques & extra-linguistiques		(1) parallélisme		
	(1) forte comparabilité	(2) forte comparabilité	(1) histoire identique	(1) article identique
			(2) histoire liée	(2) article lié
		(3) faible comparabilité	(3) aspects communs	(3) article vaguement lié
			(4) terminologie commune	
Critères extra-linguistiques (uniquement)	(2) faible comparabilité	(4) aucune comparabilité	(5) sans lien	(4) sans lien

Choix et association de critères de comparabilité

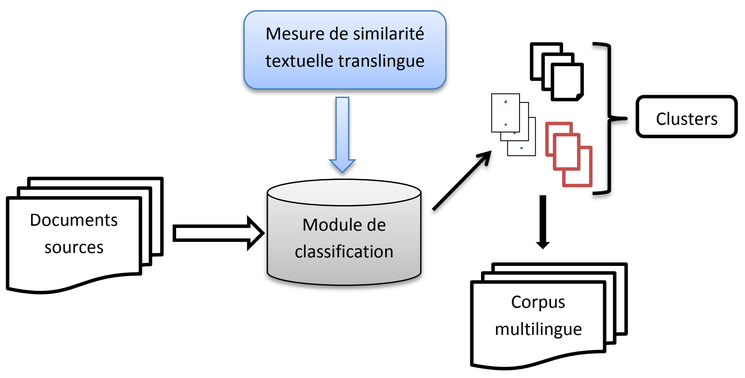


3. Construction automatique de corpus multilingues

alignement de documents



clustering de documents



4. Approches de la similarité translingue

Rupture de la barrière de langue

Paramètres \ Méthodes	Traduction		Représentation conceptuelle
	Traduction automatique	Dictionnaire	
Simplicité	–	+	–
Exhaustivité	–	–	–
Désambiguïsation	+	–	~
Proximité de sens	–	–	+

Calcul de la similarité

Mesures statistiques

Statistiques de corpus (co-occurrence de mots, «substituabilité », voisinage)

Exclusion des sens : traitement superficiel de la polysémie, synonymie, paraphrase

Mesures sémantiques

Pas ou peu d'informations en corpus (contexte)

Représentation conceptuelle des textes à partir de la structure et des descripteurs d'une ressource sémantique

Mesures hybrides

Calcul de la similarité : Cosinus, Jaccard, etc.

5. Évaluation

Évaluation intrinsèque

- Corrélation entre similarité automatique et jugement humain
- Comparaison des corrélations de plusieurs méthodes

Évaluation extrinsèque

- Apport des données obtenues dans une application
- Comparaison des apports de données issues de plusieurs méthodes

6. Conclusion

Contributions envisagées

- Description plus spécifique de la comparabilité pour l'extraction de lexiques multilingues en corpus de spécialité
- Mesure de similarité translingue hybride basée sur la représentation conceptuelle des textes
- Clustering incrémental

Retombées

- Corpus multilingue avec une comparabilité plus fine
- Application principale : extraction de néologismes multilingues
- Applications annexes : extraction de terminologies multilingues

Références

Bekavac, B., Osenova, P., Simov, K. et Tadic, M. (2004).
Making Monolingual Corpora Comparable : a Case Study of Bulgarian and Croatian.
In Proceedings of the 4th Language Resources and Evaluation Conference, pages 1187–1190, Lisbonne, Portugal.

Braschler, M. et Schäuble, P. (1998).
Multilingual Information Retrieval Based on Document Alignment Techniques.
In Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries, pages 183–197, Heraklion, Crète, Grèce.

McEnery, A. M. et Xiao, R. Z. (2007).
Parallel and Comparable Corpora : what are they up to ?
In Incorporating Corpora : Translation and the Linguist. Anderman, G. & Rogers, M., Clevedon, UK, Multilingual Matters édition.

Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E. et Temnikova, I. (2004).
Multilingual and cross-lingual news topic tracking.
In Proceedings of the 20th International Conference on Computational Linguistics, pages 959–965, Genève, Suisse.

Skadiņa, I., Aker, A., Giouli, V., Tufiş, D., Gaizauskas, R., Mieriņa, M. et Mastropavlos, N. (2010).
A Collection of Comparable Corpora for Under-resourced Languages.
In Proceedings of the Fourth International Conference Baltic HLT, pages 161–168, Riga, Latvia.

LiLPa, EA 1339 – ICube, UMR 7357, Université de Strasbourg
Mail: yapomodomkem@unistra.fr – WWW: http://lilpa.unistra.fr

Remerciements : ces travaux ont été financés dans le cadre d'un contrat doctoral IdEx 2012 de l'Université de Strasbourg.