Mining Documents and Sentiments in Cross-lingual Context

Motaz Saad

Université de Lorraine
École doctorale IAEM Lorraine

THÈSE

Fouille de documents et d'opinions multilingue

(Mining Documents and Sentiments in Cross-lingual Context)

Auteur

Motaz Saad

Présentée et soutenue publiquement le 20 janvier 2015

pour l'obtention du

Doctorat de l'Université de Lorraine

(Spécialité informatique)

Composition du jury

Rapporteurs:	Eric GAUSSIER	Professeur, Université Joseph Fourier
	Pascal PONCELET	Professeur, Université Montpellier 2
Examinateurs:	Emmanuel MORIN	Professeur, Université de Nantes
	Khalid BENALI	HDR, Université de Lorraine
Directeur de thèse:	Kamel SMAÏLI	Professeur, Université de Lorraine
Co-Encadrant:	David LANGLOIS	MCF, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Abstract

Comparable texts are a set of topic aligned documents in multiple languages, which are not necessarily translations of each other. These documents are informative because they can tell what is being said about a topic in different languages.

The aim of the thesis is to compare sentiments and emotions in comparable documents. For this, we define two objectives: the first one is to collect corpora. That is why this dissertation presents a method to retrieve comparable documents. The second objective is to study these documents by annotating them with sentiment labels, then compare sentiments in these comparable documents.

As part of the first objective, we collect comparable corpora from Wikipedia and Euronews in Arabic, English and French languages. The corpora are aligned at the document level and made available publicly for research purposes.

To retrieve and align English-Arabic comparable documents, we experiment two cross-lingual similarity measures: one is based on a bilingual dictionary, and the other is based on Latent Semantic Indexing (LSI). Experiments on several corpora show that the cross-lingual LSI (CL-LSI) measure outperforms the dictionary based measure. The CL-LSI does not need machine translation systems or morphological analysis tools, and it overcomes the problem of vocabulary mismatch between training and test documents.

Concerning the second objective, which is to analyze comparable documents collected from Internet. we collect another collection of comparable news documents from local and foreign sources. These documents are collected from the British Broadcast Corporation (BBC) and Aljazeera (JSC) news websites. Then the CL-LSI measure is used to align comparable news documents of BBC and JSC. The evaluation of BBC-JSC document alignments show that CL-LSI is not only able to align cross-lingual documents at the topic level, but also able to do so at the event level. This aligned corpus (BBC-JSC) allows us to annotate document pairs with sentiment and emotion labels, then compare the agreement of these annotations in each documents pair.

To provide English-Arabic corpus annotated with sentiment labels (subjective and objective), we propose a cross-lingual annotation method. The cross-lingual annotation method projects sentiment labels from one topic domain to another and from

one language to another. We, in fact, transfer annotations from movie reviews domain to other domains (news and talks), and from English to Arabic. The advantage of this method is that it produces resources in multiple languages without the need of a machine translation system. The resulting corpus is useful to build sentiment classifiers. We use these classifiers to annotate comparable documents with sentiment labels.

To be able to annotate English-Arabic document with emotion labels (anger, disgust, fear, joy, sadness, and surprise), we manually translate the WordNet-Affect (WNA) emotion lexicon into Arabic. We use the English-Arabic WNA emotion lexicons to annotate English-Arabic comparable documents. These annotations allow to compare the emotions in the comparable documents.

At the end, we focus on the comparison of sentiments and emotions in comparable documents. This comparison is interesting, especially when the source and the target documents are from different sources. To our knowledge this is not discussed in the literature. In this work, we inspect the pairwise agreement between sentiments expressed in the source and target comparable documents using statistical agreement measures. The agreement is inspected for each pair (source and target) of documents. We study the agreement between source and target comparable documents of Euronews corpus. We also repeat the same experiment but on BBC-JSC comparable corpus. Results indicate that BBC-JSC comparable documents diverge from each other in terms of sentiments, while source and target documents of Euronews have higher agreement than the ones of BBC-JSC corpus.

It can be concluded from this thesis that studying comparable documents is a promising research field. We provided in this thesis language independent methods to align comparable articles (CL-LSI measure) and to annotate them with sentiment labels (the cross-lingual annotation method), and the statistical agreement measures to compare sentiments and emotions in comparable document pairs.

Keywords: text mining; natural language processing; comparable corpus; cross-lingual information retrieval; cross-lingual projection; sentiment analysis

Résumé

Les documents comparables sont des textes écrits dans des langues différentes et dont le contenu concerne le même sujet. Ces documents ne sont pas nécessairement des traductions les uns des autres. L'utilité de ces documents réside dans le fait qu'on offre à des usagers de langues différentes un accès à l'information dans la langue maternelle.

Le but de cette thèse est d'étudier les documents comparables sur le plan des opinions et des sentiments. Pour ce faire, nous définissons deux objectifs : le premier consiste à collecter des corpus avec des méthodes que nous avons développées, alors que le second concerne l'exploitation des documents comparables collectés une fois étiquetés par des tags de type "sentiment" ou "opinion".

Concernant le premier objectif, nous avons collecté des corpus comparables de Wikipedia et Euronews en arabe, français et anglais. Ces corpus sont alignés au niveau du document et sont à la disposition de la communauté scientifique.

Pour récupérer et aligner les documents comparables anglais-arabe, nous expérimentons deux mesures de similarité inter-langues (IL) : la première fondée sur un dictionnaire bilingue et la seconde fondée sur l'indexation sémantique latente (LSI). Des expérimentations sur plusieurs corpus ont montré que la mesure inter-langues fondée sur la LSI (IL-LSI) a surclassé la mesure basée sur le dictionnaire. En plus, la IL-LSI n'a besoin ni de système de traduction automatique ni d'outils d'analyse morphologique. Par ailleurs, cette méthode ne souffre pas d'un manque de couverture des deux vocabulaires.

En ce qui concerne le deuxième objectif, qui est d'analyser des documents comparables recueillis à partir d'Internet. Tout d'abord, nous avons rassemblé des documents comparables d'information de sources locales et étrangères. La mesure IL-LSI est utilisée pour aligner les documents de la BBC et du site web d'Aljazeera (JSC). L'évaluation de l'alignement des documents BBC-JSC a montré que IL-LSI est non seulement capable d'aligner les documents de deux langues différentes traitant le même sujet, mais aussi traitant le même événement. Ces corpus alignés sont ensuite utilisés pour étudier les documents en termes de sentiments à identifier parmi une liste prédéfinie.

Pour fournir des corpus anglais-arabe annotés par des étiquettes de sentiments comme la joie, la colère..., nous avons proposé une méthode d'annotation qui projette les annotations des sentiments d'un domaine à un autre et d'une langue à une autre. L'avantage de cette méthode est qu'elle produit des ressources dans plusieurs langues sans avoir besoin d'un système de traduction automatique. Le corpus résultant est utile pour construire des classifieurs de sentiments. En utilisant cette méthode, nous avons projeté les annotations du domaine des critiques d'un film à d'autres domaines (les actualités et les discours), et de l'anglais à l'arabe. Ensuite, nous avons utilisé les classifieurs résultants pour annoter d'autres documents comparables avec des étiquettes de sentiments.

Pour pouvoir annoter des corpus anglais-arabe avec des étiquettes d'émotions (colère, dégoût, peur, joie, tristesse et surprise), nous traduisons manuellement WordNet-Affect (WNA) en arabe. Le lexique obtenu est ensuite utilisé pour étiqueter des documents comparables arabe-anglais.

La comparaison en termes d'opinions est intéressante lorsque le document source et cible sont exprimés dans des langues différentes. Dans ce travail, nous comparons la convergence ou la divergence des sentiments dans les paires de langue à l'aide de mesures statistiques. L'étude porte sur les corpus comparables provenant de la même source : Euronews et sur des corpus provenant de sources différentes BBC et JSC. L'expérience montre en effet que les opinions divergent entre deux sources comme la BBC et JSC alors qu'elles convergent dans les documents d'Euronews.

En conclusion, l'étude des documents comparables en termes de sentiments est très prometteuse. En effet, plusieurs applications peuvent être envisagées comme la revue de presse d'opinions. Nous avons proposé dans cette thèse des méthodes indépendantes de la langue pour aligner des articles comparables (utilisation par exemple de la méthode IL-LSI), pour les annoter en termes d'opinions et mesurer la convergence ou la divergence des opinions.

Mots-clés: fouille de textes; traitement automatique du langage naturel; corpus comparable; recherche d'information inter-langues; projection inter-langues; analyse des sentiments

Acknowledgements

Thanks to my supervisors Kamel Smaïli and David Langlois for their guidance and help and their constant support.

Thanks to my family for their patience and support.

Thanks to all my friends and colleagues for being there with me.

Contents

List of Figures

List of Tables

Abbreviations

VSM:	**V**ector **S**pace **M**odel
BOW:	**B**ag **O**f **W**ords
IR:	**I**nformation **R**etrieval
LSI:	**L**atent **S**emantic **I**ndexing
LSA:	**L**atent **S**emantic **A**nalysis
CL-IR:	**C**ross-**L**ingual **I**nformation **R**etrieval
CL-LSI:	**C**ross-**L**ingual **L**atent **S**emantic **I**ndexing
SVM:	**S**upport **V**ector **M**achines
NB:	**N**aive **B**ayes
POS:	**P**art **O**f **S**peech
MT:	**M**achine **T**ranslation
LM:	**L**anguage **M**odel
NLP:	**N**atural **L**anguage **P**rocessing
XML:	**E**xtensible **M**arkup **L**anguage

Chapter 1

Introduction

1.1 Motivation

The number of Internet users has already exceeded two billions [Miniwatts, 2012]. As a result, Internet has became the major source of knowledge. Nowadays, people have access to various information about products, news, books, movies, public services, science, etc. These unstructured knowledges are provided from various sources such as news feeds, encyclopedias, blogs, forums and social networks. But such a huge data exceeds the human capacity of understanding. Consequently, users need to use Information Retrieval (IR) technologies to find information that is relevant to their needs.

The problem appears not only due to the amount of data, but also because of the variety of languages, which makes it more challenging. The emergence of Web 2.0 technologies enlarged web contents in many languages. According to [Miniwatts, 2012], the top ten languages of the Internet (by the number of users) are English, Chinese, Spanish, Japanese, Portuguese, German, Arabic, French, Russian and Korean. In addition, Arabic is one of the most rapidly growing languages in the period from 2000 to 2011 [Miniwatts, 2012, Wikimedia, 2014]. In such case, relevant information

do exist in a foreign language, which might be not understandable by a user. For this reason, IR technologies should handle cross-lingual texts. Cross-Lingual Information Retrieval (CL-IR) methods can find relevant information in cross-lingual texts.

The CL-IR system should unify the language of queries and documents. Machine Translation (MT) systems are usually used with CL-IR systems to help users to understand the retrieved information regardless of the source language.

In multilingual web contents, many documents are comparable. Such cross-lingual documents are related to the same topic, but they are not necessarily translations of each other. The topic can be a review of a product, a book, a recipe, a biography of a person, a news event, etc. Comparable documents are informative when one is interested in what is being said about a topic in the other languages. One can also compare and inspect sentiments expressed in the comparable documents. For instance, many non-Arabic speaking journalists were interested in what was being said in Arabic posts in social media during the Egyptian revolution in 2011. During that time, automatic translation services were provided by Twitter and other organizations to understand these posts [Arthur, 2011, Diehn, 2013]. A related but more interesting use case could be as follows: a journalist can search for a topic, and a list of comparable documents can match to his query. This list can be structured according to the sentiments that are expressed in these documents. In this case, contents of comparable documents can diverge or converge in terms of sentiments. Another use case, but in a different context is when a user is interested in buying a product, that is not well-known or not used in his country. Such a user can be interested in reviews of this product, which are written in other languages. Moreover, even if the product is well-known or used in his country, he/she may be interested in whether comparable reviews are similar or different.

The amount of comparable texts in the mentioned use cases can be large. Reading and comparing all these texts requires a great deal of efforts. Therefore, comparing these documents automatically can be useful.

To check whether the sentiments diverge or converge in comparable documents, sentiments need to be identified first. The process of identifying sentiments in the text is called sentiment analysis or opinion mining [Pang and Lee, 2008]. Sentiment analysis consists in the analysis of the subjectivity or the polarity of the text. Subjectivity analysis classifies a text into subjective or objective, while polarity analysis classifies the text into positive or negative [Pang and Lee, 2008]. Sentiments need to be detected in source and target comparable documents first, then the agreement of sentiments can be compared in these documents.

1.2 Overview

This section gives an overview of this work. This dissertation provides and evaluate methods and algorithms to collect, align, and retrieve comparable documents. It also provides methods to annotate and compare the agreement of comparable documents in terms of sentiments. In Chapter 2, we review some works from the literature, which are related to these methods.

In this work, we focus on English-Arabic comparable documents. Studying English-Arabic comparable documents is interesting because sentiments expressed in these documents may diverge for some topics due the political situation in the Arabic regions associated with the emergence of so-called "Arab Spring". In Chapter 3, we collect comparable documents from Wikipedia encyclopedia[1] and Euronews[2] websites, then we align them based on links. We use the collected corpora to study comparable texts and to develop our methods that we propose in this thesis. Such kind of resources are not readily available, and will be useful for different analyses as discussed in this thesis.

[1] www.wikipedia.org
[2] www.euronews.com

In Chapter 4, we present two cross-lingual similarity measures that are used to retrieve and align cross-lingual documents. We compare the performance of these measures and choose the best one to perform the alignment to further cross-lingual documents.

In our work, we need English-Arabic sentiment resources to build classifiers that can be used to automatically annotate comparable texts. Therefore, we propose in Chapter 5 a cross-lingual annotation method to provide these resources. Such kind of annotated resources are useful because they are not available.

In Chapter 6 we describe the statistical measures that we use to compare the agreement of sentiments in comparable documents. We present experiments in this chapter on comparable news documents collected from different sources.

Figure 1.1 shows a simplified block diagram representing the main steps followed for the work discussed in this dissertation. First, we collect and align comparable documents as described in Chapter 3. Then for a given English document, we retrieve the most comparable Arabic document(s) using the cross-lingual similarity measure proposed in Chapter 4. The retrieved documents can be screened manually to make sure if they are comparable or not as shown in the figure. Next, we automatically annotate the aligned pairs with sentiment labels as described in Chapter 5. Finally, we compare the agreement of sentiments as described in Chapter 6.

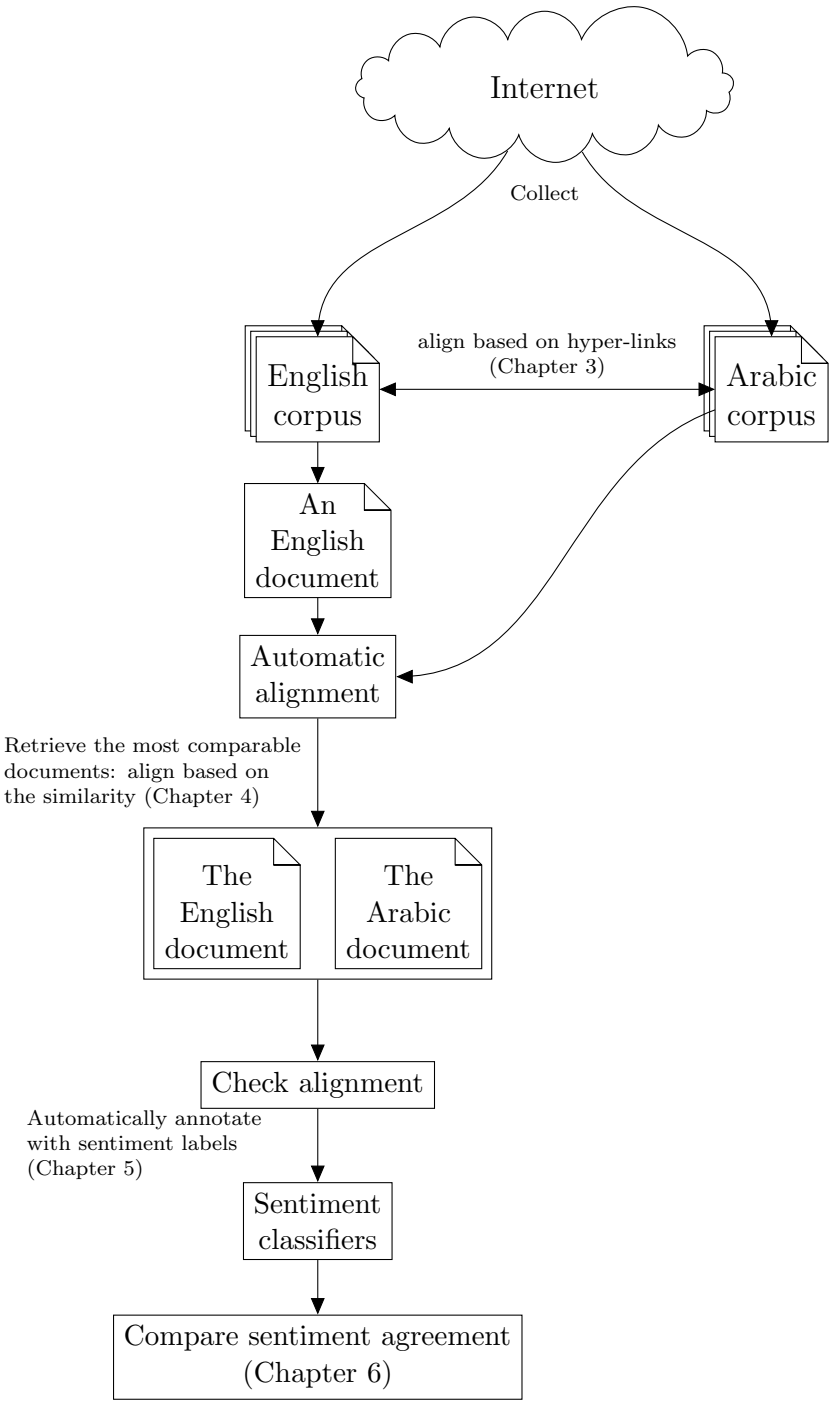

FIGURE 1.1: Work overview

Chapter 2

Related Work

This chapter reviews previous work on collecting comparable corpora, cross-lingual similarity measures, and sentiment analysis.

2.1 Comparable Corpora

A parallel corpus is a collection of aligned sentences, which are translations of each other. In contrast, a comparable corpus is a collection of topic aligned documents, which are not necessarily translations of each other.

Non-parallel (comparable) corpora can have different levels of comparability. In [Fung and Cheung, 2004], the authors proposed three levels for non-parallel corpora. These levels are *noisy-parallel*, *comparable* and *quasi-comparable* corpora. Texts in the *noisy-parallel* corpus have many parallel sentences roughly in the same order. Texts in the *comparable* corpus have topic aligned documents, which are not necessarily translations of each other. As for the *quasi-comparable* corpus, it has bilingual documents that are not necessarily related to the same topic.

Parallel and comparable corpora are useful for several tasks, such as cross-lingual text mining [Tang et al., 2011], bilingual lexicon extraction [Li and Gaussier, 2010], cross-lingual information retrieval [Knoth et al., 2011] and machine translation [Delpech, 2011]. Some of these tasks use statistical methods, which require a large amount of training data. Normally, for data-driven methods, larger amounts and better quality of data leads to better results.

Unfortunately, parallel corpora are not available for some domains and/or languages. For some language pairs in some domains, few amount of parallel corpora are available. Parallel corpora can be acquired using human translators, but this is expensive and requires a lot of efforts and time. Therefore, comparable corpora are the best alternative in this case, because they are cheaper and can produce more data. Several researchers already elaborated their methods to handle comparable corpora instead of parallel corpora. For instance, [Smith et al., 2010, Delpech, 2011] developed some techniques to improve the quality of machine translation using comparable corpora.

Comparable corpora can be obtained easily from the web. Newspaper websites and Encyclopedias are ideal for collecting comparable documents. But aligning these texts is a challenging task.

One of the attractive sources of collecting comparable corpora is the Wikipedia encyclopedia, because it covers many languages and domains. The alignment of Wikipedia documents is generally based on the document links that refer to the equivalent articles in other languages. These links are called *inter-language* links. This link refers to the corresponding articles in other languages. The form of these links is [[$languagecode$: $Title$]]. For instance, considering the English Wikipedia article, which is about "Tunisian Revolution", the *inter-language* link of the English language article is *[[en:Tunisian Revolution]]*, while the link for the French article is *[[fr:Révolution tunisienne de 2010-2011]]*. These links do exist in all the languages that the article is written in.

Many researchers collected comparable corpora from Wikipedia, for example, authors of [Otero and López, 2009] collected Spanish-Portuguese-English corpus, while authors of [Ion et al., 2010] collected English-Romanian corpus. [Otero and López, 2009] aimed to develop a cross-lingual similarity measure to extract parallel texts from Wikipedia corpus, while [Ion et al., 2010] collected their comparable corpus to improve the performance of statistical machine translation system.

Most of researchers are interested in comparable corpora because they can be used to extract parallel texts for different purposes. This interest continues to receive increased attention from the research community. For example, ACCURAT[1] project [Pinnis et al., 2012, Skadina et al., 2012] is a research project dedicated to find methods and techniques to overcome the problem of lacking linguistic resources for under-resourced languages and narrow domains. The ultimate objective is to exploit comparable data to improve the quality of machine translation for such languages. The ACCURAT project research for methods to measure comparable corpora and use them to achieve the project's objective. The project also research methods for alignment and extraction of lexical resources.

Other researchers also investigated extracting parallel texts from comparable corpora to improve machine translation. For instance, the authors in [Smith et al., 2010] used machine learning techniques to train a classifier on word-alignment features extracted by the IBM model-1 [Brown et al., 1993]. These features are extracted from a parallel corpus. The form of these features is the source-target word pairs. The classifier is then used to decide if a sentence pair is parallel or not. The authors developed a baseline for machine translation. The baseline system is trained on union of Europarl[2] and JRC-Acquis[3] parallel corpora. Then the authors merged these parallel corpora with the parallel texts extracted from a comparable corpus collected from Wikipedia.

[1] `www.accurat-project.eu`
[2] `www.statmt.org/europarl/`
[3] `http://langtech.jrc.it/JRC-Acquis.html`

The authors showed that the performance of their system has better performance than the baseline, and they concluded that the extracted parallel texts improved the performance of the machine translation.

[Abdul-Rauf and Schwenk, 2011] aimed to generate a parallel corpus from a comparable one. The objective is to generate a training data for machine translation. The authors used machine translation in combination with information retrieval (IR) system to extract parallel text. The source texts of the comparable corpus are translated using the machine translation system, then they are used as queries to the IR system to retrieve the candidate parallel texts from the target data. The method is applied on English-Arabic and English-French language pairs. The parallel corpus is extracted from Arabic, English and French Gigaword corpora, which are provided from LDC[4]. The authors investigated whether the extracted texts improves the performance of the machine translation. For that, they developed two systems. The baseline system is trained on parallel data obtain from NIST evaluation[5]. The other system is trained on the union of NIST data and extracted parallel corpus. The authors reported that the latter system has better performance than the baseline system.

To summarize the reviewed works in this section, the comparability of multilingual corpora can be in different levels [Fung and Cheung, 2004]. Researchers are mainly interested in extracting parallel texts from comparable corpora for different applications (mainly to improve the quality of machine translation for under-resourced languages [Pinnis et al., 2012, Skadina et al., 2012]). In this thesis, our interest in comparable corpora is different. We are interested in studying comparable documents in terms of sentiments. We compare the agreements of sentiments labels in comparable documents, and we investigate the convergence and divergence of the agreement.

In the next section, we review some proposed measures for cross-lingual similarity.

[4] `http://catalog.ldc.upenn.edu`
[5] `www.itl.nist.gov/iad/mig/tests/mt`

2.2 Cross-lingual Similarity Measures

A document similarity measure is a function that quantifies the similarity between two text documents [Manning and Raghavan, 2008]. This function gives a numeric score to quantify the similarity. In data mining and machine learning, Euclidean or Manhattan metrics are usually used to measure the similarity of two objects, while in text mining and information retrieval, cosine similarity is commonly used to measure the similarity of two text documents. Cosine similarity is computed as follows:

$$cosine(d_s, d_t) = \frac{d_s \cdot d_t}{\|d_s\| \times \|d_t\|} = \frac{\sum_{i=1}^{n} d_{s_i} \times d_{t_i}}{\sqrt{\sum_{i=1}^{n} (d_{s_i})^2} \times \sqrt{\sum_{i=1}^{n} (d_{t_i})^2}} \tag{2.1}$$

where d_s and d_t are document vectors. To generate a document vector, the text document is transformed into Bag Of Words (BOW), i.e., to treat the text document as a set of unstructured words.

Representing documents in a collection as BOW is called Vector Space Model (VSM) or term-document matrix. Usually, most important words are considered and less important words can be ignored. For instance, a pre-defined list of stop words (the, to, from, ...) are usually removed since they have no sense in the unstructured BOWs. In addition, stemming or lemmatization are usually applied to reduce words into their base form (root or stem). This will address the problem of word variability when estimating the similarity between documents. The number of dimensions in VSM is equal to the number of unique terms in the document collection. In large documents collection, the term-document matrix becomes sparse.

Equation (2.2) shows *term-document* matrix, which represents terms and documents of the collection. Rows of this matrix correspond to the unique terms and columns correspond to the documents in the collection. w_{ij} is the weight of the term i in the

document j in the collection. The weights can be the Term Frequency (tf) or the Term frequency-inverse document frequency (*tfidf*).

$$\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & \dots & d_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ w_{31} & w_{32} & w_{33} & \dots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \dots & w_{mn} \end{pmatrix}_{m \times n} \end{array} \tag{2.2}$$

The *tfidf* reflects how a term is important to a document in a corpus. The *tfidf* scheme increases the weight to words that appear frequently in the document, but this increase is controlled by the frequency of the word in the corpus (document frequency). Thus, common words, that appeared in the most of documents, get less weight (less discriminative) than words that appeared in some documents (more discriminative). The *tfidf* for a term t_i, in a document d_j, in a corpus C is computed as follows:

$$tfidf(t_i, d_j, C) = tf(t_i, d_j) \times idf(t_i, C) \tag{2.3}$$

where $tf(t_i, d_j)$ is the frequency of the term t_i in the document d_j, and $idf(t_i, C)$ is the frequency of documents that the term t_i appeared in. $tf(t_i, d_j)$ and $idf(t_i, C)$ are computed as follows:

$$tf(t_i, d_j) = |t_i : t_i \in d_j| \tag{2.4}$$

$$idf(t_i, C) = \log \frac{|C|}{|\{d \in C : t_i \in d\}|} \tag{2.5}$$

where $|C|$ is the corpus size.

Table 2.1 shows a sample of texts documents and Table 2.2 shows the term-document matrix for these documents collection. For instance, the BOWs of d_1 are ('NLP', 'make', 'computer', 'understand', 'human', 'language'). Table 2.3 shows the *tfidf* weights computed for each term in the collection.

TABLE 2.1: A sample of text documents

Document	Contents
d_1	NLP is to make the computer understand the human language.
d_2	NLP stands for natural language processing, NLP is useful.
d_3	Processing human language is a field of computer science, a language is a method of communication.
d_4	Information technology is a related field to computer science.
d_5	Processing data produces information.

TABLE 2.2: Term-document matrix (frequency)

Terms	d_1	d_2	d_3	d_4	d_5
NLP	1	2	0	0	0
natural	0	1	0	0	0
stands	0	1	0	0	0
language	1	1	2	0	0
information	0	0	0	1	1
produces	0	0	0	0	1
communication	0	0	1	0	0
make	1	0	0	0	0
processing	0	1	1	0	1
related	0	0	0	1	0
data	0	0	0	1	0
field	0	0	1	1	0
computer	1	0	1	1	0
understand	1	0	0	0	0
human	1	0	1	0	0
science	0	0	1	1	0
useful	0	1	0	0	0
technology	0	0	0	1	0
method	0	0	1	0	0

TABLE 2.3: Term-document matrix (*tfidf*)

Terms	d_1	d_2	d_3	d_4	d_5
NLP	0.40	0.80	0.00	0.00	0.00
natural	0.00	0.70	0.00	0.00	0.00
stands	0.00	0.70	0.00	0.00	0.00
language	0.22	0.22	0.44	0.00	0.00
information	0.00	0.00	0.00	0.40	0.40
produces	0.00	0.00	0.00	0.00	0.70
communication	0.00	0.00	0.70	0.00	0.00
make	0.70	0.00	0.00	0.00	0.00
processing	0.00	0.22	0.22	0.00	0.22
related	0.00	0.00	0.00	0.70	0.00
data	0.00	0.00	0.00	0.00	0.70
field	0.00	0.00	0.40	0.40	0.00
computer	0.22	0.00	0.22	0.22	0.00
understand	0.70	0.00	0.00	0.00	0.00
human	0.40	0.00	0.40	0.00	0.00
science	0.00	0.00	0.40	0.40	0.00
useful	0.00	0.70	0.00	0.00	0.00
technology	0.00	0.00	0.00	0.70	0.00
method	0.00	0.00	0.70	0.00	0.00

A cross-lingual similarity measure is a special case of document similarity measure, where the source and target documents are written in different languages. The cross-lingual measure can be used to identify a pair of comparable documents, i.e., to retrieve a target document for a given source document.

Many methods have been proposed to measure the similarity of comparable documents. These methods are based on bilingual dictionaries, on Cross-Lingual Information Retrieval (CL-IR), or based on Cross-Lingual Latent Semantic Indexing (CL-LSI). These methods are described in the next sections.

2.2.1 Dictionary-based Methods

In dictionary-based methods, two cross-lingual documents d_s and d_t are comparable if most of words in d_s are translations of words in d_t. A bilingual dictionary can be

used to look-up the translations of words in both documents.

Matched words in source and target documents can be weighted using binary or *tfid* weighting schemes. The similarity can be measured on binary terms ($1 \rightarrow$ term present, $0 \rightarrow$ term absent) as follows:

$$sim(d_s, d_t) = \frac{|w_s \rightarrow w_t| + |w_t \rightarrow w_s|}{|d_s| + |d_t|} \tag{2.6}$$

where $|w_s \rightarrow w_t|$ is the number of source words (w_s) that have translations in the target document (d_t), and $|w_t \rightarrow w_s|$ is the number of target words (w_t) that have translations in the source document (d_s). $|d_s| + |d_t|$ is the size (number of words) of the source and the target documents.

The similarity can also be measured on weighted terms (*tfidf*) using the cosine similarity. In this case, source and target document vectors are generated from the matched words (translation of each others) between source and target documents.

Several similarity measures based on bilingual dictionary have been proposed. Some of them consider the similarity at the corpus level [Li and Gaussier, 2010], while others consider it at the document level [Otero and López, 2011].

In [Li and Gaussier, 2010], a bilingual dictionary have been used to measure the similarity of the source and the target corpus by inspecting the translation of each word of the source vocabulary in the target vocabulary. The objective is to improve the quality of the comparable corpus to extract a bilingual lexicon of good quality. For that, the authors proposed a cross-lingual similarity (comparability) measure, and a strategy to improve the comparability of a corpus through extracting high comparable texts in iterative procedure. The authors assumes that having a good quality corpus leads extracting a good bilingual lexicon. The authors applied their work on French-English documents. The French-English bilingual dictionary used in their experiments is composed of 75K translation pairs. The authors defined the

comparability measure as the proportion of the English words that have translations (via the bilingual dictionary) in the French corpus, and the proportion of the French words that have translations in the English corpus to the whole corpus. The authors applied lemmatization on English and French words before matching them using the bilingual dictionary.

In their work, the comparability is considered at corpus level. The authors evaluated their measure on parallel corpora (Europarl[6], TREC[7] and CLEF[8]). The authors split each parallel corpus into 10 parts (P_1, P_2, P_3, ..., P_{10}). A certain portion of English documents from each part (P_i) is replaced with English documents from another corpus. In result, a new and less comparable corpus is formed (P'_i). The portion size, which is to be replaced, is increased for each P_i. Thus, there are parts with variant comparability. The authors measured the comparability for each P'_i, and they showed that the measure captured the differences between all P'_i.

The authors then used their proposed measure to improve the quality of Wikipedia comparable corpus C_w. From the English Wikipedia, the authors extracted 367k English documents from (Society) category, and from The french Wikipedia, they extracted 378k French documents from (Société) category. Their method consists in two steps. In the first step, the authors extract highly comparable parts from C_w. In the second step, the authors enrich the low comparable parts of C_w with texts from other sources.

The highly comparable corpus is extracted in the first steps as follows: documents are extracted incrementally such that their comparability is larger than a certain threshold. The process stops when the comparability of the documents rested in C_w is less than the threshold.

To enrich the remaining part of C_w, which is a low comparable corpus, the authors take the French documents separately, then their comparable documents (English) are

[6] `www.statmt.org/europarl/`
[7] `http://trec.nist.gov/`
[8] `www.clef-initiative.eu/`

fetched from an external source (corpus E). The authors repeat the same procedure with the remaining English documents in C_w. The authors added this enriched corpus to the highly comparable corpus, which is extracted in the first step, to constitute a larger highly comparable corpus.

