On the Assessment of Text Corpora*

David Pinto[1], Paolo Rosso[2], and Héctor Jiménez-Salazar[3]

[1] Faculty of Computer Science,
B. Autonomous University of Puebla, Mexico
dpinto@cs.buap.mx

[2] Natural Language Engineering Lab. - ELiRF
Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es

[3] Department of Information Technologies,
Autonomous Metropolitan University, Mexico
hgimenezs@gmail.com

Abstract. Classifier-independent measures are important to assess the quality of corpora. In this paper we present supervised and unsupervised measures in order to analyse several data collections for studying the following features: *domain broadness*, *shortness*, *class imbalance*, and *stylometry*. We found that the investigated assessment measures may allow to evaluate the quality of gold standards. Moreover, they could also be useful for classification systems in order to take strategical decisions when tackling some specific text collections.

1 Introduction

Many algorithms devoted to document categorization have been tested on classical corpora such as Reuters and 20 Newsgroups in order to determine their quality. However, up to now the relative hardness of those corpora has not been completely determined.

The relative clustering hardness of a given corpus may be of high interest, since it would be helpful to determine whether or not the usual corpora used to benchmark the clustering algorithms are hard enough.

Moreover, when dealing with raw text corpora, if it is possible to find a set of features involved in the hardness of the clustering task itself, ad-hoc clustering methods may be used in order to improve the quality of the obtained clusters. Therefore, we believe that this study would be of high benefit.

In [1], the authors attempted to determine the relative hardness of different Reuters-21578 subsets by executing various supervised classifiers. However, in their research work it is not defined any measure for determining the hardness of these corpora, neither the possible set of features that could be involved in the process of calculating the relative hardness of some corpus.

* This research work was partially supported by the CICYT TIN2006-15265-C06 project.

H. Horacek et al. (Eds.): NLDB 2009, LNCS 5723, pp. 281–290, 2010.

The aim of our proposal is to evaluate classifier-independent features which could help on determining the hardness of a given corpus. As far as we know, research work in this field nearly have been carried out in literature.

For the purpose of our investigation, we took into account four different corpus features: *domain broadness*, *shortness*, *class imbalance*, and *stylometry*. We consider that these features will be sufficient to evaluate the relative hardness of a document collection. The aim is, for instance, to agree on whether the quality of the gold standard is good enough or not.

The description of the features to be investigated together with the corresponding assessing measures is given as follows.

Domain broadness. The goal is to evaluate the broadness of a given corpus. We assume (see for instance [2]) that it is easier to classify documents belonging to very different categories, for instance "sports" and "seeds", than those belonging to very similar ones, e.g. "barley" and "corn" (Reuters-21578). The attempt is to indicate the *domain broadness* degree of a given corpus. A binary classifier would assign, respectively, the tags *wide* to the former "sports-seeds" collection and *narrow* to the latter "barley-corn" one.

Shortness. The term frequency is crucial for the majority of the similarity measures. When dealing with very short texts, the frequency of their vocabulary is very low and, therefore, the clustering algorithms have the problem that the similarity matrix has very low values. Therefore, we believe that independently of the clustering method used, the average text length of the corpus to be clustered is an important feature that must be considered when evaluating its relative difficulty. The formula introduced by Herdan [3] has extensively been used for measuring lexical richness of documents [4] such as, vocabulary richness for authorship attribution [5].

Class imbalance. The document distribution across the corpus is another feature that we consider important to take into account. There may exist different levels of difficulties depending on whether the corpus is balanced or not. This feature is even more relevant when the corpus is used with the purpose of benchmarking different classifiers, for instance in the different tasks of an international competition such as SemEval[1]. Let us suppose that the corpus is totally unbalanced and, that for some reason exists a clue of that. If so, then some participants would "wisely" force their system to obtain the least possible number of clusters in order to get the best performance (unfair for the rest of the teams). The *imbalance* degree of a given corpus is also closely-related to the external corpus validation measure used (e.g. F-Measure) and, therefore, the obtention of a single value for measuring it will clearly be of high benefit. Two research works that deal with the problem of class imbalance are the ones presented in [6] and [7]. Particularly, in the former paper it is claimed that class (category) imbalances hinder the performance of standard classifiers.

Stylometry. It refers to the linguistic style of a writer. The goal is to determine the authorship of a set of documents. Even if in our case, the aim

[1] http://nlp.cs.swarthmore.edu/semeval/

is not to attribute the authorship but to distinguish between scientific and other kind of texts. Due to the specific writing style of researchers, when the collection to be clustered is scientific then a new level of difficulty arises. This observation has its basis in domain-dependent vocabulary terms that are not considered in the pre-processing step (for example, in the elimination of stopwords phase). There have been carried out several approaches on the statistical study of writing style (stylometry) field [8]. Morover, up to now, it stills an active research area [9,10].

