Building Specialized Bilingual Lexicons Using Large-Scale Background Knowledge

Dhouha Bouamor[1], **Adrian Popescu**[1], **Nasredine Semmar**[1], **Pierre Zweigenbaum**[2]
[1] CEA, LIST, Vision and Content Engineering Laboratory, 91191
Gif-sur-Yvette CEDEX, France; firstname.lastname@cea.fr
[2]LIMSI-CNRS, F-91403 Orsay CEDEX, France; pz@limsi.fr

Abstract

Bilingual lexicons are central components of machine translation and cross-lingual information retrieval systems. Their manual construction requires strong expertise in both languages involved and is a costly process. Several automatic methods were proposed as an alternative but they often rely on resources available in a limited number of languages and their performances are still far behind the quality of manual translations. We introduce a novel approach to the creation of specific domain bilingual lexicon that relies on Wikipedia. This massively multilingual encyclopedia makes it possible to create lexicons for a large number of language pairs. Wikipedia is used to extract domains in each language, to link domains between languages and to create generic translation dictionaries. The approach is tested on four specialized domains and is compared to three state of the art approaches using two language pairs: French-English and Romanian-English. The newly introduced method compares favorably to existing methods in all configurations tested.

1 Introduction

The plethora of textual information shared on the Web is strongly multilingual and users' information needs often go well beyond their knowledge of foreign languages. In such cases, efficient machine translation and cross-lingual information retrieval systems are needed. Machine translation already has a decades long history and an array of commercial systems were already deployed, including Google Translate [1] and Systran [2]. However, due to the intrinsic difficulty of the task, a number of related problems remain open, including: the gap between text semantics and statistically derived translations, the scarcity of resources in a large majority of languages and the quality of automatically obtained resources and translations. While the first challenge is general and inherent to any automatic approach, the second and the third can be at least partially addressed by an appropriate exploitation of multilingual resources that are increasingly available on the Web.

In this paper we focus on the automatic creation of domain-specific bilingual lexicons. Such resources play a vital role in Natural Language Processing (NLP) applications that involve different languages. At first, research on lexical extraction has relied on the use of parallel corpora (Och and Ney, 2003). The scarcity of such corpora, in particular for specialized domains and for language pairs not involving English, pushed researchers to investigate the use of comparable corpora (Fung, 1998; Chiao and Zweigenbaum, 2003). These corpora include texts which are not exact translation of each other but share common features such as domain, genre, sampling period, etc.

The basic intuition that underlies bilingual lexicon creation is the *distributional hypothesis* (Harris, 1954) which puts that words with similar meanings occur in similar contexts. In a multilingual formulation, this hypothesis states that the translations of a word are likely to appear in similar lexical environments across languages (Rapp, 1995). The *standard approach* to bilingual lexicon extraction builds

[1] http://translate.google.com/
[2] http://www.systransoft.com/

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 479–489,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

on the distributional hypothesis and compares context vectors for each word of the source and target languages. In this approach, the comparison of context vectors is conditioned by the existence of a seed bilingual dictionary. A weakness of the method is that poor results are obtained for language pairs that are not closely related (Ismail and Manandhar, 2010). Another important problem occurs whenever the size of the seed dictionary is small due to ignoring many context words. Conversely, when dictionaries are detailed, ambiguity becomes an important drawback.

We introduce a bilingual lexicon extraction approach that exploits Wikipedia in an innovative manner in order to tackle some of the problems mentioned above. Important advantages of using Wikipedia are:

- The resource is available in hundreds of languages and it is structured as unambiguous concepts (i.e. articles).
- The languages are explicitly linked through concept translations proposed by Wikipedia contributors.
- It covers a large number of domains and is thus potentially useful in order to mine a wide array of specialized lexicons.

Mirroring the advantages, there are a number of challenges associated with the use of Wikipedia:

- The comparability of concept descriptions in different languages is highly variable.
- The translation graph is partial since, when considering any language pair, only a part of the concepts are available in both languages and explicitly connected.
- Domains are unequally covered in Wikipedia (Halavais and Lackaff, 2008) and efficient domain targeting is needed.

The approach introduced in this paper aims to draw on Wikipedia's advantages while appropriately addressing associated challenges. Among the techniques devised to mine Wikipedia content, we hypothesize that an adequate adaptation of Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) is fitted to our application context. ESA was already successfully tested in different NLP tasks, such as word relatedness estimation or text classification, and we modify it to mine specialized domains, to characterize these domains and to link them across languages.

