Construction de corpus multilingues : état de l'art

Manuela Yapomo [1, 2]
(1) LiLPa (Linguistique, Langues, Parole), EA 1339
(2) ICube - Laboratoire des sciences de l'Ingénieur, de l'Informatique et de l'Imagerie, UMR 7357
Université de Strasbourg
yapomodomkem@etu.unistra.fr

RÉSUMÉ

Les corpus multilingues sont extensivement exploités dans plusieurs branches du traitement automatique des langues. Cet article présente une vue d'ensemble des travaux en construction automatique de ces corpus. Nous traitons ce sujet en donnant premièrement un aperçu de différentes perceptions de la comparabilité. Nous examinons ensuite les principales approches de calcul de similarité, de construction et d'évaluation développées dans le domaine. Nous observons que Le calcul de la similarité textuelle se fait généralement sur la base de statistiques de corpus, de la structure de ressources ontologiques ou de la combinaison de ces deux approches. Dans un cadre multilingue avec l'utilisation d'un dictionnaire multilingue ou d'un traducteur automatique, de nombreux problèmes apparaissent. L'exploitation d'une ressource ontologique multilingue semble être une solution. En classification, la problématique de l'ajout de documents à la base initiale sans affecter la qualité des clusters demeure ouverte.

ABSTRACT

Multilingual document clustering : state of the art

Multilingual corpora are extensively exploited in several branches of natural language processing. This paper presents an overview of works in the automatic construction of such corpora. We address this topic by first providing an overview of different perceptions of comparability. We then examine the main approaches to similarity computation, construction and evaluation developed in the field. We notice that the measurement of the textual similarity is usually based on corpus statistics or the structure of ontological resources or on a combination of these two approaches. In a multilingual framework, with the use of a multilingual dictionary or a machine translator, many problems arise. The exploitation of a multilingual ontological ressource seems to be a worthy option. In clustering, the problem of adding documents to the initial base without affecting the quality of clusters remains open.

MOTS-CLÉS : corpus multilingues, comparabilité, similarité textuelle translingue, classification.

KEYWORDS: multilingual corpora, comparability, crosslingual textual similarity, classification.

1 Introduction

La tendance actuelle en Traitement Automatique des Langues (TAL) est au développement de méthodes translingues pour la conception et l'amélioration d'outils multilingues. De telles applications dépendent de la disponibilité de corpus multilingues en quantités considérables.

Ces corpus sont principalement de deux sortes : parallèles et comparables. Un corpus est dit parallèle s'il est constitué de textes sources et leurs traductions (McEnery et Xiao, 2007). Les corpus comparables quant à eux regroupent des documents ayant des caractéristiques communes. La difficulté qu'est l'acquisition de corpus exclusivement parallèles et l'importance des corpus comparables (démontrée empiriquement) ont favorisé le développement de méthodes de collecte de corpus comparables à grande échelle. Cependant, peu de travaux sur les standards de comparabilité de telles données ont été menés, les recherches se focalisant sur leur exploitation. Il y a donc une nécessité de développer des méthodes performantes fondées sur une définition précise de la comparabilité pour leur collecte (Su et Babych, 2012).

Cet article qui a pour objectif de présenter l'état de la recherche en acquisition et structuration de textes multilingues s'articule en 4 parties principales. Les corpus multilingues sont présentés en section 2. Nous abordons principalement la notion de comparabilité et les applications de ces corpus. La section 3 traite de leur construction en abordant les sources de collecte, les méthodes de calcul de la similarité et de construction de ces corpus. Nous examinons ensuite en section 4 les différentes approches intrinsèques et extrinsèques d'évaluation des données résultantes. Ces éléments nous permettrons de mieux identifier les limites du domaine et déterminer notre apport à la fois théorique et pratique en section 5. Enfin, nous concluons ce travail en section 6.

2 Corpus multilingues

2.1 La comparabilité en corpus multilingues

La capacité des corpus multilingues à améliorer la performance des systèmes qui y ont recours serait fortement liée à leur degré de comparabilité.

