Variations on quantitative comparability measures and their evaluations on synthetic French-English comparable corpora

Guiyao Ke, Pierre-Francois Marteau, Gildas Menier
UMR CNRS 6074 IRISA, Université de Bretagne Sud
Campus de Tohannic, 56000 Vannes, France
ke@univ-ubs.fr, marteau@univ-ubs.fr, menier@univ-ubs.fr

Abstract

Following the pioneering work by (Li and Gaussier, 2010), we address in this paper the analysis of a family of quantitative comparability measures dedicated to the construction and evaluation of topical comparable corpora. After recalling the definition of the quantitative comparability measure proposed by (Li and Gaussier, 2010), we develop some variants of this measure based primarily on the consideration that the occurrence frequencies of lexical entries and the number of their translations are important. We compare the respective advantages and disadvantages of these variants in the context of an evaluation framework that is based on the progressive degradation of the Europarl parallel corpus. The degradation is obtained by replacing either deterministically or randomly a varying amount of lines in blocks that compose partitions of the initial Europarl corpus. The impact of the coverage of bilingual dictionaries on these measures is also discussed and perspectives are finally presented.

Keywords: Comparable corpora, Comparability measures, Evaluation

1. Introduction

Parallel corpora are sets of tuples of aligned documents that are formed with texts placed alongside with their translation(s). If such resources are of great utility in particular in the field of assisted translation, they are expensive to develop and often difficult to transpose from a specialty domain to another. The notion of comparable corpora has emerged in the nineties to palliate the lack of versatility and expensiveness of parallel corpora and to offer avenues to a wider scope of applications such as multilingual terminology extraction, multilingual information retrieval or knowledge engineering (Baker, 1996), (EAGLES, 1996).

The notion of comparability between documents expressed in different languages is not easy to introduce: it is widely admitted that two documents in different languages are comparable when they share analogous criteria of composition, genre and topics. The term of comparable corpora was introduced by (Fung and Yee, 1998), (Munteanu et al., 2004) and remains quite subjective. (Déjean and Gaussier, 2002) proposed a quantitative definition of the concept of "translational" comparability (relatively suitable for machine translation) according to which "Two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called comparable if there is a significant sub-part of the vocabulary of the $\mathcal{L}_1$ language corpus, respectively $\mathcal{L}_2$ language corpus, whose translation is in the corpus of language $\mathcal{L}_2$, respectively $\mathcal{L}_1$." Recently (Li and Gaussier, 2010) have then derived a quantitative measure that is based on a bilingual translation dictionary. This measure consists essentially in inventorying the presence (in a binary way) of the translations of dictionary entries that occur into the paired documents. It depends in a non-explicit way upon jointly the coverage of the bilingual translation dictionary and the studied corpora themselves.

These authors proposed to evaluate their measure on the basis of a progressive degradation of a parallel corpus while observing the variation produced on their measure: the main idea is to check the consistency (in term of correlation with the degree of degradation) of the proposed measure when the number of direct translations of lexical entries decreases in the paired documents.

In this paper we introduce, study and evaluate two variations around this quantitative comparability measure by introducing additional information related to the number of occurrences of lexical entries and jointly to the number of occurrences of their potential translations. These new measures are presented and evaluated against the measure developed by (Li and Gaussier, 2010) while taking into account the coverage of the exploited translation dictionary. The assessment is carried out according to an improved evaluation framework.

2. Variations on a quantitative comparability measure

2.1. Comparability measure of Li and Gaussier (C_{LG})

From the definition of (Déjean and Gaussier, 2002), (Li and Gaussier, 2010) have then derived a quantitative "translational" measure that is based on a bilingual translation dictionary. This measure involves counting the number of lexical entries *gateways* for *coupling* two corpora of different languages *via* a translation lexicon. Let us consider corpus C_1 in language $\mathcal{L}_1$ and corpus C_2 in language $\mathcal{L}_2$. This measure is formally presented in the form:

$$C_{LG}(C_1, C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \sigma(w_1) + \sum_{w_2 \in WC_2 \cap WD_2} \sigma(w_2)}{|WC_1 \cap WD_1| + |WC_2 \cap WD_2|} \quad (1)$$

where : WC_i, $i \in \{1, 2\}$ is the vocabulary in language $\mathcal{L}_i$ associated with the corpus C_i ; WD_i is the set of lexical entries in language $\mathcal{L}_i$ of bilingual dictionary used appears in WC_i; $\sigma(w_i)$ is an indicator function that takes the value 1 if at least one translation of the lexical entry $w_i \in WC_i$ in

language $\mathcal{L}_i$ exists in the vocabulary associated to the other corpus, 0 otherwise.