The evaluation of the newly introduced approach is realized on four diversified specialized domains (*Breast Cancer*, *Corporate Finance*, *Wind Energy* and *Mobile Technology*) and for two pairs of languages: French-English and Romanian-English. This choice allows us to study the behavior of different approaches for a pair of languages that are richly represented and for a pair that includes Romanian, a language that has fewer associated resources than French and English. Experimental results show that the newly introduced approach outperforms the three state of the art methods that were implemented for comparison.

2 Related Work

In this section, we first give a review of the standard approach and then introduce methods that build upon it. Finally, we discuss works that rely on Explicit Semantic Analysis to solve other NLP tasks.

2.1 Standard Approach (SA)

Most previous approaches that address bilingual lexicon extraction from comparable corpora are based on the standard approach (Fung, 1998; Chiao and Zweigenbaum, 2002; Laroche and Langlais, 2010). This approach is composed of three main steps:

1. **Building context vectors**: Vectors are first extracted by identifying the words that appear around the term to be translated W_{cand} in a window of n words. Generally, association measures such as the mutual information (Morin and Daille, 2006), the log-likelihood (Morin and Prochasson, 2011) or the Discounted Odds-Ratio (Laroche and Langlais, 2010) are employed to shape the context vectors.
2. **Translation of context vectors**: To enable the comparison of source and target vectors, source vectors are translated intoto the target language by using a seed bilingual dictionary. Whenever several translations of a context word exist,

all translation variants are taken into account. Words not included in the seed dictionary are simply ignored.

3. **Comparison of source and target vectors**: Given W_{cand}, its automatically translated context vector is compared to the context vectors of all possible translations from the target language. Most often, the cosine similarity is used to rank translation candidates but alternative metrics, including the weighted Jaccard index (Prochasson et al., 2009) and the city-block distance (Rapp, 1999), were studied.

2.2 Improvements of the Standard Approach

Most of the improvements of the standard approach are based on the observation that the more representative the context vectors of a candidate word are, the better the bilingual lexicon extraction is. At first, additional linguistic resources, such as specialized dictionaries (Chiao and Zweigenbaum, 2002) or transliterated words (Prochasson et al., 2009), were combined with the seed dictionary to translate context vectors.

The ambiguities that appear in the seed bilingual dictionary were taken into account more recently. (Morin and Prochasson, 2011) modify the standard approach by weighting the different translations according to their frequency in the target corpus. In (Bouamor et al., 2013), we proposed a method that adds a word sense disambiguation process relying on semantic similarity measurement from WordNet to the standard approach. Given a context vector in the source language, the most probable translation of polysemous words is identified and used for building the corresponding vector in the target language. The most probable translation is identified using the monosemic words that appear in the same lexical environment.

On specialized French-English comparable corpora, this approach outperforms the one proposed in (Morin and Prochasson, 2011), which is itself better than the standard approach. The main weakness of (Bouamor et al., 2013) is that the approach relies on WordNet and its application depends on the existence of this resource in the target language. Also, the method is highly dependent on the coverage of the seed bilingual dictionary.

2.3 Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) is a method that maps textual documents onto a structured semantic space using classical text indexing schemes such as TF-IDF. Examples of semantic spaces used include Wikipedia or the Open Directory Project but, due to superior performances, Wikipedia is most frequently used. In the original evaluation, ESA outperformed state of the art methods in a word relatedness estimation task.

Subsequently, ESA was successfully exploited in other NLP tasks and in information retrieval. Radinsky and al. (2011) added a temporal dimension to word vectors and showed that this addition improves the results of word relatedness estimation. (Hassan and Mihalcea, 2011) introduced Salient Semantic Analysis (SSA), a development of ESA that relies on the detection of salient concepts prior to mapping words to concepts. SSA and the original ESA implementation were tested on several word relatedness datasets and results were mixed. Improvements were obtained for text classification when comparing SSA with the authors' in-house representation of the method. ESA has weak language dependence and was already deployed in multilingual contexts. (Sorg and Cimiano, 2012) extended ESA to other languages and showed that it is useful in cross-lingual and multilingual retrieval task. Their focus was on creating a language independent conceptual space in which documents would be mapped and then retrieved.

