Word Co-occurrence Counts Prediction for Bilingual Terminology Extraction from Comparable Corpora

Amir Hazem and **Emmanuel Morin**
Laboratoire d'Informatique de Nantes-Atlantique (LINA)
Université de Nantes, 44322 Nantes Cedex 3, France
{Amir.Hazem, Emmanuel.Morin}@univ-nantes.fr

Abstract

Methods dealing with bilingual lexicon extraction from comparable corpora are often based on word co-occurrence observation and are by essence more effective when using large corpora. In most cases, specialized comparable corpora are of small size, and this particularity has a direct impact on bilingual terminology extraction results. In order to overcome insufficient data coverage and to make word co-occurrence statistics more reliable, we propose building a predictive model of word co-occurrence counts. We compare different predicting models with the traditional *Standard Approach* (Fung, 1998) and show that once we have identified the best procedures, our method increases significantly the performance of extracting word translations from comparable corpora.

1 Introduction

Using comparable corpora for bilingual lexicon extraction is becoming more and more a matter of interest, especially because of the easier availability of this kind of corpora comparing to parallel ones. Many researchers proposed a variety of approaches (Fung, 1995; Rapp, 1999; Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Morin et al., 2007; Laroche and Langlais, 2010, among others). While different improvements were achieved, the starting point remains words'co-occurrences as they represent the observable evidence that can be distilled from a corpus. Hence, frequency counts for word pairs often serve as a basis for distributional methods. The main assumption underlying bilingual lexicon extraction is: two words are more likely to be a translation of each other if they share the same lexical contexts (Fung, 1998). The most popular approach named, the *Standard Approach* (Fung and Mckeown, 1997; Rapp, 1999), makes use of this assumption to perform bilingual lexicon extraction. While good results on single word terms (SWTs) can be obtained from large corpora of several million words (80% for the top 10-20 (Fung and Mckeown, 1997), 91% accuracy for the top 3 (Cao and Li, 2002)). Results drop significantly using specialized small corpora (60% for the top 20 (Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Morin et al., 2007).

The reliability of co-occurrence counts greatly relies on the amount of data. Clearly, the larger the training corpus, the more representative it is likely to be, and thus the more reliable the statistics of words. Therefore, the number and distribution of types in the available small sample are not reliable estimators (Evert and Baroni, 2007). This latter fact motivates the necessity of an alternative to the unreliable counts especially when using small specialized comparable corpora. Statistical NLP often deals in the prediction of variables ranging from text categories to linguistic structures to novel utterances. If large specialized comparable corpora are not available, one way to approach this problem and to make co-occurrence counts more reliable, is to use prediction models of word co-occurrence counts based on large training datasets. Corpus data from the general domain such as newspapers, for instance, is abundant and can be easily used for training.

The main contribution of this paper is to investigate different word co-occurrence prediction models for the task of bilingual terminology extraction from comparable corpora. Our aim is to make the observed word co-occurrence counts in small specialized comparable corpora more reliable by re-estimating their probabilities. For that purpose we explore different models such as the linear regression often used to model data using

International Joint Conference on Natural Language Processing, pages 1392–1400,
Nagoya, Japan, 14-18 October 2013.

linear predictor functions, the mean average word co-occurrence increase and the Good-Turing estimator. All of the predicting models rely on the observed counts of word co-occurrence in a training dataset of small and large corpora from the general domain. While prediction is widely used in NLP, to our knowledge no investigation of co-occurrence prediction for the task of bilingual terminology extraction from comparable corpora has been addressed so far. We show that using our method as a pre-processing step of the *Standard Approach*, leads to significant improvements on the performance of bilingual terminology extraction.

In the remainder of this paper, we present in section 2 the related work on bilingual lexicon extraction from comparable corpora. Then, we introduce in section 3 the *Standard Approach* used as baseline. Section 4 describes our method and the different predicting models of co-occurrence counts. Section 5 describes the different linguistic resources used in our experiments. Section 6 evaluates the contribution of the predicting models on the quality of bilingual terminology extraction through different experiments. We discuss our findings in section 7 and finally conclude in section 8.

