Fast and Accurate Language Detection in Short Texts using Contextual Entropy

Edgar Chávez[1], Moisés García[2], Jesus Favela[1]

[1] CICESE, Ensenada, México {elchavez,favela}@cicese.mx
[2] Universidad Michoacana, México moises@fie.umich.mx

Abstract. In this work we address the problem of Language identification (LI) on short segments of text. The central idea is to compute the entropy of a document in different contexts and assign it to the category where the entropy is maximal. Only word distributions are needed for the task, no other training is done. For LI the contexts are the languages, and classification is done by just evaluating the high order entropy of the text. Our results show that the language of the text, in the challenging case of short texts, can be accurately identified, matching state of the art approaches reported in the literature. Our method is also fast, given its simplicity, it is easy to code and needs no training, aside from the estimation of words distributions for each language, if not already available.

1 Introduction

The information in social media is growing exponentially. An important part of this corpus is made of short texts, these messages tend to be written in an informal tone, using different languages and are often not grammatically correct.

The problem of language identification (LI) is interesting on its own. However, in many practical applications LI can be seen as step of a more complex text processing task. Accurate LI can facilitate the use of background information about the language and the use of more specialized natural language processing approaches dealing with a collection or a stream of texts [10].

There is widespread misconception of LI being a "solved task", generally as a result of isolated experiments over homogeneous datasets[5, 12]. Part of the motivation for this paper is to draw attention to the fact that, as a field, we are still a long way off from perfect LI of web documents, mainly due to the small size of the texts and the number of languages available on the Web.

The popularity of social media, including Twitter and social networking sites, has generated research on social media data analysis, such as opinion mining. Accurate LI on short and grammatically-ill text messages is thus, required to support advances in these areas.

We consider LI as a text classification problem –the assignment of natural language texts to one or more predefined categories (languages in this case) – Text classification poses many challenges for inductive learning methods since there can be millions of word features. The resulting classifiers, however, have

many advantages: they are easy to build and update, they depend only on information that can be easily obtained (i.e., samples of items that are within the same categories), they can be customized to specific categories of interest to individuals, and they allow users to smoothly tradeoff precision and recall depending on their task.

In this work, we propose the use of an approach for text classification based on the Contextual Entropy (CE) [4], with a modification that considers that all words in each category (language) have equal probability. Thus, we compute the contextual entropy value using different word distributions, one for each document class, then classify the document as belonging to the class which maximizes the entropy (or a higher moment of it).

The rest of the paper is organized as follows. In section 2 we define the problem of language identification. In section 3, we define and explain contextual entropy. Section 4, presents experimental results comparing the performance of our approach to LI with previous results. Finally, section 5 contains our conclusions and suggestions for future work.

2 Language Identification

LI is a supervised learning task, particularly a plain single label multi-class classification [10]. Given some historical or training data in which for each text t there exists a label l, the language in which this text is written, the goal is to learn a model such that given some previously unseen text, it can identify, as accurately as possible, in which language this text is written. Some cases to classify text in LI are: a) When a text written partly in one language and partly in some other language and someone would like to get both labels as an output. b) Language groups or any other dependencies between the language labels. Case (a) can be solved by chopping the text in small portions (for example by segmenting the text in phrases and classifying each phrase independently). Case (b) refers to languages that have many common words, like italian, spanish, portugues, etc. Both problems can be solved by aiming at short text analysis, as we do in this paper.

A recent study [11] established that the best known technique to address this problem is to classify documents according to rank order statistics over character n-gram sequences between a document and a global language profile [1]. Other statistical approaches applied to LI include Markov models over n-gram frequency profiles [3], dot products of word frequency vectors [2], and string kernels in support vector machines [7].

In contrast to purely statistical methods, linguistically-motivated models for LI have also been proposed, such as the use of stop word lists [6], where a document is classified according to its degree of overlap with lists for different languages. This latter idea is taken into account in the development of our approach.

