# A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora

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

## Abstract

Smoothing is a central issue in language modeling and a prior step in different natural language processing (NLP) tasks. However, less attention has been given to it for bilingual lexicon extraction from comparable corpora. If a first work to improve the extraction of low frequency words showed significant improvement while using distance-based averaging (Pekar et al., 2006), no investigation of the many smoothing techniques has been carried out so far. In this paper, we present a study of some widely-used smoothing algorithms for language n-gram modeling (Laplace, Good-Turing, Kneser-Ney...). Our main contribution is to investigate how the different smoothing techniques affect the performance of the standard approach (Fung, 1998) traditionally used for bilingual lexicon extraction. We show that using smoothing as a pre-processing step of the standard approach increases its performance significantly.

## 1 Introduction

Cooccurrences play an important role in many corpus based approaches in the field of natural-language processing (Dagan et al., 1993). They represent the observable evidence that can be distilled from a corpus and are employed for a variety of applications such as machine translation (Brown et al., 1992), information retrieval (Maarek and Smadja, 1989), word sense disambiguation (Brown et al., 1991), etc. In bilingual lexicon extraction from comparable corpora, frequency counts for word pairs often serve as a basis for distributional methods, such as the standard approach (Fung, 1998) which compares the cooccurrence profile of a given source word, a vector of association scores for its translated cooccurrences (Fano, 1961; Dunning, 1993), with the profiles of all words of the target language. The distance between two such vectors is interpreted as an indicator of their semantic similarity and their translational relation. If using association measures to extract word translation equivalents has shown a better performance than using a raw cooccurrence model, the latter remains the core of any statistical generalisation (Evert, 2005).

As has been known, words and other type-rich linguistic populations do not contain instances of all types in the population, even the largest samples (Zipf, 1949; Evert and Baroni, 2007). Therefore, the number and distribution of types in the available sample are not reliable estimators (Evert and Baroni, 2007), especially for small comparable corpora. The literature suggests two major approaches for solving the data sparseness problem: smoothing and class-based methods. 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. Generally, smoothing methods not only prevent zero probabilities, but they also attempt to improve the accuracy of the model as a whole (Chen and Goodman, 1999). Class-based models (Pereira et al., 1993) use classes of similar words to distinguish between unseen cooccurrences. The relationship between given words is modeled by analogy with other words that are in some sense similar to the given ones. Hence, class-based models provide an alternative to the independence assumption on the cooccurrence of given words $w_1$ and $w_2$: the more frequent $w_2$ is, the higher estimate of $P(w_2|w_1)$ will be, regardless of $w_1$.

*Proceedings of the 6th Workshop on Building and Using Comparable Corpora*, pages 24–33,
Sofia, Bulgaria, August 8, 2013. ©2013 Association for Computational Linguistics

Starting from the observation that smoothing estimates ignore the expected degree of association between words (assign the same estimate for all unseen cooccurrences) and that class-based models may not structure and generalize word cooccurrence to class cooccurrence patterns without losing too much information, (Dagan et al., 1993) proposed an alternative to these latter approaches to estimate the probabilities of unseen cooccurrences. They presented a method that makes analogies between each specific unseen cooccurrence and other cooccurrences that contain similar words. The analogies are based on the assumption that similar word cooccurrences have similar values of mutual information. Their method has shown significant improvement for both: word sense disambiguation in machine translation and data recovery tasks. (Pekar et al., 2006) employed the nearest neighbor variety of the previous approach to extract translation equivalents for low frequency words from comparable corpora. They used a distance-based averaging technique for smoothing (Dagan et al., 1999). Their method yielded a significant improvement in relation to low frequency words.

Starting from the assumption that smoothing improves the accuracy of the model as a whole (Chen and Goodman, 1999), we believe that smoothed context vectors should lead to better performance for bilingual terminology extraction from comparable corpora. In this work we carry out an empirical comparison of the most widely-used smoothing techniques, including additive smoothing (Lidstone, 1920), Good-Turing estimate (Good, 1953), Jelinek-Mercer (Mercer, 1980), Katz (Katz, 1987) and kneser-Ney smoothing (Kneser and Ney, 1995). Unlike (Pekar et al., 2006), the present work does not investigate unseen words. We only concentrate on observed cooccurrences. We believe it constitutes the most systematic comparison made so far with different smoothing techniques for aligning translation equivalents from comparable corpora. We show that using smoothing as a pre-processing step of the standard approach, leads to significant improvement even without considering unseen cooccurrences.