2 Related Work

The distributional hypothesis which states that words with similar meaning tend to occur in similar contexts, has been extended to the bilingual scenario (Fung, 1998; Rapp, 1999). Hence, using comparable corpora, a translation of a source word can be found by identifying a target word with the most similar context. A popular method often used as a baseline is the *Standard Approach* (Fung, 1998). It consists of using the bag-of-words paradigm to represent words of source and target language by their context vector. After word contexts have been weighted using an association measure (the point-wise mutual information (Fano, 1961), the log-likelihood (Dunning, 1993), the discounted odds-ratio (Laroche and Langlais, 2010)), the similarity between a source word's context vector and all the context vectors in the target language is computed using a similarity measure (cosine (Salton and Lesk, 1968), Jaccard (Grefenstette, 1994)...). Finally, the translation candidates are ranked according to their similarity score.

Many variants of the *Standard Approach* have been proposed. They can differ in context representation (window-based, syntactic-based) (Morin et al., 2007; Gamallo, 2008), corpus characteristics (small, large, general or domain specific...)(Chiao and Zweigenbaum, 2002; Déjean et al., 2002; Morin et al., 2007), type of words to translate (single word terms (SWTs) or multi-word terms (MWTs))(Rapp, 1999; Daille and Morin, 2005), words frequency (less frequent, rare...)(Pekar et al., 2006), etc.

There exist other approaches for bilingual lexicon extraction. Déjean et al. (2002) introduce the *Extended Approach* to avoid the insufficient coverage of the bilingual dictionary required for the translation of source context vectors. A variation of the latter method based on centroid is proposed by Daille and Morin (2005). Haghighi et al. (2008) employ dimension reduction using canonical component analysis (CCA) and Rubino and Linares (2011) propose a multi-view approach based on linear discriminant analysis (LDA) among others.

3 Standard Approach

The *Standard Approach* is based on words co-occurrence vectors. The basic idea is to go through a corpus and to count the number of times $n(c,t)$ each context word c occurs within a window of a certain size w around each target word t. According to (Fung and Mckeown, 1997; Fung, 1998; Rapp, 1999), the *Standard Approach* can be carried out as follows:

For a source word to translate w_i^s, we first build its context vector $v_{w_i^s}$. The vector $v_{w_i^s}$ contains all the words that co-occur with w_i^s within windows of n words. Let's denote by $cooc(w_i^s, w_j^s)$ the co-occurrence value of w_i^s and a given word of its context w_j^s. The process of building context vectors is repeated for all the words of the target language. An association measure such as the point-wise mutual information (Fano, 1961), the log-likelihood (Dunning, 1993) or the discounted odds-ratio (Laroche and Langlais, 2010) is used to score the strength of correlation between a word and all the words of its context vector. The context vector $v_{w_i^s}$ is projected into the target language $v^t_{w_i^s}$. Each word w_j^s of $v_{w_i^s}$ is translated with help of a bilingual dictionary D. If w_j^s is not present in D, w_j^s is discarded. Whenever the bilingual dictionary provides several translations for a word, all the entries are considered but weighted according to their frequency in the target language (Morin et al., 2007). A similarity measure is used to score each target word w_i^t, in the target language with respect to the translated context vector, $v_{w_i^s}^t$. Usual measures of vector similarity include the cosine similarity (Salton and Lesk, 1968) or the weighted Jaccard index (Grefenstette, 1994) for instance. The candidate translations of the word w_i^s are the target words ranked following the similarity score.