This comparability measure is called "translational" because it is relatively well suited to assisted translation task.

2.2. Towards a quantitative definition of thematic comparability

As previously mentioned, the C_{LG} measure takes no account neither of the number of occurrences of the lexical entries in documents nor of the number of their translations. However, according to previous works such as in (Rossignol and Sébillot, 2003), a theme is generally characterized by keywords that are frequent inside the theme itself and quite discriminant comparatively to other themes. This leads us to consider, for the construction of a quantitative "thematic" comparability measure, the frequencies of word occurrences inside documents as well as their degree of ambiguity (estimated via the existing number of possible translations in the translation dictionary).

We therefore propose a definition of a "thematic" comparable corpus which is a set of multilingual documents that deals with a same theme. In particular, the (discriminative) terms characterizing the domain are expected to be frequent in the corpus with low ambiguity. We propose a quantitative and operational definition of a "thematic" comparability measure as follows: Two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called "thematically" comparable if:
- on the one hand there is a significant subset of the vocabulary of the $\mathcal{L}_1$ language corpus, respectively $\mathcal{L}_2$ language corpus, whose translation is in the corpus of language $\mathcal{L}_2$, respectively $\mathcal{L}_1$.
- on the other hand, the terms of the related vocabulary subsets must be such that the ratio between their frequency of occurrence and their number of translations is the largest possible (namely, they have to be frequent and lowly ambiguous).

From these definitions, we propose two variants of the C_{LG} measure that explicitly involve these two improvements with the expectation that their inclusion will produce in some experimental situations a positive effect (in particular when considering the classification and the clustering of thematic bilingual documents).

2.2.1. Two "thematic" comparability measures, variants of the C_{LG} measure

The two variants highlight symmetrically between source language and target language the following three factors: the number of occurrences of lexical entries w in the vocabulary of the source language corpus, the number of translations in the bilingual dictionary and the presence of at least one of their translations in vocabulary in the target language corpus.

Let $A_{1|2}$, A_1, $A_{2|1}$, A_2 be defined as follows:

$$A_{1|2} = \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{W(w_1, \mathcal{C}_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right)$$
$$A_1 = \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{W(w_1, \mathcal{C}_1)}{\tau(w_1, WD_1)} \right)$$
$$A_{2|1} = \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{W(w_2, \mathcal{C}_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right)$$
$$A_2 = \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{W(w_2, \mathcal{C}_2)}{\tau(w_2, WD_2)} \right)$$

where $W(w_i, \mathcal{C}_i)$ is either a weight *tf*; $\tau(w_i, WD_i)$ is the translation numbers of the lexical entry w_i of the corpus $\mathcal{C}_i$ in the translation dictionary WD_i. $\sigma(w_i)$ is defined as above ($\sigma(w_i) = 1$ if at least one translation of the lexical entry $w_i \in WC_i$ in language $\mathcal{L}_i$ exists in the vocabulary associated with the other corpus, 0 otherwise.).

1. The first variant, C_{VA_1}, is defined as follows:

$$C_{VA_1} = \frac{1}{2} \cdot \left(\frac{A_{1|2}}{A_1} + \frac{A_{2|1}}{A_2} \right) \quad (2)$$

2. The second variant, C_{VA_2}, is defined as follows:

$$C_{VA_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2} \quad (3)$$

These two variants are very close to each another. They differ mainly on how the symmetrization is performed. Basically, the first variant is similar to an arithmetic average while the second variant relates to a weighted average.

3. Evaluation protocol

Our experiments focuse on the English and French languages and follow the protocol proposed in (Li and Gaussier, 2010) although some improvements and complementary tests have been added. This protocol is built on the principle of a gradual degradation of a parallel corpus by means of deterministic replacements of blocks of lines of text. We completed the protocol by developing a non-deterministic approach for block replacements in order to evaluate the impact of the replacement procedure on the estimation of the capability of the comparability measures to discriminate between successive levels of degradation.