Some open ESA topics related to bilingual lexicon creation include: (1) the document representation which is simply done by summing individual contributions of words, (2) the adaptation of the method to specific domains and (3) the coverage of the underlying resource in different language.

3 ESA for Bilingual Lexicon Extraction

The main objective of our approach is to devise lexicon translation methods that are easily applicable to a large number of language pairs, while preserving the overall quality of results. A subordinated objective is to exploit large scale background multilingual knowledge, such as the encyclopedic content available in Wikipedia. As we mentioned, ESA

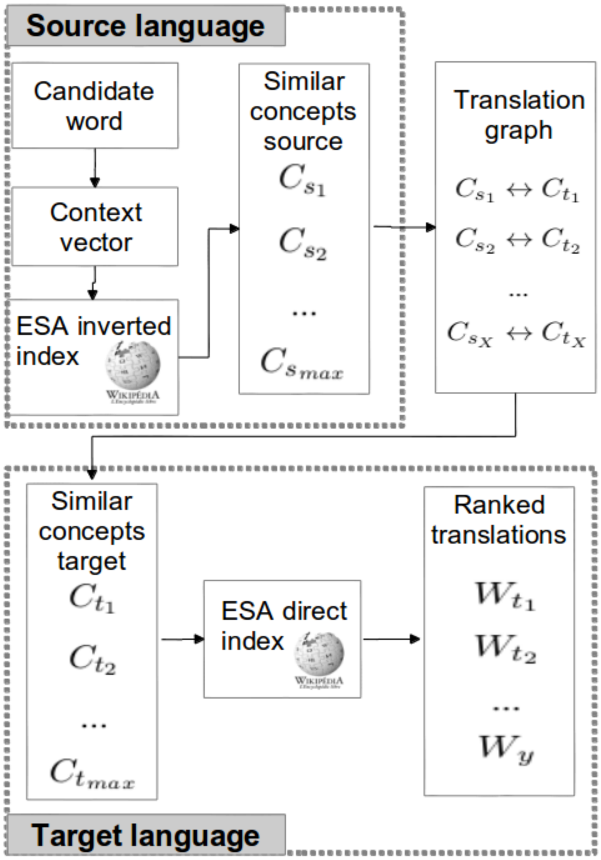

Figure 1: Overview of the Explicit Semantic Analysis enabled bilingual lexicon extraction.

(Gabrilovich and Markovitch, 2007) was exploited in a number of NLP tasks but not in bilingual lexicon extraction.

Figure 1 shows the overall architecture of the lexical extraction process we propose. The process is completed in the following three steps:

1. Given a word to be translated and its context vector in the source language, we derive a ranked list of similar Wikipedia concepts (i.e. articles) using the ESA inverted index.
2. Then, a translation graph is used to retrieve the corresponding concepts in the target language.
3. Candidate translations are found through a statistical processing of concept descriptions from the ESA direct index in the target language.

In this section, we first introduce the elements of the original formulation of ESA necessary in our approach. Then, we detail the three steps that compose the main bilingual lexicon extraction method illustrated in Figure 1. Finally, as a complement to the main method we introduce a measure for domain word specificity and present a method for extracting generic translation lexicons.

3.1 ESA Word and Concept Representation

Given a semantic space structured using a set of M concepts and including a dictionary of N words, a mapping between words and concepts can be expressed as the following matrix:

$$\begin{array}{cccc}
w(W_1, C_1) & w(W_2, C_1) & ... & w(W_N, C_1) \\
w(W_1, C_2) & w(W_2, C_2) & ... & w(W_N, C_2) \\
... & ... & ... & \\
w(W_1, C_M) & w(W_2, C_M) & ... & w(W_N, C_M)
\end{array}$$

When Wikipedia is exploited concepts are equated to Wikipedia articles and the texts of the articles are processed in order to obtain the weights that link words and concepts. In (Gabrilovich and Markovitch, 2007), the weights w that link words and concepts were obtained through a classical TF-IDF weighting of Wikipedia articles. A series of tweaks destined to improve the method's performance were used and disclosed later[3]. For instance, administration articles, lists, articles that are too short or have too few links are discarded. Higher weight is given to words in the article title and more longer articles are favored over shorter ones. We implemented a part of these tweaks and tested our own version of ESA with the Wikipedia version used in the original implementation. The correlation with human judgments of word relatedness was 0.72 against 0.75 reported by (Gabrilovich and Markovitch, 2007). The ESA matrix is sparse since the N size of the dictionary, is usually in the range of hundreds of thousands and each concept is usually described by hundreds of distinct words. The direct ESA index from Figure 1 is obtained by reading the matrix horizontally while the inverted ESA index is obtained by reading the matrix vertically.