3 Contextual Entropy

In previous work [4] we have shown that entropy is effective in text categorization for formal as well as for intuitive reasons. The entropy is a statistic that depends both on the object itself (the text) and the context (the vocabulary distribution). This fact is strong, specially because it accounts for the property we are seeking: Putting the text in the right context. Formally measuring the amount of information in the process of communication was introduced by Shannon's work [9], the amount of information in the text can be determined by the value of entropy, according to equation 1.

$$H(X) = -\sum_i p_i log(p_i) \tag{1}$$

Where X is a discrete variable, and $\{p_i\}$ are the probability of taking the value i. The above equation 1 plays a central role in information theory as a measure of information, choice and uncertainty [8]. Shannon suddenly jumps from the description of information production to information itself, choice, and uncertainty. H measures the amount of information transmitted in the communication process as a product of the selection of one out of several possible messages, reflecting the occurrence of one event out of several events in the information source.

Higher order moments of the entropy has been shown to be more accurate for the task, and Contextual Entropy, defined according to equation 2 [4] optimizes for a certain value of n. This optimal classification value need to be estimated empirically.

$$H(c)(X) = -\sum_i p(c)_i \log^n(p(c)_i) \tag{2}$$

In equation 2, $H(c)$ is the entropy computed in context c, and X is a a document, $p(c)_i$ is the probability of the word i in context c. The value n is empirically obtained. The document is assigned to the language maximizing the higher order entropy. In figure 1 the entropy with highest value computed correspond to the distribution that belongs the test document.

To show how the value n affects the computation of equation 2, consider that each element in the $\mathbf{W}$ distribution is equally probable, i.e. have the same value given by $p = \frac{1}{|W|}$. If we have multiple distributions ($\mathbf{W}_m$) and they give close entropy values, when we increase n and compute H the net effect is an increase in the gap values. Figure 2 shows the effect of the computed values for H for various values of n . The results better for some specific values of n.

3.1 Classification

The classification can be speed up with a rather simple procedure. We only need to verify common/unique words in each one of the distributions and compute the cumulative contextual entropy accordingly. Lets see the details. Let W_A be

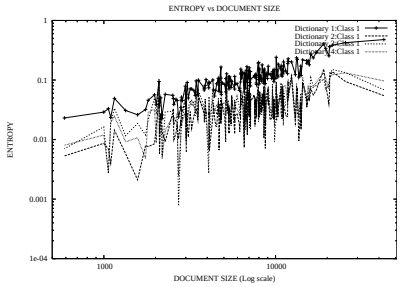

Fig. 1. Entropy computed using training documents of the class k and different distributions $\mathbf{W}_m$. The distributions were obtained from m classes (1 to 4) of the training data as example.

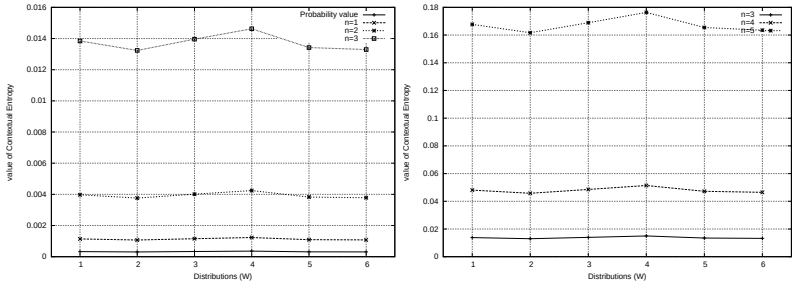

Fig. 2. Increasing n from equation 2.

the distribution of class A, and let W_B be the distribution of class B, such that $P_i = \{p_1, p_2, \ldots, p_d\}$ represent the probability distribution of the words w_d obtained from class i, $i = \{A, B\}$, in our approach all p_d are equal, for $0 \leq p_d \leq 1$, d is the number of words in the dictionary W, and

$$\sum_d p_d = 1$$

Let a new document $\mathbf{x} = \{x_1, x_2, \ldots, x_l\}$ where x_l is the word w_d in the document $\mathbf{x}$. First we compute the entropy using P_A and P_B.

$$H_A(\mathbf{x}) = \sum_l p_l log^n(p_l)$$

$$H_B(\mathbf{x}) = \sum_l p_l log^n(p_l)$$

Then we obtain the $max(H_i)$ and decide the language to which the new document belongs. Figure 3 shows a flow diagram that summarizes the process for LI.