3.1. Evaluation measure

3.1.1. Empirical golden standard reference

The empirical reference measure is built on the basis of the percentage of degradation of the Europarl parallel corpus (Koehn, 2005). For example, if we consider 100 lines per block, for each block and for each test, we obtain a vector of 101 values (starting from 0% replacement to 100% line replacements). We thus obtain an empirical referential measure which serves as a gold standard, characterized by a vector (0%, 1%, 2%...100%) of $N = 101$ coordinates.

3.1.2. Confronting a comparability measure to the empirical reference

To determine the adequacy/inadequacy degree of a measure to the empirical reference, we use the Pearson correlation coefficient. It estimates the correlation degree between a comparability measure X and the empirical reference Y as follows:

$$r_p = \frac{\sum_{n=1}^{N}(X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^{N}(X_n - \bar{X})^2}\sqrt{\sum_{n=1}^{N}(Y_n - \bar{Y})^2}} \quad (4)$$

Among other correlation estimators, the Pearson's correlation coefficient is generally used when random variables X and Y are assumed to follow normal distributions. In the absence of specific known counter argument against it, this coefficient seems to be an acceptable compromise.

3.1.3. Coverage rate

The dictionary and vocabulary coverage rates are parameters that greatly influence the comparability measures as will be shown in the experiments. We define them as follows:

- we define the coverage rate of a dictionary D with respect to a lexicon V associated to a corpus by the quantity $T_D = \frac{|V \cap D|}{|V|}$.
- we define the coverage rate of a lexicon V associated to a corpus with respect to a dictionary D by the quantity $T_V = \frac{|V \cap D|}{|D|}$.

3.2. Preprocessing and evaluation principles

3.2.1. Preprocessing

We have exploited two corpora: a parallel corpus "French-English Europarl (EP) corpus" (Koehn, 2005) and an English corpus "Associated Press corpus (AP)". These corpora are lemmatized by exploiting the TreeTagger (Schmid, 1994) (Schmid, 2009) then segmented into sentences (one sentence per line). We finally obtain three documents, each containing several millions of lines: a parallel French document EPF, a parallel English document EPE and an English document AP.

Furthermore, we used a bilingual dictionary available under the reference ELRA-M0033 (ELRA, 2013) that contains 243,580 pairs of French and English entries, divided into 110,541 English entries and 109,196 French entries.

3.2.2. Evaluation principles

Following the work by (Li and Gaussier, 2010), we have partitioned the parallel corpus Europarl in five different ways by selecting a variable number of lines: 1000 lines, 10000 lines, 100000 lines and 1428000 lines (i.e. the whole corpus Europarl). Each element of the partitions is then divided into 10 blocks, each containing the same number of lines (100 lines, 1000 lines, 10000 lines and 142 800 lines). A progressive degradation of the initial corpus is then carried out as shown in Figure 1 to derive comparable corpora of varying level of comparability. We then evaluate the comparability measures between aligned blocks of lines.

We conducted two series of experiments which distinguish by the block replacement mode: deterministic or random. For each series, three different tests are carried out according to the principles described below. The evaluation of the comparability measures consists in quantifying the correlation, when the the Europarl corpus is progressively degraded: basically, the observed decay of the measures are compared in terms of correlation with the expected decay of the empirical measure serving as a *gold standard* reference.

3.2.3. Deterministic replacement

For the first test, we build the corpus referenced as *GAd* by replacing deterministically a given number of lines from a block (the number of lines depends on the degradation percentage of the parallel corpus $0\%, 1\% \cdots 100\%$) by the same number of lines from another block. The permutation of the blocks is predefined and deterministic, e.g. block 1 is exchanged with block 6, block 2 is exchanged the block 7, etc.

For the second test, we build the corpus referenced as *corpus GBd*, by replacing deterministically a given number of lines from a block (the number of lines depends on the degradation percentage of the desired parallel corpus) by the same number of lines extracted from the corpus *AP*.

For the third test, we build the corpus referenced as *GCd*, by first replacing all the lines of a block by all the lines of another block (for example, block 1 becomes block 6 and block 2 becomes block 7, etc). Then, afterwards, and in each block, a number of lines (that depends on the degradation percentage of the Europarl corpus) are replaced deterministically by the same number of lines extracted from the corpus *AP*.

3.2.4. Random replacement

For the first test, we build the corpus referenced as *GAa* by randomly replacing (according to a uniform law) a given number of lines, depending on the degradation percentage of the Europarl corpus, by the same number of lines extracted (without replacement to ensure that the replacements relate systematically to different lines) from the remaining (not yet drawn) lines of the Europarl corpus.