In the remainder of this paper, we present in Section 2, the different smoothing techniques. The steps of the standard approach and our extended method are then described in Section 3. Section 4 describes the experimental setup and our resources. Section 5 presents the experiments and comments on several results. We finally discuss the results in Section 6 and conclude in Section 7.

## 2 Smoothing Techniques

Smoothing describes techniques for adjusting the maximum likelihood estimate of probabilities to reduce more accurate probabilities. The smoothing techniques tend to make distributions more uniform. In this section we present the most widely used techniques.

### 2.1 Additive Smoothing

The Laplace estimator or the additive smoothing (Lidstone, 1920; Johnson, 1932; Jeffreys, 1948) is one of the simplest types of smoothing. Its principle is to estimate probabilities $P$ assuming that each unseen word type actually occurred once. Then, if we have $N$ events and $V$ possible words instead of :

$$\mathrm{P(w)} = \frac{\mathrm{occ(w)}}{\mathrm{N}} \tag{1}$$

We estimate:

$$\mathrm{P_{addone}(w)} = \frac{\mathrm{occ(w)} + 1}{\mathrm{N} + \mathrm{V}} \tag{2}$$

Applying Laplace estimation to word's cooccurrence suppose that : if two words cooccur together $n$ times in a corpus, they can cooccur together $(n + 1)$ times. According to the maximum likelihood estimation (MLE):

$$\mathrm{P(w_{i+1}|w_i)} = \frac{\mathrm{C(w_i, w_{i+1})}}{\mathrm{C(w_i)}} \tag{3}$$

Laplace smoothing:

$$\mathrm{P^*(w_{i+1}|w_i)} = \frac{\mathrm{C(w_i, w_{i+1})} + 1}{\mathrm{C(w_i)} + \mathrm{V}} \tag{4}$$

Several disadvantages emanate from this method:

1. The probability of frequent n-grams is underestimated.

2. The probability of rare or unseen n-grams is overestimated.

3. All the unseen n-grams are smoothed in the same way.
4. Too much probability mass is shifted towards unseen n-grams.

One improvement is to use smaller added count following the equation below:

$$\mathrm{P}^*(\mathrm{w_{i+1}}|\mathrm{w_i}) = \frac{\delta + \mathrm{C}(\mathrm{w_i}, \mathrm{w_{i+1}})}{\delta|\mathrm{V}| + \mathrm{C}(\mathrm{w_i})} \tag{5}$$

with $\delta \in ]0, 1]$.

### 2.2 Good-Turing Estimator

The Good-Turing estimator (Good, 1953) provides another way to smooth probabilities. It states that for any n-gram that occurs $r$ times, we should pretend that it occurs $r^*$ times. The Good-Turing estimators use 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 (assuming 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:

$$\mathrm{r}^* = (\mathrm{r} + 1)\frac{\mathrm{N_{r+1}}}{\mathrm{N_r}} \tag{6}$$

### 2.3 Jelinek-Mercer Smoothing

As one alternative to missing n-grams, useful information can be provided by the corresponding (n-1)-gram probability estimate. A simple method for combining the information from lower-order n-gram in estimating higher-order probabilities is linear interpolation (Mercer, 1980). The equation of linear interpolation is given below:

$$\mathrm{P_{int}}(\mathrm{w_{i+1}}|\mathrm{w_i}) = \lambda \mathrm{P}(\mathrm{w_{i+1}}|\mathrm{w_i}) + (1 - \lambda)\mathrm{P}(\mathrm{w_i}) \tag{7}$$