The computational complexity for training the model in our approach is given by $O(c|D|)$, where c is the number of clases, and $D = \{D_1, D_2, \ldots, D_k\}$ denote

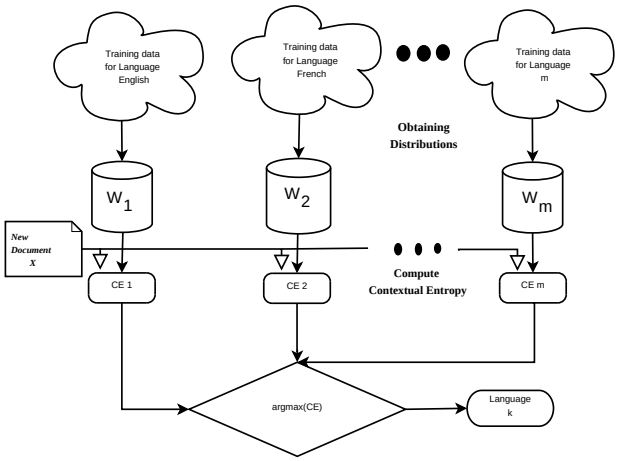

Fig. 3. The CE of document $\mathbf{x}$ is calculated for each distribution W_i or information sources in the system. The class k with highest entropy coincides with to the document $\mathbf{x}$ class.

a training set of k documents where each document has been assigned a label from the classes label set. To classify a new document $\mathbf{x}$, we use only the words of the document $\mathbf{x}$ to compute the entropy in each class c, so the classifcation complexity is $O(cL(x))$, with $L(x)$ the length of the document to be classified.

4 Experimental Results

4.1 Data Sets

The dataset provided by [10], consist of 9066 tweets for 6 languages of at most 140 bytes. The six languages are English, French, German, Spanish, Dutch and Italian. These are languages we have sufficient knowledge of to identify. Note that Spanish, French and Italian presents a challenge as these languages contain a lot of similar word extensions.

4.2 Experiments

In order to compare the accuracy of the methods reported in [10], we train a model with one part of the data called train. We vary the size of the train data to establish the influence of the corpus' size. We use 5%, 10%, 25% and 50% of the entire dataset stratified per language and sampled uniformly.

In all cases we compare the mean accuracy of the CE for $n = 2$ and $n = 5$. The mean accuracies are obtained by taking the mean of 50 different repeated random sub-sampling validation experiment runs.

The main results are presented in table 1 which shows, averaged over 50 experiments runs, the accuracies and standard deviations for CE using $n = 2$ and $n = 5$, LIGA and N-gram reported by [10].

Table 1. Accuracy averaged over 50 runs.

Train Docs.	LIGA	N-gram	CE n=2	CE n=5
5%	94.9 ± 0.8	87.5 ± 1.5	97.1 ± 2.5	97.3 ±2.0
10%	96.4 ± 0.5	90.6 ± 1.0	97.9 ± 1.3	98.2 ±0.5
25%	97.3 ± 0.5	92.5 ± 0.9	98.6 ± 0.9	99.0 ±0.4
50%	97.5 ± 0.5	93.1 ± 0.8	98.9 ± 0.5	99.2 ±0.1

To show the performance of the classifier respect to n, another experiment was carried out using 50% of the documents for training and the rest to test the classifier at different values of n in the equation for entropy contextual. Table 2 and figure 4 present the averaged accurate of this experiment.