2.1.1 Définition de la comparabilité de documents multilingues

Plusieurs travaux mentionnent le besoin d'une définition de la comparabilité et formulent leur compréhension de celle-ci. Su et Babych (2012) mesurent la comparabilité de textes à leur potentiel d'extraction de segments parallèles et d'amélioration de la performance de systèmes de traduction automatique. Li *et al.* (2011) quant à eux considèrent deux textes ou corpus comme comparables s'ils ont une partie non négligeable de vocabulaire en commun, la principale application étant l'extraction de lexiques bilingues. Leturia *et al.* (2009) soutiennent que la définition de la comparabilité ne peut être dissociée de l'application ciblée et du type de corpus souhaité. La conception de la similarité varierait donc d'un objectif à l'autre. Elle serait également influencée par le type de corpus qui peut être général ou de spécialité. Nous pensons qu'il en est de même pour la source de collecte de documents. Les critères de comparabilité de documents obtenus de Wikipédia (Paramita *et al.*, 2012) par exemple peuvent différer de ceux de documents venant d'un domaine d'articles de presse. Cette mesure de comparabilité est définie par les critères qui la composent.

2.1.2 Choix et association de critères de comparabilité

Dans le cadre d'une application particulière, les paramètres de comparabilité n'ont pas la même préséance. Certains travaux ne prennent en compte que des critères linguistiques, d'autres des critères purement extralinguistiques et d'autres encore font usage de paramètres des deux types.

– En se focalisant principalement sur le contenu, Steinberger *et al.* (2002) mesurent la similarité de documents en comparant les représentations des contenus de documents obtenues au moyen des descripteurs d'un thesaurus. Pour l'extraction de terminologies, Goeuriot *et al.* (2009) ajoutent au thème et au domaine, le type de discours comme critère d'homogénéité. Le type de discours (science ou science populaire) est identifié à travers la structure et les aspects modaux et lexicaux des textes. Pour la même application, Leturia *et al.* (2009) se basent sur la similarité de domaine.
– Pour ce qui est de l'exploitation de critères extralinguistiques, Resnik (1999) identifie des documents parallèles sur le Web à l'aide de la structure de leurs pages. Utsuro *et al.* (2002) quant à eux déterminent la comparabilité d'articles de presse en fonction de leurs dates de publications.
– La majorité des travaux combine à la fois des critères des deux types. Ainsi, les travaux de Baradaran Hashemi *et al.* (2010) se basent sur le sujet et la date de publication pour l'obtention de documents comparables pour la traduction de requêtes en recherche d'information interlingue (RII). Aker *et al.* (2012) exploitent d'une part le sujet et d'autre part, les entités nommées et dates de publication pour une application de traduction automatique. Ils mettent l'accent sur l'importance des métadonnées dans cet exercice.

Nous observons que la tendance, indépendamment de l'application, est de prendre en compte comme critères de similarité soit le thème seul qui peut être modélisé de diverses manières soit le thème accompagné d'un ou de plusieurs autres paramètres. L'utilisation exclusive de critères extralinguistiques fournit des résultats moins bons. Pour rendre compte de la comparabilité, plusieurs échelles de comparabilité ont été développées dans la littérature. Nous établissons dans le tableau 1 une correspondance entre les différents niveaux de ces échelles.

Seuls Skadiņa *et al.* (2010) traitent de la collecte de corpus multilingues avec la mention de textes parallèles. Les autres travaux se limitent aux corpus comparables. Les échelles les plus et moins granulaires sont respectivement celles de Braschler et Schäuble (1998) à 5 niveaux et Bekavac *et al.* (2004) à 2 niveaux. Excepté cette dernière échelle, aucune autre ne considère les critères extralinguistiques comme seules composantes de la comparabilité même la plus légère. Nous remarquons qu'il n'y a pas de consensus quant aux différents degrés de comparabilité en corpus multilingues, sachant qu'à l'exception de Skadiņa *et al.* (2010), ces travaux se basent uniquement sur des articles de presse. Ces échelles de comparabilité étant établies pour le jugement humain, il se pose également le problème de leur adaption à la similarité automatique.

Le degré de comparabilité de documents multilingues jouerait un rôle crucial dans leurs applications.

2.2 Applications des corpus multilingues

L'importance des corpus en général et notamment des corpus multilingues s'observe dans plusieurs domaines du TAL. De nombreux travaux portent sur l'extraction de segments parallèles

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

TABLE 1 – Niveaux de comparabilité en corpus multilingues

à partir de corpus multilingues. Concernant l'extraction de terminologies ou de lexiques multilingues des textes, l'approche générale (Fung et Yee, 1998) consiste en la représentation des mots des textes en vecteurs de contextes. La traduction d'un mot source dans la langue cible est identifiée par le repérage de vecteurs similaires ou équivalents dans les données cibles. Le projet TTC [1] (Blancafort *et al.*, 2010) a abouti à la création d'un outil de génération automatique de terminologies bilingues à partir de corpus comparables dans plusieurs langues pour la traduction automatique. Pour cette même application, Bin *et al.* (2010) agrandissent un corpus parallèle anglais-chinois avec des phrases parallèles extraites de corpus comparables de brevets d'invention. Des résultats encourageants sont obtenus par le système de traduction automatique entraîné et testé avec le corpus parallèle obtenu. Ion (2012) propose un outil d'extraction de segments parallèles des corpus comparables pour enrichir des modèles de traduction statistique.