From 367k and 378k English and French Wikipedia documents, the authors extracted 20k highly comparable English-French document pairs. They used 0.3 as a threshold for the comparability degree. The rest of C_w is enriched from external corpora. The external English corpora are Los Angeles Times[9] (LAT94) and Glasgow Herald[10] (GH95), while the external French corpora are Le Monde[11] (MON94), SDA French[12] 94/95 (SDA94/94). After enrichment process, the total size of the highly comparable corpus is 54k English-French document pairs. To evaluate the comparability of extracted corpora (highly comparable corpus and enriched corpus), the authors applied their comparability measure on the original Wikipedia corpus C_w and the high comparable corpus. They found that the highly comparable and enriched corpora are more comparable than the original Wikipedia corpus (C_w).

Finally, after obtaining the highly comparable corpus, the authors investigated different methods to extract bilingual lexicon from this corpus. The authors showed that the lexicons extracted from the highly comparable corpus have better precisions compared to the lexicon extracted from low comparable corpus.

Later, the authors proposed another method in [Li et al., 2011] to improve corpus comparability. This method is proposed for the same purpose, which is to extract bilingual dictionary. Their approach is based on clustering cross-lingual documents to improve the comparability of the corpus. The authors assumes that having a homogeneous groups of cross-lingual documents (clusters), leads to extract a better quality lexicon. Their cross-lingual clustering algorithm uses the comparability

[9] `www.latimes.com`
[10] `www.heraldscotland.com`
[11] `www.lemonde.fr`
[12] SDA French Swiss news agency data – 1994-1995

measure, which they previously proposed in [Li and Gaussier, 2010], as a similarity measure.

The authors used hierarchical clustering algorithm to cluster French-English documents that have related contents into groups. The hierarchical clustering algorithm builds a hierarchy of clusters (a tree of clusters composed of sub-clusters). The authors set a threshold (node depth from the root) to select hight quality sub-clusters to form hight comparable clusters. The remaining clusters can be enriched using external sources as described previously in [Li and Gaussier, 2010]. The authors used cross-lingual clustering method to extract a highly comparable corpus from the same comparable corpus (Wikipedia), which is described in [Li and Gaussier, 2010]. The authors investigated different methods to extract bilingual lexicon from the highly comparable corpus, and they showed that the extracted lexicons have better precision than the previous method proposed in [Li and Gaussier, 2010].

In [Otero and López, 2011], the authors also proposed a comparability measure for cross-lingual documents. They reported that the measure can be useful to extract comparable corpora or to extract bilingual lexicon. The authors considered internal links of Wikipedia articles as a vocabulary. Internal links in Wikipedia articles are links to other Wikipedia articles. To explain how an internal link can be used as a vocabulary term, consider the English Wikipedia article "Napoléon". We find the term "French Revolution" is mentioned in this article because it is a related topic. The term "French Revolution" has a link to another Wikipedia article (English), which describes the "French Revolution". Therefore, "French Revolution" is considered by the authors as a vocabulary term in the English article "Napoléon", because it links to another internal article in Wikipedia. Similarly, the french article has the term "Révolution", which links to another Wikipedia article (French).

The authors consider these terms, which have internal links, as important terms in the documents. Thus, the source and the target documents are composed of internal links terms. These terms are translated by Wikipedia inter-language links. In

[Otero and López, 2011], the authors considered that two documents are comparable if they have most of these common (translation of each others) internal links. The authors applied the proposed measure on Portuguese, Spanish, and English aligned documents from Wikipedia. They repeated the experiment on non-aligned documents, and they showed that the measure capture the differences (degree of comparability) between the aligned and non-aligned corpora.

To summarize the reviewed works in this section, dictionary based methods for cross-lingual similarity are proposed at corpus level [Li and Gaussier, 2010] and at document level [Otero and López, 2011]. The main purpose is to extract high comparable corpora. [Li and Gaussier, 2010] searched for translations of source words in the target corpus and vice-versa (at corpus level), while [Otero and López, 2011] considered words that have internal links as important words, and they matched these source and target words via bilingual dictionary.

The drawbacks of the dictionary based approach are the dependency on the bilingual dictionaries, which are not always available, and the necessity to use morphological analyzers for inflected languages. Moreover, word-to-word dictionary translations without considering the context can lead to many errors because of the multiplicity of senses (translation ambiguity), and because the text is considered only as a bag of independent words.

In our work, we investigate a cross-lingual document similarity measured based on WordNet bilingual dictionary for English-Arabic documents. The similarity is measure in our work at document level. Further, two weighting schemes (binary and ($tfidf$) are investigated.

2.2.2 CL-IR Methods

Information Retrieval (IR) consists in finding documents in a large collection that are relevant to a given query [Manning and Raghavan, 2008]. Finding comparable

documents is a very similar task to Information Retrieval (IR). To find comparable documents, the query is the whole document instead of keywords, and the task then is to find the most similar (relevant) document(s) to a given one.

The IR system usually computes a numeric score that quantifies how each document in the collection matches the query, and ranks the documents according to this value [Manning and Raghavan, 2008]. A document is ranked by estimating the cosine similarity with the query. The top ranked documents are then shown to the user. The IR system normally indexes text documents by representing them as vectors in the VSM.

Cross-Lingual Information Retrieval (CL-IR) is a special case of IR, where the language of the query is different from the language of the documents. In this case, the CL-IR system should unify the language of queries and documents. Therefore, the system should help users to understand the results regardless of the source language.

In (CL-IR) methods, either queries or documents can be translated using a Machine Translation (MT) system. Then, classical IR tools, which are based on VSM, can be used to identify comparable articles. The drawback of this approach is the dependency on the MT system, which affects the performance of the IR system. Moreover, the MT system needs to be developed first if it is not available for the desired language.

Researchers usually translate queries into the language of the indexed documents instead of translating the whole document collection. This is because the computational cost of translating queries is far less than the cost of translating all indexed documents. Thus, the IR system indexes the target documents and the source documents are translated by the MT system. For example, in [Aljlayl et al., 2002], the authors translated Arabic queries using a machine translation system to retrieve English documents. The authors experimented several query lengths, and they found that the shorter the query is, the better the precision and recall are. The authors

reported that the limitation of this approach is the quality of the machine translation system.

[Larkey et al., 2002] investigated the impact of several Arabic stemming techniques (stemming and light stemming) on monolingual (Arabic) IR and English-Arabic CL-IR. The authors used bilingual dictionary for translating English-Arabic queries. They found that Arabic light stemming technique leads to best precision and recall for monolingual and cross-lingual retrieval. The authors pointed out that the results in monolingual IR is better than the results in CL-IR. This is maybe due to ambiguity that occurs when translating words using bilingual dictionary. Arabic stemming and light stemming are described in details in Chapter 3.

In [Farag et al., 2011], the authors addressed the problem of English-Arabic CL-IR. Their system uses machine translation to translate user queries. The authors proposed an interface that allows the user to interactively choose between possible translations for his query. Each possible translation is described with contextual information in the user's language. This description is extracted from a parallel corpus. The authors conducted a questionnaire on a set of users to evaluate this method. The outcome of this study is that this method provides clearer and easier approach than the non-interactive interface for CL-IR.

On the other hand, [Fujii and Ishikawa, 2000] considered that MT has prohibitive computational cost. Therefore, they proposed a method, which degrades this cost by avoiding the translation of all documents. This method only translates queries and their top N relevant documents. First, source queries are translated into the target language. Then for each query, top N relevant target documents are translated back into the source language. Finally, the translated target documents are re-indexed and re-ranked using the IR system of their original source queries.

[Ture, 2013] reported that translating only queries or only documents is not helpful because this makes a user to deal with contents in a language that he cannot understand. The objective of this work is to improve integration between search (IR) and

translation (MT) techniques to produce better quality output for the user.

The author proposed two solutions to improve the integration between IR and MT: the first one is "search to translate" and the second one is "translate to search". The first solution searches to extract bilingual data from Wikipedia to train better translation models. The second solution integrates the MT into cross-lingual search process. The integration improves the performance of the MT system for query translation such that the translated queries improve the IR process. In other words, the author used IR to improve MT, then used MT to improve IR.

The authors in [Ture et al., 2012] combined three statistical machine translation techniques for CL-IR. They investigated three statistical machine translation techniques to translate words: context-independent translation, phrase-dependent contexts, and sentence-dependent contexts. The authors retrieved Arabic, Chinese and French documents using English queries. The authors implemented a linear combination of the three translation techniques. The authors evaluated the method on an English-Arabic dataset from TREC2002[13], an English-Chinese dataset from NTCIR-8[14], and an English-French dataset from CLEF2006[15]. They compared the combined techniques with the standalone technique. The authors found that the optimal combination depends on the corpus.

To summarize the reviewed works in this section, researchers use translation approach (machine translation or bilingual dictionary) combined with classical IR system to address the problem of CL-IR. Some of them reported that the approach is limited with the quality of the translation [Aljlayl et al., 2002], while other addressed the best morphological analysis techniques to improve the overall performance [Larkey et al., 2002]. Another researcher addressed the computational cost of translating the whole document collection [Fujii and Ishikawa, 2000]. On the other

[13] `http://trec.nist.gov/data/qa/t9_qadata.html`
[14] `http://research.nii.ac.jp/ntcir/ntcir-ws8/ws-en.html`
[15] `http://www.clef-initiative.eu/edition/clef2006`

hand, [Ture et al., 2012, Ture, 2013] addressed the problem of improving the performance of MT to have better retrieval results. [Ture et al., 2012] investigated different machine translation techniques for better for CL-IR, while [Ture, 2013] proposed to integrate IR and MT by tuning the MT system to improve the IR results.

In our work, we investigate different approaches for CL-IR. We investigate bilingual dictionary approach and machine translation approach for cross-lingual documents retrieval, but we investigate machine translation approach with Latent Semantic Indexing (LSI) method. We further compare the results of these approaches.

2.2.3 CL-LSI Methods

As introduced earlier, document similarity can be estimated on term level or on semantically related terms. Generally, semantic similarity is a metric that quantifies the likeness (similarity) of documents or terms based on the meaning or semantics of the contents. Semantic similarity can be measured based on a pre-defined ontology, which specifies the distance between concepts, or can be measured using statistical methods, which correlate terms and contexts from a text corpus. One of the methods in the latter category is Latent Semantic Indexing (LSI).

LSI is also referred in the literature as Latent Semantic Analysis (LSA). LSI is designed to solve the problems of the ordinary IR system, which depends on lexical matching [Dumais, 2007]. For instance, some irrelevant information can be retrieved because some literal words have different meanings. On the contrary, some relevant information can be missed because there are different way to describe the object. The VSM in ordinary IR system is sparse (most of elements are zero), while LSI space is dense. In the sparse VSM, terms and documents are loosely coupled, while terms and documents in LSI space are coupled (correlated) with certain weights. VSM is usually high dimensional space for large document collection, while LSI is a reduced space. The number of dimension is LSI is lower than the number of unique terms in

the document collection. The cosine similarity can be noisy or inaccurate in sparse and high dimensional space.

Besides using LSI for IR, it has been used also for tasks rather than IR, such as document clustering [Wei et al., 2008], text summarization [Yeh et al., 2005], citation and link analysis [Jin et al., 2004].

In LSI, the term-document matrix (m terms $\times$ n documents) is decomposed by Singular Value Decomposition (SVD) into three matrices (see Equation (2.7)) [Deerwester et al., 1990]:

1. The term matrix (U), which is an $m \times k$ matrix, where k is the reduced dimension.
2. The diagonal matrix (S), which is an $k \times k$ matrix of the singular values.
3. The document matrix (V^T), which is an $k \times n$ matrix.

$$\overset{\text{terms}}{\left(\begin{array}{c} \overset{\text{docs}}{\text{X}} \end{array}\right)_{m \times n}} = \left(\text{U}\right)_{m \times k} \left(\text{S}\right)_{k \times k} \left(V^T\right)_{k \times n} \tag{2.7}$$

U and V^T are the left and right singular vectors respectively, while S is a diagonal matrix of singular values. Each column vector in U maps terms in the corpus into a single concept of semantically related terms that are grouped with similar values in U.

k is the reduced concept space in LSI. [Landauer et al., 1998, Dumais, 2007] reported that the optimal value of k to perform SVM is between 100 and 500. That depends

on the task and the nature of data. Thus, one can determine the optimal value of k between 100 and 500 experimentally.

Considering the term-document matrix presented in Table 2.2, the result of applying SVD ($X = USV^T$) for this matrix is shown in Tables 2.4 and 2.5. Table 2.4 represents the document matrix (V^T) in the LSI space, while Table 2.5 represents the term matrix (U) in the LSI space.

Table 2.4 shows the document matrix (V^T) in the LSI space. In this example, we choose to project the term-document matrix (19 dimension) into 3 dimension ($k = 3$) in the LSI space. The table shows the weights of the concepts (C_i) for each document in the LSI space. We use the LSI implementation in Gensim package [Rehurek and Sojka, 2010] to generate this example.

TABLE 2.4: Term-document matrix in LSI space

Concepts	d_1	d_2	d_3	d_4	d_5
C_1	1.74	1.91	2.92	1.03	0.46
C_2	-0.34	-2.12	1.05	1.48	0.10
C_3	0.84	-0.72	0.60	-1.16	-1.39

Table 2.5 shows the term matrix (U) in the LSI space. Each term has a specific weight in the concept (C_i), which specify how much the term contributes to the concept. For instance, for C_2, the most important terms are 'language', 'computer', etc, while the terms 'NLP' and 'useful' are unimportant.

TABLE 2.5: Concepts of the LSI space

Concepts	Terms
C_1	(language, 0.574), (computer 0.345), (field, 0.239), (NLP, 0.336), (processing, 0.320), (human, 0.282), (science, 0.239), (method, 0.177), (communication, 0.177), (natural, 0.115)
C_2	(field, 0.319), (science, 0.319), (computer, 0.275), (information, 0.200), (technology, 0.187), (related, 0.187), (stands, -0.267), (natural, -0.267), (useful, -0.267), (NLP, -0.578),
C_3	(human, 0.295), (language, 0.271), (make, 0.172), (understand, 0.172), (produces, -0.284), (data, -0.284), (related, -0.237), (technology, -0.237) (processing, -0.309) (information, -0.521)

LSA assumes that words and documents form a joint Gaussian model. Documents and words in LSA are projected to a reduced space using SVD. Similar words and document are mapped closer to each others in the LSI pace. LSA can capture synonyms but it can not address polysemy. Synonyms are the words or phrases that have exact or near exact meaning. For instance, the following term couples are synonyms: (happy, content), (shut, close), (beautiful, attractive), (intelligent, smart). In the contrary, a polyseme is a word or a phrase that has many possible meanings. For instance, following terms have multiple meaning: (wood: a piece of a tree; a geographical area with many trees), (head: part of the body; a person in charge of an organization).

To address the limitation of LSA, the authors in [Blei et al., 2003] extended the LSA to a probabilistic version called Latent Dirichlet Allocation (LDA). In LDA, documents are represented as random mixtures of latent topics, these topics are probabilistic distributions over words. In other words, every topic is a distribution over different words, and every document is a distribution over different topics. Therefore, LDA is a generative model and it specifies a per-document topic distribution. LDA is useful for topic modeling, on the other hand, LSA is useful to map similar documents and words into a reduced feature space (model concepts).

Some researchers compared LSA and LDA for different tasks. For instance, in [Biro et al., 2008], the authors conducted a comparative study of using LSA and LDA for text classification. The authors applied the two methods on a corpus from the open directory project[16]. The corpus is composed of 350K English documents categorized into 8 categories. The corpus is split into 80% for training and 20% for test. The training data is used to build the LSA and the LDA models. The authors applied Support Vector Machines and C4.5 decision tree classification algorithms. Their results indicate that, in terms of F-Measure, SVM outperforms C4.5 algorithm, and LDA outperforms LSA model.

[16] `www.dmoz.org/`

In [Cui et al., 2011], the authors addressed the following question. How similar are the LSA's concepts to LDA's topics? Are the most important LSA's concepts and LDA's topics mostly similar? Are clusters produced by the two methods the same? The authors stated that, in order to determine which method may be most suitable for a given analysis task, a direct comparison and interpretation of results should be carried out. The comparison, understanding and interpreting results is a challenging job. This is why the authors proposed a human consumable view (called *TopicView*) instead of statistical comparison to compare the differences between LSA and LDA.

TopicView is an application designed to visually compare and to explore the differences between LSA and LDA. In *TopicView*, the user can view concepts/topics and documents relationships. The authors presented some case studies on synthetic and real-world corpora.

The synthetic corpus (called *alphabet*) is composed of 26 clusters containing 10 documents each. Each cluster consists of documents made up exclusively of terms starting with the same letter. The real-world corpus (called *DUC*) composed of 298 news documents categorized into 30 clusters. This corpus is collected from the Associated Press and New York Times in 2003[17].

The authors demonstrated using *TopicView* that LSA clusters *alphabet* corpus very well, while LDA is unable to partition the data. The authors concluded that LSA clustered the documents correctly. The authors used the tool to demonstrate that the relationships between LSA's concepts and LDA's topics are weak for *alphabet* corpus. The authors pointed out that these outcomes are expected on *alphabet* synthetic dataset.

As for *DUC* corpus, the tool showed that LSA and LDA are similar in defining clusters, but they are different how connections between documents in the clusters are modeled. The connections between documents in the same cluster by LSA model

[17] `http://duc.nist.gov`

are stronger than the ones by the LDA model. The authors demonstrated that both LSA and LDA provide different but useful characteristics and view of the data.

The authors used *TopicView* to infer the following general conclusions. LSA concepts provide good summarizations over broad groups of documents, while LDA topics are focused on smaller groups. LDA's topics are good for labeling clusters using most probable words than LSA's concepts, while LSA model does not include extraneous connections between disparate topics identified by LDA.

In our work, we use LSA method for cross-lingual similarity measure because our aim is to map similar documents and words across languages closer to each others into a reduced feature space.

In Cross-Lingual Latent Semantic Indexing (CL-LSI) methods, documents are represented as vectors like in CL-IR method, but these vectors are further transformed into another reduced vector space like in LSI. Then, one can compute the cosine between vectors in this new space to measure the similarity between them. LSI method has already been used for CL-IR in [Littman et al., 1998]. In this approach, the source and the target documents are concatenated into one document and then LSI learns the links between source and target words. The CL-LSI method is described in detail in Section 4.3 in Chapter 4.

The advantage of CL-LSI is that it does not need morphological analyzers or MT systems. Moreover, it overcomes the problem of vocabulary mismatch between queries and documents. [Dumais, 2007] compared between the performance of CL-IR and CL-LSI, and the authors showed that LSI outperforms the vector space model for IR.

Many works have been done on retrieval of document pairs written in various languages using CL-LSI. For example, [Berry and Young, 1995] worked on

Greek-English documents and [Littman et al., 1998] worked on French-English documents, Spanish-English [Evans et al., 1998, Carbonell et al., 1997], Portuguese-English [Orengo and Huyck, 2003], Japanese-English [Jiang and Littman, 2000, Mori et al., 2001], while [Muhic et al., 2012] worked on several European languages.

In [Littman et al., 1998], the authors compared the performance of retrieving documents based on machine translation and based on CL-LSI. They applied their work on Hansard parallel corpus[18], which is the proceedings of the Canadian parliament. The corpus is composed of 2.4K French-English paragraphs. They divided this corpus into training and test pairs. Each paragraph pair in the training set is concatenated into one document, while test paragraphs are kept separated. The training set is used to create the CL-LSI space. French and English test paragraphs are mapped into the CL-LSI space separately. Each English test paragraph is used as a query to retrieve the exactly one relevant French paragraph (its translation).

The authors repeated the same experiment, but using French paragraphs as queries and English paragraphs as targets. The CL-LSI was able to find about 98.4% of pairs for the given queries. Moreover, the authors investigated if French-English word overlap has an impact on the results. To do this, they repeated the experiment, but they added the prefix "F" to French words and "E" to English words. As a result, French and English texts have no word overlaps. The result of this experiment is comparable to the first one. The CL-LSI was able to find 98.9% pairs of the queries.

In addition, the authors investigated if machine translation is sufficient for cross-lingual retrieval. They replicated the same experiment using the machine translation system. In this experiment, a monolingual LSI space is created using the English training documents. Then the test documents are mapped into the LSI space, then the French test documents (queries) are translated into English using a machine translation system. The system was able to find 99.4% of query matches. Based

[18] `www.isi.edu/natural-language/download/hansard`

on experiments, the authors concluded that machine translation can be sufficient for cross-lingual retrieval.

Finally, the authors investigated retrieving documents out of the domain of Hansard corpus. In this experiment, the CL-LSI is trained on Hansard corpus, while the test consists of French-English parallel documents from yellow page's domain. The CL-LSI was able to find 22.8% of pairs of the queries. The performance increased to 63.8% when some training texts were added from yellow page's domain. These additional training data are generated using a machine translation system.

In [Muhic et al., 2012], the authors used CL-LSI for cross-lingual similarity measure, and applied it on Europarl parallel corpus[19] (European Parliament Proceedings) and aligned documents from Wikipedia in several European languages. All aligned sentences (seven languages) of the Europarl parallel corpus are merged into one sentence. Similarly, all aligned documents (seven languages) of the Wikipedia corpus are merged into one document. The resulting term-document matrix D indexes terms of all languages. The authors transformed D into LSI space using the same approach as [Littman et al., 1998]. The authors reported that pair document retrieval is not symmetric in all languages for Europarl and Wikipedia corpora, and it depends on the language pairs. Since the result were not symmetric for all languages even for parallel corpus, then we can assume that the performance of CL-LSI can be language dependent.

In contrast, [Fortuna and Shawe-Taylor, 2005] considered the use of the LSI method as a weakness because it requires an aligned parallel corpus. The authors compared cross-lingual IR and text classification models, which are built using two sets: human generated corpus (set A) and MT generated corpus (set B). Their method has been used to retrieve/classify documents from the Hansard corpus (Canadian Parliament records in French and English languages). The authors built one CL-LSI model from the set A and another one from the set B. The authors showed that when the training

[19] `www.statmt.org/europarl`

and the test sets are from the same domain, the CL-LSI models of set A and B have similar performance. But when the domains of training and test sets are different, then the CL-LSI model of set B has better performance.

The authors reported that the main advantage of using the MT system is that it allows to generate a training set relevant to the target domain of the application. Despite of this advantage, a question arises about the quality of MT in the domain of the application, and whether it is sufficient to produce an accurate cross-lingual LSI space. They addressed this question, but ignored the fact that the MT system itself needs to be developed first, and requires a human generated parallel corpus. This contradicts with the first claim, which is the problem of CL-LSI is that it requires a parallel corpus. In addition, we show in this thesis that CL-LSI works well with comparable corpora.

The authors in [Cimiano et al., 2009] conducted an empirical comparison between Latent Semantic Indexing (LSI) with Explicit Semantic Indexing (ESI) methods for CL-IR. The former is based on the Singular Values Decomposition (SVD) of the terms-by-documents matrix that is used to find the latent semantics, and the latter indexes the text using given external concepts or knowledge-base. The concepts are explicit in ESI, while the concepts are latent in LSA. In other words, the ESI's concepts are defined explicitly, while LSI's concepts are extracted implicitly (latent) from the corpus. In their work, the external knowledge-base is a corpus of Wikipedia articles, where each article's title is considered as a concept.

The authors applied ESI and LSI methods on two parallel test corpora: the first corpus is collected from the "Journal of European Community", and the second one is collected from legislative documentations of the European Union. The authors built two ESI spaces using the training documents of the parallel corpora.

The authors conducted experiments on English, French and German languages. The ESI and LSI models are trained on the training parts of parallel corpora. The authors reported that ESI outperforms LSI for CL-IR task. The authors repeated the same

experiment, ESI and LSI models are trained on aligned documents from Wikipedia corpus. The result in this experiment is that both ESI and LSI have the same performance. The authors claimed that ESI is preferable because the availability of aligned Wikipedia corpus is a restriction, but this is not true because Wikipedia articles can be aligned using inter-language links as we achieve in this thesis.

To summarize the works reviewed in this section, CL-LSI can achieve good results and it has better performance than CL-IR [Dumais, 2007, Littman et al., 1998] for cross-lingual document retrieval. Machine translation can be sufficient for cross-lingual document retrieval using LSI [Littman et al., 1998, Fortuna and Shawe-Taylor, 2005], but the benefit of CL-LSI is that it does not need machine translation and it has a competitive performance. The CL-LSI method is language independent, but the performance is language dependent, i.e., it depends on the language pair [Muhic et al., 2012]. To achieve better results when using CL-LSI, it is recommended that the training and test documents are from close domains [Littman et al., 1998, Fortuna and Shawe-Taylor, 2005]. ESI and LSI have the same performance when a corpus of aligned document from Wikipedia is available [Cimiano et al., 2009]. To our knowledge, English-Arabic document retrieval using CL-LSI is not addressed in the literature. In this thesis we investigate cross-lingual document retrieval using CL-LSI but for Arabic-English language pair. We also investigate the performance of retrieving documents using machine translation approach, and compare it to CL-LSI. In our work we investigate training CL-LSI using two types of corpora (parallel and comparable).

2.3 Sentiment Analysis

Sentiment analysis (or opinion mining) consists in identifying the subjectivity or the polarity of a text [Pang and Lee, 2008]. Subjectivity analysis includes the classification of a given text into subjective (*e.g., I will buy this amazing book!*) or objective

(*e.g., The new edition of the book is released*) labels, while polarity analysis aims to classify the text into positive (*e.g., The image quality is good!*) or negative (*e.g., The battery life is very short!*).

Subjective (or opinionated) text conveys opinions, while an objective text represents facts. Regarding polarity, positive and negative texts are opinionated. These opinions can be like/dislike, satisfied/unsatisfied, happy/sad, etc. Opinions can be related to a product, a book, a movie, news story, etc.

Some texts may contain mixed sentiments (*e.g., "I like this camera, but it is very expensive!"* or *"The phone will be available for order starting from the next week, but the price is prohibitively expensive!"*), and even humans may disagree among themselves about sentiment labels of these texts. This is why automatically analyzing subjectivity of the texts is a challenging task.

Popular methods for sentiment analysis are lexicon based or corpus based [Pang and Lee, 2008]. The lexicon based method uses a pre-annotated lexicon, which is usually composed of terms and the corresponding scores that represent the subjectivity or polarity of terms. The lexicon based method uses string matching techniques between texts and the annotated lexicon, and calculates the average score of the matched words.

The most common publicly available sentiment lexicons are WordNet-Affect [Valitutti, 2004] and SentiWordNet [Baccianella et al., 2010], which are the extensions of WordNet [Miller and Fellbaum, 1998]. Additionally, SenticNet [Cambria et al., 2010] is a knowledge-based extension of the aforementioned lexicons. All these lexicons are built using a semi-automatic method, which starts with an initial set of WordNet synonyms (synsets). First, this set is manually annotated, and then it is expanded iteratively based on synset relationships using a semi-supervised method.

SentiWordNet is evaluated by [Baccianella et al., 2010] on a manually annotated dataset, and authors report that this lexicon can be reliable for sentiment analysis. However, authors of [Chaumartin, 2007] report some incorrect entries in the lexicon. SenticNet is also evaluated by [Cambria et al., 2012] on a manually annotated dataset, and the authors conclude that SenticNet can be used for sentiment analysis.

The corpus based approach is also popular for sentiment analysis. It uses a pre-annotated corpus and machine learning algorithms to train classifiers to automatically classify a given text. The task is considered as a text classification problem. The most commonly used features are n-grams and POS tags, and the most commonly used classifiers are Support Vector Machines (SVM) and Naive Bayes (NB) [Pang and Lee, 2008]. One of the publicly available resources is a collection of movie reviews, which are pre-annotated with subjectivity labels [Pang and Lee, 2004], and polarity labels [Pang and Lee, 2005].

The work of [Pang et al., 2002] assessed three machine learning algorithms (Maximum Entropy, NB, and SVM) to classify movie reviews that are collected in [Pang and Lee, 2005] into positive and negative labels. The authors reported that the classification into positive or negative classes does not perform as well as classification into transitional categories such as sport, health, news, etc. Consequently, the authors have shown that sentiment classification is more challenging task. Each review of the movie corpus has a positive or negative label. The authors extracted different features from each review of this corpus. So the training instance is the extracted features associated with class labels. The authors assessed combining different features for this task. These combinations are: "uni-grams", "bi-grams", "uni-grams + bi-grams" and "uni-grams + POS tags". The authors reported that SVM algorithm outperforms the other algorithms, and the best feature combinations are "bi-grams" and "uni-grams + bi-grams".

[Pang and Lee, 2004] proposed to use a 2-stage classification. The first stage classifies texts into subjective and objective, and filters out the objective portion, while keeping the subjective portion for the second stage of classification, which classifies the subjective texts into positive and negative. The idea is to prevent the polarity classifier from considering irrelevant texts. The authors showed that discarding objective texts improves the accuracy of polarity classification (from 82% to 86%).

Some researchers combine lexicon and corpus based methods to enrich the extracted features to improve the accuracy of sentiment classification [Dang et al., 2010, Hamouda and Rohaim, 2011]. The idea is to extract sentiment and some linguistic features, such as n-grams and POS tags, from the corpus and to add the sentiment score from the lexicon as a feature value. Then classifiers are trained on such combined features to improve the accuracy of the classifiers.

All the resources in the reviewed works above are in English language. In the following review, we revise the state-of-the-art of sentiment resources in Arabic language. Then, we review how researchers tackle the problem of generating resources for under-resourced languages using the English resources.

Sentiment resources in Arabic language are very limited. [Rushdi-Saleh et al., 2011] assessed sentiment resources in Arabic language, and they concluded that resources are scarce. Due to this fact, the authors collected a small corpus of movie reviews in Arabic, which is composed of 500 (250 positive and 250 negative) reviews. Despite the fact that the corpus is small, the authors investigated some machine learning techniques for sentiment classification on this corpus. Namely, they applied NB and SVM algorithms on uni-gram, bi-gram and tri-gram features extracted from the corpus. Their conclusion is that bi-gram and tri-gram features are better than uni-gram features, and SVM classifier outperforms NB.

[Abdul-Mageed et al., 2011] developed a manually annotated corpus in Arabic language. Two human annotators labeled newspaper documents with objective, subjective-positive, subjective-negative labels. In addition, the authors manually

created a polarity lexicon. In their work, the authors investigated 2-stage sentiment classification task on this corpus, where they first classify the text into subjective and objective, then the subjective text is classified into positive and negative. Unfortunately, these resources are not available publicly for other researchers.

[Abdul-Mageed and Diab, 2012] presented a multi-genre sentiment corpus in Arabic language. The corpus is annotated by two methods; by trained annotators and by crowd-sourcing annotators. The authors provided annotation guidelines with examples to the trained and crowd-sourcing annotators. The objective is to investigate the impact of the annotation guidelines on the two types of annotators. The authors reported the difficulties that the two types of annotators experienced in the labeling process. Their conclusion is that sentiment labeling is fuzzy and the annotators should be well-trained to have a reliable annotation. The authors precised the annotation guidelines, which incorporate reliable labeling. Unfortunately, even this corpus is not available publicly for other researchers.

The authors in [Aly and Atiya, 2013] collected book reviews written in Arabic from Goodreaders[20] social network. The corpus consist of about 63K book reviews. Each review have a rating of 1 to 5 stars. The authors considered the reviews with rating 4 or 5 stars as positive and the reviews with 1 or 2 stars as negative. The reviews with 3 stars are considered as neutral. The authors reported that the majority of reviews have positive label. To allow other researchers to compare their results with this corpus, the authors provided a standard split of the corpus into training and testing subsets. This corpus and the standard splits are publicly available. The authors used this corpus for polarity classification and rate prediction. The authors investigated SVM and NB classifiers on 1-gram, 2-gram and 3-gram text features. The authors reported that SVM classifier outperforms NB and the 3-gram are the best features.

[Abdulla et al., 2014] conducted an extended study on sentiment analysis in Arabic texts. The authors collected sentiment corpus from Yahoo-maktoob forum[21]. The

[20] `www.goodreads.com`
[21] `http://maktoob.yahoo.com`

collected corpus is composed of 5K reviews related to four domains: art, politics, science & technology and social. The corpus is annotated manually with polarity labels. The authors applied NB and SVM algorithms on this corpus, and they reported that SVM outperforms NB algorithm. In addition, the authors reviewed other studies related to sentiment analysis in Arabic texts, and they reported that most of such studies use in-house collected data from social networks and annotated with polarity labels. Unfortunately, the corpus collected by the authors is not available publicly at this moment.

Since most of the sentiment resources (lexicons and corpora) are available in English language, one should build the resources for other languages from scratch, or adapt English resources using machine translation systems [Denecke, 2008, Ghorbel, 2012]. However, [Brooke et al., 2009, Ghorbel, 2012] reported that creating new resources from scratch is better than using a MT system to generate them.

To our knowledge, comparing sentiments across languages is only addressed by [Bautin et al., 2008], while the rest reviewed works just adopt the English resources into other languages using machine translation, and debate whether machine translation is sufficient to capture sentiments or not.

In [Bautin et al., 2008], the authors used machine translation systems to translate texts of eight languages into English, then they applied sentiment analysis on the translated texts. The authors investigated whether machine translation is sufficient to capture sentiments in the translated texts. The authors used a sentiment analysis system called Lydia [Lloyd, 2006] to analyze sentiments based on named entities. For a given named entity (a person, a city, etc.), the system computes the sentiment score. The score quantifies the polarity of stories related to the named entity in news collected in a time period. The score calculation in Lydia system is done using a pre-defined sentiment lexicon.