$\lambda$ is the confidence weight for the longer n-gram. In general, $\lambda$ is learned from a held-out corpus. It is useful to interpolate higher-order n-gram models with lower-order n-gram models, because when there is insufficient data to estimate a probability in the higher order model, the lower-order model can often provide useful information. Instead of the cooccurrence counts, we used the Good-Turing estimator in the linear interpolation as follows:

$$\mathrm{c^*_{int}}(\mathrm{w_{i+1}}|\mathrm{w_i}) = \lambda \mathrm{c}^*(\mathrm{w_{i+1}}|\mathrm{w_i}) + (1 - \lambda)\mathrm{P}(\mathrm{w_i}) \tag{8}$$

### 2.4 Katz Smoothing

(Katz, 1987) extends the intuitions of Good-Turing estimate by adding the combination of higher-order models with lower-order models. For a bigram $w_{i-1}^{i}$ with count $r = c(w_{i-1}^{i})$, its corrected count is given by the equation:

$$\mathrm{c_{katz}}(\mathrm{w_{i-1}^{i}}) = \begin{cases} \mathrm{r}^* & \text{if } \mathrm{r} > 0 \\ \alpha(\mathrm{w_{i-1}})\mathrm{PML}(\mathrm{w_i}) & \text{if } \mathrm{r} = 0 \end{cases} \tag{9}$$

and:

$$\alpha(\mathrm{w_{i-1}}) = \frac{1 - \sum_{\mathrm{w_i : c(w_{i-1}^{i}) > 0}} \mathrm{P_{katz}}(\mathrm{w_{i-1}^{i}})}{1 - \sum_{\mathrm{w_i : c(w_{i-1}^{i}) > 0}} \mathrm{PML}(\mathrm{w_{i-1}})} \tag{10}$$

According to (Katz, 1987), the general discounted estimate $c^*$ of Good-Turing is not used for all counts $c$. Large counts where $c > k$ for some threshold $k$ are assumed to be reliable. (Katz, 1987) suggests $k = 5$. Thus, we define $c^* = c$ for $c > k$, and:

$$\mathrm{c}^* = \frac{(\mathrm{c} + 1)\frac{\mathrm{N_{c+1}}}{\mathrm{N_c}} - \mathrm{c}\frac{\mathrm{(k+1)N_{k+1}}}{\mathrm{N_1}}}{1 - \frac{\mathrm{(k+1)N_{k+1}}}{\mathrm{N_1}}} \tag{11}$$

### 2.5 Kneser-Ney Smoothing

Kneser-Ney have introduced an extension of absolute discounting (Kneser and Ney, 1995). The estimate of the higher-order distribution is created by subtracting a fixed discount D from each non-zero count. The difference with the absolute discounting smoothing resides in the estimate of the lower-order distribution as shown in the following equation:

$$r = \begin{cases} \frac{\mathrm{Max(c(w_{i-n+1}^{i}) - D, 0)}}{\sum_{w_i} c(w_{i-n+1}^{i})} & \text{if } \mathrm{c(w_{i-n+1}^{i})} > 0 \\ \alpha(\mathrm{w_{i-n+1}^{i-1}})\mathrm{P_{kn}}(\mathrm{w_i}|\mathrm{w_{i-n+2}^{i-1}}) & \text{if } \mathrm{c(w_{i-n+1}^{i})} = 0 \end{cases} \tag{12}$$

where $r = \mathrm{P_{kn}}(\mathrm{w_i}|\mathrm{w_{i-n+1}^{i-1}})$ and $\alpha(w_{i-n+1}^{i-1})$ is chosen to make the distribution sum to 1 (Chen and Goodman, 1999).

## 3 Methods

In this section we first introduce the different steps of the standard approach, then we present our extended approach that makes use of smoothing as a new step in the process of the standard approach.

### 3.1 Standard Approach

The main idea for identifying translations of terms in comparable corpora relies on the distributional hypothesis [1] that has been extended to the bilingual scenario (Fung, 1998; Rapp, 1999). If many variants of the standard approach have been proposed (Chiao and Zweigenbaum, 2002; Hervé Déjean and Gaussier, 2002; Morin et al., 2007; Gamallo, 2008)[among others], they mainly differ in the way they implement each step and define its parameters. The standard approach can be carried out as follows:

**Step 1** 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 cooccur with $w_i^s$ within windows of $n$ words. Lets denote by $cooc(w_i^s, w_j^s)$ the cooccurrence 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.