Réaliser ces applications nécessite d'avoir une quantité considérable de documents multilingues et donc des méthodes performantes pour leur collecte automatique.

3 Construction automatique de corpus multilingues

La procédure de construction de corpus multilingues présente 3 aspects principaux.
– les sources à partir desquelles les documents sont initialement extraits (section 3.1).
– la mesure de similarité permettant d'évaluer la comparabilité entre deux textes ou entre des textes et des groupes de textes (section 3.2).
– et enfin la méthode de construction de corpus elle-même pouvant prendre la forme de crawling, de RII ou de classification (section 3.3).

1. TTC : *Terminology Extraction, Translation Tools and Comparable Corpora*. `http://ttc.syllabs.com/`

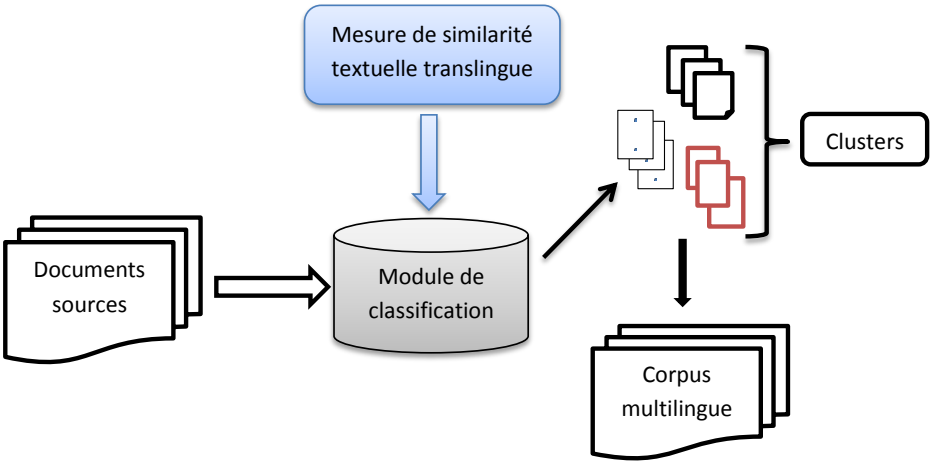

FIGURE 1 – Architecture générale d'un système de classification de documents

La méthode de construction se sert de la mesure de similarité développée pour obtenir des documents hautement comparables et/ou parallèles de la source. Cette procédure développée sous la présente section est illustrée par la figure 1. Dans cette article, nous nous orientons vers la méthode de classification, également illustrée dans la figure 1, pour la construction d'un corpus multilingue.

3.1 Sources d'acquisition

3.1.1 Collections existantes

Les recherches antérieures en compilation de corpus multilingues ont été principalement faites sur la base de corpus de recherche existants. Dans leur objectif de produire un corpus comparable, Bekavac *et al.* (2004) exploitent des sous-corpus monolingues d'un corpus de référence bulgare et croate dont ils alignent les documents comparables. Talvensaari *et al.* (2007) utilisent deux collections de documents monolingues, pour créer un corpus comparable suédois-anglais. Ces données ont l'avantage d'être prêtes pour l'utilisation puisqu'il n'est pas nécessaire de développer des méthodes supplémentaires pour la collecte comme c'est le cas avec le Web. Cependant, la variété des données reste un problème auquel le Web apporte un début de solution.

3.1.2 Le Web

Le Web est une source de plus en plus sollicitée pour la collecte de corpus en tous genres. De nombreux travaux en acquisition de corpus multilingues sur le Web se focalisent sur les articles de presse. Aker *et al.* (2012) tirent profit de la représentation des documents en clusters et de la disponibilité des titres à travers les flux RSS qu'offre *Google News* pour extraire des articles

comparables pour les paires de langues anglais-allemand et anglais-grec. Wikipédia aussi est utilisé pour cette application grâce à sa structure ou à ses liens inter-langues qui mettent en relation les articles de différentes langues sur des thèmes identiques ou liés. Ion *et al.* (2010) compilent un corpus comparable roumain-anglais par l'identification d'articles ayant les mêmes titres. Paramita *et al.* (2012) identifient les articles comparables à l'aide des *dumps* [2] et des liens inter-langues sous Wikipedia. Certaines études ne restreignent pas leur collecte à des domaines spécifiques. C'est le cas des travaux de Talvensaari *et al.* (2008) et Leturia *et al.* (2009) qui conçoivent respectivement un crawler thématique et un moteur de recherche pour obtenir des corpus de spécialité de l'Internet.