4 Method

4.1 Basic Idea

We start from the assumption that words that co-occur together more often than by chance in a small corpus, should have the same behaviour in a bigger corpus with higher co-occurrence values. Our aim is to estimate the increasing values. We choose to observe co-occurrence counts using a training dataset. Table 1 shows the increase of word co-occurrence counts in corpus of different sizes. Let's denote by w_i and w_j two given words. We take as a starting point a corpus of 500,000 words in English, French and Spanish then, for each couple of words (w_i, w_j) that occur together, we observe their co-occurrence count variation in corpus of 1, 2 and 5 million words per language. For instance, if $coocc(w_i, w_j) = 5$ in the English corpus of 500,000 words, we observe $coocc(w_i, w_j)$ in the English corpus of 1, 2 and 5 million words and observe how much this value increases.

Table 1 can be read as follows: Let's take $coocc_{En} > 1$, we can see in Table 1 that there is an increase of 37.19% of words that co-occur more than 1 time in the corpus of 1 million words, 57.06 % in the corpus of 2 million words and 73.02% in the corpus of 5 million words. The observations of Table 1 confirm that word co-occurrence counts increase in most cases (97.34% for $coocc_{Es} > 4$, 98.87% for $coocc_{Fr} > 4$...)

4.2 Co-occurrence Counts Estimation

Let's denote by $E_S = \{v_S^1, v_S^2, ..., v_S^n\}$ the set of the observed co-occurrence counts in a small training corpus. Our aim is to estimate the expected co-occurrence counts $E_L = \{v_L^1, v_L^2, ..., v_L^n\}$ in a large corpus. To do so, one intuitive way for estimation is the mean average increase (MAI) of each co-occurrence count. A more effective model that has proven its efficiency is linear regression. For that reason, we decided to use linear regression ($LReg$) for prediction. In statistical NLP, smoothing techniques for n-gram models have been addressed in a number of studies (Chen and Goodman, 1999). We chose to apply the simple Good-Turing estimator (Good, 1953) as it is an appropriate way to estimate word co-occurrence counts. We finally present a naive model based on the maximum (Max) and the mean average count ($Mean$) of observed word co-occurrence counts in a small and large training datasets.

#Co-occ	$1m$	$2m$	$5m$
$coocc_{En} > 0$	16.13	30.04	47.06
$coocc_{En} = 1$	10.99	23.39	40.66
$coocc_{En} > 1$	37.19	57.06	73.02
$coocc_{En} > 2$	57.50	77.50	88.54
$coocc_{En} > 3$	68.83	85.39	92.63
$coocc_{En} > 4$	77.41	90.79	95.65
$coocc_{En} > 5$	82.11	92.70	96.28
$coocc_{Fr} > 0$	17.74	30.50	47.76
$coocc_{Fr} = 1$	12.55	24.04	41.58
$coocc_{Fr} > 1$	47.84	67.79	83.39
$coocc_{Fr} > 2$	69.28	86.95	95.30
$coocc_{Fr} > 3$	80.81	93.68	98.00
$coocc_{Fr} > 4$	87.93	96.76	98.87
$coocc_{Fr} > 5$	91.30	97.84	99.14
$coocc_{Es} > 0$	18.64	35.50	51.27
$coocc_{Es} = 1$	13.15	28.55	44.91
$coocc_{Es} > 1$	40.99	63.60	76.92
$coocc_{Es} > 2$	60.93	82.93	91.42
$coocc_{Es} > 3$	71.60	89.40	94.80
$coocc_{Es} > 4$	78.91	93.74	97.03
$coocc_{Es} > 5$	83.13	95.06	97.34

Table 1: Word co-occurrence counts increase (%) in corpus of different sizes on the English, French and Spanish Newspapers