**Step 2** An association measure such as the pointwise 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.

**Step 3** The context vector $v_{w_i^s}$ is projected into the target language $v_{w_i^s}^t$. Each word $w_j^s$ of $v_{w_i^s}$ is translated with the 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).

**Step 4** 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 (WJ) (Grefenstette, 1994) for instance. The candidate translations of the word $w_i^s$ are the target words ranked following the similarity score.

### 3.2 Extended Approach

We aim at investigating the impact of different smoothing techniques for the task of bilingual terminology extraction from comparable corpora. Starting from the assumption that word cooccurrences are not reliable especially for small corpora (Zipf, 1949; Evert and Baroni, 2007) and that smoothing is usually used to counteract this problem, we apply smoothing as a preprocessing step of the standard approach. Each $cooc(w_i^s, w_j^s)$ is smoothed according to the techniques described in Section 2. The smoothed cooccurrence $cooc^*(w_i^s, w_j^s)$ is then used for calculating the association measure between $w_i^s$ and $w_j^s$ and so on (steps 2, 3 and 4 of the standard approach are unchanged). We chose not to study the prediction of unseen cooccurrences. The latter has been carried out successfully by (Pekar et al., 2006). We concentrate on the evaluation of smoothing techniques of known cooccurrences and their effect according to different association and similarity measures.

## 4 Experimental Setup

In order to evaluate the smoothing techniques, several resources and parameters are needed. We present hereafter the experiment data and the parameters of the standard approach.

### 4.1 Corpus Data

The experiments have been carried out on two English-French comparable corpora. A specialized corpus of 530,000 words from the medical domain within the sub-domain of 'breast cancer' and a specialize corpus from the domain of 'wind-energy' of 300,000 words. 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. For the breast cancer corpus, we have selected the documents from the Elsevier website[2] 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. We collected 130 documents in French and 118 in

[1] words with similar meaning tend to occur in similar contexts

[2] www.elsevier.com

English. As summarised in Table 1, The comparable corpora comprised about 6631 distinct words in French and 8221 in English. For the wind energy corpus, we used the *Babook* crawler (Groc, 2011) to collect documents in French and English from the web. We could only obtain 50 documents in French and 65 in English. As the documents were collected from different websites according to some keywords of the domain, this corpus is more noisy and not well structured compared to the breast cancer corpus. The wind-energy corpus comprised about 5606 distinct words in French and 6081 in English.

| | Breast cancer | Wind energy |
|---|---|---|
| $Tokens_S$ | 527,705 | 307,996 |
| $Tokens_T$ | 531,626 | 314,551 |
| $\|S\|$ | 8,221 | 6,081 |
| $\|T\|$ | 6,631 | 5,606 |

Table 1: Corpus size

### 4.2 Dictionary

In our experiments we used the French-English bilingual dictionary ELRA-M0033 of about 200,000 entries[3]. It contains, after linguistic preprocessing steps and projection on both corpora fewer than 4000 single words. The details are given in Table 2.

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

Table 2: Dictionary coverage

### 4.3 Reference Lists

In bilingual terminology extraction from specialized comparable corpora, the terminology reference list required to evaluate the performance of the alignment programs is often composed of 100 single-word terms (SWTs) (180 SWTs in (Hervé Déjean and Gaussier, 2002), 95 SWTs in (Chiao and Zweigenbaum, 2002), and 100 SWTs in (Daille and Morin, 2005)). To build our reference lists, we selected only the French/English pair of SWTs which occur more than five times in each part of the comparable corpus. As a result of filtering, 321 French/English SWTs were extracted (from the UMLS[4] meta-thesaurus.) for the breast cancer corpus, and 100 pairs for the wind-energy corpus.

### 4.4 Evaluation Measure

Three major parameters need to be set to the standard approach, namely the similarity measure, the association measure defining the entry vectors and the size of the window used to build the context vectors. (Laroche and Langlais, 2010) carried out a complete study of the influence of these parameters on the quality of bilingual alignment. As a similarity measure, we chose to use Weighted Jaccard Index (Grefenstette, 1994) and Cosine similarity (Salton and Lesk, 1968). The entries of the context vectors were determined by the log-likelihood (Dunning, 1993), mutual information (Fano, 1961) and the discounted Odds-ratio (Laroche and Langlais, 2010). We also chose a 7-window size. 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 et al., 2008) which represents the quality of the system.