[3] `https://github.com/faraday/wikiprep-esa/wiki/roadmap`

Terme	Concepts
action	*évaluation d'action, communisme, actionnaire activiste, socialisme, dévelopement durable ...*
déficit	*crise de la dette dans la zone euro, dette publique, règle d'or budgétaire, déficit, trouble du déficit de l'attention ...*
cisaillement	*taux de cisaillement, zone de cisaillement, cisaillement, contrainte de cisaillement, viscoanalyseur ...*
turbine	*ffc turbine potsdam, turbine à gaz, turbine, urbine hydraulique, cogénération ...*
cryptage	*TEMPEST, chiffrement, liaison 16, Windows Vista, transfert de fichiers ...*
protocole	*Ad-hoc On-demand Distance Vector, protocole de Kyoto, optimized link state routing protocol, liaison 16, IPv6 ...*
biopsie	*biopsie, maladie de Horton, cancer du sein, cancer du poumon, imagerie par résonance magnétique ...*
palpation	*cancer du sein, cellulite, examen clinique, appendicite, ténosynovite ...*

Table 1: The five most similar Wikipedia concepts to the French terms *action[share], déficit[deficit], cisaillement[shear], turbine[turbine], cryptage[encryption], biopsie[biopsie]* and *palpation[palpation]* and their context vectors.

3.2 Source Language Processing

The objective of the source language processing is to obtain a ranked list of similar Wikipedia concepts for each candidate word (W_{cand}) in a specialized domain. To do this, a context vector is first built for each W_{cand} from a specialized monolingual corpus. The association measure between W_{cand} and context words is obtained using the Odds-Ratio (defined in equation 5). Wikipedia concepts in the source language C_s that are similar to W_{cand} and to a part of its context words are extracted and ranked using equation 1.

$$Rank(C_s) = (10 * \max(Odds^{W_{cand}}_{W_{s_i}}) * w(W_{cand}, C_s)) + \sum_{i=1}^{n} Odds^{W_{cand}}_{W_{s_i}} * w(W_{s_i}, C_s) \quad (1)$$

where $\max(Odds^{W_{cand}}_{W_{s_i}})$ is the highest Odds-Ratio association between W_{cand} and any of its context words W_{s_i}; the factor 10 was empirically set to give more importance to W_{cand} over context words; $w(W_{cand}, C_s)$ is the weight of the association between W_{cand} and C_s from the ESA matrix; n is the total number of words W_{s_i} in the context vector of W_{cand}; $Odds^{W_{cand}}_{W_{s_i}}$ is the association value between W_{cand} and W_{s_i} and $w(W_{s_i}, C_s)$ are the weights of the associations between each context word W_{s_i} and C_s from the ESA matrix. The use of contextual information in equation 1 serves to characterize the candidate word in the target domain.

In table 1, we present the five most similar Wikipedia concepts to the French terms *action, déficit, cisaillement, turbine, cryptage, biopsie* and *palpation* and their context vectors. These terms are part of the four specialized domains we are studying here. From observing these examples, we note that despite the difference between the specialized domains and word ambiguity (words *action* and *protocole*), our method has the advantage of successfully representing each word to be translated by relevant conceptual spaces.

3.3 Translation Graph Construction

To bridge the gap between the source and target languages, a concept translation graph that enables the multilingual extension of ESA is used. This concept translation graph is extracted from the explicit translation links available in Wikipedia articles and is exploited in order to connect a word's conceptual space in the source language with the corresponding conceptual space in the target language. Only a part of the articles have translations and the size of the conceptual space in the target language is usually smaller than the space in the source language. For instance, the French-English translation graph contains 940,215 pairs of concepts while the French and English Wikipedias contain approximately 1.4 million articles, respectively 4.25 million articles.

