Building comparable corpora from social networks

Malek Hajjem[†], Maroua Trabelsi[*], Chiraz Latiri[*]

LISI[†]
INSAT,Centre Urbain Nord Tunis; Tunisia Université de Tunis Carthage, Faculté des sciences de Tunis; Tunisia[*]
malek.hajjem@gmail.com, trabelsimarou@live.com ,chiraz.latiri@gnet.tn

Abstract

Working with comparable corpora becomes an interesting alternative to rare parallel corpora in different natural language tasks. Therefore many researchers have accentuated the need of large quantities of such corpora and the need to work on their construction. In this paper, we highlight the interest and usefulness of textual data mining in social networks. We propose the extraction of tweets from the microblog Twitter in order to construct a comparable corpus. This work aims to develop a new method for the construction of comparable corpus from twitter that could be used in forthcoming work to construct a bilingual dictionary, using text mining approach.

Keywords: Social networks, Text mining , Comparable corpora, Comparability metrics

1. Introduction

Social networks are dynamic structures formed by individuals or organizations. They have been developed and diversified on the web allowing large audiences to express their thoughts and reactions throughout multiple platforms such as blogs, micro-blogs, facebook and wikis in various languages. Recently, social networks have even played an instrumental role in popular revolution, social movement and participated to entire governmental policy changes (Eltantawy and Wiest, 2011). As a result, a large multilingual collection of posts became publicly available. This has made text mining in social networks the subject of many recent researches. In this work, we conduct an exploratory study of the construction of a multilingual resource from these new modes of communication. In fact, multilingual corpora are useful in different areas such as multilingual text mining, bilingual lexicons extraction, cross-lingual information retrieval and machine translation. Multilingual corpora are either **parallel** corpora: corpus that contains source text and their translations (McEnery and Xiao, 2007), or **comparable** corpora : collections of documents in the same or in different languages made up of similar texts.

Although parallel corpora are very effective and used, they have several disadvantages: firstly, their language coverage remains insufficient. Besides, parallel texts freely available are few. They are expensive to produce as they need human translation. Then, comparable corpora are the best alternative, because they are less expensive and more productive. It is clearly easier to find document collections with similar topics in multiple languages than to find parallel corpora (Talvensaari et al., 2007). However, it remains to note that researchers like (Morin et al., 2006) and (Li et al., 2011) are more interested by the exploitation of comparable corpora than creating new methods for their automatic construction.

Our work consists in analyzing and exploiting the huge data text from Twitter in order to build a comparable corpus. Our goal is proving the feasibility of the new method for the construction of comparable corpus using tweets. We focus, in this work, on Arabic and French language seeing that there are few Arabic/other languages pair comparable corpora. For that, we decided to collect French/Arabic tweets about **Arab Spring** posted from May 2013 to September 2013 and to calculate a comparability measure (CM) between collected posts.

In fact, a comparability metric is a key issue in field of building comparable corpora. Its function is to estimate the quality of corpus built on similar topics and different languages. Recent works refer to three ways to calculate comparability measures:

- **Statistical measures**: they are based on the quantity of the common vocabulary. It includes (Li and Gaussier, 2010) who used a translation table and (Su and Babych, 2012) who used a bilingual dictionary, given a comparable corpus P consisting of a source part Ps, and a target part Pt, the degree of comparability of P is defined as the expectation of finding the translation of any given source/target words in the target/source corpus vocabulary. Regarding (Yapomo et al., 2012), their work described a CLIR- based method to calculate similarity between texts. We cite also (Saad et al., 2013) who have proposed two different comparability measures based on binary and cosinus similarity measures. Their work is closer to (Li and Gaussier, 2010). Unlike (Li and Gaussier, 2010), their work was based on the bilingual dictionary Open Multilingual WordNet (OMWN) word alignment.

- **Semantic measures**: they are based on the exploitation of semantic resources to calculate word similarity and still basically used for a monolingual collection (Corley and Mihalcea, 2005). This measure can be adapted to multilingual environment by using resources like global wordNet[1].

- **Hybrid measures**: they are based on the use of both information from corpora and a semantic resource such as the work of (Mohammad et al., 2007) who presented the idea of estimating semantic distance in one language using a knowledge source in another.

[1] http://www.globalwordnet.org

Concerning our work, we discuss in Section 4, the result of two different statistical comparability measures applied to our collected corpus from twitter, which are based on binary and cosine similarity measures. Our work is close to (Saad et al., 2013) who proposed the same comparability measure for Wikipedia corpus. Moreover, (Saad et al., 2013) used a bilingual dictionary, we propose to use machine translation (MT). In fact, MT seems to be more appropriate with the noisy nature of data processed (twitter data). The rest of the paper is organized as follows. We first present some related work in the next section. Section 3 introduces our proposed approach. In section 4, we discuss different evaluations used in this work. Finally, conclusions and some prospects are stated.