4.2.1 Mean Average Increase

Results shown in Table 1 lead to an intuitive model which consists of the estimation of the mean average increase of each co-occurrence count. To estimate E_L we use a training corpus divided in two sets of small (500,000 words) and large (10 million words) corpus. Hence, we estimate the increasing value for each co-occurrence pair count. Let's denote by:
$\mathrm{E_S^1} = \{\mathrm{coocc_S^1(w_i, w_j) = 1, i \in [1, N], j \in [1, M]}\}$ the set of co-occurrence pairs of count 1 observed in a small corpus and by:
$\mathrm{E_L^o} = \{\mathrm{coocc_L^1(w_i, w_j) = o_{ij}, i \in [1, N], j \in [1, M]}\}$ the set of co-occurrence pairs of count o_{ij} observed in a large corpus. The mean average increase MAI_1 for 1 count co-occurrence pairs is:

$$\mathrm{MAI_1} = \frac{1}{|\mathrm{E_S^1}|}\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{M}}(\mathrm{coocc_L^1(w_i,w_j)} - \mathrm{coocc_S^1(w_i,w_j)}) \quad (1)$$

The generalized formula for a given pair co-occurrence count k is:

$$\mathrm{MAI_k} = \frac{1}{|\mathrm{E_S^k}|}\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{M}}(\mathrm{coocc_L^k(w_i,w_j)} - \mathrm{coocc_S^k(w_i,w_j)}) \quad (2)$$

4.2.2 Good-Turing Estimator

Smoothing techniques (Good, 1953) are often used to better estimate probabilities when there is insufficient data to estimate probabilities accurately. They tend to make distributions more uniform, by adjusting low probabilities such as zero probabilities upward, and high probabilities downward. The Good-Turing estimator (Good, 1953) states that for any n-gram that occurs r times, we should pretend that it occurs r^* times. The Good-Turing estimator uses the count of things you have seen once to help estimate the count of things you have never seen. In order to compute the frequency of words, we need to compute N_c, the number of events that occur c times (assumes that all items are binomially distributed). Let N_r be the number of items that occur r times. N_r can be used to provide a better estimate of r, given the binomial distribution. The adjusted frequency r^* is then:

$$\mathrm{r^*} = (\mathrm{r}+1)\frac{\mathrm{N_{r+1}}}{\mathrm{N_r}} \quad (3)$$

The function r^* is applied to all the observed co-occurrence counts of the test data.

4.2.3 Linear Regression

Starting from the observations in Table 1, thanks to linear regression we attempt to model the relationship between the first variable which corresponds to the co-occurrence distribution of words in the small corpus known as the explanatory variable, and the second variable which corresponds to the co-occurrence distribution of words in the large corpus known as the dependent variable. Before applying the linear regression we want to ensure that there is a correlation between the two variables; to do so, we apply the correlation coefficient as presented in Table 2:

Cor	$1m$	$2m$	$5m$
cor_{En}	0.933	0.894	0.788
cor_{Fr}	0.924	0.899	0.872
cor_{Es}	0.904	0.854	0.801

Table 2: Word co-occurrence counts correlation between corpus of 500,000 words and corpus of different sizes (1 million, 2 million and 5 million words) on the English, French and Spanish Newspaper

We can see according to Table 2 that there is a strong correlation of word co-occurrence counts across corpora of different sizes. Let's denote by f the linear function of explanatory variables. We use in our case one explanatory variable X that corresponds to the set of word co-occurrence counts in a small corpus.

- $Y = \beta_1 X + \beta_0$
- For each x of X: $f(x) = \beta_1 x + \beta_0$

By applying linear regression to our training dataset we obtain the following equations:

For the English corpus we obtain:

$$Y_{1m} = 1.742X - 0.686$$

$$Y_{2m} = 3.184X - 2.008$$

$$Y_{5m} = 5.997X - 3.967$$

For the French corpus we obtain:

$$Y_{1m} = 1.802X - 0.673$$

$$Y_{2m} = 3.104X - 1.773$$

$$Y_{5m} = 7.167X - 5.137$$

Where Y_{1m} for instance, corresponds to the linear regression function learned from the training corpus of 1 million words.