3.4 Target Language Processing

The third step of the approach takes place in the target language. Using the translation graph, we select the 100 most similar concept translations (threshold determined empirically after preliminary experiments) from the target language and use their direct ESA representations in order to retrieve potential translations for the candidate word W_{cand} from source language. These candidate translations W_t are ranked using equation 2.

$$Rank(W_t) = (\sum_{i=1}^{n} \frac{w(W_t, C_{t_i})}{avg(C_{t_i})}) * log(count(W_t, \mathbf{S})) \quad (2)$$

with $w(W_t, C_{t_i})$ is the weight of the translation candidate W_T for concept C_{t_i} from the ESA matrix in the target language; $avg(C_{t_i})$ is the average TF-IDF score of words that appear in C_{t_i}; **S** is the set of similar concepts C_{t_i} in the target language and $count(W_t, \mathbf{S})$ accounts for the number of different concepts from **S** in which the candidate translation W_T appears.

The accumulation of weights $w(W_t, C_{t_i})$ follows the way original ESA text representations are calculated (Gabrilovich and Markovitch, 2007) and $avg(C_{t_i})$ is used in order to correct the bias of the TF-IDF scheme towards short articles. $log(count(W_t, \mathbf{S}))$ is used to favor words that are associated with a larger number of concepts. log weighting was chosen after preliminary experiments with a wide range of functions.

3.5 Domain Specificity

In previous works, ESA was usually exploited in generic tasks that did not require any domain adaptation. Here we process information from specific domains and we need to measure the specificity of words in those domains. The domain extraction is seeded by using Wikipedia concepts (noted C_{seed}) that best describes the domain in the target language. For instance, in English, the *Corporate Finance* domain is seeded with *https://en.wikipedia.org/wiki/Corporate_finance*. We extract a set of 10 words with the highest TF-IDF score from this article (noted SW) and use them to retrieve a domain ranking of concepts in the target language $Rank_{dom}(C_t)$ with equation 3.

$$Rank_{dom}(C_t) = (\sum_{i=1}^{n} w(W_{t_i}, C_t) * w(C_{seed}, W_{t_i})) * count(SW, C_t) \quad (3)$$

where n is size of the seed list of words (i.e. 10 items), $w(W_{t_i}, C_t)$ is the weight of the domain words in the concept C_t ; $w(C_{seed}, W_{t_i})$ is the weight of W_{t_i} in C_{seed}, the seed concept of the domain, and $count(SW, C_t)$ is the number of distinct seed words from SW that appear in C_t.

The first part of equation 3 sums up the contributions of different words from SW that appear in C_t while the second part is meant to further reinforce articles that contain a larger number of domain keywords from SW.

Domain delimitation is performed by retaining articles whose $Rank_{dom}(C_t)$ is at least 1% or the score of the top $Rank_{dom}(C_t)$ score. This threshold was set up during preliminary experiments. Given the delimitation obtained with equation 3, we calculate a domain specificity score ($specif_{dom}(W_t)$) for each word that occurs in the domain (equation 4). $specif_{dom}(W_t)$ estimates how much of a word's use in an underlying corpus is related to a target domain.

$$specif_{dom}(W_t) = \frac{DF_{dom}(W_t)}{DF_{gen}(W_t)} \quad (4)$$

where DF_{dom} and DF_{gen} stand for the domain and the generic document frequency of the word W_t.

$specif_{dom}(W_t)$ will be used to favor words with greater domain specificity over more general ones when several translations are available in a seed generic translation lexicon. For instance, the French word *action* is ambiguous and has English translations such as *action*, *stock*, *share* etc. In a general case, the most frequent translation is *action* whereas in a *corporate finance* context, *share* or *stock* are more relevant. The specificity of the three translations, from highest to lowest, is: *share*, *stock* and *action* and is used to rank these potential translations.

3.6 Generic Dictionaries

Generic translation dictionaries, already used by existing bilingual lexicon extraction approaches, can also be integrated in the newly proposed approach. The Wikipedia translation graph is transformed into a translation dictionary by removing the disambiguation marks from ambiguous concept titles, as well as lists, categories and other administration pages. Moreover, since the approach does not handle multiword units, we retain only translation pairs that are composed of unigrams in both languages. When existing, unigram redirections are also added in each language.

The obtained dictionaries are incomplete since: (1) Wikipedia focuses on concepts that are most often nouns, (2) specialized domain terms often do not have an associated Wikipedia entry and (3) the translation graph covers only a fraction of the concepts available in a language. For instance, the resulting translation dictionaries have 193,543 entries for French-English and 136,681 entries for Romanian-English. They can be used in addition to or instead of other resources available and are especially useful when there are only few other resources that link the pair of languages processed.