For the second test, we build the corpus referenced as *GBa* by randomly replacing (according to a uniform law) a given number of lines that depends on the degradation percentage of the Europarl corpus, by the same number of lines drawn (without replacement) from the document *AP*.

For the third test, we build the corpus referenced as *GCa*, by first replacing all the lines of a block by the same number of lines drawn from the complement set of lines of the Europarl corpus (without replacement). Then afterwards, within each block, we perform the random replacement according to a uniform law of a given number of lines (that depends on the degradation percentage of the Europarl corpus) by the same number of lines drawn from the corpus *AP* without replacement.

Hence, for these two series of three tests, the average comparability degree of our degraded corpora decreases, in principle, from $GA_{d|a}$ to $GB_{d|a}$ down to $GC_{d|a}$. Finally,

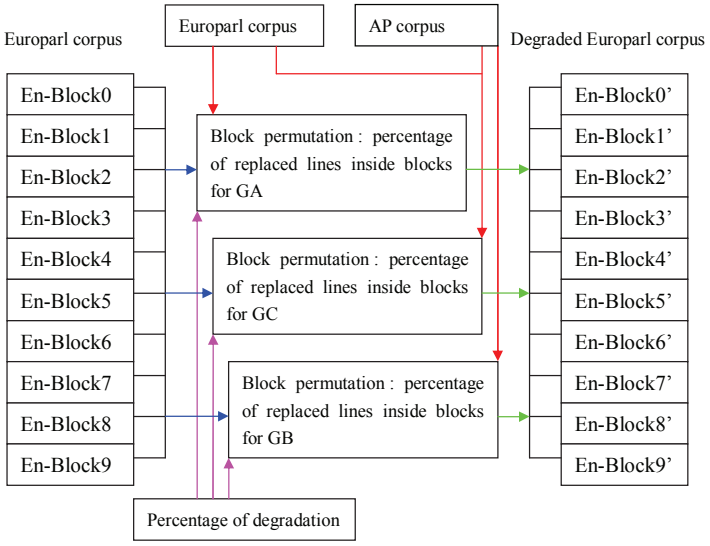

Figure 1: Europarl corpus partitioning and progressive degradation by deterministic or random line replacements

the evaluation of the comparability measures consists in quantifying the correlation between their observed decrease and the expected decrease according to the *empirical gold standard* measure that estimates the degree of degradation of the Europarl corpus.

4. Experiments

4.1. Influence of the block size of the corpus partitions

We study below the average correlations and their standard deviations between the comparability measures and the *empirical gold standard* reference when the block size expressed in number of lines varies into the set $\{10^2, 10^3, 10^4, 10^5\}$.

The figure 2 shows, for the two alternative modes of block replacement, that the measure C_{LG} is more in adequacy with the empirical reference according to the Pearson's correlation coefficient on corpora *GA* than its two variants C_{VA_1} et C_{VA_2}, particularly for large block sizes. For corpora *GB*, the three measures reach almost the same level of correlation with respect to the empirical reference. Finally, on corpora *GC*, both of the variants C_{VA_1} et C_{VA_2} appear to be more robust than C_{LG}, mainly for small block sizes. However, the random replacement procedure seems to improve for all the measures the correlation with the *empirical golden standard* reference, both in average and in standard deviation.

4.2. Influence of the coverage rates (dictionary and lexicons)

We investigate hereinafter the influence of the coverage rates (dictionary and lexicons) on the average correlations with respect to the *empirical golden standard* reference obtained on corpora that are degraded by deterministic or random replacements, this for the three measures C_{LG}, C_{VA_1} and C_{VA_2}. Figure 3 introduces these average correlations for the *dicElra* dictionary and for the two alternative replacement modes, random and deterministic.

It can be noticed in Figure 3 a better average correlation for the C_{LG} measure on the corpus *GA*, while the variants C_{VA_1} and C_{VA_2} show their correlations reducing drastically on the same corpus as the dictionary coverage decreases. However on the *GB* and *GC* corpora, the two variants show to be better correlated than the C_{LG} measure although a slight decrease in average correlation occurs for all three measures when the vocabulary coverage rate is very low. These results are similar for the two replacement modes, deterministic or random.