4.2.4 Mean and Max Models

As shown in Table 1, co-occurrence counts increase automatically when corpus size increases. A straightforward and maybe naive process is to select the observed counts of co-occurrence pairs in the training large corpus as the new estimation values. Hence, using the mean process, each co-occurrence pair count can be estimated as follows:

$$\mathrm{Mean_k} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{count(k, i)} \qquad (4)$$

Where k is the observed count in the small corpus and i is the observed count in the large corpus of a given words pair. In the same way, using the max process, each co-occurrence pair count is estimated as follows:

$$\mathrm{Max_k} = \frac{1}{\mathrm{N}} \mathrm{MAX}_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{count(k, i)} \qquad (5)$$

5 Linguistic Resources

In order to evaluate the prediction techniques, several linguistic resources are needed. We present hereafter the comparable corpora, the bilingual dictionary and the reference lists used in our experiments.

5.1 Corpus Data

Experiments have been carried out on two English-French comparable corpora. A specialized corpus of 1 million words from the medical domain within the sub-domain of breast cancer and a specialized corpus from the domain of wind energy of 600,000 words.

	Breast cancer	Wind energy
$Tokens_S$	500,000	300,000
$Tokens_T$	500,000	300,000
$\|S\|$	8,221	6,081
$\|T\|$	6,631	5,606

Table 3: Corpus size

For the breast cancer corpus, we have selected the documents from the Elsevier website[1] in order to obtain an English-French specialized comparable corpora. We have automatically selected the documents published between 2001 and 2008 where the title or the keywords contain the term ’cancer du sein’ in French and ’breast cancer’ in English. For the wind energy corpus, we used the *Babook* crawler (Groc, 2011) to collect documents in French and English from the web. As the documents were collected from different websites according to some keywords of the domain, this corpus is more noisy and less well structured comparing to the breast cancer corpus. The two bilingual corpora have been normalized through the following linguistic pre-processing steps: tokenization, part-of-speech tagging, and lemmatization. The function words have been removed and the words occurring once (i.e. hapax) in the French and the English parts have been discarded. As summarized in Table 3, The breast cancer corpus comprised about 8,221 distinct words in English ($|S|$) and 6,631 distinct words in French ($|T|$). The wind energy corpus comprised about 6,081 distinct words in English ($|S|$) and 5,606 distinct words in French ($|T|$).

5.2 Dictionary

We used in our experiments the French-English bilingual dictionary ELRA-M0033 of about 200,000 entries[2]. It contains, after linguistic pre-processing steps and projection on both corpora less than 4000 distinct words. The details are given in Table 4.

	Breast cancer	Wind energy
$\|ELRA_S\|$	3,573	3,459
$\|ELRA_T\|$	3,670	3,326

Table 4: Dictionary coverage

5.3 Reference Lists

To build our reference lists, we selected only the English/French pair of single-word terms (SWTs) which occur more than five times in each part of the comparable corpus. As a result of filtering, 321 English/French SWTs were extracted (from the UMLS[3] meta-thesaurus) for the breast cancer corpus and 100 pairs for the wind energy corpus. The small size of the reference lists can be explained by the fact that small specialized comparable corpora contain a limited set of specialized terms. We can also notice that in bilingual terminology extraction from specialized comparable corpora, the terminology reference list is often composed of 100 SWTs (180 SWTs in (Déjean et al., 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)).

[1] www.elsevier.com

[2] ELRA dictionary has been done by Sciper in the Technolangue/Euradic project

[3] http://www.nlm.nih.gov/research/umls

5.4 Training Dataset

Predicting models such as linear regression or the Good-Turing estimator need a large training corpus to estimate the adjusted co-occurrences. For that purpose, we chose a training corpus composed of two sets. A small set of 500,000 words and a large set of 10 million words. We selected the documents published in 1994 from newspapers *Los Angeles Times/Le Monde*.

6 Experiments and Results

The baseline in our experiments is the *Standard Approach* (Fung, 1998) which is often used for comparison (Pekar et al., 2006; Gamallo, 2008; Prochasson and Morin, 2009), etc. In this section, we first give the parameters of the *standard approach*, than we present the results of the experiments conducted on the two corpora presented above: 'Breast cancer' and 'Wind energy'.