L'extraction de documents comparables et parallèles des sources exposées ci-dessus requiert une mesure de similarité.

3.2 Mesures de similarité translingue

L'acquisition de documents multilingues est actuellement réalisée au moyen de mesures de similarité. Ces mesures peuvent être de trois sortes : statistiques, sémantiques et éventuellement hybrides.

3.2.1 Mesures statistiques

Des méthodes basées sur le vocabulaire commun et la recherche d'information peuvent être utilisées pour calculer la similarité de textes.

– Selon la méthode basée sur le vocabulaire commun, la similarité de deux textes se mesure à la quantité de mots qu'ils ont en communs. Dans un contexte multilingue, cette mesure repose sur la quantité d'équivalents de traduction que partagent deux textes, obtenue par l'utilisation d'un dictionnaire bilingue ou d'un traducteur automatique. Li et Gaussier (2010) et Su et Babych (2012) développent des mesures de comparabilité en corpus ou textes multilingues à travers l'utilisation de dictionnaires bilingues. Su et Babych (2012) va plus loin dans le prétraitement de données pour contourner les problèmes d'ambiguïté et des différentes formes d'un mot par l'annotation en parties du discours et la stemmatisation.
– Une seconde approche consiste dans la conversion de documents sources en requêtes au moyen de techniques d'extraction de mots clés. Yapomo *et al.* (2012) et Talvensaari *et al.* (2007) utilisent respectivement les mesures TF-IDF [3] et [4] pour obtenir des listes de mots clés représentatives de textes sources. Les documents pertinents sont ceux dont les vecteurs se rapprochent des vecteurs de requêtes. Les requêtes sont traduites dans la langue des documents cibles pour une similarité translingue.

Ces techniques sont certes les moins coûteuses par l'utilisation principale de statistiques de corpus mais l'exclusion d'informations sémantiques soulève plusieurs problèmes. Puisque la similarité est calculée principalement sur la base de mots identiques, synonymie et paraphrase ne sont généralement pas prises en compte. Les limites de cette méthode sont exacerbées dans un contexte multilingue avec la traduction de textes ou de requêtes. Le calcul de la similarité translingue

2. `http://dumps.wikimedia.org/`
3. Term frequency-inverse document frequency (Ramos, 2003)
4. Relative Average Term Frequency (Pirkola *et al.*, 2002)

peut être affecté par le mauvais choix de traductions candidates dans un dictionnaire. De plus, la couverture limitée des dictionnaires/systèmes de traduction en termes de mots nouveaux, de spécialité, ou encore d'unités polylexicales représente aussi un inconvénient considérable. Les mesures sémantiques sont une alternative à cette méthode.

3.2.2 Mesures sémantiques

Plusieurs études exploitent la structure de ressources sémantiques (Leacock et Chodorow, 1998; Jiang et Conrath, 1997) pour calculer la similarité de mots. Corley et Mihalcea (2005) proposent une adaptation de ces méthodes de mesure de la similarité lexicale monolingue à partir de WordNet à la similarité de segments textuels. Les segments textuels sont représentés par leurs mots pleins groupés en catégories grammaticales dont la similarité est mesurée par WordNet. La valeur de similarité entre deux phrases est obtenue par la moyenne des scores de similarité des paires de mots de même catégorie. La spécificité de mots est aussi prise en compte.

A notre connaissance, Il existe peu ou pas d'études utilisant cette approche pour une similarité translingue. Dans cet objectif, une ressource multilingue, à l'exemple de global wordNet [5] ou BabelNet [6] peut être utilisée. L'efficacité de cette méthode est étroitement liée à la qualité de la ressource sémantique utilisée. La structure de WordNet en hiérarchies différentes pour chaque partie du discours limite la portée des similarités qui ne peuvent être qu'intracatégorielles. De plus, la faible composition des hiérarchies autres que celle des noms remet en cause la fiabilité des valeurs de similarité obtenues pour des mots appartenant aux autres catégories grammaticales (verbes, adjectifs et adverbes). Aussi, dans une ressource multilingue, le problème du déséquilibre entre les sous-réseaux sémantiques des différentes langues impliquées se pose. L'absence d'informations sur le contexte pourrait également fausser la liaison des mots de documents aux concepts d'une ressource sémantique : d'où l'introduction de l'approche hybride.