The rest of this paper is structured as follows. The following section describes the measures we have used in the study of assessment of the quality of text corpora. Section 3 presents the evaluation of standard corpora used in the task of categorization. Finally, the conclusions are given.

2 Corpus Assessment Measures

The supervised vs. unsupervised nature way of measuring each of the mentioned corpus features is very important. Some measures evaluate the gold standard of the target corpus and, therefore, they are devoted to evaluate the classification given by the "experts". The other evaluations are meant to be obtained without any knowledge of the distribution of the documents and, therefore, they may be used to either evaluate general features of the collection or to improve, for instance, clustering results from an unsupervised viewpoint.

In the following sub-sections we present both, the supervised and unsupervised versions of the previously introduced corpus features.

2.1 Domain Broadness Evaluation Measures

In this approach, we assume (see for instance [2]) that it is easier to classify documents belonging to very different categories, for instance "sports" and "seeds", than those belonging to very similar ones, e.g. "barley" and "corn" (Reuters-21578). The attempt is to indicate the *domain broadness* degree of a given corpus. A binary classifier would assign, respectively, the tags *wide* to the former "sports-seeds" collection and *narrow* to the latter "barley-corn" one.

Using statistical language modeling. The first approach presented for the assessment of gold standards makes use of Statistical Language Modeling (SLM) in order to calculate probabilities of sequences of words in the different classes of a gold standard and, thereafter, to determine the domain broadness degree of the corresponding corpus by using two different variants, namely supervised and unsupervised.

SLM is commonly used in different natural language application areas such as machine translation, part-of-speech tagging, information retrieval, etc [11,12,13]. However, it has been originally known by its use in speech recognition (see for instance [14]) which stills the most important application area.

Informally speaking, the goal of SLM consists in building a statistical language model in order to estimate the distribution of words/strings of natural language. The calculated probability distribution over strings S of length n, also called n-grams, attempts to reflect the relative frequency in which S occurs as a sentence. In this way, from a text-based perspective, such a model tries to capture the writing features of a language in order to predict the next word given a sequence of them.

In our particular case, we have considered that every hand-tagged category of a given corpus would have a language model. Therefore, if this model is very similar to the rest of them which were calculated from the remaining categories, then we could affirm that the corpus is narrow domain. Our proposal approaches also in an unsupervised way the problem of determining the domain broadness of a given corpus by calculating language models for v partitions of the corpus without any knowledge about the expert document categorization. However, due to the fact that the perplexity is by definition dependent on the text itself, we should make sure that the text chosen is representative of the entire corpus [15].

Given a corpus made up of k categories $C = \{C_1, C_2, \cdots, C_k\}$, we obtain the language model of all the categories except C_i ($\bar{C}_i$) and, thereafter, we compute the perplexity (PP) of the obtained language model with respect to the model of C_i. That is, we use the category C_i as a test corpus and the remaining ones as a training corpus in a leave one out process. Formally, the *Supervised* Language Modeling Based (SLMB) approach for determining the domain broadness degree of the corpus C may be obtained as shown in Eq. (1).

$$SLMB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(PP(C_i|\bar{C}_i) - \mu(PP(C|\bar{C}_i))\right)^2} \qquad (1)$$

where

$$\mu(PP(C|\bar{C}_i)) = \frac{\sum_{i=1}^{k} PP(C_i|\bar{C}_i)}{k} \qquad (2)$$

The *Unsupervised* Language Modeling Based (ULMB) approach for assessing the domain broadness of a text corpus is computed as follows. Given a corpus C splitted into subsets C_i' of l documents, we calculate the perplexity of the language model of C_i' with respect to the model of a training corpus composed of all the documents not contained in C_i' ($\bar{C}_i'$). Formally, given $\bar{C}_i' \bigcup C_i' = C$ such as $\bar{C}_i' \bigcap C_i' = \emptyset$ and $k =$Integer($\frac{|C|}{|C_i'|}$) with $|C_i'| \approx l$, the *unsupervised* broadness degree of a text corpus C may be obtained as shown in Eq. (3).

$$ULMB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(PP(C_i'|\bar{C}_i') - \mu(PP(C|\bar{C}_i'))\right)^2} \qquad (3)$$

where

$$\mu(PP(C|\bar{C}_i')) = \frac{\sum_{i=1}^{k} PP(C_i'|\bar{C}_i')}{k} \qquad (4)$$

Using vocabulary dimensionality. This measure of calculating the domain broadness of a corpus assumes that those subsets belonging to a narrow domain will share the maximum number of vocabulary terms compared with the subsets which do not. In case of a wide domain corpus, it is expected that the standard deviation of vocabularies obtained from subsets of this corpus is greater than the one of a narrow domain corpus. We formalise the above mentioned idea as follows.