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{m_i} \sum_{m_i=1}^{k} P(R_{ik}) \qquad (13)$$

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).

### 4.5 Baseline

The baseline in our experiments is the standard approach (Fung, 1998) without any smoothing of the data. The standard approach is often used for comparison (Pekar et al., 2006; Gamallo, 2008; Prochasson and Morin, 2009), etc.

### 4.6 Training Data Set

Some smoothing techniques such as the Good-Turing estimators need a large training corpus to

[3] ELRA dictionary has been created by Sciper in the Technolangue/Euradic project

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

estimate the adjusted cooccurrences. For that purpose, we chose a training general corpus of 10 million words. We selected the documents published in 1994 from the 'Los Angeles Times/Le Monde' newspapers.

## 5 Experiments and Results

We conducted a set of three experiments on two specialized comparable corpora. We carried out a comparison between the standard approach (SA) and the smoothing techniques presented in Section 2 namely : additive smoothing (Add1), Good-Turing smoothing (GT), the Jelinek-Mercer technique (JM), the Katz-Backoff (Katz) and kneser-Ney smoothing (Kney). Experiment 1 shows the results for the breast cancer corpus. Experiment 2 shows the results for the wind energy corpus and finally experiment 3 presents a comparison of the best configurations on both corpora.

### 5.1 Experiment 1

Table 3 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 = 27.9%. We can also notice that for this configuration, only the Additive smoothing significantly improves the performance of the standard approach with a MAP = 30.6%. The other smoothing techniques even degrade the results. The second observation concerns the Odds-Cos parameters where none of the smoothing techniques significantly improved the performance of the baseline (SA). Although Good-Turing and Katz-Backoff smoothing give slightly better results with respectively a MAP = 25.2 % and MAP = 25.3 %, these results are not significant. The most notable result concerns the PMI-COS parameters. We can notice that four of the five smoothing techniques improve the performance of the baseline. The best smoothing is the Jelinek-Mercer technique which reaches a MAP = 29.5% and improves the Top1 precision of 6% and the Top10 precision of 10.3%.

| | SA | Add1 | GT | JM | Katz | Kney | |
|---|---|---|---|---|---|---|---|
| P1 | 15.5 | 17.1 | 18.7 | **21.5** | 18.7 | 05.3 | PMI-Cos |
| P5 | 31.1 | 32.7 | 32.0 | **38.3** | 33.9 | 13.4 | |
| P10 | 34.5 | 37.0 | 37.0 | **44.8** | 38.0 | 15.2 | |
| MAP | 22.6 | 24.8 | 25.6 | **29.5** | 25.9 | 09.1 | |
| P1 | 15.8 | 16.1 | 16.8 | 14.6 | **17.1** | 09.0 | Odds-Cos |
| P5 | **34.8** | 33.6 | 34.2 | 33.0 | 33.9 | 19.6 | |
| P10 | 40.4 | **41.7** | 39.8 | 38.3 | 40.1 | 25.2 | |
| MAP | 24.8 | 24.4 | 25.2 | 23.3 | **25.3** | 14.1 | |
| P1 | 20.2 | **22.4** | 14.6 | 14.6 | 14.6 | 16.2 | Log-Jac |
| P5 | 35.8 | **40.5** | 27.7 | 26.7 | 26.7 | 29.9 | |
| P10 | 42.6 | **44.2** | 34.2 | 33.3 | 33.0 | 33.9 | |
| MAP | 27.9 | **30.6** | 21.4 | 21.2 | 21.2 | 22.9 | |

Table 3: Results of the experiments on the "Breast cancer" corpus (except the Odds-Cos configuration, the improvements indicate a significance at the 0.05 level using the Student t-test).