2. Related work

2.1. Data sources of comparable corpora

Comparable corpora can be obtained easily from multilingual textual contents. Initially comparable corpora were made from newspapers, in this case the corpora does not target a particular area and cover different topics (Fung and McKeown, 1997), (Rapp, 1999). Scientific articles are considered as an interesting source for comparable corpora, because they cover many languages and topics. For example (Déjean and Gaussier, 2002) built a comparable corpus composed of medical records. For their part, (Chiao, 2004) used specialized websites in the medical field (CISMeF [2] for French corpora and CliniWeb [3] for the English corpora) rather than using general search engines.

Comparable corpora can also be acquired from the web, which is considered as large source of data. Among the studies that have used the web, we cite (Issac et al., 2001) which built a corpus based on syntactic and semantic criteria from the web. (Goeuriot, 2009) has built comparable corpus for language pairs with great linguistic distance (Japanese/French) based on an automatic classification system.

Other approaches like (Laroche and Langlais, 2010), (Rebout, 2012), (Sellami et al., 2013) and (Saad et al., 2013) work on the online encyclopedia, wikipedia to extract comparable articles. Recently, work such as (Gotti et al., 2013), invested in automatic translation of tweets, they exploit the great potential of tweets published by canadian government agencies and organizations to construct a bilingual tweet feeds used to create a tunnig and training material for Statistical Machine Translation. (Jehl et al., 2012) also focused on automated translation of microblogging messages, they provide a bilingual sentence pair data from twitter in English and Arabic about Arab spring for training SMT system.

2.2. Construction methods of comparable corpora

Construction methods allow the acquisition and structuring of multilingual data. They depend on the selected data sources :

- **Thematic crawling** or focused crawling is a method adopted for automatic construction of comparable corpora from the web. It consists in using links between pages to collect documents. This method was used by (Talvensaari et al., 2008) to extract English-Spanish-German comparable corpora mined from the web and concentrate on a specific domain. Thematic crawling has as objective to minimize the number of pages which are not related with the area studied. We note that even if the web is a large volume of data, the automatic acquisition of comparable corpora is still a challenging task.
- **Cross-language information retrieval** is a method which is an independent method from the web. It consists in using the translated keywords of a source collection as a query to the target collection. It was operated by (Talvensaari et al., 2007) who have proposed a new approach using CLIR to extract Swedish-English comparable corpus. In this approach, the keywords were extracted using the RATF [4] measure. Their translations are executed as query on the target collection by the Indri[5] information retrieval system. This method may extract pertinent documents from the target collection but it has a disambiguation issue in the choice of the best translation of keywords.
- **Clustering** is defined as the distribution of a set of texts in groups according their similarity and without a priori knowledge. It has been used by (Li et al., 2011) to obtain bilingual clusters from a part of an initial corpus. This part includes texts above a minimum threshold of similarity that will be used to form a comparable corpora. The same procedure is reproduced on the rest of the corpus. This method of construction is simple and organized but it can be slow.

3. Proposed approach to construct comparable corpora from Twitter

3.1. Textual data collection

In this section, we present our textual data collection extracted from the popular social network Twitter. Twitter is an online social networking and microblogging service that allows users to send and read Twitter messages (tweets), limited to 140 characters. An important role was played by Twitter in the socio-political events, such as the Arabic spring, the theme of our corpus. In fact, since the Arabic revolutions, this media presents itself as a vehicle for the voice of politicians, artists, and especially young people.

This choice of source data was made because of the massive volume of data posted on twitter and available through the Twitter API which allows queries against specific topics. Also, Twitter data can respect criteria of comparability like theme, date proximity and document length.

Tweets about the Arab spring were retrieved using Twitter's Search Api [6] feature which is offered by Twitter to give developers access to tweet data servers. The search API is focused on relevance and not completeness. It usually serves only tweets from the past week.

[2] www.chu-rouen.fr/cismef

[3] www.ohsu.edu/cliniweb

[4] Relative Average Term Frequency

[5] http://www.lemurproject.org/indri.php

[6] http://dev.twitter.com

This API can filter tweets based on queries. For example, to retrieve tweets that report on the movement Occupy Wall Street, you have just to use the keywords that describe this movement and specify the language/period of this movement. After collecting the data, we specify in the following subsection the various forms of data preprocessing performed on the collected corpus.