4 Evaluation

The performances of our approach are evaluated against the standard approach and its developments proposed by (Morin and Prochasson, 2011) and (Bouamor et al., 2013). In this section, we first describe the data and resources we used in our experiments. We then present differents parameters needed in the implementation of the different methods tested. Finally, we discuss the obtained results.

4.1 Data and Resources

Comparable corpora

We conducted our experiments on four French-English and Romanian-English specialized comparable corpora: *Corporate Finance, Breast Cancer, Wind Energy* and *Mobile Technology*. For the Romanian-English language pair, we used Wikipedia to collect comparable corpora for all domains since they were not already available. The Wikipedia corpora are harvested using a category-based selection. We consider the topic in the source language (for instance *Cancer Mamar* [*Breast Cancer*]) as a query to Wikipedia and extract all its subtopics (i.e., sub-categories) to construct a domain-specific *category tree*. Then, based on the constructed tree, we collect all Wikipedia articles belonging to at least one of these categories and use *inter-language links* to build the comparable corpora.

Domain	FR	EN
Corporate Finance	402,486	756,840
Breast Cancer	396,524	524,805
Wind Energy	145,019	345,607
Mobile Technology	197,689	144,168

Domain	RO	EN
Corporate Finance	206,169	524,805
Breast Cancer	22,539	322,507
Wind Energy	121,118	298,165
Mobile Technology	200,670	124,149

Table 2: Number of *content words* in the comparable corpora.

Concerning the French-English pair, we followed the strategy described above to extract the comparable corpora related to the *Corporate Finance* and *Breast Cancer* domains since they were otherwise unavailable. For the two other domains, we used the corpora released in the TTC project[4]. All corpora were normalized through the following linguistic preprocessing steps: tokenization, part-of-speech tagging, lemmatization, and function word removal. The resulting corpora[5] sizes are presented in Table 2. The size of the domain corpora vary within and across languages, with the corporate finance domain being the richest in both languages. In Romanian, *Breast Cancer* is particularly small, with approximately 22,000 tokens included. This variability will allow us to test if there is a correlation between corpus size and quality of results.

Bilingual dictionary

The seed generic French-English dictionary used to translate French context vectors consists of an in-house manually built resource which contains approximately 120,000 entries. For Romanian-

[4] `http://www.ttc-project.eu/index.php/releases-publications`

[5] Comparable corpora will be shared publicly

Domain	FR-EN	RO-EN
Corporate Finance	125	69
Breast Cancer	96	38
Wind Energy	89	38
Mobile Technology	142	94

Table 3: Sizes of the evaluation lists.

English, we used the generic dictionary extracted following the procedure described in Subsection 3.6.

Gold standard

In bilingual terminology extraction from comparable corpora, a reference list is required to evaluate the performance of the alignment. Such lists are usually composed of around 100 single terms (Hazem and Morin, 2012; Chiao and Zweigenbaum, 2002). Reference lists[6] were created for the four specialized domains and the two pairs of languages. For the French-English, reference words from the *Corporate Finance* domain were extracted from the glossary of bilingual micro-finance terms[7]. For *Breast Cancer*, the list is derived from the MESH and the UMLS thesauri[8]. Concerning *Wind Energy* and *Mobile Technology*, lists were extracted from specialized glossaries found on the Web. The Romanian-English gold standard was manually created by a native speaker starting from the French-English lists. Table 3 displays the sizes of the obtained lists. Reference terms pairs were retained if each word composing them appeared at least five times in the comparable domain corpora.

4.2 Experimental setup

Aside from those already mentioned, three parameters need to be set up: (1) the window size that defines contexts, (2) the association measure that measures the strength of the association between words and the (3) similarity measure that ranks candidate translations for state of the art methods. Context vectors are defined using a seven-word window which approximates syntactic dependencies. The association and the similarity measures (Discounted Log-Odds ratio (equation 5) and the cosine similarity) were set following Laroche and Langlais (2010), a comprehensive study of the influence of these parameters on the bilingual alignment.

$$\textit{Odds-Ratio}_{disc} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})} \quad (5)$$

where O_{ij} are the cells of the 2×2 contingency matrix of a token s co-occurring with the term S within a given window size.