4.3. Ability of the comparability measures to discriminate successive degradation levels

To estimate the ability of the comparability measures to discriminate between the successive degradation levels of the Europarl parallel corpus, whether they are deterministic or random, we use the following discrimination measure:

$$\Delta(i) = \frac{|\sigma_i+\sigma_{i+1}+2\cdot(m_i-\sigma_i/2-(m_{i+1}+\sigma_{i+1}/2))|}{\sigma_i+\sigma_{i+1}} = \frac{2\cdot|m_i-m_{i+1}|}{\sigma_i+\sigma_{i+1}}$$

where m_i and σ_i are the mean and standard deviation of the comparability measures associated with the degradation levels (from 0%, 1%, $\cdots 100\%$) of the corpus Europarl indexed by $i \in \{1, \cdots, 101\}$. In practice, we observe that $\forall i, m_i >= m_{i+1}$ and thus the absolute value is not required. $\Delta(i) \in [0, \infty[$ is still high especially as the deviation between the successive average comparabilities is high and the sum of associated standard deviations is low. Thus, the higher function $\Delta(i)$ is, the better is discriminated degradation level i by the comparability measure.

The figure 4 shows for the three measures C_{LG}, C_{VA_1} and C_{VA_2}, applied on the three types of degraded corpora (*GA*, *GB* and *GC*) the average value and standard deviation of the discrimination measure Δ depending on the coverage of the *dicElra* dictionary. Here, as well, we find that the C_{VA_1} and C_{VA_2} variants are less discriminative than the measure C_{LG} on the corpus *GA* especially for low coverage rate. On