Given a corpus C (with vocabulary $V(C)$) which is made up of k categories C_i, the *Supervised* Vocabulary Based (SVB) measure for the domain broadness of C may be written as shown in Eq. (5).

$$SVB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|V(C_i)| - |V(C)|}{|C|}\right)^2} \tag{5}$$

The Unsupervised version of the Vocabulary-Based (UVB) domain broadness evaluation measure would be useful when the gold standard is not available. Since the categories are unknown, we could then use each document (n) instead of the corpus categories (k). The *unsupervised* broadness evaluation measure (based on vocabulary dimensionality) of a corpus C made of n documents $(D_1, ..., D_n)$ may be written as shown in Eq. (6).

$$UVB(C) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|V(D_i)| - |V(C)|}{|C|}\right)^2} \tag{6}$$

2.2 Shortness-Based Evaluation Measures

These evaluation measures assess features derived from the length of a text. Given a corpus C made up of n documents D_i, we present two *unsupervised* text length-based evaluation measures which take into account the level of shortness [3]. We directly calculated the arithmetic mean of Document Lengths (DL) and Vocabulary Lengths (VL) as shown in Eq. (7) and (8), respectively.

$$DL(C) = \frac{1}{n}\sum_{i=1}^{n}|D_i| \tag{7}$$

$$VL(C) = \frac{1}{n}\sum_{i=1}^{n}|V(D_i)| \tag{8}$$

2.3 Class Imbalance Degree Evaluation Measure

The class *imbalance* degree is an important feature that must be considered when corpora are categorized, since according to the imbalance degree there could exist different levels of difficulty [6]. This feature is even more relevant when the corpus is used for benchmarking different classifiers. Let us suppose that the corpus is totally unbalanced and, that for some reason there exist some clue of that. This fact could lead some participants to force their system to obtain the least possible number of clusters in order to get the best performance. In these conditions it would be quite difficult to carry out a fair evaluation and, therefore, to determine which is(are) the best system(s).

Given a corpus C (made of n documents) with a pre-defined gold standard composed of k classes (C_i), the Expected Number of Documents per Class is assumed to be: $ENDC(C) = \frac{n}{k}$.

The *supervised* Class Imbalance (CI) evaluation measure is calculated as the standard deviation of C with respect to the expected number of documents per class in the gold standard as shown in Eq. (9).

$$CI(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\Big(|C_i| - ENDC(C)\Big)^2} \tag{9}$$

2.4 Stylometric-Based Evaluation Measure

The aim of this measure is to determine whether a corpus is written with the same linguistic style or not.

For the analysis of slylometry introduced here, we make use of the Zipf law [16]. Formally, given a corpus C with vocabulary $V(C)$, we may calculate the probability of each term t_i in $V(C)$ as shown in Eq. (10) and the expected Zipfian distribution of terms as shown in Eq. (11). We used the classic version of the Zipf's law and, therefore, s was set to 1.

$$P(t_i) = \frac{freq(t_i, C)}{\sum_{t_i \in V(C)} freq(t_i, C)} \tag{10}$$

$$Q(t_i) = \frac{1/i^s}{\sum_{r=1}^{|V(C)|} 1/r^s} \tag{11}$$

The *unsupervised* Stylometric Evaluation Measure (SEM) of C is obtained by calculating the asymmetrical Kullback-Leibler distance of the term frequency distribution of C with respect to its Zipfian distribution, as shown in Eq. (12).

$$SEM(C) = \sum_{t_i \in V(C)} P(t_i) \log \frac{P(t_i)}{Q(t_i)} \tag{12}$$

A summary of the presented assessment corpus measures is given in Table 1.