3.2. Preprocessing of collected corpora

After collecting the data we have employed a number of preprocessing techniques. This phase is a succession of three steps, the result of each step will be used by the next. The three steps are performed on each of the Arabic and French corpus separately.

First, we have eliminated special characters and numbers of each collection to just obtain the textual content of tweets (for example remove the names of users, the punctuation, smileys, etc). Second we have eliminated redundancy by deleting retweets. Retweets is a copy of someone else's tweet broadcasted by a second user to their followers, they do not generally add any new information (McMinn et al., 2013).

The last step is the morpho-syntactic labeling of the tweet corpora. This task associates to each word of the collected corpora a label which recapitulates its morpho-syntactic proprieties in the text. Morpho-syntactic labeling has been made in this step using TreeTagger (Schmid, 1994) for French tweets and MADA (Habash and Rambow, 2005) for Arabic .

3.3. Normalisation of tweet corpora

The variety of linguistic phenomena existing in the textual data and the lack of conventions and spelling standards in social networks require a phase of standardization. In fact, building comparable corpora from this media raises a number of challenges.

Indeed, the recovered data could not be used directly. The writing style, used in social networks and microblogs, is sometimes incomprehensible. The users frequently make spelling and grammar mistakes and create short texts that are difficult to analyze. Our normalisation process is focused, in this work, on the French collection. It was based on a spellcheking approach for normalising short text as works that have been conducted on normalising social media in French language were scarce except some attempts like (Fairon et al., 2006), (Yvon, 2008) and (Beaufort et al., 2010). Our implementation involves the following steps:

- First, we have used a short text messages (SMS) dictionary[7] which covers global spelling mistakes used with SMS and their standard lexical forms. In other words, it provide translations from SMS expressions to plain language expressions. This dictionary was used to identify candidate token (OOV) for lexical normalisation. We note that the coverage of SMS dictionary used, was incapable to identify all OOV words in tweet corpora. For the purpose of this work we have employed a personalized dictionary manually built from a training corpus collected through topsy [8] a tweet search engine. This, personalized dictionary check the OOV words of tweet relative to our theme corpora (for example: manif→ manifestation, mvt→mouvement, jan→janvier).
- Second, our two dictionaries were automatically applied to the corpus, then ill-formed words were transformed to their standard format.

3.4. Description of the built corpus

The constructed textual resource is an Arabic/French bilingual corpus consisting of a total of 52000 tweets which were published on Twitter's public message board during May 2013 to September 2013. We collected tweets that contained the keywords respectively in Table 1 and Table 2. The tweets are then subjected to Pre-processing and standarization resulted in a total of **20025** tweets in Arabic and **20023** tweets in French .

Keywords	Translation	number of tweets
Printemps arabe	Arab spring	4003
Révolution arabe	Arabic revolution	110
Syrie	Syria	9110
Egypte	Egypt	4600
Révolution tunisienne	Tunisian revolution	2200

Table 1: Number of French tweets by keywords (after Preprocessing)

Keywords	Transliteration	Number of tweets
الرّبيع العربيّ	Alrrabyς Alςrbyy	14285
الثّورة العربيّة	Alθwrħ Alςrbyyħ	20
سوريا	swryA	2100
مصر	mSr	2300
الثّورة التونسيّة	Alθwrħ Altwnsyyħ	1320

Note: The transliteration consists on writing Arabic with latin caracters to help non Arabic speakers to read Arabic. In this paper, Arabic orthographic transliteration is presented in the HSB scheme (Habash et al., 2007): (in alphabetical order)

ي و ه ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ت ب ا

A b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y

and the additional letters: ’ ء, Â أ, Ă إ, Ā آ, ŵ ؤ, ŷ ئ, ى, ħ ة, ý ى.

Table 2: Number of Arabic tweets by keywords (after Preprocessing)

4. Evaluation of the comparability

As we stated, comparability is the key concept in the process of building comparable corpora. However, there has been no widely accepted definition of comparability (Liu and Zhang, 2013). Even if, tweets that talk about the same event in the same period but in different languages were extracted, thus respecting comparability's criteria, we need to evaluate similarity between the Arabic and French data collected from Twitter. During this step, methods based on word frequency have been processed to measure corpus homogeneity between French and Arabic collections. In fact,

[7] http ://www.langagesms.com/dictionnaire.html

[8] http://topsy.com/tweets

comparability is defined according to an application. As we aim to use our corpora in extracting bilingual lexicons, these methods are the best alternative because they are generally focused on the amount of common vocabulary in the document. So, the comparability measures used in our approach are statistical measures based on CLIR. Two information retrieval models were considered: binary and vector space model (cosine similarity).