3.2.3 Mesures hybrides

Les recherches s'orientent également vers des mesures de similarité hybrides qui par l'utilisation d'informations en corpus et d'une ressource sémantique, tirent avantage de chacune des approches ci-dessus.

Partant de l'hypothèse selon laquelle le sens d'un mot se détermine en contexte, Mohammad *et al.* (2007) proposent une méthode de calcul de la distance sémantique translingue des mots à travers la comparaison de leurs *profils distributionnels de concepts*. Les profils distributionnels de concepts sont composés des sens de mots non-ambigüs environnants qui permettent d'inférer le sens du mot cible. Pour calculer la similarité de paires de mots allemands, des profils distributionnels de concepts sont construits pour ces paires à partir de leur contextes d'occurrence et d'un thesaurus anglais. Le corpus allemand est mis en correspondance avec le thesaurus anglais à travers un lexique bilingue. Nous pensons que l'adaptation de cette méthode à la similarité de segments plus larges que le mot fournirait de bon résultats. L'utilisation unique d'un thesaurus multilingue sans un recours à des lexiques multilingues est aussi envisageable. Ainsi, Steinberger *et al.* (2002) exploitent les descripteurs du thesaurus multilingue EUROVOC [7]. Ils calculent par ce moyen la

5. http://www.globalwordnet.org
6. http://lcl.uniroma1.it/babelnet/
7. http://eurovoc.europa.eu/

similarité de documents de différentes langues à partir de leurs représentations conceptuelles. Un inconvénient de cette approche est l'effort considérable consacré à l'annotation manuelle d'une collection d'apprentissage en concepts du thesaurus pour l'annotation automatique de nouveaux documents à comparer.

Dans notre but de classifier des documents multilingues, nous nous orientons vers cette approche. Sa nature hybride réside dans le fait qu'elle permette de comparer des documents à l'aide d'une ressource sémantique et d'informations en corpus sur les mots dans les documents à comparer. La mesure de similarité ainsi définie est le principal élément pris en compte par les techniques de construction de corpus multilingues.

3.3 Méthodes de construction de corpus multilingues

Nous abordons dans cette section les principales approches d'acquisition et de structuration de documents multilingues. L'approche Web que nous aborderons est le crawling thématique. Celles pouvant être réalisées indépendamment du Web sont la RII et le clustering.

3.3.1 Le crawling thématique

Une méthode d'extraction de corpus du Web est le crawling qui consiste à utiliser les liens entre les pages pour collecter les documents. Les crawlers thématiques ont été développés pour identifier les sections du Web pertinentes par rapport à un thème donné . Talvensaari *et al.* (2008) utilisent cette méthode pour compiler un corpus comparable de spécialité anglais-espagnol-allemand. Les données de départ sont un ensemble d'URLs reflétant un sujet donné. Les pages correspondantes sont extraites et celles dont les liens figurent dans ces pages initiales sont également visités et prises en compte si le lien thématique avec les pages initiales peut être établi. L'identification du domaine dans les pages candidates se fait sur la base de terminologies collectées séparément pour chaque langue.

Puisque nous n'envisageons pas de développer des techniques d'extraction de documents de l'Internet nous nous intéressons aux approches indépendantes du Web ci-dessous.

3.3.2 Recherche d'information interlingue

Partant de collections de différentes langues, l'approche de RII est aussi utilisée pour collecter des textes comparables. Elle consiste en l'obtention à partir d'une collection source, de mots clés qui sont ensuite traduits et utilisés comme des requêtes exécutées sur la collection cible pour obtenir les documents souhaités. Talvensaari *et al.* (2007) proposent une approche de RII pour la construction d'un corpus comparable suédois-anglais. Les mots clés sont extraits des documents sources par la RATF. Leurs traductions sont exécutées comme requêtes sur la collection cible par le système de recherche d'information Indri qui fait partie du projet Lemur [8]. Les documents obtenus avec cette technique sont appariés ou classés en fonction de leurs scores de similarité avec un document source alors que la méthode de clustering dans la section suivante permet d'obtenir des *clusters* ou groupes de documents.

8. www.lemurproject.org

3.3.3 Clustering de documents

Le clustering de documents se définit comme la répartition d'un ensemble de textes dans des groupes selon leurs traits de similarité sans connaissances a priori. Les documents ayant des caractéristiques communes devraient apparaître dans le même *cluster* (Montalvo *et al.*, 2006). Réciproquement, les documents non- ou peu similaires devraient appartenir à des clusters distincts.