Table 1. Text corpora assessing measures

Short name	Description	Category	Approach
$SLMB(C)$	Language model perplexity	Broadness	Supervised
$ULMB(C)$	Language model perplexity	Broadness	Unsupervised
$SVB(C)$	Vocabulary of categories	Broadness	Supervised
$UVB(C)$	Vocabulary of document	Broadness	Unsupervised
$DL(C)$	Document length	Shortness	Unsupervised
$VL(C)$	Vocabulary size	Shortness	Unsupervised
$CI(C)$	Document distribution	Imbalance	Supervised
$SEM(C)$	Zipfian based distribution	Stylometric	Unsupervised

3 Computational Study of Testbed Corpora

The experiments were carried out over the following corpora: *WebKB* [17], *CICLing-2002* [18], *hep-ex* [7], *20 Newsgroups* (20NG) and the *R8* and *R52* subsets of the Reuters-21578 text categorization collections. Moreover, the 100 corpora which compose the *WSI-SemEval* collection were also used [19].

In order to assess how well each assessment measure performs, we have correlated the automatic ranking of the corpora (according to each measure) with respect to an expert manual ranking.

We have considered that there are no tied ranks and we have not made any assumptions about the frequency distribution of the evaluation measures. Moreover, the equi-distance between the different corpora evaluation value cannot be justified and, therefore, the correlation was calculated by means of the *Kendall tau* (τ) rank correlation coefficient [20].

$$\tau = \frac{2 \cdot P}{(e \cdot (e-1))/2} - 1 \tag{13}$$

where e is the number of items, and P is the number of concordant pairs obtained as the sum, over all the items, of those items ranked after the given item by both rankings.

This coefficient value lies between -1 and 1, and high values imply a high agreement between the two rankings. Therefore, if the agreement (disagreement) between the two rankings is perfect, then the coefficient will have the value of 1 (-1).

Tables 2 and 3 illustrate all the evaluation measures with the corresponding obtained value for each one of the evaluated corpus. Aside of each measure we may also see the associated manual ranking which is used to evaluate their performance.

The correlation results (see Table 4) show a high agreement between the automatic and manual corpus rankings for each one of the analysed measures (over 109 corpora). The lowest value (0.56) was obtained for two unsupervised measures ($ULMB(C)$ and $UVB(C)$). Therefore, we consider to have a good

Table 2. Corpus assessment measures for domain broadness

Corpus	$SLMB(C)$	$ULMB(C)$	$SVB(C)$	$UVB(C)$
CICLing-2002	38.9 / 1	63.6 / 1	1.73 / 1	2.70 / 1
hep-ex	298.2 / 2	93.8 / 2	2.75 / 2	3.07 / 2
WSI-SemEval	195.0 / 3	130.6 / 3	1.80 / 3	3.06 / 3
WebKb-Training	262.3 / 5	628.6 / 5	0.50 / 5	1.77 / 5
WebKb-Test	337.4 / 4	218.9 / 4	0.44 / 4	1.60 / 4
R52-Training	627.6 / 9	143.1 / 9	4.38 / 9	4.62 / 9
R52-Test	565.8 / 8	177.5 / 8	4.58 / 8	4.82 / 8
R8-Training	603.9 / 7	135.9 / 7	3.67 / 7	4.76 / 7
R8-Test	545.7 / 6	134.6 / 6	3.84 / 6	4.89 / 6
20NG-Training	694.4 / 11	400.2 / 11	5.23 / 11	6.08 / 11
20NG-Test	786.0 / 10	455.4 / 10	5.21 / 10	6.05 / 10

Table 3. Corpus assessment measures (stylometry, shortness and class imbalance)

Corpus	$SEM(C)$	$DL(C)$	$VL(C)$	$CI(C)$
CICLing-2002	0.301 / 11	70.5 / 7	48.4 / 7	0.036 / 3
hep-ex	0.271 / 10	46.5 / 1	36.8 / 1	0.280 / 11
WSI-SemEval	0.448 / 9	59.6 / 2	50.3 / 2	0.226 / 10
WebKb-Training	0.231 / 8	133.7 / 9	77.1 / 9	0.096 / 6
WebKb-Test	0.227 / 7	136.2 / 8	79.4 / 8	0.097 / 7
R52-Training	0.159 / 5	70.3 / 6	43.1 / 6	0.067 / 4
R52-Test	0.120 / 2	64.3 / 4	39.7 / 4	0.068 / 5
R8-Training	0.142 / 4	66.3 / 5	41.2 / 5	0.171 / 9
R8-Test	0.098 / 1	60.1 / 3	37.3 / 3	0.169 / 8
20NG-Training	0.154 / 6	142.7 / 11	84.3 / 11	0.004 / 1
20NG-Test	0.144 / 3	138.7 / 10	83.2 / 10	0.005 / 2

trade-off between the unsupervised characteristic and the relatively low Kendall tau value obtained by $ULMB$ and UVB. In particular, the time needed for calculating the $ULMB$ measure in huge collections may be prohibitive, but this issue may be alleviated by using sampling over the complete data set.