The F-measure of the Top 20 results (F-Measure@20), which measures the harmonic mean of precision and recall, is used as evaluation metric. Precision is the total number of correct translations divided by the number of terms for which the system returned at least one answer. Recall is equal to the ratio between the number of correct translation and the total number of words to translate (W_{cand}).

4.3 Results and discussion

In addition to the basic approach based on ESA (denoted ESA), we evaluate the performances of a method so-called Dico$_{Spec}$ in which the translations are extracted from a generic dictionary and a method we called ESA$_{Spec}$ which combine ESA and Dico$_{Spec}$. DICO$_{Spec}$ is based on the generic dictionary we presented in subsection 3.6 and proceeds as follows: we extract a list of translations for each word to be translated from the generic dictionary. The domain specificity introduced in subsection 3.5 is then used to rank these translations. For instance, the french term *port* referring in the *Mobile Technology* domain, to the system that allows computers to receive and transmit information is translated into *port* and *seaport*. According to domain specificity values, the following ranking is obtained: the English term *port* obtain the highest specificity value (0.48). *seaport* comes next with a specificity value of 0.01. In ESA$_{Spec}$, the translations set out in the translations lists proposed by both ESA and the generic dictionary are weighted according to their domain specificity values. The main intuition behind this method is that by adding the information about the domain specificity, we obtain a new ranking of the bilingual extraction results.

The obtained results are displayed in table 4. The comparison of state of the art method shows that BA13 performs better than STAPP and MP11 for French-English and has comparable performances

[6] Reference lists will be shared publicly

[7] `http://www.microfinance.lu/en/`

[8] `http://www.nlm.nih.gov/`

	Method	F-Measure@20			
		Breast Cancer	Corporate Finance	Wind Eenrgy	Mobile Technology
a) FR-EN	STAPP	0.49	0.17	0.08	0.06
	MP11	0.55	0.33	0.24	0.05
	BA13	0.61	0.37	0.30	0.24
	Dico$_{spec}$	0.50	0.20	0.36	0.25
	ESA	0.74	0.50	0.83	0.72
	ESA$_{spec}$	**0.81**	**0.56**	**0.86**	**0.75**

	Method	F-Measure@20			
		Breast Cancer	Corporate Finance	Wind Eenrgy	Mobile Technology
b) RO-EN	STAPP	0.21	0.13	0.08	0.16
	MP11	0.21	0.13	0.08	0.16
	BA13	0.21	0.14	0.08	0.17
	Dico$_{spec}$	0.44	0.11	0.21	0.16
	ESA	0.76	0.17	**0.58**	0.53
	ESA$_{spec}$	**0.78**	**0.24**	**0.58**	**0.55**

Table 4: Results of the specialized dictionary creation on four specific domains, two pairs of languages.Three state of the art methods were used for comparison: STAPP is the standard approach, MP11 is the improvement of the standard approach introduced in (Morin and Prochasson, 2011), BA13 is a recent method that we developed (Bouamor et al., 2013). Dico$_{spec}$ exploits a generic dictionary, combined with the use of domain specificity (see Subsection 3.5). ESA stands for the ESA based approach introduced in this paper (see Figure 1). ESA$_{spec}$ combines the results of Dico$_{spec}$ and ESA.

for RO-EN. Consequently, we will use BA13 as the main baseline for discussing the newly introduced approach. The results presented in Table 4 show that ESA$_{spec}$ clearly outperforms the three baselines for the four domains and the two pairs of languages tested. When comparing ESA$_{spec}$ to BA13 for French-English, improvements range between 0.19 for *Corporate Finance* and 0.56 for *Wind Energy*. For RO-EN, the improvements vary from 0.1 for *Corporate Finance* to 0.5 for *Wind Energy*. Also, except for the *Corporate Finance* domain in Romanian, the performance variation across domains is much smaller for ESA$_{spec}$ than for the three state of the art methods. This shows that ESA$_{spec}$ is more robust to domain change and thus more generic.

The results obtained with ESA are significantly better than those obtained with Dico$_{spec}$ and ESA$_{spec}$, their combination, further improves the results. The main contribution to ESA$_{spec}$ performances comes from ESA, a finding that validates our assumption that the adequate use of a rich multilingual resource such as Wikipedia is appropriate for specialized lexicon translation. Dico$_{spec}$ is a simple method that ranks the different meanings of a candidate word available in a generic dictionary but its average performances are comparable to those of BA13 for FR-EN and higher for RO-EN. This finding advocates for the importance of good quality generic dictionaries in specialized lexicon translation approaches. However, it is clear that such dictionaries are far from being sufficient in order to cover all possible domains. There is no clear correlation between domain size and quality of results. Although richer than the other three domains, *Corporate Finance* has the lowest associated performances. This finding is probably explained by the intrinsic difficulty of each domain. When passing from FR-EN to RO-EN the average performance drop is more significant for BA13 than for the ESA based methods. The result indicates that our approach is more robust to language change.