6.1 Experimental Setup

Using the *Standard Approach*, three major parameters need to be set:

1. The size of the window used to build the context vectors (Morin et al., 2007; Gamallo, 2008)

2. The association measure (the log-likelihood (Dunning, 1993), the point-wise mutual information (Fano, 1961), the discounted odds-ratio (Laroche and Langlais, 2010)...)

3. The similarity measure (the weighted Jaccard index (Grefenstette, 1994), the cosine similarity (Salton and Lesk, 1968),...)

Laroche and Langlais (2010) carried out a complete study of the influence of these parameters on the quality of bilingual lexicon extraction from comparable corpora. To build the context vectors we chose a 7-window size. The entries of the context vectors were determined by the log-likelihood, the point-wise mutual information and the discounted odds-ratio. As similarity measure, we chose to use the weighted Jaccard index and the cosine similarity. Other combinations of parameters were assessed but the previous parameters turned out to give the best performance.

We note that *Top k* means that the correct translation of a given word is present in the k first candidates of the list returned by the *Standard Approach*. We use also the mean average precision *MAP* (Manning and Schutze, 2008) which represents the quality of the system.

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{m_i=1}^{k}P(R_{ik}) \qquad (6)$$

where $|Q|$ is the number of terms to be translated, m_i is the number of reference translations for the i^{th} term (always 1 in our case), and $P(R_{ik})$ is 0 if the reference translation is not found for the i^{th} term or $1/r$ if it is (r is the rank of the reference translation in the translation candidates).

6.2 Results

We conducted a set of two experiments on two specialized comparable corpora. We carried out a comparison between the *Standard Approach* (SA) and the different prediction models presented in section 4.2 namely: the maximum model (Max), the mean model ($Mean$), the linear regression model ($LReg$), The Good-Turing estimator (GT) and the mean average increase model (MAI). Experiment 1 shows the results on the breast cancer corpus and experiment 2 those of the wind energy corpus.

Table 5 shows the results of the experiments on the breast cancer corpus. The first observation concerns the *Standard Approach* (SA). The best results are obtained using the Log-Jac parameters with a MAP of 27.9%. We can also notice that for this configuration, none of the prediction models improve the performance of the *Standard Approach*. On the contrary, they even degrade the results. The second observation concerns the Odds-Cos parameters where the naive $Mean$, Max and MAI models are under the baseline. The best score is obtained by the $LReg$ model with a MAP of 27.6%. The most notable result concerns the PMI-Cos parameters. We can notice that four of the five techniques improve the performance of the baseline. The best prediction model is the Max technique which reaches a MAP of 27.2% and improves the Top1 precision of 4.8% and the Top10 precision of 6.6%.

Table 6 shows the results of the experiments on the wind energy corpus. Generally the results follow the same behaviour as the previous experiment. The best results of the *Standard Approach* are obtained using the Log-Jac parameters with a MAP of 25.7%. Here also, none of the prediction models improve the performance of the *Standard Approach*. About the Odds-Cos parameters,

	SA	Max	Mean	LReg	MAI	GT	
P1	15.5	**20.2**	13.7	18.0	18.6	18.6	PMI-Cos
P5	31.1	**35.8**	28.3	**35.8**	34.2	32.0	
P10	34.5	41.1	32.7	**42.0**	38.3	37.0	
MAP	22.6	**27.2**	20.3	26.7	26.4	25.6	
P1	15.8	15.5	11.8	**19.9**	13.7	16.8	Odds-Cos
P5	**34.8**	30.2	28.6	34.2	27.7	34.2	
P10	40.4	36.7	35.5	**41.7**	33.0	39.8	
MAP	24.8	22.9	19.8	**27.6**	20.9	25.2	
P1	**20.2**	06.5	16.5	15.5	09.9	14.6	Log-Jac
P5	**35.8**	15.5	33.9	28.6	21.4	27.7	
P10	**42.6**	20.5	38.3	37.3	26.7	34.2	
MAP	**27.9**	11.6	24.6	22.6	15.6	21.4	

Table 5: Results of the experiments on the 'Breast cancer' corpus (the improvements indicate a significance at the 0.05 level using Student's t-test).