All the assessment measures were compiled in a on-line system which we have made available for all interested researchers[2]. Therefore, the assessment measures may be used not only to evaluate other corpora but to compare the results with the standard corpora already evaluated and presented in this research work.

Table 4. Correlation between the automatically and manually obtained ranking

Assessment measure	τ **value**
$SLMB(C)$	0.82
$ULMB(C)$	0.56
$SVB(C)$	0.67
$UVB(C)$	0.56
$SEM(C)$	0.86
$DL(C)$	0.96
$VL(C)$	0.78
$CI(C)$	1.00

4 Conclusions

We have presented a set of corpus evaluation measures that may be used to either, evaluate gold standards or to make decisions a priori when, for instance, clustering particular kinds of text collections such as, narrow domain short-text corpora [18].

All the proposed measures were executed over several corpora in order to determine their evaluation capability. We ranked each corpus according to the evaluation value given by the corresponding measure and, thereafter, we calculated the Kendall tau correlation coefficient in order to determine the correlation

[2] http://nlp.dsic.upv.es:8080/watermarker; http://nlp.cs.buap.mx/watermarker

degree between the automatically and the manually obtained ranking. Our findings indicate a strong agreement of all the evaluation measures with respect to the manual ranking.

The developed quality corpus analysis system would allow researches in different fields of linguistics and computational linguistics to easily assess their corpora with respect to the aforementioned corpus features.

References

1. Debole, F., Sebastiani, F.: An analysis of the relative hardness of Reuters-21578 subsets. Journal of the American Society for Information Science and Technology 56(6), 584–596 (2005)
2. Wibowo, W., Williams, H.: On using hierarchies for document classification. In: Proc. of the Australian Document Computing Symposium, pp. 31–37 (1999)
3. Herdan, G.: Type-Token Mathematics: A Textbook of Mathematical Linguistics. Mouton & Co., The Hague (1960)
4. Tweedie, F.J., Baayen, R.H.: How variable may a constant be?: Measures of lexical richness in perspective. Computers and the Humanities 32(5), 323–352 (1998)
5. Hoover, D.L.: Another perspective on vocabulary richness. Computers and the Humanities 37(2), 151–178 (2004)
6. Japkowicz, N.: The class imbalance problem: Significance and strategies. In: Proc. of the 2000 International Conference on Artificial Intelligence (IC-AI 2000), vol. 1, pp. 111–117 (2000)
7. Montejo-Ráez, A.: Automatic text categorization of documents in the High Energy Physics domain. Phd thesis, Granada University, Spain (2006)
8. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Applied Intelligence 19(1-2), 109–123 (2004)
9. Can, F., Patton, J.M.: Change of writing style with time. Computers and the Humanities 38(1), 61–82 (2004)
10. Hoover, D.L.: Corpus stylistics, stylometry, and the styles of henry james. Style 41(2), 174–203 (2007)
11. Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J.: Large language models in machine translation. In: Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 858–867 (2007)
12. Màrquez, L., Padró, L.: A flexible pos tagger using an automatically acquired language model. In: Proc. of the 35th annual meeting on Association for Computational Linguistics, pp. 238–245 (1997)
13. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Research and Development in Information Retrieval, pp. 275–281 (1998)
14. Bahl, L.R., Jelinek, E., Mercer, R.L.: A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 5(2), 179–190 (1983)
15. Brown, P.F., Pietra, V.J.D., de Souza, P.V., Lai, J.C., Mercer, R.L.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–479 (1992)
16. Zipf, G.K.: Human behaviour and the principle of least effort. Addison-Wesley, Reading (1949)
17. Cardoso-Cachopo, A., Oliveira, A.: Combining LSI with other classifiers to improve accuracy of single-label text categorization. In: First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning - EWLSATEL 2007 (2007)

18. Pinto, D., Benedí, J.M., Rosso, P.: Clustering narrow-domain short texts by using the Kullback-Leibler distance. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 611–622. Springer, Heidelberg (2007)
19. Agirre, E., Soroa, A.: Semeval-2007 task 2: Evaluating word sense induction and discrimination systems. In: Proc. of the 4th International Workshop on Semantic Evaluations - SemEval 2007, pp. 7–12. Association for Computational Linguistics (2007)
20. Kendall, M.: A new measure of rank correlation. Biometrika 30, 81–89 (1938)