| | SA | Add1 | GT | JM | Katz | Kney | |
|---|---|---|---|---|---|---|---|
| P1 | 07.0 | 14.0 | 14.0 | **21.0** | 16.0 | 09.0 | PMI-Cos |
| P5 | 27.0 | 32.0 | 31.0 | **37.0** | 30.0 | 17.0 | |
| P10 | 37.0 | 42.0 | 43.0 | **51.0** | 44.0 | 28.0 | |
| MAP | 17.8 | 23.6 | 22.9 | **30.1** | 24.2 | 14.1 | |
| P1 | 12.0 | **17.0** | 12.0 | 12.0 | 12.0 | 06.0 | Odds-Cos |
| P5 | 31.0 | **35.0** | 31.0 | 32.0 | 28.0 | 16.0 | |
| P10 | 38.0 | **44.0** | 36.0 | 39.0 | 35.0 | 21.0 | |
| MAP | 21.8 | **26.5** | 19.8 | 20.8 | 19.7 | 11.1 | |
| P1 | 17.0 | **22.0** | 13.0 | 13.0 | 13.0 | 14.0 | Log-Jac |
| P5 | 36.0 | **38.0** | 27.0 | 27.0 | 27.0 | 29.0 | |
| P10 | 42.0 | **50.0** | 37.0 | 38.0 | 38.0 | 39.0 | |
| MAP | 25.7 | **29.7** | 20.5 | 21.3 | 21.3 | 22.9 | |

Table 4: Results of the experiments on the "Wind Energy" corpus (except the Odds-Cos configuration, the improvements indicate a significance at the 0.05 level using the Student t-test).

### 5.2 Experiment 2

Table 4 shows the results of the experiments on the wind energy corpus. Generally the results exhibit the same behaviour as the previous experiment. The best results of the standard approach are obtained using the Log-Jac parameters with a MAP = 25.7%. Here also, only the Additive smoothing significantly improves the performance of the standard approach with a MAP = 39.7%. The other smoothing techniques also degrade the results. About the Odds-Cos parameters, except the additive smoothing, here again none of the smoothing techniques significantly improved the performance of the baseline. Finally the most remarkable result still concerns the PMI-COS parameters where the same four of the five smoothing techniques improve the performance of the baseline. The best smoothing is the Jelinek-Mercer technique which reaches a MAP = 30.1% and improves the Top1 and and the Top10 precisions by 14.0%.

### 5.3 Experiment 3

In this experiment, we would like to investigate whether the smoothing techniques are more efficient for frequent translation equivalents or less frequent ones. For that purpose, we split the breast cancer reference list of 321 entries into two sets of translation pairs. A set of 133 frequent pairs named : *High-test set* and a set of 188 less frequent pairs called *Low-test set*. The initial reference list of 321 pairs is the *Full-test set*. We consider frequent pairs those of a frequency higher than 100. We chose to analyse the two configurations that provided the best performance that is : Log-Jac and Pmi-Cos parameters according to the *Full-test*, *High-test* and *Low-test* sets.

Figure 1 shows the results using the Log-Jac configuration. We can see that the additive smoothing always outperforms the standard approach for all the test sets. The other smoothing techniques are always under the baseline and behave approximately the same way. Figure 2 shows the results using the PMI-COS configuration. We can see that except the Kneser-Ney smoothing, all the smoothing techniques outperform the standard approach for all the test sets. We can also notice that the Jelinek-Mercer smoothing improves more notably the *High-test* set.