4.1. Binary measure of comparability

In binary measure, the source and target (Arabic and French) collections are represented as a bag of words. In this case, the degree of comparability reflects the absence or presence of keywords (or index) translation from the source vocabulary (respectively target) in the target vocabulary (respectively source).

To extract the index of the two collections (Arabic and French), we have used the Lemur[9] information retrieval system. The resulting indexes are translated with an online MT system[10]. Finally, we have verified the absence/presence of index terms in each collection, in other words, we have calculated the degree of comparability of our corpus in a binary way as follows.

Given a corpus P with a source language L_s and a target language L_c, the binary function $trans(W_s, d_t)$ returns 1 if the translation of a Word from the source vocabulary W_s was found in the target vocabulary d_t and 0 in the other case. Thus, bin-DC for the source and target documents is calculated as follows:

$$binDC(d_s, d_t) = \frac{\sum_{w_s \in d_s} trans(w_s, d_t)}{|d_s|}$$

We note that, $binDC(d_s, d_t)$ and $binDC(d_t, d_s)$ are not symmetrical (Saad et al., 2013), our work was based on the following formula for measuring the total comparability of our comparable corpus :

$$\frac{binDC(d_s,d_t)+binDC(d_t,d_s)}{2}$$

For this measure based on Boolean information retrieval model(bin-DC) the comparability degree is between [0-1]: 1 strongly parallel, 0 neither parallel nor comparable.

4.2. Vector measure of comparability

In the vector information retrieval model, a document is represented as a vector in the vector space. Each vector's document is compound indexing terms. The coordinates of a vector represents the weight of each term. The similarity measure is usually the cosine of the angle that separates the two vectors (Boubekeur-Amirouche, 2008). To represent documents in the vector space model (VSM), we have built the source and target vectors with the following method: we extracted indexes with lemur. The resulting index (in source langage) was translated with MT and ran against the target collection with the Lemur retrieval system based its cosine similarity as retrieval model which uses the idf weighting model to convert documents to vectors. For this second measure which is based on vector model (cosine-DC) the similarity measure logically should therefore be between [-1, 1]:-1 totally opposed, 1 exactly the same and 0 independent. As vectors in our case represent the weight of words in tweets. Since weights of words are always positive values, then the cosine measures ranges also from 0 to 1.

4.3. Results and discussion

To illustrate the evolution of the degree of comparability depending on the amount of data retrieved from *Twitter*, we have created from the French and Arabic corpus several sets containing variable data rates between the first 10% of the corpus and the entire corpus. Then we have calculated the comparability between these datasets through the Boolean model of information retrieval in both Arabic → French and French → Arabic .

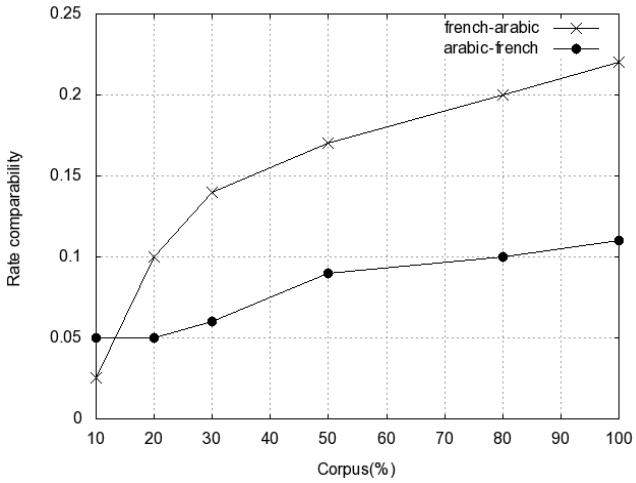

Figure 1: Evolution of comparability with the amount of data

The curve of figure 1 shows that the degree of comparability is proportional to the amount of data in the source and target corpus. As this corpus is exploited automatically, increasing the size of data ensures a lexical coverage. The more vocabulary is used, the more comparability improves.

Measures	bin-DC	cosine-DC
Degree of comparability	0.17	0.22

Table 3: measurement of comparability results

Table 3 summarizes the results of two measures of comparability : bin-DC in the boolean method and cosineDC in the vector method. The results show that the measure of comparability cosine-DC is better than bin-DC. This result was expected since the measure based on vector model includes weighting of terms unlike the Boolean model that uses a binary weighting.

Our experimental results of comparability measures are

[9] http://www.lemurproject.org/

[10] http://www.bing.com/translator

promising and show that our corpora has a comparability feature especially if we compare our results with (Saad et al., 2013) who had used articles from Wikipedia which is considered as user content, less noisy then our textual data, and found close results (0.11 for binary measure and 0.18 for vector measure).