The authors of [Bautin et al., 2008] conducted three experiments. In the first experiment authors collected news in eight languages (Arabic, Chinese, English, French,

German, Italian, Japanese, Korean, and Spanish) and translated non-English texts into English using Machine Translation (MT) systems. Then polarity scores are computed for the common entities in the eight languages. These entities are analyzed in all languages in temporal manner. Namely, the sentiment scores are analyzed for each day of ten days (1-10 May 2007). The authors reported that sentiment scores of the common entities are significantly correlated in the eight languages. The authors also tried to interpret spikes and drops of some sentiment scores in ten days by explaining some "positive" and "negative" news happened in the world in that period.

In the second experiment, the authors analyzed sentiments in JRC-Acquis parallel corpus[22], which is the European Union law applicable in the union countries. Non-English texts of this corpus are also translated into English using MT systems. Named entities are analyzed in all languages. The authors reported that the scores of the named entities are correlated in the parallel texts.

In the third experiment, the authors investigated the impact of machine translation on the results of their experiment. They conducted this investigation on Spanish language because of the availability of two machine translation systems for Spanish language. This experiment is carried out on the news corpus for ten days period. The authors computed correlation of sentiment scores of the two translations of the Spanish text. Their conclusion is that the correlation of sentiment scores of each day separately vary in the ten days period, but the overall scores are more correlated. The authors believed that despite MT makes some serious errors in translations, it can be sufficient to capture sentiments.

The rest reviewed works below just adopt the English resources into other languages using machine translation, and debate whether machine translation is sufficient to capture sentiments or not.

The work of [Brooke et al., 2009] explored the adaption of English resources for sentiment analysis of Spanish texts. The authors examined two approaches: the first

[22] `http://langtech.jrc.it/JRC-Acquis.html`

one is to build and annotate resources in the Spanish language from scratch, the second one is to use a MT system to translate English sentiment resources into Spanish. The authors compared a sentiment analysis method that uses the translated resources and the resources that they built from scratch. The authors reported that translation has a disruptive effect on the performance of sentiment analysis. Moreover, it is time and effort consuming to translate the lexicon and the corpus. Therefore, the authors concluded that it is worthy to build resources from scratch. The authors also concluded that the best approach for long-term improvement is through creating language-specific resources.

For example, [Denecke, 2008] introduced a method for sentiment analysis for non-English text. The method translates non-English texts into English using machine translation systems, then SentiWordNet lexicon [Esuli and Sebastiani, 2006] is used for sentiment analysis in the translated documents. The method is applied on movie reviews in German language collected from Amazon. The pipeline of their method is as follows. First, the language of the given text is identified using a language classifier. Then the document is translated into English using a suitable machine translation system for that language. Finally, the polarity of the translated document is computed based on SentiWordNet lexicon.

In [Ghorbel, 2012], the authors collected French corpus, and they translated it into English using two means: with bilingual dictionaries (WordNet [Miller and Fellbaum, 1998] and EuroWordNet [Vossen, 1998]), and with a MT system. The objective for that is to be able to use the SentiWordNet sentiment lexicon, which is in English language. The authors collected 650 French posts from Infrarouge online forum, which is a Swiss TV program. The forum discusses political, social and economic issues. The collected documents (posts) are annotated manually with positive and negative labels. The authors developed a baseline system by training a SVM algorithm on features extracted from the collected corpus. These features are uni-grams and the POS tags (adjective, adverb, noun and verb). More precisely, the

feature values are binary, i.e., they indicate the presence or absence of words or POS tags in the document.

The authors compared this baseline system with two other systems. The first system uses SVM algorithm trained on English translations of the French corpus, which is translated by a MT system. The second system trains the SVM algorithm on English translations of the corpus, which is translated by the bilingual dictionary. In these two systems, the authors added polarity scores of each POS tag that appeared in the document. The polarity information is extracted from SentiWordNet. The authors argued whether translation preserve sentiment, and they investigated the answer in their experiment. The authors reported that neither WordNet translation nor machine translation system significantly improved the performance with respect to the baseline system. The authors supported this with the fact that the quality of the translations is insufficient.

In [Wan, 2009], the authors used the co-training algorithm for cross-lingual sentiment classification. The co-training algorithm is a bootstrapping method that exploits labeled data to increase the annotated data incrementally. The authors have two datasets: labeled reviews written in English (L_{en}), and unlabeled reviews written in Chinese U_{ch}. The MT system is used to translate the English labeled corpus into Chinese (L_{ch}), and the unlabeled Chinese corpus into English (U_{en}).

The authors used the co-training algorithm as follows. A SVM classifier is trained on L_{en}, and another SVM classifier is trained on L_{ch}. The English classifier (SVM_{en}) is used to label U_{en} corpus. The co-training algorithms selects annotated examples such that the class distribution is balanced and the annotation is confident. Similarly, the Chinese classifier (SVM_{ch}) is used to label the U_{ch}. The new English and Chinese labeled data is added to original labeled corpus (L_{en} and L_{ch}). This process can be repeated iteratively until enough amount of labeled data is generated. Then, in the classification phase, SVM_{ch} is applied on the Chinese test reviews to predict the polarity label. The Chinese test reviews are translated into English using a MT

system, then SVM_{ch} is applied to the translated reviews to predict the class label. Finally, the average of SVM_{en} and SVM_{ch} classifiers determines the class label of the Chinese test reviews. The authors experimented several iterations and reported that the best classification accuracy is achieved after 40 iterations.

In [Wei and Pal, 2010], the authors used the same idea as in [Wan, 2009], but they proposed to use Structural Correspondence Learning (SCL) for improving the adoption of cross-lingual data, that is, to remove the noise that is produced by the machine translation. Basically, the SCL semi-supervised learning method was proposed by [Ando and Zhang, 2005] for domain adoption of the unlabeled data. Given two datasets: a labeled data (A) of domain (X), and unlabeled data (B) of domain (Y). The SCL methods adopts a classifier trained on the dataset (A) to the domain (Y), which is the domain of the unlabeled data (B). The idea of SCL is to find the pivot features, which behave in a similar manner in both source and target domains.

[Wei and Pal, 2010] used this idea but they considered that the source domain is the source language, and the target domain is the target language. The authors kept only the pivot features and discarded the other features to train the classifier. The authors applied the method on the same dataset of [Wan, 2009], and they reported that their method outperforms the co-training algorithm that is proposed by [Wan, 2009].

[Balamurali et al., 2012] defined the cross-lingual sentiment analysis as using a classifier trained on a corpus written in the source language, to predict the label of texts written in the target language. The authors presented an alternative approach using WordNet synset links to avoid machine translation. The WordNet is a lexical database that groups English words that have the same sense into sets called synset. Each synset have an identifier. WordNet is expanded to other languages by adding words that represent the same synset in the other languages. The authors extracted the WordNet synset features from the text by replacing words by their synset identifiers. The classifier that is trained on these features is language independent because

these identifiers are common in all WordNets. Thus, the classifier which is trained on the source corpus can be used to classify a text in the target language.

The authors applied their method on Hindi and Marathi language pair, which are widely spoken Indian languages. The authors developed a baseline system based on word translations using bilingual dictionary. The authors reported that the classifier that is trained on WordNet synset features outperforms the one that is trained on word-features.

The authors in [Demirtas and Pechenizkiy, 2013] investigated whether machine translation improves cross-lingual polarity classification. The authors have three labeled corpora; movie reviews written in English, movie reviews written in Turkish, product reviews written in Turkish.

The authors conducted two experiments. In the first experiment they expanded the Turkish corpus by translating the English corpus using machine translation, then they investigated whether the new data improves classification accuracy. In the second experiment, the authors used co-training with machine translation to improve the classification accuracy.

The authors concluded in the first experiment that expanding the training data using texts generated by machine translation does not necessarily improve the classification accuracy. The authors explained that the reason for that is the people who wrote the comments are from different cultures and backgrounds, and not because the quality of the machine translation. As for the second experiment, the authors showed that the co-training algorithm improves the classification accuracy when the unlabeled data comes from the same domain, and the algorithm does not improve the accuracy when the unlabeled data comes from different domains.

To summarize the reviewed work in this section, sentiment analysis is a challenging task, popular methods are based on per-annotated corpora or lexicons. Researchers already investigated several combination of sentiment features (n-grams, POS tags,

predefined scores). Studies related to multilingual sentiment analysis mainly focus on producing resources for under-resourced languages. The main approach for that is to use machine translation. The researchers further argue whether the machine translation is sufficient to capture sentiments.

In this thesis, we address the problem of producing sentiment resources differently. We produce resources in the target language using cross-lingual projection.

We further compare sentiments across languages. Our work is different from [Bautin et al., 2008], who use machine translation to translate all news documents in non-English languages into English, then they calculate the average sentiment scores, then the correlation of sentiments across languages is computed. In our work, we do not use machine translation, and we compare the agreement of sentiment labels of aligned comparable news documents collected from different sources.

2.4 Emotion Identification

Emotion identification is the automatic detection of emotions that are expressed in a text. It is useful for many applications such as market analysis, text-to-speech synthesis, and human-computer interaction [Pang and Lee, 2008].

The basic six human emotions, which are reported in a psychological study by [Ekman, 1992], are widely adopted in emotion identification [Pang and Lee, 2008]. These emotions are *anger, disgust, fear, joy, sadness,* and *surprise.*

There are various studies that addressed emotions identification in texts. For example, the work of [Zhe and Boucouvalas, 2002] introduces a text-to-emotions engine that generates expressive images of user's emotions in chat conversations. The authors reported that the system can be useful for real time gaming and real time communication systems, where transmitting video is restricted due to low bandwidth of the connection.

TABLE 2.6: A sample of emotion words in WNA lexicon

Anger	Disgust	Fear	Joy	Sadness	Surprise
umbrage	distasteful	apprehensive	cheerful	desolate	marvelous
pique	foul	hardhearted	smile	dispirit	stunned
huffy	recalcitrance	affright	pride	pitiful	terrific
aggression	nauseate	horrify	friendly	weariness	surprise
pestering	disgustful	fear	amorous	glum	amazing
torture	nauseating	dreadful	compassion	despondent	wonderful
mad	distastefully	isolation	carefree	suffer	terrifically
angry	ugly	afraid	happy	unhappy	incredible

In [Alm et al., 2005], the authors used a supervised method to predict the emotions in kid's fairy tales. The authors pointed that the system can be used for expressive rendering of text-to-speech synthesis of a narrative system. They adopted the basic six human emotions of [Ekman, 1992]. The authors considered the problem as multi-class classification task. They annotated 207 kid stories using human annotators. Feature extracted from this corpus include uni-grams and absence or presence of emotion words, which are determined using the emotion lexicon created by [Strapparava and Mihalcea, 2007]. The authors trained NB classifier on these features to predict the emotions in texts. They pointed that the initial results are encouraging, but the techniques need to be tuned, and data need to be extended to tackle the problem.

[Strapparava and Mihalcea, 2007] created the WordNet-Affect (WNA) emotions lexicon by annotating a subset of English WordNet. Each entry (synonym set or synset) of the WNA is annotated with one of the six basic emotions. Table 2.6 shows some sample of words in WNA emotions lexicon.

WNA emotion lexicon was basically developed for *indirect emotion* identification shared task. Indirect emotion identification consists in identifying emotions that are expressed indirectly in the text. The task is challenging because it is uneasy to anticipate the emotional state if emotions written in the text. For example, the emotional state expressed in *"seeing a lion"* is ambiguous. The text can indirectly refer to *joy* or *fear* emotions. That depends on its context. The context *"I saw a big*

lion in the national zoo" can refer indirectly to *joy* emotion, but the context *"While I was exploring the African savanna, a lion came across to me"* can refer indirectly to *fear* emotion.

The result of the systems that participate in the shared task are reported in [Strapparava and Mihalcea, 2010]. The authors pointed out that the task is very difficult, where the best F-Measure score was 0.17.

The WNA can be used as emotion lexicon to detect emotions in texts, or it can be used to extract emotion features. Extracted features can be used to train machine learning algorithms to detect emotions [Alm et al., 2005, Aman and Szpakowicz, 2007].

In addition, WNA can be also used to develop emotion resources for non-English languages. For instance, [Bobicev et al., 2010] translated WordNet-Affect from English into Romanian and Russian languages using a bilingual dictionary. Also [Torii et al., 2011] developed a Japanese WNA from the English one by crossing synsets-IDs with the Japanese WordNet. The authors evaluated the lexicon for emotion identification on an in-house Japanese corpus. They compared the precision and recall of identifying emotions with and without applying morphological analysis on the text. They reported that applying morphological analysis improves the results by 4.1%.

To summarize the reviewed works in this section, basic six human emotions (*anger, disgust, fear, joy, sadness, surprise*) has been adopted by several researchers. Researchers focus mainly on developing system to identify emotions for different applications such as text-to-speech synthesis, and human-computer interaction. Researchers also focused on developing emotion resources for low-resourced languages. In this thesis, we translate WNA emotions lexicon from English into Arabic manually. To our knowledge, comparing emotions in comparable documents have not been addressed in the literature. In our work, we compare the agreement of emotion labels in comparable documents collected from different sources.

Chapter 3

Collected and Used Corpora

3.1 Introduction

This chapter presents the collected and used corpora. In our work, we use English-Arabic parallel corpora, and we collect English-Arabic-French comparable corpora.

We need the parallel corpus for several reasons: (a) to compare the application of the proposed methods on comparable as well as parallel corpora, (b) to study the degree of similarity of comparable texts as compared to the degree of similarity of parallel texts, and (c) to use parallel texts to transfer sentiment annotation from English to Arabic. The parallel corpora come from several different domains, and they are ideal to test our methods on different genres of texts.

English-French-Arabic comparable corpora are not available; Therefore, collecting such resources is one of the contributions in this thesis. In addition, the work of this dissertation needs such dataset to study comparable texts, and to develop and test our proposed methods for aligning, retrieving, and annotating comparable texts and compare them in terms of sentiments and emotions. Moreover, such resources can be useful for several applications such as cross-lingual text mining, bilingual lexicon extraction, cross-lingual information retrieval and machine translation.

We collect the comparable corpora from two sources: Euronews website[1], and Wikipedia encyclopedia[2]. Comparable and parallel corpora are described in detail in Sections 3.3 and 3.4. The collected Wikipedia corpus is made publicly available online[3] for research purposes.

It should be noted that we collected the comparable corpora in Arabic, English and French languages, but we focus on English-Arabic language pair for alignment, retrieval, annotation and comparison tasks.

Before describing the used and collected corpora, we briefly introduce some characteristics of the Arabic language in the next section.

3.2 Arabic Language

Arabic language is used by about 422M people in the Middle East, North Africa and the Horn of Africa [UNESCO, 2012]. Arabic words are derived from roots, which can be composed of three, four or five letters. Triliteral root is the most common one. About 80% of Arabic roots are triliteral [Khoja and Garside, 1999, Sawalha and Atwell, 2008]. Words can be derived from a root by adding prefixes, infixes or suffixes.

Arabic is a highly inflected language [Habash, 2010]. Table 3.1 presents some examples of inflected terms of the Arabic language. The table shows several different English words, that are related to the same root in Arabic. Therefore, for an English-Arabic NLP task, applying rooting on Arabic words may lead to lose the meaning of Arabic words against the corresponding English words.

[1] `www.euronews.com`
[2] `www.wikipedia.org`
[3] Corpus is publicly available at `http://sf.net/projects/crlcl`

TABLE 3.1: Examples of some inflected terms in Arabic language

Arabic word	Meaning	Description	Root
كاتب *kātb*	author	name of the subject	كتب *ktb*
يكتب *yktb*	he writes	the verb	كتب *ktb*
كتاب *ktāb*	book	the outcome of the verb	كتب *ktb*
مكتبة *mktbh*	library	where the verb takes place	كتب *ktb*
مكتب *mktb*	office	the place of the verb (to write)	كتب *ktb*
يطير *yṭyr*	he flies	the verb	طير *ṭyr*
طائر *ṭāʾyr*	bird	name of the subject	طير *ṭyr*
طيار *ṭyār*	pilot	name of the subject	طير *ṭyr*
طائرة *ṭāʾyrh*	airplane	name of the subject	طير *ṭyr*

Unlike English terms that are isolated, certain Arabic terms can be agglutinated (words or terms are combined) [Habash, 2010]. For instance, the Arabic item وسيعطيك *wsyʿṭyk* corresponds to "and he will give you" in English. In Arabic, usually the definite article الـ *āl* "the" and pronouns are connected to the words.

Arabic words have different forms depending on gender (masculine and feminine). For example, the English word "travelers" corresponds to مسافرون *msāfrwn* in masculine form, and مسافرات *msāfrāt* in feminine form. Word forms in Arabic may change according to its grammatical case. For instance, مسافرون *msāfrwn* is in nominative form, while its accusative/genitive form is مسافرين *msāfryn* .

Besides the singular and plural forms, Arabic words have the dual form. Singular form refers to one person or thing, dual form refer to two persons or things, and plural form refers to three or more persons or things.

Verb conjugation in Arabic is derived according to person, number, gender, and tense. Table 3.2 shows some examples of conjugating the verb كتب *ktb* . The table shows the conjugation for the first person (1) (I and we), second person (2) (you), and the third person (3) (he, she, it, one, and they). The table also shows the conjugation for the masculine (m), feminine (f), singular, dual and plural forms.

TABLE 3.2: Verb conjugation of the root كتب *ktb* (to write)

Person	Perfect	Meaning	Imperfect	Meaning
Singular 1	كَتَبْتُ *katabtu*	I wrote	أَكْتُبُ *ʾaaktubu*	I write
Singular 2m	كَتَبْتَ *katabta*	you wrote	تَكْتُبُ *taktubu*	you write
Singular 2f	كَتَبْتِ *katabti*	you wrote	تَكْتُبِينَ *taktubiyna*	you write
Singular 3m	كَتَبَ *kataba*	he wrote	يَكْتُبُ *yaktubu*	he write
Singular 3f	كَتَبَتْ *katabat*	she wrote	تَكْتُبُ *taktubu*	she write
Dual 2	كَتَبْتُمَا *katabtumaā*	you wrote	تَكْتُبَانِ *taktubaāni*	you write
Dual 3m	كَتَبَا *katabaā*	they wrote	يَكْتُبَانِ *yaktubaāni*	they write
Dual 3f	كَتَبَتَا *katabataā*	they wrote	تَكْتُبَانِ *taktubaāni*	they write
Plural 1	كَتَبْنَا *katabnaā*	we wrote	نَكْتُبُ *naktubu*	we write
Plural 2m	كَتَبْتُمْ *katabtum*	you wrote	تَكْتُبُونَ *taktubuwna*	you write
Plural 2f	كَتَبْتُنَّ *katabtuna*	you wrote	تَكْتُبْنَ *taktubna*	you write
Plural 3m	كَتَبُوا *katabuwā*	they wrote	يَكْتُبُونَ *yaktubuwna*	they write
Plural 3f	كَتَبْنَ *katabna*	they wrote	يَكْتُبْنَ *yaktubna*	they write

Common methods to analyze words in Arabic language is rooting [Khoja and Garside, 1999, Taghva et al., 2005] and light stemming [Larkey et al., 2007]. Rooting removes the word's prefix, suffix and infix, then converts it into the root form, while light stemming just removes the word's prefix and suffix. Table 3.3 shows some examples, where words are analyzed using rooting and light stemming methods.

Light stemming have better performance than rooting for several Arabic NLP tasks such as text classification [Saad, 2010], text clustering [Ghanem, 2014], information retrieval [Larkey et al., 2007], and measuring texts similarity [Froud et al., 2012].

TABLE 3.3: Methods of morphology analysis for some Arabic words

Word	Meaning	Prefix	Infix	Suffix	Light Stem
Words inflected from the root كتب *ktb* (to write)					
المكتبة *ālmktbh*	the library	الـ *āl*	مـ *m*	ة *h*	مكتب *mktb*
الكاتب *ālkātb*	the author	الـ *āl*	ا *ā*	-	كاتب *kātb*
الكتاب *ālktāb*	the book	الـ *āl*	ا *ā*	-	كتاب *ktāb*
يكتب *yktb*	he writes	ي *y*	-	-	كتب *ktb*
Words inflected from the root سفر *sfr* (to travel)					
المسافرون *ālmsāfrwn*	the travelers	الـ *ālm*	ا *ā*	ون *wn*	مسافر *msāfr*
المسافرين *ālmsāfryn*	the travelers	الـ *ālm*	ا *ā*	ين *yn*	مسافر *msāfr*
سيسافر *sysāfr*	he will travel	سي *sy*	ا *ā*	-	سافر *sāfr*
سافرت *sāfrt*	she traveled	-	ا *ā*	ت *t*	سافر *sāfr*

For an English-Arabic NLP task, applying rooting on Arabic words may lead to lose the meaning of Arabic words against the corresponding English words [Saad et al., 2013].

To recapitulate, Arabic language have very different characteristics from English language. Several consideration should be taken into account when doing Arabic or Arabic-English NLP tasks. This makes the task more challenging.

3.3 Comparable Corpora

This section describes the comparable corpora that we collect from two sources: Euronews website[4], and Wikipedia encyclopedia[5].

We align the collected texts at the document level. That means, for Euronews corpus, the aligned articles are related to the same news story, and for Wikipedia corpus the aligned articles are related to the same context. For example, the English Wikipedia

[4] `www.euronews.com`

[5] `www.wikipedia.org`

article "Olive oil" is aligned to the French article "Huile d'olive", and to the Arabic article "زيت زيتون". In the next two sections, we described our collected comparable corpora.

3.3.1 Wikipedia Comparable Corpus

Wikipedia is an open source encyclopedia written by contributors in several languages. Anyone can edit and write Wikipedia articles. Therefore, articles are usually written by different authors. Some of Wikipedia articles of some languages are translations of the corresponding English versions, and others are written independently. Wikipedia provide a free copy of all available contents of the Encyclopedia (articles, revisions, discussion of contributors). These copies are called dumps[6]. Because Wikipedia contents change with time, the dumps are provided regularly every month. Wikipedia dumps can be downloaded in XML format. Our Wikipedia corpus is extracted by parsing Wikipedia dumps of December 2011, which are composed of 4M English, 1M French, and 200K Arabic articles.

English Wikipedia started in 2001 with 2.7K articles, and French Wikipedia started in in the same year with 895 articles, while Arabic Wikipedia started in 2003 with 655 articles. Table 3.4 shows the rank of English, French and Arabic Wikipedias according to the number of articles [Wikimedia, 2014]. To judge these ranks, we need to compare the number of speakers and articles in each language. By August 2014, 335M English speakers added 4.7M articles, 456M French speakers added 1.5M articles, and 422M Arabic speakers added 315K articles. Thus, Arabic people added a very few articles compared to other language speakers. Despite that, the growth rate of Arabic articles is the highest compared to English and French [Wikimedia, 2014]. Figure 3.1 shows the growth rate of English, French, and Arabic Wikipedias from Jan. 2014 to Aug. 2014.

[6] `dumps.wikimedia.org`

TABLE 3.4: A list of English, French and Arabic Wikipedias ordered by number of articles (August 2014)

Rank	Language	Started in	Number of articles
1	English	2001	4.7M
5	French	2001	1.5M
22	Arabic	2003	315K

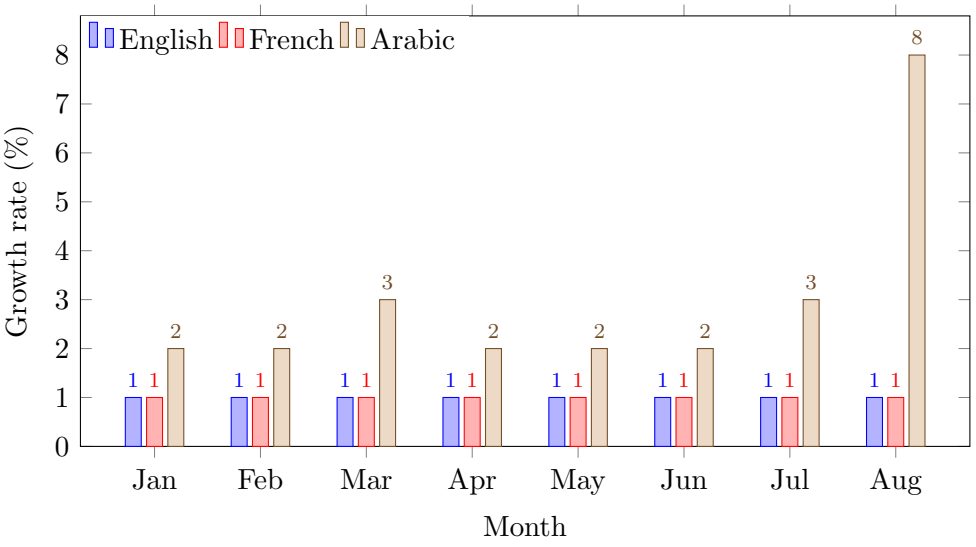

FIGURE 3.1: Growth rate of English, French, and Arabic Wikipeidas from Jan. 2014 to Aug. 2014 [Wikimedia, 2014]

Another indicator of the growth of a language in Wikipedia is the *article depth*, which is a rough indicator of the encyclopedia's collaborative quality, showing how frequently its articles are updated [Wikipedia, 2014]. Table 3.5 shows the depth indicator rank for English, French, and Arabic articles.

TABLE 3.5: Wikipedia article depth (August 2014)

Rank	Language
1	English
4	Arabic
5	French

Arabic, French, and English comparable articles are extracted based on inter-language links. In a given Wikipedia article written in a specific language, "inter-language links" refer to the corresponding articles in other languages. The form of these links

is [[*languagecode* : *Title*]]. For example, the inter-language links of the English language article "Rain" are shown in Figure 3.2.

FIGURE 3.2: The form of inter-language links of Wikipedia

```
[[ar:مطر]]
[[de:Regen]]
[[es:Lluvia]]
[[fr:Pluie]]
[[en:Rain]]
...
```

Using inter-language links for a given Wikipedia articles, We can select the titles of Wikipedia documents in other languages and extract them and link (align) them together. Thus, the extracted articles are aligned at article level. That means the three comparable articles are related to the same topic (context). We denote the extracted corpus as Arabic-French-English Wikipedia Corpus (AFEWC). The following steps describe our approach to extract and align comparable articles from Wikipedia dumps. These steps are applied for each English article in Wikipedia dump files.

1. If the English article contains Arabic and French inter-language links, then extract the French and Arabic titles.
2. Search by titles for the three comparable articles in the Wikipedia dump, and then extract them.
3. Extract the plain-text of the three comparable articles from wiki-markup.
4. Write comparable articles in plain-texts and xml files.

The extracted information includes article's title and wiki markup. From wiki markup, we extract the article's summary (abstract), categories, and the plain text. Examples of generic categories are sport, economics, religion, etc. Examples of specific categories are 'Nobel Peace Prize laureates', 'cooking oils', etc. All the aligned articles are structured in XML files. We also keep the wiki-markup for the aligned

articles because it can be useful to extract additional information later such as info boxes, image captions, etc.

Wikipedia December 2011 dumps contain about 4M English articles, 1M French articles, and 200K Arabic articles. In total, we extracted and aligned about 40K comparable articles. Corpus information is presented in Table 3.6, where $|D|$ is the number of articles, $|S|$ is the number of sentences, $|W|$ is the number of words, $|V|$ is the vocabulary size, $\bar{S}$ is the average number of sentences per article, and $\bar{W}$ is average words per article. A sample of Wikipedia comparable articles is presented in Appendix A.1. It can be noted from Table 3.6 that the number of sentences of Arabic articles is less than the number of sentences of English and French articles. Maybe this is because Arabic Wikipedia is still evolving as mentioned in the beginning of this section.

Table 3.6: Wikipedia comparable corpus (AFEWC) characteristics

	English	French	Arabic
$\|D\|$	40K	40K	40K
$\|S\|$	4.8M	4.7M	1.2M
$\|W\|$	91.3M	57.8M	22M
$\|V\|$	2.8M	1.9M	1.5M
$\bar{S}$	119	69	30
$\bar{W}$	2.2K	1.4K	548

Since our Wikipedia comparable corpus have the articles and their abstracts, then it can be useful for automatic text summarization applications.

3.3.2 Euronews Comparable Corpus

Euronews is a multilingual news TV channel, which aims to cover world news from a pan-European perspective[7]. News stories are also posted on the website. Euronews is available now in many European languages as well as in Arabic. English and French news services started in 1993, while Arabic started in 2008.

[7] `www.euronews.com`

Euronews corpus is extracted by parsing the html files of articles collected from Euronews website. Each English document has a hyperlink to the corresponding Arabic and French articles. We align comparable articles using these hyperlinks. Then, html tags are stripped for the three comparable articles, and stored in plain text files. Category information is also included in the plain text files. Euronews categories are: cinema, corporate, economy, Europe, hi-tech, interview, markets, science, and world.

Euronews corpus contains about 34K comparable articles as shown in Table 3.7. The average number of sentences is almost the same in English, French and Arabic documents.

A sample of Euronews comparable articles are presented in Appendix A.2. The sample documents are related to "level of corruption in world's country". The articles are mostly translations of each other. Comparing the English article with the Arabic and the French ones, we find that the Arabic article has an additional paragraph, which reports the situation of the corruption in some Arab countries, while the French article gives some additional detail about the situation of France.

Table 3.7: Euronews comparable corpus characteristics

	English	French	Arabic
$\|D\|$	34K	34K	34K
$\|S\|$	744K	746K	622K
$\|W\|$	6.8M	6.9M	5.5M
$\|V\|$	232K	256K	373K
$\bar{S}$	21	21	17
$\bar{W}$	198	200	161

This corpus is interesting because as we discussed above, articles are mostly translation of others. So the corpus is comparable and near-parallel at the same time. Is is interesting to discover the result of measuring document similarities and compare it with Wikipedia corpus. In addition, it is interesting to have a comparable corpus in news domain and compare its result with the encyclopedia domain.

3.4 Parallel Corpora

In this thesis we work on several corpus in order to measure the robustness of the proposed methods. In this section we describe the parallel corpora that we use in this dissertation.

We need the parallel corpora in our work because it can be baseline for the cross-lingual similarity measure, so we can compare it with our comparable corpora. The parallel corpora are useful for the cross-lingual annotation method, which is described in Chapter 5. The parallel corpora is used as baseline when we compare the agreement of sentiments in comparable documents, as we describe in Chapter 6

The parallel corpora come from several different domains, and they are ideal to test our methods on different genres of texts.

Table 3.8 shows the characteristics of the parallel corpora that we use in this work. $|S|$ is the number of sentences, $|W|$ is the number of words, and $|V|$ is the vocabulary size. The table also shows the domain of each corpus. The parallel corpora are: AFP[8], ANN[9], ASB[10] [Ma and Zakhary, 2009], Medar[11], NIST [NIST, 2010], UN [Rafalovitch and Dale, 2009], TED[12] [Cettolo et al., 2012], OST[13] [Tiedemann, 2012] and Tatoeba[14] [Tiedemann, 2012]. The corpora are collected from different sources and present different genres of text.

Note that OST is a collection of movie subtitles translated and uploaded by contributors. These contributors are just ordinary persons and they are not qualified translators. Therefore, the quality of the translations may vary from one to another.

[8] `www.afp.com`
[9] `www.annahar.com`
[10] `www.assabah.com.tn`
[11] `www.medar.info`
[12] `www.ted.com`
[13] `www.opensubtitles.org`
[14] `www.tatoeba.org`

TABLE 3.8: Parallel Corpora characteristics

Corpus	$\|S\|$	$\|W\|$		$\|V\|$	
		English	Arabic	English	Arabic
Newspapers					
AFP	4K	140K	114K	17K	25K
ANN	10K	387K	288K	39K	63K
ASB	4K	187K	139K	21K	34K
Medar	13K	398K	382K	43K	71K
NIST	2K	85K	64K	15K	22K
United Nations Resolutions					
UN	61K	2.8M	2.4M	42K	77K
Talks					
TED	88K	1.9M	1.6M	88K	182K
Movie Subtitles					
OST	2M	31M	22.4M	504K	1.3M
Other					
Tatoeba	1K	17K	13K	4K	6K
Total	2.3M	37M	27.5M	775K	1.8M

As can be noted from Table 3.8, in all parallel corpora, English texts have more words than Arabic ones. The reason is that certain Arabic terms can be agglutinated, while English terms are isolated, as described in Section 3.2. In contrast, the vocabulary of Arabic texts is larger than the vocabulary of the English one. This is because Arabic is a highly inflected language, as described in Section 3.2.

We merge parallel news corpora (AFP, ANN, ASB, Medar and NIST) into one corpus, and we call it *parallel-news*. We merge these corpora because they are from the same domain (newspapers). The characteristics of this corpus are presented in Table 3.9.

TABLE 3.9: *parallel-news* corpus characteristics

$\|S\|$	$\|W\|$		$\|V\|$	
	English	Arabic	English	Arabic
34K	1.2M	0.9M	83K	141K

To recapitulate this chapter, we described some characteristics of Arabic language. We pointed out that some consideration should be taken into account for Arabic and English-Arabic NLP task.