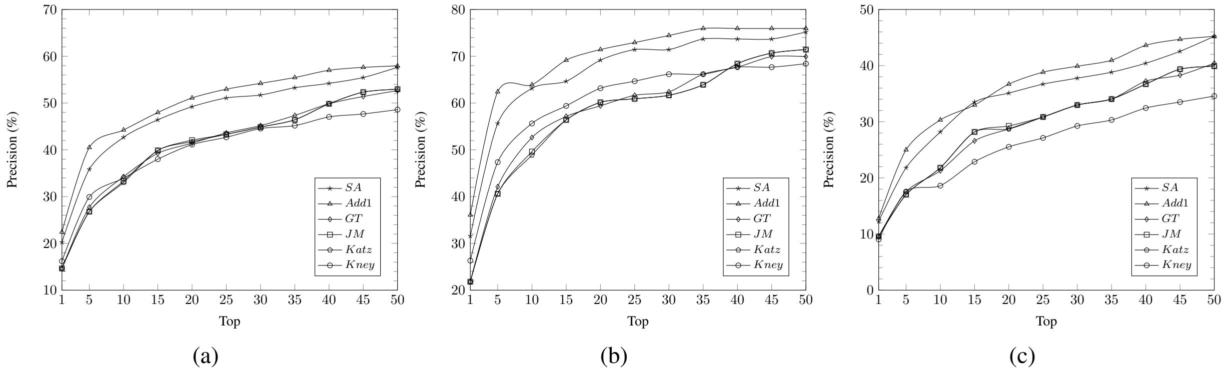
## 6 Discussion

Smoothing techniques are often evaluated on their ability to predict unseen n-grams. In our experiments we only focused on smoothing observed cooccurrences of context vectors. Hence, the previous evaluations of smoothing techniques may not always be consistent with our findings. This is for example the case for the additive smoothing technique. The latter which is described as a poor estimator in statistical NLP, turns out to perform well when associated with the Log-Jac parameters. This is because we did not consider unseen cooccurences which are over estimated by the Add-one smoothing. Obviously, we can imagine that adding one to all unobserved cooccurrences would not make sense and would certainly degrade the results. Except the add-one smoothing, none of the other algorithms reached good results when associated to the Log-Jac configuration. This is certainly related to the properties of the log-likelihood association measure. Additive smoothing has been used to address the problem of rare words aligning to too many words (Moore, 2004). At each iteration of the standard Expectation-Maximization (EM) procedure all the translation probability estimates are smoothed by adding virtual counts to uniform probability distribution over all target words. Here also additive smoothing has shown interesting results. According to these findings, we can consider the additive smoothing as an appropriate technique for our task.

Concerning the Odds-Cos parameters, although there have been slight improvements in the add-one algorithm, smoothing techniques have shown disappointing results. Here again the Odds-ratio association measure seems to be incompatible with re-estimating small cooccurrences. More investigations are certainly needed to highlight the reasons for this poor performance. It seems that smoothing techniques based on discounting does not fit well with association measures based on contingency table. The most noticeable improvement concerns the PMI-Cos configurations. Except Kneser-Ney smoothing, all the other techniques showed better performance than the standard approach. According to the results, point-wise mutual information performs better with smoothing techniques especially with the linear interpolation of Jelinek-Mercer method that combines high-order (cooccurrences) and low-order (unigrams) counts of the Good-Turing estimations. Jelinek-Mercer smoothing counteracts the disadvantage of the point-wise mutual information which consists of over estimating less frequent words. This latter weakness is corrected first by the Good-Turing estimators and then by considering the low order counts. The best performance was obtained with $\lambda = 0.5$.

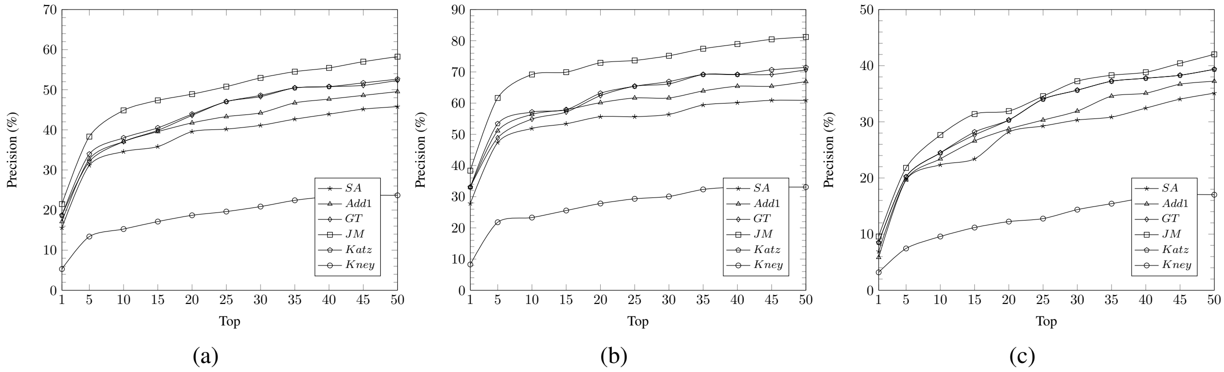
Smoothing techniques attempt to improve the accuracy of the model as a whole. This particularity has been confirmed by the third experiment where we noticed the smoothing improvements for both reference lists, that is the *High-test* and *Low-test* sets. This latter experiment has shown that smoothing observed cooccurrences is useful for all frequency ranges. The difference of precision between the two test lists can be explained by the fact that less frequent words are harder to translate.