Pour créer des clusters de documents, Li *et al.* (2011) utilisent l'approche agglomérative ascendante. Ils obtiennent des clusters bilingues à partir d'une partie d'un corpus initial. Cette partie regroupe des textes au-dessus d'un seuil minimum de similarité qui serviront à former le corpus comparable. La même procédure est reproduite sur la partie restante du corpus initial par l'intégration de données externes et la création éventuelle de nouveaux clusters. Ertöz *et al.* (2003) utilisent quant à eux l'algorithme de clustering *Shared Nearest Neighbour (SNN)*. Selon cette méthode, deux documents ont plus de chance d'appartenir au même cluster s'ils ont en commun un nombre élevé de voisins. Ils la comparent à la méthode de *K-means* selon laquelle un document appartient à un cluster s'il est proche d'un nombre moyen de documents dans ce cluster (Ertöz *et al.*, 2003).

Nous privilégions le clustering qui va plus loin qu'un simple appariement ou alignement de documents similaires en organisant les documents en groupes.

4 Évaluation

4.1 Évaluation intrinsèque

La valeur d'une méthode de compilation de corpus multilingues peut être estimée à la qualité des données qui en résultent. Il s'agit d'une évaluation intrinsèque. La méthode courante est la comparaison des scores de similarité attribués automatiquement à ceux attribués manuellement. Une échelle de similarité est alors utilisée pour le jugement humain (voir section 2.1.2). Une corrélation faible entre ces deux types de résultats signifie généralement une mauvaise performance du système automatique étant donné que le jugement humain tient lieu de référence.

Pour évaluer leur méthode de calcul de la similarité prenant uniquement en compte les titres des articles, Aker *et al.* (2012) comparent la qualité des alignements obtenus avec celle des alignements produits lorsque le contenu entier de l'article est utilisé. Ils examinent en outre la correspondance entre les résultats de ces méthodes automatiques et la norme que sont les résultats humains. Li et Gaussier (2010) réalisent l'ensemble de l'évaluation automatiquement contournant ainsi l'effort d'annotation manuelle. Le corpus de référence est constitué par des corpus dont la comparabilité est graduellement réduite par l'import de données externes. La corrélation entre les scores de similarité des documents du corpus de référence ainsi construit et ceux obtenus automatiquement est calculée par le coefficent de Pearson. Steinberger *et al.* (2002) utilisent comme critère d'évaluation la capacité de leur méthode à identifier des documents parallèles en leur attribuant les scores de similarité les plus élevés.

4.2 Évaluation extrinsèque

Nous avons vu en section 2.2, les applications des corpus multilingues en TAL. La qualité des données multilingues obtenues est déterminée par leur apport dans ces applications. C'est le cas de l'étude de Talvensaari *et al.* (2007) dans laquelle le corpus comparable obtenu est utilisé comme un thésaurus de similarité accompagné d'un outil de traduction pour améliorer la traduction des requêtes et par ricochet la performance d'un système de RII. Bin *et al.* (2010) entraînent et testent un système de traduction automatique avec un corpus de phrases parallèles obtenu à partir d'un corpus comparable. Li *et al.* (2011) évaluent la qualité du corpus comparable obtenu dans l'application d'extraction de lexiques bilingues. Ils examinent en outre l'apport des lexiques bilingues obtenus en RII.

Des variantes de ces méthodes d'évaluation sont dans un cadre intrinsèque, la comparaison des résultats de plusieurs méthodes automatiques au jugement humain sur les mêmes données (Mihalcea *et al.*, 2006). Pour une évaluation extrinsèque, cela reviendrait à utiliser les données multilingues obtenues d'une même source par différentes techniques de similarité dans une même application et à comparer leurs apports.

Nous avons abordé le sujet de la construction de corpus multilingues en passant en revue quelques principes théoriques et les principales méthodes développées dans ce domaine. Il convient à présent de situer l'apport envisagé dans cette état de l'art.

5 Contributions envisagées

Dans cette partie, nous présentons les limites de la littérature et soulignons quelques perspectives futures notamment celles que nous envisageons de réaliser.