We further presented in this chapter our collected comparable corpora. These comparable corpora are collected from different sources and different domains. The corpora are collected from Wikipedia encyclopedia and from Euronews news agency. We also described the parallel corpora that we use in this thesis. The parallel corpus is useful for us because we use it as a baseline corpus to compare it with our collected comparable corpora.

Chapter 4

Cross-lingual Similarity Measures

In this chapter, we present two cross-lingual similarity measures: based on bilingual dictionary and based on Latent Semantic Indexing (LSI). We use these measure for cross-lingual document retrieval, i.e., to retrieve the target document using the source document as a query. We evaluate these methods on parallel and comparable corpora. We further compare and discuss the performance of the two measures.

In the dictionary based method, we developed a new morphological analysis technique for Arabic words. We further investigate the best morphological analysis technique to match English-Arabic words.

To our knowledge, LSI method has been applied for several language pairs, but Arabic language has not been addressed yet. In this chapter, we investigate LSI method for English-Arabic document retrieval.

Finally, we use best one of the two cross-lingual similarity measures to align further comparable documents collected from sources other than Wikipedia and Euronews. In this chapter, we align news documents collected from the British broadcasting corporation and Al-Jazeera news agencies.

4.1 Introduction

As we mentioned earlier, texts in the *parallel* corpus have aligned sentences, which are translations of each other, while texts in the *comparable* corpus have topic aligned documents, which are not necessarily translations of each other.

Similarity of comparable documents can vary, and we need a measure that can identify the degree of similarity of these documents. A similarity measure is a real-valued function that quantifies the likeness of the meaning or the contents of two documents. The function estimates the distance between two units of text (terms, sentences, paragraphs, documents, or concepts) through numerical representations of the text documents.

This chapter presents our similarity measures for English-Arabic documents. These measures can identify the degree of similarity of two cross-lingual documents, and they can be used to align and retrieve comparable documents. The value of these measures range from 1 (exactly similar) to 0 (not similar).

We present two similarity measures in this chapter: the first one is based on a bilingual dictionary, and the second one is based on the Cross-Lingual Latent Semantic Indexing (CL-LSI) method. In the following sections, we present our measures, our experiment setup, then we discuss and compare the results.

4.2 Cross-lingual Similarity Using Bilingual Dictionary

Our dictionary based method, uses multi-WordNet bilingual dictionary [Bond and Paik, 2012], to measure the similarity of comparable documents. This method requires the source and target texts to be represented as vectors of matched words. Source and target words are matched using the bilingual dictionary.

For inflected language, bilingual dictionaries usually do not cover all word variations, so morphological analysis is applied on words to improve the dictionary coverage.

Word weights can be either binary (1 or 0 to indicate the presence or absence of the translation in the target document) or numeric represented by the term frequency-inverse document frequency (*tfidf*) of words in the document.

For binary weighting scheme we propose a binary measure, and for *tfidf* weighting scheme we propose a cosine measure. For a given source document d_s and target document d_t, the binary measure counts the words in d_s which have translations in d_t and then normalizes these counts by the vector size, while the cosine measure computes the cosine similarity between source and target vectors which are represented by the *tfidf* of the matching words of d_s and d_t.

The binary measure uses the function $trans(w_s, d_t)$, which returns 1 if a translation of the source word w_s is found in the target document d_t, and 0 otherwise. The similarity using the binary measure can be computed as follows:

$$bin(d_s, d_t) = \frac{\sum_{w_s \in d_s \cap V_s} trans(w_s, d_t)}{|d_s \cap V_s|} \tag{4.1}$$

where V_s is the source vocabulary of the bilingual dictionary, d_s and d_t are the source and target documents considered as bags of words. Because $bin(d_s, d_t)$ is not symmetric, the actual value used for measuring the comparability between d_s and d_t is as follows:

$$\frac{bin(d_s, d_t) + bin(d_t, d_s)}{2} \tag{4.2}$$

Cosine similarity is a measure of similarity, between two vectors in a vector space, that measures the cosine of the angle between the two vectors. Source and target texts can be represented as vectors where the value of each dimension corresponds to

weights/features (for e.g. *tfidf*) associated to the matched words in the documents. This representation is generally referred to as a Vector Space Model (VSM). Given two vectors d_s and d_t of n attributes representing the source and target documents, the cosine similarity $cosine(d_s, d_t)$ between these documents is computed as follows:

$$cosine(d_s, d_t) = \frac{d_s \cdot d_t}{\|d_s\| \times \|d_t\|} = \frac{\sum_{i=1}^{n} d_{s_i} \times d_{t_i}}{\sqrt{\sum_{i=1}^{n} (d_{s_i})^2} \times \sqrt{\sum_{i=1}^{n} (d_{t_i})^2}} \tag{4.3}$$

To represent cross-lingual documents in the VSM, we build the source and target vectors as follows: using a bilingual dictionary, for each translation $w_s \leftrightarrow w_t$ in this dictionary, define one attribute of the vectors. For the source vector this attribute is equal to the *tfidf* of w_s (0 if w_s is not in the source document), and for the target vector this attribute is equal to the *tfidf* of w_t (0 if w_t is not in the target document).

The term frequency-inverse document frequency (*tfidf*) for a term t_i, in a document d_j, in a corpus C is computed as follows:

$$tfidf(t_i, d_j, C) = tf(t_i, d_j) \times idf(t_i, C) \tag{4.4}$$

where $tf(t_i, d_j)$ is the frequency of the term t_i in the document d_j, and $idf(t_i, C)$ is the frequency of documents that the term t_i appeared in. $tf(t_i, d_j)$ and $idf(t_i, C)$ are computed as follows:

$$tf(t_i, d_j) = |t_i : t_i \in d_j| \tag{4.5}$$

$$idf(t_i, C) = \log \frac{|C|}{|\{d \in C : t_i \in d\}|} \tag{4.6}$$

where $|C|$ is the corpus size.

Open Multilingual WordNet (OMWN) bilingual dictionary [Bond and Paik, 2012] is used in our work to match the source and the target words. OMWN is available in many languages including Arabic and English. OMWN has 148K English words, which are extracted from English WordNet [Miller and Fellbaum, 1998] and 14K Arabic words, which are extracted from Arabic WordNet [Black et al., 2006]. Synonym words are grouped into sets called synsets. These synsets help to identify possible translations from source to target.

To match words in the source and the target texts, each word is looked up in the bilingual dictionary. Before that, morphological analysis is applied on words to increase the coverage of dictionary between source and target texts. Also stopwords and punctuation are removed from all the texts before matching words.

There are many word reduction techniques for English and Arabic languages. For English, stemming and lemmatization are widely used in the community. Stemming [Porter, 2001] prunes a word into a stem, which is a part of the word, and may not be in the dictionary, while lemmatization [Miller and Fellbaum, 1998] retrieves the dictionary form (lemma) of an inflected word.

As for Arabic, rooting [Khoja and Garside, 1999, Taghva et al., 2005] or light stemming [Larkey et al., 2007] are widely used techniques. Rooting removes the word's prefix, suffix and infix, then converts it to the root form, while light stemming just removes the word's prefix and suffix. As discussed in Section 3.2, Arabic is a highly inflected language. Thus, applying rooting leads to lose the meaning of Arabic words against the corresponding English words.

In order to increase English-Arabic word matching using the bilingual dictionary, we have developed a new reduction technique for Arabic words, which combines rooting and light stemming techniques. We name this technique as *morphAr*. The idea is to try to reduce Arabic words by applying light stemming first, and then applying rooting. If the stem is found in the dictionary, then its translations are returned, otherwise the translations of the root are returned.

We have two reduction techniques for English (stemming and lemmatization) and three techniques for Arabic (light stemming, rooting and *morphAr*). To determine the best combination of these techniques, we conducted an experiment using each technique separately, also inspecting the percentage of words that are Out Of Vocabulary (OOV). This experiment is applied on AFP, ANN, ASB, TED, UN parallel corpora, which are described in Section 3.4. The OOV rate is computed as follows:

$$\frac{1}{2} \times \left(\frac{|w_s^{oov}|}{|d_s|} + \frac{|w_t^{oov}|}{|d_t|}\right) \tag{4.7}$$

where d_s is the source document, d_t is the target document, $|d|$ is the word count in the document and $|w^{oov}|$ is the count of the words that are OOV (not found in the dictionary).

Word matching rate is the count of source and target words that are translation of each others in the source and the target documents ($|w_s \leftrightarrow w_t|$), normalized by source and target document sizes. It is computed as follows:

$$\frac{|w_s \leftrightarrow w_t|}{|d_s| + |d_t|} \tag{4.8}$$

Figure 4.1 presents the OOV rate for each word reduction technique separately. If we consider word reduction techniques for each language separately, then rooting for Arabic and lemmatization for English have the lowest OOV rate as shown in Figure 4.1. But we do not aim to just reduce OOV independently for each language. Instead, we aim to increase matching rate of source and target words by finding the appropriate translation for these words using the bilingual dictionary. In addition, as we discussed in Chapter 3, using rooting in Arabic language leads to lose the corresponding meaning in the English language.

The word matching rates for different combinations of the word reduction techniques in both Arabic and English are presented in Figure 4.2. It can be noted that *morphAr*

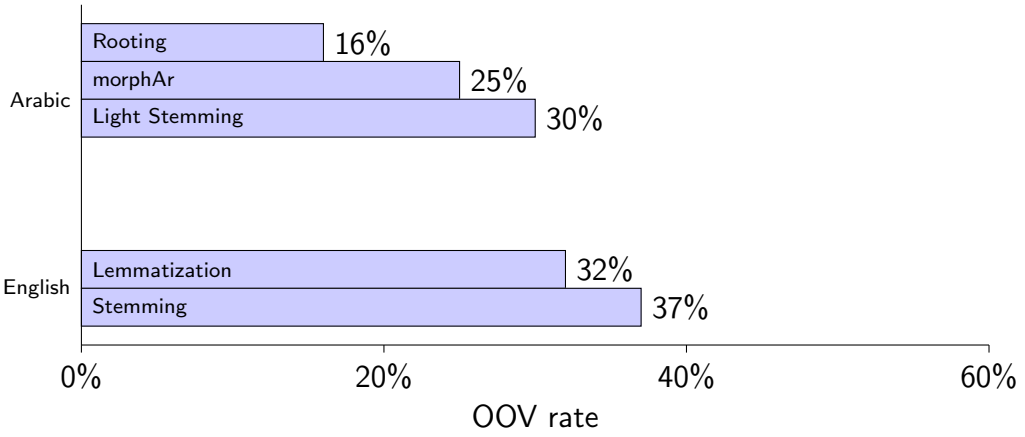

FIGURE 4.1: OOV rate using different word reduction techniques for Arabic and English parallel corpus

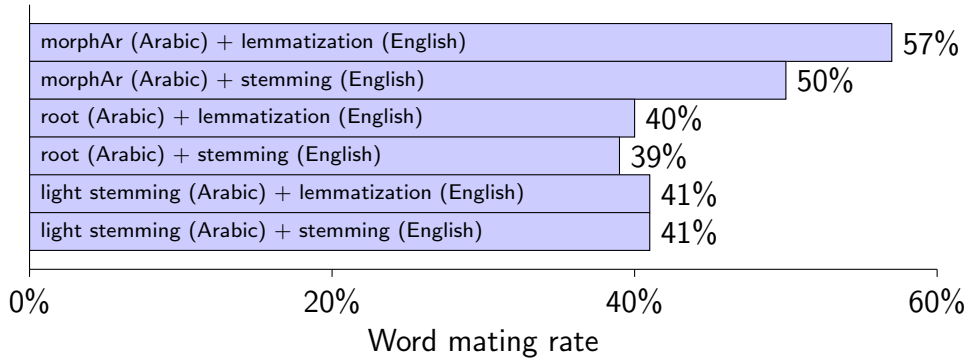

FIGURE 4.2: Word matching rate of combined Arabic-English word reduction techniques using the bilingual dictionary

for Arabic and lemmatization for English lead together to the best coverage (best matching rate). Therefore, we use this combination of techniques in our experiments.

4.2.1 Results

To evaluate the performance of the Dictionary based methods we conducted experiments on parallel and comparable corpora. We select a random sample of 100 English-Arabic sentences from each parallel corpus, and random sample of 100 English-Arabic documents from each comparable corpus. Parallel and comparable corpora are described in Sections 3.4 and 3.3 respectively.

Each source text (English) is used as a query to retrieve exactly one relevant target text (Arabic). The experiment is conducted at the sentence level for parallel corpora and at the document level for comparable corpora. In other words, for parallel corpora, the source sentence is used as a query to retrieve its translation in the target language. For comparable corpora, the source document is used as a query to retrieve its target pair.

The retrieving process is done as follows: for a given source text, we compute the similarity to all of the 100 target texts, then we select the top-n target texts according to the similarity values. Then, we check if the corresponding target text is in the 1-top list (first recall R@1), in the 5-top list (fifth recall R@5), and in the 10-top list (tenth recall R@10).

Tables 4.1 and 4.2 present the recall results for parallel and comparable corpora respectively, using the binary measure (Dict-bin) and the cosine measure (Dict-cos). Dict-bin is computed using the Formula 4.2 and Dic-cos is computed using the formula 4.3. The tables show that cosine measure achieves better results than the binary measure in terms of recall scores. The recall scores of each measure depend on the corpus as shown in the results. For binary measure, the best R@1 is achieved on Tatoeba corpus, and for cosine measure, the best R@1 is achieved on NIST corpus. However, the recall scores are still limited for both measures, and for both parallel and comparable corpora. This is due to the limitations of the dictionary and the morphological tools. Besides that, word-to-word translations based on dictionaries can lead to many errors (translation ambiguity).

4.2.2 Conclusion

In this section we proposed two dictionary based similarity measures for cross-lingual documents.

TABLE 4.1: Recall of retrieving parallel documents using Dict-bin and Dict-cos measures

Corpus	Method	*R*@1	*R*@5	*R*@10
Newspapers				
AFP	Dict-bin	0.11	0.24	0.32
	Dict-cos	0.30	0.69	0.80
ANN	Dict-bin	0.11	0.24	0.37
	Dict-cos	0.39	0.72	0.78
ASB	Dict-bin	0.12	0.32	0.42
	Dict-cos	0.38	0.72	0.83
Medar	Dict-bin	0.09	0.22	0.32
	Dict-cos	0.35	0.71	0.82
NIST	Dict-bin	0.16	0.32	0.44
	Dict-cos	0.46	0.78	0.84
United Nations Resolutions				
UN	Dict-bin	0.13	0.30	0.32
	Dict-cos	0.32	0.63	0.74
Talks				
TED	Dict-bin	0.22	0.45	0.56
	Dict-cos	0.41	0.76	0.82
Movie Subtitles				
OST	Dict-bin	0.25	0.43	0.53
	Dict-cos	0.31	0.55	0.62
Other				
Tatoeba	Dict-bin	0.26	0.44	0.51
	Dict-cos	0.45	0.69	0.77

TABLE 4.2: Recall of retrieving comparable documents using Dict-bin and Dict-cos measures

Corpus	Method	*R*@1	*R*@5	*R*@10
Euronews	Dict-bin	0.03	0.13	0.19
	Dict-cos	0.20	0.52	0.62
AFEWC	Dict-bin	0.08	0.18	0.25
	Dict-cos	0.24	0.46	0.57

We proposed a new morphological analysis technique (*MorphAr*) for Arabic words to match them with English words. We experimentally investigated different combination of English-Arabic morphological analysis techniques to determine the best one to match English-Arabic words. We found that *MorphAr* for Arabic and lammatization for English lead together to best matching rate for English and Arabic words.

We compared Dict-bin and Dict-cos measure, and the results showed that Dict-cos measure gives better results than the binary measure. However, the dictionary based method has limited performance due to the limitations of the bilingual dictionaries and the morphological analysis tools. Moreover, word-to-word matching based on dictionaries can lead to many errors.

In the next section, we use another cross-lingual similarity measure based on Cross-Lingual Latent Semantic Indexing (CL-LSI). We also compare the dictionary based measure with the CL-LSI based measure.

4.3 Cross-lingual Similarity Using CL-LSI

In this section we present a cross-lingual similarity measure based on the Cross-Lingual Latent Semantic Indexing (CL-LSI). This method can be used to align, retrieve and compare cross-lingual documents. The advantage of this method is that it does not need bilingual dictionaries, morphological analyzers or machine translation systems. Moreover, this method overcomes the problem of vocabulary mismatch between queries and documents.

In our work, we use the same approach as [Littman et al., 1998], but we apply it on Arabic-English documents. Moreover, [Littman et al., 1998] used parallel corpus to train the CL-LSI, whereas we use both parallel and comparable corpora for training.

Let us consider a term-document matrix X that describes the weights of terms that occur in a collection of documents as shown in 4.9. Rows of this matrix correspond to the terms and columns correspond to the documents in the collection.

$$X = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & \dots & d_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \dots & w_{2n} \\ w_{31} & w_{32} & w_{33} & \dots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & w_{m3} & \dots & w_{mn} \end{pmatrix}_{m \times n} \end{array} \tag{4.9}$$

In LSI, the term-document matrix (m terms $\times$ n documents) is decomposed by Singular Value Decomposition (SVD) into three matrices; (1) the term matrix (U), which is an $m \times k$ matrix, where k is the reduced dimensions. Each column vector in U maps terms in the corpus into a single concept of semantically related terms that are grouped with similar values. (2) a diagonal matrix (S), which is an $k \times k$ matrix of singular values. (3) the document matrix (V^T), which is an $k \times n$ matrix [Deerwester et al., 1990] (See 4.10). U and V^T are the left and right singular vectors respectively, while S is a diagonal matrix of singular values. k is the reduced concept space in LSI. [Landauer et al., 1998, Dumais, 2007] reported that the optimal value of k to perfom SVM is between 100 and 500. That depends on the task and the nature of data. Thus, one can determine the optimal value of k between 100 and 500 experimentally.

$$\overset{\text{terms}}{\left(\begin{array}{c} \overset{\text{docs}}{\text{X}} \end{array} \right)_{m \times n}} = \left(\begin{array}{c} \text{U} \end{array} \right)_{m \times k} \left(\begin{array}{c} \text{S} \end{array} \right)_{k \times k} \left(\begin{array}{c} V^T \end{array} \right)_{k \times n} \tag{4.10}$$

For monolingual LSI, X is an $m \times n$ matrix that represents a monolingual corpus consisting of n documents, and m terms as shown in (4.9).

The matrix X for cross-lingual LSI is shown in (4.11). The source and the target texts are concatenated into one document. Therefore, each d_i^u is the concatenation of the source text d_i^s (Arabic) and its target (English) d_i^t. Consequently, X describes a bilingual corpus that consists of n cross-lingual documents, l Arabic terms, and m English terms. I.e., X is an $(l+m) \times n$ matrix.

$$X = \begin{array}{c} \\ t_1^s \\ t_2^s \\ t_3^s \\ \vdots \\ t_l^s \\ t_1^t \\ t_2^t \\ t_3^t \\ \vdots \\ t_m^t \end{array} \begin{array}{c} \begin{array}{ccccc} d_1^u & d_2^u & d_3^u & \dots & d_n^u \end{array} \\ \begin{pmatrix} w_{11}^s & w_{12}^s & w_{13}^s & \dots & w_{1n}^s \\ w_{21}^s & w_{22}^s & w_{23}^s & \dots & w_{2n}^s \\ w_{31}^s & w_{32}^s & w_{33}^s & \dots & w_{3n}^s \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{l1}^s & w_{l2}^s & w_{l3}^s & \dots & w_{ln}^s \\ w_{11}^t & w_{12}^t & w_{13}^t & \dots & w_{1n}^t \\ w_{21}^t & w_{22}^t & w_{23}^t & \dots & w_{2n}^t \\ w_{31}^t & w_{32}^t & w_{33}^t & \dots & w_{3n}^t \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{m1}^t & w_{m2}^t & w_{m3}^t & \dots & w_{mn}^t \end{pmatrix}_{(l+m)\times n} \end{array} \tag{4.11}$$

X can be used to describe parallel or comparable corpus. For a parallel corpus, each d_i^u represents a pair of parallel sentences, while for a comparable corpus, it represents a pair of comparable documents.

w_{ij} in the matrices 4.11 and 4.9 are the term frequency-inverse document frequency (*tfidf*) weights. See Section 4.2 for details about how *tfidf* is computed.

The term-document matrix as formulated in Equation 4.11, enables LSI to learn the relationship between terms, which are semantically related within the same language and between two languages.

This method helps us to achieve our objective to retrieve comparable articles. We concatenate source and target texts of a training corpus, then the LSI model is trained

on these texts. The source and target test texts are then projected in the LSI space separately. Each source test text is used as a query and compared to all target test texts to find the target pair which corresponds to the source text. We describe this process in the next section in detail.

4.3.1 Experiment Procedure

In this section, we describe how LSI matrices are built and how they are used to retrieve comparable articles.

Building LSI Matrices

The method below describe how LSI matrices are built:

1. Split English-Arabic corpora into training (90%) and test (10%) subsets.
2. Use Arabic training corpus to create the matrix X as in Equation 4.9. Then apply LSI to obtain USV^T; the monolingual LSI matrix (AR-LSI) is shown in Figure 4.3.
3. Use English-Arabic training corpus to create the matrix X as in Equation 4.11. Then apply LSI to obtain USV^T; the cross-lingual LSI matrix (CL-LSI) is shown in Figure 4.4.

As we mentioned earlier, the optimal value of k of the USV^T for AR-LSI and CL-LSI can be chosen experimentally. To choose this value, we follow the experience of [Landauer et al., 1998, Dumais, 2007], who report that the optimal value of k to perform SVM is between 100 and 500. We conducted several experiments in order to determine the best rank for AR-LSI and CL-LSI, and we found that the dimension 300 optimizes the similarity for the parallel corpus. Therefore, we use this dimension in all our experiments.

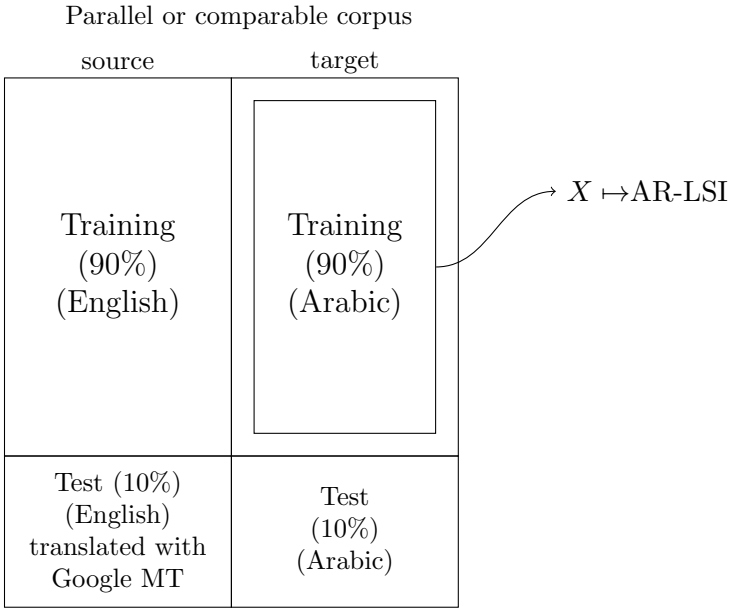

FIGURE 4.3: The monolingual (Arabic) LSI model (AR-LSI)

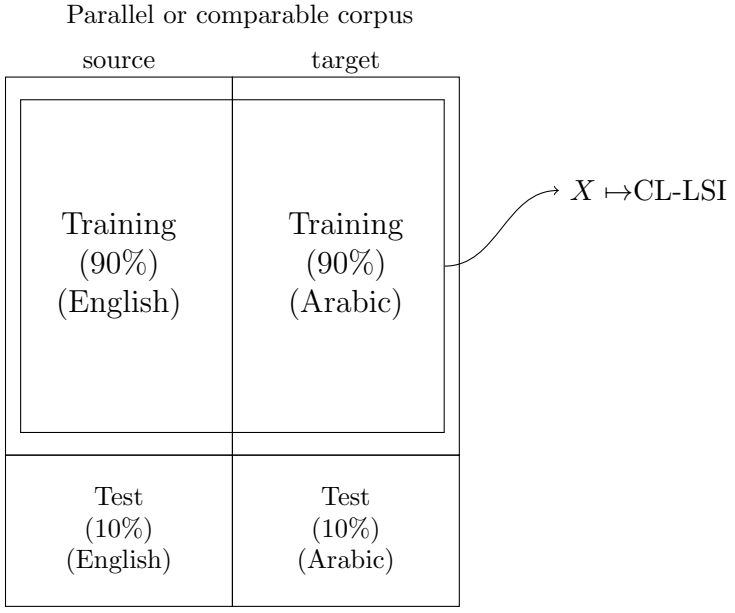

FIGURE 4.4: The cross-lingual LSI model (CL-LSI)

In our work, we used the implementation of LSI in the Gensim package [Rehurek and Sojka, 2010] to build AR-LSI and CL-LSI models.

Retrieving Text Pairs

The test corpus is composed of n pairs of English (e_i) and Arabic (a_j) texts (aligned at the sentence level in parallel corpus and at the document level in comparable corpus). i.e., each text pair consists in the source and target sentences in the parallel corpus, while it consists in the source and the target documents in the comparable corpus.

For a given source text e_i (English), the task is to retrieve the target text a_j (Arabic). More precisely, each source text (English) is used as a query to retrieve exactly one relevant target text (Arabic). This procedure is applied at the sentence level for parallel corpora and at the document level for comparable corpora. In other words, for parallel corpora, the source sentence is used as a query to retrieve its translation in the target language. For comparable corpora, the source document is used as a query to retrieve its target pair.

The procedure involves the following steps: all English and Arabic texts are preprocessed by removing punctuation marks, stopwords (common words) and words that appeared less than three times (low-frequency words) in the corpus. The source text is compared with all target texts and then the most similar target texts are selected. Both source and target texts are mapped into LSI space. The cosine similarity value is computed for the vectors in the LSI space.

The method is applied on parallel and comparable corpora described in Sections 3.4 and 3.3 respectively.

Algorithms 1 and 2 describe the method to retrieve the most similar a_j to an English document e_i using AR-LSI and CL-LSI respectively.

Algorithm 1 takes the English test corpus C_e and the Arabic test corpus C_a. All Arabic documents of C_a are transformed into AR-LSI (built from the Arabic training corpus) space. Then, each English document e_i is translated into Arabic using Google

MT service[1]. Next, the translated document a_{e_i} is transformed into AR-LSI space. Then the most similar Arabic documents are retrieved from Arabic corpus.

Algorithm 1: Retrieving Arabic documents using AR-LSI

Input: C_e: English corpus, C_a: Arabic corpus, n: number of docs to retrieve

1 $C'_e \leftarrow \varnothing$; $C'_a \leftarrow \varnothing$;
2 **foreach** *doc* a_j *in* C_a **do**
3 $\quad a'_j \leftarrow a_j^t U S^{-1}$; put a'_j in C'_a // map a_j into AR-LSI
4 **foreach** *doc* e_i *in* C_e **do**
5 $\quad a_{e_i} \leftarrow$ translate(e_i) // translate e_i into Arabic
6 $\quad a'_{e_i} \leftarrow a_{e_i}^t U S^{-1}$ // map a_{e_i} into AR-LSI
7 $\quad R \leftarrow$ retrieve(a'_{e_i}, C'_a, n) // retrieve top-n similar docs to e'_i from C'_a
8 $\quad$ evaluate(R) // check if a'_i is in R

Retrieving Arabic documents using CL-LSI is done in a similar manner as AR-LSI, but machine translation service is not used. Algorithm 2 describes how CL-LSI is used to retrieve Arabic documents that are comparable to an English document. Algorithm 2 also takes the English test corpus C_e and the Arabic test corpus C_a. All documents in C_e and C_a are transformed into CL-LSI (built from the English-Arabic training corpus) space. Each e_i is used as a query to retrieve the target pair from C_a using *retrieve* procedure.

Algorithm 2: Retrieving Arabic documents using CL-LSI

Input: C_e: English corpus, C_a: Arabic corpus, n: number of docs to retrieve

1 $C'_e \leftarrow \varnothing$; $C'_a \leftarrow \varnothing$;
2 **foreach** *doc* a_j *in* C_a **do**
3 $\quad a'_j \leftarrow a_j^t U S^{-1}$; put a'_j in C'_a // map a_j into CL-LSI
4 **foreach** *doc* e_i *in* C_e **do**
5 $\quad e'_i \leftarrow e_i^t U S^{-1}$ // map e_i into CL-LSI
6 $\quad R \leftarrow$ retrieve(e'_i, C'_a, n) // retrieve top-n similar docs to e'_i from C'_a
7 $\quad$ evaluate(R) // check if a'_i is in R

The difference between Algorithm 1 and 2 is the highlighted lines in the both algorithms. The English document in Algorithm 1 is translated into Arabic first, then it

[1] http://translate.google.com

is mapped to LSI space, while the English document in Algorithm 2 is mapped into the LSI space directly. Machine translation is needed in Algorithm 1 because the LSI model is monolingual, but machine translation is not needed in Algorithm 2 because the LSI model is cross-lingual.

The *retrieve* function takes the source document d_s, the target corpus C_t, and the number of documents to retrieve (n). The source document d_s is compared with all documents in the target corpus C_t. The procedure then returns the top n most similar documents.

Procedure retrieve(d_{s_i}, C_t, n)

Input: d_{s_i}: source doc, C_t: target corpus, n: number of docs to retrieve

$R \leftarrow \varnothing$; `// a list of retrieved docs`
foreach *doc* d_{t_j} *in* C_t **do**
    $sim \leftarrow \cos(d_{s_i}, d_{t_j})$ `// compute the similarity to all target docs`
    put (j, sim) in R;
`sort(`R`)` `// sort` R `in descending order according to` sim `values`
return *top* n *elements of* R;

The evaluation in Algorithms 1 and 2 is done as follows: given e_i, a_j is considered to be correctly retrieved if and only if $i = j$. In other words, for each source document, we consider there is exactly one relevant document (its pair). The list R is composed of indexes of retrieved documents. The condition $(i = j)$ is checked in the top-1 (recall at 1 or R@1), top-5 (recall at 5 or R@5), and top-10 (recall at 10 or R@10) of the list R. The performance measure is defined as the percentage of a_i, which are correctly retrieved in R@1, R@5, R@10 lists, among all e_i.

4.3.2 Results

Retrieving Parallel Documents

The results of retrieving parallel documents (at the sentence level) using AR-LSI and CL-LSI are presented in Table 4.3. The table also shows the results of the same

TABLE 4.3: Recall of retrieving parallel documents using AR-LSI, CL-LSI and Dict-cos methods

Corpus	Method	R@1	R@5	R@10
Newspapers				
AFP	AR-LSI	0.98	1.00	1.00
	CL-LSI	0.99	1.00	1.00
	Dict-cos	0.30	0.69	0.80
ANN	AR-LSI	0.88	0.96	0.98
	CL-LSI	0.88	0.96	0.98
	Dict-cos	0.39	0.72	0.78
ASB	AR-LSI	0.91	0.96	0.99
	CL-LSI	0.92	0.98	0.98
	Dict-cos	0.38	0.72	0.83
Medar	AR-LSI	0.77	0.92	0.97
	CL-LSI	0.81	0.97	0.99
	Dict-cos	0.35	0.71	0.82
NIST	AR-LSI	0.82	0.94	0.96
	CL-LSI	0.83	0.91	0.94
	Dict-cos	0.46	0.78	0.84
United Nations Resolutions				
UN	AR-LSI	0.97	1.00	1.00
	CL-LSI	0.98	1.00	1.00
	Dict-cos	0.32	0.63	0.74
Talks				
TED	AR-LSI	0.57	0.74	0.83
	CL-LSI	0.57	0.82	0.87
	Dict-cos	0.41	0.76	0.82
Movie Subtitles				
OST	AR-LSI	0.39	0.61	0.72
	CL-LSI	0.33	0.76	0.85
	Dict-cos	0.31	0.55	0.62
Other				
Tatoeba	AR-LSI	0.60	0.75	0.79
	CL-LSI	0.47	0.72	0.78
	Dict-cos	0.45	0.69	0.77

corpora using Dict-cos measure described in Section 4.2 for comparison purpose. The results are for the same 100 random sample, which are selected in Section 4.2.

As shown in Table 4.3, it is not easy to get a general conclusion about the performance of LSI since it depends on the nature of the corpus and on the desired recall (R@1,

R@5 or R@10). For example, for AFP, ASB, NIST, Medar, and UN corpora, CL-LSI is slightly better than AR-LSI for R@1. In contrast, for OST and Tatoeba, AR-LSI is better than CL-LSI. The performance of the CL-LSI is equal to, or better than the AR-LSI in 6 out of 9 of corpora for R@1.

We checked the significance of differences of the results using McNemar's test [McNemar, 1947]. The conclusion is that they are not significantly different. Therefore, both approaches obtain mostly similar performance. However, it should be noted that the CL-LSI does not require a MT system. Therefore, we can affirm that the CL-LSI is competitive compared to the AR-LSI.

The performance of AR-LSI and CL-LSI approaches on OST corpus is poor because of the nature of this corpus. The OST corpus is composed of subtitles that are translated by many users as mentioned in Section 3.4.