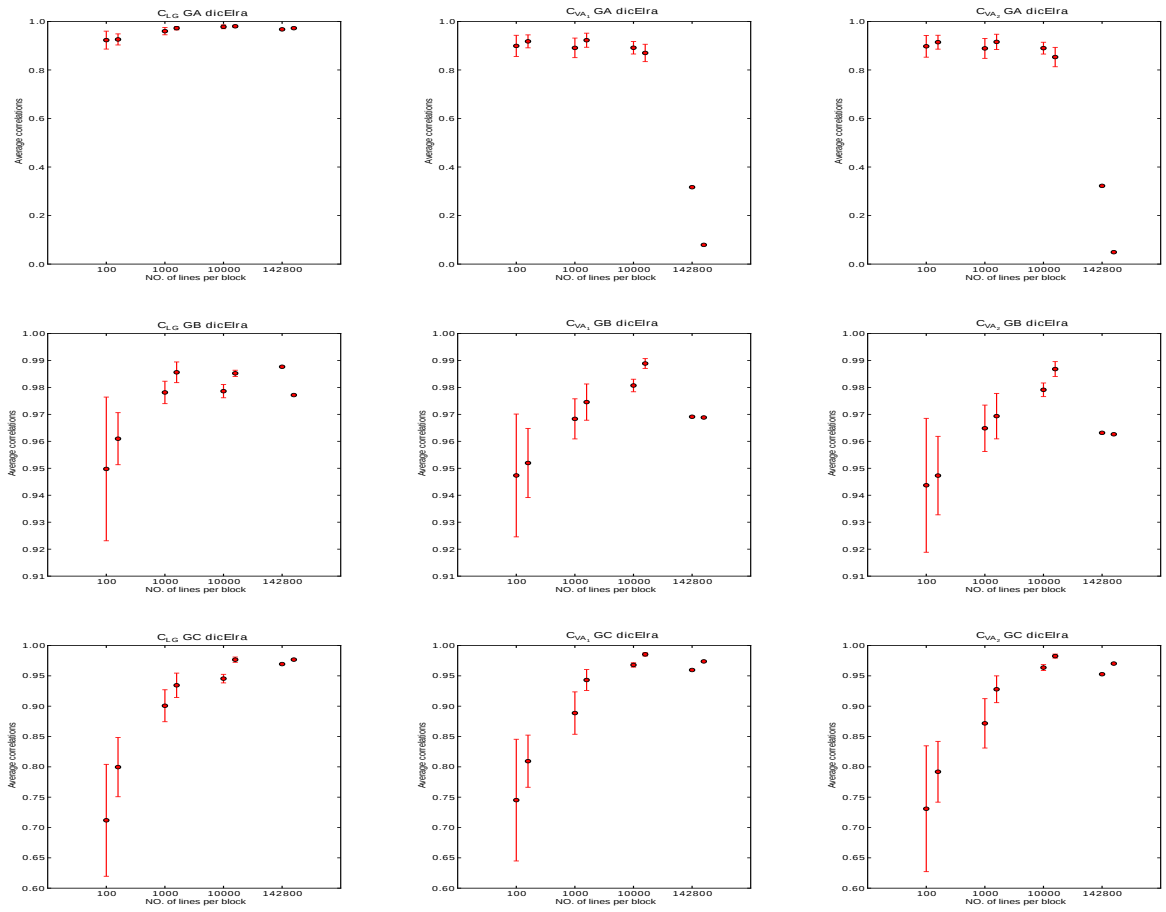

Figure 2: Influence of the block size of the corpus partitions on the average correlations for the three comparability measures with respect to the golden standard empirical reference. The two alternative modes of line replacement are shown on the figure for each block size with a slight horizontal shift: deterministic replacement is shown with a left shift and random replacement is shown with a right shift.

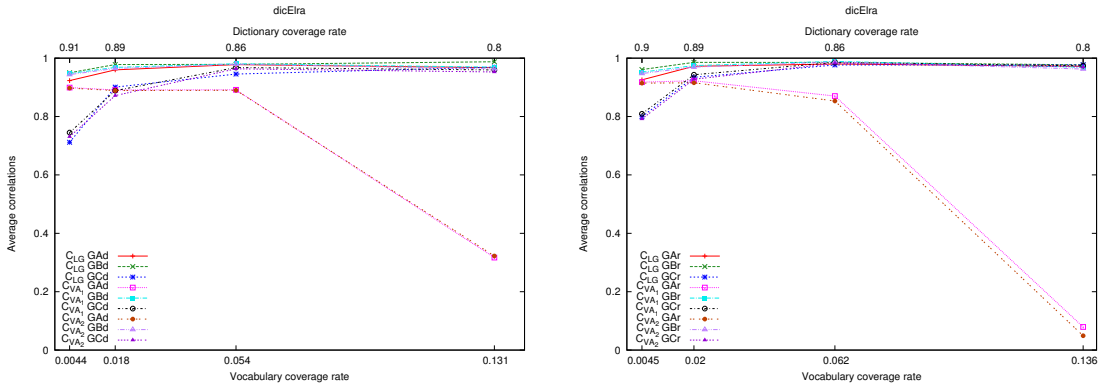

Figure 3: Influence of the coverage rates on the average correlations of the comparability measures with respect to the *empirical golden standard* reference for the dictionary *dicElra*: for the degraded corpora obtained using deterministic replacements (left sub-figure), and for the degraded corpora obtained using random replacements (right sub-figure).

the corpus *GB*, the measures have very similar correlation levels, especially when dictionary coverage rate are high. Finally, on the corpus *GC*, the variants are slightly more robust, especially for high dictionary coverage. Note that in most cases the average discrimination capability of all measures increases when the average dictionary coverage rate decreases.

5. Analysis and conclusions

The results obtained from our various experiments show that the measure C_{LG} and its C_{VA_1}, C_{VA_2} variants are relatively similar in terms of their correlation with respect to the *empirical golden standard* measure defined in the scope of our evaluation protocol. Though it is clear that the measure C_{LG} is better correlated with the golden standard mea-