5. Conclusion and Future Work

Despite the popularity of twitter, we note that few researches have been conducted on the construction of corpora based on tweets. This is due to a number of issues associated with the construction of Twitter corpora, including restrictions on the distribution of the tweets, which prevents us to make our corpus available. In this work, we created an Arabic-French comparable corpus, which is, to the best of our knowledge, the first comparable corpus collected from Twitter. We created the corpus of tweets extracted through the Twitter API based on their topic similarities and close publication dates. Experimental results showed that our calculated comparability measures capture a similarity degree for our comparable corpus. In the future we will improve the normalisation step and we will try to treat a larger tweet corpus. We aim also to improve the comparability evaluation. In closing, building comparable corpus from twitter isn't an end in itself; our goal is to exploit this corpus for bilingual extraction in future works.

6. References

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, and Cédrick Fairon. 2010. Une approche hybride traduction/correction pour la normalisation des SMS. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada.

Fatiha Boubekeur-Amirouche. 2008. *Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets*. Ph.D. thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.

Yun-Chuang Chiao. 2004. *Extraction lexicale bilingue à partir de textes médicaux comparables: application à la recherche d'information translangue*. Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hervé Déjean and Eric Gaussier. 2002. Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

Nahed Eltantawy and Julie Wiest. 2011. The arab spring| social media in the egyptian revolution: Reconsidering resource mobilization theory. *International Journal of Communication*, 5(0).

Cédrick Fairon, Jean René, and Sébastien Paumier Klein. 2006. Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation. Presses universitaires de Louvain, Louvain-la-Neuve. Cahiers du Cental, 3.2.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Lorraine Goeuriot. 2009. *Découverte et caractérisation des corpus comparables spécialisés*. Ph.D. thesis, Université de Nantes.

Fabrizio Gotti, Philippe Langlais, and Atefeh Farzindar. 2013. Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*, Atlanta, Georgia, June. Association for Computational Linguistics, Association for Computational Linguistics.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580. Association for Computational Linguistics.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Fabrice Issac, Thierry Hamon, Christophe Fouqueré, Lorne Bouchard, and Louisette Emirkanian. 2001. Extraction informatique de données sur le web. *Revue DistanceS*, 5(2):195–210.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 410–421, Stroudsburg, PA, USA. Association for Computational Linguistics.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING*, pages 617–625. Tsinghua University Press.

Bo Li and Eric Gaussier. 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.

Bo Li, Éric Gaussier, Emmanuel Morin, Amir Hazem, et al. 2011. Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In *18e Conférence sur le Traitement Automatique des Langues Naturelles*, volume 1, pages 211–222.

Sa Liu and Chengzhi Zhang. 2013. Termhood-based comparability metrics of comparable corpus in special domain. In *Proceedings of the 13th Chinese Conference on Chinese Lexical Semantics*, CLSW'12, pages 134–144, Berlin, Heidelberg. Springer-Verlag.

A. M. McEnery and R. Z. Xiao, 2007. *Parallel and comparable corpora: What are they up to?* Translating Europe.

Multilingual Matters. The PDF offprint will be provided when available.

Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information; Knowledge Management*, CIKM '13, pages 409–418, New York, NY, USA. ACM.

Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance.

Emmanuel Morin, Béatrice Daille, et al. 2006. Comparabilité de corpus et fouille terminologique multilingue. *Traitement Automatique des Langues*, 47(1):113–136.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics.

Lise Rebout. 2012. *L'extraction de phrases en relation de traduction dans Wikipédia*. Mémoire présenté à la Faculté des arts et des sciences.Université de Montréal en vue de l'obtention du grade de Maitre de sciences (M.Sc.) en informatique.

Motaz Saad, David Langlois, and Kamel Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. In *V International Conference on Corpus Linguistics. University of Alicante, Spain*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Rahma Sellami, Fatiha Sadat, and Lamia Hadrich Belguith. 2013. Traduction automatique statistique à partir de corpus comparables : application aux couples de langues arabe-français. In *CORIA*, pages 431–440.

Fangzhong Su and Bogdan Babych. 2012. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 10–19, Avignon, France, April. Association for Computational Linguistics.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1):4.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. 2008. Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445.

Manuela Yapomo, Gloria Corpas, and Ruslan Mitkov. 2012. Clir-and ontology-based approach for bilingual extraction of comparable documents. In *The 5th Workshop on Building and Using Comparable Corpora*, page 121.

François Yvon. 2008. Réorthographier des sms. LIMSI.