	SA	Max	Mean	LReg	MAI	GT	
P1	07.0	13.0	10.0	**18.0**	15.3	14.0	PMI-Cos
P5	27.0	34.0	30.0	**37.0**	33.0	31.0	
P10	37.0	**46.0**	36.0	**46.0**	43.0	43.0	
MAP	17.8	23.1	19.2	**28.0**	25.0	22.9	
P1	12.0	09.0	06.0	**14.0**	10.0	12.0	Odds-Cos
P5	31.0	20.0	27.0	**32.0**	25.0	31.0	
P10	38.0	26.0	39.0	**40.0**	33.0	36.0	
MAP	21.8	15.7	17.0	**23.3**	18.0	19.8	
P1	**17.0**	09.0	18.0	15.0	18.0	13.0	Log-Jac
P5	**36.0**	16.0	30.0	31.0	29.0	27.0	
P10	**42.0**	22.0	45.0	36.0	36.0	37.0	
MAP	**25.7**	14.0	25.1	22.9	23.7	20.5	

Table 6: Results of the experiments on the 'Wind energy' corpus (the improvements indicate a significance at the 0.05 level using Student's t-test).

here again the naive $Mean$, Max and MAI models are under the baseline. We can also notice that GT is slightly under the *standard approach*. The best score is obtained by the $LReg$ model with a MAP of 23.3%. Finally, the most remarkable result still concerns the PMI-Cos parameters where the same four of the five predicting techniques improve the performance of the baseline. The best prediction model is the $LReg$ technique which reaches a MAP of 28.0% and improves the Top1 precision of 11.0% and the Top10 precision of 10.2%.

7 Discussion

The aim of this work was to propose and contrast different word co-occurrence prediction approaches: naive or intuitive (Max, $Mean$ and MAI) and more sophisticated ($LReg$ and GT) aiming to improve bilingual terminology extraction. Our approach can be used as a pre-processing step of the *Standard Approach* by applying a prediction function to word co-occurrence counts.

According to the experimental results, the first observation is that the *Standard Approach* performs better when using the log-likelihood measure comparatively to the discounted odds-ratio and the point-wise mutual information measures. This supposes that the log-likelihood provides a better estimation of word co-occurrence counts. The log-likelihood measures significance (i.e. the amount of evidence against the null hypothesis) and is known to be more robust against low expected frequencies (Dunning, 1993). The lower performance of the *Standard Approach* when using the point-wise mutual information is certainly due to the over-estimation of low frequencies. In practical applications, PMI was found to have a tendency to assign inflated scores to low-frequency word pairs. Thus, even a single co-occurrence of two words might result in a fairly high association score. The discounted odds-ratio has shown lower results when compared to log-likelihood unlike its better performance as shown in Laroche and Langlais (2010). This is certainly due to the multiple parameters and resources of the *Standard Approach* and also the cosine similarity measure which is sensitive to context vector size. In our experiments, we did not investigate this parameter as it is not the matter of our study. We considered the whole context vector of each word.

According to the PMI-Cos configuration, the baseline is consistently outperformed by every prediction model (except $Mean$ on the breast cancer experiment). The good results of the proposed methods when associated to the PMI-Cos configuration suggest that the over-estimation of infrequent counts of PMI is skimmed by the prediction function. This finding can be considered as a new way to counterbalance the low-frequency bias of PMI. The best prediction approach shown in the experiments is Max with a MAP of 27.2%, followed by $LReg$ with a MAP of 26.7% on the breast cancer corpus. Nevertheless, in the wind energy corpus $LReg$ performed substantially better than Max with a MAP of 28.0% while Max reaches 23.1% only. Even the lower performance of MAI and GT, they also provide significant improvements.