Table 2. Accuracy averaged for CE over 20 runs for 50% of documents for trainig at diferent values of n.

n value	Accuracy	Variance
2	98.92436	0.53479
3	99.09331	0.26163
4	99.11911	0.19739
5	99.22973	0.14092
6	99.24737	0.15320
7	99.29478	0.17751
8	99.25101	0.18417
9	99.09037	0.41025
10	99.11566	0.45383
11	99.15769	0.34583
12	99.11285	0.43935
13	99.02434	0.57851
14	98.75165	0.96047
15	98.44259	1.65902

In figure 4 we can see the behavior of CE classifier to different values of n, that as in [4] the best results are obtaining on the first values of n.

5 Conclusions and future work

We presented a new approach for language identification using contextual entropy. Our empirical analysis shows that CE is better than state of the art tech-

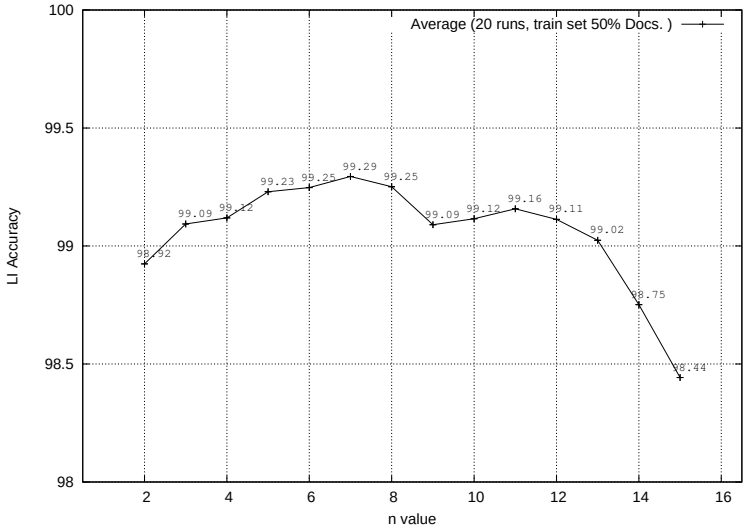

Fig. 4. Average accurate at some values of n to the CE classifier.

niques used in LI. Even more interesting is the change of perspective, seeking the overlap in the distribution that considers each word equally probable.

The computational complexity for LI of a new document using CE is $O(cL(x))$, where c is the number of classes, and $L(x)$ is the length of the document to be classified.

Our preprocessing step consists only in obtaining the different words in the training set to generate the distribution of each language and count the words, we don't need another text operation.

We are currently working in filtering the vocabulary to increase the intra-class entropy as well as decreasing the inter-class entropy. We are also working on other applications of text categorization that may benefit from our method, such as sentiment analysis, mood detection and language identification in multilingual documents.

References

1. William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.
2. Marc Damashek et al. Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199):843–848, 1995.
3. Ted Dunning. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University, 1994.
4. Moises Garcia, Hugo Hidalgo, and Edgar Chavez. Contextual entropy and text categorization. In *Web Congress, 2006. LA-Web'06. Fourth Latin American*, pages 147–153. IEEE, 2006.

5. Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew Mackinlay. Reconsidering language identification for written language resources. In *Proceedings of LREC2006*, pages 485–488, 2006.
6. Stephen Johnson. Solving the problem of language recognition. Technical report, University of Leeds, 1993.
7. Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language identification based on string kernels. In *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*, volume 2, pages 926–929. IEEE, 2005.
8. Marcin J. Schroeder. An alternative to entropy in the measurement of information. *Entropy*, 6:388–412, 2004.
9. Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948.
10. Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In *Proceedings of the Twentieth Belgian Dutch Conference on Machine Learning (Benelearn 2011)*, pages 27–34, 2011.
11. NICTA VRL. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25, 2014.
12. Fei Xia, William D. Lewis, and Hoifung Poon. Language id in the context of harvesting language data off the web. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, pages 870–878, 2009.