Comme nous l'avons observé en section 2.1.1, la comparabilité se définit dans les limites d'une application. En extraction de lexiques bilingues, elle se mesure généralement à la quantité de vocabulaire que des documents ont en commun (Li *et al.*, 2011). La comparabilité en termes de vocabulaire est-elle suffisante ? Dans un cadre de spécialité où terminologie et vocabulaire n'ont pas la même préséance, l'injection de connaissances du domaine ne serait-elle pas nécessaire à la comparabilité ? Nous prévoyons d'évaluer cette hypothèse de la comparabilité et éventuellement de l'affiner dans notre objectif de concevoir une mesure de similarité permettant d'obtenir des données hautement homogènes. Les mesures de similarité textuelle ont été largement explorées dans un cadre monolingue avec des méthodes basées sur des statistiques de corpus ou sur la structure de ressources ontologiques. Afin d'élaborer une mesure de similarité textuelle translingue et hautement sémantique, nous partirons de corpus multilingues faiblement comparables pour en extraire des sous corpus de meilleure qualité. Notre approche consistera dans la représentation conceptuelle de documents par des descripteurs d'un thesaurus en s'aidant des contextes d'occurrences des mots dans ces documents. Ceci permettra une meilleure attribution des concepts aux documents. Le calcul de la similarité textuelle se fera entre ces représentations. La prise en compte de critères supplémentaires permettra de parfaire la mesure de similarité pour une meilleure classification de documents. Nous adoptons comme approche de construction de corpus multilingues, le clustering qui nous permettra de former des sous corpus ou clusters à partir de notre collection de départ. Les techniques existantes de clustering ne traitent généralement pas de la mise à jour des clusters. Comment permettre l'intégration continuelle de nouveaux

textes à la base ? Le problème de clustering du flux de données restant ouvert, nous pensons que le clustering incrémental (Kurtz, 2012) serait une solution appropriée. Le corpus multilingue obtenu par cette méthode sera évalué en extraction de lexiques, plus précisément de néologismes multilingues. A notre connaissance, aucune étude dans le domaine n'a été entreprise dans cet objectif.

6 Conclusion

Dans cet article, nous avons abordé la construction de corpus multilingues sous plusieurs aspects. Nous avons pu constater que la notion de comparabilité et les applications de corpus multilingues en traitement automatique des langues sont étroitement liées. En effet, la définition de la comparabilité devrait se limiter dans un cadre applicatif donné. Les mesures de similarité textuelles ont généralement suivies les approches statistiques et sémantiques utilisées en majorité dans un contexte monolingue. Les approches hybrides multilingues sont un domaine de la similarité textuelle peu exploré. Pour ce qui est des techniques de clustering de documents existantes, entre autres les méthodes agglomérative hiérarchique, k-means et SNN, elles effectuent une classification ponctuelle et ne résolvent pas le problème de l'ajout permanent de textes à une base. Au vu des limites identifiées, nous prévoyons de fournir une description plus fine de la comparabilité pour l'extraction de lexiques multilingues. Nous envisageons également d'adapter une/plusieurs des méthodes de classification existante(s) au clustering incrémental. La mesure hybride développée déterminera la similarité de textes à travers leurs représentations conceptuelles pour l'application d'extraction de néologismes multilingues.

Références

Aker, A., Kanoulas, E. et Gaizauskas, R. (2012). A Light Way to Collect Comparable Corpora from the Web. *In Proceedings of LREC 2012*, pages 21–27, Istanbul, Turquie.

Baradaran Hashemi, H., Shakery, A. et Faili, H. (2010). Creating a Persian-English Comparable Corpus. *In Proceedings of the 2010 international conference on Multilingual and multimodal information access evaluation*, pages 27–39, Padoue, Italie.

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.

Bin, L., Jiang, T., Chow, K. et Benjamin K., T. (2010). Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT. *In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 42–49, La Valette, Malte.

Blancafort, H., Daille, B., Gornostay, T., Heid, U., Sharoff, S. et Méchoulam, C. (2010). TTC : Terminology Extraction, Translation Tools and Comparable Corpora. *In Proceedings of EURALEX 2010*, pages 263–268, Leeuwarden/Ljouwert, Pays-Bas.

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.

CORLEY, C. et MIHALCEA, R. (2005). Measuring the Semantic Similarity of Texts. *In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan, USA.

ERTÖZ, L., STEINBACH, M. et KUMAR, V. (2003). Finding Topics in Collections of Documents : A Shared Nearest Neighbor Approach. *Clustering and Information Retrieval*, 11:83–103.

FUNG, P. et YEE, L. Y. (1998). An IR approach for Translating New Words from Nonparallel, Comparable Texts. *In Proceedings of the 17th International Conference on Computational linguistics-Volume 1*, pages 414–420.