The results show that MT can be sufficient for cross-lingual retrieval. However, to investigate the effect of the performance of the MT system on the performance of the AR-LSI, we run an experiment to simulate a perfect MT system. This is done by retrieving an Arabic document by providing the same document as a query. In other words, source and target documents are the same. This experiment is done on all corpora and the results of R@1 is 1.0 for each corpus of the parallel corpora. This result reveals the lack of robustness of AR-LSI according to the MT system's performance.

We compare our method with the cosine measure of the dictionary-based method (Dict-cos) presented in Section 4.2. As shown in Table 4.3, both LSI methods achieve better results than Dict-cos method for all corpora. It can be concluded that LSI method is better and more robust than Dict-cos since it does not need any dictionary or morphological analysis, and it is language independent.

Finally, we compare our LSI result to the results in [Littman et al., 1998]. The authors of the papers worked on French-English document retrieval using LSI. They

applied the method on Hansard parallel corpus[2], which is the proceedings of the Canadian parliament. UN corpus in our work is close (in terms of domain) to Hansard corpus. Therefore, the results can be compared. [Littman et al., 1998] reported 0.98 of R@1 on English-French texts of Hansard corpus and, we achieved a similar result on English-Arabic texts of the UN corpus using the CL-LSI method.

Retrieving Comparable Documents

The same experimental protocol as described in Section 4.3.1 is applied to retrieve the documents of comparable corpus. The difference is that the CL-LSI matrix is built using the training part of the comparable corpus. The objective of this experiment is to investigate using comparable corpora for training CL-LSI to retrieve cross-lingual documents. In this experiment, the source document (English is used as a query to retrieve its target comparable document (exactly one relevant Arabic document).

Results of retrieving comparable documents (at the document level) using CL-LSI are presented in Table 4.4. The table shows the recall scores of the CL-LSI method on Euronews and AFEWC comparable corpora. The recall of CL-LSI on Euronews corpus is better than on AFEWC corpus. This could be due to the fact that Euronews articles are mostly translations of each other, while Wikipedia articles are not necessarily translations of each other as mentioned in Section 3.3.

TABLE 4.4: Recall of retrieving comparable documents using CL-LSI method

Corpus	R@1	R@5	R@10
Euronews	0.95	1.00	1.00
AFEWC	0.68	0.98	1.00

From Tables 4.4 and 4.3, it can be noted that CL-LSI can retrieve the target information at the document level and at the sentence level respectively with almost the same performance.

[2] `www.isi.edu/natural-language/download/hansard`

Comparing Corpora

We use the CL-LSI method in order to study the comparability of parallel and comparable corpora. We achieve this by computing the average cosine (avg(cos)) for all the pair articles in the test parts of these corpora. For each corpus, the CL-LSI matrix is built from the training part, and used to compute the avg(cos) for the test part. This experiment is done on parallel and comparable corpora. Statistics of comparability are presented in Table 4.5. It is not easy to get a general conclusion from the result about the value of avg(cos) for parallel or comparable corpora. The result shows that value of avg(cos) does not depend on the type of the corpus (parallel or comparable), but it depends on the nature and the quality of the corpus.

Table 4.5: Statistics of comparability using CL-LSI

Corpus	avg(cos)
Parallel	
AFP	0.65
ANN	0.53
ASB	0.54
Medar	0.43
NIST	0.45
UN	0.73
TED	0.29
tatoeba	0.28
Comparable	
Euronews	0.73
AFEWC	0.36

However, the average similarities presented in Table 4.5 confirm the results presented in Tables 4.3 and 4.4. I.e., corpora that have high avg(cos) values also have high recall scores. For instance, corpora that have the highest (avg(cos)) are UN and Euronews corpora, and their R@1 scores are 0.98 and 0.95 respectively. In the contrary, corpora that have the lowest (avg(cos)) are TED and tatoeba corpora, and their R@1 scores are 0.57 and 0.47 respectively. R@1 scores for UN and Euronews corpora are the highest among the other corpora, while it is the lowest for TED and tatoeba corpora.

We conducted another experiment to compare the average pairwise similarity for aligned and non-aligned test corpus. The objective is to investigate the ability of CL-LSI similarity measure to distinguish the difference (degree of comparability) between the aligned and non-aligned corpus. This experiment is done on UN and Euronews corpora. Non-aligned corpus is generated simply by shuffling the order of the source texts in the corpus. This shuffling makes the sources texts to be not aligned to the their targets.

Table 4.6 shows that the average pairwise similarity for the non-aligned test corpus is lower than the aligned test corpus. The result shows that CL-LSI captured the difference between aligned and non-aligned corpus. So CL-LSI can be used to study the degree of comparability in cross-lingual corpora, and to distinguish between aligned and non-aligned corpora.

TABLE 4.6: Average Similarity of aligned vs. not aligned corpus using CL-LSI

Corpus	Aligned	Non-aligned
UN	0.73	0.07
Euronews	0.73	0.09

4.3.3 Conclusion

In this section we described a method that can be used to measure the similarity of cross-lingual documents. This method is based on LSI, which we used in two ways: monolingual (AR-LSI) and cross-lingual (CL-LSI). The first method needs to use a machine translation system in order to translate the source into the language of the target text, while the second method merges the training data of both languages. In the test step, the comparison is done between vectors of the same type.

We applied these methods on several corpora and the results showed that the CL-LSI can be competitive to the AR-LSI. The advantage of CL-LSI is that it does need machine translation. The results also showed that the method can be used to retrieve comparable pairs. Both CL-LSI and AR-LSI achieved better results than

the dictionary based method. In addition, LSI methods do not need morphological analysis tools or bilingual dictionaries, and they are language independent.

In this section, we used CL-LSI to retrieve comparable documents of Euronews and Wikipedia corpora, but in the next section we use it to align cross-lingual documents collected from other different sources.

4.4 Aligning Comparable Documents Collected From Different Sources

One of the objectives of this dissertation is to study comparable documents, which are collected from the Internet. The aim is to inspect if the documents are comparable or not. In case they are comparable, then we study another level of comparability, which is to inspect the agreement of sentiments expressed in these documents. Thus, the pairs of comparable documents can be organized in according to their agreement or disagreement of sentiments and emotions that are expressed in these documents.

A potential application can be for customers, who are interested in product reviews that are written in foreign languages. These reviews can be arranged to the customer according to agreement or disagreement of sentiments that are expressed in these reviews.

Another application, but in the news media domain, is the comparison of news articles. For instance, a journalist can be interested in what is being said about an event in the foreign media. Finding comparable news document pairs allows the journalist to compare these documents from different perspectives such as expressed sentiments and emotions. To analyze reviews and sentiments in this manner it is required to have aligned documents. In this Section, we focus on aligning English-Arabic news articles collected from the Internet.

The English news are collected from the British Broadcast Corporation (BBC) website[3], and the Arabic news are collected from Al-Jazeera (JSC) website[4]. The objective is to be able to study comparable news articles that come from local and foreign sources.

We use the CL-LSI method that is presented in Section 4.3 to align cross-lingual news articles. The task is to align English-Arabic news articles that are related to the same news story or event. In other words, for a given source document, the objective is to retrieve and align the most relevant target document (same news story) and not all similar documents. For example, if the English document is related to "elections in France", we want to retrieve the Arabic document that is related to the same news story, and not any other news article related to "elections". Therefore, this task is more challenging compared to the work in the previous section. This inspects the ability of CL-LSI to perform automatic alignment at the event level. In the previous section we used CL-LSI to retrieve English-Arabic comparable articles of the Euronews corpus. These articles come from the same news agency (Euronews[5]). But in this section, we automatically align BBC-JSC news articles, which come from different news agencies. Moreover, the validation in the method earlier was automatic because Euronews corpus is already aligned, but here we validate the alignment manually. In the next section, we describe the methodology and present the experimental results.

4.4.1 The Proposed Method

The CL-LSI approach needs a parallel or comparable corpus for training as described in Section 4.3. The term-document matrix X of the corpus can be represented as shown in Equation 4.11 described in Section 4.3.

[3] www.bbc.com/news
[4] www.aljazeera.net
[5] www.euronews.com

To build the CL-LSI matrix, we use Euronews comparable corpus, which is described in Section 3.3, to produce X according to Equation 4.11. Then, we apply LSI to obtain USV^T. All text documents are preprocessed by removing punctuation marks, stopwords (common words), and words that appeared less than three times (low-frequency words) in the corpus.

As mentioned earlier, the objective is to align English articles, which are collected from BBC news website, with Arabic news articles, which are collected from JSC website. First, we crawl BBC and JSC websites to collect news articles published in 2012 and 2013 using httrack tool[6]. The articles of the BBC-JSC corpus are then split into several sub-corpora. Each sub-corpus is composed of news articles that are published in a specific month. Consequently, we obtain 24 sub-corpora for each language as shown in Figure 4.5. The number of articles in each month-corpus ranges between 70 and 300.

Algorithm 3: Aligning English-Arabic documents

Input: C_e: English corpus, C_a: Arabic corpus
Result: top N aligned articles

$L \leftarrow \varnothing$; // a list of aligned documents
$C'_e \leftarrow \varnothing$; $C'_a \leftarrow \varnothing$;
foreach *document* e_i *in* C_e **do**
    $e'_i \leftarrow e_i^t U S^{-1}$; // map e_i into CL-LSI
    put e'_i in C'_e;
foreach *document* a_j *in* C_a **do**
    $a'_j \leftarrow a_j^t U S^{-1}$; // map a_j into CL-LSI
    put a'_j in C'_a;
foreach *document* e'_i *in* C'_e **do**
    $(a_j, sim) \leftarrow$ `align`(e'_i, C'_a);
    put (e_i, a_j, sim) in L;
`sort`(L) // sort L in descending order according to sim values
Select top N elements from L;

Each BBC sub-corpus and its corresponding JSC sub-corpus are provided to the CL-LSI to perform the automatic alignment as shown in Figure 4.5.

[6] www.httrack.com

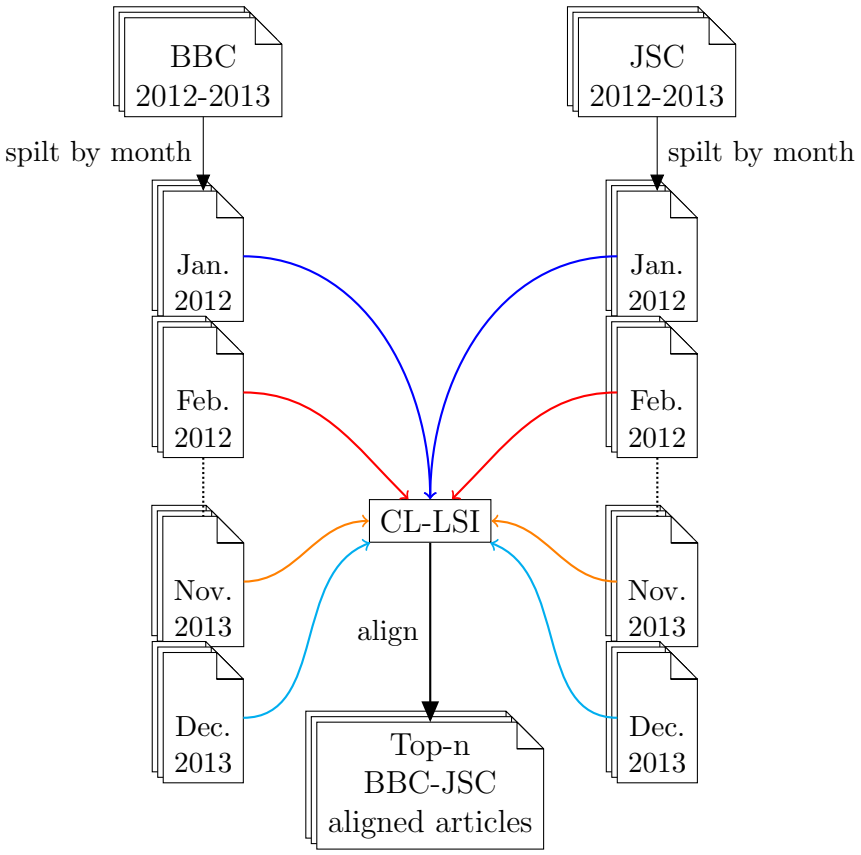

FIGURE 4.5: Automatic alignment of BBC and JSC news stories

Procedure align(e'_i, C'_a)

Input: e'_i, C'_a
Output: a_j, sim

```
L ← ∅; // a list of candidate Arabic document
sim_max ← 0;
foreach document a'_j in C'_a do
    sim ← cos(e'_i, a'_j); // compare a'_j to all documents in C'_a
    if sim > sim_max then
        sim_max ← sim;
        a ← a_j;
return a, sim_max;
```

The alignment steps are described in Algorithm 3. The approach we propose aligns English and Arabic documents of a month-corpus. The process is repeated for each

month. To align an English document (e_i) and an Arabic document (a_j) of a month-corpus, first the English C_e month-corpus and the Arabic C_a month-corpus are provided to the Algorithm which, maps English e_i and Arabic a_i into the CL-LSI space. Then, the algorithm aligns each English document to the most relevant Arabic documents. The aligned articles with their similarity value (a_i, e_i, sim) are added to the list L, which is sorted later in descending order according to the similarity value.

The *align* procedure that is called in the algorithm takes the English document e'_i and the Arabic corpus C'_a. Then, the procedure computes the similarity between e'_i and all a'_j of C'_a and returns a'_j that has the highest similarity value.

The output of Algorithm 3 is a list of top-n most similar document pairs. If the aligned document pairs are related to the same story (checked manually), then they are considered to be correctly aligned. Otherwise, they are considered to be misaligned. The list of top-n most similar document pairs, is checked by hand to make sure that document pairs are correctly aligned. We remind that the objective of the experiment is to align news articles such that they are related to the same news story or event, and not to retrieve the articles sharing the same generic topic. This handwork is done on the top-15 article pairs retrieved from each month-corpus. The total number of documents to be validated is 360 article pairs. In the next section we present the results of our method.

4.4.2 Results

Experimental results are presented in Figures 4.6 and 4.7. The figures show the accuracy of alignment of the top-15 most similar documents of each month of the 24 month-corpus corresponding to the years 2012–2013. The accuracy of the alignment is defined as the number of cross-lingual articles, which are correctly aligned, divided by the total number of articles.

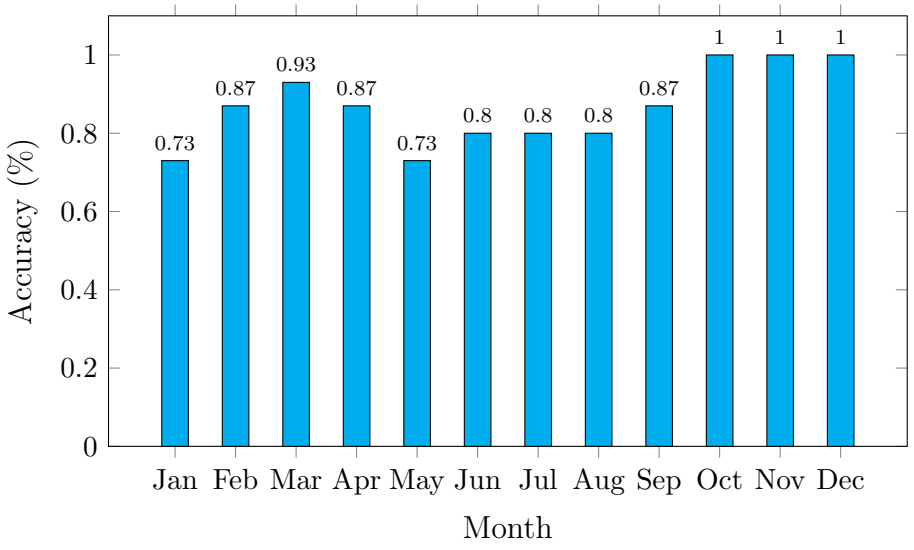

FIGURE 4.6: Accuracy of articles alignment for year 2012

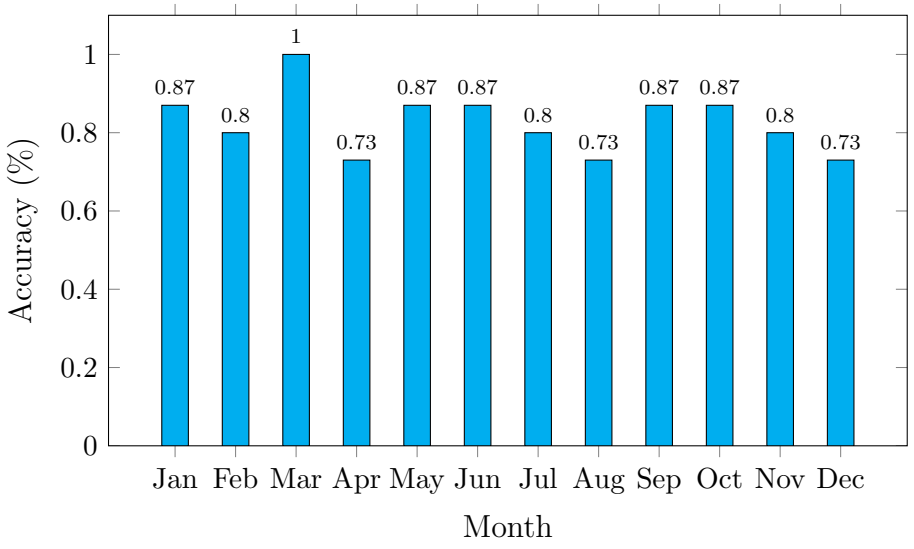

FIGURE 4.7: Accuracy of articles alignment for year 2013

The ranges of similarity values of the top-15 aligned articles for the years 2012 and 2013 are shown in Figures 4.8 and 4.9. The figures show the minimum and the maximum of similarity values for each month. For 2012, the maximum value is 0.86 and the minimum is 0.45. For 2013, the maximum value is 0.89 and the minimum is 0.26. It can be noted from the figures that the similarity ranges (minimum and maximum values) are close to each others for all months in 2012 and 2013 except for Jan., Feb., Apr. and May 2013. This is may be due to the nature of crawled articles

for each month, where the crawling tool may miss some articles in the crawling process.

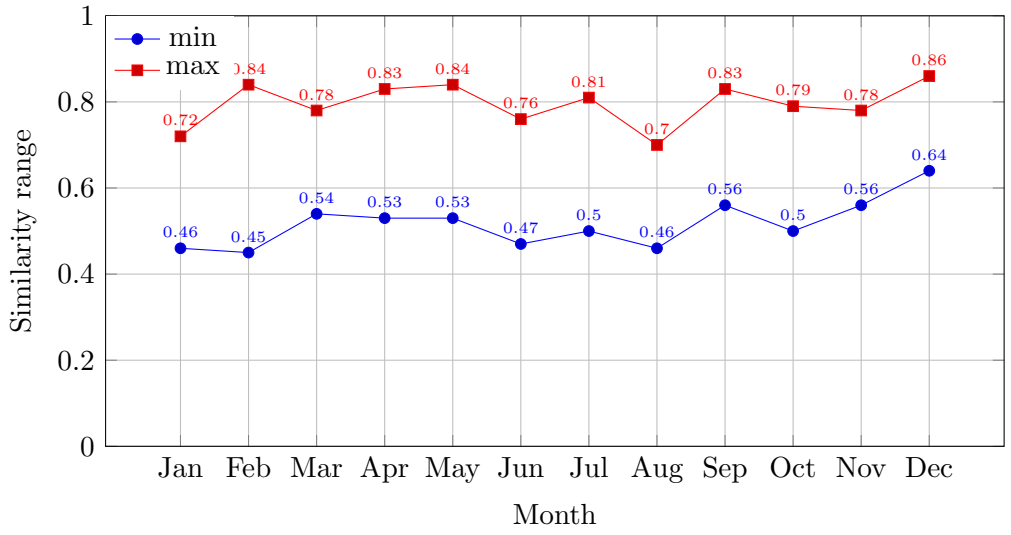

FIGURE 4.8: Similarity ranges of the top-15 similar documents of BBC-JSC of the year 2012

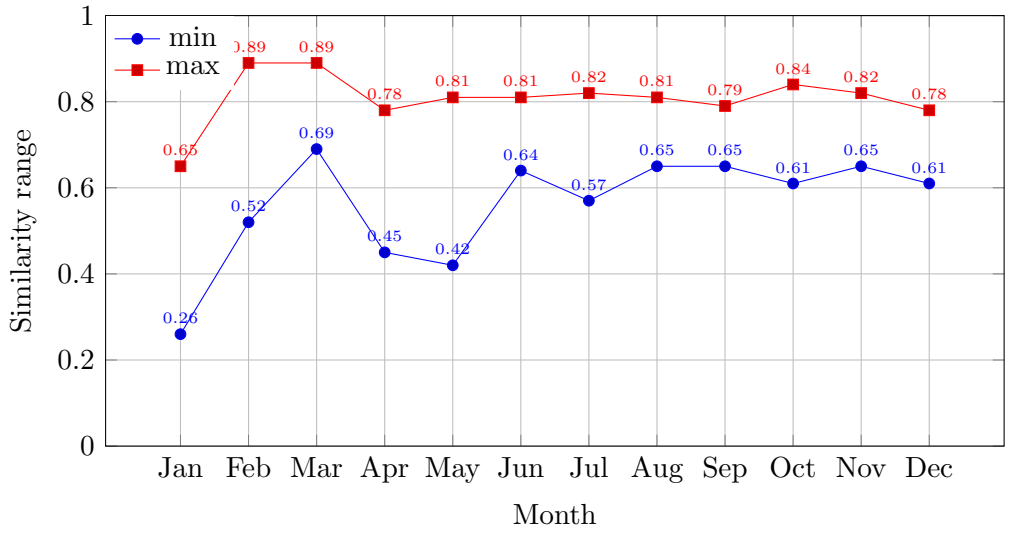

FIGURE 4.9: Similarity ranges of the top-15 similar documents of BBC-JSC of the year 2013

The accuracy of correctly aligned documents is 0.85 (305 out of 360). We carried out more investigations about misaligned articles during the validation process. We found that they are all related to the same topic domain, but they are not related to the same news story or event. The investigation reveals that some of these articles are

misaligned despite of high similarity. The reason is that they are related to the same event, but this event happened in different countries. For instance, one of misaligned news articles were related to the elections, but the English article was related to the elections in Bulgaria, while the Arabic article was related to elections in Pakistan. We conducted a search for "elections in Bulgaria" in our JSC collection, but we could not find any news article that is related to elections in Bulgaria. We also found that some of these stories are local news, which are covered only by either JSC or BBC. Besides that, it should also be noted that the crawling tool sometimes cannot crawl all the web pages from the website. This is why for some months, some news stories could not be found either in the BBC or JSC collections.

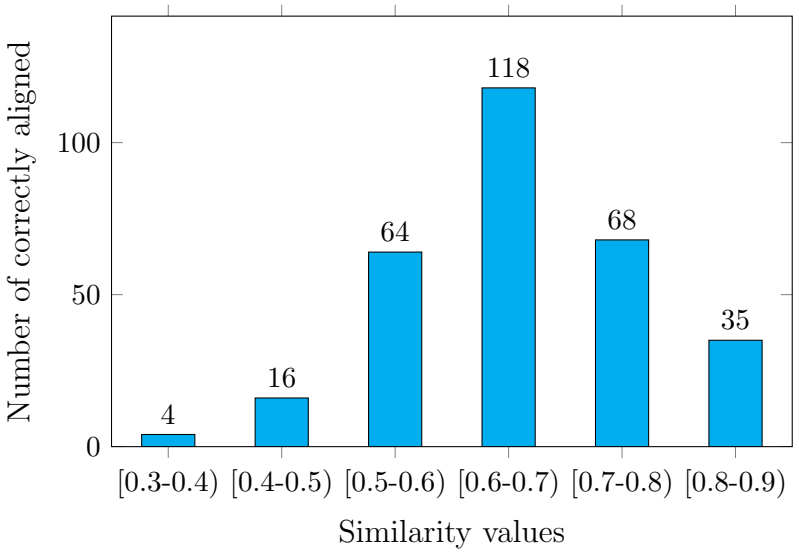

FIGURE 4.10: Similarity values vs. number of correctly aligned articles

Figure 4.10 shows the number of correctly aligned articles vs. their similarity values. The similarity values in this figure are divided into intervals. The number of correctly aligned articles increases as the similarity value increases, up to the interval $[0.6-0.7)$, then it decreases for higher similarity values. The interpretation might be as follows: when the similarity is low, the articles are mostly related to the same topic but not the same news story. As the similarity increases, the likelihood for the aligned articles to be related to the same news story increases up to a certain value, then it normally

decreases again. This is because it is unlikely to find news articles written by different news agencies, that have a high similarity value at the same time.

At the end, we got 305 documents of the BBC-JSC corpus, for which the alignments are checked by hand. These documents will be used in our experiment in Chapter 6 to compare cross-lingual news based on opinions and emotions. The goal is to investigate sentiment and emotion agreements of news articles that come from the Arab media with the ones that come from the English media.

4.5 Conclusion

In this chapter, we presented two cross-lingual similarity measures: one is based on bilingual dictionary and the second is based on the CL-LSI. The experiments showed that the performance of the CL-LSI method is better than the performance of the dictionary based method. Moreover, the advantages of the CL-LSI are that it overcomes the problem of vocabulary mismatch between queries and documents, and it does not need machine translation between source and target texts.

We further showed in this chapter that CL-LSI is able to not only align cross-lingual documents collected from the same source based on topics, but it can also align cross-lingual news articles collected from different sources based on events. The results showed that 85% of cross-lingual articles are correctly aligned. Also we demonstrated that CL-LSI method can be reliable to retrieve and align cross-lingual news.

BBC-JSC corpus, which is obtained in this chapter will be used later in our work to study the agreement of sentiments and emotions that are expressed in comparable news documents in Arab and English media.

In the next chapter, we propose a cross-lingual method to annotate English-Arabic parallel texts with sentiment and emotion labels. We use this annotated corpus to

build sentiment classifiers. We use these classifiers to automatically annotate English-Arabic comparable documents with sentiment labels.

Chapter 5

Cross-lingual Sentiment Annotation

5.1 Introduction

The aim of this chapter is to annotate parallel corpus with sentiment labels. The idea is to train a classifier on a corpus (A) of domain (X), and then to apply it on a parallel corpus (B) of domain (Y). Because B is parallel, the annotations can be transferred to other target languages. These generated resources are useful when there are no sentiment resources in the target languages, and when the resources of domain Y are not available in the source language.

As outlined in the introduction, most of state-of-the-art focus only on creating sentiment resources for low-resourced languages by building these resources from scratch or by adapting English resources using machine translation systems. On the other hand, many authors have argued whether machine translation preserves sentiments [Denecke, 2008, Ghorbel, 2012]. In this chapter, we present a new method of creating sentiment resources in multiple languages. We propose a cross-lingual annotation method that can be transferred across topic domains and across languages. We use

this method to create resources in English and Arabic languages in various domains. Next, we use these resources to build sentiment classifiers that can be used to automatically annotate English-Arabic articles. The advantage of this method is that it does not require a machine translation system. Sentiment resources in English-Arabic languages are not available; therefore, creating these resources is one of the contributions of this thesis.

Further in this chapter, we use language models to show that the cross-lingual annotation method can transfer the annotation reliably.

5.2 Cross-lingual Sentiment Annotation Method

Our proposed cross-lingual annotation method is described in Figure 5.1. Given a corpus (A) of domain (X) written in a source language and annotated with given labels. We split this corpus into training and testing subsets (step 1). The training part is used to train a classifier (step 2). This classifier can be preliminary validated by taking a subset from corpus (B) (step 3), and annotate it by a human annotator (step 4), then this subset is used to validate the classifier (step 5). Next, this classifier (source language and domain X) is used to automatically annotate the source texts of the corpus B (domain Y), and the labels are projected to the target texts (step 6). Projecting labels means to give the same label of the source text to target texts. Thus, the labels for the parallel corpus B are generated. The new contribution of this method is that it can generate annotated resources in multiple languages without need of machine translation systems.

The source and the target documents of the corpus (B) are used to train classifiers of domain (Y) (step 7). The source classifier can be further validated by the test set of the corpus (A) (step 8).

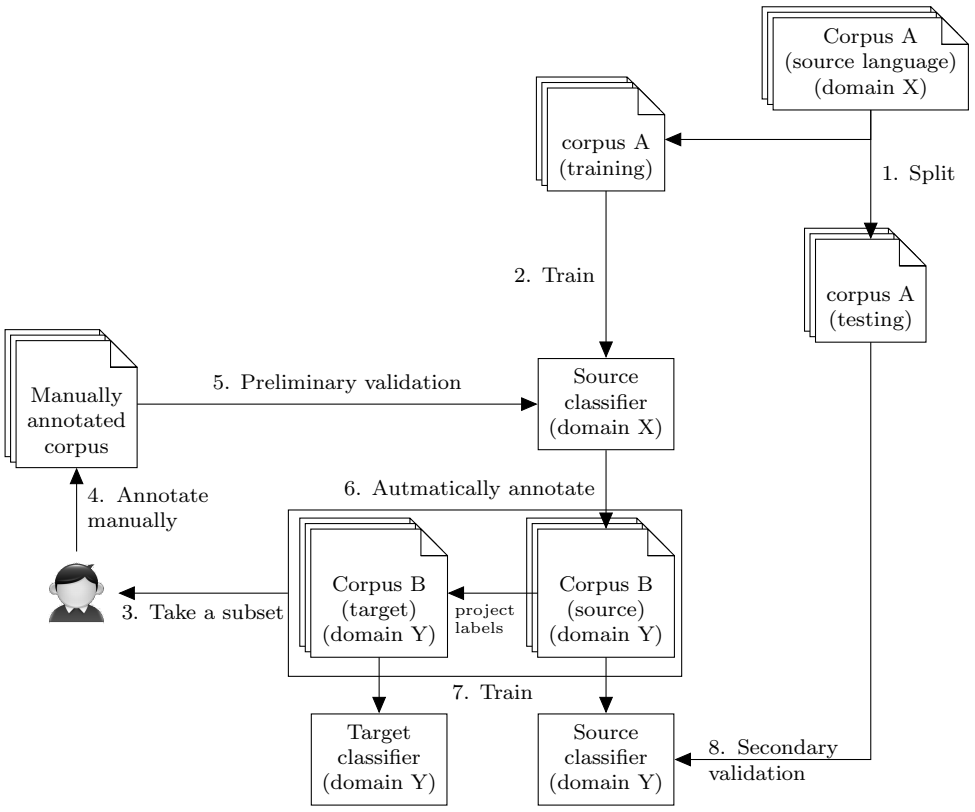

FIGURE 5.1: Cross-lingual annotation method

5.3 Experimental Setup

In our experiment, the initial corpus (A) is composed of movie reviews written in English. The corpus (B) is English-Arabic parallel corpus in various domains: newspapers, talks, United Nations resolutions. See Section 3.4 for details.

The corpus (A) is a collection of movie reviews written in English [Pang and Lee, 2004]. This corpus is pre-annotated with subjective and objective labels. The corpus is composed of 5,000 subjective and 5,000 objective sentences. The authors collected the subjective reviews from the Rotten Tomatoes website[1], and the objective reviews from IMDb plot summaries[2]. Rotten Tomatoes website is launched in 1998 and it is dedicated to film reviews. The website is widely known as a worldwide film review aggregator for important critics. The website also enable

[1] www.rottentomatoes.com
[2] www.imdb.com

users to review and discuss films. IMDb stands for Internet Movie Database, and launched in 1990. It is on-line database of information related to films, TV programs, actors, plot summaries, etc. A plot summary tells the main things that happened in the film. It is a brief description of the story of the film.