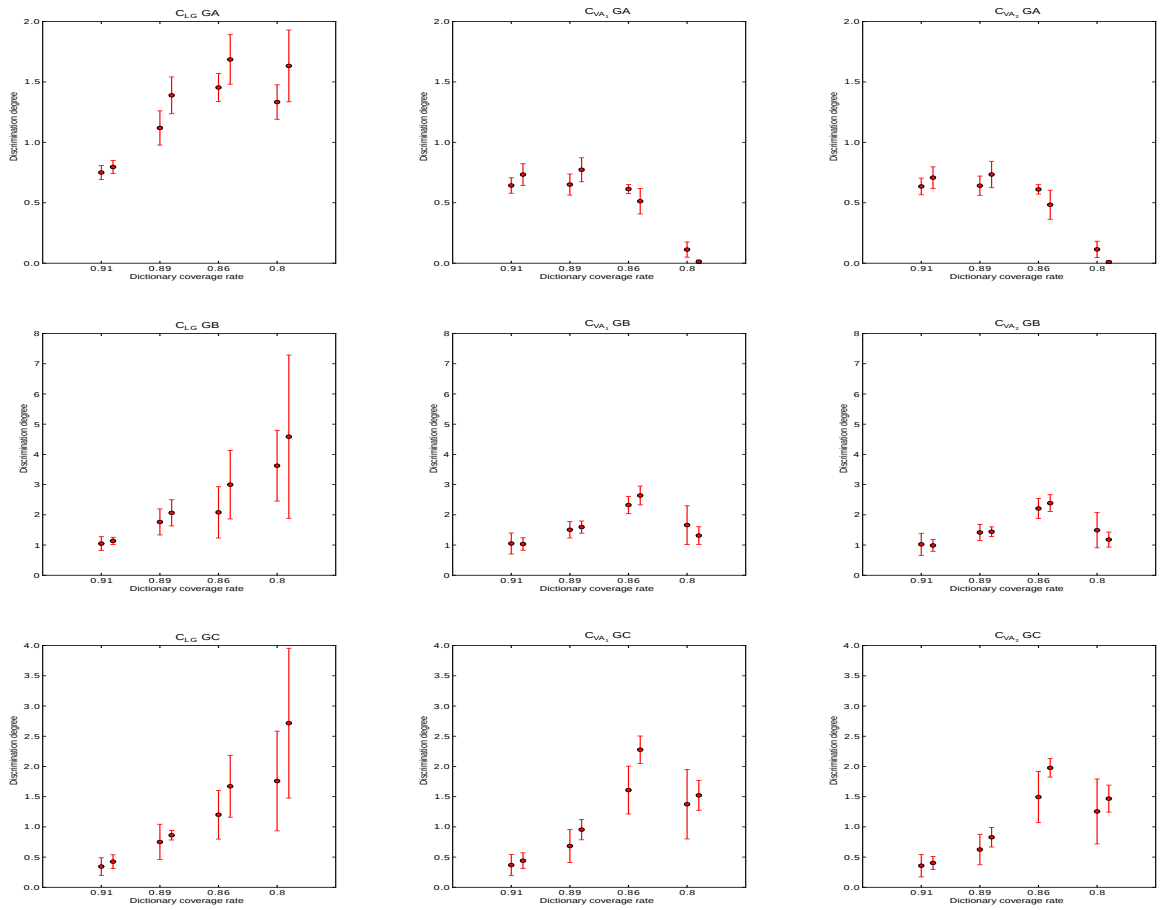

Figure 4: Ability of the comparability measures to discriminate between the degradation degrees of the corpus Europarl: means and standard deviations of $\Delta(.)$ depending on the dictionary coverage rate *dicElra* exploited on the corpus produced by deterministic (left shifts) and random replacements (right shifts).

sure on corpora *GAd* and *GAa* closest to the initial Europarl corpus, however, we cannot conclude that C_{LG} is better than C_{VA_1}, C_{VA_2} on corpora *GAd* and *GAa* because the replacement is done within the Europarl parallel corpus, so when the size of lines is very big, even if we replace 100% of lines, the blocks remain quite comparable. That is why when we use C_{VA_1}, C_{VA_2} to calculate the comparability, the values are very near (that leads to be no linear when compared with the golden standard measure), so in some sense, we can say C_{VA_1}, C_{VA_2} are quite reasonable; while the C_{VA_1}, C_{VA_2} variants are slightly more robust when the measures confront more real life corpora *GCd* and *GCa* that are the farthest from the Europarl corpus and without doubt the closest to *noisy* corpora such as those obtained when harvesting the Web for instance. On the intermediate corpora *GBd* and *GBa*, the three measures achieve comparable correlation levels with respect to our *empirical golden standard* measure.

The correlation degrees of these measures increase when the number of lines per block increases, especially for corpora *GC* (the increase in correlation is of more than 20% between the configuration characterized by 100 lines per block and the configuration characterized by $142,800$ lines per block). For example, for two documents containing about 100 lines each, if the comparability value is greater than 0.7, the two documents are likely to be very similar according to their covered topics while for two documents containing over than 1000 lines each, the comparability value would be greater than 0.8, to assert the same expected degree of comparability. According to this result, we can elaborate a reasonably stable reference for assessing documents comparability that will nevertheless depends on the document size (in number of sentences) to determine whether two documents are sufficiently comparable or not for a given task.