In our experiments, none of the proposed algorithms reached good results while associated to the Log-Jac configuration. This is certainly related to the properties of the log-likelihood association measure. While the prediction models tend to increase small co-occurrence counts, this can lead to the overrating of infrequent words renders the ranking of the log-likelihood measure useless. Concerning the Odds-Cos parameters, although there were slight improvements on the $LReg$ algorithm, other methods have shown disappointing results. Here again the Odds-ratio association measure seems to be not compatible with re-estimating co-occurrence counts. More investigations are certainly needed to highlight the reasons of this poor performance. It seems that prediction functions do not fit well with association measures based on contingency table.

The most noticeable improvement concerns the PMI-Cos configuration. Aside from the $Mean$ method, all the other techniques have shown better performance than the *Standard Approach*. According to the empirical results, point-wise mutual information performs better with Max and $LReg$ techniques. Furthermore and as has been pointed out above, prediction models seem to be an alternative to the low-frequency bias of the point-wise mutual information. It is our hope that the present work may provide a starting point to co-occurrence prediction on comparable corpora as an alternative to unreliable counts. The next step is to explore more complex prediction models such as nonlinear regression that intuitively should fit better than a simple linear regression and to contrast our prediction function with the various suggested heuristics for correcting PMI bias.

8 Conclusion

In this paper, we have described and compared different prediction models for the task of bilingual terminology extraction from comparable corpora. Our belief is that word co-occurrence counts prediction can be an alternative to the unreliable counts observed in small corpora. The results demonstrate the viability of the proposed approach using the PMI-Cos configuration. If more investigation is certainly needed for the Odds-Cos and Log-Jac configurations, the empirical results of our proposition suggest that predicting word co-occurrence counts is an appropriate way to improve the accuracy of the *Standard Approach* in small specialized comparable corpora.

Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-12-CORD-0020.

References

Y. Cao and H. Li. 2002. Base noun phrase translation using web data and the em algorithm. *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Tapei, Taiwan*, pages 127–133.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

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 (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Béatrice Daille and Emmanuel Morin. 2005. French-English Terminology Extraction from Comparable Corpora. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCLNP'05)*, pages 707–718, Jeju Island, Korea.

Hervé Déjean, Fatia Sadat, and Éric Gaussier. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 218–224, Tapei, Taiwan.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Stefan Evert and Marco Baroni. 2007. zipfr: Word frequency modeling in r. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic.

Robert M. Fano. 1961. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC97)*, page 192202, Hong Kong.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries From a non-Parallel English-Chinese Corpus. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'95)*, pages 1–16, Langhorne, PA, USA.

Pascale Fung. 1998. A Statistical View on Bilingual Lexicon Extraction: From ParallelCorpora to Nonparallel Corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, pages 1–16, Langhorne, PA, USA.

Otero Gamallo. 2008. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, pages 19–26, Marrakech, Marroco.

I. J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Clément De Groc. 2011. Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *Proceedings of The IEEE-WICACM International Conferences on Web Intelligence*, pages 497–498, Lyon, France.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 771–779, Columbus, Ohio, USA.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Raghavan; Manning, Christopher D.; Prabhakar and Hinrich Schutze. 2008. Introduction to information retrieval. Cambridge University Press.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Viktor Pekar, Ruslan Mitkov, Dimitar Blagoev, and Andrea Mulloni. 2006. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Emmanuel Prochasson and Emmanuel Morin. 2009. Anchor points for bilingual extraction from small specialized comparable corpora. *TAL*, 50(1):283–304.

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 (ACL'99)*, pages 519–526, College Park, MD, USA.

Raphaël Rubino and Georges Linares. 2011. A multi-view approach for term translation spotting. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'11)*, pages 29–40, Tokyo, Japan.

Gerard Salton and Michael E. Lesk. 1968. Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.