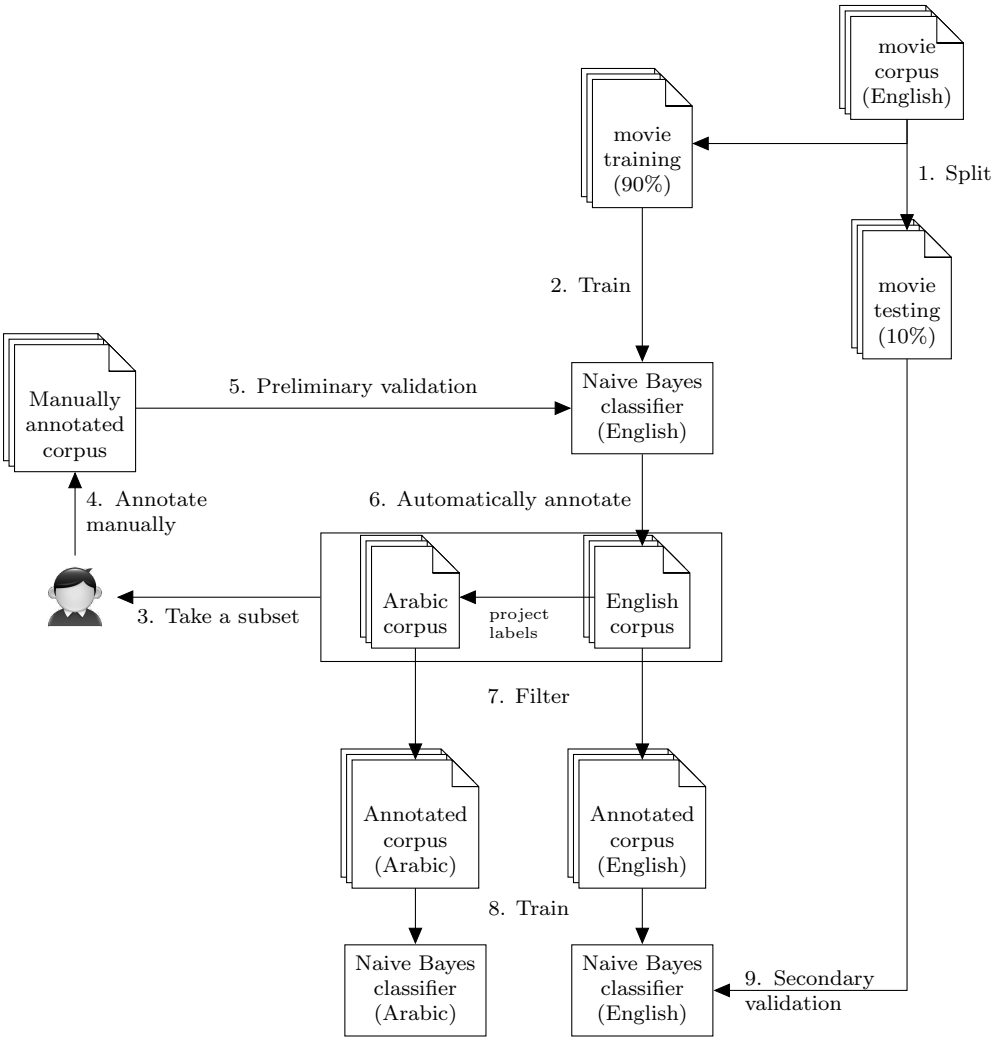

FIGURE 5.2: Experiment setup of the cross-lingual annotation to label Parallel corpus with sentiments

In our experiment, we use the method described in Figure 5.1. The details of our experimental setup of cross-lingual annotation is described in Figure 5.2. First, the movie corpus is split into training (90%) and testing (10%) (step 1). The training set of the movie corpus is used to train a Naive Bayes classifier (step 2). The classifier is trained on the combination of 1-gram, 2-gram and 3-gram features extracted from

each review of the training set of the movie corpus. We use n-gram features because it can represent words or sequence of words that can be used to express sentiments. The training instance of a review is presented as a set of n-gram features (f_k) associated with a class label (c_i) as follows $((f_1, f_2, f_3, \cdots, f_k), c_i)$.

The Naive Bayes classifier uses following Formula to assign a label to a given text.

$$classify(T) = \underset{c}{argmax} P(c) \prod_{k=1}^{n} P(f_k|c) \tag{5.1}$$

The text T is represented as a set of features f_k. These features are the most frequent n-grams that occur in the text. These are binary features (1 if the n-gram occurs in the document and 0 otherwise). Because features are generated from all possible n-grams, we discard n-gram features that occurred less than three times (low-frequency n-grams) in the training corpus, which results in keeping 10K out of 1.2M n-gram features of the movie corpus.

Table 5.1 shows the accuracy and the F-Measure ($F1$) of the movie classifier tested on the testing set of the movie corpus. $F1$ is the weighted average of the precision and recall, and it is computed as follows:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5.2}$$

where P is the precision and R is the recall. P is the ability of the classifier not to label as positive a sample that is negative, while R is the ability of the classifier to find all the positive samples. P and R are computed as follows:

$$P = \frac{tp}{tp + fp} \tag{5.3}$$

$$R = \frac{tp}{tp + fn} \tag{5.4}$$

where tp is true positive, fp is false positive, fp is false positive and fn is false negative.

The accuracy of classification is computed as follows:

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{5.5}$$

As shown in table 5.1, the accuracy of the movie classifier is 0.926, and subjective/objective $F1$ scores are 0.926/0.927 respectively.

TABLE 5.1: Accuracy and $F1$ scores of the movie classifier

Accuracy	Subjective $F1$	Objective $F1$
0.926	0.926	0.927

Some examples of the most informative subjective and objective n-gram features of the movie classifier are presented in Table 5.2. The n-gram feature is considered as more informative (or discriminative) if it appears in a larger number of texts associated with a single class label. As can be noted from the table, subjective n-gram features are expressions that the reviewer uses to express what he/she thinks about the movie, while objective n-gram features are used by the reviewer to describe the events of the movie. For instance, the 2-gram feature *"she is"* is usually used to tell what a women did in the film, i.e., it tell what happened in the story. This is why it is objective feature. In contrary, the 1-gram feature *"I"* is usually used to tell the personal opinions such as *"I think . . . "*, *"I am . . . "*. This is why the 1-gram feature *"I"* is subjective.

To ensure that the classifier built from movie domain can correctly identify the subjectivity of sentences of parallel corpus of another domain, which is composed of news and non-news domains, a preliminary evaluation is done on a subset of 330 sentences

TABLE 5.2: Examples of the most informative subjective and objective 1-gram, 2-gram and 3-gram features of the movie classifier

Subjective	Objective
I	order
me	decides
fans	discover
entertaining	led
interesting	kill
entertainment	to kill
even if	she is
but it	with her
is so	his family
if it	one day
it is not	the story of

selected randomly from the parallel corpora (step 3), and annotated manually by a human annotator (step 4). News sentences are selected from AFP, ANN, ASB, Medar, and NIST corpora, while non-news sentences are selected from TED, UN, and Tatoeba corpora. This annotated material is described in Table 5.3.

TABLE 5.3: The subset of parallel sentences manually annotated with sentiment labels by a human annotator (step 4 of Figure 5.2)

Corpus	Subjective	Objective
News	112	101
Non-News	60	57
Total	330	

Then, in step 5, the movie classifier is validated by the subset that is presented in Table 5.3. The corresponding results are described in Table 5.4. The classification accuracy of news sentences is 0.718, subjective $F1$ is 0.717, and objective $F1$ is 0.720, while the classification accuracy of non-news sentences is 0.658, subjective $F1$ is 0.667, and objective $F1$ is 0.649.

TABLE 5.4: Evaluation on the manually annotated parallel sentences (step 5 of Figure 5.2)

Corpus	Accuracy	Subjective $F1$	Objective $F1$
News	0.718	0.717	0.720
Non-News	0.658	0.667	0.649

As illustrated in Table 5.4, the classifier built from movie reviews can classify sentences from other domains: it can detect subjective and objective sentences from news corpus and from non-news corpus. In other words, the classifier does not model the genres of the corpus (news and non-news), but it can successfully distinguish between objectivity and subjectivity. We also note that this classifier (trained on movie corpus) performs better on news texts than non-news texts.

In step 6, the movie classifier is used to automatically annotate sentences of each corpus of the parallel corpora. Sentences are annotated only if the certainty of the annotation is above 0.8 (step 7 of Figure 5.2).

The statistics of annotations (after steps 7 and 8 of Figure 5.2) are shown in Table 5.5. The table presents the percentage of the annotated sentences with respect to the corpus. The table also shows the class distribution of the annotated sentences for each corpus (steps 7 and 8 of Figure 5.2).

Table 5.5: Parallel corpus annotation (steps 7 and 8 of Figure 5.2)

Corpus	Annotated sentences, %	Subjective, %	Objective, %
Newspapers			
AFP	90.6	9.4	90.6
ANN	89.9	18.6	81.4
ASB	91.7	17.8	82.2
Medar	89.6	25.8	74.2
NIST	79.4	20.6	79.4
United Nations resolutions			
UN	89.6	15.7	84.3
Talks			
TED	88.7	74.8	25.2
Other			
Tatoeba	86.4	59.3	40.7
Total	81.0	45.0	55.0

As shown in Table 5.5, 81% of all sentences are annotated (45% of them are subjective and 55% are objective). It can be noted from the table that AFP, ANN, ASB, Medar and NIST corpora are mostly objective. This is because these corpora correspond to

newspaper articles. Apparently UN resolutions are mostly objective. On the other hand, TED corpus is mostly subjective. It is not surprising, as TED talks contain unconstrained speech, and the authors frequently express their own opinions about the presented topics. Tatoeba corpus is mostly subjective too. This may be due to the nature of this corpus (sentences entered by contributors).

The source and the target documents of parallel annotated corpus are used to train Naive Bayes sentiment classifiers (step 8). Thus, we have English and Arabic sentiment classifiers.

The English classifier resulted from the previous step is validated on the test set of the movie corpus (step 9). The aim is to validate the classifier across domains. The results are presented in Table 5.6.

TABLE 5.6: Validating the English classifier by the of movie corpus (step 9 of Figure 5.2)

Accuracy	Subjective $F1$	Objective $F1$
0.79	0.81	0.74

It can be concluded from the results in Tables 5.4 and 5.6 that the movie classifier can classify texts of the parallel corpus (various domain), and the English classifier (resulted from step 8) can classify movie reviews. Comparing results of Table 5.1 and 5.6, it can be noticed that, as expected, the classification accuracy is better when the training and test from the same domain (movie reviews). On the other hand, the classification accuracy is less when the training is from news and non-news domain and the test is from movie reviews domain.

5.4 Statistical Language Models of Opinionated texts

Subjective and objective texts are very different in terms of writing style. Consequently, statistical language models built from subjective texts may differ from the ones built from objective texts.

A statistical language model represents the sequence of words by a probability distribution [Jurafsky and Martin, 2009]. It assigns higher probability for sequences that are "frequently observed" than for "rarely observed" or "ungrammatical" sequences. Opinionated language models can be useful for some applications where users express their sentiments, such as in human-computer interaction [Pang and Lee, 2008].

In this section we investigate the link between language models and cross-lingual annotation. We perform additional evaluation of the cross-lingual annotation method by using statistical language models. The test verifies if the cross-lingual annotation method can transfer the annotation across domains and across languages.

In this section, we conduct three experiments. The first experiment investigate whether subjective and objective texts are different in terms of writing style. In the second experiment we inspect whether language models built from the annotated corpus fit to movie text and vice-versa. In the third experiment, we investigate the subjectivity of the comparable corpus by two ways: by using annotation and by using language models.

5.4.1 Opinionated Language Models

In this experiment, we build Language Models (LM) from the annotated parallel corpus (the output of steps 3 and 4 of Figure 5.2 in Section 5.2). This corpus is split

into 90% for the training and 10% for the testing. Training corpus information is presented in Table 5.7.

TABLE 5.7: Word count and vocabulary size of the annotated parallel corpus

	Words	Vocabulary
English	4.4M	138K
Arabic	3.7M	264K

For each language, there is a subjective training corpus (S_{train}), an objective training corpus (O_{train}), a subjective test corpus (S_{test}) and an objective test corpus (O_{test}). A 3-gram language model is built on S_{train} (called SLM1), and another one on O_{train} (called OLM1). SRILM toolkit [Stolcke, 2002] is used to build language models, by using Kneser-Ney discounting method. The vocabulary for SLM1 and OLM1 is made up of the union of words of S_{train} and O_{train}, and composed at most of 138K English words and 264K Arabic words as presented in Table 5.7.

One measure for evaluating statistical language models is perplexity (PPL). Given a test word sequence ($W =< w_1 w_2 w_3 \cdots w_N >$), the perplexity of a language model is defined as follows:

$$PPL = exp^{-\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|w_1 w_2 w_3 \cdots w_{i-1})} \tag{5.6}$$

The lower the value of perplexity, the better the model is. Perplexity can determine how good the language model is, and can be used to compare the language models. The language model which has the lower perplexity is the better one [Jurafsky and Martin, 2009].

Figures 5.3, 5.4, 5.5 and 5.6 show the perplexity of subjective and objective language models (SLM1 and OLM1) on subjective S_{test} and objective O_{test} test sets for several vocabulary sizes. We investigate several vocabulary size because English and Arabic texts have different characteristics for their vocabulary. English words are isloated, while Arabic terms can be agglutinated. Each vocabulary is made up of the most

frequent words. Figures 5.3 and 5.5 show the test of language models on English subjective and objective texts, while Figures 5.4 and 5.6 show the test of language models on Arabic subjective and objective texts. The number on the right side of each curve presents the value of the last point.

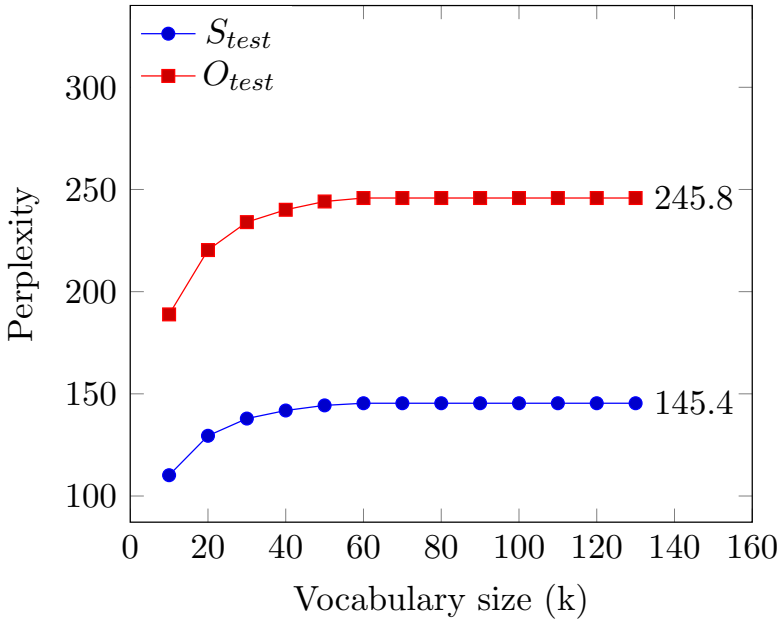

FIGURE 5.3: Perplexity of SLM1 on English subjective and objective test texts

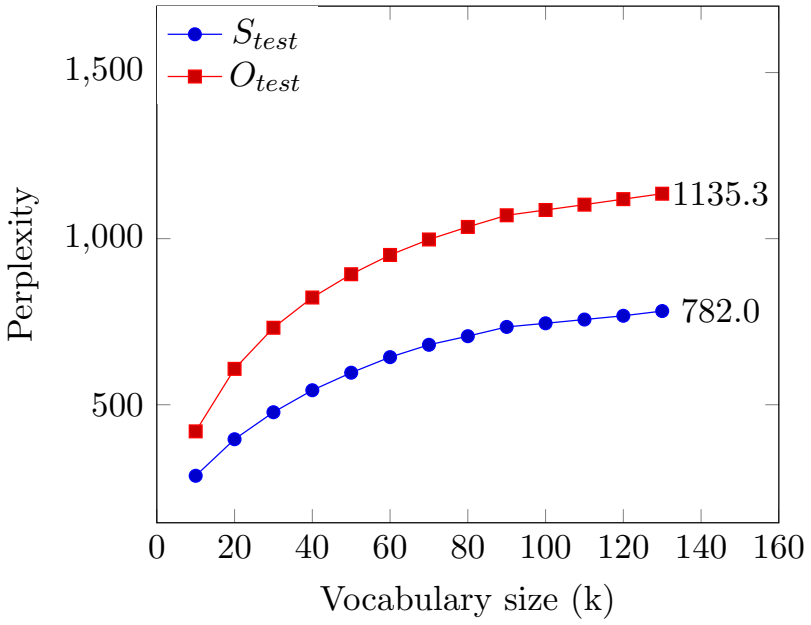

FIGURE 5.4: Perplexity of SLM1 on Arabic subjective and objective test texts

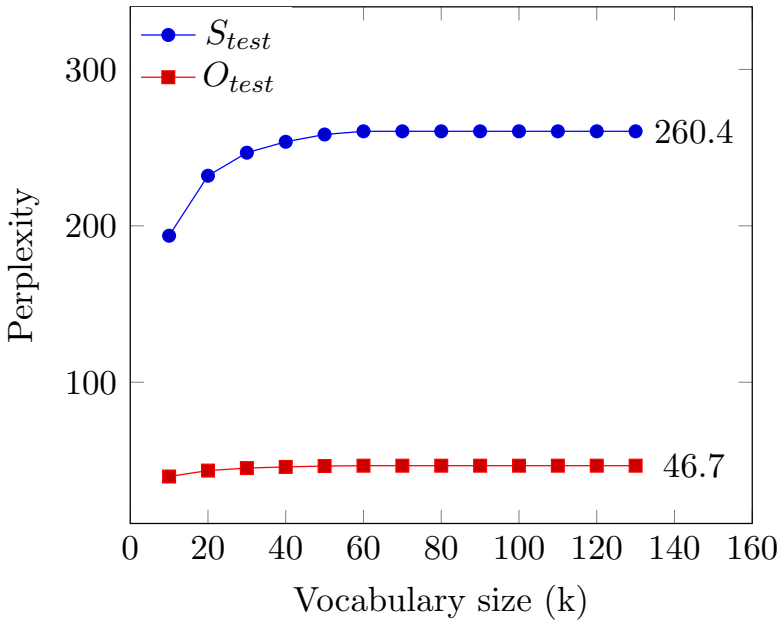

FIGURE 5.5: Perplexity of OLM1 on English subjective and objective test texts

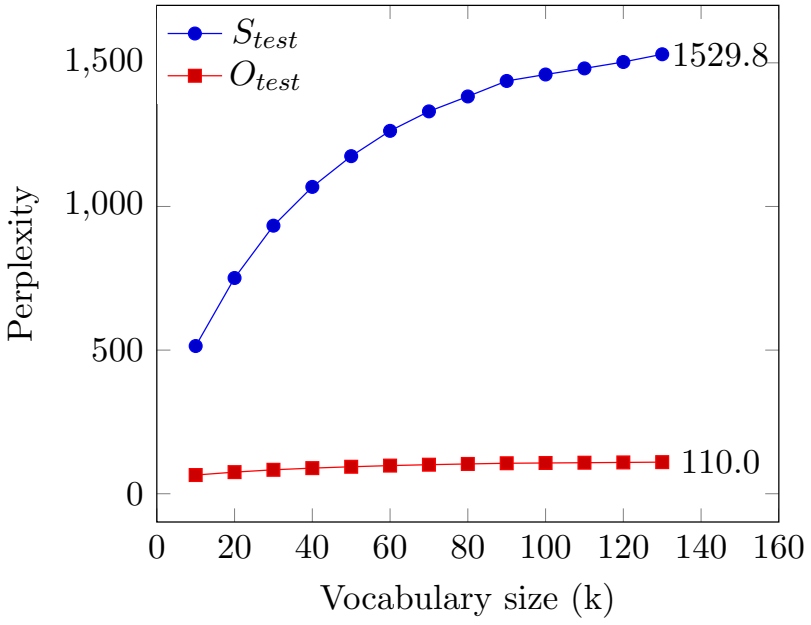

FIGURE 5.6: Perplexity of OLM1 on Arabic subjective and objective test texts

To have a closer look to the results, we present the perplexity results of the four figures for the vocabulary size (10K) in Table 5.8.

It can be noted that the perplexity for Arabic texts (Figures 5.4 and 5.6) is larger than

TABLE 5.8: Perplexity of SLM1 and OLM1 language models on English and Arabic subjective and objective test texts (vocabulary size = 10k)

	English		Arabic	
	SLM1	OLM1	SLM1	OLM1
S_{test}	110.17	193.68	286.07	514.19
O_{test}	188.80	39.73	419.80	65.10

the perplexity for English texts (Figures 5.3 and 5.5). According to a study about statistical language modeling for many languages conducted by [Meftouh et al., 2010], Arabic is one of the languages that has high perplexity, because it is agglutinative and highly inflected language. Second, it can be noted that SLM1 and OLM1 models fit better to subjective and objective texts respectively. Indeed, subjective (respectively objective) language models applied to objective (respectively subjective) test text lead to bad performance. It can be concluded from the last result that subjective and objective corpora are very different. Table 5.8 also confirms all the conclusions inferred above.

5.4.2 Testing Language Models across domains

In this experiment, we inspect the language models built using texts from different domain. We test SLM1 and OLM1 language models on the movie test corpus. In other words, we compute the perplexity language models built from the annotated parallel corpus (various domains) on the movie texts. In addition, we build subjective and objective language models from the movie corpus. These language models are called SLM2 for the subjective part, and OLM2 for the objective part. Then we test them on the annotated parallel corpus.

If subjective and objective language models fit better to subjective and objective test texts respectively, then we can conclude that the annotation method is good and reliable. We also aim in this experiment to inspect if the difference between objective and subjective texts is repeated across various topics (movie reviews vs. others).

To summarize, there are four language models in this experiment: SLM1 and OLM1 from the previous section, which are trained on subjective and objective parts of annotated parallel corpus, and SLM2 and OLM2, which are trained on subjective and objective training parts of movie corpus. All the language models are built using 10K most frequent words vocabulary.

Table 5.9: Perplexity of SLM1 and OLM1 language models

		Models	
Test corpus	Test type	SLM1	OLM1
Movie	Subjective	368.8	609.7
	Objective	507.8	442.2
Parallel	Subjective	110.2	193.7
	Objective	188.9	39.7

Table 5.10: Perplexity of SLM2 and OLM2 language models

		Models	
Test corpus	Test type	SLM2	OLM2
Movie	Subjective	354.3	688.9
	Objective	805.2	379.5
Parallel	Subjective	456.3	643.5
	Objective	900.1	687.9

Table 5.9 shows the perplexity of SLM1 and OLM1 on the movie and the parallel test corpora, while Table 5.10 shows the perplexity of SLM2 and OLM2 on the movie and the parallel test corpora. It can be noted from the results that the perplexities are higher when the models and the test corpora are from different domain topics (movie reviews vs. UN resolutions, newspapers, and talks). It can also be noted that the perplexity of OLM1 for objective part of parallel test corpus is low. This happens due to the UN corpus being mostly objective, in this corpus numerous sentences contain common parts (for example "taking note of the outcome of the..."); and these common parts are distributed between training and test. We also note that SLM1 and OLM1 fit better to subjective and objective parts of the movie corpus respectively, and SLM2 and OLM2 fit better to subjective and objective parts of the parallel corpus respectively. Moreover, it can be noted that subjective test corpus (respectively objective) does not fit to objective language model (respectively subjective).

It can be concluded that language models built from the parallel corpus can fit to the movie corpus. Namely, the distinction of subjective and objective text is stable across different topics. Additionally, it can be concluded that the annotation of the parallel corpus is good and reliable, because subjective (respectively objective) models built from parallel corpus have better perplexity on subjective (respectively objective) movie corpus. Finally, this experiment confirms the results of the previous section, which concludes that subjective and objective texts are distinct in terms of writing style.

5.4.3 Testing Language Models on Comparable Corpora

In this experiment, opinionated language models built using the parallel corpus (the output of steps 3 and 4 of Figure 5.2 in Section 5.2) are inspected on AFEWC and Euronews comparable corpora that are described in Section 3.3. The aim of this experiment is to explore the subjectivity of AFEWC and Euronews using two methods: annotating with the Naive Bayes classifier, and testing with language models. We also want to verify whether the results of annotation and language model tests accord with each other or not. We take a random subset of sentences of these corpora for our experiments (about 30K words).

Using the Naive Bayes classifier built in step 5 of Figure 5.2, we annotate each sentence in AFEWC and Euronews with subjective or objective labels. The distribution of labels for Euronews and AFEWC comparable corpora is presented in Table 5.11. The table shows the percentage of subjective and objective sentences in the subset of the comparable corpora. As can be seen from the table, for both comparable corpora, and for both languages, there are more objective sentences than the subjective ones. This is coherent because Wikipedia and news mostly tend to be objective.

To confirm this objectivity of the comparable corpora, we test SLM1 and OLM1 language models on AFEWC and Euronews comparable corpora. Figures 5.7 and

TABLE 5.11: Subjective and objective sentences distribution for the subsets of the comparable corpora (30K words)

Corpus	Subjective, %	Objective, %
AFEWC English	24	76
AFEWC Arabic	18	82
Euronews English	23	77
Euronews Arabic	11	89

5.8 show the perplexity of models evaluated on English and Arabic texts of AFEWC corpus respectively, and Figures 5.9 and 5.10 show the perplexity of models evaluated on English and Arabic texts of Euronews corpus respectively.

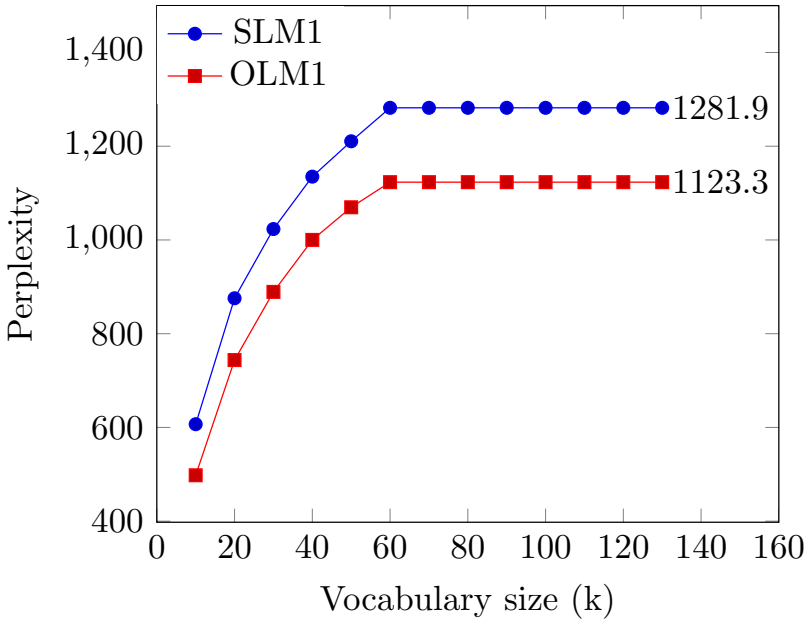

FIGURE 5.7: Perplexity of SLM1 and OLM1 models on English AFEWC corpus (30K words)

We first observe from the figures that the perplexity values in all tests are high. This is maybe because language models are built from different corpora and domains than the test texts. We also observe that OLM1 fits better to English and Arabic AFEWC and Euronews comparable corpora (has lower perplexity value) than SLM1.

The classification results (Table 5.11) and the test of language models (Figures 5.7, 5.8, 5.9 and 5.10) confirm the distinction between subjective and objective texts.

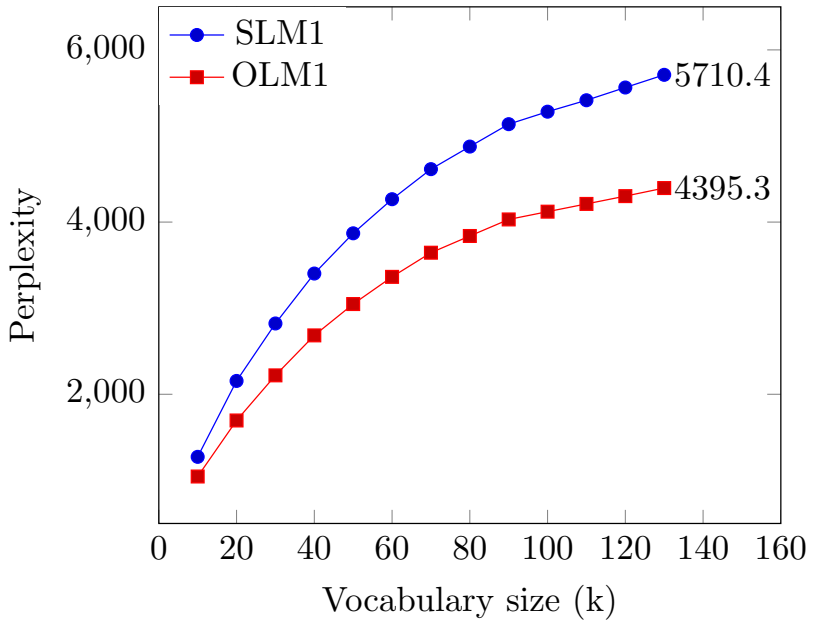

FIGURE 5.8: Perplexity of SLM1 and OLM1 models on Arabic AFEWC corpus (30K words)

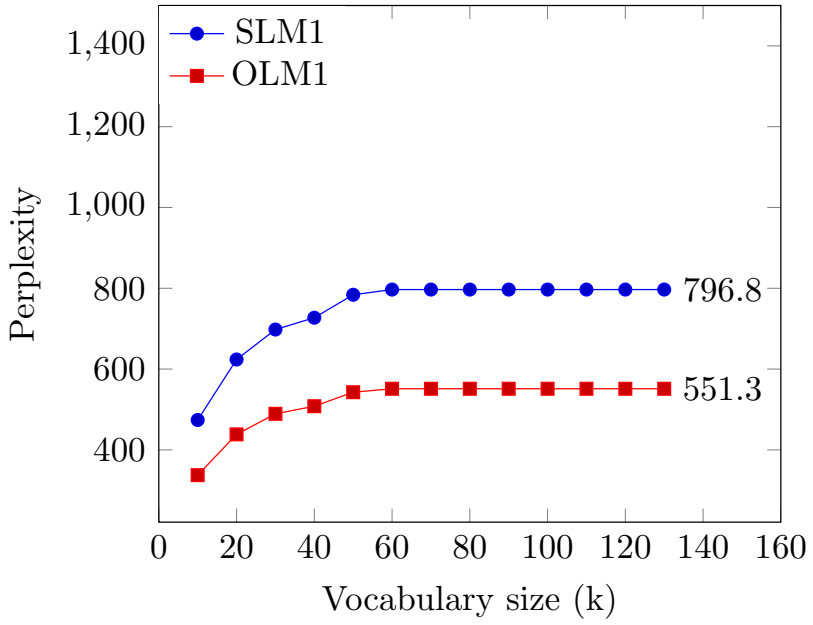

FIGURE 5.9: Perplexity of SLM1 and OLM1 models on English Euronews corpus (30K words)

The results achieved with the two methods accord with each other. This leads us to confirm that comparable corpora Euronews and AFEWC are mostly objective.

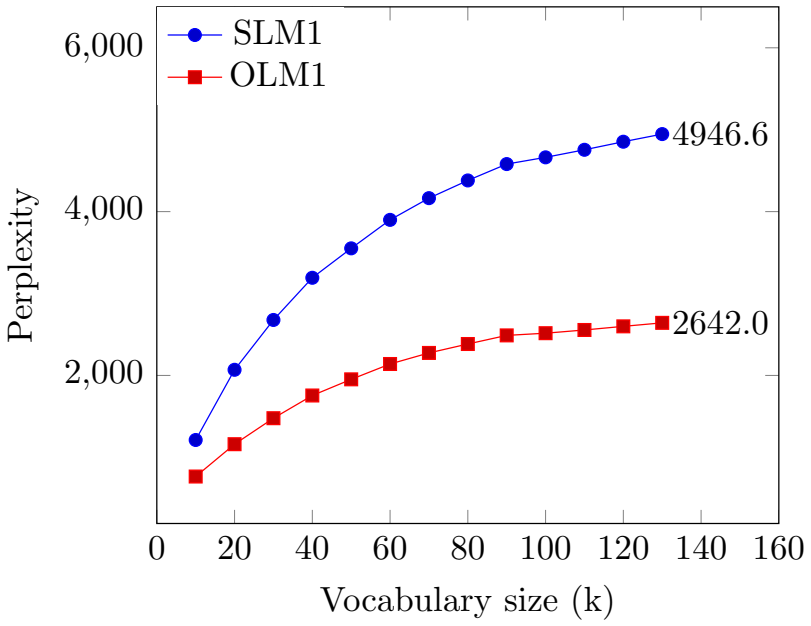

FIGURE 5.10: Perplexity of SLM1 and OLM1 models on Arabic Euronews corpus (30K words)

5.5 Conclusion

We proposed a method for cross-lingual annotation with subjective and objective labels. The cross-lingual annotation method provided annotated resources in various domains for English and Arabic languages. These resources can be used to build classifiers for annotating English and Arabic texts.

Using statistical language models, we showed that subjective and objective texts are statistically different in terms of writing style. We verified this by testing with models trained on corpora of different genres. Moreover, the results achieved with the Naive Bayes classifier and with the language models highlighted that our comparable corpora (AFEWC and Euronews: data extracted from Wikipedia and from news website in English and Arabic) are more objective than subjective.

In the next chapter, Naive Bayes classifiers will be used in to annotate English-Arabic news. Then, their annotation will be compared to each others.

Chapter 6

Comparing Sentiments and Emotions in Comparable Documents

6.1 Introduction

As outlined in Chapter 2 that describes the related work, the major interest in comparable texts consists in using them to extract parallel sentences. As for multilingual sentiment analysis, the major interest is to create sentiment resources for low-resourced languages, or to demonstrate whether machine translation is sufficient to capture sentiments. To our knowledge, comparing sentiments or emotions in comparable texts is not addressed in the literature. Comparing sentiments or emotions in comparable texts consists in inspecting the agreement of the expressed sentiments or emotions in the source and the target texts. The new contribution in this chapter is that we compare comparable documents in terms of sentiments and emotions.

Potential application is comparing customer reviews of a product, which are written in different languages. Another application is a tool for a journalist that is interested

in comparing and structuring comparable news in terms of sentiments and emotions. We focus in this chapter on English-Arabic comparable news documents.

For a given pair of English-Arabic documents annotated with sentiments and emotions, one can ask the following question: do these documents convey the same sentiments or emotions? The answer can be provided by a measure of the agreement of the annotations of English and Arabic comparable texts.

In this chapter, we study the agreement of annotations in three types of news corpora: a parallel news corpus, and two comparable corpora (Euronews and BBC-JSC). In the next sections, we first describe the used agreement measures. Then we present the method and the experimental results of comparing news documents in terms of sentiments and emotions.