In addition, the estimation of the ability of a measure to discriminate between successive degradation levels of the parallel corpus we have proposed seems also an interesting comparison criterion to take into account. According to this criterion, the measure C_{LG} performs better on corpora *GA*, while the C_{VA_1} and C_{VA_2} variants seem more discriminative on corpora *GC* and slightly better also on corpora *GB* given the lower variances observed on this criterion for the two variants.

The random or deterministic replacement procedures used to progressively degrade the Europarl corpus seem to have a fairly significant impact on our results. The deterministic degradation replacement protocol proposed by (Li and Gaussier, 2010) generates, in general, a decrease in average of the correlations of the evaluated three measures with the

gold standard measure, as well as an increase of the standard deviation, especially on corpora that are far away from the parallel Europarl corpus (i.e. *GB* et *GC*). This leads to prefer the random replacement mode to the deterministic replacement mode.

In terms of perspective, on the one hand, we will try to improve the correlation of the comparability measures with the empirical reference by means of word sense disambiguation. On the other hand, we will exploit and evaluate these comparability measures for harvesting the web for topical comparable corpora production and related subtasks, namely co-classification and co-clustering of bilingual topical data. Another issue is the consolidation of thematic similarity measures by the means of comparability. In (Ke et al., 2013) we have clearly shown that a mixing model of comparability with monolingual thematic similarities improved greatly the co-classification or co-clustering of bilingual thematic documents. We have furthermore shown that the two comparability measure variants we have proposed are more suited to these tasks than the C_{LG} measure that has been developed mostly for translation purposes. Hereinafter, we present, as an example, the effect of combining similarity and comparability on a 1-NN classification task when using the three comparability measures.

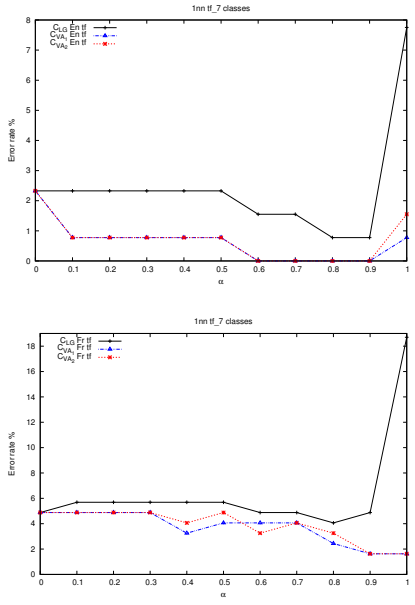

Figure 5: Effect of combining similarity and comparability on a 1-NN classification task. The accuracy of the classification is given for the three comparability measures (C_{LG} +/black, C_{VA_1} */red and C_{VA_2} triangle/blue) when the mixing parameter α varies in $[0, 1]$ (0: no comparability, 100% similarity; 1: 100% comparability, no similarity); top figure shows the classification of English documents while the down figure shows the classification of French document.

The figure 5 shows that the combination of similarity and comparability has a significant effect on all the three comparability measures, especially for our two variants C_{VA_1} and C_{VA_2} by lowering about 3 percent the error rate for jointly the English language and the French language. The measure C_{LG} can also improved the accuracy when choosing a *good* α value. However, the improvement is much lower and less stable than for the two variants we have proposed. This experiment justifies the "thematic" designation that we used to characterize the two variants by opposition of "translational" designation that characterize the original C_{LG} measure.

6. Acknowledgements

This work has been partially funded by the French National Research Agency (ANR-METRICC project).

7. References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam & Philadelphia: John Benjamins.

Déjean, H. and Gaussier, E. (2002). Une nouvelle approche a l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Numéro spécial, corpus alignés:1–22.

EAGLES. (1996). Expert advisory group on language engineering standards guidelines: http://www.ilc.pi.cnr.it/eagles96/browse.html. Technical report, EAGLES.

ELRA. (2013). European language resources association, `http://catalog.elra.info`.

Fung, P. and Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 414–420, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ke, G., Marteau, P.-F., and Ménier, G. (2013). Improving the clustering or categorization of bi-lingual data by means of comparability mapping. Technical report, October.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING*, pages 644–652.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272.

Rossignol, M. and Sébillot, P. (2003). Extraction statistique sur corpus de classes de mots-clés thématiques. *TAL. Traitement automatique des langues*, 44(3):217–246.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schmid, H. (2009). TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.