5 Conclusion

We have presented a new approach to the creation of specialized bilingual lexicons, one of the central building blocks of machine translation systems. The proposed approach directly tackles two of the major challenges identified in the Introduction. The scarcity of resources is addressed by an adequate exploitation of Wikipedia, a resource that is available in hundreds of languages. The quality of automatic translations was improved by appropriate domain delimitation and linking across languages, as well as by an adequate statistical processing of concepts similar to a word in a given context.

The main advantages of our approach compared to state of the art methods come from: the increased number of languages that can be processed, from the smaller sensitivity to structured resources and the appropriate domain delimitation. Experimental validation is obtained through evaluation with four different domains and two pairs of languages which shows consistent performance improvement. For French-English, two languages that have rich associated Wikipedia representations, performances are very interesting and are starting to approach those of manual translations for three domains out of four (F-Measure@20 around 0.8). For Romanian-English, a pair involving a language with a sparser Wikipedia representation, the performances of our method drop compared to French-English . However, they do not decrease to the same extent as those of the best state of the art method tested. This finding indicates that our approach is more general and, given its low language dependence, it can be easily extended to a large number of language pairs.

The results presented here are very encouraging and we will to pursue work in several directions. First, we will pursue the integration of our method, notably through comparable corpora creation using the data driven domain delimitation technique described in Subsection 3.5. Equally important, the size of the domain can be adapted so as to find enough context for all the words in domain reference lists. Second, given a word in a context, we currently exploit all similar concepts from the target language. Given that comparability of article versions in the source and the target language varies, we will evaluate algorithms for filtering out concepts from the target language that have low alignment with their source language versions. A final line of work is constituted by the use of distributional properties of texts in order to automatically rank parts of concept descriptions (i.e. articles) by their relatedness to the candidate word. Similar to the second direction, this process involves finding comparable text blocks but rather at a paragraph or sentence level than at the article level.

References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Context vector disambiguation for bilingual lexicon extraction. In *Proceedings of the 51st Association for Computational Linguistics (ACL-HLT)*, Sofia, Bulgaria, August.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5. Association for Computational Linguistics.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2003. The effect of a general lexicon in corpus-based identification of french-english medical word translations. In *Proceedings Medical Informatics Europe, volume 95 of Studies in Health Technology and Informatics*, pages 397–402, Amsterdam.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Parallel Text Processing*, pages 1–17. Springer.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Alexander Halavais and Derek Lackaff. 2008. An Analysis of Topical Coverage of Wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440.

Z.S. Harris. 1954. Distributional structure. *Word*.

Samer Hassan and Rada Mihalcea. 2011. Semantic relatedness using salient semantic analysis. In *AAAI*.

Amir Hazem and Emmanuel Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings, 8th international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May.

Azniah Ismail and Suresh Manandhar. 2010. Bilingual lexicon extraction from comparable corpora using in-domain terms. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 481–489. Association for Computational Linguistics.

Audrey Laroche and Philippe Langlais. 2010. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, Beijing, China, Aug.

Emmanuel Morin and Béatrice Daille. 2006. Comparabilité de corpus et fouille terminologique multilingue. In *Traitement Automatique des Langues (TAL)*.

Emmanuel Morin and Emmanuel Prochasson. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *Proceedings, 4th Workshop on Building and Using Comparable Corpora (BUCC)*, page 27–34, Portland, Oregon, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Emmanuel Prochasson, Emmanuel Morin, and Kyo Kageura. 2009. Anchor points for bilingual lexicon extraction from small comparable corpora. In *Proceedings, 12th Conference on Machine Translation Summit (MT Summit XII)*, page 284–291, Ottawa, Ontario, Canada.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322. Association for Computational Linguistics.

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*, ACL '99, pages 519–526. Association for Computational Linguistics.

P. Sorg and P. Cimiano. 2012. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data Knowl. Eng.*, 74:26–45, April.