6.2 Agreement Measures

The measure that we use in this work is called Inter-annotator agreement. It is defined as the degree of agreement or homogeneity between annotators [Artstein and Poesio, 2008]. The terms inter-annotator, inter-rater, and inter-coder are used interchangeably in the literature.

Normally, inter-annotator agreement is used in machine learning when there is no validation set to evaluate the accuracy of the classifier.

Inter-annotator agreement is based on the following assumption: if annotations are consistent, then annotators implicitly have similar understanding of the annotation guidelines, which describe how to make the annotations. Consequently, the annotation scheme is expected to perform consistently under these guidelines. In other words, the annotation scheme is reliable if the annotations are consistent (agree to each others).

In our work, the motivation for studying the agreement of comparable news documents is that if two persons disagree with each others when describing a news event, then it is difficult for a third person to understand what really happened, or maybe the third person think that both sources are interesting for him because they carry different view points. The idea is to automatically detect the agreement or disagreement between news stories written by different news agencies in different languages. The inspection of agreement of sentiments and emotions in comparable news documents can be seen as follows. Two news agencies A and B try to cover a news event. Each agency writes a document to describe and/or comment on this event. If A and B have the same perspective on the news event, then they will have some degree of agreement in their documents. Otherwise, they will diverge in terms of sentiments and emotions that are expressed in their documents.

We use inter-annotator agreement to inspect the agreement between sentiments or emotions in the annotated English and Arabic comparable documents. The parallel news corpus is expected to have a "perfect agreement". If near perfect agreement is achieved for parallel corpus, one can claim that the model is reliable, and it can be used to inspect the agreement between comparable documents. In our work, we develop an annotation scheme (a classifier), we prove that it is reliable for the parallel texts, and then we use this annotation scheme (the classifier) to inspect the agreement between the comparable texts.

Inter-annotator agreement can be calculated using statistical measures, such as Cohen's Kappa (k) [Cohen, 1960] or Krippendorff's alpha (α) [Krippendorff, 1980]. Unlike the simple percentage agreement calculation, statistical measures take into account the agreement that occurred by chance.

Cohen's Kappa (k) can be used for the cases where two annotators classify the data into two categories [Artstein and Poesio, 2008], and it is calculated as follows:

$$k = \frac{A_o - A_e}{1 - A_e} \tag{6.1}$$

where A_o is the observed agreement, and A_e is the expected agreement occurred by chance. A_e is calculated as follows:

$$A_e = p(l_1|c_1) \times p(l_2|c_1) + p(l_1|c_2) \times p(l_2|c_2) \tag{6.2}$$

where $p(l_i|c_j)$ is the probability that the coder j annotates data with the label i.

The range of k is between -1 and 1. Various researchers have interpreted the value of k measure in different ways as shown in Figure 6.1. For example, in [Landis and Koch, 1977], the authors divided the scale into detailed intervals, where the perfect agreement is assumed if $k > 0.8$, 0.6-0.8 as substantial agreement, 0.4-0.6 as moderate, 0.2-0.4 as fair, 0.0-0.2 as slight, and < 0 as no agreement.

In [Krippendorff, 1980] the authors discard the agreement if the k value is below 0.67. The k value between 0.67 and 0.8 is considered as tentative agreement, while k above 0.8 means good agreement.

In [Green, 1997, Fleiss et al., 2013] the authors consider that k value that is below 0.4 is low/poor agreement, from 0.4 to 0.75 is fair/good agreement, and above 0.75 is high/excellent agreement.

The interpretation of the scale depends on the task. Normally, the agreement for objective annotations is higher than for the subjective annotations [Fort, 2011]. Examples of objective annotation tasks are POS tagging, syntactic annotation, and phonetic transcription. Examples of subjective annotation tasks are lexical semantic (subjective interpretation), discourse annotation, and subjectivity analysis. According to [Fort, 2011], the agreement for objective tasks can be in the range 0.93-0.95, while for subjective tasks it is in the range 0.67-0.70.

Krippendorff's alpha (α) [Krippendorff, 1980] has been widely used in computational linguistics domain to measure the agreement of corpus annotation tasks [Artstein and Poesio, 2008]. α can be applied when the data are annotated by any

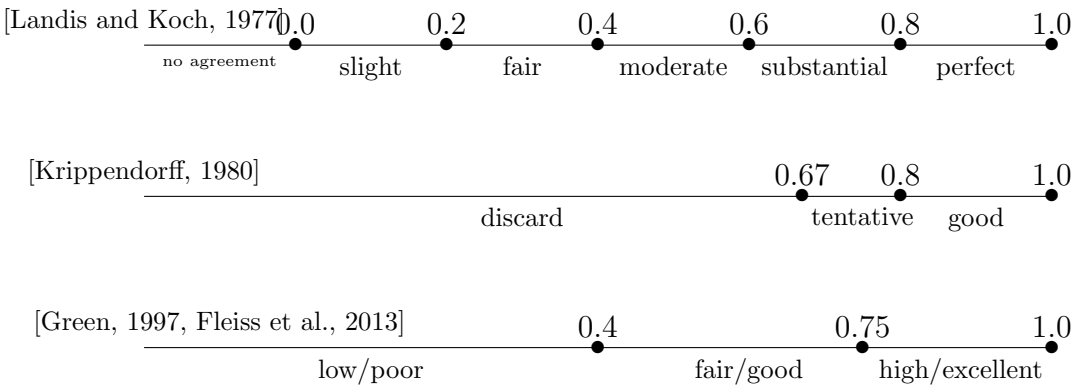

FIGURE 6.1: Kappa interpretation

number of annotators, and when this data belongs to any number of categories [Artstein and Poesio, 2008].

Krippendorff's α considers the variance (S) of annotations to inspect their agreement as follows:

$$\alpha = 1 - \frac{S_w^2}{S_t^2} \tag{6.3}$$

where S_t^2 is the *total variance*, and S_w^2 is the *within variance*. See [Krippendorff, 1980, Artstein and Poesio, 2008] for details.

The range of the value of α is between -1 and 1. The value can be considered as perfect agreement when α is close to 1, and as disagreement when $\alpha < 0$.

In this work, comparing sentiments and emotions in comparable documents is done as follows:

1. Automatically annotate comparable documents with sentiment and emotion labels.

2. Inspect the agreement between labels of comparable documents using k and α agreement measures.

In the next sections, we describe in detail how automatic annotation is done for comparable documents, then we present the experimental results.

Experiments are done on a parallel news corpus, and two comparable corpora (Euronews and BBC-JSC). The parallel-news and Euronews corpora are described in Chapter 3. We take a subset (10%) from parallel-news and Euronews to achieve our experiment. Thus, parallel-news is composed of 3.4K sentences and Euronews corpus is composed of 4.2K comparable. Regarding BBC-JSC corpus, it is composed of 305 comparable documents (see Section 4.4). In all corpora, each source text of these corpora is aligned to the target one. First we make annotation for the source and the target texts, and then we compare their annotations. The annotation is made at the sentence level for parallel-news corpus, and at the document level for Euronews and BBC-JSC corpora.

6.3 Comparing Sentiments in Comparable Documents

The automatic sentiments annotation is done using the classifiers, which are built in step 5 of Figure 5.2 (described in Section 5.2). Each pair of comparable documents is annotated with subjective and objective labels using these classifiers.

Table 6.1 shows the average pairwise subjective and objective agreement (k and α) calculated for each pair of documents from parallel-news, Euronews and BBC-JSC corpora.

Table 6.1: Subjective and objective agreement of news documents

Corpus	k	α
parallel-news	0.7642	0.7634
Euronews	0.2854	0.1866
BBC-JSC	0.0550	-0.1423

It can be noted from Table 6.1 that, as expected, the agreement between source and target texts of the parallel-news corpus can be considered as nearly perfect/good or high according to the interpretations presented in Figure 6.1. The parallel-news corpus also has the highest agreement scores among other corpora.

The Euronews corpus comes in the second place in terms of agreements, and BBC-JSC corpus has the lowest agreement among the other corpora.

For BBC-JSC corpus that we collected, the results reveal that BBC documents diverge from JSC documents in terms of subjectivity. These results also show that Euronews documents have higher degree of agreement compared to BBC-JSC. This is maybe because Euronews documents are mostly translations of each other as mentioned in Section 3.3, and they are written by the same news agency, while BBC-JSC documents are written by different news agency.

6.4 Comparing Emotions in Comparable Documents

To identify emotions in comparable documents, we use WordNet-Affect (WNA) emotion lexicon [Strapparava and Mihalcea, 2007], which is a subset of English WordNet. Each entry (synset) in this lexicon is annotated with one of six emotions (anger, disgust, fear, joy, sadness, and surprise), which are considered as the basic human emotions according to the psychological study conducted in [Ekman, 1992].

To be able to use this lexicon on Arabic texts, we manually translated it into Arabic. Table 6.2 describes the English and the Arabic lexicons, where $|syn|$ is the number of synsets (synonym words are grouped into sets and called synsets), and $|w|$ is the number of words associated with each emotion label. We use these lexicons to identify emotions in English-Arabic comparable documents.

TABLE 6.2: English-Arabic WordNet-Affect emotions lexicon

Emotion	$\|syn\|$	English $\|w\|$	Arabic $\|w\|$
anger	127	351	748
disgust	19	83	155
fear	82	221	425
joy	227	543	1156
sadness	123	259	522
surprise	28	94	201
Total	606	1551	3207

To identify whether an emotion is expressed in a text or not using WNA lexicon, the text is first converted into a list of bag-of-words (BOW). To improve the matching between the BOW of the text and words in the lexicons, lemmatization [Miller and Fellbaum, 1998] for English texts and light stemming [Larkey et al., 2007, Saad and Ashour, 2010] for Arabic texts are applied. Each term in the BOW list is checked with the emotion lexicon. If the term is matched with an emotion word in the lexicon, then the emotion label of that word is extracted from the lexicon, and the text is annotated with that label. In the end, the text is associated with six labels that indicate the presence or the absence of emotions that are expressed in the text.

Before using this lexicon to achieve our objective, we need to investigate its performance for identifying emotions. For this purpose, we select a 100 random sentences from news-parallel corpus. Each sentence is annotated with emotion labels by a human reader. The human annotator reads each sentence and check the emotions that are expressed in text, then annotates the text with the corresponding emotion labels. For instance, for the sentence *"Shock and deep sadness in the country due to the sudden death of President"*, then the human annotator annotate this sentences with *surprise* and *sadness* labels. These sentences are also annotated automatically using the WNA lexicon as described above. The automatic annotation is then compared to the human annotation. The evaluation is presented in Table 6.3. The table shows the accuracy, precision (P), recall (R), and the F-Measure ($F1$). The accuracy of emotion identification using WNA lexicon ranges between 0.85 and 1.0, while the

F-Measure ($F1$) ranges between 0.81 and 1.0. As can be seen in the results, the WNA lexicon can be reliable for emotion identification task.

TABLE 6.3: Evaluating WNA emotion lexicon

Emotion	accuracy	P	R	$F1$
anger	0.91	0.95	0.70	0.81
disgust	0.98	1.00	0.75	0.86
fear	0.97	1.00	0.80	0.89
joy	0.85	0.85	0.79	0.82
sadness	0.98	0.86	1.00	0.92
surprise	1.00	1.00	1.00	1.00

Using the WNA English and Arabic emotion lexicons, we automatically annotate each pair of documents in parallel-news, Euronews and BBC-JSC corpora. Then, the pairwise label agreement is calculated for each pair of documents.

Table 6.4 shows the average pairwise agreement between each emotion category in English-Arabic documents. As can be seen from the results, the parallel-news corpus has the highest agreement score among the other corpora for all emotions. This is expected since this corpus is parallel. In addition, the agreement degree can be considered as good or near perfect according to the scale interpretation that is described in Figure 6.1.

As shown in Table 6.4, agreement scores for Euronews are higher than for BBC-JSC for all emotions. This shows that emotions expressed in BBC diverge from the ones in JSC for the same news stories in our dataset. This may be because Euronews documents come from the same agency, while BBC-JSC documents are written by different agencies.

TABLE 6.4: Average pairwise agreement of each emotion category in the English-Arabic news documents

Corpus	k	α
anger		
parallel-news	0.8399	0.8400
Euronews	0.4237	0.4173
BBC-JSC	0.1812	0.1748
disgust		
parallel-news	0.9074	0.9074
Euronews	0.2856	0.2798
BBC-JSC	0.0232	0.0016
fear		
parallel-news	0.8074	0.8073
Euronews	0.3148	0.3073
BBC-JSC	0.1721	0.1662
joy		
parallel-news	0.7678	0.7672
Euronews	0.2940	0.2916
BBC-JSC	0.0904	0.0900
sadness		
parallel-news	0.8321	0.8321
Euronews	0.4050	0.4003
BBC-JSC	0.2113	0.1867
surprise		
parallel-news	0.8182	0.8181
Euronews	0.2066	0.2044
BBC-JSC	0.1039	0.0998

6.5 Conclusion

In this chapter, we compared English-Arabic comparable news documents based on sentiments and emotions. The comparison is done in two steps: annotate comparable documents automatically, and inspect the agreement of annotations between annotated documents.

We provided in this chapter a new method to study comparable documents by comparing sentiments and emotions in these document and inspect the agreement of

between sentiments and emotions in the comparable documents. We believe that our work will open new research directions for studying comparable corpora.

Results according to agreement scale demonstrated that parallel news documents have high agreement scores, while Euronews documents have some degree of agreement, and BBC-JSC documents have the lowest agreement among other corpora. The results revealed that, for our collected corpora, news documents coming from the same news agency have a higher agreement degree compared to the documents from different news agencies. That is, they diverge from each other in terms of sentiment and emotion agreement.

Chapter 7

Conclusion and Future Work

In this dissertation, we provided methods for collecting, retrieving and aligning comparable documents. We also proposed a method to compare sentiments and emotions in comparable documents.

First, we collected comparable corpora from Wikipedia and Euronews in Arabic, English and French languages. The corpora are aligned at the document level. Wikipedia corpus is made available publicly for research purposes.

Then, we investigated two cross-lingual similarity measures to retrieve and align English-Arabic comparable documents. The first measure is based on bilingual dictionary, and the second measure is based on Latent Semantics Indexing (LSI). The experiments on several corpora showed that the Cross-Lingual LSI (CL-LSI) measure outperformed the dictionary based measure. These conclusions are based on several corpus, parallel and comparable, and from different sources, different natures. The advantage of CL-LSI is that it needs neither bilingual dictionaries nor morphological analysis tools. Moreover, it overcomes the problem of vocabulary mismatch between documents.

Moreover, we also collected English-Arabic comparable news documents from local and foreign sources. The English documents are collected from the British Broadcast

Corporation (BBC), while the Arabic documents are collected from Aljazeera (JSC) news websites. These corpora are collected for analyzing sentiments and emotions in comparable documents from different sources.

After, we used CL-LSI similarity measure to align BBC-JSC news documents. The evaluation of the alignment has shown that CL-LSI is not only able to align cross-lingual documents at the topic level, but also it is able to do this at the event level.

To build the English-Arabic corpus annotated with subjective and objective labels, we proposed a novel cross-lingual annotation method. This method projects sentiment annotations from one topic domain to another and from one language to another. We also studied the subjectivity and the objectivity in the texts using statistical language modeling. We have shown in our experiments that subjective and objective texts are statistically different in terms of writing style. In addition, using the language models, we have showed that the cross-lingual annotation method projected the annotation successfully across domains and across languages. The advantage of the proposed cross-lingual annotation method is that it produces the annotated resources in multiple languages without need for a machine translation system. We used the resulting corpus to build sentiment classifiers. We use these classifiers to annotate comparable documents with sentiment labels.

Finally, we compared sentiments and emotions in comparable documents. The results are interesting, especially when the source and the target documents come from different sources. The comparison is done by inspecting the pairwise agreement of sentiments and emotions expressed in the source and the target comparable documents using statistical agreement measures.

We studied the agreement of sentiments in comparable news corpora (Euronews and BBC-JSC). The experiments show that BBC-JSC documents diverge from each other in terms of sentiments and emotions, while the Euronews corpus has a higher agreement, and most of the parallel-news corpus documents express the same sentiments and emotions.

Studying comparable documents is a promising research field. The contribution in this thesis is that we provided language independent methods to study comparable documents in different aspects (similarity, sentiments and emotions). We provided in this thesis language independent methods to align comparable articles (CL-LSI measure) and a novel method to annotate them with sentiment labels (the cross-lingual annotation method), and novel method to compare sentiments and emotions in comparable document pairs using statistical agreement measures.

In future, we will collect and study comparable corpora in various Arabic dialects. These corpora will be collected from social media. This means, alignment, annotation, and sentiment analysis tasks, that are done in this thesis, will be extended from English-Arabic languages to English-Arabic-Dialects. In addition, we will collect and study comparable news documents from other local and foreign source than BBC and JSC.

Moreover, we will use the CL-LSI method to align the cross-lingual texts at the sentence level (in this thesis it was applied at the document level). The parallel sentences will be extracted from Euronews and AFEWC comparable corpora. These parallel sentences can be useful to train machine translation systems. We will study the impact of different text preprocessing techniques, such as stemming, lemmatization and linguistic features, on the CL-LSI method. In addition, other semantic analysis methods, such as Probabilistic LSI (or LDA), will be investigated to improve the overall performance of the system.

In addition, more advanced methods for sentiment and emotion annotation will be developed. Recall that we used Naive Bayes classifiers trained on 3-gram features for sentiment annotation, and lexicon based method for emotion annotation. In the future, we will elaborate these systems to improve the overall annotation performance. Domain adaptation methods will be used to adapt domains in our cross-lingual annotation method. Other methods for statistical agreement, such as Point-wise Mutual

Information (PMI), will be used in the future to compare the agreement of sentiments and emotions between comparable documents.

A longterm future work also includes collecting and manually annotating sentiment and emotion resources written in Arabic language. These resources will help to build better sentiment and emotions classifiers for Arabic texts.

Appendix A

Samples of Comparable Documents

This section presents a sample of English-Arabic comparable documents.

A.1 Samples from Wikipedia Corpus

A.1.1 Wikipedia English Article

Napoleon

Napoleon Bonaparte (French: Napoléon Bonaparte, Corsican: Napoleone Buonaparte; 15 August 1769 – 5 May 1821) was a French military and political leader who rose to prominence during the latter stages of the French Revolution and its associated wars in Europe.

As Napoleon I, he was Emperor of the French from 1804 to 1814 and again in 1815. He implemented a wide array of liberal reforms across Europe, including the abolition of feudalism and the spread of religious toleration. His legal code in France, the Napoleonic Code, influenced numerous civil law jurisdictions worldwide. Napoleon is remembered for his role in leading France against a series of coalitions in the

Napoleonic Wars. He won the large majority of his battles and seized control of most of continental Europe in a quest for personal power and to spread the ideals of the French Revolution. Widely regarded as one of the greatest commanders in history, his campaigns are studied at military academies worldwide. He remains one of the most studied political and military leaders in all of history.

Napoleon was born in Corsica in a family of noble Italian ancestry that had settled in Corsica in the 16th century. He spoke French with a heavy Corsican accent. Well-educated, he rose to prominence under the French First Republic and led successful campaigns against the enemies of the French revolution who set up the First and Second Coalitions, most notably his campaigns in Italy.

He took power in a coup d'état in 1799 and installed himself as First Consul. In 1804 he made himself emperor of the French people. He fought a series of wars—the Napoleonic Wars—that involved complex coalitions for and against him. After a streak of victories, France secured a dominant position in continental Europe, and Napoleon maintained the French sphere of influence through the formation of extensive alliances and the elevation of friends and family members to rule other European countries as French vassal states.

The Peninsular War (1807–14) and the French invasion of Russia in 1812 marked major military failures. His Grande Armée was badly damaged and never fully recovered. In 1813, the Sixth Coalition defeated his forces at the Battle of Leipzig and his enemies invaded France. Napoleon was forced to abdicate and go in exile to the Italian island of Elba. In 1815 he escaped and returned to power, but he was finally defeated at the Battle of Waterloo in June 1815. He spent the last 6 years of his life in confinement by the British on the island of Saint Helena. An autopsy concluded he died of stomach cancer but there has been debate about the cause of his death, and some scholars have speculated he was a victim of arsenic poisoning.

A.1.2 Wikipedia French Article

Napoléon Ier

Napoléon Ier, né le 15 août 1769 à Ajaccio, et mort le 5 mai 1821 sur l'île Sainte-Hélène, est le premier empereur des Français, du 18 mai 1804 au 6 avril 1814 et du 20 mars 1815 au 22 juin 1815. Second enfant de Charles Bonaparte et Letitia Ramolino, Napoléon Bonaparte est un militaire, général dans les armées de la Première République française, née de la Révolution, commandant en chef de l'armée d'Italie puis de l'armée d'Orient. Il parvient au pouvoir en 1799 par le coup d'État du 18 brumaire et est Premier consul jusqu'au 2 août 1802, puis consul à vie jusqu'au 18 mai 1804, date à laquelle il est proclamé empereur par un sénatus-consulte suivi d'un plébiscite. Enfin il est sacré empereur en la cathédrale Notre-Dame de Paris le 2 décembre 1804 par le pape Pie VII.

En tant que général en chef et chef d'état, Napoléon tente de briser les coalitions montées et financées par le Royaume de Grande-Bretagne et qui rassemblent depuis 1792 les monarchies européennes contre la France et son régime né de la Révolution. Il conduit pour cela les armées françaises d'Italie au Nil et d'Autriche à la Prusse et à la Pologne : ses nombreuses et brillantes victoires (Arcole, Rivoli, Pyramides, Marengo, Austerlitz, Iéna, Friedland), dans des campagnes militaires rapides, disloquent les quatre premières coalitions. Les paix successives, qui mettent un terme à chacune de ces coalitions, renforcent la France et donnent à son chef, Napoléon, un degré de puissance jusqu'alors rarement égalé en Europe lors de la paix de Tilsit (1807).

Il réorganise et réforme durablement l'État et la société. Il porte le territoire français à son extension maximale avec 134 départements en 1812, transformant Rome, Hambourg, Barcelone ou Amsterdam en chefs-lieux de départements français. Il est aussi président de la République italienne de 1802 à 1805, puis roi d'Italie de 1805 à 1814, mais également médiateur de la Confédération suisse de 1803 à 1813 et protecteur de la Confédération du Rhin de 1806 à 1813. Ses victoires lui permettent d'annexer

à la France de vastes territoires et de gouverner la majeure partie de l'Europe continentale en plaçant les membres de sa famille sur les trônes de plusieurs royaumes : Joseph sur celui de Naples puis d'Espagne, Louis sur celui de Hollande, Jérôme sur celui de Westphalie et son beau-frère Joachim Murat à Naples. Il crée également un duché de Varsovie, sans oser restaurer formellement l'indépendance polonaise, et soumet temporairement à son influence des puissances vaincues telles que le Royaume de Prusse et l'Empire d'Autriche.

Objet, dès son vivant, d'une légende dorée comme d'une légende noire, il doit sa très grande notoriété à son habileté militaire, récompensée par de très nombreuses victoires, et à sa trajectoire politique étonnante, mais aussi à son régime despotique et très centralisé ainsi qu'à son ambition qui se traduit par des guerres d'agression très meurtrières (au Portugal, en Espagne et en Russie) avec des centaines de milliers de morts et blessés, militaires et civils pour l'ensemble de l'Europe. Il tente également de renforcer le régime colonial français d'Ancien Régime en outre-mer, en rétablissant en particulier l'esclavage en 1802 ce qui provoque la guerre de Saint-Domingue (1802-1803) et la perte définitive de cette colonie, tandis que les Britanniques s'assurent le contrôle de toutes les autres colonies entre 1803 et 1810. Cet ennemi britannique toujours invaincu s'obstinant à financer des coalitions de plus en plus générales, les Alliés finissent par remporter des succès décisifs en Espagne (bataille de Vitoria) et en Allemagne (bataille de Leipzig) en 1813. L'intransigeance de Napoléon devant ces sanglants revers lui fait perdre le soutien de pans entiers de la nation française tandis que ses anciens alliés ou vassaux se retournent contre lui. Amené à abdiquer en 1814 après la prise de Paris, capitale de l'Empire français, et à se retirer à l'île d'Elbe, il tente de reprendre le pouvoir en France lors de l'épisode des Cent-Jours en 1815. Capable de reconquérir son empire sans coup férir, il amène pourtant la France dans une impasse devant sa mise au ban de l'Europe, avec la lourde défaite de Waterloo qui met fin à l'Empire napoléonien et assure la restauration de la dynastie des Bourbons. Sa mort en exil à Sainte-Hélène sous la garde des Anglais, fait l'objet de nombreuses controverses.

Une tradition romantique fait de Napoléon l'archétype du grand homme appelé à bouleverser le monde. C'est ainsi que le comte de Las Cases, auteur du Mémorial de Sainte-Hélène tente de présenter Napoléon au parlement britannique dans une pétition rédigée en 1818. Élie Faure, dans son ouvrage Napoléon, qui a inspiré Abel Gance, le compare à un "prophète des temps modernes". D'autres auteurs, tel Victor Hugo, font du vaincu de Sainte-Hélène le "Prométhée moderne". L'ombre de "Napoléon le Grand" plane sur de nombreux ouvrages de Balzac, Stendhal, Musset, mais aussi de Dostoïevski, de Tolstoï et de bien d'autres encore. Par ailleurs, un courant politique français émerge au xixe siècle, le bonapartisme, se revendiquant de l'action et du mode de gouvernement de Napoléon.

A.1.3 Wikipedia Arabic Article

نابليون الأول

نابليون بونابرت الأول هو قائد عسكري وحاكم فرنسا وملك إيطاليا وإمبراطور الفرنسيين، عاش خلال أواخر القرن الثامن عشر وحتى أوائل عقد العشرينيات من القرن التاسع عشر. حكم فرنسا في أواخر القرن الثامن عشر بصفته قنصلاً عامًا، ثم بصفته إمبراطورًا في العقد الأول من القرن التاسع عشر، حيث كان لأعماله وتنظيماته تأثيرًا كبيرًا على السياسة الأوروبية.

وُلد نابليون في جزيرة كورسيكا لأبوين ينتميان لطبقة أرستقراطية تعود بجذورها إلى إحدى عائلات إيطاليا القديمة النبيلة. ألحقه والده آكارلو بونابرتْ، المعروف عند الفرنسيين باسم آشارل بونابرتْ بمدرسة بريان العسكرية. ثم التحق بعد ذلك بمدرسة سان سير العسكرية الشهيرة، وفي المدرستين أظهر تفوقًا باهرًا على رفاقه، ليس فقط في العلوم العسكرية وإنما أيضًا في الآداب والتاريخ والجغرافيا. وخلال دراسته اطلع على روائع كتّاب القرن الثامن عشر في فرنسا وجلّهم، حيث كانوا من أصحاب ودعاة المبادئ الحرة. فقد عرف عن كثب مؤلفات فولتير ومونتسكيو وروسو، الذي كان أكثرهم أثرًا في تفكير الضابط الشاب.

أنهى دروسه الحربية وتخرج في سنة ٥٨٧١م وعين برتبة ملازم أول في سلاح المدفعية التابع للجيش الفرنسي الملكي. وفي سنة ٥٩٧١ أعطى له فرصة الظهور، ليظهر براعته لأول مرة في باريس نفسها حين ساهم في تعضيد حكومة الإدارة وفي القضاء على المظاهرات التي قام بها الملكيون، تساعدهم العناصر المحافظة والرجعية. ثم عاد في سنة ٧٩٧١ ودعم هذه الحكومة ضد توجه أن تكون فرنسا ملكية دستورية فبات منذ هذا التأريخ السند الفعلي لها ولدستور سنة ٥٩٧١. بزغ نجم بونابرت خلال عهد الجمهورية الفرنسية الأولى، عندما عهدت إليه حكومة الإدارة بقيادة حملتين عسكريتين موجهتين ضد ائتلاف الدول المنقضة على فرنسا. وفي سنة ٩٩٧١، قام بعزل حكومة الإدارة وأنشأ بدلا منها حكومة مؤلفة من ٣ قناصل، وتقلد هو بنفسه منصب القنصل الأول؛ ثم سعى في إعلان نفسه إمبراطورا وتم له هذا بعد ٥ سنوات بإعلان من مجلس الشيوخ الفرنسي. خاضت الإمبراطورية الفرنسية نزاعات عدة خلال العقد الأول من القرن التاسع عشر، عرفت باسم الحروب النابليونية، ودخلت فيها جميع القوى العظمى في أوروبا. أحرزت فرنسا انتصارات باهرة في ذلك العهد، على جميع الدول التي قاتلتها، وجعلت لنفسها مركزا رئيسيا في أوروبا القارية، ومدت أصابعها في شؤون جميع الدول الأوروبية تقريبا، حيث قام بونابرت بتوسيع نطاق التدخل الفرنسي في المسائل السياسية الأوروبية عن طريق خلق تحالفات مع بعض الدول، وتنصيب بعض أقاربه وأصدقائه على عروش الدول الأخرى.

شكل الغزو الفرنسي لروسيا سنة ٢١٨١م نقطة تحول في حظوظ بونابرت، حيث أصيب الجيش الفرنسي خلال الحملة بأضرار وخسائر بشرية ومادية جسيمة، لم تمكن نابليون من النهوض به مرة أخرى بعد ذلك. وفي سنة ٣١٨١، هزمت قوات الائتلاف السادس الجيش الفرنسي في معركة الأمم؛ وفي السنة اللاحقة اجتاحت هذه القوات فرنسا ودخلت العاصمة باريس، وأجبرت نابليون على التنازل عن العرش، ونفوه إلى جزيرة ألبا. هرب بونابرت من منفاه بعد أقل من سنة، وعاد ليتربع على عرش فرنسا، وحاول مقاومة الحلفاء واستعادة مجده السابق، لكنهم هزموه شر هزيمة في معركة واترلو خلال شهر يونيو من عام ٥١٨١م. استسلم بونابرت بعد ذلك للبريطانيين، الذين نفوه إلى جزيرة القديسة هيلانة، المستعمرة البريطانية، حيث أمضى السنوات الست الأخيرة من حياته. أظهر تشريح جثة نابليون أن وفاته جاءت نتيجة لإصابته بسرطان المعدة، على الرغم من أن كثيرا من العلماء يقولون بأن الوفاة جاءت بسبب التسمم بالزرنيخ.

تدرس حملات نابليون العسكرية في العديد من المدارس الحربية حول العالم، وعلى الرغم من أن الآراء منقسمة حوله، حيث يراه معارضوه طاغية جبارا أعاد الحكومة لإمبراطورية ووزع المناصب والألقاب على أسرته ودخل مغامرات عسكرية دمرت الجيش، فإن محبيه يرونه رجل دولة وراعيا للحضارة، إذ ينسب إليه القانون المدني الفرنسي، المعروف باسم قانون نابليون، الذي وضع الأسس الإدارية والقضائية لمعظم دول أوروبا الغربية، والدول التي خضعت للاستعمار والانتداب الفرنسي في العصور اللاحقة.

A.2 Samples from Euro-news Corpus

A.2.1 Euro-news English Article

Transparency International warns 2/3rds of states "corrupt"

03-12-2013

The fairtrade watchdog Transparency International reports in its latest global roundup that two-thirds of the 177 countries it surveys are below average in the corruption stakes.

The report highlights continuing political pressure resulting in market distortions caused by bribery, cronyism, a lack of accountability and inadequate legal systems.

"We have two-thirds of all countries, 177 countries in total, where we can see that they score under 50 points, which means they are below the average, and corruption in those countries, two thirds of the 177, are still to be seen in a very critical situation," says Transparency's German head Edda Muller. Top of the class is Denmark, followed by New Zealand and Finland. The USA was only 19th, while among the major European economies France, (22), again lagged behind Germany, (12), and the UK, (14).

Greece's position improved, but it is still a lowly 80th, a potential turn-off for investors and the worst in Europe. Somalia, Afghanistan and North Korea tied for last place, unchanged from last year.

A.2.2 Euro-news French Article

Corruption : l'Espagne recule dans l'index de Transparency International

03-12-2013

Selon l'indice de la corruption dans le monde, calculé par l'ONG Transparency International et publié mardi, le Danemark est le pays le plus honnête, à égalité avec la Nouvelle Zélande. Ces deux pays jugé les moins corrompus en 2013, étaient déjà les deux meilleurs de la liste de 177 pays établie par l'ONG basée en Allemagne.

"Nous avons deux tiers des 177 pays qui sont sous les 50 points, explique Edda Muller, la présidente de Transparency International, ce qui veut dire qu'ils sont sous la note moyenne. La corruption dans ces pays : les deux tiers des 177, est dans une situation très critique".

Les trois premiers du classement sont le Danemark, la Nouvelle Zélande et la Finlande. La Grèce gagne quelques places mais reste 80ème et dernier pays européen.

La surprise c'est l'Espagne en proie aux affaires et qui passe de la trentième à la quarantième place. En règle générale les pays du Nord de l'Europe font figure d'élèves modèles. L'enquête de Transparency International n'est pas un classement selon le niveau de corruption mais elle mesure la perception de la corruption en raison du secret qui entoure les pratiques les moins avouables.