In statistical NLP, smoothing techniques for n-gram models have been addressed in a number of studies (Chen and Goodman, 1999). The ad-



(a) (b) (c)

Figure 1: A set of three figures on the breast cancer corpus for the Log-Jac configuration : (a) Full-test set ; (b) High-test set; and (c) Low-test set.



(a) (b) (c)

Figure 2: A set of three figures on the breast cancer corpus for the PMI-COS configuration : (a) Full-test set ; (b) High-test set; and (c) Low-test set.

ditive smoothing that performs rather poorly has shown good results in our evaluation. The Good-Turing estimate which is not used in isolation forms the basis of later techniques such as Back-off or Jelinek-Mercer smoothing, two techniques that generally work well. The good performance of $Katz$ and $JM$ on the PMI-Cos configuration was expected. The reason is that these two methods have used the Good-Turing estimators which also achieved good performances in our experiments. Concerning the Kneser-Ney algorithm, surprisingly this performed poorly in our experiments while it is known to be one of the best smoothing techniques. Discounting a fixed amount in all counts of observed cooccurrences degrades the results in our data set. We also implemented the modified Knener-ney method (not presented in this paper) but this also performed poorly. We conclude that discounting is not an appropriate method for observed cooccurrences. Especially for point-wise mutual information that over-estimates low frequencies, hense, discounting low cooccurrences will increase this over-estimation.

## 7 Conclusion

In this paper, we have described and compared the most widely-used smoothing techniques for the task of bilingual lexicon extraction from comparable corpora. Regarding the empirical results of our proposition, performance of smoothing on our dataset was better than the baseline for the Add-One smoothing combined with the Log-Jac parameters and all smoothing techniques except the Kneser-ney for the Pmi-Cos parameters. Our findings thus lend support to the hypothesis that a re-estimation process of word cooccurrence in a small specialized comparable corpora is an appropriate way to improve the accuracy of the standard approach.

## Acknowledgments

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

## References

Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pages 264–270, California, USA.

Brown, P. F., Pietra, V. J. D., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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

Chiao, Y.-C. and Zweigenbaum, P. (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.

Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

Dagan, I., Marcus, S., and Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. In *Proceedings of the 31ST Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 164–171, Ohio, USA.

Daille, B. and Morin, E. (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.

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

Evert, S. (2005). *The statistics of word cooccurrences : word pairs and collocations*. PhD thesis, University of Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.

Evert, S. and Baroni, M. (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.

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

Fung, P. (1998). A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora 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.

Gamallo, O. (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.

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

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

Groc, C. D. (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.

Hervé Déjean and Gaussier, É. (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.

Jeffreys, H. (1948). *Theory of Probability*. Clarendon Press, Oxford. 2nd edn Section 3.23.

Johnson, W. (1932). Probability: the deductive and inductive problems. *Mind*, 41(164):409–423.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

Kneser, R. and Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proceedings of the 20th International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, pages 181–184, Michigan, USA.

Laroche, A. and Langlais, P. (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.

Lidstone, G. J. (1920). Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192.

Maarek, Y. S. and Smadja, F. A. (1989). Full text indexing based on lexical relations an application: Software libraries. In *SIGIR*, pages 198–206, Massachusetts, USA.

Manning, D. C., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Mercer, L. ; Jelinek, F. (1980). Interpolated estimation of markov source parameters from sparse data. In *Workshop on pattern recognition in Practice*, Amsterdam, The Netherlands.

Moore, R. C. (2004). Improving ibm word alignment model 1. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 518–525, Barcelona, Spain.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (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.

Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266.

Pereira, F. C. N., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31ST Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 183–190, Ohio, USA.

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

Rapp, R. (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.

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

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology*. Addison-Wesley.