GOEURIOT, L., MORIN, E. et DAILLE, B. (2009). Compilation of Specialized Comparable Corpora in French and Japanese. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, pages 55–63, Suntec, Singapore.

ION, R. (2012). PEXACC : A Parallel Sentence Mining Algorithm from Comparable Corpora. *In Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2181–2188, Istanbul, Tuquie.

ION, R., TUFIS, D., BOROS, T., CEAUSU, A. et STEFANESCU, D. (2010). On-Line Compilation of Comparable Corpora and their Evaluation. *In FASSBL7*, pages 29–33, Dubrovnik, Crotia.

JIANG, J. J. et CONRATH, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997*, Taipei, Taiwan.

KURTZ, C. (2012). Une distance hiérarchique basée sur la sémantique pour la comparaison d'histogrammes nominaux. *In Actes de Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance*, pages 77–88, Bordeaux, France.

LEACOCK, C. et CHODOROW, M. (1998). Combining Local Context and WordNet Similarity for Word Wense Identification. *In WordNet : An electronic lexical database*, page 265–283. Fellbaum, C., Cambridge, MA, MIT Press édition.

LETURIA, I., SAN VICENTE, I. et SARALEGI, X. (2009). Search Engine Based Approaches for Collecting Domain-specific Basque-English Comparable Corpora from the Internet. *In Proceedings of the Fifth Web as Corpus Workshop*, pages 53–61, Donostia-San Sebastian, Basque Country, Spain.

LI, B. et GAUSSIER, E. (2010). Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, pages 644–652, Beijing, China.

LI, B., GAUSSIER, E., MORIN, E. et HAZEM, A. (2011). Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. *In 18ème conférence sur le Traitement Automatique des Langues Naturelles*, Montpellier, France.

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.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *In Proceedings of the 21st national conference on Artificial intelligence*, volume 1, pages 775–780, Boston, MA, USA.

MOHAMMAD, S., GUREVYCH, I., HIRST, G. et ZESCH, T. (2007). Cross-lingual Distributional Profiles of Concepts for Measuring Semantic Distance. *In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 571–580, Prague, République tchèque.

MONTALVO, S., MARTINEZ, R., CASILLAS, A. et FRESNO, V. (2006). Multilingual Document Clustering : an Heuristic Approach Based on Cognate Named Entities. *In Proceedings of the 21st International Conference on Computational Linguistics*, volume 44, pages 1145–1152, Sydney, Australie.

PARAMITA, M., CLOUGH, P., AKER, A. et GAIZAUSKAS, R. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 790–797, Istanbul, Tuquie.

PIRKOLA, A., LEPPÄNEN, E. et JÄRVELIN, K. (2002). The RATF Formula (Kwok's formula) : Exploiting Average Term Frequency in Cross-language Retrieval. *Information Research*, 7(2):7–2.

RAMOS, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. *In Proceedings of the First Instructional Conference on Machine Learning*, Piscataway, NJ USA.

RESNIK, P. (1999). Mining the Web for Bilingual Text. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA.

SKADIŅA, I., AKER, A., GIOULI, V., TUFIŞ, D., GAIZAUSKAS, R., MIERIA, 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.

STEINBERGER, R., POULIQUEN, B. et HAGMAN, J. (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus Eurovoc. *In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 415–424, Mexico, Mexique.

SU, F. et BABYCH, B. (2012). Measuring Comparability of Documents in Non-parallel Corpora for Efficient Extraction of (Semi-)Parallel Translation Equivalents. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 10–19, Avignon, France.

TALVENSAARI, T., LAURIKKALA, J., JÄRVELIN, K., JUHOLA, M. et KESKUSTALO, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM TOIS*, 25(1):4.

TALVENSAARI, T., PIRKOLA, A., JÄRVELIN, K., JUHOLA, M. et LAURIKKALA, J. (2008). Focused Web Crawling in the Acquisition of Comparable Corpora. *IR*, 11(5):427–445.

UTSURO, T., HORIUCHI, T., CHIBA, Y. et HAMAMOTO, T. (2002). Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-lingually Relevant News Articles onWWW News Sites. *In Proccedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA) : From Research to Real Users*, pages 165–176, Tiburon, CA, USA.

YAPOMO, M., CORPAS, G. et MITKOV, R. (2012). CLIR- and Ontology-based Approach for Bilingual Extraction of Comparable Documents. *In Proceedings of the 5th Workshop on Building and Using Comparable Corpora*, pages 121–125, Istanbul, Turquie.