La France, 22ème, n'occupe que le 10ème rang en Europe. L'ONG considère que le bilan des lois votées en France en 2013 en matière de transparence et de lutte contre la corruption est "globalement positif" mais s'interroge sur leur mise en oeuvre effective.

A.2.3 Euro-news Arabic Article

الشفافية الدولية : اليونان أسوأ الأوروبيين والإمارات أفضل العرب

٠٣-١٢-٢٠١٣

منظمة الشفافية الدولية قالت إن ثلثي البلدان المدرجة في مؤشر مدركات الفساد للعام ألفين وثلاثة عشر هي في حالة حرجة للغاية.

المنظمة التي تضم مائة وسبعة وسبعين دولة، صنفت معظم البلدان دون درجة الخمسين، حيث تعكس الصفر أعلى مستويات الفساد، والمائة أرفع مقام في النزاهة.

إدا مولر مديرة المنظمة :

ثلثا البلدان المشمولة في هذه القائمة هي تحت خمسين نقطة، ما يعني أنها تحت المستوى المتوسط، والفساد في تلك البلدان يعتبر في حالة حرجة للغاية.

الدانمارك تصدرت قمة الدول الشفافة في العالم تليها نيوزيلندا ثم فنلندا، بينما تردت اليونان إلى المرتبة الثمانين لتكون الأسوأ في دول الإتحاد الأوروبي، وحلت الصومال في ذيل القائمة.

عربياً تصدرت دولة الإمارات العربية المتحدة حيث جاءت في المرتبة السادسة والعشرين، تلتها دولة قطر في المرتبة الثامنة والعشرين.

بين الاقتصادات الكبرى، لم يتغير وضع الولايات المتحدة التي حلت في المرتبة التاسعة عشرة، والصين في المرتبة الثمانين ، بينما تراجعت اليابان نقطة واحدة إلى المرتبة الثامنة عشرة.

A.2.4 English Translation of Euro-news Arabic Article

Transparency International: Greece is the worst of Europeans and Emirates is the best of Arabs.

03-12-2013

Transparency International said that two-thirds of the countries included in the Corruption Perception Index for the year two thousand and thirteen are in a very critical condition.

The organization, which includes one hundred and seventy-seven countries, classified most countries below fifty, where zero reflects the highest levels of corruption, and one hundred is the highest status of integrity.

Edda Muller, director of the organization:

Two-thirds of the countries included in this list are below fifty points, which means they are below the average, and corruption in those countries is in a very critical condition.

Denmark is the top transparent country in the world, followed by New Zealand and Finland, while Greece has degraded to the rank eighty to be the worst in the European Union countries, and Somalia is placed in the bottom of the list.

As for the Arab countries, United Arab Emirates is top country in rank twenty-sixth, followed by Qatar in the rank twenty-eighth.

Among major economies, nothing has changed where the U.S. ranked in nineteen position, and China ranked eighty, while Japan degraded one point to ranked eighteen.

A.3 Samples from BBC-JSC corpus

A.3.1 BBC-JSC English Article

Guantanamo Taliban inmates 'agree to Qatar transfer'

10 March 2012

Five senior Taliban fighters held at Guantanamo Bay have agreed to be moved to custody in Qatar as part of a peace plan, Afghan government officials say.

The US administration has not approved the transfer but is considering it as an incentive for the militants to enter negotiations in Afghanistan.

None of the five inmates is accused of directly killing Americans. They would be reunited with their families in Qatar, which is playing an increasing role in negotiations.

They reportedly agreed to the transfer when they met Afghan government officials who visited the US prison this week on a mission from President Hamid Karzai. Correspondents point out that the visit to the US prison on Cuba would not have been possible without US approval.

Aside from the aim of ending the war in Afghanistan, the prospect of transferring Taliban detainees proves once again that, more than three years after he promised to close it, Guantanamo Bay remains a thorn in President Obama's side, the BBC's Jonathan Blake reports from Washington.

If the president pursues this strategy, though, he will need support from wary politicians in Congress, our correspondent says.

Many there see a transfer of what they call the most dangerous inmates at Guantanamo as a step too far, he adds.

'No decision'

Ibrahim Spinzada, a senior President Karzai aide, visited Guantanamo on Monday, according to Reuters.

Both he and Shahida Abdali, a senior Afghan security official, also visited the US this week, the White House said without giving details.

"We are hopeful this will be a positive step towards peace efforts," Mr Karzai's spokesman, Aimal Faizi, told Reuters news agency.

Asked about the transfer plan, White House spokeswoman Caitlin Hayden said: "The United States has not decided to transfer any Taliban officials from Guantanamo Bay. "We are not in a position to discuss ongoing deliberations or individual detainees, but our goal of closing Guantanamo is well established and widely understood."

The spokeswoman pointed out that any decision on transfers would be undertaken in accordance with US law and in consultation with Congress.

US officials are hoping that President Obama can announce the establishment of fully fledged political talks between the Karzai government and the Taliban at a Nato summit in May. The international peacekeeping mission in Afghanistan is due to finish at the end of 2014. A total of 171 detainees were still being held at Guantanamo as of this month, Reuters reports.

In January, the Taliban announced it was opening a political office in Qatar. Afghan Foreign Minister Zalmay Rasool is expected to visit the Gulf state this month for talks with government officials on reconciliation with the Taliban.

A.3.2 BBC-JSC Arabic Article

قادة طالبان بغوانتانامو ينقلون إلى قطر

١٠ مارس ٢٠١٢

وافق خمسة من كبار معتقلي حركة طالبان في سجن غوانتانامو الأميركي على تسليمهم لدولة قطر في حال الإفراج عنهم. وقالت مصادر أفغانية رسمية إن خطوة كهذه من شأنها أن تسهم في تقدم التفاوض بين الحركة وحكومة الرئيس الأفغاني حامد كرزاي برعاية الولايات المتحدة. يأتي بينما أعلن عن زيارة وزير خارجية أفغانستان إلى قطر عقب الإعلان عن التوجه نحو افتتاح مكتب لطالبان في الدوحة.

وقال أيمل فيضي الناطق باسم الرئيس الأفغاني إن خمسة من كبار قادة طالبان المعتقلين في سجن غوانتانامو وافقوا على نقلهم إلى العاصمة القطرية فور الإفراج عنهم.

وأكد الناطق أن خطوة كهذه ستسهم في دفع عملية السلام الأفغانية إلى الأمام، في إشارة إلى التفاوض الذي تشجع عليه واشنطن بين طالبان وحكومة كرزاي. وأضاف فيضي آنأمل أن هذا سيكون خطوة إيجابية باتجاه جهود السلامْ.

وذكرت وكالة رويترز أن من بين السجناء الخمسة الذين قد ينقلون الى قطر محمد فضل، وهو معتقل آشديد الخطورةْ قيل إنه مسؤول عن قتل الآلاف من الشيعة الأفغان في الفترة بين عامي ١٩٩٨ و ٢٠٠١.

ومن بينهم أيضا نور الله نوري وهو قائد عسكري كبير سابق، وعبد الحق واثق نائب وزير الاستخبارات الأسبق، وخير الله خير خوا وهو وزير داخلية سابق.

وفد يزور

وكان وفد من الحكومة الأفغانية قد زار المعتقل العسكري الأميركي في خليج غوانتانامو الأسبوع الماضي للحصول على موافقة خمسة معتقلين من طالبان قد ينقلون قريبا إلى قطر. وقالت مصادر على دراية بهذا الأمر إن الوفد الذي زار مركز الاعتقال بكوبا يوم الاثنين الماضي، ضم إبراهيم سبين زاده كبير مستشاري كرزاي للسياسة الخارجية.

وقالت مصادر حكومية في العاصمة كابل إن سبين زاده ومسؤول الأمن الأفغاني البارز شهيد عبدلي زارا الولايات المتحدة الأسبوع الماضي. وقال البيت الأبيض إن وزارة الدفاع الأميركية رفضت التعقيب على زيارة غوانتانامو.

وطالبت حكومة كرزاي بأن يعطي الأعضاء الخمسة البارزون في حكومة طالبان السابقة موافقتهم قبل نقلهم إلى قطر.

ويأتي نقل هؤلاء المعتقلين ضمن سلسلة من إجراءات تستهدف إظهار حسن النية، بحيث إذا تمكن دبلوماسيون أميركيون من التغلب على العقبات المتبقية، تبدأ أول مفاوضات سياسية جوهرية بشأن الصراع في أفغانستان منذ الإطاحة بحكومة طالبان عام ٢٠٠١ في الغزو الذي قادته الولايات المتحدة.

وبعد عام من الكشف عن مبادرة حكومة الرئيس باراك أوباما للسلام، فإنها ربما تطرح قريبا للولايات المتحدة فرصة للوساطة في نهاية للصراع الذي بدأ ردا على هجمات ١١ سبتمبر ٢٠٠١ واستمر لمدة عقد من الزمن بتكاليف مالية وبشرية باهظة.

ويطرح متمردو طالبان شرطا مسبقا لكل تفاوض الإفراجَ عن عناصرهم المحتجزين في سجن غوانتانامو الأميركي في كوبا. لكن واشنطن تريد منهم التخلي عن العنف قبل أن تجري أي مفاوضات معهم.

زيارة وترتيبات وأوضح جنان موسى زي المتحدث باسم وزير الخارجية الأفغاني أن الحكومة الأفغانية وافقت على إعادة المعتقلين الأفغان في غوانتانامو إلى عائلاتهم إذا كان هذا ما يرغبون فيهْ.

ولا يزال نحو عشرين أفغانيا بينهم خمسة من قيادات طالبان في حكومتهم السابقة (١٩٩٦-٢٠٠١)- أسرى في غوانتانامو.

وكانت الحكومة الأفغانية قد أعلنت في وقت سابق اليوم أن وزير الخارجية زلماي رسول سيزور قطر خلال أقل من عشرة أيام ليبحث مع المسؤولين الحكوميين قضية المصالحة مع حركة طالبان.

وأعلنت طالبان في يناير الثاني الماضي أنها ستفتح مكتبا سياسيا في قطر، مشيرة إلى أنها ربما تكون مستعدة للانخراط في مفاوضات يرجح أن تمنحها مناصب بالحكومة الأفغانية أو سيطرة رسمية على معظم معقلها التاريخي في جنوب أفغانستان.

وقال محللون أفغان إن الزيارة تأتي في إطار مساعي السلام التي تبذلها الحكومة الأفغانية، وإنها ستناقش موضوع المكتب السياسي.

وكانت الولايات المتحدة وأفغانستان قد وافقتا على فتح المكتب السياسي في قطر رغم أن حكومة كرزاي ترددت في البداية، وكانت تفضل إقامته في السعودية أو تركيا. وقد سحبت أفغانستان سفيرها في قطر بعدما شكّت في استبعادها من جهود السلام، غير أنها عادت لتليين موقفها بعد ذلك. وتسعى الإدارة الأميركية إلى دعم عملية السلام قبل سحب القوات الأجنبية من أفغانستان عام ٢٠١٤.

A.3.3 English Translation of BBC-JSC Arabic Article

Taliban leaders in Guantanamo transferred to Qatar

10 March 2012

Five inmates of Taliban senior leaders, held at Guantanamo in the U.S., agreed to be transferred to Qatar in case they are released. An official Afghan source said that such a move would contribute to progress the negotiations between Taliban and the Afghan government of the president Hamid Karzai, under sponsorship of the United States. Meanwhile, the minister of Foreign Affairs of Afghanistan announced a visit to Qatar after the announcement of the trend towards the opening of a Taliban bureau in Doha.

He Oiml Faidi, the spokesman of Afghan president, said that five senior leaders of the Taliban detainees held at Guantanamo Bay have agreed to be transferred to the capital of Qatari immediately after their release.

The spokesman confirmed that such a move will contribute to push the Afghan peace process forward, in reference to the negotiation encouraged by Washington between the Taliban and the Karzai government. Faidi added: "We hope that this will be a positive step towards peace efforts."

Reuters news agency reported that one of the five prisoners, who had transferred to Qatar, is Mohamed Fadl, a "very dangerous" detainee. It has been said that he is responsible for the killing of thousands of Afghan Shiites in the period between 1998 and 2001.

Also among these prisoners: Nour Allah Nouris who is a former military senior commander, and Abdul Haq Wathiq Naeb El-haq who is a former intelligence minister, and Khairallah Khair Khoa who is a former interior minister.

Delegation to visit

A delegation from the Afghan government has visited the U.S. military prison at Guantanamo Bay last week for approval of five Taliban detainees may soon be transferred to Qatar. Knowledgeable Sources reported that the delegation who visited the detention center in Cuba on Monday, included Ibrahim Spin Zadeh, senior adviser to Karzai's foreign policy.

Government sources said in the capital Kabul that Spin Zadeh and a senior Afghan security responsible, Shahid Abdali, visited the United States last week. The White House said that the U.S. Department of Defence refused to comment on the visit Guantanamo.

Karzai's government has demanded the five prominent prominent in the former Taliban government to give their agreement before being transferred to Qatar.

The transfer of these detainees comes with a series of actions aim to show the goodwill, so that if U.S. diplomats managed to overcome the remaining obstacles, the first substantial political negotiations starts on the conflict in Afghanistan since the overthrow of the Taliban government in 2001 in the invasion that lead by the United States.

A year later to the revealing of the initiative of the government of President Barack Obama for peace, It is probably that the United States will provide an opportunity to be a broker at end to the conflict, which began in response to the attacks of September 11, and lasted for a decade, with huge cost in terms of finical and human.

Taliban rebels pose a prerequisite for each negotiate on releasing their cadres held in the U.S. prison at Guantanamo Bay in Cuba. But Washington ask them to give up violence before conducting any negotiations with them.

Visit arrangements

Jinan Moussa Zi, the spokesman of Afghan Foreign Minister, stated that "the Afghan government has agreed to rejoin the Afghan detainees at Guantanamo with their families if that is what they want."

There are about twenty Afghans - including five leaders of the former Taliban government in (1996-2001) - still prisoners at Guantanamo.

The Afghan government has announced earlier today that Foreign Minister Zalmai Rassoul will visit Qatar in less than ten days to discuss with government officials the issue of reconciliation with the Taliban.

Taliban announced last January that it would open a political bureau in Qatar, pointing that it might be willing to engage in negotiations that likely grant them positions in the Afghan government or official control over most historic stronghold in southern Afghanistan.

Afghan Analysts said that the visit comes within the context of the peace efforts being made by the Afghan government, and it will discuss the political bureau.

The United States and Afghanistan have agreed to open the political bureau in Qatar despite the fact that the Karzai government hesitated at the beginning, and they prefer to open it in Saudi Arabia or Turkey. Afghanistan has withdrawn its ambassador to Qatar, after the doubts to be excluded from the peace efforts, but they returned to soften its stance afterwards. The U.S. administration is seeking to support the peace process before the withdrawal of foreign forces from Afghanistan in 2014.

Bibliography

[Abdul-Mageed and Diab, 2012] Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914. Citeseer.

[Abdul-Mageed et al., 2011] Abdul-Mageed, M., Diab, M. T., and Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 587–591. Association for Computational Linguistics.

[Abdul-Rauf and Schwenk, 2011] Abdul-Rauf, S. and Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine translation*, pages 1–35.

[Abdulla et al., 2014] Abdulla, N., Al-Ayyoub, M., and Al-Kabi, M. (2014). An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence*, 1(1):103–113.

[Aljlayl et al., 2002] Aljlayl, M., Frieder, O., and Grossman, D. (2002). On Arabic-English Cross-Language Information Retrieval: Machine Translation Approach. In *Machine Readable Dictionaries and Machine Translation," ACM Tenth Conference on Information and Knowledge Managemen (CIKM)*, pages 295–302. ACM Press.

[Alm et al., 2005] Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the*

Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 579–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Aly and Atiya, 2013] Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *ACL (2)*, pages 494–498.

[Aman and Szpakowicz, 2007] Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In Matoušek, V. and Mautner, P., editors, *Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer Berlin Heidelberg.

[Ando and Zhang, 2005] Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

[Arthur, 2011] Arthur, C. (2011). Google and twitter launch service enabling egyptians to tweet by phone. `http://www.theguardian.com/technology/2011/feb/01/google-twitter-egypt`, [Online; accessed 21-August-2014].

[Artstein and Poesio, 2008] Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[Balamurali et al., 2012] Balamurali, A., Joshi, A., and Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for Indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73–82, Mumbai, India. The COLING 2012 Organizing Committee.

[Bautin et al., 2008] Bautin, M., Vijayarenu, L., and Skiena, S. (2008). International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

[Berry and Young, 1995] Berry, M. W. and Young, P. G. (1995). Using latent semantic indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429.

[Biro et al., 2008] Biro, I., Benczur, A., Szabo, J., and Maguitman, A. (2008). A comparative analysis of latent variable models for web page classification. In *Latin American Web Conference, 2008. LA-WEB '08.*, pages 23–28.

[Black et al., 2006] Black, W., Elkateb, S., and Vossen, P. (2006). Introducing the Arabic WordNet project. In *Proceedings of the Third International WordNet Conference*, pages 295–300.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

[Bobicev et al., 2010] Bobicev, V., Maxim, V., Prodan, T., Burciu, N., and Angheluş, V. (2010). Emotions in words: Developing a multilingual wordnet-affect. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 375–384. Springer Berlin Heidelberg.

[Bond and Paik, 2012] Bond, F. and Paik, K. (2012). A Survey of WordNets and their Licenses. In *6th Global WordNet Conference (GWC2012)*, page 64–71.

[Brooke et al., 2009] Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In *International Conference RANLP*, pages 50–54.

[Brown et al., 1993] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

[Cambria et al., 2012] Cambria, E., Havasi, C., and Hussain, A. (2012). SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS Conference*.

[Cambria et al., 2010] Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). SenticNet: A publicly available semantic resource for opinion mining. *Artificial Intelligence*, pages 14–18.

[Carbonell et al., 1997] Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R. D., Geng, Y., and Lee, D. (1997). Translingual information retrieval: A comparative evaluation.

[Cettolo et al., 2012] Cettolo, M., Girardi, C., and Federico, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

[Chaumartin, 2007] Chaumartin, F.-R. (2007). Upar7: a knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 422–425, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Cimiano et al., 2009] Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, IJCAI'09, pages 1513–1518, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Cohen, 1960] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

[Cui et al., 2011] Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., and Tong, X. (2011). Textflow: Towards better understanding of evolving topics in text. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2412–2421.

[Dang et al., 2010] Dang, Y., Zhang, Y., and Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *Intelligent Systems, IEEE*, 25(4):46–53.

[Deerwester et al., 1990] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

[Delpech, 2011] Delpech, E. (2011). Evaluation of terminologies acquired from comparable corpora: an application perspective. In *Proceedings of the 18th International Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 66–73.

[Demirtas and Pechenizkiy, 2013] Demirtas, E. and Pechenizkiy, M. (2013). Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, pages 9:1–9:8, New York, NY, USA. ACM.

[Denecke, 2008] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512.

[Diehn, 2013] Diehn, S. (2013). Social media use evolving in Egypt. `http://www.dw.de/social-media-use-evolving-in-egypt/a-16930251`, [Online; accessed 21-August-2014].

[Dumais, 2007] Dumais, S. (2007). LSA and information retrieval: Getting back to basics. *Handbook of latent semantic analysis*, pages 293–321.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

[Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

[Evans et al., 1998] Evans, D. A., Handerson, S. K., Monarch, I. A., Pereiro, J., Delon, L., and Hersh, W. R. (1998). Mapping vocabularies using latent semantics.

[Farag et al., 2011] Farag, A., Nürnberger, A., and Nitsche, M. (2011). Supporting arabic cross-lingual retrieval using contextual information. In *Proceedings of the Second International Conference on Multidisciplinary Information Retrieval Facility*, IRFC'11, pages 30–45, Berlin, Heidelberg. Springer-Verlag.

[Fleiss et al., 2013] Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

[Fort, 2011] Fort, K. (2011). Corpus Linguistics : Inter-Annotator Agreements.

[Fortuna and Shawe-Taylor, 2005] Fortuna, B. and Shawe-Taylor, J. (2005). The use of machine translation tools for cross-lingual text mining. In *Proceedings of the Workshop on Learning with Multiple Views, 22nd ICML*.

[Froud et al., 2012] Froud, H., Lachkar, A., and Ouatik, S. A. (2012). Stemming versus light stemming for measuring the simitilarity between arabic words with latent semantic analysis model. In *Information Science and Technology (CIST), 2012 Colloquium in*, pages 69–73.

[Fujii and Ishikawa, 2000] Fujii, A. and Ishikawa, T. (2000). Applying machine translation to two-stage cross-language information retrieval. In White, J., editor, *Envisioning Machine Translation in the Information Future*, volume 1934 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg.

[Fung and Cheung, 2004] Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1051. Association for Computational Linguistics.

[Ghanem, 2014] Ghanem, O. (2014). Evaluating the effect of preprocessing in arabic documents clustering. Master's thesis, Computer Engineering Dept., Islamic University of Gaza, Palestine.

[Ghorbel, 2012] Ghorbel, H. (2012). Experiments in cross-lingual sentiment analysis in discussion forums. In Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., and Guéret, C., editors, *Social Informatics*, volume 7710 of *Lecture Notes in Computer Science*, pages 138–151. Springer Berlin Heidelberg.

[Green, 1997] Green, A. M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual Conference of SAS Users Group*, San Diego, USA.

[Habash, 2010] Habash, N. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

[Hamouda and Rohaim, 2011] Hamouda, A. and Rohaim, M. (2011). Reviews classification using sentiwordnet lexicon. *Journal on Computer Science and Information Technology (OJCSIT)*, 2(1):120–123.

[Ion et al., 2010] Ion, R., Tufis, D., Boros, T., Ceausu, A., and Stefanescu, D. (2010). On-line compilation of comparable corpora and their evaluation. *FASSBL7*, page 29.

[Jiang and Littman, 2000] Jiang, F. and Littman, M. L. (2000). Approximate dimension equalization in vector-based information retrieval. In *ICML*, pages 423–430.

[Jin et al., 2004] Jin, X., Zhou, Y., and Mobasher, B. (2004). Web usage mining based on probabilistic latent semantic analysis. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 197–205. ACM.

[Jurafsky and Martin, 2009] Jurafsky, D. and Martin, J. (2009). *Speech & language processing*. Pearson Education.

[Khoja and Garside, 1999] Khoja, S. and Garside, R. (1999). Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

[Knoth et al., 2011] Knoth, P., Zilka, L., and Zdrahal, Z. (2011). Using explicit semantic analysis for cross-lingual link discovery. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 2–10.

[Krippendorff, 1980] Krippendorff, K. (1980). *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc.

[Landauer et al., 1998] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

[Landis and Koch, 1977] Landis, R. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

[Larkey et al., 2002] Larkey, L., Ballesteros, L., and Connell, M. (2002). Improving stemming for arabic information retrieval: light stemming and co-occurrence analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–282. ACM.

[Larkey et al., 2007] Larkey, L., Ballesteros, L., and Connell, M. (2007). Light stemming for arabic information retrieval. In *Arabic computational morphology*, pages 221–243. Springer.

[Li and Gaussier, 2010] Li, B. and 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. Association for Computational Linguistics.

[Li et al., 2011] Li, B., Gaussier, E., and Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 473–478.

[Littman et al., 1998] Littman, M. L., Dumais, S. T., and Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In

Grefenstette, G., editor, *Cross-Language Information Retrieval*, volume 2 of *The Springer International Series on Information Retrieval*, pages 51–62. Springer US.

[Lloyd, 2006] Lloyd, L. (2006). *Lydia: A System for the Large Scale Analysis of Natural Language Text*. PhD thesis, Stony Brook, NY, USA. AAI3239002.

[Ma and Zakhary, 2009] Ma, X. and Zakhary, D. (2009). Arabic newswire english translation collection. Linguistic Data Consortium, Philadelphia.

[Manning and Raghavan, 2008] Manning, C. D. and Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

[McNemar, 1947] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

[Meftouh et al., 2010] Meftouh, K., Laskri, M. T., and Smaïli, K. (2010). Modeling Arabic Language using statistical methods. *Arabian Journal for Science and Engineering*, 35(2C):69–82.

[Miller and Fellbaum, 1998] Miller, G. and Fellbaum, C. (1998). Wordnet: An electronic lexical database. MIT Press Cambridge.

[Miniwatts, 2012] Miniwatts (2012). World Internet Statistics – Top Ten Internet Languages. `http://www.internetworldstats.com/stats7.htm`, [Online; accessed 1-July-2014].

[Mori et al., 2001] Mori, T., Kokubu, T., and Tanaka, T. (2001). Cross-lingual information retrieval based on lsi with multiple word spaces. In *In Proceedings of the 2nd NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*. Citeseer.

[Muhic et al., 2012] Muhic, A., Rupnik, J., and Skraba, P. (2012). Cross-lingual document similarity. In *Information Technology Interfaces (ITI), Proceedings of the ITI 2012 34th International Conference on*, pages 387–392.

[NIST, 2010] NIST, M. I. G. (2010). NIST 2008/2009 open machine translation (OpenMT) evaluation. Linguistic Data Consortium, Philadelphia.

[Orengo and Huyck, 2003] Orengo, V. M. and Huyck, C. (2003). Portuguese-english experiments using latent semantic indexing. In *Advances in Cross-Language Information Retrieval*, pages 147–154. Springer.

[Otero and López, 2009] Otero, P. and López, I. (2009). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 21–25.

[Otero and López, 2011] Otero, P. and López, I. (2011). Measuring comparability of multilingual corpora extracted from wikipedia. In *Iberian Cross-Language Natural Language Processings Tasks (ICL)*, page 8.

[Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

[Pang and Lee, 2005] Pang, B. and Lee, L. (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[Pinnis et al., 2012] Pinnis, M., Ion, R., Stefanescu, D., Su, F., Skadina, I., Vasiljevs, A., and Babych, B. (2012). Accurat toolkit for multi-level alignment and information extraction from comparable corpora. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 91–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Porter, 2001] Porter, M. (2001). Snowball: A language for stemming algorithms.

[Rafalovitch and Dale, 2009] Rafalovitch, A. and Dale, R. (2009). United nations general assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT Summit XII*, volume 13, pages 292–299.

[Rehurek and Sojka, 2010] Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

[Rushdi-Saleh et al., 2011] Rushdi-Saleh, M., Martín-Valdivia, T., Ureña-López, A., and Perea-Ortega, J. (2011). OCA: Opinion corpus for Arabic. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):2045–2054.

[Saad, 2010] Saad, M. (2010). The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification. Master's thesis, Computer Engineering Dept., Islamic University of Gaza, Palestine.

[Saad and Ashour, 2010] Saad, M. and Ashour, W. (2010). Arabic Morphological Tools for Text Mining. In *EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*, pages 112–117. European University of Lefke, Cyprus.

[Saad et al., 2013] Saad, M., Langlois, D., and Smaïli, K. (2013). Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. *Procedia - Social and Behavioral Sciences*, 95(0):40 – 47. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

[Sawalha and Atwell, 2008] Sawalha, M. and Atwell, E. (2008). Comparative evaluation of arabic language morphological analysers and stemmers. In *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics (Poster Volume))*, pages 107–110. Coling 2008 Organizing Committee.

[Skadina et al., 2012] Skadina, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita, M. L., and Pinnis, M. (2012). Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.*

[Smith et al., 2010] Smith, J., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics.

[Stolcke, 2002] Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 70–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Strapparava and Mihalcea, 2010] Strapparava, C. and Mihalcea, R. (2010). Annotating and identifying emotions in text. In Armano, G., Gemmis, M., Semeraro, G., and Vargiu, E., editors, *Intelligent Information Access*, volume 301 of *Studies in Computational Intelligence*, pages 21–38. Springer Berlin Heidelberg.

[Taghva et al., 2005] Taghva, K., Elkhoury, R., and Coombs, J. (2005). Arabic stemming without a root dictionary. In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, volume 1, pages 152–157. IEEE.

[Tang et al., 2011] Tang, G., Xia, Y., Zhang, M., Li, H., and Zheng, F. (2011). CLGVSM: Adapting Generalized Vector Space Model to Cross-lingual Document Clustering. In *IJCNLP*, pages 580–588.

[Tiedemann, 2012] Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Torii et al., 2011] Torii, Y., Das, D., Bandyopadhyay, S., and Okumura, M. (2011). Developing japanese wordnet affect for analyzing emotions. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 80–86. Association for Computational Linguistics.

[Ture, 2013] Ture, F. (2013). *Searching to Translate and Translating to Search: When Information Retrieval Meets Machine Translation*. PhD thesis, Graduate School of the University of Maryland, College Park.

[Ture et al., 2012] Ture, F., Lin, J., and Oard, D. (2012). Combining statistical translation techniques for cross-language information retrieval. In *COLING*, pages 2685–2702.

[UNESCO, 2012] UNESCO (2012). World arabic language day - 18 december 2012. `http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day/`. [Online; accessed 8-July-2014].

[Valitutti, 2004] Valitutti, R. (2004). WordNet-Affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

[Vossen, 1998] Vossen, P. (1998). Eurowordnet: building a multilingual database with wordnets for european languages. *The ELRA Newsletter*, 3(1):7–10.

[Wan, 2009] Wan, X. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 235–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wei and Pal, 2010] Wei, B. and Pal, C. (2010). Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 258–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wei et al., 2008] Wei, C.-P., Yang, C. C., and Lin, C.-M. (2008). A latent semantic indexing-based approach to multilingual document clustering. *Decision Support Systems*, 45(3):606–620.

[Wikimedia, 2014] Wikimedia (2014). Wikipedia statistics. `http://stats.wikimedia.org/EN/TablesArticlesTotal.htm`, [Online; accessed 25-August-2014].

[Wikipedia, 2014] Wikipedia (2014). Wikipedia article depth. `http://meta.wikimedia.org/wiki/Wikipedia_article_depth`, [Online; accessed 25-August-2014].

[Yeh et al., 2005] Yeh, J.-Y., Ke, H.-R., Yang, W.-P., and Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, 41(1):75 – 95. An Asian Digital Libraries Perspective.

[Zhe and Boucouvalas, 2002] Zhe, X. and Boucouvalas, A. (2002). Text-to-emotion engine for real time internet communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs*, pages 164–168.

List of Publications

[Saad et al.(2014a)Saad, Langlois, and Smaïli] Motaz Saad, David Langlois, and Kamel Smaïli. Cross-lingual semantic similarity measure for comparable articles. In *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings*, pages 105–115. Springer International Publishing, 2014a. doi: 10.1007/978-3-319-10888-9_11. URL `http://dx.doi.org/10.1007/978-3-319-10888-9_11`.

[Saad et al.(2014b)Saad, Langlois, and Smaïli] Motaz Saad, David Langlois, and Kamel Smaïli. Building and Modelling Multilingual Subjective Corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014b. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

[Saad et al.(2013a)Saad, Langlois, and Smaïli] Motaz Saad, David Langlois, and Kamel Smaïli. Comparing Multilingual Comparable Articles Based On Opinions. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 105–111, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W13-2513`.

[Saad et al.(2013b)Saad, Langlois, and Smaïli] Motaz Saad, David Langlois, and Kamel Smaïli. Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities. *Procedia - Social and Behavioral Sciences*, 95(0):

40 – 47, 2013b. ISSN 1877-0428. doi: http://dx.doi.org/10.1016/j.sbspro.2013.10.620. URL http://www.sciencedirect.com/science/article/pii/S1877042813041402. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Titre: Fouille de documents et d'opinions multilingue

Résumés

L'objectif de cette thèse est d'étudier les sentiments dans les documents comparables. Premièrement, nous avons recueillis des corpus comparables en anglais, français et arabe de Wikipédia et d'Euronews, et nous avons aligné ces corpus au niveau document. Nous avons en plus collecté des documents d'informations des agences de presse locales et étrangères dans les langues anglaise et arabe. Les documents en anglais ont été recueillis du site de la BBC, ceux en arabe du site d'Al-Jazzera. Deuxièmement, nous avons présenté une mesure de similarité coss-linguistique des documents dans le but de récupérer et aligner automatiquement les documents comparables. Ensuite, nous avons proposé une méthode d'annotation cross-linguistique en termes de sentiments, afin d'étiqueter les documents source et cible avec des sentiments. Enfin, nous avons utilisé des mesures statistiques pour comparer l'accord des sentiments entre les documents comparables source et cible. Les méthodes présentées dans cette thèse ne dépendent pas d'une paire de langue bien déterminée, elles peuvent être appliquées sur toute autre couple de langue.

Mots-clés: fouille de textes; traitement automatique du langage naturel; corpus comparable; recherche d'information inter-langues; projection inter-langues; analyse des sentiments

---

Title: Mining Documents and Sentiments in Cross-lingual Context

Abstract

The aim of this thesis is to study sentiments in comparable documents. First, we collect English, French and Arabic comparable corpora from Wikipedia and Euronews, and we align each corpus at the document level. We further gather English-Arabic news documents from local and foreign news agencies. The English documents are collected from BBC website and the Arabic document are collected from Al-jazeera website. Second, we present a cross-lingual document similarity measure to automatically retrieve and align comparable documents. Then, we propose a cross-lingual sentiment annotation method to label source and target documents with sentiments. Finally, we use statistical measures to compare the agreement of sentiments in the source and the target pair of the comparable documents. The methods presented in this thesis are language independent and they can be applied on any language pair.

Keywords: text mining; natural language processing; comparable corpus; cross-lingual information retrieval; cross-lingual projection; sentiment analysis